

# Introduction to R

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 3rd, 2025

# Plan for today

- Introduction to R
- Data management
- Loops
- Graphics

# Outline

- 1 Introduction to R
- 2 Importing Data to R
- 3 Description of Data
- 4 Modifying Data
- 5 Loops and Flow Controls in R
- 6 Saving Your Work
- 7 Graphics
  - Histogram
  - Box plot
  - Scatter Plot
  - Line plot

# Introduction to R

- R is a programming *language* and a programming *environment*.
- It is Free! Developed by users under a GNU license.
- Runs on a variety of platforms including Windows, Unix and MacOS.  
You can even get it for Android.
- Allows for fast implementation of new methods by user demand through packages.
- R has state-of-the-art graphics capabilities.

# Advantages of R

Frank Harrel (my highlighting):

- "One point that hasn't been made very explicitly is one of the greatest advantages of R:

**Getting your work done better and in less time.**

Hundreds of companies hire a multitude of SAS programmers to write code in an archaic language, the SAS macro language. I believe there is a real cost savings from R because of its value as a **data analysis, data manipulation, and graphics** environment. Instead of programming using an indirect syntax manipulation environment (SAS macros), in R you can program in a dynamic data-sensitive framework".

That was more than 10 years ago. Things have progressed since...

# Base R

- Base R and most R packages are available for download at the Comprehensive R Archive Network (CRAN).
- <http://www.cran.r-project.org>
- Base R includes basic data management, analysis and graphics tools.
- For non-specialized tasks, Base R is all you need.
- Specialized tasks may be handled by *packages*.
- We will download, install and use many packages during the course.
- Packages are not all very well-documented (depends on the contributor).
- Want to be sure about what your program does?
  - Use well-established packages only;
  - or write your own code.

# RStudio

- You can work directly in R.
- Many prefer another front end (GUI, Graphical User Interface).
- We will use RStudio.
- Download from <http://www.posit.co/>

# RStudio

- The GUI **RStudio** has 4 windows.
- One for writing the commands (the "script").
  - Use script for reproducibility.
- One for results and interactive use.
- One for plots, help and packages.
- One showing which objects are resident in the **R** memory.

# RStudio

The screenshot shows the RStudio interface with the following components:

- Code Editor:** Displays R code related to a quiz question about probability calculations involving log(2), log(5), and log(6).
- Spatial Data Viewer:** Shows a list of spatial datasets in the environment, including coastlines, countries, Europe, trapezoids, and various types of polygons.
- Console:** Displays a series of blank lines, indicating the user has cleared the console or is in a state where no output is visible.
- Environment View:** Lists the objects in the global environment, including the spatial datasets and their respective class and sizes.
- Help Sidebar:** Provides links to R Resources, Manuals, Reference, Packages, and Miscellaneous Material, along with links to R documentation and support forums.
- Taskbar:** Shows the Windows taskbar with multiple open applications like Internet Explorer, FileMaker Pro, and Microsoft Word.

# R as a calculator

```
2+2
```

```
[1] 4
```

```
(2*5)+(12/3)-(2^3)
```

```
[1] 6
```

```
exp(log(1))
```

```
[1] 1
```

```
sqrt(25)
```

```
[1] 5
```

```
log(2*2)
```

```
[1] 1.3863
```

```
log(2)+log(2)
```

```
[1] 1.3863
```

# Writing commands in R

- Commands are separated by either a new line or ;
- R is case sensitive: id is a different name than ID.
- The character # at the beginning of a line shows that the text in this line is a comment. I.e. the text is not executed.
- Help can be found on the internet or in R by writing ? followed by the function you want to help about:

```
?plot
```

- or, in RStudio, highlight the expression and press F1.

# Objects in R

- Both data and output from analyses are stored as **objects** (if stored);
- Some times, output is just displayed on the screen, and you need to *assign* the object to an identifier to keep it (see below).
- In fact, everything in the R memory is stored in **objects**.
- An object could be a vector, a matrix or a data frame.
- Values are assigned to objects using the assignment operator `<-`
- The operator `=` also works, but it is **not recommended** as it may confuse assignments with default values and logical expressions.
- We can see the objects of the current R session memory in **RStudio**, or by using the function `ls()`

```
a <- 2+5
```

```
A <- 10
```

```
ls()
```

```
[1] "a" "A"
```

# Generating a sequence

- Specify the first and last values separated by a colon.
- Otherwise use `seq()`

```
0:10
```

```
[1] 0 1 2 3 4 5 6 7 8 9 10
```

```
15:5
```

```
[1] 15 14 13 12 11 10 9 8 7 6 5
```

```
seq(from = 0, to = 1.2, by = 0.1)
```

```
[1] 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0 1.1 1.2
```

```
x <- seq(from = 0, to = 1.5, length = 11); x
```

```
[1] 0.00 0.15 0.30 0.45 0.60 0.75 0.90 1.05 1.20 1.35 1.50
```

# Generating repeats using rep()

```
rep(8, 5)
```

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

```
rep(1:4, each = 2)
```

```
[1] 1 1 2 2 3 3 4 4
```

```
rep(1:4, each = 2, times = 3)
```

```
[1] 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4 1 1 2 2 3 3 4 4
```

# Functions in R

We assign a simple function to the identifier f:

```
>f<-function(x){x^2}; f(2)  
[1] 4
```

A function of two variables:

```
>f<-function(x,pow){x^pow}; f(2,2)  
[1] 4
```

A function with a default value:

```
>f<-function(x,pow=2){x^pow}; f(2,2); f(2);f(2,3)  
[1] 4  
[1] 4  
[1] 8
```

# Functions in R

We have already used many functions with and without default values:

- `"+"(2,2)`
- `sqrt(25)`
- `log(2)`
- `ls()`
- `":"(0,10)`
- `seq(from=0.1,to=1.2,by=0.1)`
- `rep(1:4,each=2,time=3)`

Many applications in R are built up as functions. You can see default arguments in the help files. Example: `log`.

# Data structures in R: Singles

- Logical, e.g:

```
> TRUE  
[1] TRUE  
  
> 1==2  
[1] FALSE
```

- Single numbers, e.g:

```
> 1  
[1] 1  
  
> 1.2  
[1] 1.2
```

- Character, e.g:

```
> "5"  
[1] "5"  
  
> "abc"  
[1] "abc"
```

# Data structures in R: Vectors

Constructed via the concatenate function `c()`.

- Vector of numbers, e.g:

```
> c(1,1.2,pi,exp(1))  
[1] 1.000000 1.200000 3.141593 2.718282
```

- We can have vectors of other things too, e.g:

```
> c(TRUE,1==2)  
[1] TRUE FALSE  
> c("a","ab","abc")  
[1] "a" "ab" "abc"
```

- But not combinations, e.g:

```
> c("a",5,1==2)  
[1] "a" "5" "FALSE"
```

Note that R just turned everything into characters!

# Data structures in R: Matrices

- Columns of same type and same length:

```
> matrix(c(1,2,3,4,5,6)+pi,nrow=2)
[,1] [,2] [,3]
[1,] 4.141593 6.141593 8.141593
[2,] 5.141593 7.141593 9.141593
```

```
> matrix(c(1,2,3,4,5,6)+pi,nrow=2)<6
[,1] [,2] [,3]
[1,] TRUE FALSE FALSE
[2,] TRUE FALSE FALSE
```

# Data structures in R: Data frames

- Same length of columns but different types; spread-sheet data.
- Created from reading in data from external files;
- or by using the function `data.frame()` on a set of vectors.

```
> data.frame(treatment=c("active","active","placebo"),
+ bp=c(80,85,90))
  treatment   bp
1    active    80
2    active    85
3 placebo    90
```

- Compare to a matrix created with the `cbind()` command:

```
> cbind(treatment=c("active","active","placebo"),bp=c(80,85,90))
  treatment   bp
[1,] "active" "80"
[2,] "active" "85"
[3,] "placebo" "90"
```

# Data structures in R: Lists

- Different length of columns and different types.
- Most general object type.

```
> list(a=1,b="abc",c=c(1,2,3),d=list(e=matrix(1:4,2), f=function(x){x^2}))  
$a  
[1] 1  
$b  
[1] "abc"  
$c  
[1] 1 2 3  
$d  
$d$e  
     [,1] [,2]  
[1,]    1    3  
[2,]    2    4  
$d$f  
function (x)  
{  
  x^2  
}
```

- The objects returned from many of the built-in functions in R are fairly complicated lists.

# Importing Data to R

- The easiest is to use data saved as text files.
- Usually values in text files are separated, or delimited, by tabs or commas.
- First tell R where you want to find your data using the command `setwd()`.
- Check that all went according to plan with `getwd()`.

```
setwd("C:/Users/ANST/Teaching/02935 January 2025")
getwd()
```

```
[1] "C:/Users/ANST/Teaching/02935 January 2025"
```

# Importing Data to R

- The function `read.table()` can be used to read data saved as text.
- Wrappers: `read.csv()`, `read.csv2()` and `read.delim()`.
- Notice the option `sep = .`.
- We are assigning the loaded data to objects.
- If you have an Excel sheet, then save as text.

```
Births.tab <- read.table("Data/Births.txt", header = TRUE, sep = "\t")  
  
# SHORT FORM TO READ TAB SEPARATED  
Births.delim <- read.delim("Data/Births.txt")  
  
# SHORT FORM TO READ ; SEPARATED  
Births.csv2 <- read.csv2("Data/Births.csv")  
  
# SHORT FORM TO READ , SEPARATED  
Births.csv1 <- read.csv("Data/Births.csv1")
```

# Importing Data using RStudio

- In the Objects Window, click "Import Dataset"

The screenshot shows a RStudio interface with several windows open:

- Environment**: Shows variables: L0 (500 obs. of 8 variables), a (500 obs. of 9 variables), x (500 obs. of 9 variables), y (500 obs. of 9 variables), and Functions (Function(x, pos = 2)).
- Data View**: Displays a data frame named "abc" with 11 rows and 7 columns. The columns are: gender, milkplus, weanstart, height, weight, weaning, age, gender, and id.
- Code Editor**: Shows R code for reading data from "births.txt" and "SPSS.sav", creating a matrix "x", and defining a function "f".
- Plots**: A histogram of variable "x" is visible.
- Session View**: Shows the command history, including file paths like "C:/.../Jan2010/Introdat" and "C:/.../Jan2010/Introdat/abc.csv".
- Help**: A help window for the "abc" dataset is open.
- File Explorer**: Shows the project structure with files like "abc.R", "abc.Rproj", "abc.csv", "abc.RData", and "abc.Rmd".

# Importing Data using RStudio

- In the Objects Window, click "Import Dataset"

The screenshot shows the RStudio interface with the following details:

- File Menu:** File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Project Area:** my\_statisticapp, Minitab, weight.height, dapt.Rnw, Temp.Rnw, ImportDataset.R, cdc.R.
- Environment Pane:** Shows several datasets:
  - birthro.csv: 500 obs. of 9 variables
  - birtrho.csv2: 500 obs. of 9 variables
  - birtrho.delim: 500 obs. of 9 variables
  - birtrho.tab: 500 obs. of 9 variables
  - cdc: 207 obs. of 9 variables
- Import Dataset Dialog Box:**
  - Name:** birth
  - Input File:** birtrho.csv
  - Encoding:** Automatic
  - Heading:** Yes (radio button selected)
  - Row names:** Automatic
  - Separator:** Comma
  - Decimal:** Commas
  - Quote:** Double-quote (")
  - Comment:** None
  - Data Frame:** Shows the first few rows of the data frame.
  - Import, Cancel:** Buttons at the bottom right.
- Code Editor (Terminal):**

```
> #AMNT/Undervejning/Kurser/phd kurser i basal statistik Jan2019/Introdag
> #verden er en dataset med 207 observationer og 9 variable
> treatment_bp
[1] "active"
[2] "control"
[3] "placebo"
[4] "rest"
[5] "treatments"
[6] "treatments"
[7] "treatments"
[8] "treatments"
[9] "treatments"
[10] "treatments"
[11] "treatments"
[12] "treatments"
[13] "treatments"
[14] "treatments"
[15] "treatments"
[16] "treatments"
[17] "treatments"
[18] "treatments"
[19] "treatments"
[20] "treatments"
[21] "treatments"
[22] "treatments"
[23] "treatments"
[24] "treatments"
[25] "treatments"
[26] "treatments"
[27] "treatments"
[28] "treatments"
[29] "treatments"
[30] "treatments"
[31] "treatments"
[32] "treatments"
[33] "treatments"
[34] "treatments"
[35] "treatments"
[36] "treatments"
[37] "treatments"
[38] "treatments"
[39] "treatments"
[40] "treatments"
[41] "treatments"
[42] "treatments"
[43] "treatments"
[44] "treatments"
[45] "treatments"
[46] "treatments"
[47] "treatments"
[48] "treatments"
[49] "treatments"
[50] "treatments"
[51] "treatments"
[52] "treatments"
[53] "treatments"
[54] "treatments"
[55] "treatments"
[56] "treatments"
[57] "treatments"
[58] "treatments"
[59] "treatments"
[60] "treatments"
[61] "treatments"
[62] "treatments"
[63] "treatments"
[64] "treatments"
[65] "treatments"
[66] "treatments"
[67] "treatments"
[68] "treatments"
[69] "treatments"
[70] "treatments"
[71] "treatments"
[72] "treatments"
[73] "treatments"
[74] "treatments"
[75] "treatments"
[76] "treatments"
[77] "treatments"
[78] "treatments"
[79] "treatments"
[80] "treatments"
[81] "treatments"
[82] "treatments"
[83] "treatments"
[84] "treatments"
[85] "treatments"
[86] "treatments"
[87] "treatments"
[88] "treatments"
[89] "treatments"
[90] "treatments"
[91] "treatments"
[92] "treatments"
[93] "treatments"
[94] "treatments"
[95] "treatments"
[96] "treatments"
[97] "treatments"
[98] "treatments"
[99] "treatments"
[100] "treatments"
[101] "treatments"
[102] "treatments"
[103] "treatments"
[104] "treatments"
[105] "treatments"
[106] "treatments"
[107] "treatments"
[108] "treatments"
[109] "treatments"
[110] "treatments"
[111] "treatments"
[112] "treatments"
[113] "treatments"
[114] "treatments"
[115] "treatments"
[116] "treatments"
[117] "treatments"
[118] "treatments"
[119] "treatments"
[120] "treatments"
[121] "treatments"
[122] "treatments"
[123] "treatments"
[124] "treatments"
[125] "treatments"
[126] "treatments"
[127] "treatments"
[128] "treatments"
[129] "treatments"
[130] "treatments"
[131] "treatments"
[132] "treatments"
[133] "treatments"
[134] "treatments"
[135] "treatments"
[136] "treatments"
[137] "treatments"
[138] "treatments"
[139] "treatments"
[140] "treatments"
[141] "treatments"
[142] "treatments"
[143] "treatments"
[144] "treatments"
[145] "treatments"
[146] "treatments"
[147] "treatments"
[148] "treatments"
[149] "treatments"
[150] "treatments"
[151] "treatments"
[152] "treatments"
[153] "treatments"
[154] "treatments"
[155] "treatments"
[156] "treatments"
[157] "treatments"
[158] "treatments"
[159] "treatments"
[160] "treatments"
[161] "treatments"
[162] "treatments"
[163] "treatments"
[164] "treatments"
[165] "treatments"
[166] "treatments"
[167] "treatments"
[168] "treatments"
[169] "treatments"
[170] "treatments"
[171] "treatments"
[172] "treatments"
[173] "treatments"
[174] "treatments"
[175] "treatments"
[176] "treatments"
[177] "treatments"
[178] "treatments"
[179] "treatments"
[180] "treatments"
[181] "treatments"
[182] "treatments"
[183] "treatments"
[184] "treatments"
[185] "treatments"
[186] "treatments"
[187] "treatments"
[188] "treatments"
[189] "treatments"
[190] "treatments"
[191] "treatments"
[192] "treatments"
[193] "treatments"
[194] "treatments"
[195] "treatments"
[196] "treatments"
[197] "treatments"
[198] "treatments"
[199] "treatments"
[200] "treatments"
[201] "treatments"
[202] "treatments"
[203] "treatments"
[204] "treatments"
[205] "treatments"
[206] "treatments"
[207] "treatments"

```

# Importing Data from other stats programs

- We can read data from a series of other statistical software packages using the package `foreign`. A similar package is `Hmisc`, but we will stick to `foreign`.

```
# INSTALL AN EXTRA PACKAGE
install.packages("foreign")

# ACTIVATE THE PACKAGE
library("foreign")

SPSS_Data <- read.spss("Data/SPSS_Data.sav", to.data.frame = TRUE)

Stata_Data <- read.dta("Data/string.dta")
```

- For SAS data not in XPORT format, use the `sas7bdat` package.

# Look at Your Data

There are several ways to look at the data (or parts of the data).

*# FIRST FEW OBSERVATIONS*

```
head(Births.tab)
```

	id	bweight	lowbw	gestwks	preterm	matage	hyp	sex	sexalph
1	1	2974	0	38.52	0	34	0	2	female
2	2	3270	0	NA	NA	30	0	1	male
3	3	2620	0	38.15	0	35	0	2	female
4	4	3751	0	39.80	0	31	0	1	male
5	5	3200	0	38.89	0	33	1	1	male
6	6	3673	0	40.97	0	33	0	2	female

# Look at Your Data

# LAST FEW OBSERVATIONS

```
tail(Births.tab)
```

		id	bweight	lowbw	gestwks	preterm	matage	hyp	sex	sexalph
495	495	2968	0	41.01		0	34	0	1	male
496	496	2852	0	38.45		0	28	0	2	female
497	497	3187	0	38.03		0	38	1	1	male
498	498	3054	0	38.50		0	26	0	2	female
499	499	3178	0	39.92		0	31	0	2	female
500	500	2918	0	37.97		0	31	0	1	male

# VARIABLE NAMES

```
names(Births.tab)
```

```
[1] "id"          "bweight"     "lowbw"       "gestwks"    "preterm"    "matage"     "hyp"  
[8] "sex"         "sexalph"
```

# VIEW THE DATA IN A NEW WINDOW; OFTEN THE HEAD WILL DO THOUGH

```
View(Births.tab)
```

# Missing values

- In real life examples it is very common to have missing values.
- In R missing values are coded as **NA** (not available).
- In your Excel file leave missing values blank, do not set them to 99 or 999.

	id	bweight	lowbw	gestwks	preterm	matage	hyp	sex	sexalph
1	1	2974	0	38.52		0	34	0	2 female
2	2	3270	0	NA		NA	30	0	1 male

# Accessing Observations

- Data are (usually) stored in a data frame object.
- Observations are the rows.
- Variables, either numerical or categorical, are the columns.
- We can access individual rows, columns and cells in the data frame.
- For this, we use the bracket operator: `object[row, column]`.

# Accessing Observations

```
# A SINGLE CELL
Births.tab[345, 4]

[1] 38.55

# LEAVING OUT A COLUMN NUMBER INDICATES THAT ALL COLUMNS
# ARE CHOSEN. HERE ALL COLUMNS IN ROW 224
Births.tab[224 , ]

  id bweight lowbw gestwks preterm matage hyp sex sexalph
224 224     3216      0   39.94      0     38    1    1    male
```

# Accessing Observations

*# LEAVING OUT A ROW NUMBER INDICATES THAT ALL ROWS ARE CHOSEN*

# HERE ALL ROWS IN COLUMN 5

Births.tab[5]

# Accessing Observations

```
# USE RANGES, ROWS 15 TO 18 COLUMNS 1 TO 4  
Births.tab[15:18, 1:4]
```

		id	bweight	lowbw	gestwks
15	15	3662	0	39.23	
16	16	3035	0	38.96	
17	17	3351	0	39.35	
18	18	3804	0	38.99	

# Accessing Observations

Variables can be accessed directly using their name, either with the \$ operator (`object$variable`), the name (`object[, "variable"]`), or the column number (`object[, k]`).

```
# GET THE BIRTH WEIGHT FOR CHILD 26 TO 36
Births.tab$bweight[26:36]
[1] 3585 3798 3164 3739 1780 4022 3942 2887 2391 3911 3509

Births.tab[26:36, "bweight"]
[1] 3585 3798 3164 3739 1780 4022 3942 2887 2391 3911 3509

Births.tab[26:36,2]
[1] 3585 3798 3164 3739 1780 4022 3942 2887 2391 3911 3509
```

## Subsetting using the `c()` function

- The concatenate function `c()` concatenates the arguments into a vector. It can be used for many things; one is to access non-sequential rows and columns from a data frame.

```
# GET COLUMNS 2, 5, 7, 8, 9 FOR ROW 33  
Births.tab[33, c(2, 5, 7:9)]
```

```
bweight preterm hyp sex sexalph  
33     2887      0   0    1    male
```

```
# GET bweight, preterm and sexalph FOR ROW 71  
Births.tab[71, c("bweight", "preterm", "sexalph")]
```

```
bweight preterm sexalph  
71     3189      0    male
```

# Variable Names

If we want to change the variable names we can use `names()`.

```
# NEW VARIABLE NAMES
names(Births.tab) <- c("ID", "Bweight", "LowBW", "GestWks",
                        "Preterm", "Matage", "Hyp", "Sex", "Sexalph")

# JUST THE FIRST VARIABLE NAME
names(Births.tab)[1] <- c("ID_new")

# CHECK HOW IT WENT
names(Births.tab)

[1] "ID_new"   "Bweight"  "LowBW"    "GestWks"  "Preterm"  "Matage"   "Hyp"
[8] "Sex"       "Sexalph"

# RESETTING
names(Births.tab)[1] <- c("ID")
```

# Saving/Exporting data

- We can save the data to a textfile using either `write.table()` for a tab separated file, or `write.csv()`/`write.csv2()` for a comma/semicolon separated file (with `"."`and `","`as decimal point, respectively).

```
write.table(Births.tab, file = "Birth_new.txt",
            sep = "\t", na = ".", row.names= FALSE)

write.csv2(Births.tab, file = "Birth_new.csv")
```

## Exercise: Protein Consumption

- Open **RStudio** and set the working directory to where you want to keep the data for the course today.
- Import the data **Protein.xlsx** to R.
- Look at the data.
- What is the protein consumption in Denmark?
- Look at the protein consumption from red meat alone.
- Look at the protein consumption in Denmark, Norway, Sweden from red meat, white meat and eggs.
- Rename the variables “RedMeat” and “WhiteMeat” to “Red” and “White”.
- Save the new version of the protein data as a tab delimited text file “**Protein2.txt**”.

# Description of Data

We are still looking at the data set with birth weights for 500 children. Using the function `str()` we can see a description of what our data frame contains (the structure).

```
str(Births.tab)
```

```
'data.frame': 500 obs. of  9 variables:  
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...  
 $ bweight: int  2974 3270 2620 3751 3200 3673 3628 3773 3960 3405 ...  
 $ lowbw   : int  0 0 0 0 0 0 0 0 0 0 ...  
 $ gestwks: num  38.5 NA 38.2 39.8 38.9 ...  
 $ preterm: int  0 NA 0 0 0 0 0 0 0 0 ...  
 $ matage  : int  34 30 35 31 33 33 29 37 36 39 ...  
 $ hyp     : int  0 0 0 0 1 0 0 0 0 0 ...  
 $ sex     : int  2 1 2 1 1 2 2 1 2 1 ...  
 $ sexalph: chr "female" "male" "female" "male" ...
```

# Description of Data: Birth weights

- `str()` supplies all information about an object. **Much is redundant**, do not use `str()` for report writing but for your own overview only.
- We see that we have a data frame with 500 observations and 9 variables.
- Some are integers but “gestwks” is numeric.
- The variable “sexalph” is a character variable.
- Note that “sexalph” and “sex” describes the same thing. But R does not interpret the character values as group labels.
- Factor: Grouping with informative labels. We can convert “sexalph” to a factor using `as.factor()`.

# Description of Data: Birth weights

```
# TELL R THAT sexalph IS A FACTOR
Births.tab$sexalph <- as.factor(Births.tab$sexalph)
levels(Births.tab$sexalph)

[1] "female" "male"

str(Births.tab)

'data.frame': 500 obs. of 9 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ bweight: int  2974 3270 2620 3751 3200 3673 3628 3773 3960 3405 ...
 $ lowbw   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ gestwks: num  38.5 NA 38.2 39.8 38.9 ...
 $ preterm: int  0 NA 0 0 0 0 0 0 0 0 ...
 $ matage  : int  34 30 35 31 33 33 29 37 36 39 ...
 $ hyp     : int  0 0 0 0 1 0 0 0 0 0 ...
 $ sex     : int  2 1 2 1 1 2 2 1 2 1 ...
 $ sexalph: Factor w/ 2 levels "female","male": 1 2 1 2 2 1 1 2 1 2 ...
```

# Description of Data: Birth weights

```
# TELL R THAT sex IS A FACTOR WITH SPECIFIC LEVELS
Births.tab$sex <- factor(Births.tab$sex, labels =c("Male","Female"))
levels(Births.tab$sex)

[1] "Male" "Female"

str(Births.tab)

'data.frame': 500 obs. of  9 variables:
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ bweight: int  2974 3270 2620 3751 3200 3673 3628 3773 3960 3405 ...
 $ lowbw   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ gestwks: num  38.5 NA 38.2 39.8 38.9 ...
 $ preterm: int  0 NA 0 0 0 0 0 0 0 0 ...
 $ matage  : int  34 30 35 31 33 33 29 37 36 39 ...
 $ hyp     : int  0 0 0 0 1 0 0 0 0 0 ...
 $ sex     : Factor w/ 2 levels "Male","Female": 2 1 2 1 1 2 2 1 2 1 ...
 $ sexalph: Factor w/ 2 levels "female","male": 1 2 1 2 2 1 1 2 1 2 ...
```

# Descriptive Statistics

- There are some simple functions for summary statistics in R.
- Very common extractor functions are `mean()`, `sd()`, `median()`, `max()` and `min()`.

```
mean(Births.tab$bweight)
```

```
[1] 3136.9
```

```
sd(Births.tab$bweight)
```

```
[1] 637.45
```

```
median(Births.tab$bweight)
```

```
[1] 3188.5
```

```
max(Births.tab$bweight)
```

```
[1] 4553
```

```
min(Births.tab[, 2])
```

```
[1] 628
```

# The Summary Function

- The function `summary()` can be used with many objects in R.
- When used on a data frame we get all the main summary statistics.

```
# SUMMARY OF THE DATA FRAME
```

```
summary(Births.tab)
```

	<code>id</code>	<code>bweight</code>	<code>lowbw</code>	<code>gestwks</code>
Min. :	1	Min. : 628	Min. :0.00	Min. :24.7
1st Qu.:	126	1st Qu.:2862	1st Qu.:0.00	1st Qu.:37.9
Median :	250	Median :3188	Median :0.00	Median :39.1
Mean :	250	Mean :3137	Mean :0.12	Mean :38.7
3rd Qu.:	375	3rd Qu.:3551	3rd Qu.:0.00	3rd Qu.:40.1
Max. :	500	Max. :4553	Max. :1.00	Max. :43.2
			NA's :10	
	<code>preterm</code>	<code>matage</code>	<code>hyp</code>	<code>sex</code>
Min. :	0.000	Min. :23	Min. :0.000	1:264 female:236
1st Qu.:	0.000	1st Qu.:31	1st Qu.:0.000	2:236 male :264
Median :	0.000	Median :34	Median :0.000	
Mean :	0.129	Mean :34	Mean :0.144	
3rd Qu.:	0.000	3rd Qu.:37	3rd Qu.:0.000	
Max. :	1.000	Max. :43	Max. :1.000	
NA's :	10			
	<code>sexalph</code>			

# Summaries

- We may only want summaries for some of the data, e.g. babies with birth weight < 2900g.
- We subset the data and then summarize as before:

```
summary(Births.tab[Births.tab$bweight<2900,])
```

	<b>id</b>	<b>bweight</b>	<b>lowbw</b>	<b>gestwks</b>
Min.	: 3	Min. : 628	Min. :0.000	Min. :24.7
1st Qu.	:146	1st Qu.:2120	1st Qu.:0.000	1st Qu.:35.5
Median	:254	Median :2580	Median :0.000	Median :37.4
Mean	:250	Mean :2355	Mean :0.441	Mean :36.6
3rd Qu.	:359	3rd Qu.:2741	3rd Qu.:1.000	3rd Qu.:38.5
Max.	:496	Max. :2894	Max. :1.000	Max. :41.4
			NA's :2	
	<b>preterm</b>	<b>matage</b>	<b>hyp</b>	<b>sex</b>
Min.	:0.000	Min. :24	Min. :0.000	1:63 female:73
1st Qu.	:0.000	1st Qu.:31	1st Qu.:0.000	2:73 male :63
Median	:0.000	Median :34	Median :0.000	
Mean	:0.403	Mean :34	Mean :0.243	
3rd Qu.	:1.000	3rd Qu.:37	3rd Qu.:0.000	
Max.	:1.000	Max. :43	Max. :1.000	
NA's	:2			

# Group Summaries

- Data may be separated by groups.
- Suppose that we want to calculate the mean birth weight for boys and girls (many ways to do this).
- We will use the `tapply()` function to apply the `mean` function to the two levels of “sexalph”.
- `tapply(<variable to summarize>, <variable to group by>, <function to use>)`.

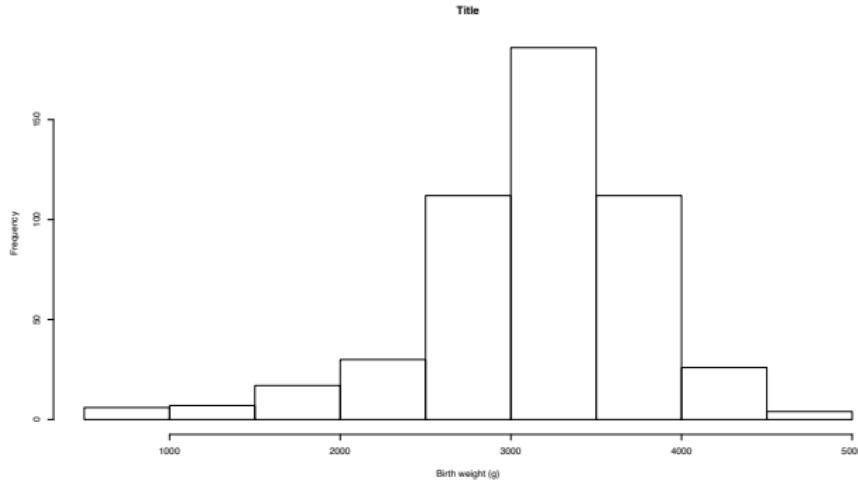
```
# MEAN BIRTH WEIGHT FOR BOYS AND GIRLS
tapply(Births.tab$bweight, Births.tab$sexalph, mean)

female      male
3032.831  3229.902
```

# Histogram

Often it is easier to get an impression of a distribution using plots.  
Histograms are typically used for continuous variables.

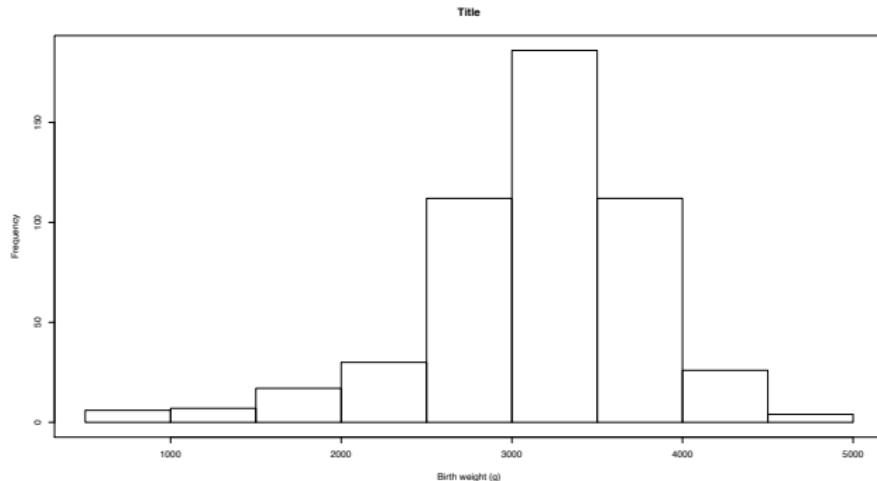
```
hist(Births.tab$bweight, main = "Title", xlab = "Birth weight (g)")
```



# Histogram

Often it is easier to get an impression of a distribution using plots.  
Histograms are typically used for continuous variables. Here with a box on.

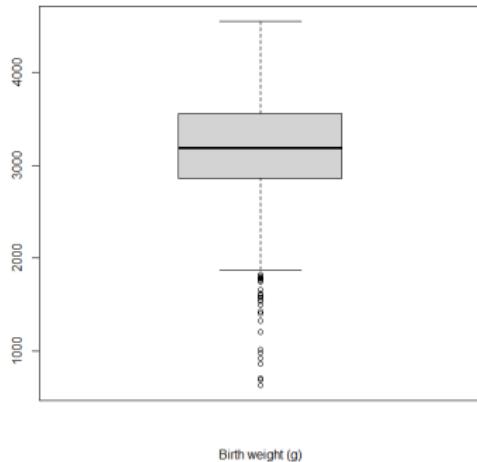
```
hist(Births.tab$bweight, main = "Title",  
      xlab = "Birth weight (g)")  
box()
```



# Boxplot

Boxplots show the median, upper, lower quartiles and potentially extreme values.

```
boxplot(Births.tab$bweight, xlab = "Birth weight (g)")
```



# Exercise: Descriptive Statistics 1

- Import the dataset cdc.csv into R.
- Look at the data.
- Describe the structure of the data.
- Change the variable “smoke100” to a factor.
- Summarize the variables in the dataset.
- Calculate the mean age by general health.
- Draw a histogram of age.
- Draw a boxplot of age.

# Modifying Data

We will concentrate on how to modify and rearrange our data.

- Data can be sorted with the `order` function.
- `order` can sort the Birth.tab data by “sex”, and then by “bweight”.
- The `order` function returns a vector of sorted indices, which we apply to the rows of the unsorted data frame to get a sorted version.

```
Birth_sort <- Births.tab[order(Births.tab$sex, Births.tab$bweight), ]
```

```
head(Birth_sort)
```

		id	bweight	lowbw	gestwks	preterm	matage	hyp	sex	sexalph	
253	253	253	693	1	30.71		1	34	1	1	male
226	226	226	981	1	27.99		1	29	1	1	male
181	181	181	1019	1	28.04		1	31	1	1	male
22	22	22	1203	1	32.80		1	39	0	1	male
312	312	312	1500	1	35.27		1	34	0	1	male
313	313	313	1595	1	30.52		1	33	1	1	male

## Creating new variables and deleting old

New variables can be added to a data frame.

```
# ADD A VARIABLE TO DATA FRAME  
Births.tab$log_bweight <- log(Births.tab$bweight)
```

Columns can be dropped from a data frame (log birth weight is column 10):

```
Births.tab <- Births.tab[ , -10]
```

```
# CREATE A VARIABLE AS A SEPARATE OBJECT  
log_bweight <- log(Births.tab$bweight)
```

Objects can be removed from the R memory (cleaning up):

```
rm(log_bweight)
```

## Grouping the values of a variable using cut

You might want to group a continuous variable e.g. mother's age (matage) into the groups: ]20-30], ]30-35], ]35-40], ]40-45]:

```
Births.tab$agegrp <- cut(Births.tab$matage,  
                           breaks = c(20, 30, 35, 40, 45))  
summary(Births.tab[ , c("matage", "agegrp")])
```

	matage	agegrp
Min.	:23	(20,30]: 99
1st Qu.	:31	(30,35]: 215
Median	:34	(35,40]: 174
Mean	:34	(40,45]: 12
3rd Qu.	:37	
Max.	:43	

# Creating new variables: RowSums

- Often we want to form new variables from other variables.
- For example, we might want to calculate a total score from sub scores.
- We can sum variables using `rowSums`. Related functions are: `rowMeans`, `colSums`, `colMeans`.
- Notice the option `na.rm`.
- If we take a row sum where one of the values is missing then the row sum is set to missing `na.rm= FALSE`.
- If we want to ignore missing values and calculate a sum of the non missing then `na.rm= TRUE`.
- `rowSums`, `rowMeans`, `colSums` and `colMeans` are wrappers of `sapply`, ie. t.ex. `colMeans(x)` is the same as `sapply(x,mean)`. `sapply` can be used with many other functions.

# Creating new variables: RowSums

```
# NEW VARIABLE SCORE SUMMING PRETERM, LOWBW AND HYP  
Births.tab$score <- rowSums(Births.tab[ ,c(3,5,7)], na.rm = FALSE)
```

*# REMOVE MISSING*

```
Births.tab$scoreRM <- rowSums(Births.tab[ ,c(3,5,7)], na.rm = TRUE)
```

```
head(Births.tab)
```

	id	bweight	lowbw	gestwks	preterm	matage	hyp	sex	sexalph	score	scoreRM
1	1	2974	0	38.52	0	34	0	2	female	0	0
2	2	3270	0	NA	NA	30	0	1	male	NA	0
3	3	2620	0	38.15	0	35	0	2	female	0	0
4	4	3751	0	39.80	0	31	0	1	male	0	0
5	5	3200	0	38.89	0	33	1	1	male	1	1
6	6	3673	0	40.97	0	33	0	2	female	0	0

## Exercise

- Import the data cdc.
- Height is in inches, weight and wtdesire are in pounds. Generate new variables in cm and kg (use Google to find conversion factors).
- Make a factor with 4 *approximately* equally sized groups from the weight variable.
- Sort the data by gender and age.
- Calculate the average of the weight and desired weight for each subject.
- Calculate the mean weight and wtdesire (in kg) for each level of "genhlth" with the **by** function.

## Split Data: Subset

- Sometimes we may need to split our data.
- In the Births data we may need to split the data into boys and girls.
- We can use the `subset()` function and assign the new data sets to separate R objects.
- Notice `==` (logical expression). We are not assigning a value to “sex”, but asking whether “sex is equal to 1”.

```
Births.Male <- subset(Births.tab, sex == 1)  
Births.Female <- subset(Births.tab, sex == 2)
```

- Alternatively the bracket operator:

```
Births.Male <- Births.tab[Births.tab$sex == 1,]  
Births.Female <- Births.tab[Births.tab$sex == 2,]
```

# Subset

- Often data sets come with a lot of variables and we only want to use a few.
- Similarly to the bracket operator [ the function `subset()` can be used to select the variables we want.
- Notice the `select` option. This is needed to say that we want a subset of columns (on the previous slide it was rows).
- Notice that we do not need quotes in `select`.

```
# SELECT 3 VARIABLES
Births.new <- subset(Births.tab, select = c(id, bweight, sex))
```

## Aggregating data

- Sometimes we want to make a new dataframe as a summary of the original dataframe on the basis of factor levels.
- Below we want to make a new dataframe with the mean birthweight for combinations of preterm and sex.

```
PreSex <- aggregate(Births.tab$bweight,
                      by = list(Preterm = Births.tab$preterm,
                                Sex = Births.tab$sexalph), mean)
```

PreSex

	Preterm	Sex	x
1	0	female	3191.6
2	1	female	2020.8
3	0	male	3361.2
4	1	male	2321.6

## Add rows: rbind()

- Data are collected for subgroups of subjects and saved in separate objects.
- The separate objects are appended (stacked) to create a single object.
- This will give an error message if the number of columns differs.

```
# APPEND
Births.Both <- rbind(Births.Male, Births.Female)
dim(Births.Both)

[1] 500 11

dim(Births.Male)

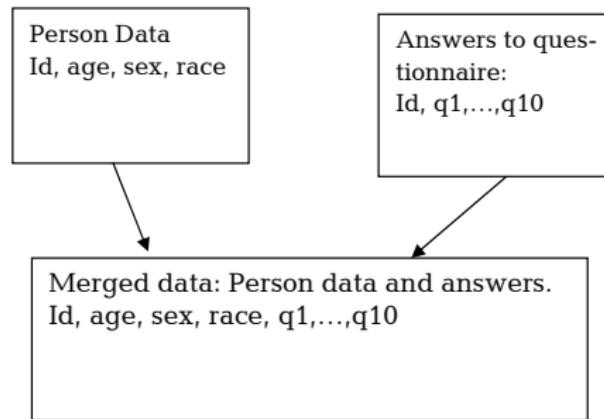
[1] 264 11

dim(Births.Female)

[1] 236 11
```

## Add variables: merge()

Often you have data in several data sets and want to combine the data sets by merging using one or more variables as *key variables*. Adding variables to a master data set.



# Merge

We have two data sets with a key variable "id". One with background information and one set with blood pressure measurements.

```
agesex <- read.delim("agesex.txt")
agesex

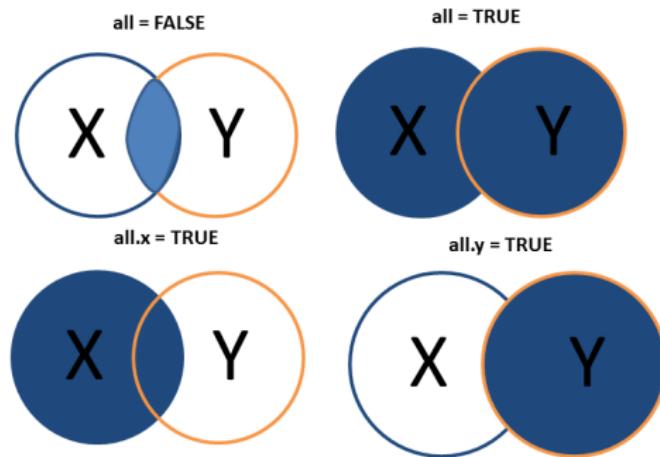
  id age sex
1 99  43   m
2 100 47   f
3 101 NA   f
4 102 67   m
```

```
bp <- read.delim("bp.txt")
bp

  id visit bp
1 100    1 180
2 100    2 160
3 100    3 155
4 101    1 160
5 102    1 120
6 102    2 140
7 103    1 135
```

## 4 Different Merges

- In the `merge` function we will look at 4 of the options.
- We have `merge(x, y, by = "key variable", all = TRUE, <all = FALSE, all.x = TRUE, all.y = FALSE>)`.
- Here `x` and `y` are data frames



# Merging all=FALSE

```
merge_small <- merge(agesex, bp, by = "id", all = FALSE)  
merge_small
```

	id	age	sex	visit	bp
1	100	47	f	1	180
2	100	47	f	2	160
3	100	47	f	3	155
4	101	NA	f	1	160
5	102	67	m	1	120
6	102	67	m	2	140

# Merging all=TRUE

```
merge_large <- merge(agesex, bp, by = "id", all = TRUE)  
merge_large
```

	id	age	sex	visit	bp
1	99	43	m	NA	NA
2	100	47	f	1	180
3	100	47	f	2	160
4	100	47	f	3	155
5	101	NA	f	1	160
6	102	67	m	1	120
7	102	67	m	2	140
8	103	NA	<NA>	1	135

## Merging all.x=TRUE

```
merge_x <- merge(agesex, bp, by = "id", all.x = TRUE)  
merge_x
```

	id	age	sex	visit	bp
1	99	43	m	NA	NA
2	100	47	f	1	180
3	100	47	f	2	160
4	100	47	f	3	155
5	101	NA	f	1	160
6	102	67	m	1	120
7	102	67	m	2	140

# Merging all.y=TRUE

```
merge_y <- merge(agesex, bp, by = "id", all.y = TRUE)  
merge_y
```

	id	age	sex	visit	bp
1	100	47	f	1	180
2	100	47	f	2	160
3	100	47	f	3	155
4	101	NA	f	1	160
5	102	67	m	1	120
6	102	67	m	2	140
7	103	NA	<NA>	1	135

# Counting the Missing Observations: The `is.na()` and `sum()` functions

- Suppose that we want to count the number of missing observations.
- The function `is.na` returns a logical vector that is **TRUE** when a value is missing and **FALSE** otherwise.

```
is.na(merge_y$sex)  
[1] FALSE FALSE FALSE FALSE FALSE FALSE TRUE
```

```
#COUNT MISSING FOR ONE VARIABLE  
sum(is.na(merge_y$sex))
```

```
[1] 1
```

```
#COUNT FOR DATA FRAME  
colSums(is.na(merge_y))
```

id	age	sex	visit	bp
0	2	1	0	0

# Loops in R

- In R, the **for loop** is used perform a repetitive task for each element in a set.
- Example:
  - Given a set of integers 1:3:
  - Let a variable  $i$  run through the set and print  $i + i$ :

```
for(i in 1:3) {  
  cat(i, "+", i, "=", i+i, "\n")  
}
```

- Output:

```
1 + 1 = 2  
2 + 2 = 4  
3 + 3 = 6
```

# Flow Control: if and if else Statements

- if statement:

```
for(i in 1:3){  
  if (i==2) cat("This index is even:", "\n")  
  cat(i, "\n")  
}  
  
1  
This index is even:  
2  
3
```

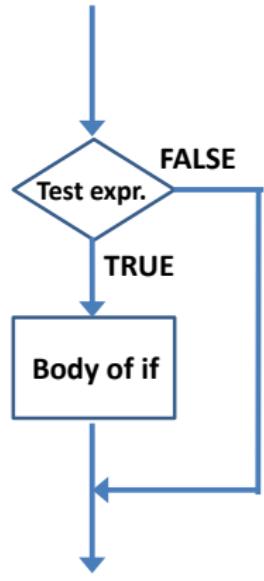
- if else statement:

```
for(i in 1:3){  
  if (i==2) cat("The index is 2", "\n") else  
    cat("The index is not 2", "\n")  
}
```

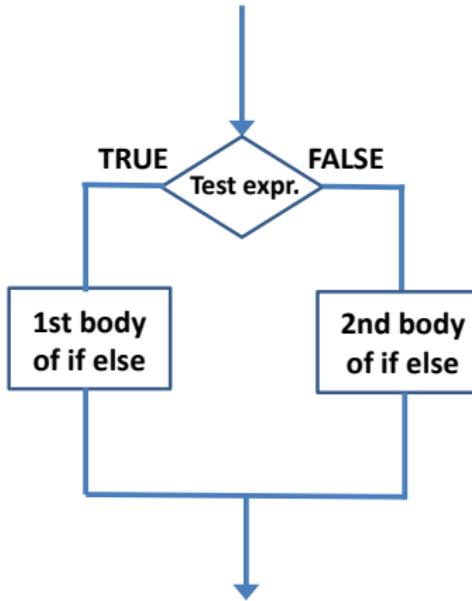
```
The index is not 2  
The index is 2  
The index is not 2
```

# Flow Charts for if and if else Statements

Enter if statement



Enter if else statement



# while and repeat Loops

## The **while** loop:

- `while(condition) expression`

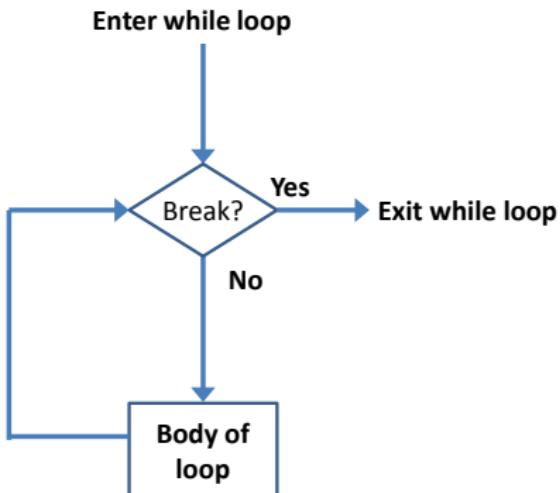
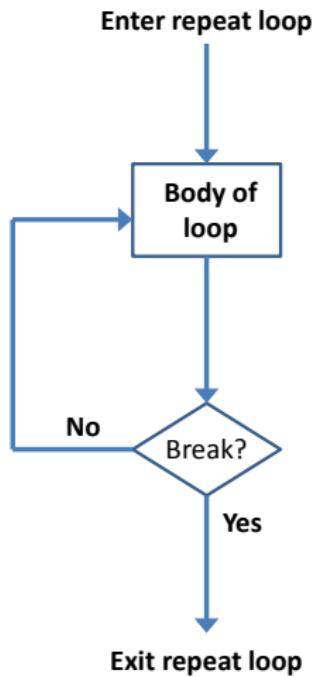
## The **repeat** loop:

- `repeat expr`

The repeat loop has to be exited manually. Flow controls:

- **next**: Halts the current iteration and advances to the next immediately;
- **break**: Exits the loop.

# Flow Charts for the while and repeat loops



# while Loop Example

## Storing of machine parts

```
k<-0 # number of big parts (>2)
y<-abs(rnorm(1000)) # simulated part size
i<-0 # index of parts
# loop:
while(k<3 & i<1000){
  i<-i+1
  temp<-y[i]
  k<-k+(temp>2)
}
i
```

```
[1] 42
```

## repeat Loop Example

### Selecting persons without blue or yellow eyes

```
eye.colors<-c("brown","blue","green","yellow","grey")
eyecolor<-data.frame(personId=1:100,color=
                      sample(eye.colors,100,rep=T))

i<-0
list.of.ids<-numeric(0) # patient ID list
#loop:
repeat {
  i<-i+1
  if(eyecolor$color[i]=="yellow" |
     eyecolor$color[i]=="blue") next
  list.of.ids<-c(list.of.ids,eyecolor$personId[i])
  if(i==100 | length(list.of.ids)==20) break
}
list.of.ids
```

```
[1]  5  6  7  9 10 11 12 14 15 18 19 20 21 22 23 24 25 28 29 30
```

# Saving your work

- Saving your script
- Saving your workspace

Always save your script - do it often if you work in [Rstudio](#).

- Reasons for saving your workspace:
  - Extensive data creations will be there next time you open your workspace.
  - Objects created 'on the fly' (not in your script) will be there.
- Reasons for not saving your workspace:
  - With a well-written script, you can recreate your analysis in seconds, unless you work with huge amounts of data.
  - Edited and saved data where editions have been forgotten may cause havoc on your results.
  - Left-over objects created for various purposes may enter your calculations unintentionally due to the structure of R's search path.

# Saving your work

How to save your work:

- Script: Click on the script and press 'save' in Rstudio and the plain R GUI.
- Workspace: Click on the command prompt and press 'save'. Alternatively, use the `save.image()` function
- Both: Accept when asked after terminating Rstudio or the plain R GUI.

## Exercise

- Data on exercise habits for the subjects in the cdc data set are found in the data `cdc_exer.csv`. Load and describe these data.
- Merge the cdc data set with the `cdc_exer` data set using `id` as the key variable. How many subjects in the study did not have exercise information?
- List the identifiers for the subjects without exercise information.
- Make a new dataframe with average age for combinations of gender and general health.

# Visualizing Data is Important

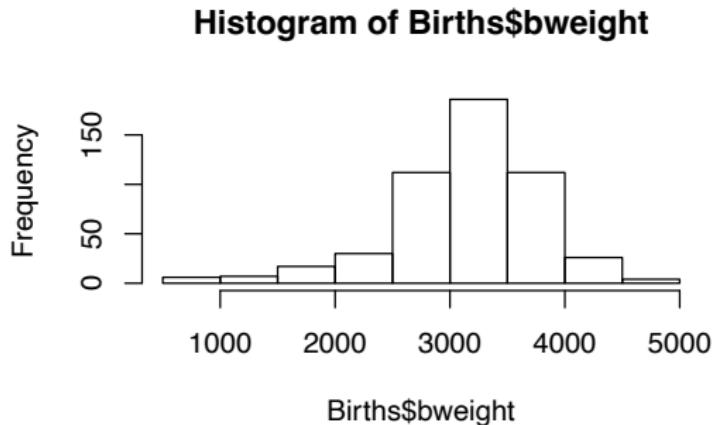
- Whenever we want to analyze data the first thing we should do is to **have a look** at it.
- How are the observations spread out?
- What are the most common values?
- Are there any unusual observations?
- Are there any relationships between variables?

This session will not tell you all about graphics in R but get you going.

# A Basic Histogram

- Common way to examine the distribution of a continuous variable.
- The range of the variable is by default divided into equal-width intervals (bins). Plots the number of observations in each bin (unless you specify otherwise).

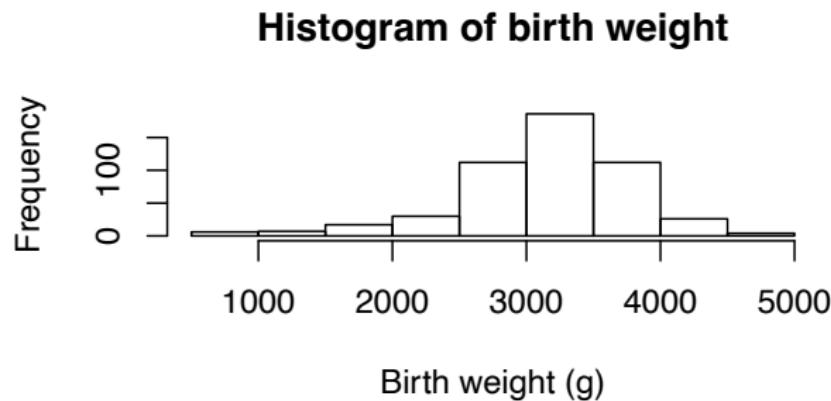
```
hist(Births$bweight)
```



# Histogram with a few options

- Note that R automatically has created axis **labels** and a heading.
- To modify axis labels we set the options **xlab** and **ylab**.
- The heading is set in the option **main**.

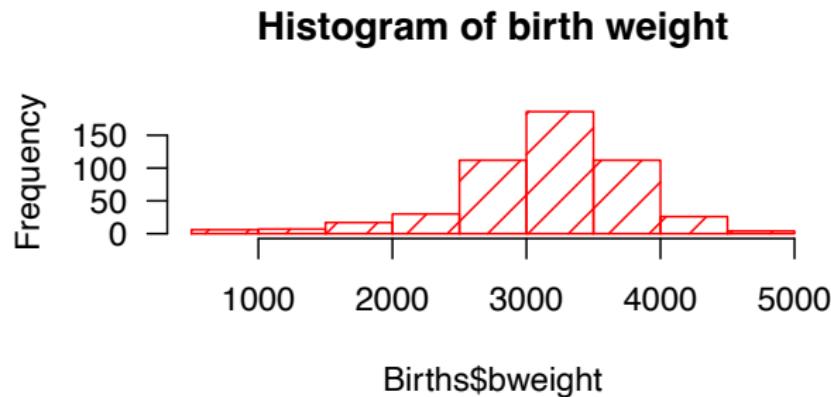
```
hist(Births$bweight, xlab = "Birth weight (g)",  
      main = "Histogram of birth weight")
```



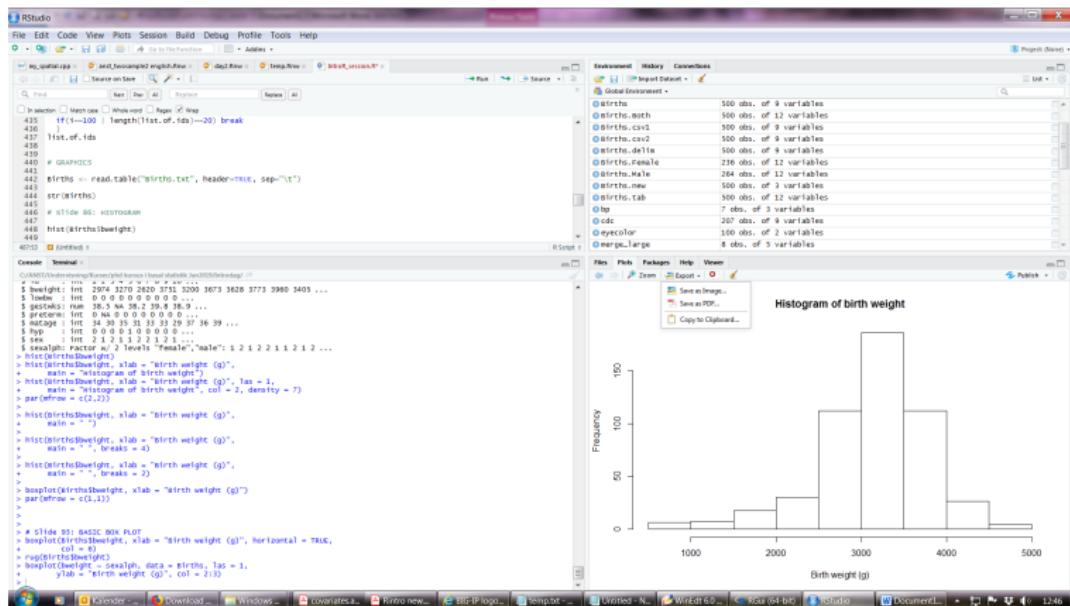
# Histogram with more options

- We could type ?hist to find more options to customize the histogram.
- The available colours are coded as numbers or one can write `col = "red"`
- If we want shading we can try the `density` function.
- The angle of the numbers on the axes is set by the option `las`.

```
hist(Births$bweight, las = 1, main = "Histogram of birth weight",
      col = 2, density = 7)
```



# How to get your plot from RStudio



# Writing to a graphics device

```
pdf("my.histogram.pdf")
hist(Births$bweight, las = 1, main = "Histogram of birth weight",
      col = 2, density = 7)
dev.off()
```

```
svg("my.histogram.svg")
hist(Births$bweight, las = 1, main = "Histogram of birth weight",
      col = 2, density = 7)
dev.off()
```

```
png("my.histogram.png")
hist(Births$bweight, las = 1, main = "Histogram of birth weight",
      col = 2, density = 7)
dev.off()
```

- Options can be specified; see the help files.

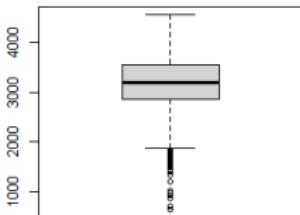
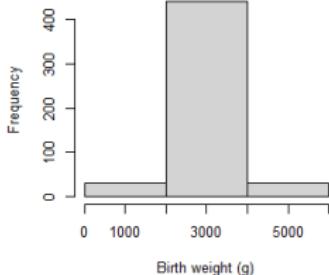
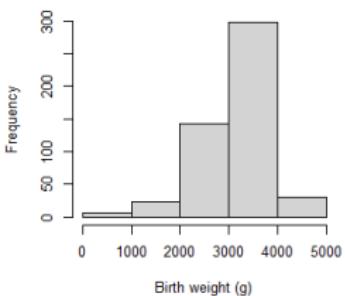
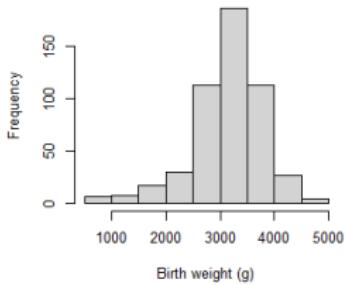
# Exercise

- Load the cdc data set.
- Make a histogram of the weight in kg.
- Add your own title and x-axis label.
- Try different colours and shadings.
- Copy your favorite histogram into a document in e.g. Word.

# A Basic Box Plot

- Box plots show some distributional properties very clearly.
- Box plots show a measure of the location (the median line).
- The spread of the distribution (the length of the box and whiskers).
- Skewness as asymmetry in the upper and lower parts of the box and whisker length.
- We use the function `boxplot(variable)`. Adding labels to the axes and colours is done as for `hist`.

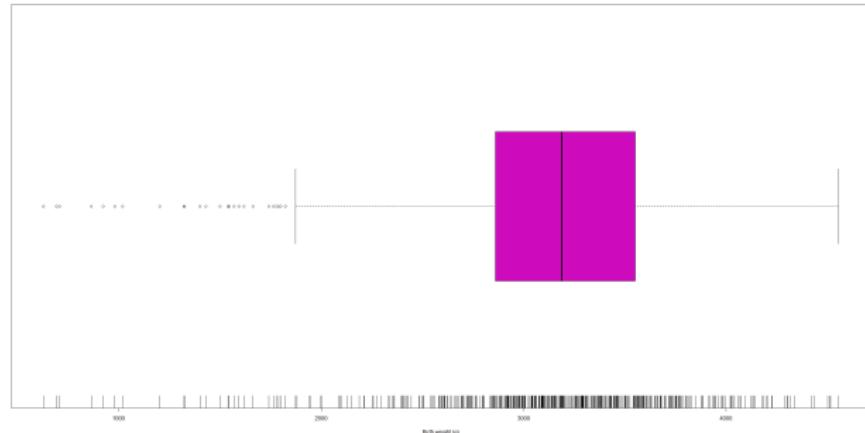
# Histograms and a Box Plot



# A Basic Box Plot

- When describing data we can even add the observations to the plot.
- Notice the function `rug` shows the observations.

```
boxplot(Births$bweight, xlab = "Birth weight (g)", horizontal = TRUE,  
        col = 6)  
rug(Births$bweight)
```

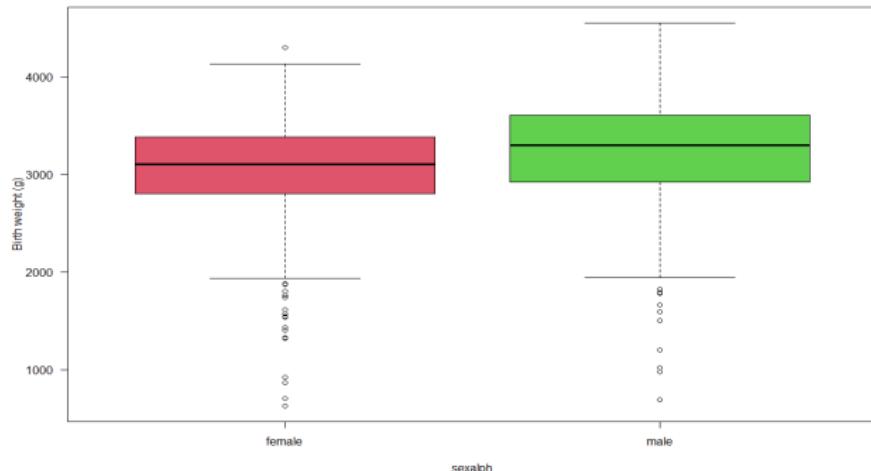


## Box Plot for Groups

A useful feature is that we can make box plots for different groups next to each other for comparison. Notice the option `data = Births`.

*# BOX PLOT FOR BOYS AND GIRLS*

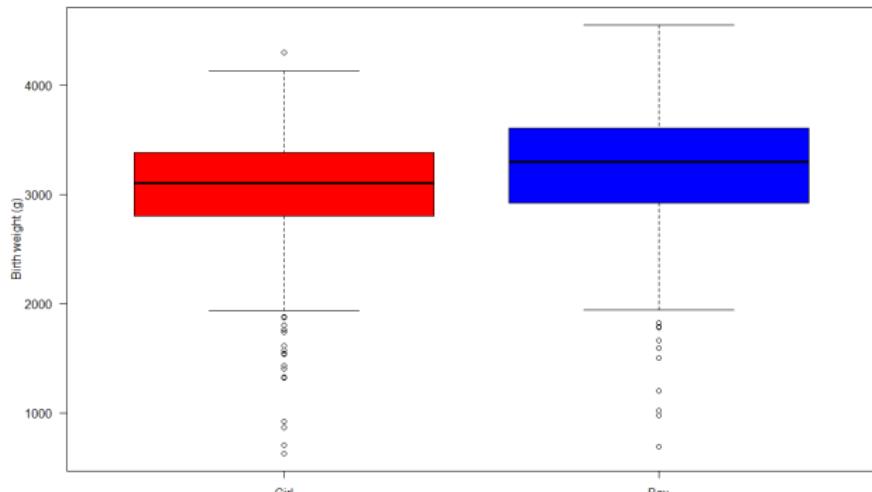
```
boxplot(bweight ~ sexalph, data = Births, las = 1,  
        ylab = "Birth weight (g)", col = 2:3)
```



# Box Plot for Groups

Set our own axis. Notice `xaxt = "n"`.

```
# BOX PLOT WHERE WE WANT TO MAKE OUR OWN AXIS
boxplot(bweight ~ sexalph, data = Births, las = 1,
         ylab = "Birth weight (g)", col = c("red", "blue"), xaxt = "n")
axis(1 ,at = c(1,2), labels = c('Girl', 'Boy'))
```



## Exercise

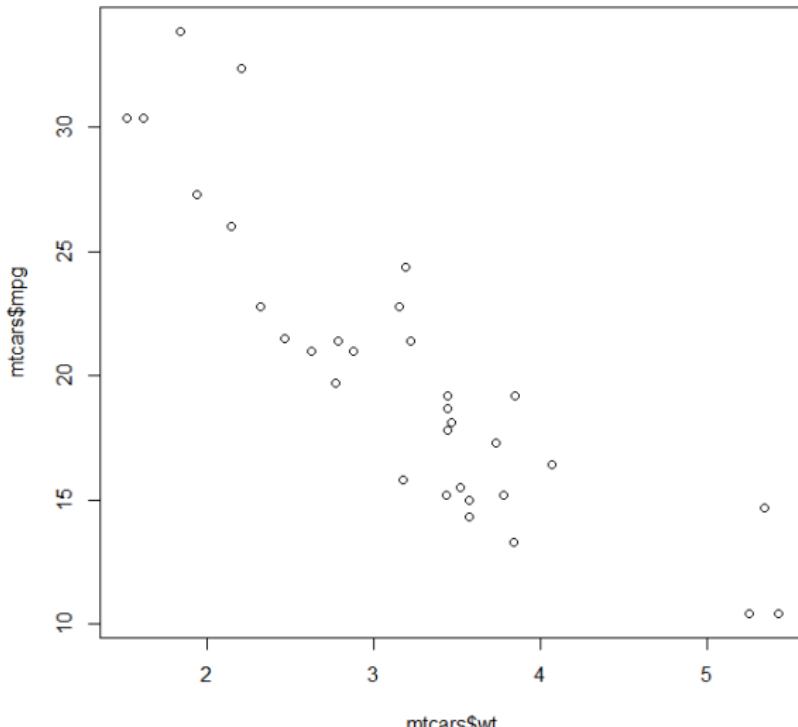
- Load the cdc data set.
- Make a box plot of the weight in kg showing individual observations. Add your own titles and colours.
- Make a box plot for each level of "genhlth", where each box is a different colour. How were the plots sorted?
- Make a new variable genhlth.num with 5 levels where 1 is "poor", 2 is "fair", 3 is "good", 4 is "very good" and 5 is "excellent".
- Make a new box plot for each level of genhlth.num with labels on the x-axis.
- Does this box plot show any pattern?

# The Basic Scatter Plot

- The scatter plot is the standard graph for examining the relationship between two continuous variables.
- The `plot(x,y)` function is used to create scatter plots. Where (x,y) are the points we want to plot.
- We will look at the relationship between car weight (lbs/1000) and miles per gallon for 32 cars.

```
plot(mtcars$wt, mtcars$mpg)
```

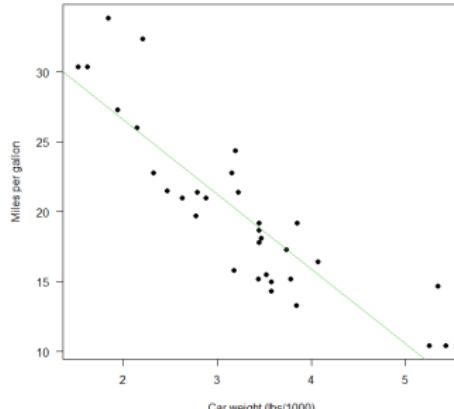
# The Basic Scatter Plot



# The Scatter Plot

- We can customize the scatter plot similar to before.
- The function `abline` adds a straight line to the plot.
- When we write `abline(lm(mpg ~ wt))` we get the best fitting line.

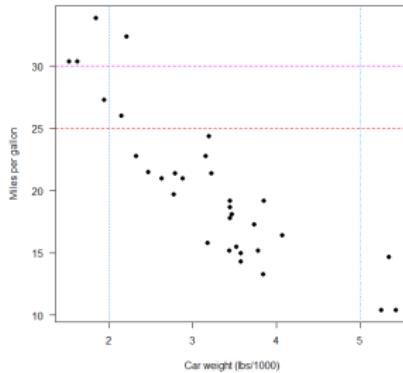
```
plot(mtcars$wt, mtcars$mpg, xlab = "Car weight (lbs/1000)",  
      ylab = "Miles per gallon", las = 1, pch = 19)  
abline(lm(mtcars$mpg ~ mtcars$wt), lty = 1, col = 3)
```



# abline

- The function `abline` can also add reference lines to a plot.
- A horizontal line, e.g. at 25 and 30 `abline(h = c(25, 30))`
- A vertical line, e.g. at 2 and 5 `abline(v = c(2, 5))`

```
plot(mtcars$wt, mtcars$mpg, xlab = "Car weight (lbs/1000)",  
     ylab = "Miles per gallon", las = 1, pch = 19)  
abline(h = c(25, 30), col = c("red", "magenta"), lty = 2)  
abline(v = c(2, 5), col = 4:5, lty = 3:4)
```

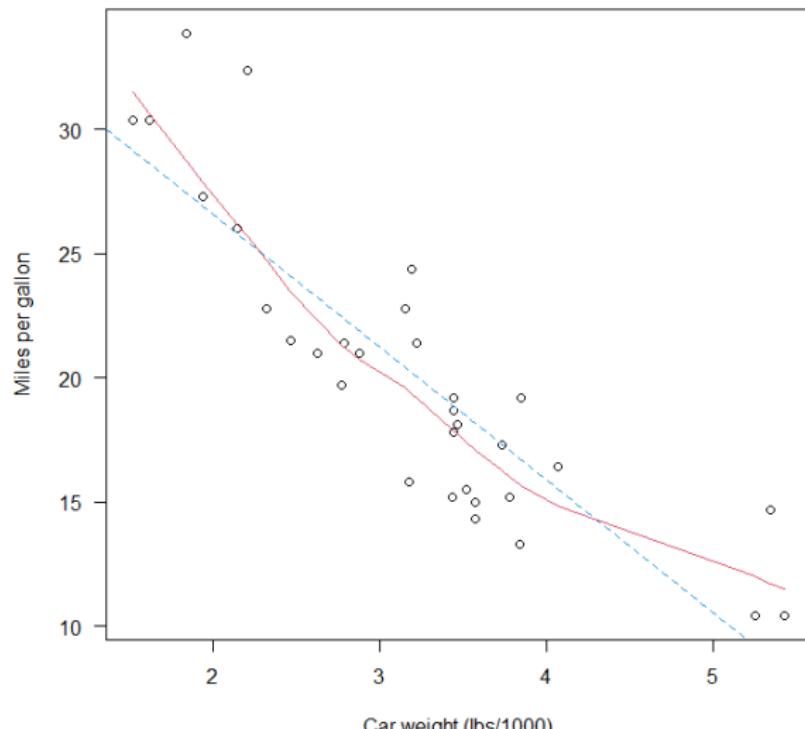


## Add a smoothed line

Perhaps we do not think the association is linear and try a nonparametric smoothed line.

```
plot(mtcars$wt, mtcars$mpg, xlab = "Car weight (lbs/1000)",  
      ylab = "Miles per gallon", las = 1)  
abline(lm(mtcars$mpg ~ mtcars$wt), lty = 2, col = 4)  
lines(lowess(mtcars$wt, mtcars$mpg), lty = 1, col = 2)
```

# Add a smoothed line

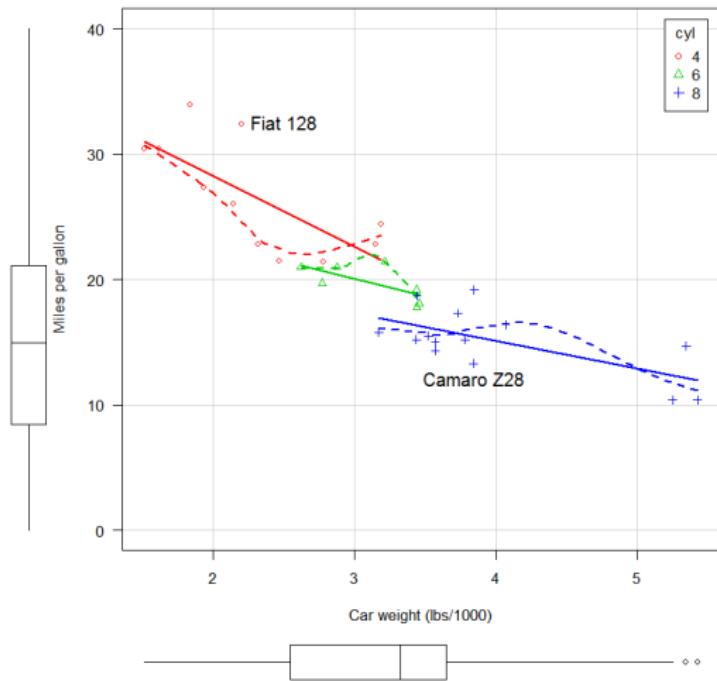


# Enhanced graph procedures: Scatter plot example from the "car" package

```
scatterplot(mpg ~ wt | cyl, data = mtcars, ylim = c(0,40),  
           xlab = "Car weight (lbs/1000)",  
           ylab = "Miles per gallon", las = 1,  
           legend = list(coords="topright"),  
           id = list(method="identify"),  
           boxplots = "xy", col=2:4)
```

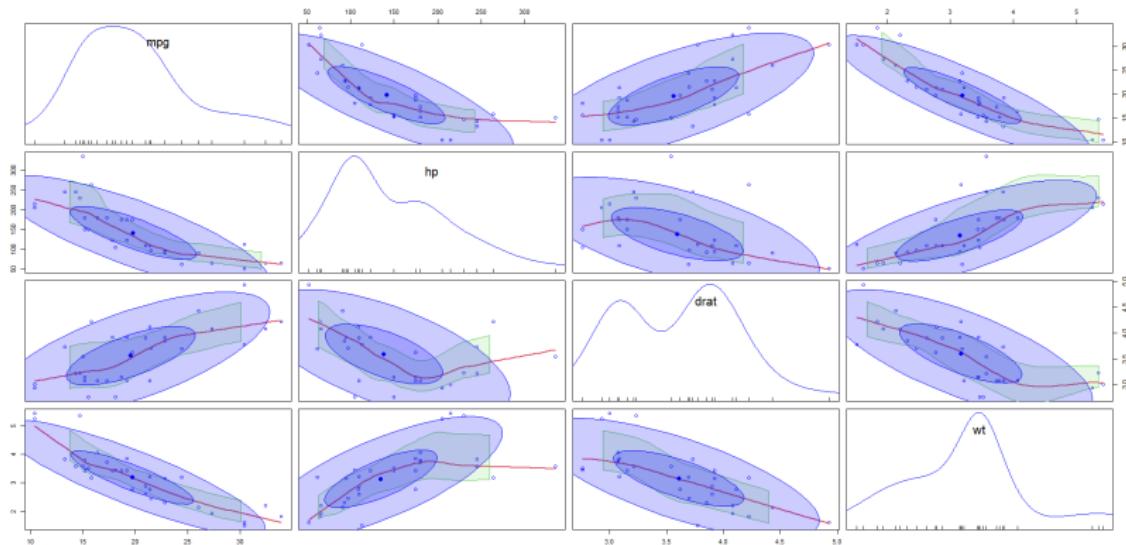
- Here we want to plot miles per gallon versus weight for cars that have 4, 6 or 8 cylinders. We write this as `mpg ~ wt | cyl`.
- By default we get different colours for groups and both a linear and a smoothed line.
- A legend is included in the top right corner of the plot.
- The option `id = list(method='identify')` means that points can be identified by mouse clicks.
- Box plots of miles per gallon and weight included ("xy"option for both axes).
- More possibilities: `?scatterplot`.

# The resulting scatter plot



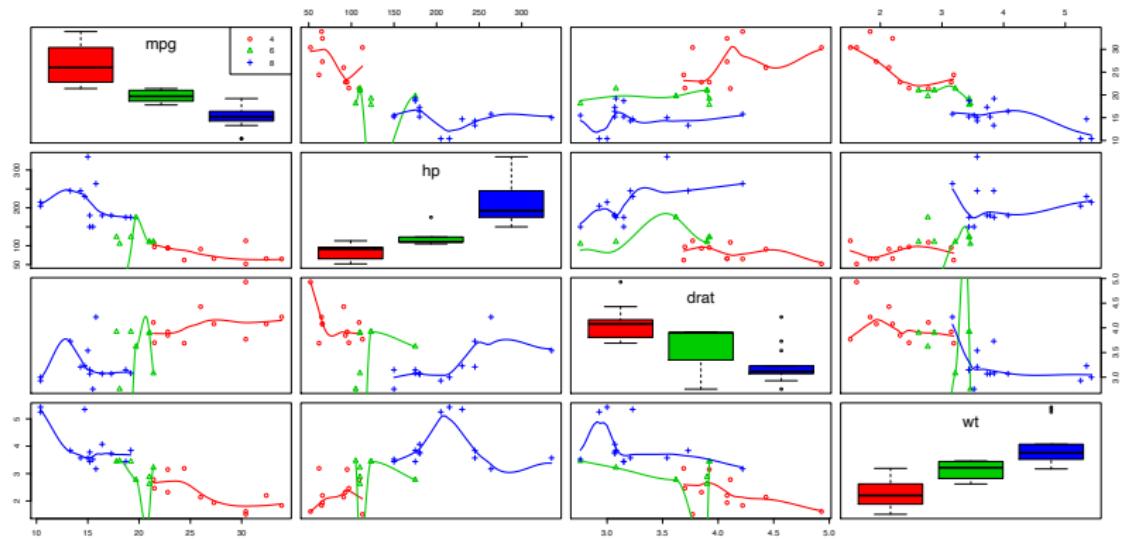
# A scatter plot matrix from the "car" package

```
scatterplotMatrix(~ mpg + hp + drat + wt, data = mtcars,  
smooth = list(col.smooth=2,col.spread=3,  
lty.smooth=1),  
regLine = FALSE,ellipse = TRUE)
```



# A scatter plot matrix with boxplots

```
scatterplotMatrix(~ mpg + hp + drat + wt | cyl,
  diagonal = list(method="boxplot"),
  smooth = list(lty.smooth=1),
  regLine = FALSE, col = 2:6,data = mtcars )
```



## Exercise

- Load the cdc data set.
- Make a basic scatter plot of weight and desired weight with “nice” axes and labels.
- Add a smoothed line to the plot.
- Make a scatter plot of weight and desired weight for each level of “genhlth” in separate plots, with straight and smoothed lines.
- Install the package “car”.
- Make a scatter plot of weight and desired weight for each level of “genhlth” in one plot.

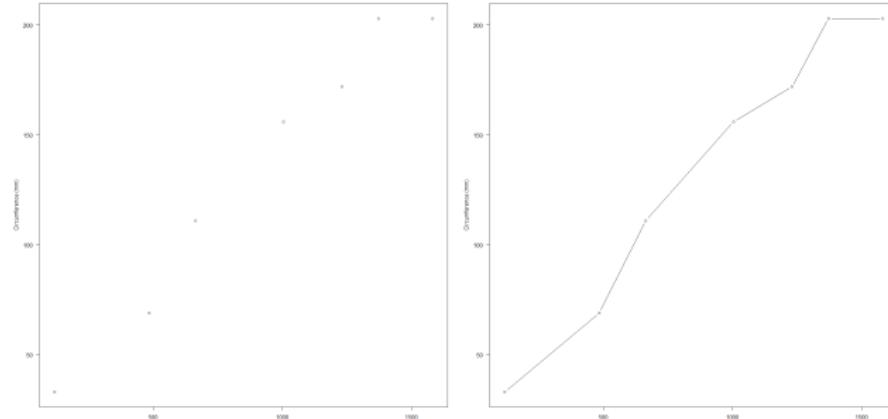
# Exercise

- Load the Protein data set.
- Make a scatter plot matrix of “RedMeat”, “WhiteMeat”, “Eggs”, “Milk” and “Fish”. Can you see any patterns in the protein intake?
- Also make a scatter plot matrix with box plots in the diagonal.

# A Line Plot

Connecting points in a scatter plot from left to right. Here the growth of a tree. Notice the option `type = "b"` meaning points joined by lines.

```
par(mfrow=c(1,2)),
plot(TreeA$age, TreeA$circumference, xlab = "Age (days)",
      ylab = "Circumference (mm)", las = 1)
plot(TreeA$age, TreeA$circumference, type = "b", xlab = "Age (days)",
      ylab = "Circumference (mm)", las = 1)
```

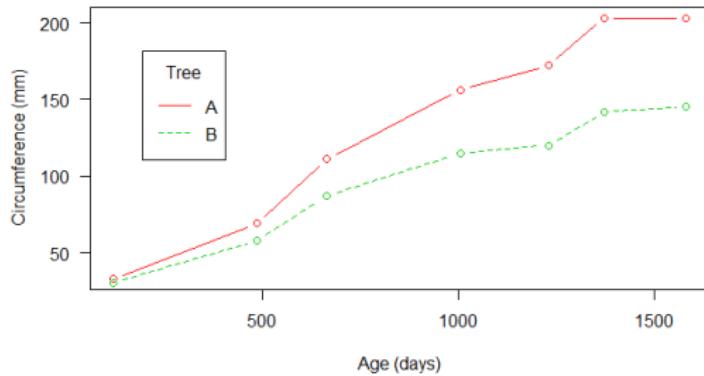


# Difference between `plot()` and `lines()` functions

- We have seen both the `plot` and the `lines` functions.
- The `plot` function creates a new graph. It is a **high-level** plotting function.
- The `lines` function adds information to an existing graph but it cannot produce its own graph. It is a **low-level** plotting function.
- A **high-level** plotting function can (often) be converted to a **low-level** plotting function with the option `ADD=TRUE`.
- Usually `lines` will be used after a **high-level** plotting function (such as `plot`) has produced a graph.

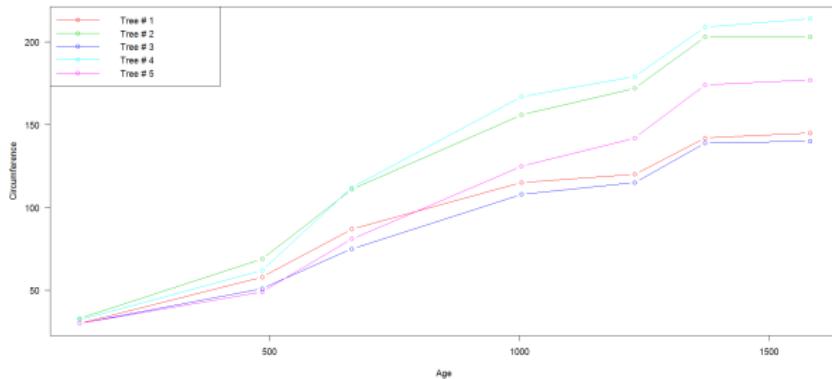
# A line plot and a legend

```
plot(TreeA$age, TreeA$circumference, type = "b", lty = 1,  
     xlab = "Age (days)", ylab = "Circumference (mm)", las = 1, col= 2)  
lines(TreeB$age, TreeB$circumference, type = "b", col = 3, lty = 2)  
legend(locator(1), # we will place it with a mouse click  
       legend = c("A","B"), title = "Tree",  
       lty = 1:2, col= 2:3)
```



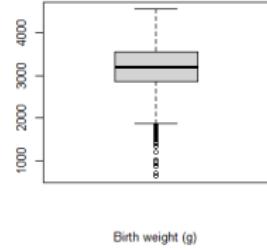
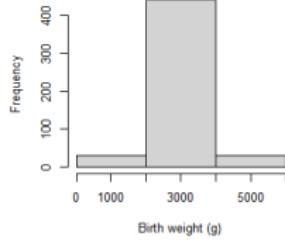
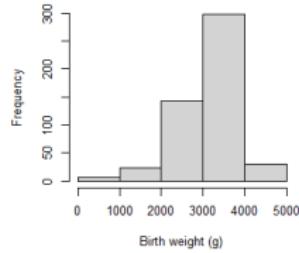
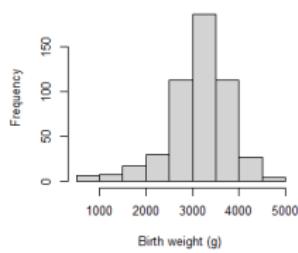
# Plot of growth of 5 trees

```
treedata<-Orange  
plot(treedata$age,treedata$circumference,pch=" ",xlab = 'Age',  
      ylab='Circumference',las=1)  
for(i in 1:5){lines(treedata$age[treedata$Tree==i],  
                     treedata$circumference[treedata$Tree==i],type="b",col=i+1)}  
legend(x="topleft",paste("Tree #",1:5),pch=1,lty=1,col=2:6)
```



# Layout of several plots on one graph

Several plots on one graph:



Use the option `par(mfrow = c(2, 2))`, and back to one plot `par(mfrow = c(1, 1))`.

# Layout of several plots on one graph

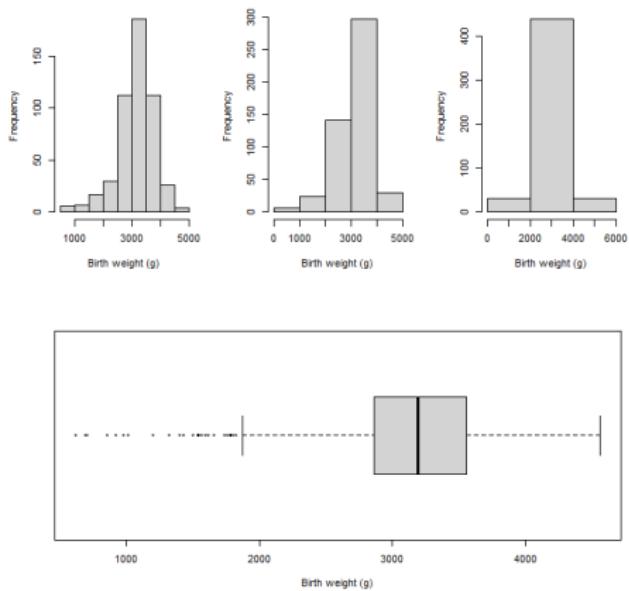
The `layout()` function:

- indicate in matrix form which part of the plot area that you wish to belong to which graph

```
> layout.matrix<-matrix(c(1:3,rep(4,3)),nrow=2,ncol=3,byrow=T)
> layout.matrix
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    4    4    4
```

# Layout of several plots on one graph

```
>layout(layout.matrix)
```



# Principal Components Analysis

January 10th, 2025

Anders Stockmarr  
Section for Statistics and Data Analysis, DTU  
anst@dtu.dk

DTU Compute  
Department of Applied Mathematics and Computer Science

---

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$
$$\Theta^{\sqrt{17}} + \Omega \int_a^b \delta e^{i\pi} =$$
$$\infty = \{2.71828182845904523536028747135266249$$
$$\Sigma \gg,$$
$$!$$

# Programme

- Monday: Statistical Inference, the t-test
- Tuesday: Simple and Multiple Regression
- Wednesday: ANOVA, ANCOVA, and Linear Models
- Thursday: Categorical Data, Writing Statistical Reports, Logistic regression
- Friday: Repeated Measurements, **Principal Component Analysis**

# Contents:

1. Introduction.
2. Example: Wine data variable separation
3. Principal Components Analysis
4. PCA of the Wine Data
5. Diagnostics
6. Example: Jam
7. Example: Horse.

# Introduction

# PCA: Modeling Data

- In all previous parts of the course, data has been subdivided into two classes:
  - Response variables;
  - explanatory variables.
- The data were analyzed in order to identify impact of the explanatory variable on the response variables.
- What if the target of analysis isn't to identify effects;
- But to **uncover the structure** of a complicated set of data?

This is the essence of the use of

## Principal Components Analysis - PCA

# Main example: The Wine data

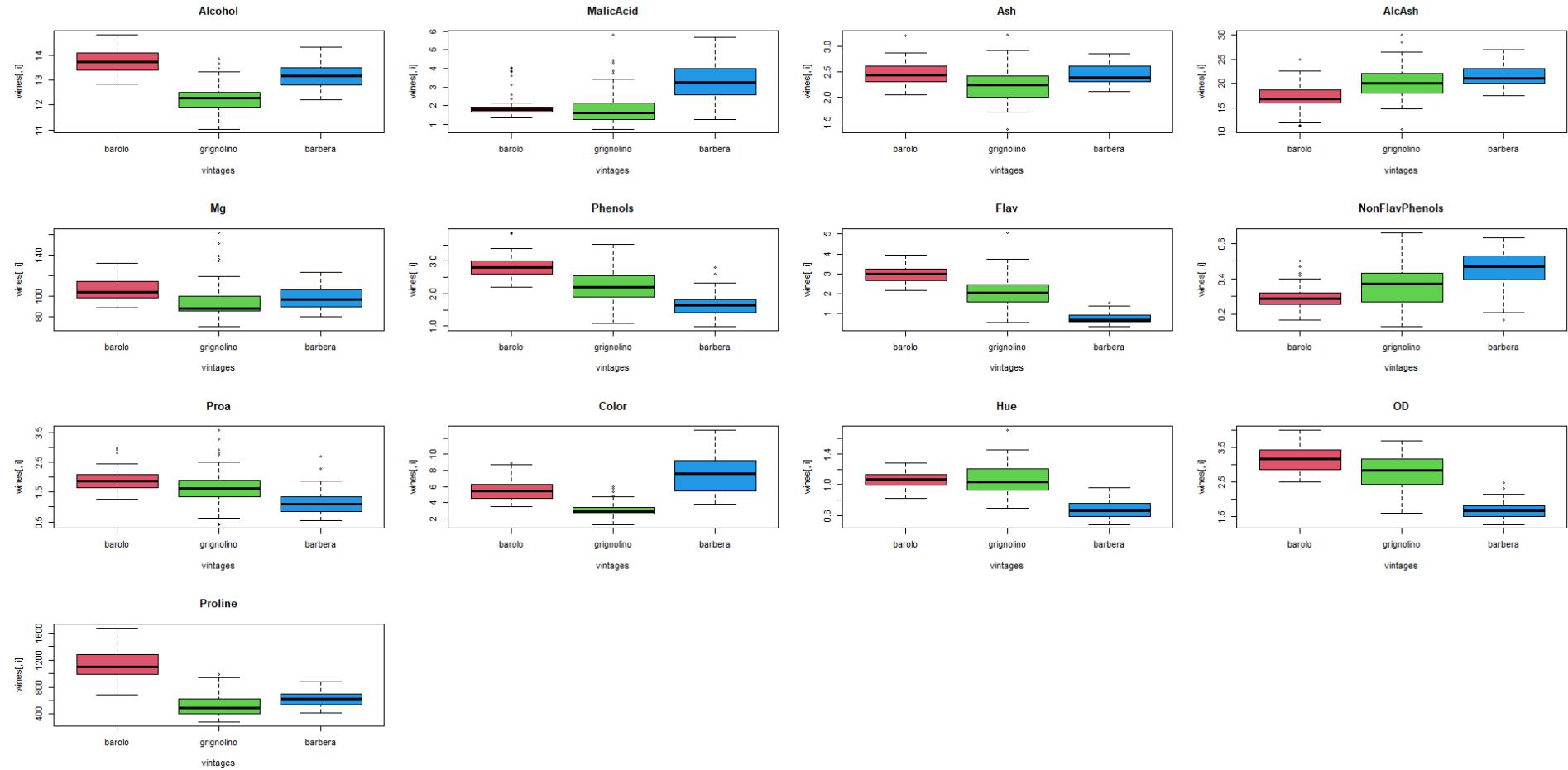
```
load("Data/Winedata.Rdata")
```

- A dataset consisting of n=178 Italian wines. Of these, 59 are Barolo wines, 71 are Grignolino wines, and 48 are Barbera wines. Two sub-elements: wines and vintages (Barolo, Grignolino or Barbera).

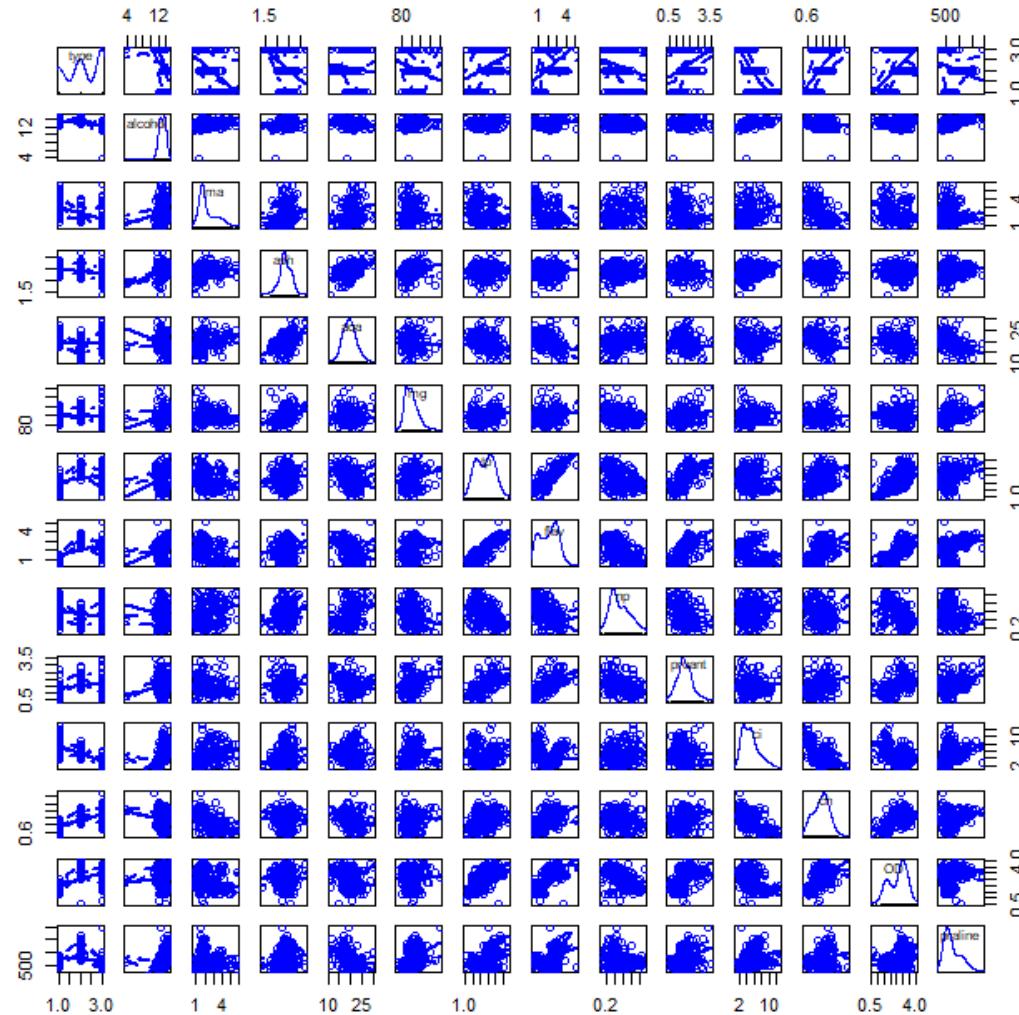
- 13 characteristics of the wines:

- |                              |   |
|------------------------------|---|
| 1) Alcohol                   | 8) Nonflavanoid phenols: NonFlavPhenols |
| 2) MalicAcid                 | 9) Proanthocyanins: Proa                |
| 3) Ash                       | 10) Color intensity: Color              |
| 4) Alkalinity of ash: AlcAsh | 11) Color hue: Hue                      |
| 5) Magnesium: Mg             | 12) OD280/OD315 protein measurement: OD |
| 6) Total Phenols: Phenols    | 13) Proline (amino acid): Proline       |
| 7) Flavanoids: Flav          |   |

# Variation of Vintages



# The Wine Data



# Separation of Wines

- What causes wines to be different?
- With the 13 characteristics, we can distinguish wines through differences in the characteristics. But are all 13 characteristics necessary? Some may be redundant.
- If we can identify scales (linear combinations of the characteristics) where the characteristics vary the most, we can also find a scale that differentiates optimally between the wines.

# Separation of Wines

- The variance covariance matrix of the wine characteristics:

	Alcohol	MalicAcid	Ash	AlcAsh	Mg	Phenols	Flav	NonFlavPhenols	Proa	Color	Hue	OD	Proline
Alcohol	<b>0.66</b>	0.09	0.05	-0.84	3.14	0.15	0.19		-0.02	0.06	1.03	-0.01	0.04
MalicAcid	0.09	<b>1.25</b>	0.05	1.08	-0.87	-0.23	-0.46		0.04	-0.14	0.64	-0.14	-0.29
Ash	0.05	0.05	<b>0.08</b>	0.41	1.12	0.02	0.03		0.01	0.00	0.16	0.00	0.00
AlcAsh	-0.84	1.08	0.41	<b>11.15</b>	-3.97	-0.67	-1.17		0.15	-0.38	0.15	-0.21	-0.66
Mg	3.14	-0.87	1.12	-3.97	<b>203.99</b>	1.92	2.79		-0.46	1.93	6.62	0.18	0.67
Phenols	0.15	-0.23	0.02	-0.67	1.92	<b>0.39</b>	0.54		-0.04	0.22	-0.08	0.06	0.31
Flav	0.19	-0.46	0.03	-1.17	2.79	0.54	<b>1.00</b>		-0.07	0.37	-0.40	0.12	0.56
NonFlavPhenols	-0.02	0.04	0.01	0.15	-0.46	-0.04	-0.07		<b>0.02</b>	-0.03	0.04	-0.01	-0.04
Proa	0.06	-0.14	0.00	-0.38	1.93	0.22	0.37		-0.03	<b>0.33</b>	-0.03	0.04	0.21
Color	1.03	0.64	0.16	0.15	6.62	-0.08	-0.40		0.04	-0.03	<b>5.37</b>	-0.28	-0.71
Hue	-0.01	-0.14	0.00	-0.21	0.18	0.06	0.12		-0.01	0.04	-0.28	<b>0.05</b>	0.09
OD	0.04	-0.29	0.00	-0.66	0.67	0.31	0.56		-0.04	0.21	-0.71	0.09	<b>0.50</b>
Proline	164.57	-67.55	19.32	-463.36	1769.16	98.17	155.45		-12.20	59.55	230.77	17.00	69.93
													<b>99166.72</b>

# Separation of Wines

- To avoid scaling problems, we must scale the data to the same scale.
- The `scale` function in R subtracts the mean and divide by the sd:

$$X^{scaled} = \frac{X - \text{mean}(X)}{\text{sd}(X)}$$

```
round(var(scale(wines)), digits=2)
```

	Alcohol	MalicAcid	Ash	AlcAsh	Mg	Phenols	Flav	NonFlavPhenols	Proa	Color	Hue	OD	Proline	
Alcohol	<b>1.00</b>	0.09	0.21	-0.31	0.27	0.29	0.24		-0.16	0.14	0.55	-0.07	0.07	0.64
MalicAcid	0.09	<b>1.00</b>	0.16	0.29	-0.05	-0.34	-0.41		0.29	-0.22	0.25	-0.56	-0.37	-0.19
Ash	0.21	0.16	<b>1.00</b>	0.44	0.29	0.13	0.12		0.19	0.01	0.26	-0.07	0.00	0.22
AlcAsh	-0.31	0.29	0.44	<b>1.00</b>	-0.08	-0.32	-0.35		0.36	-0.20	0.02	-0.27	-0.28	-0.44
Mg	0.27	-0.05	0.29	-0.08	<b>1.00</b>	0.21	0.20		-0.26	0.24	0.20	0.06	0.07	0.39
Phenols	0.29	-0.34	0.13	-0.32	0.21	<b>1.00</b>	0.86		-0.45	0.61	-0.06	0.43	0.70	0.50
Flav	0.24	-0.41	0.12	-0.35	0.20	0.86	<b>1.00</b>		-0.54	0.65	-0.17	0.54	0.79	0.49
NonFlavPhenols	-0.16	0.29	0.19	0.36	-0.26	-0.45	-0.54		<b>1.00</b>	-0.37	0.14	-0.26	-0.50	-0.31
Proa	0.14	-0.22	0.01	-0.20	0.24	0.61	0.65		-0.37	<b>1.00</b>	-0.03	0.30	0.52	0.33
Color	0.55	0.25	0.26	0.02	0.20	-0.06	-0.17		0.14	-0.03	<b>1.00</b>	-0.52	-0.43	0.32
Hue	-0.07	-0.56	-0.07	-0.27	0.06	0.43	0.54		-0.26	0.30	-0.52	<b>1.00</b>	0.57	0.24
OD	0.07	-0.37	0.00	-0.28	0.07	0.70	0.79		-0.50	0.52	-0.43	0.57	<b>1.00</b>	0.31
Proline	0.64	-0.19	0.22	-0.44	0.39	0.50	0.49		-0.31	0.33	0.32	0.24	0.31	<b>1.00</b>

# Separation of Wines

- Let us consider the correlation matrix:

```
X<-var(scale(wines))
```

The sum of the standardized variances:

```
sum(diag(X))  
[1] 13
```

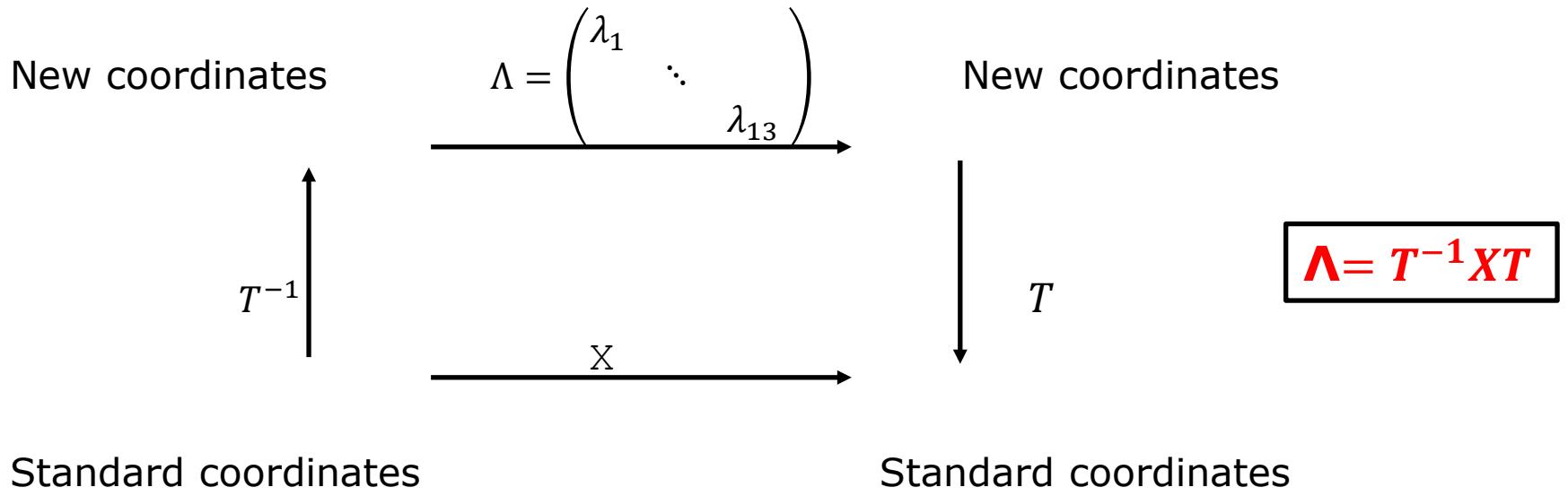
Of course – there are 13 variables.

Original question: In which direction (scale) do wines data vary the most?  
Let us start by representing the data in a set of coordinates where no correlation is present, to get an overview not disturbed by correlations.

# Separation of Wines

- No correlations means that  $\mathbf{X}$  is represented by a diagonal matrix in these directions;
- in other words that the new coordinates  $\mathbf{T}$  consists of eigenvectors for  $\mathbf{X}$ ; solution to the equation

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{v}$$



# Separation of Wines

- Eigenvectors in R:

```
T<-eigen(X)$vectors
```

The inverse of T is equal to the matrix transpose t(T):

```
Lambda<-t(T)%*%X%*%T
```

```
round(Lambda, digits=2)
```

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,] 4.71 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[2,] 0.00 2.5 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[3,] 0.00 0.0 1.45 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[4,] 0.00 0.0 0.00 0.92 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[5,] 0.00 0.0 0.00 0.00 0.85 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[6,] 0.00 0.0 0.00 0.00 0.00 0.64 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[7,] 0.00 0.0 0.00 0.00 0.00 0.00 0.55 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[8,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.35 0.00 0.00 0.00 0.00 0.00 0.0
[9,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.29 0.00 0.00 0.00 0.00 0.0
[10,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00 0.0
[11,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.23 0.00 0.00 0.0
[12,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.17 0.00 0.0
[13,] 0.00 0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.1 0.0
```

# Separation of Wines

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,] 4.71  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[2,] 0.00  2.5 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[3,] 0.00  0.0 1.45 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[4,] 0.00  0.0 0.00 0.92 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[5,] 0.00  0.0 0.00 0.00 0.85 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[6,] 0.00  0.0 0.00 0.00 0.00 0.64 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[7,] 0.00  0.0 0.00 0.00 0.00 0.55 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[8,] 0.00  0.0 0.00 0.00 0.00 0.00 0.35 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[9,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.29 0.00 0.00 0.00 0.00 0.0
[10,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00 0.0
[11,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.23 0.00 0.00 0.0
[12,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.17 0.00 0.0
[13,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.1
```

- It turns out that this matrix **provides us with the answer to our problem**:
- Any (normed) linear combination of the 13 standardized characteristics will also be a (normed) linear combination of the 13 uncorrelated eigenvalues;
- A little consideration shows that because of this, the variance of any normed linear combination can never exceed the maximum variance of the eigenvectors – **4.71**.
- The solution is thus **the first eigenvector**:  $\mathbb{T}[, 1]$ .

# Separation of Wines

- The combination of the scaled data that varies the most:

```
round(T[,1], digits=2)
[1] -0.14  0.25  0.00  0.24 -0.14 -0.39 -0.42  0.30 -0.31  0.09 -0.30
[12] -0.38 -0.29
```

- Thus the most varying combination of the scaled data is

$$\begin{aligned} & -0.14 * \widetilde{\text{Alcohol}} + 0.25 * \widetilde{\text{MalicAcid}} + 0 * \widetilde{\text{Ash}} + 0.24 * \widetilde{\text{AlcAsh}} - 0.14 * \widetilde{\text{Mg}} - 0.39 * \widetilde{\text{Phenols}} \\ & - 0.42 * \widetilde{\text{PhenolsFlav}} + 0.30 * \widetilde{\text{NonFlavPhenols}} - 0.31 * \widetilde{\text{Proa}} + 0.09 * \widetilde{\text{Color}} - 0.30 * \widetilde{\text{Hue}} \\ & - 0.38 * \widetilde{\text{OD}} - 0.29 * \widetilde{\text{Proline}} \end{aligned}$$

Where the  $\sim$  versions are the scaled variables, with the mean subtracted and divided by the standard deviation.

- This is the scale that we want to look at, when we want to maximize the separation of wines.

# Separation of Wines

```
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12] [,13]
[1,] 4.71  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[2,] 0.00  2.5 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[3,] 0.00  0.0 1.45 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[4,] 0.00  0.0 0.00 0.92 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[5,] 0.00  0.0 0.00 0.00 0.85 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[6,] 0.00  0.0 0.00 0.00 0.00 0.64 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[7,] 0.00  0.0 0.00 0.00 0.00 0.00 0.55 0.00 0.00 0.00 0.00 0.00 0.00 0.0
[8,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.35 0.00 0.00 0.00 0.00 0.00 0.0
[9,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.29 0.00 0.00 0.00 0.00 0.0
[10,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.25 0.00 0.00 0.00 0.0
[11,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.23 0.00 0.00 0.0
[12,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.17 0.00 0.0
[13,] 0.00  0.0 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.1
```

- **Further conclusions:**

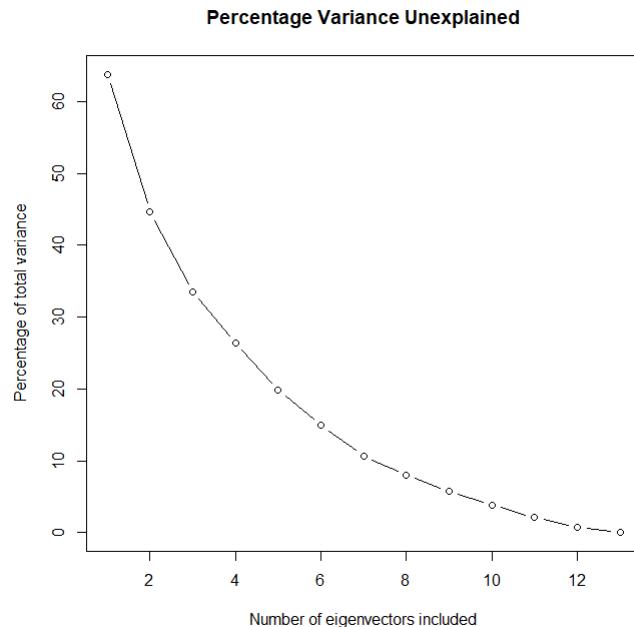
- The scale, uncorrelated with the first eigenvector, that varies the most, is exactly the 2<sup>nd</sup> eigenvector  $\mathbb{T}[ , 2 ]$ , with variance **2.5**. And so on...
- The total variation after the coordinate shift is unchanged:

```
sum(diag(Lambda))  
[1] 13
```

- Also note that the contribution from the 13th eigenvector is only 0.1/13, **0.7%**

# Separation of Wines

```
plot(100*(13-cumsum(diag(Lambda)))/13,type="b",
  main="Percentage Variance Unexplained",
  xlab='Number of eigenvectors included',
  ylab='Percentage of total variance')
```



Eigenvectors	% variance explained
1	36
2	55
3	67
4	74
5	80
6	85
7	89
8	92
9	94
10	96
11	98
12	99
13	100

# Principal Component Analysis

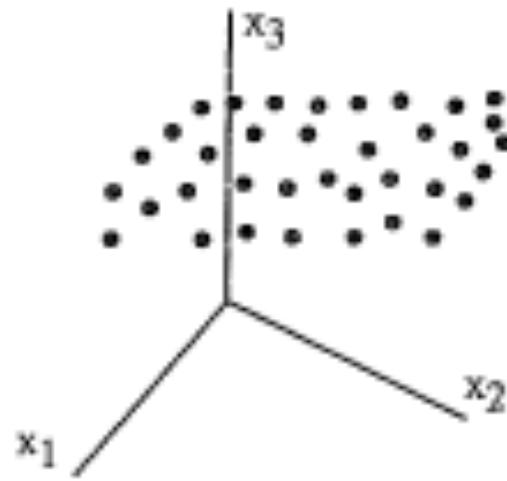
- PCA is a method to handle many variables, which are mutually correlated.
- PCA seeks to identify underlying dimensions in a data material, and to estimate the relationship between these.
- PCA may be used as a data reducing method, often in relation to multiple regression.
- PCA reduces the number of explanatory variables to a lesser number of "principal components", with (we hope) nearly as much of the variation as the initial variables.

**What PCA does is essentially the contents of the preceding slides!**

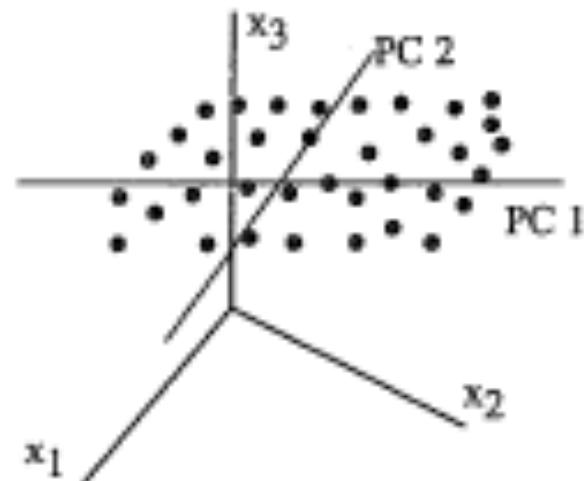
- The eigenvectors on the previous slides are **exactly** the Principal Components.

# Identification of Lower- Dimensional Spaces

*Figure 3.10 Data swarm  
(quasi-planar)*



*Figure 3.11 Data swarm with  
2 PCs*



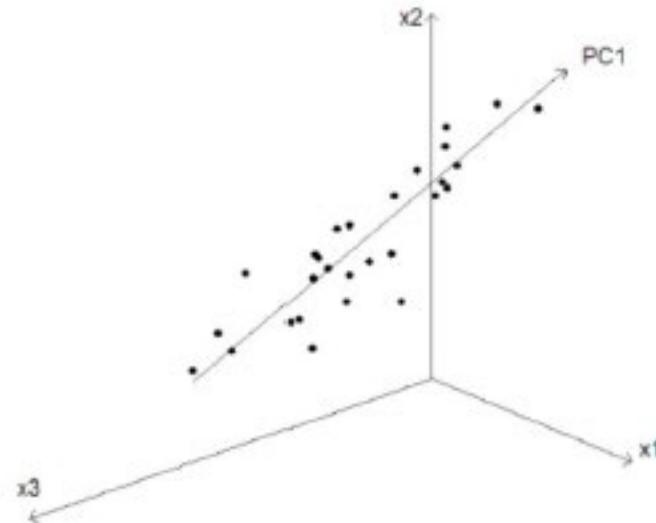
# The First Principal Component – PC1

We look for the direction that **explains as much as possible of the variation in the data**. We assume here 3 variables and n=28:

$$p_1 = t_{11}X_1 + t_{21}X_2 + t_{31}X_3,$$

where  $\sum_{j=1}^3 t_{1j}^2 = 1$ .

- p are the "scores";
- t are the "loadings".



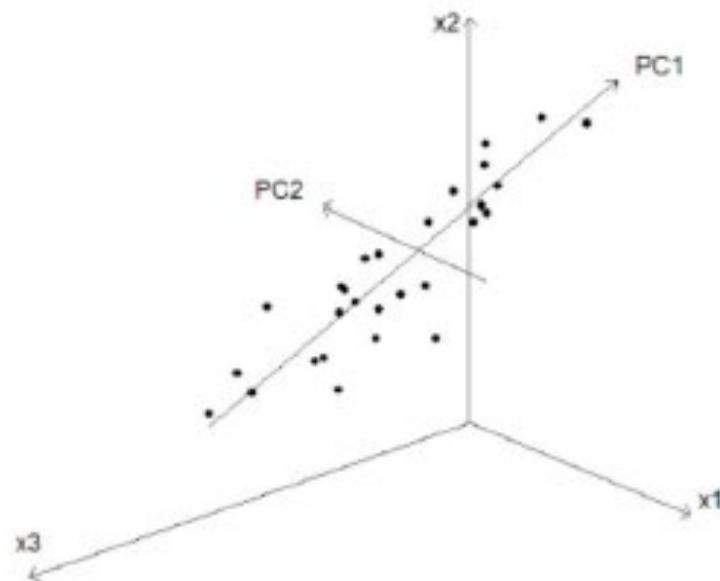
# The Second Principal Component – PC2

We consider the plane perpendicular to **PC1**, and find the linear combination that explains the 2nd most variation:

$$\mathbf{p}_2 = t_{12}\mathbf{X}_1 + t_{22}\mathbf{X}_2 + t_{32}\mathbf{X}_3,$$

where  $\sum_{j=1}^3 t_{2j}^2 = 1$

PC1 and PC2 are orthogonal.



# PCA of the Wines Data

```
wines.PC<- PCA(scale(wines))
names(wines.PC)
[1] "scores"          "loadings"        "var"            "totalvar"
[5] "centered.data"
```

```
summary(wines.PC)
```

PCA model of a mean-centered matrix of 178 by 13  
Number of PCs to cover 90 percent of the variance: 8

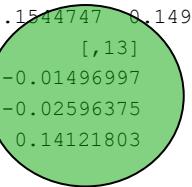
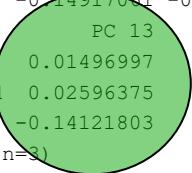
	Var	Cumul. var.
PC 1	36.198848	36.19885
PC 2	19.207490	55.40634
PC 3	11.123631	66.52997
PC 4	7.069030	73.59900
PC 5	6.563294	80.16229
PC 10	1.930019	96.16972

- Lets take a look at the components

# PCA of the Wines Data – the Loadings

- The **loadings** are the *coordinates of the principal components*:

```
> head(wines.PC$loadings,n=3)
      PC 1       PC 2       PC 3       PC 4       PC 5       PC 6
Alcohol -0.144329395 0.4836515 -0.20738262  0.0178563 -0.26566365 0.2135386
MalicAcid 0.245187580 0.2249309  0.08901289 -0.5368903  0.03521363 0.5368138
Ash     0.002051061 0.3160688  0.62622390  0.2141756 -0.14302547 0.1544747
      PC 7       PC 8       PC 9       PC 10      PC 11      PC 12
Alcohol -0.05639636 0.39613926 -0.50861912  0.21160473  0.22591696 -0.26628645
MalicAcid 0.42052391 0.06582674  0.07528304 -0.30907994 -0.07648554  0.12169604
Ash    -0.14917061 -0.17026002  0.30769445 -0.02712539  0.49869142 -0.04962237
      PC 13
Alcohol 0.01496997
MalicAcid 0.02596375
Ash     -0.14121803
> head(T, n=3)
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
[1,] -0.144329395 -0.4836515 -0.20738262 -0.0178563  0.26566365 0.2135386 0.05639636
[2,]  0.245187580 -0.2249309  0.08901289  0.5368903 -0.03521363 0.5368138 -0.42052391
[3,]  0.002051061 -0.3160688  0.62622390 -0.2141756  0.14302547 0.1544747 0.14917061
      [,8]      [,9]      [,10]      [,11]      [,12]      [,13]
[1,]  0.39613926  0.50861912  0.21160473 -0.22591696 -0.26628645 -0.01496997
[2,]  0.06582674 -0.07528304 -0.30907994  0.07648554  0.12169604 -0.02596375
[3,] -0.17026002 -0.30769445 -0.02712539 -0.49869142 -0.04962237  0.14121803
```



- The principal components are only identified up to a sign change.

# PCA of the Wines Data – the Scores

- The **scores** are the *new coordinates* of the (scaled) wines data *relative to the principal components*:

```
head(wines.PC$scores,n=3)
  PC 1      PC 2      PC 3      PC 4      PC 5      PC 6      PC 7
[1,] -3.307421 1.4394023 -0.1652728 0.2150246 0.6910933 0.2232504 0.59474883
[2,] -2.203250 -0.3324551 -2.0207571 0.2905387 -0.2569299 0.9245123 0.05362434
[3,] -2.509661 1.0282507 0.9800541 -0.7228632 -0.2503270 -0.5477310 0.42301218
  PC 8      PC 9      PC 10     PC 11     PC 12     PC 13
[1,] -0.06495586 -0.6396384 1.0180840 0.4502932 0.5392891439 -0.066052305
[2,] -1.02153432  0.3079780 0.1592521 0.1422560 0.3871456499  0.003626273
[3,]  0.34324787  1.1745213 0.1130420 0.2858665 0.0005819316  0.021655423
> head(scale(wines)%*%T,n=3)
      [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] -3.307421 -1.4394023 -0.1652728 -0.2150246 -0.6910933 0.2232504 -0.59474883
[2,] -2.203250  0.3324551 -2.0207571 -0.2905387  0.2569299 0.9245123 -0.05362434
[3,] -2.509661 -1.0282507 0.9800541  0.7228632  0.2503270 -0.5477310 -0.42301218
      [,8]   [,9]  [,10]  [,11]  [,12]  [,13]
[1,] -0.06495586  0.6396384 1.0180840 -0.4502932 0.5392891439 0.066052305
[2,] -1.02153432 -0.3079780 0.1592521 -0.1422560 0.3871456499 -0.003626273
[3,]  0.34324787 -1.1745213 0.1130420 -0.2858665 0.0005819316 -0.021655423
```

- Note the same sign changes

# PCA of the Wines data – var, totalvar and centered.data

```
wines.PC$var
```

PC 1	PC 2	PC 3	PC 4	PC 5	PC 6	PC 7	PC 8
4.7058503	2.4969737	1.4460720	0.9189739	0.8532282	0.6416570	0.5510283	0.3484974
PC 9	PC 10	PC 11	PC 12	PC 13			
0.2888799	0.2509025	0.2257886	0.1687702	0.1033779			

```
wines.PC$totalvar
```

```
[1] 13
```

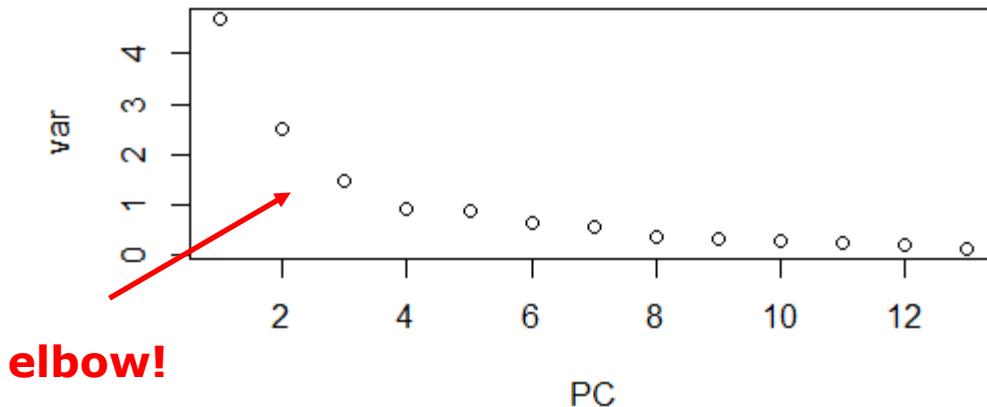
```
wines.PC$centered.data
```

```
[1] TRUE
```

We recognize the eigenvalues of the matrix  $\Lambda$ , and the sum of these. Lastly an indicator that we have ‘done the right thing’ (in this case).

# Selecting the Number of Principal Components – the Skree Plot

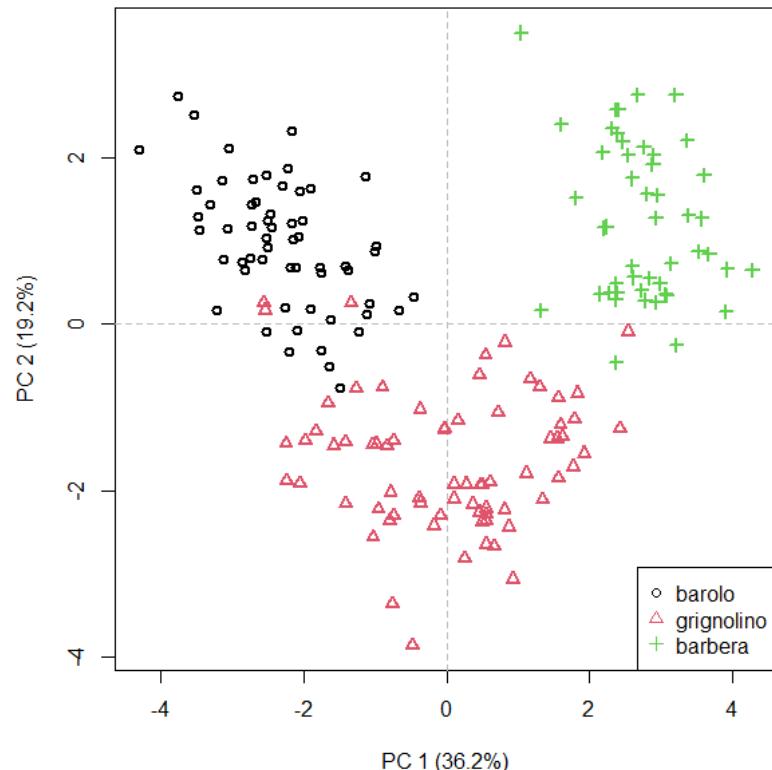
```
plot(1:13,wines.PC$var,xlab="PC",ylab="var")
```



- Rules of thumb:
  - i. You select a number of principal components where the 'elbow' of the graph is.
  - ii. You usually (but not always) only select principal components with a variance greater than 1 – if the value is lower, the PC explain less than one average ordinary observation.
  - iii. You fix the amount of total variation that you need explained – t.ex. 80%.

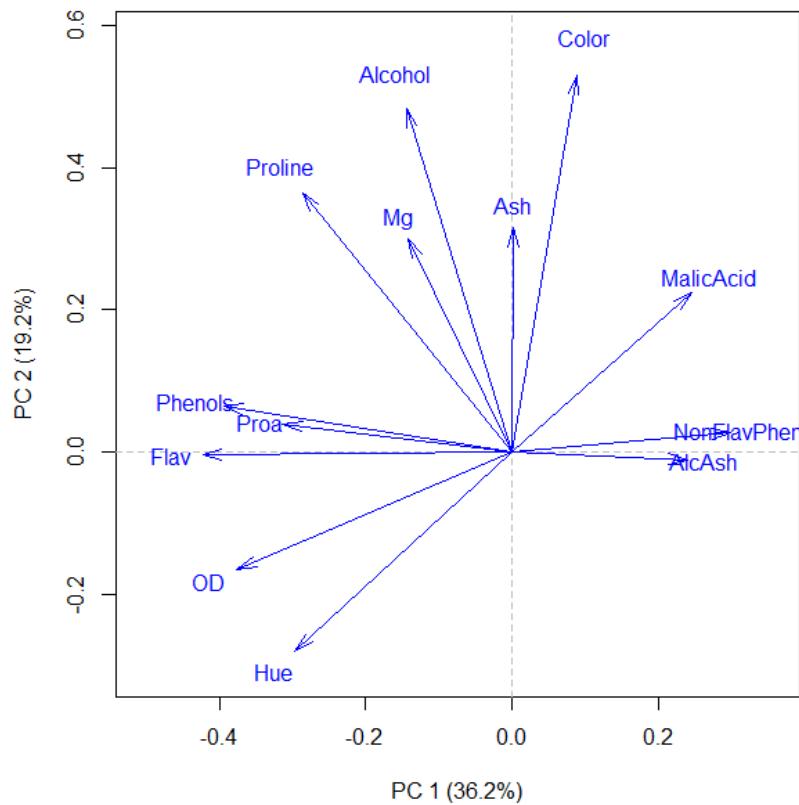
# The Score plot

```
scoreplot(wines.PC, col = vintages, pch= as.numeric(vintages), lwd=2)
legend("bottomright", levels(vintages), col=1:3,pch=1:3)
```



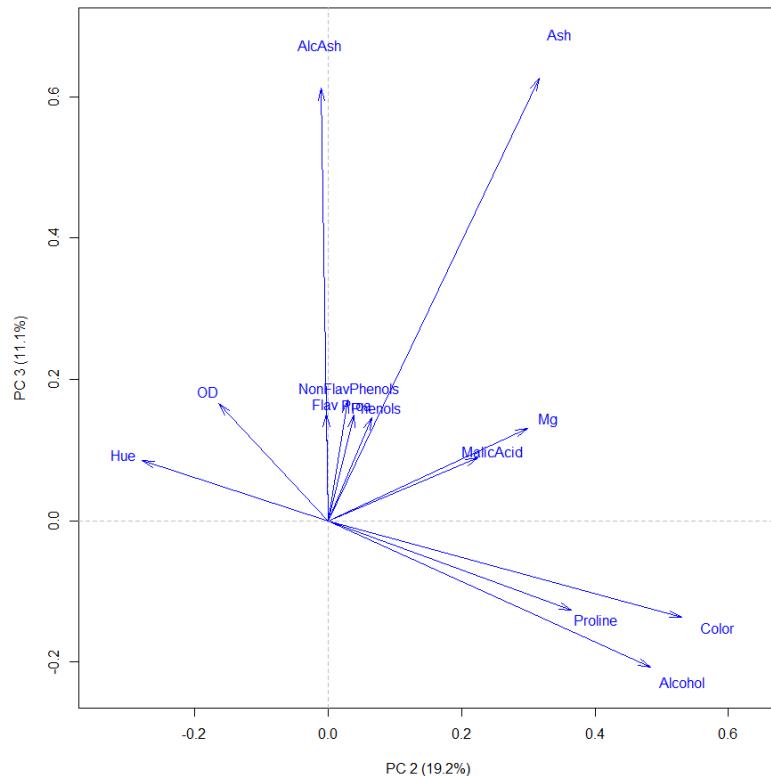
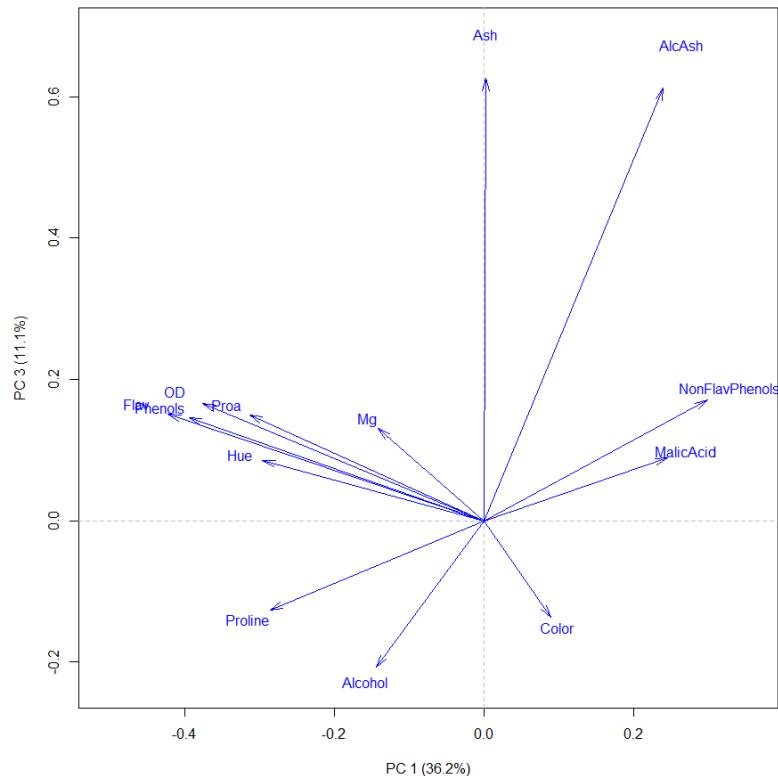
# The Loading plot

```
loadingplot(wines.PC, show.names= TRUE)
```



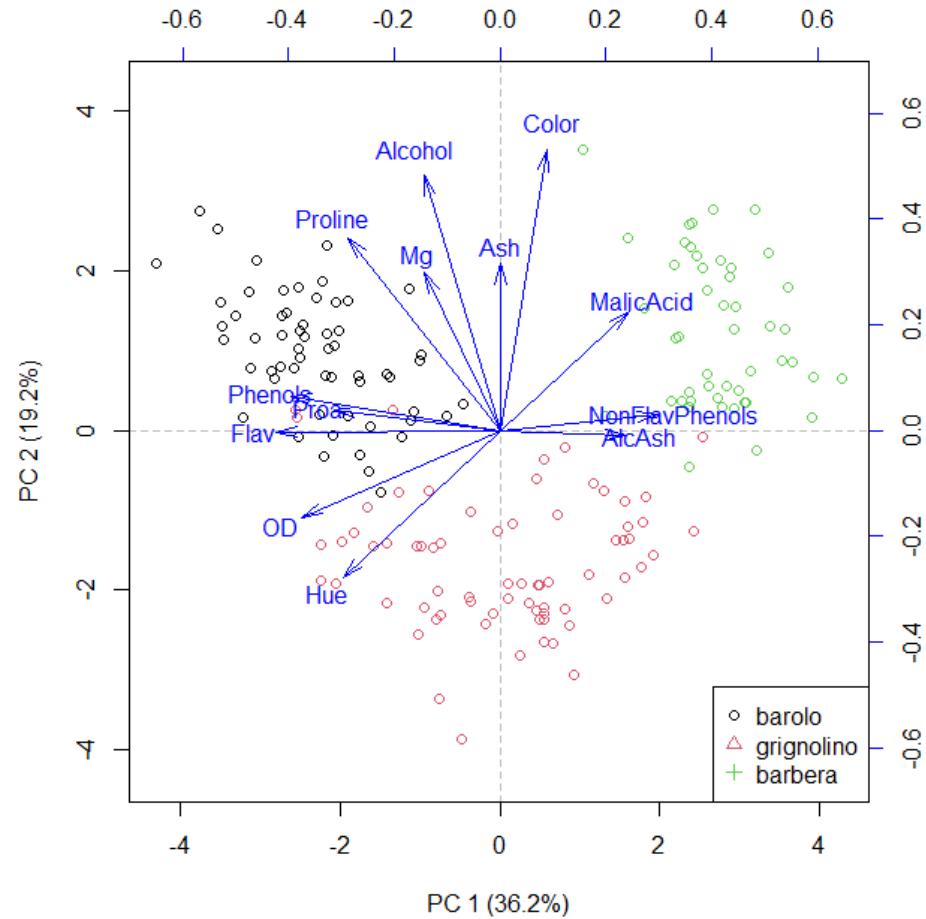
# Higher Order Loading Plots

```
par(mfrow=c(1,2))
loadingplot(wines.PC, pc=c(1,3), show.names= TRUE)
loadingplot(wines.PC, pc=c(2,3), show.names= TRUE)
```



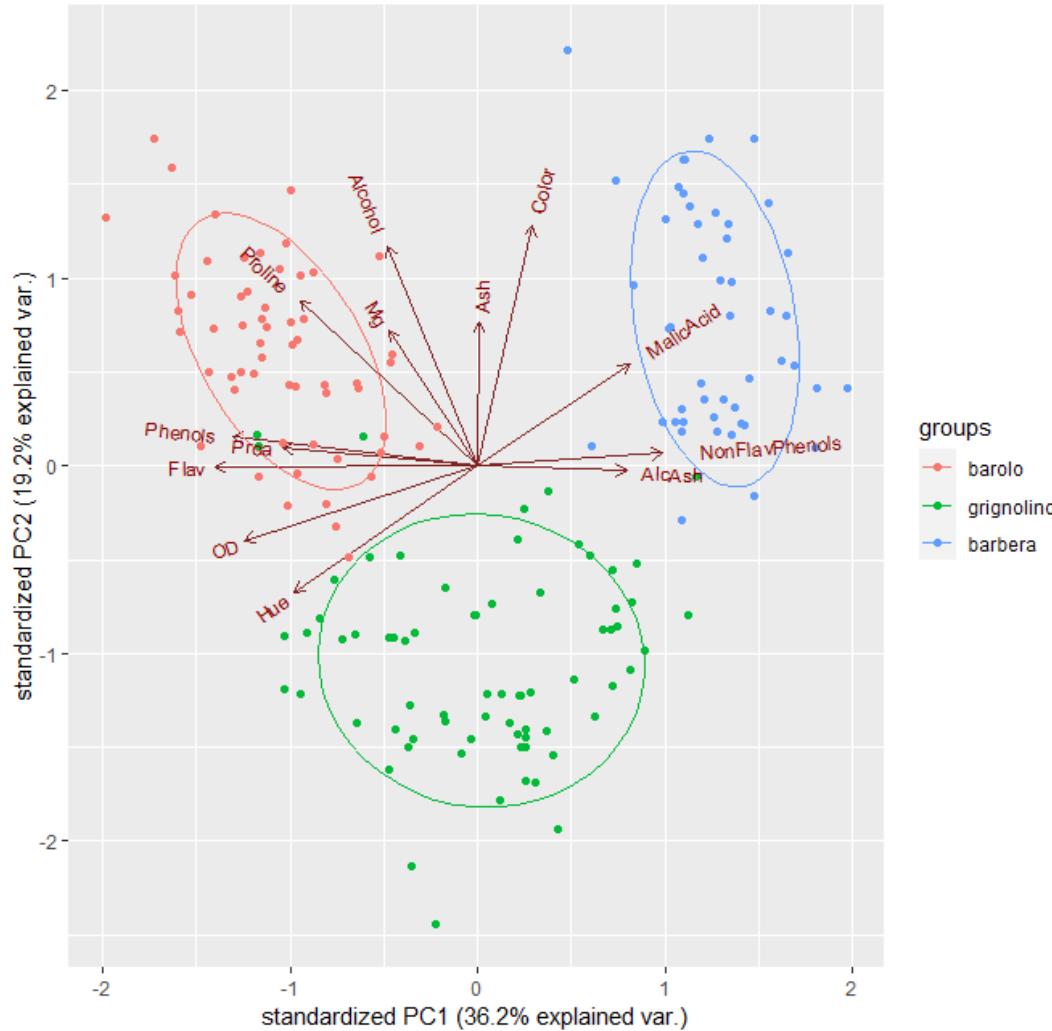
# The Biplot

```
biplot(wines.PC, score.col = vintages, show.names = "loadings")  
legend("bottomright", levels(vintages), col=1:3, pch=1:3)
```



# Alternative Biplot

```
ggbiplot(prcomp(scale(wines)), groups=vintages, ellipse=T)
```



# Diagnostic Plots – Residuals and Leverage

1. Large residual, far from the PCA space but central

1

Residual = orthogonal distance

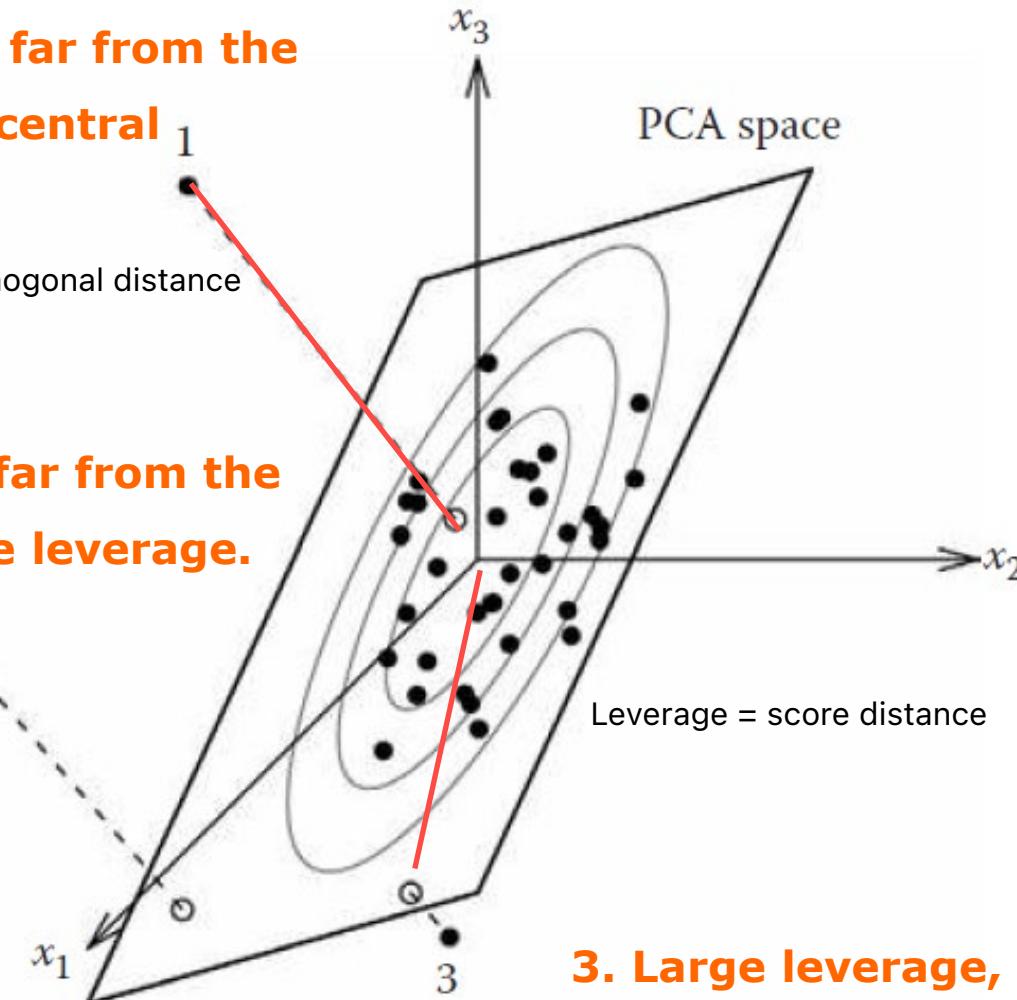
2. Large residual, far from the PCA space, large leverage.

2

Leverage = score distance

3

3. Large leverage, small residual

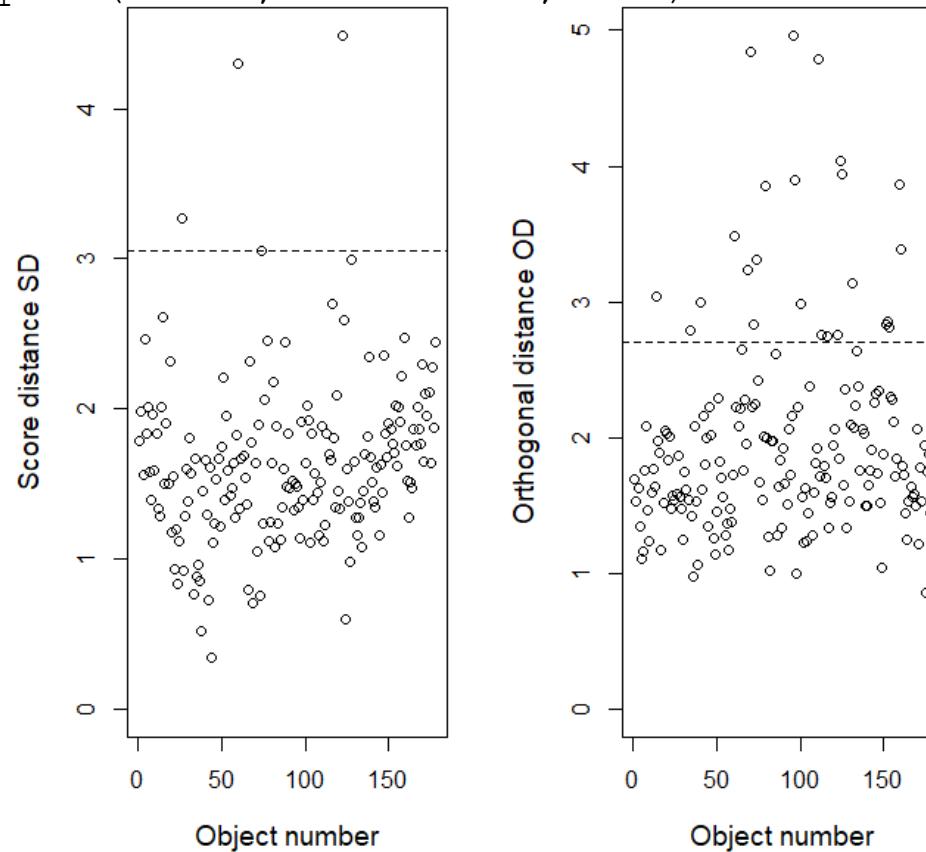


# Diagnostic Plots

Plotting distances for 3 principal components, leverage (left) and residual (right):

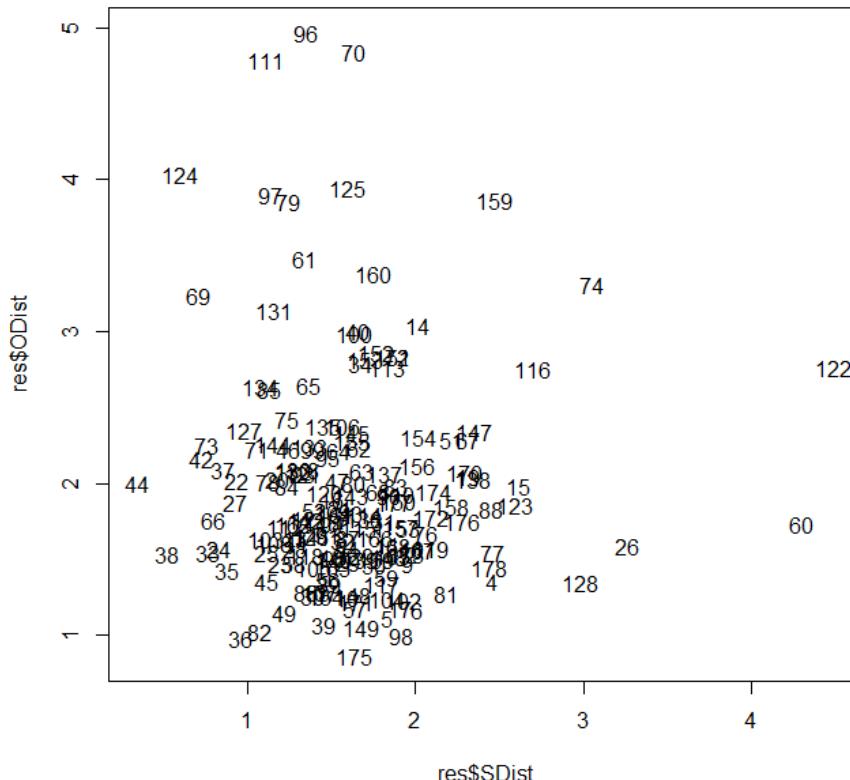
```
wines.PCA<- princomp(wines, cor = TRUE)
```

```
res<-pcaDiagplot(wines, wines.PCA, a=3)
```



# Diagnostic Plots

```
par(mfrow=c(1, 1))  
plot(res$SDist, res$ODist, type="n")  
text(res$SDist, res$ODist, labels=as.character(1:178))
```



No points have high leverage and also high residual

# Application Areas for PCR

Areas with complex, correlated data structures:

- Quantitative Finance;
- Neuroscience
- Spectroscopy(UV-VIS, x-ray, IR, NIR, NMR etc.)
- Questionnaires
- Image data
- Etc.

# Example: Jam

- Results from a taste survey for jam. 12 observations of raspberry jam, where berries are picked at four locations C1, C2, C3, C4, and harvested at three time points H1, H2, H3.
- A response variable Y is an averaged preference score from 0 to 9, given by 114 representative consumers.
- Explanatory variables are twelve sensoric variables, where trained sensoric panel members have evaluated these:

1) REDNESS	Redness	7) SOURNESS	Sourness
2) COLOUR	Colour intensity	8) BITTERNE	Bitterness
3) SHININES	Shininess	9) OFF.FLAV	Off-flavour
4) R.SMELL	Raspberry smell	10) JUICINES	Juiciness
5) R.FLAV	Raspberry taste	11) THICKNES	Thickness
6) SWEETNES	Sweetness	12) CHEW:RES	Chewing resistance

# Performing PCA

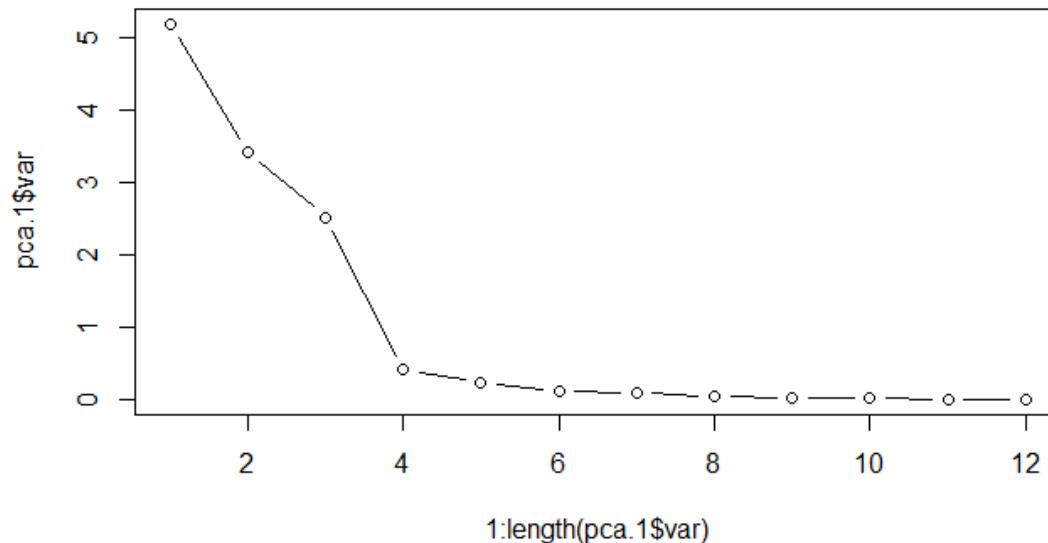
```
jam<-read.table("Data/Jam.txt", header=TRUE, quote="\r")  
# first column is names, last column is outcome:  
pca.1<- PCA(scale(jam[ , -c(1,14)]))  
summary(pca.1)
```

PCA model of a mean-centered matrix of 12 by 12  
Number of PCs to cover 90 percent of the variance: 3

	Var	Cumul. var.
PC 1	43.11404052	43.11404
PC 2	28.45577680	71.56982
PC 3	20.92326359	92.49308
PC 4	3.41038319	95.90346
PC 5	1.92478143	97.82825
PC 10	0.05286637	99.99396
>		

# Example: Jam - Skree Plot

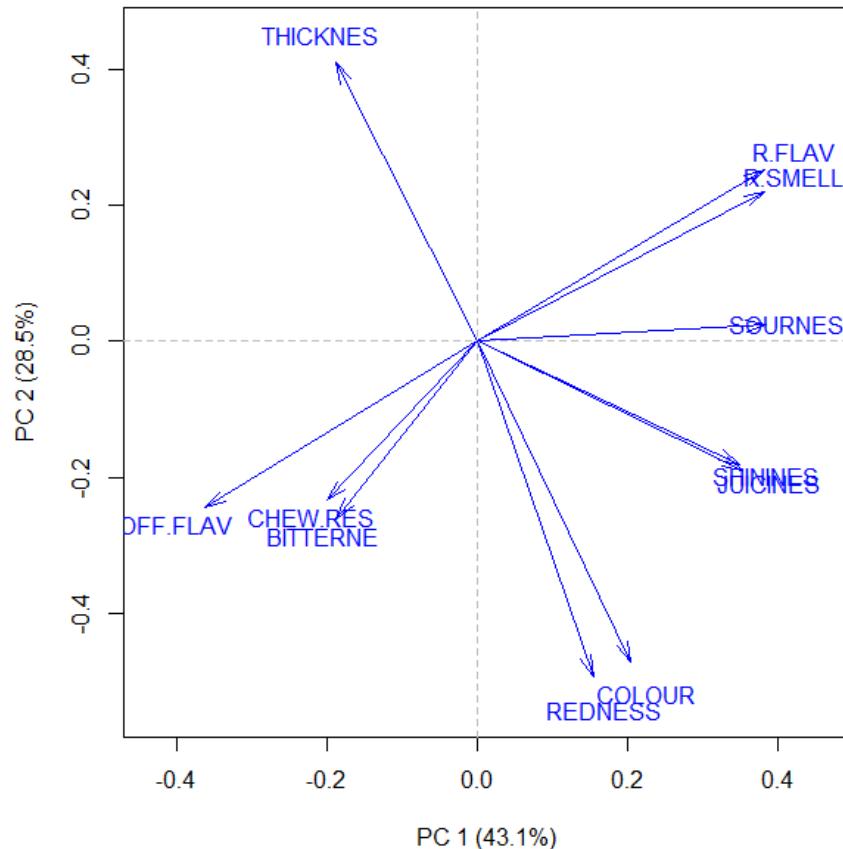
```
plot(1:length(pca.1$var), pca.1$var, type="b")
```



- We choose 3 principal components

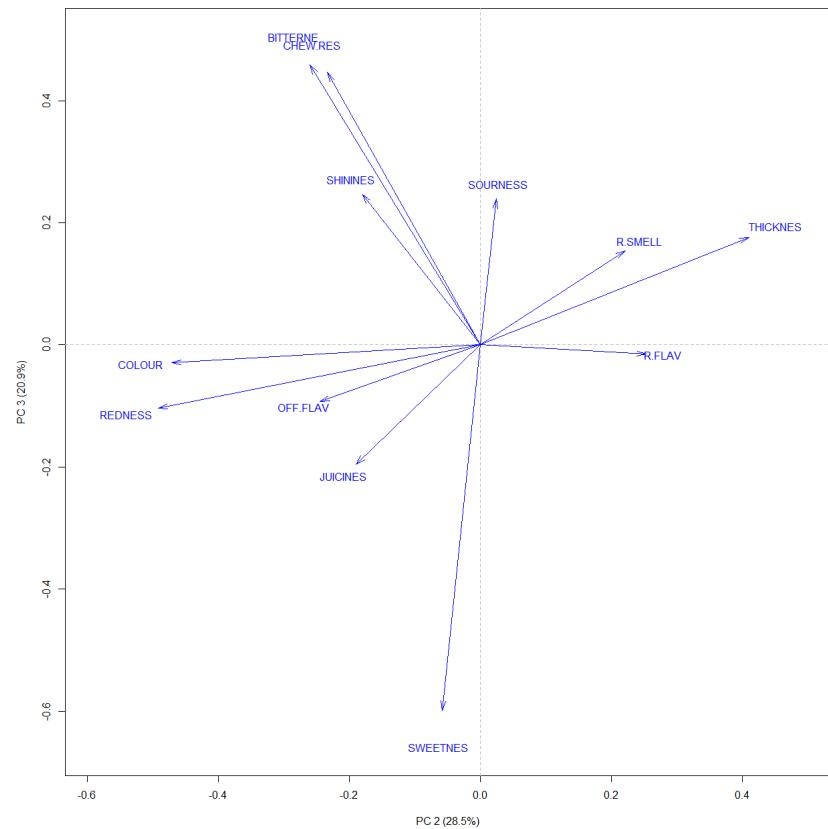
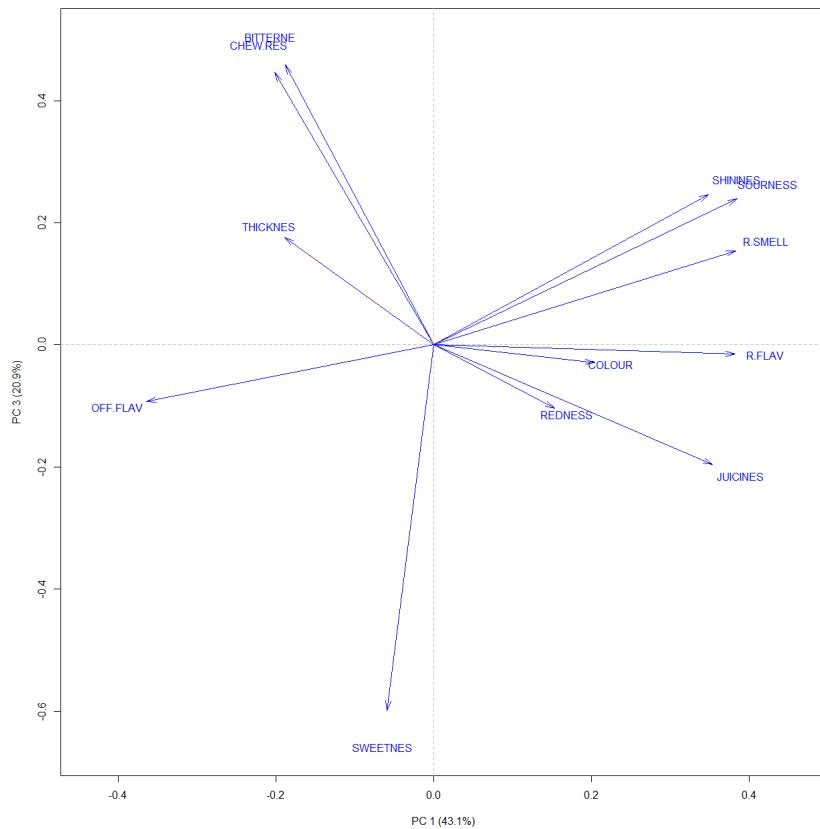
# Example: Jam – Loading plot

```
loadingplot(pca.1, show.names= TRUE)
```



# Example: Jam – Loading plot

```
par(mfrow=c(1,2))
loadingplot(pca.1, pc=c(1,3), show.names= TRUE); loadingplot(pca.1, pc=c(2,3), show.names= TRUE)
par(mfrow=c(1,1))
```



# Example: Jam – Loadings. Labelling the PCs



```
>loadings(pca.1)
```

Loadings:

	PC 1	PC 2	PC 3
REDNESS	0.154	-0.492	-0.104
COLOUR	0.204	-0.472	-0.029
SHININES	0.349	-0.180	0.246
R.SMELL	0.383	0.221	0.154
R.FLAV	0.382	0.253	-0.015
SWEETNES	-0.059	-0.058	-0.599
SOURNESS	0.385	0.024	0.239
BITTERNE	-0.188	-0.260	0.459
OFF.FLAV	-0.364	-0.245	-0.093
JUICINES	0.353	-0.190	-0.195
THICKNES	-0.189	0.409	0.176
CHEW.RES	-0.201	-0.234	0.446

**PC1:** Raspberry Feeling

**PC2:** Looks and consistency

**PC3:** Taste experience;  
sweetness (-) /  
bitterness (+)/ chewing  
resistance (+)

# PCR – Principal Component Regression

## Linear Regression of Response on PCs

- We wish to regress the Response variable  $Y$ , the averaged preference score, on scores of Raspberry feeling, Looks and Consistency, and Taste Experience
- We will skip the selection procedure for number of PCAs. In PCR, the selection of the number of PCs needs to involve the response through cross-validation. We will stick to 3 PCs.

# PCR – Principal Component Regression

## Linear Regression of Response on PCs

- We use the three PCs as explanatory variables in a multiple regression model, with  $Y$  as response.
- We thus consider the following model:

$$Y_i = \beta_0 + \beta_1 \cdot PC1_i + \beta_2 \cdot PC2_i + \beta_3 \cdot PC3_i + \varepsilon_i$$

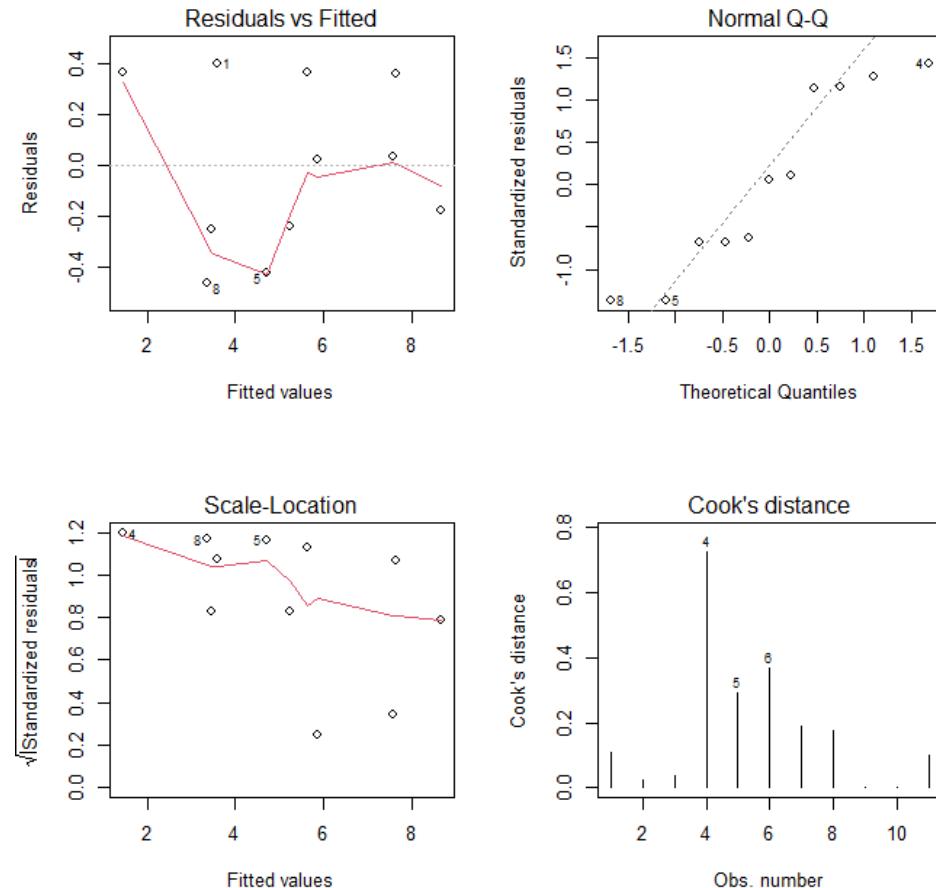
where  $Y_i$  is the preference, and  $PC1_i$ ,  $PC2_i$  and  $PC3_i$  are the principal component scores: our new, known and labeled explanatory variables. Finally, we have  $\varepsilon_i \sim N(0, \sigma^2)$ .

# PCR – Principal Component Regression Linear Regression of Response on PCs

Leaving out an outlier:

```
scores<-pca.1$scores
analysis<-lm(jam$PREFEREN[-12] ~
  . , as.data.frame(scores[-12 ,1:3]))

par(mfrow=c(2,2))
plot(analysis,which=1:4)
par(mfrow=c(1,1))
```



# PCR – Principal Component Regression

## Linear Regression of Response on PCs

Model Reduction:

```
drop1(analysis,test="F")
Single term deletions

Model:
jam$PREFEREN[-12] ~ `PC 1` + `PC 2` + `PC 3`
  Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>           1.093 -17.396
`PC 1`   1     0.5154  1.609 -15.148   3.2998   0.1121
`PC 2`   1   31.1115 32.205  17.816 199.2083 2.126e-06 ***
`PC 3`   1   15.7846 16.878  10.709 101.0694 2.066e-05 ***

analysis<-lm(jam$PREFEREN ~ . , as.data.frame(scores[,2:3]))
drop1(analysis,test="F")
Single term deletions

Model:
jam$PREFEREN ~ `PC 2` + `PC 3`
  Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>           5.148 -4.1546
`PC 2`   1     30.169 35.317 16.9536  52.739 4.754e-05 ***
`PC 3`   1     17.879 23.028 11.8215  31.255 0.0003383 ***
```

# PCR – Principal Component Regression

## Linear Regression of Response on PCs

```
> summary(analysis)$coef
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.0166667	0.2183356	22.976864	2.662062e-09
`PC 2`	-0.8962075	0.1234079	-7.262157	4.753719e-05
`PC 3`	-0.8045862	0.1439174	-5.590609	3.383175e-04

- Looks and Consistency and Taste Experience has an impact on the preference, while Raspberry Feeling does not;
- PC2: Evaluators appreciate **Redness** and **Colour**, but not **Thickness** of berries.
- PC3: Evaluators appreciate **Sweetness**, but not **Bitterness** and **Chewing Resistance**.

# Example: Image Analysis



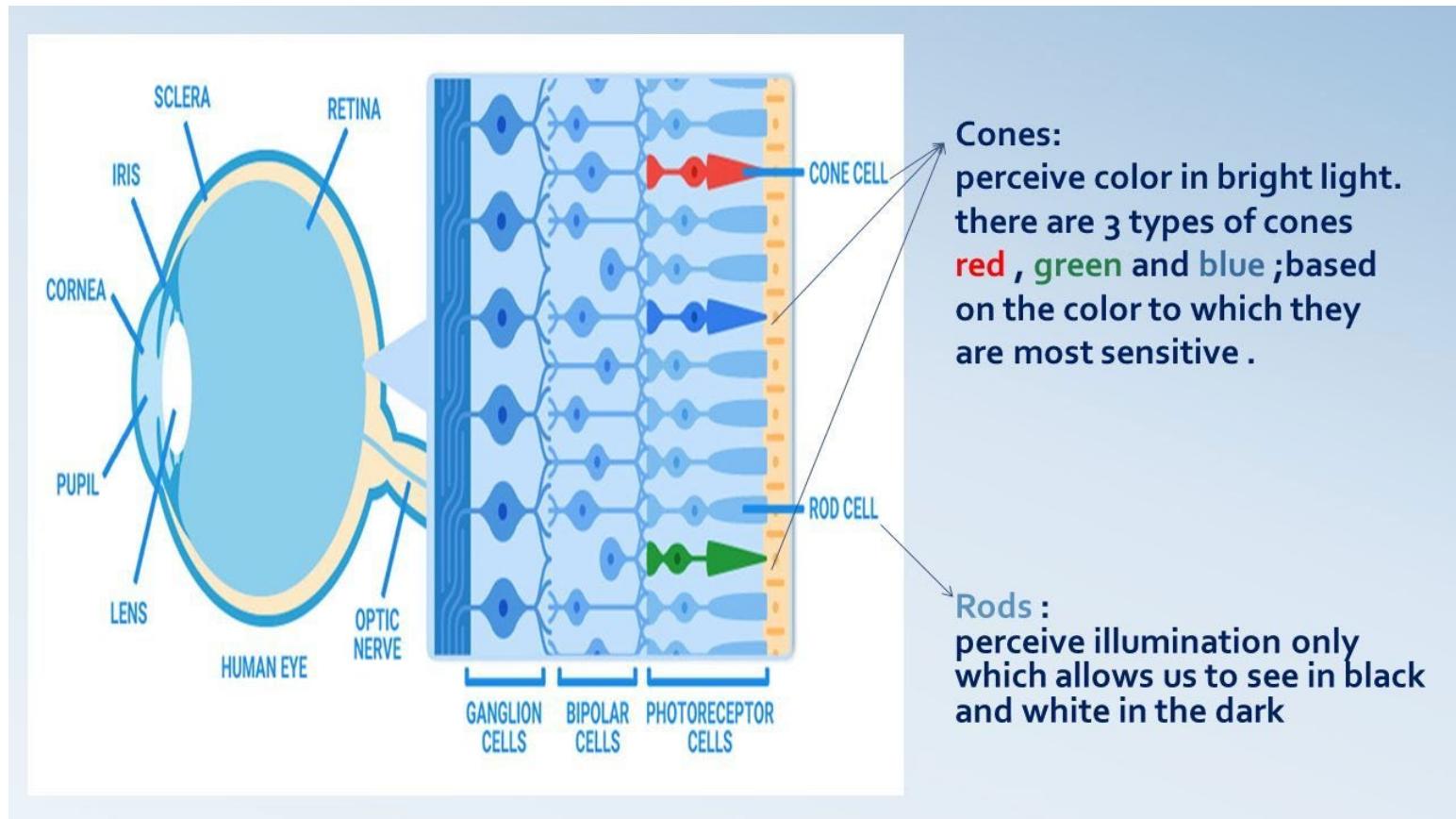
# Horse Analysis

```
horse <- readJPEG("Data/horse.jpg")  
  
ncol(horse)  
[1] 480  
  
nrow(horse)  
[1] 341  
  
#480*341 pixels - 480*341*3=491.040 numbers  
  
# array with 3 layers:  
str(horse)  
num [1:341, 1:480, 1:3] 0.796 0.796 0.796 0.796 0.796 ...
```

- Why 3 layers? RGB-coding; **Red-Green-Blue.**

# PCA Example: Image compression

Color pictures are coded with an intensity of **Red, Green and Blue** :



# Horse Analysis

```
str(horse)
num [1:341, 1:480, 1:3] 0.796 0.796 0.796 0.796 0.796 ...
```

```
red <- horse[, , 1]
green <- horse[, , 2]
blue <- horse[, , 3]
```

Running PCA on red, green and blue  
(with `prcomp()` for compatibility):



```
horse.red.pca <- prcomp(red, center = FALSE)
horse.green.pca <- prcomp(green, center = FALSE)
horse.blue.pca <- prcomp(blue, center = FALSE)
```

Gather PCA objects in one list:

```
rgb.pca <- list(horse.red.pca, horse.green.pca, horse.blue.pca)
```

# Horse Analysis

Indexing after the amount of principal components (max is the number of rows in the picture – 341):

```
>index<-c(3,6,9,12,15,18,50,100)
```

The function below reconstructs the picture from the first  $i$  principal components;

'x' are the principal components, 'rotation' are the loadings (see `?prcomp`), such that what is returned is the inversion from scores to standard coordinates (`scores%*%T-1`) which constitute the reconstruction. If T is all the PCs, the original picture is obtained.

```
my.reconstruct<-function(j) {  
  return( j$x[,1:i] %*% t(j$rotation[,1:i]))  
}
```

Create picture based on the first  $i$  principal components:

```
for (i in index) {  
  pca.picture <- sapply(rgb.pca,my.reconstruct,simplify = 'array')  
  writeJPEG(pca.picture,  
            paste("picture/horse_compressed_",i,"_components.jpg", sep = ""))  
}
```

# Horse Analysis – Results

3 PCs



6 PCs



9 PCs



12 PCs



15 PCs



18 PCs



50 PCs



100 PCs



Original

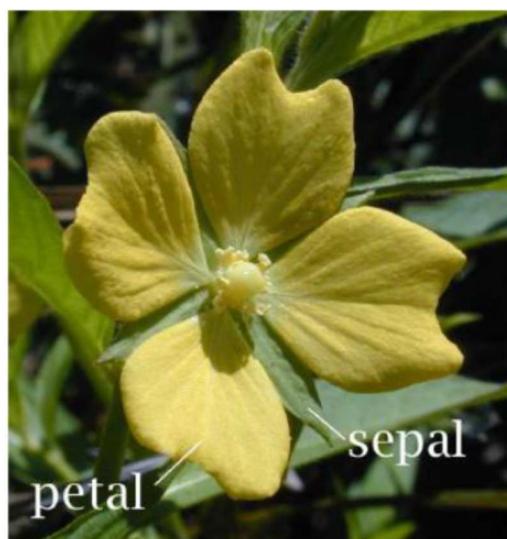


# Horse Analysis – Conclusion

- Further immediate use of PCs above 100 does not change the picture quality. It is hard to tell the difference to the original with 341 components...
- The compression rate for the picture with 100 PCs is 58% (ratio of bit sizes).
- One can use the picture with 100PCs, and save 58% storage.

## Iris data

The dataset “Fishers Iris data” is a classical dataset used in many examples. The dataset consists of 150 observations/objects, 50 Iris Setosa, 50 Iris versicolor and 50 Iris virginica. The flowers on these plant have been measured (mm). The measured variables are sepal length, sepal width, petal length and petalwidth (4 variables). The original hypothesis was that versicolor was a hybrid of the two other species.



1. Read in the data iris.txt.
2. Make descriptive plots.
3. Carry out a PCA on scaled data.
  - a. How many principal components are need for an adequate description of the variation in data?
  - b. Describe how you chose the number of principal components.
4. How can the principal components be interpreted?
5. Can the principal components be used to distinguish between the sorts?
6. Investigate if the model fits well: look at residuals and influential observations (leverage)

# Statistical Inference

Anders Stockmarr

Course developers: Anders Stockmarr, Helle Rootzen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 6th, 2025

## Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Outline

- Introduction
- Background
- Summary Statistics
- Statistical Modelling
- Estimation
- Test
- Power

# An Introductory Experiment

- How much time did you spend on social media yesterday? Guess?

# Example: Low Birth Weight

```
BWTdata <- read.csv2("Data/lowlwt.txt")  
head(BWTdata)
```

	ID	LOW	AGE	LWT	RACE	SMOKE	PTL	HT	UI	FTV	BWT
1	85	0	19	182	2	0	0	0	1	0	2523
2	86	0	33	155	3	0	0	0	0	3	2551
3	87	0	20	105	1	1	0	0	0	1	2557
4	88	0	21	108	1	1	0	0	1	2	2594
5	89	0	18	107	1	1	0	0	1	0	2600
6	91	0	21	124	3	0	0	0	0	0	2622

- Rows are *observations*
- Columns are *variables*

Hosmer & Lemeshow data 2000

## Example: Low Birth Weight - Statistical Analysis

The variable that we want to analyse is **BWT**. When we look at data, what can we say about the underlying statistical model that gave rise to these data?

- A normal model with parameters mean  $\mu$  and variance  $\sigma^2$ ?
- A log-normal model with similar parameters?
- something else?

Once we settle upon a (reasonable) model, we can proceed to conduct *statistical Inference*

# Example: Low Birth Weight - Statistical Analysis

- **Estimation:** With these 189 Birth weights, what can we say about the two unknown parameters  $\mu$  and  $\sigma^2$ ?  
Guesses on values, precision on guesses, possible dependencies on variables.
- **Test:** Does the mean  $\mu$  depend on Age? Does the variance  $\sigma^2$  depend on Race?
- **Prediction:** Given values of Age and Race, what expectation will we have to the birthweight? What will be the uncertainty of our expectation?

# Statistical Inference - Study planning

When planning a study, the planned statistical inference is a key element:

- Formulate a scientific question that you wish to answer with your study
  - Is the birth weight different for smoking and non-smoking mothers?
  - If so, how big is the difference?
  - These questions should be answered through statistical inference.

# Statistical Inference - Study planning

When planning a study, the planned statistical inference is a key element:

- Formulate a scientific question that you wish to answer with your study
  - Is the birth weight different for smoking and non-smoking mothers?
  - If so, how big is the difference?
  - These questions should be answered through statistical inference.
- Choosing the subjects
  - Who? Inclusion and Exclusion criteria. Availability?
  - How many?
  - these choices need to be made to enable the study to provide relevant results that may be used for inference on the scientific question.

# Statistical Inference - Study planning

When planning a study, the planned statistical inference is a key element:

- Formulate a scientific question that you wish to answer with your study
  - Is the birth weight different for smoking and non-smoking mothers?
  - If so, how big is the difference?
  - These questions should be answered through statistical inference.
- Choosing the subjects
  - Who? Inclusion and Exclusion criteria. Availability?
  - How many?
  - these choices need to be made to enable the study to provide relevant results that may be used for inference on the scientific question.
- What information are we going to collect? A *suitable response variable (birth weight)*. *Primary explanatory variables (smoking)* *Variables that may affect the outcome variable and/or cloud effects of the main explanatory variable (Age, gender, region etc)*.
- *All of these choices are made to be able to conduct relevant statistical inference*

# Background

- Probability - outcomes and sample spaces.
- Frequency
- Subjective probability
- Continuous variables

## Background - Probability

A *probability* is relative to an event. It is a number between 0 and 1, indicating uncertainty about if that event occurs or not.

- Probability of 0.5 - large uncertainty (toss of a coin)
- Probability close to 0 or 1 - small uncertainty (winning in lotto; surviving until tomorrow)
- *Outcomes* - elements of the *sample space*
- Coin toss: sample space  $\{Heads, Tails\}$
- Birth weight in grams - sample space  $(0; \infty)$

## Example: Toss of a Die

Tossing a die: Sample space and outcome probability

Outcome	1	2	3	4	5	6	Landing on the Edge
Probability	1/6	1/6	1/6	1/6	1/6	1/6	0

Altering tools of the experiment: Put lead opposite the six eyes, a common trick for sharpers:

Outcome	1	2	3	4	5	6	Landing on the Edge
Probability	2/15	2/15	2/15	2/15	2/15	1/3	0

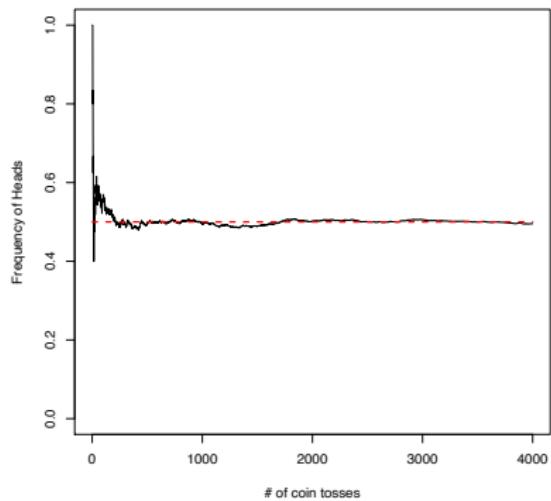
Altering circumstances of the experiment: Throw on a garden table:

Outcome	1	2	3	4	5	6	landing on the Edge
Probability	1/9	1/9	1/9	1/9	1/9	1/9	1/3

# Frequency

- The frequency is the relative proportion of an event.
- In contrast to a probability, a frequency contains an element of randomness.
- The **Law of Large Numbers**: If an experiment is repeated many times without changing the circumstances, the frequency of an event will converge to the probability.
- The **Frequency Interpretation** of probabilities: Limit values of frequencies

# Example: Frequency of Heads in Coin Tosses



- Frequency converges to 0.5.

# Subjective probabilities

- Reflect persons (subjects) individual assessment of probabilities from the own impression of circumstances and effects.
- Should NOT be confused with true probability.
- Example: Gut feeling, rule of thumb.
- Example: I think that the chance that FC Copenhagen will win the Danish soccer league is 80%, as they have been playing really well lately (when writing slides, FC Copenhagen was 1<sup>st</sup> in the Danish soccer league).

# Subjective probabilities

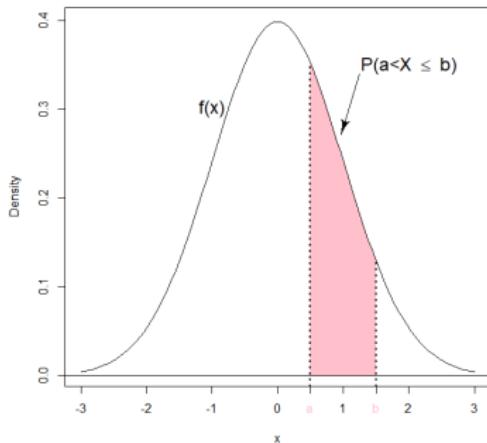
- Reflect persons (subjects) individual assessment of probabilities from the own impression of circumstances and effects.
- Should NOT be confused with true probability.
- Example: Gut feeling, rule of thumb.
- Example: I think that the chance that FC Copenhagen will win the Danish soccer league is 80%, as they have been playing really well lately (when writing slides, FC Copenhagen was 1<sup>st</sup> in the Danish soccer league).
- Typically, subjective probabilities contain an element of *subject bias*.

# Probabilities and Continuous Variables

- What does a probability mean when we are talking about a continuous variable  $X$ , with no upper limit to the number of possible values, like  $X = \text{birth weight}$ ?
- The sample space is  $(0; \infty)$ . Each possible outcome, say 3545 g, has probability 0 of occurring.
- In this case we represent the probability distribution by a **density function**, and discuss the probability of an interval;

$$P(a < X \leq b)$$

# Probabilities and Continuous Variables

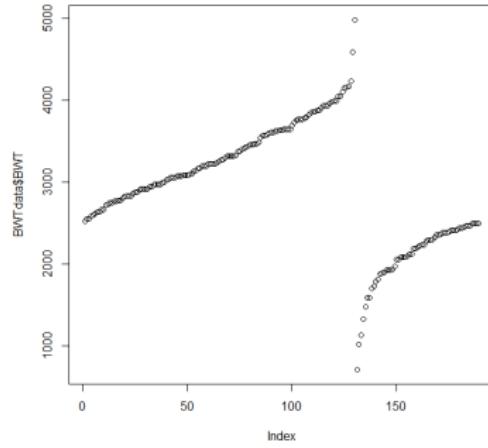


- Note that  $\lim_{\varepsilon \rightarrow 0} P(X \in [x - \varepsilon; x + \varepsilon]) / 2\varepsilon = f(x)$
- Thus the density function is a limit of (normalized) probabilities.

# Summary Statistics

- the first thing that we want to do is to plot the data:

```
plot(BWTdta$BWT)
```



Data are ordered according to low birthweight (< 2500 g), and birthweight.

# Summary Statistics - Location

After having plotted our data, we may want to get a further overview by calculating some simple statistics. Assume that we have observations  $y_1, \dots, y_n$  of a continuous variable  $Y$ .

The **location**, or **centre** of our data:

- Empirical mean:  $\bar{y} = \frac{1}{n} (y_1 + \dots + y_n)$
- Median: The middle observation when data are sorted and  $n$  is odd.  
The average of the two central observations when  $n$  is even.

# Summary Statistics - Measures of Variation

The empirical variance (or standard deviation) is a measure of how much the observations are spread out

- Empirical Variance  $s^2$ :

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$$

- Standard deviation (SD)  $s$

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

- Percentiles: the median is the 50% percentile.

# Summary Statistics - Measures of Variation

Percentile (and quartiles): Sort the data from smallest to largest:

**2.5% percentile**: First observation for which at least 2.5% of the observations are smaller or equal to, and at most 97.5% are larger.

**25% percentile**: First observation for which at least 25% of the observations are smaller or equal to, and at most 75% are larger.

**50% percentile**: First observation for which at least 50% of the observations are smaller or equal to, and at most 50% are larger.

**75% percentile**: First observation for which at least 75% of the observations are smaller or equal to, and at most 25% are larger.

**97.5% percentile**: First observation for which at least 97.5% of the observations are smaller or equal to, and at most 2.5% are larger.

# Summary Statistics - Measures of Variation

Empirical percentiles (and quartiles): Sort the data from smallest to largest:

**$p$ -percentile  $q_p$** : First observation for which at least the fraction  $p$  of the observations are smaller, and at most the fraction  $1 - p$  are larger

- The **quartiles** are  $q_{0.25}, q_{0.5}, q_{0.75}$ .  $q_{0.5}$  is usually equated with the median.
- Inter quartile range **IQR**:  $q_{0.75} - q_{0.25}$ . This is the size of the box in a box plot.
- $q_{0.025}$  and  $q_{0.975}$  spans an interval where **95% of the observations lie within**.

# Summary Statistics in R

```
mean(BWTdata$BWT)
```

```
[1] 2944.656
```

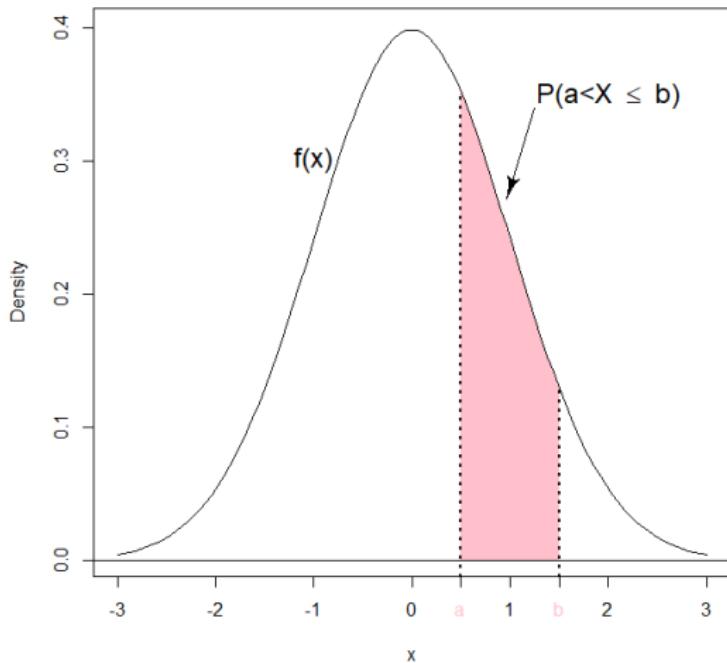
```
summary(BWTdata$BWT)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
709	2414	2977	2945	3475	4990

```
quantile(BWTdata$BWT,probs=seq(0,1,by=0.1),type=2)
```

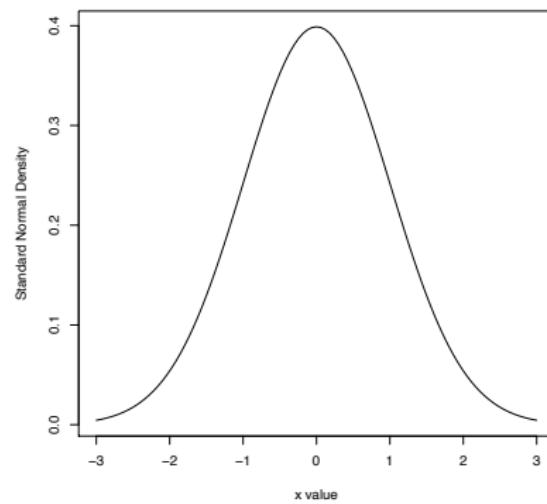
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
709	1970	2325	2495	2778	2977	3175	3374	3629	3884	4990

# Statistical Models from Probability Densities



# The Standard Normal Density

$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ , with mean= 0 and SD= 1.



# The Importance of the Normal Distribution

- **The Central Limit Theorem :** Averages of a large numbers of observations are approximately normally distributed, irrespectively of the (common) distribution that you start out with.

# The Importance of the Normal Distribution

- **The Central Limit Theorem :** Averages of a large numbers of observations are approximately normally distributed, irrespectively of the (common) distribution that you start out with.
- Because of this, it turns out that the Normal distribution is often a good approximation to real life distributions (perhaps after a transformation with the log, the square root or other...).

# The Importance of the Normal Distribution

- **The Central Limit Theorem :** Averages of a large numbers of observations are approximately normally distributed, irrespectively of the (common) distribution that you start out with.
- Because of this, it turns out that the Normal distribution is often a good approximation to real life distributions (perhaps after a transformation with the log, the square root or other...).
- The structure of the Normal distribution is mathematically tractable, and software has been developed for a lot of situations.

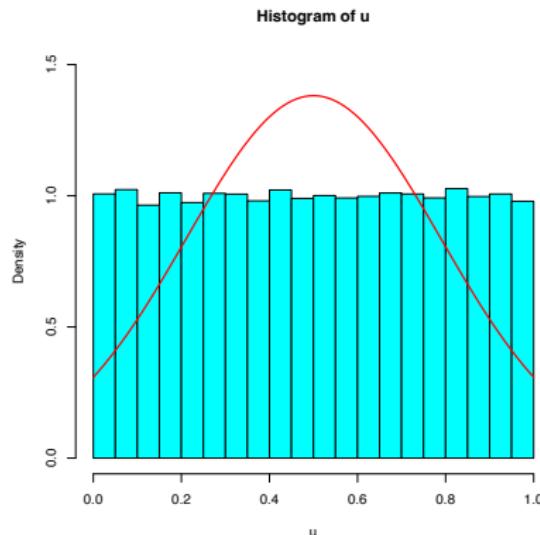
# A Non-Normal Distribution

A non-normal distribution could be the uniform distribution between 0 and 1;  $f(x) = 1, 0 < x < 1$ . Histogram of  $u$ :

```
u<-runif(2500)
```

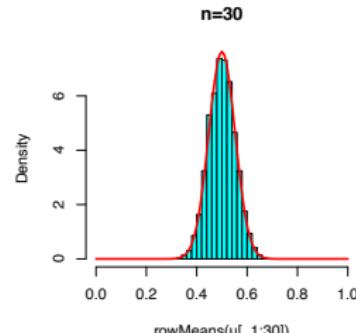
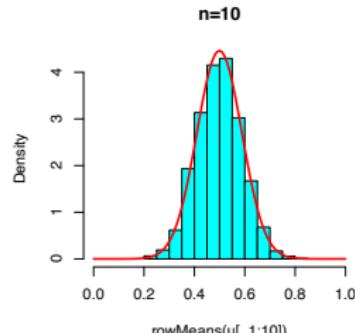
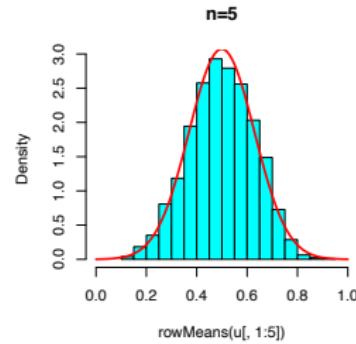
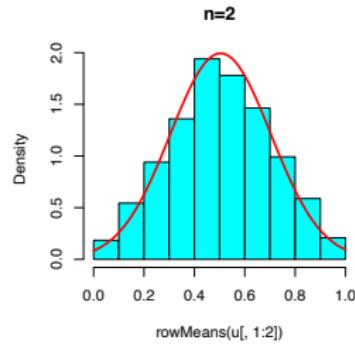
```
hist(u, col="cyan", probability=T, breaks=20, ylim=c(0,1.5))
```

```
curve(dnorm(x,mean=mean(u),sd=sd(u)),0,1,add=T,col="red",lwd=2)
```



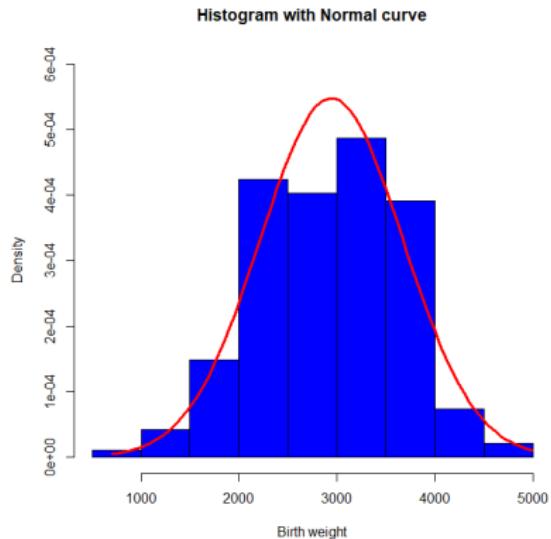
# A Non-Normal Distribution

Histograms for means of uniformly distributed variables:



## Normality Check

With normality, a data histogram should resemble normality. Birth weight data:

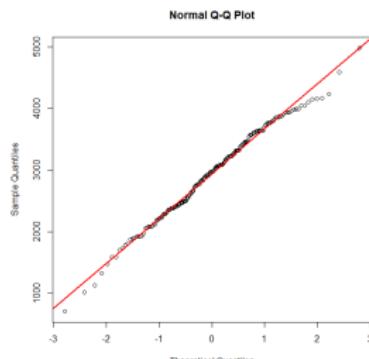


May be difficult to detect deviations from normality from a histogram.

# Normality Check

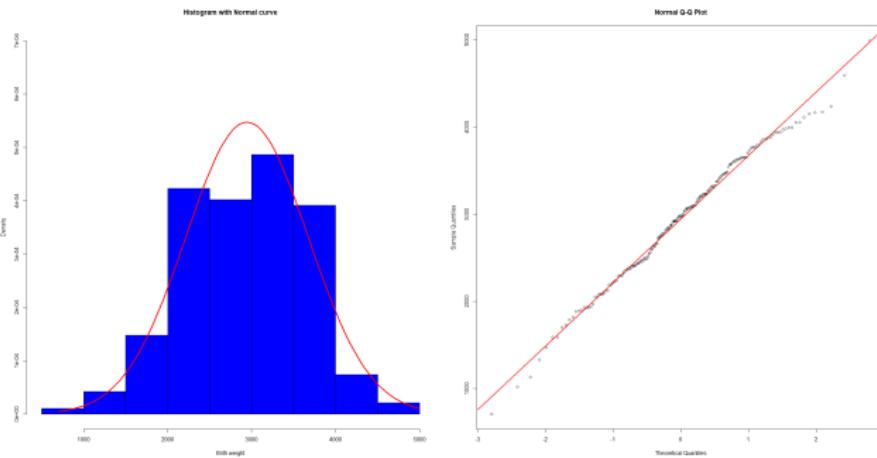
- Much better to use a *quantile-quantile plot* (qq-plot). Here, observed percentiles are plotted against percentiles from the normal distribution.
- If the empirical distribution is normal, a more or less straight line.
- If the data are non-normal, some deviation from a straight line should occur.

```
qqnorm(BWTdata$BWT)
lines((-3):3, ((-3):3)*sd(BWTdata$BWT)+mean(BWTdata$BWT) ,
      type="l", col="red", lwd=2)
```



# Statistical Model for the Birthweight data

Let  $Y_1, \dots, Y_{189}$  be the 189 registered birth weights in our dataset. We will assume that the 189 variables are **independent and normally distributed** with a common mean  $\mu$  and common variance  $\sigma^2$ .



# Statistical Models in General

In general, a statistical model consists of

- A family of distributions. In our example the 1-dimensional normal distributions (but there are others);
- the parameters that typically parametrizes the family of distributions.  
Eg. our example: The mean and variance in a normal distribution;

$$(\mu, \sigma^2) \in \mathbf{R} \times (0; \infty)$$

or the probability of ‘Heads’ in a coin toss:  $p \in [0; 1]$ .

Parameter interpretation: Effects/associations; e.g. the decrease in mpg per lbs/1000 weight.

# Exercise 1

- Set your working directory to where you keep your data for today.
- Load the cars dataset mtcars.txt.
- Describe the data
- Make plots of the variable miles per gallon, "mpg".
- Calculate summary statistics for mpg.
- Where would we expect most of the observations to be found?
- Calculate IQR and 0.025, 0.975 percentiles.

## Exercise 2

- Load the cars data set mtcars.txt.
- can we assume that miles per gallon are normally distributed?
- The variable “am” is 0 for cars with automatic transmission and 1 for cars with manual. Make a boxplot for the two levels of am.
- can we assume that miles per gallon are normally distributed for each level of am?

# Estimation

- Based on observations  $X_1, \dots, X_n$ , independent and each having a density function  $f_\theta(x)$ , we want to choose the parameters that fits our data the best.
- The density function measures the limiting probability of the data; The simultaneous (limiting) data probability is given by

$$f_\theta(X_1, \dots, X_n) := \prod_{i=1}^n f_\theta(X_i)$$

- Let's view this as a function of the parameter  $\theta$ , rather than the data  $X$ :

$$L_X(\theta) := f_\theta(X_1, \dots, X_n) = \prod_{i=1}^n f_\theta(X_i)$$

We choose the parameter that **maximizes the probability of the data**; ie. we maximize  $L_X(\theta)$  (or  $\ell_X(\theta) = \log(L_X(\theta))$ ).

# Estimation in the Birth Weights Example

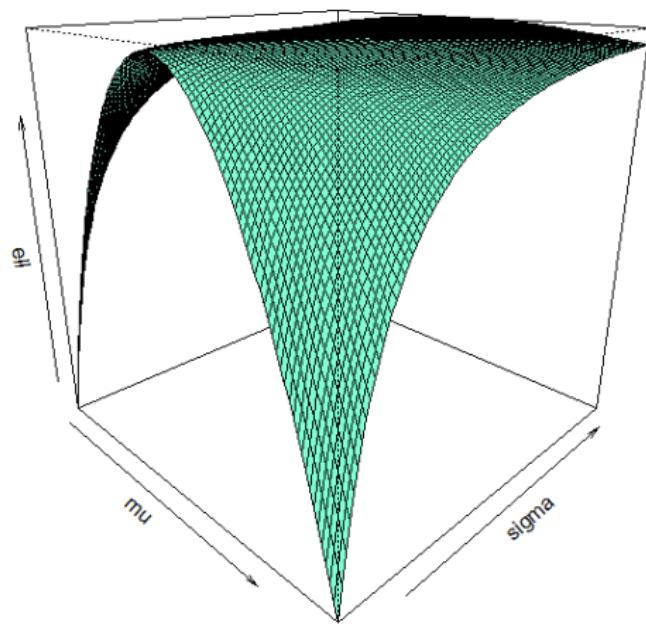
This is also called the one-sample problem (will be treated in detail tomorrow). 189 samples assumed from the same normal distribution:

$$Y_i \sim N(\mu, \sigma^2)$$

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|Y - \mu\|^2$$

# Birth Weights Example: The Log Likelihood Function

The log-likelihood function



# Estimation in the Birth Weights Example

$$\begin{aligned}\ell(\mu, \sigma^2) &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|Y - \mu\|^2 \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (n\bar{Y}^2 - 2n\bar{Y}\mu + n\mu^2)\end{aligned}$$

For fixed  $\sigma^2$  this is a 2nd order polynomial in  $\mu$  - easy to maximize!

- maximization point:

$$(\hat{\mu}, \hat{\sigma}^2) = (\bar{Y}, \text{Var}(Y))$$

- $\text{Var}(Y) = \frac{1}{n} \sum (Y_i - \bar{Y})^2$  is downwards biased as an estimator of  $\sigma^2$ ; the mean is not  $\sigma^2$ . We use the slightly bigger estimator  $s^2 = \frac{n}{n-1} \text{Var}(Y)$  defined on slide 19, to arrive at an unbiased estimate. Thus, theoretical mean and variance is estimated by empirical mean and variance.
- values from R:

```
>Y<-BWTdata$BWT; mean(Y); var(Y)
[1] 2944.656
[1] 531473.7
```

# Uncertainty of an Estimate

- We have  $\hat{\mu} = 2944.7g$ ; this is our best guess on the parameter, but (likely) not the true value. How uncertain is our estimate?
- Just as  $Y_i$  has a distribution, so does  $\hat{\mu}$ , as it is nothing but a constructed random variable.
- We can use the standard deviation of this distribution to construct measures of uncertainty. We call this the standard error of the estimate (or the Standard Error of the Mean (SEM)).

# The Standard Error of the Mean (SEM):

The Standard Error of the Mean is calculated as

$$SEM = SD(\bar{Y}) = \frac{SD(Y_1)}{\sqrt{n}}$$

SEM gets smaller as  $\sqrt{n}$  when  $n$  increases.

```
> SEM<-sd(BWTdata$BWT)/sqrt(length(BWTdata$BWT))  
> SEM  
[1] 53.02858
```

# Confidence Intervals

The interval

$$[\hat{\mu} - 1.97SEM; \hat{\mu} + 1.97SEM]$$

is a stochastic interval that (in this case) has a 95% probability of containing the true value  $\mu$ .

More on this later.

# Statistical Tests

Could we use a simpler model? Could one or more parameters in the chosen model be of a known value (often 0)? We want to attempt to simplify our model, through a statistical test of hypotheses, where we decide if a given hypothesis is supported by the data at hand.

Examples of hypotheses:

- Is the mean birth weight 3000g ( $\mu = 3000$ )?
- Is the mean birth weight the same for smokers and non-smokers ( $\mu_1 = \mu_2$  or  $\mu_1 - \mu_2 = 0$ )?
- Are the miles per gallon independent of the weight of the car (slope  $\beta = 0$ )?

# Tests Statistics

- We test a hypothesis through a **test statistic**. A test statistic measures the discrepancy between a hypothesis and an already accepted model (say, a normality assumption).
- An extreme test statistic says that the hypothesis fits the data badly.
- We will be studying whether an observed test statistic is more extreme than what could be expected by chance, assuming that the hypothesis is correct.

# A Test Statistic in the Birth Weight Example

We would like to test

$H_0$ : The mean of a birth weight is 3000g.

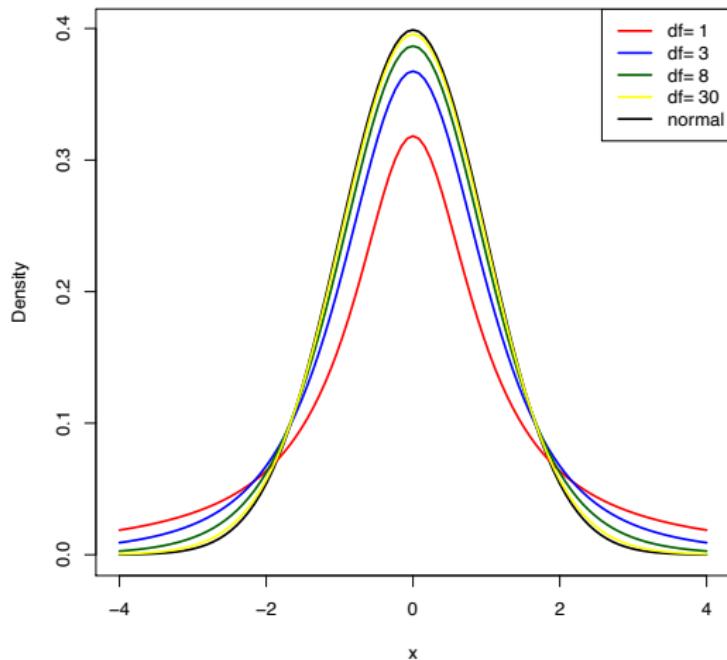
An obvious idea would be to assess whether our mean estimate,  $\hat{\mu} = \bar{Y}$  is close to 3000, ie. whether  $\bar{Y} - 3000$  is close to 0, relative to its uncertainty. Under  $H_0$  we have:

$$\bar{Y} \sim N\left(3000, \frac{\sigma^2}{n}\right), \quad \text{ie. } \frac{\bar{Y} - 3000}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0, 1)$$

We do not know  $\sigma^2$ , but we have an estimate  $s^2 = 531473.7$ . Replacing  $\sigma^2$  with its estimate  $s^2$  increases the uncertainty:

$$T = \frac{\bar{Y} - 3000}{\sqrt{\frac{s^2}{n}}} = \frac{\bar{Y} - 3000}{SEM} \sim t_{n-1}$$

# Densities from t-distributions



# A Test Statistic in the Birth Weight Example

We now have the test statistic

$$t = \frac{2944.7 - 3000}{53.03} = -1.0437$$

Is the value  $-1.0437$  extreme in a t-distribution with 188 degrees of freedom? We should calculate the probability

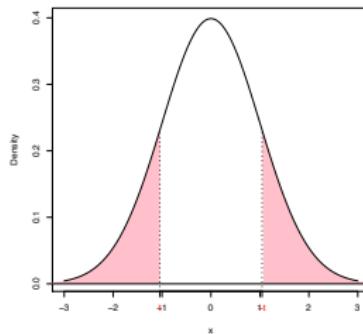
$$P(|t| > 1.0437)$$

This is straight forward in R:

```
> 2*pt(-1.0437,188)
[1] 0.2979644
```

# p value

- The **p-value** is the probability of observing what we have seen, or something worse (more extreme) if  $H_0$  is true.
- If what we have seen is very unlikely ( $p < 0.05$ ), We will not put faith in the hypothesis and **reject**  $H_0$ .
- We will not reject  $H_0$  for the Birth weight data ( $p=0.30$ ). *Statistical Inference*: The data *conforms* with a mean of 3000g at the 5% test level. But that doesn't mean that 3000g is the correct mean, only that the data can't support a rejection of it.



# Significance Level

- If the p-value is below 0.05, we say that the test is *significant at the 5% level*, and we reject the hypothesis  $H_0$
- If the p-value is above 0.05, we say that the test is *insignificant at the 5% level*, and we accept the hypothesis  $H_0$ .
- The choice of the threshold 5%, the *significance level*, is a standard, but not a convention that is applicable in all settings. Both the p-value and the significance level should normally be reported.

# Confidence Interval

The interval reported earlier:

$$[\hat{\mu} - 1.97SEM; \hat{\mu} + 1.97SEM] \approx [2840.05; 3049.264]$$

is a **standard confidence interval**:  $\hat{\mu} \pm qSEM$  where  $q$  is the 0.975 percentile in the t-distribution with 188 degrees of freedom. For high degrees of freedom,  $q$  will converge to 1.96. The 1.97 in the formula is rounded; the interval to the right is for the exact  $q$  value.

- Our best guess on  $\mu$  is thus 2944.66g, but we cannot reject that it may be between 2840.05 and 3049.264.
- For normally distributed data, the probability that this stochastic interval contains the true parameter  $\mu$  is 0.95 if the model is correct.
- For non-normal data, the probability is approximately 95% (The Central Limit Theorem), barring weird cases where the CLT does not apply (Cauchy distributed data etc.).

# A Test Statistic in the Birth Weight Example

The whole process in one go:

```
> t.test(BWTdata$BWT, mu=3000)
```

One Sample t-test

```
data: BWTdata$BWT
t = -1.0437, df = 188, p-value = 0.298
alternative hypothesis: true mean is not equal to 3000
95 percent confidence interval:
2840.049 3049.264
sample estimates:
mean of x
2944.656
```

# 'Exact' Confidence Intervals

- So-called **Exact confidence intervals** consists of the parameter values that a given test will accept as parameter value.
- A different but not necessarily more precise concept. Very handy in some situations (binomial distributions, more on this on Friday).
- The drawback is that the concept is dependent on a specific test.
- However, for the normal distribution, a standard CI and an 'exact' CI from the t-test **coincides**:

```
t.test(BWTdata$BWT, mu=2840.05)
t.test(BWTdata$BWT, mu=3049.264)
```

both give borderline significant p-values EXACTLY equal to 0.05.

## Exercise 3

- Load the mtcars data.
- Calculate the mean and SD for miles per gallon (mpg).
- Calculate a 95% CI for mean mpg.
- In which interval will you expect the true mean to be found?

## Exercise 4

- Load the mtcars data.
- Test the hypothesis that the mean mpg is 22.
- Which values would be acceptable at a 1% test level?

# Type 1 and Type 2 Error

When we are testing hypotheses, we can make (one of) two different types of errors:

**Type I:** Reject  $H_0$  when  $H_0$  is true.

**Type II:** Fail to reject  $H_0$  when the alternative  $H_1$  is true.

Standard notation:

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta$$

# Type 1 and Type 2 Errors

	Reject $H_0$	Fail to reject $H_0$
$H_0$ is true	Type I error ( $\alpha$ )	Correct acceptance of $H_0$ ( $1 - \alpha$ )
$H_0$ is false	Correct rejection of $H_0$ ( $1 - \beta$ )	Type II error ( $\beta$ )

$1 - \beta$  is called the **power**, and is the probability of rejecting a false hypothesis.

Difficulty:  $H_0$  can be wrong in many ways!

Usually one looks at different possible scenarios (what if  $\mu$  really was 4000g, with what probability could I detect that? How about 3500g etc.).

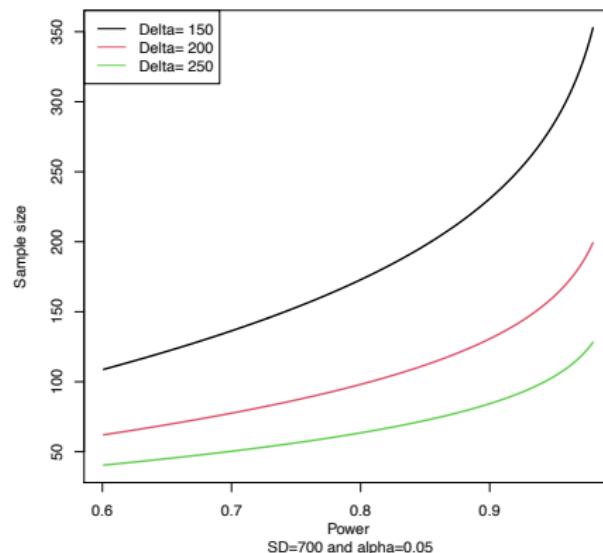
## Planning a Study: The Power

Suppose that you have decided on a test statistic. One need to assign values to four of the following five quantities to be able to calculate the last:

- The sample size  $n$ .
- The significance level  $\alpha$  of the test.
- A change in mean that you would want to detect ( $\mu_0 - \mu_1$ ).
- The population standard deviation  $\sigma$ .
- The power  $1 - \beta$ .

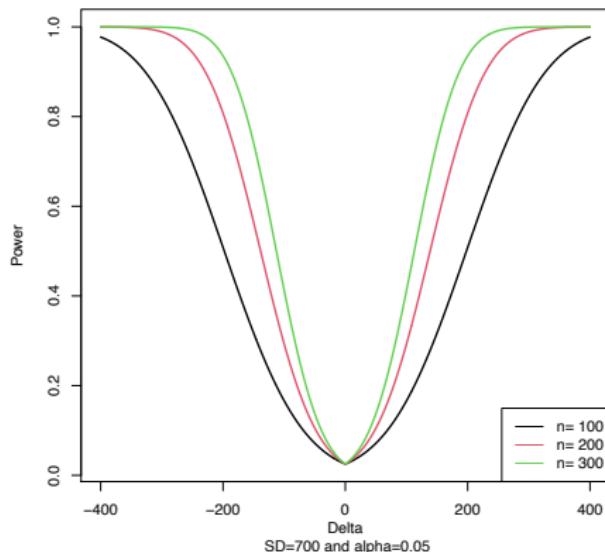
# The Power Function

The sample size depends on the power, and the difference one wants to detect:



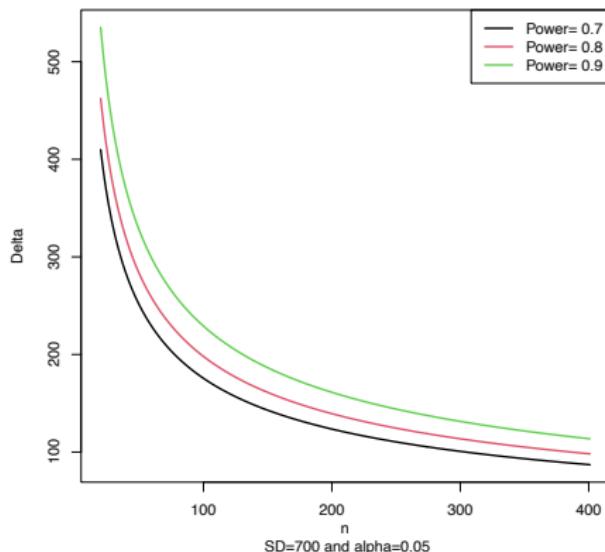
# The Power Function

The power depends on the sample size, and the difference one wants to detect:



# The Power Function

The difference that you can expect to detect depends on the sample size, and the power for your detection:



# Planning a Study: The Power

Important note:

- The power is calculated **before the study is carried out**, in the planning phase. The main reason is to find the study sample size (should it be 10, 100 or 10000....)
- Once the study has been completed, report confidence intervals.
- The power calculations presented here are for the simple t-test. In general, power calculations follow the same principles, but may be **much** more complicated. Powers, sample sizes and minimum differences may be best found through simulations of statistical models.

# Power Calculations for the t-test in R

If you wish to plan a study with a power of 0.8,  $\alpha = 0.05$  to detect a difference of 250, where you expect the  $sd = 750$ , then you will need  $n = 73$  subjects:

```
>power.t.test(power=0.8,delta=250,sd=750 , type='one.sample')
```

One-sample t test power calculation

```
    n = 72.58407
    delta = 250
    sd = 750
    sig.level = 0.05
    power = 0.8
    alternative = two.sided
```

## Power Calculations for the t-test in R

If you wish to plan a study with 150 subjects,  $\alpha = 0.05$  to detect a difference of 100, where you expect the  $sd = 750$ , then the power will be 37%:

```
>power.t.test(n=150,delta=100,sd=750 , type='one.sample')
```

One-sample t test power calculation

```
      n = 150
      delta = 100
      sd = 750
      sig.level = 0.05
      power = 0.3678721
      alternative = two.sided
```

# Power Calculations for the t-test in R

If you wish to plan a study with 150 subjects and a power of 0.8,  $\alpha = 0.05$ , where you expect the  $sd = 750$ , then the minimum difference that you can detect with such power will be  $delta = 173$ :

```
>power.t.test(n=150, power=0.8, sd=750 , type='one.sample')
```

One-sample t test power calculation

```
    n = 150
    delta = 172.677
    sd = 750
    sig.level = 0.05
    power = 0.8
    alternative = two.sided
```

# How to Set Up the Analysis

- Explore the form of the data: Make plots of the following types:
  - Histogram
  - Box plots
  - Scatter plots
- Find preliminary values of centes and deviations etc: Descriptive Statistics like
  - Tables
  - Summary Statistics
- Then: Well prepared, proceed to **Analyses** (main focus for the rest of the course):
  - Select model
  - Estimation
  - Test

# Steps in a Statistical Analysis

- **Estimation:** Which parameter values fit the observations best? How certain are we of our estimates?
- **Model check :** Are the assumptions on the underlying model fulfilled? Logically this should come first, but for practical reasons it comes after estimation.
- **Simplifying the model (test):** Is there a more simple model that fits the data nearly as well?

In practice, one can move back and forth between the first two steps a number of times, before a satisfying model is found.

## Exercise 5

In a one-sample setting with  $\alpha = 0.05$ :

- Calculate the sample size to get a power of 80% when trying to detect a difference of 2, when  $SD=6$  is expected.
- Calculate the power in a study planned to include 40 subjects, if we want to detect a difference of 3 and expect  $SD=6$ .
- What difference can we detect in a study with power of 80% , 45 subjects and  $SD=4$ ?

# T-test

Anders Stockmarr

Course developers: Anders Stockmarr, Helle Rootzen, Elisabeth Wreford  
Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 6th, 2024

## Plan for this week

Monday Statistical inference, and the **t-test**

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Outline

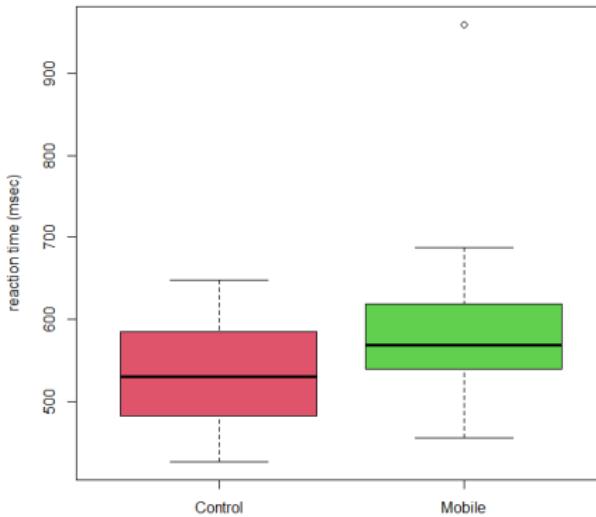
- Mobile Phone Example
- Exercise: MTcars example
- Paired Data
- Exercises - Mobile Phones

# Mobile Phone Example

- A study to investigate whether mobile phone use impairs drivers' reaction times
- 64 students randomly assigned to two groups (mobile phone or control).
- in a simulated driving situation, the participants were instructed to press the "brake" when they saw a red light flash.
- The mobile phone group were having a conversation, while the control group listened to radio.
- We want to investigate whether the reaction differs between the two groups.

```
Mobile.phone <- read.delim("Data/Mobiltelefon.txt")
```

# Mobile Phone Example

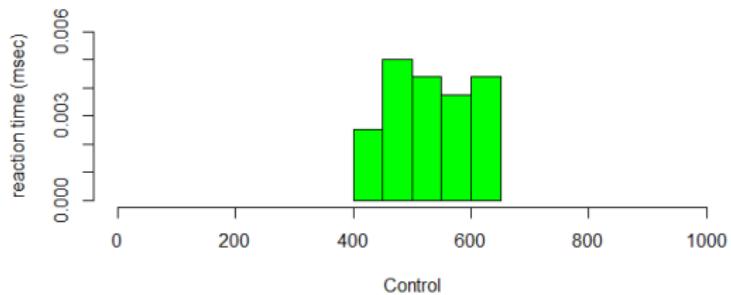
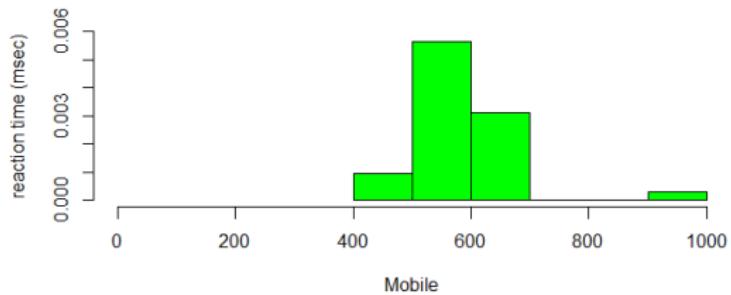


- What does this box plot show?

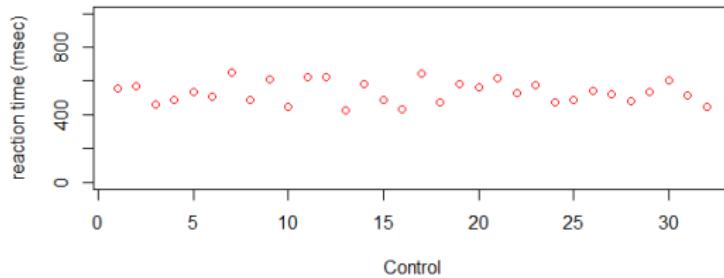
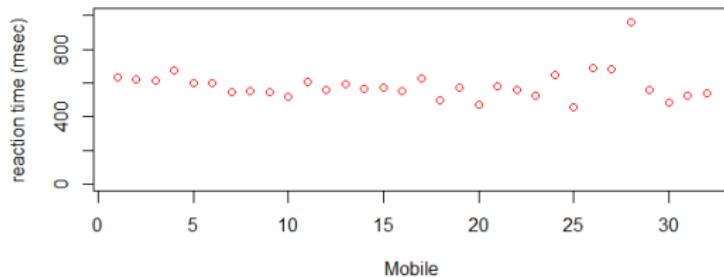
# How To Set Up the Analysis

- Draw
  - Histogram
  - Box plot
  - Scatter plot
- Descriptive Statistics
  - Tables
  - Summary Statistics
- Analyses
  - Select model
  - Estimation
  - Test

# Mobile Phone Example - Histogram



# Mobile Phone Example - Scatter plot



# Mobile Phone Example - Summary Statistics

- We have 64 observations of two variables: Time and Group (Mobile/Control).

```
by(Mobile$Time, Mobile$Group, summary)
```

Mobile\$Group: Control

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
426.0	483.5	530.0	533.6	585.2	648.0

---

Mobile\$Group: Mobile

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
456.0	540.5	569.0	585.2	617.0	960.0

```
by(Mobile$Time, Mobile$Group, sd)
```

Mobile\$Group: Control

[1] 65.35998

---

Mobile\$Group: Mobile

[1] 89.64606

# Statistical Model - Two Groups

## Model:

Two groups with (possibly) different normal distributions of reaction times:

Mobile phone group:  $Y_{1i} \sim N(\mu_1, \sigma_1^2), \quad i = 1, \dots, 31$

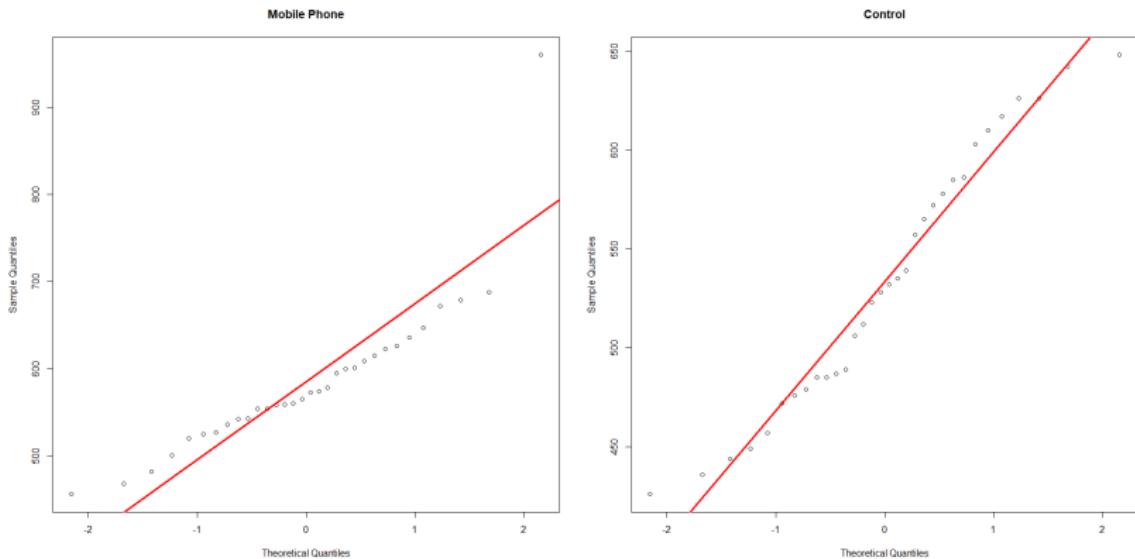
Control group:  $Y_{2i} \sim N(\mu_2, \sigma_2^2), \quad i = 1, \dots, 32$

## Assumptions:

- Normality as described - **how could this be violated?**
- Independence: all observations are independent - **how could this be violated?**
- Representativity: students represent a random sample - **how could this be violated?**

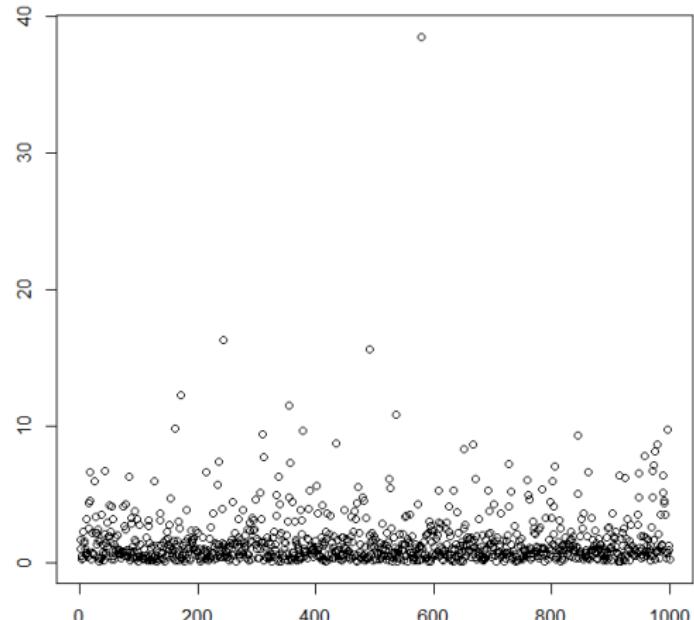
Hypotheses:  $H_0 : \mu_1 = \mu_2 \quad \text{vs.} \quad H_1 : \mu_1 \neq \mu_2$

# Normality assumption



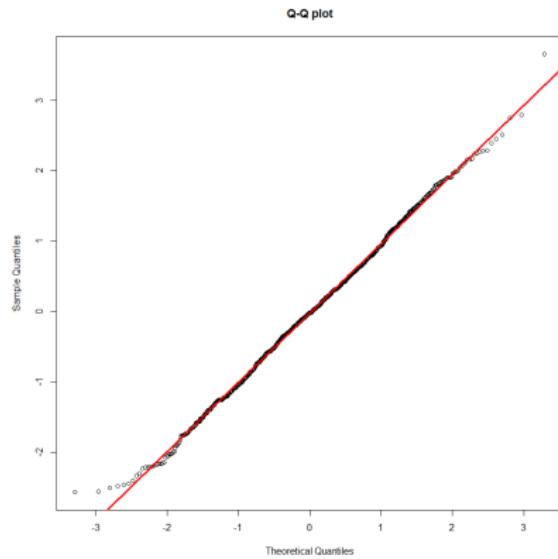
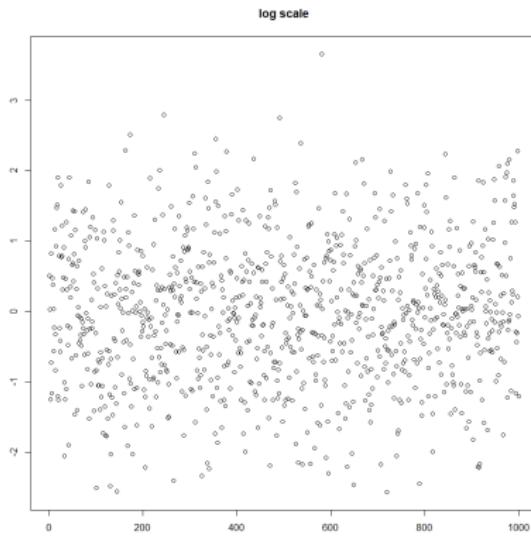
# Normality assumption - Outliers

Is there an outlier here?



# Normality assumption - Outliers

No outlier - normality on the log scale. Probability of reaching max is 12.2%, not cause for dismissal.



## Normality assumption - Outliers

Not so for the Mobile data:

```
Y<-Mobile$Time[Mobile$Group=="Mobile"]
max(Y)
[1] 960
2*(1-pnorm(960-mean(Y),sd=sd(Y))^length(Y))
[1] 0.0009284201
```

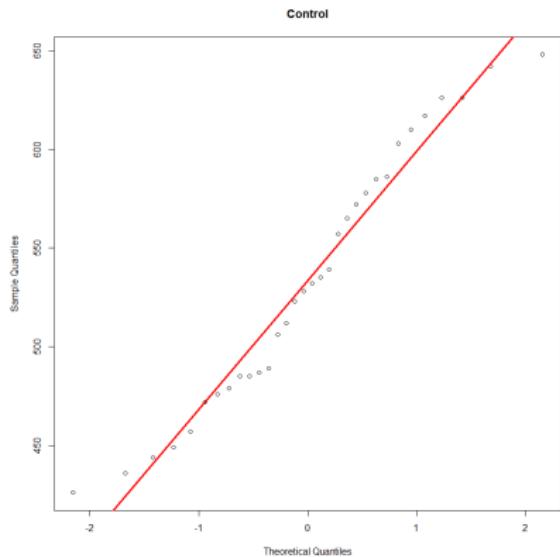
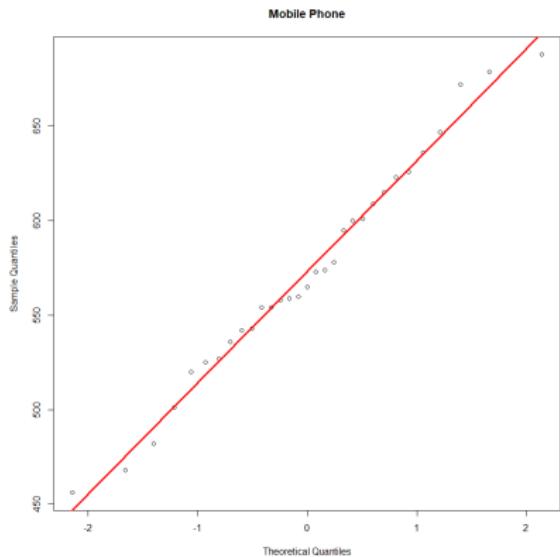
Without the variance-inflating observation:

```
Y<-Y[which(Y<900)]
2*(1-pnorm(960-mean(Y),sd=sd(Y))^length(Y))
[1] 1.584403e-09
```

Both numbers point towards an outlier.

# Normality assumption

Leaving out the outlier in data:



## Test of hypothesis $H_0$ vs. $H_1$

We use the *Welch t-test*, accounting for possibly unequal variances, and leaving out the outlier from the Mobile Phone group:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\widehat{s_e}(\bar{Y}_1 - \bar{Y}_2)} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

*Satterthwaite approximation* to the number of degrees of freedom  $\nu$ :

$$\nu = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}$$

## Test of hypothesis $H_0$ vs. $H_1$

- We have observed  $T = 39.50/15.67 = 2.52$
- The approximate degrees of freedom are found as  $\nu = 60.69052$ .
- Values critical for  $H_0$  are numerically large values. The p-value is the probability of observing something more critical than the actual observation of T.
- calculating the p-value in R:

```
2*(1-pt(T,df=nu))
[1] 0.01432928
```
- The p-value is thus below the standard test level of  $\alpha = 0.05$ . At the 0.05 test level, the data do not support that the control group and the Mobile Phone group have similar reaction times ( $p=0.01$ ).

## Estimated Difference

- We estimate the difference in reaction times as follows:

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{Y}_1 - \bar{Y}_2 = 573.0968 - 533.5938 = 39.5030 \text{ msec.}$$

- What is the uncertainty of this estimate?

$$\widehat{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\widehat{var}(\bar{Y}_1) + \widehat{var}(\bar{Y}_2)} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} = 15.6667$$

# Confidence Interval for Estimated Difference

Standard confidence interval:

$$\hat{\mu} \pm q_{0.975} \times sd(\hat{\mu})$$

where  $q_{0.975}$  is the 97.5% percentile in the relevant t-distribution. In our case, with  $\nu = 60.69$  which gives  $q_{0.975} = 1.9998$ :

$$CI(\mu_1 - \mu_2) = [39.50 - 2 * 15.67; 39.50 + 2 * 15.67] = [8.17; 70.83]$$

Compare with the tighter approximative interval, where we use normal uncertainty of 1.96 rather than the  $t_\nu$  uncertainty of 2:

$$[39.50 - 1.96 * 15.67; 39.50 + 1.96 * 15.67] = [8.80; 70.21]$$

A fairly good approximation here.

## t-test in R

```
t.test(Y1,Y2)
```

Welch Two Sample t-test

data: Y1 and Y2

t = 2.5215, df = 60.691, p-value = 0.01433

alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:

8.172203 70.833845

sample estimates:

mean of x mean of y

573.0968 533.5938

# Similar Standard Deviation

- In many situations it makes sense to have an extra model assumption:

$$\sigma_1^2 = \sigma_2^2$$

ie. the variation in the two groups are identical. The model in this case is thus

Mobile phone group:  $Y_{1i} \sim N(\mu_1, \sigma^2), \quad i = 1, \dots, 32$

Control group:  $Y_{2i} \sim N(\mu_2, \sigma^2), \quad i = 1, \dots, 32$

# Similar Standard Deviation

- The assumption of similar standard deviation model leads to a different test statistic, where the empirical variances are pooled:

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 3877.745$$

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{s^2}{n_1} + \frac{s^2}{n_2}}} = 2.52$$

## Similar Standard Deviation

- This case of equal variances is much simpler, and no approximations to the number of degrees of freedom for the t-test is needed: It is  $n_1 + n_2 - 2 = 61$ . The test provides a higher power than the Welch t-test, because the model has one less parameter to estimate.
- However, if the difference in variance is considerable, the similar variance t-test may be misleading. Without thorough investigations, the Welch version of the t-test should be used. In particular, for small sample sizes, it may be difficult to detect differences in variation with sufficient strength.
- In the present case we have estimates  $s_1^2 = 3470.424$  and  $s_2^2 = 4271.926$ . The data does not support that these values should be different ( $p=0.31$ ). More on this on Thursday.

# Similar Standard Deviation

The t-test with assuming equal variances:

```
t.test(Y1,Y2,var.equal=TRUE)
```

Two Sample t-test

data: Y1 and Y2

t = 2.5172, df = 61, p-value = 0.01447

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

8.123071 70.882978

sample estimates:

mean of x mean of y

573.0968 533.5938

# Sensitivity Analysis

- We found an outlier in the Mobile Phone group, with a reaction time of 960 msec - nearly a second!
- We removed the outlier, as it clearly wasn't comparable to the remaining data - likely a student that wasn't up for the task and was thinking about something else.
- Without the removal of the outlier, we would be violating the normality assumption, and the t-test would no longer be valid.

# Sensitivity Analysis

- **Problem:** Are we testing on an idealized population, without the proper association to reality?
- To supplement our analysis, we will investigate how to include the outlier in an analysis, **to see if the presence affect our conclusion.**
- The interpretation is that we allow for a fraction not being observant at all, not being "up for the task".

# Sensitivity Analysis

- **Test statistic:** We still use the t-test statistic, but this time we INCLUDE the outlier in the mobile phone group  $T_1$ :

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{\hat{s}e(\bar{Y}_1 - \bar{Y}_2)} = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{\frac{S_1^2}{n_1-1} + \frac{S_2^2}{n_2-1}}} = 2.63$$

- Estimated group difference: 51.59. Much bigger than without the outlier, but the variance is also increased.
- The assumption of normality is seriously violated however, and we have to resort to other means to evaluate the statistic T.

## Sensitivity Analysis - the permutation test

- Consider the Null hypothesis that we wish to test:

$$H_0 : \mu_1 = \mu_2$$

- Under  $H_0$ , the mean in the two groups is identical.
- Thus, if we resample our reaction times in two new groups, the two groups will still theoretically have the same mean.
- We use this technique to investigate the sampling variation may be the cause of the difference in means.

# Sensitivity Analysis - the permutation test

- Strategy:
  - resample the 64 data points in two new random groups.
  - calculate the test statistic  $T$  for the two new groups.
  - Compare the new  $T$  statistic with the original, to see if it is bigger.
  - Repeat the above a large number of times.
  - Use the fraction of  $T$  statistics bigger than the original as the p-value, as this simulates the probability of getting a more extreme result than our original due to sampling variation.

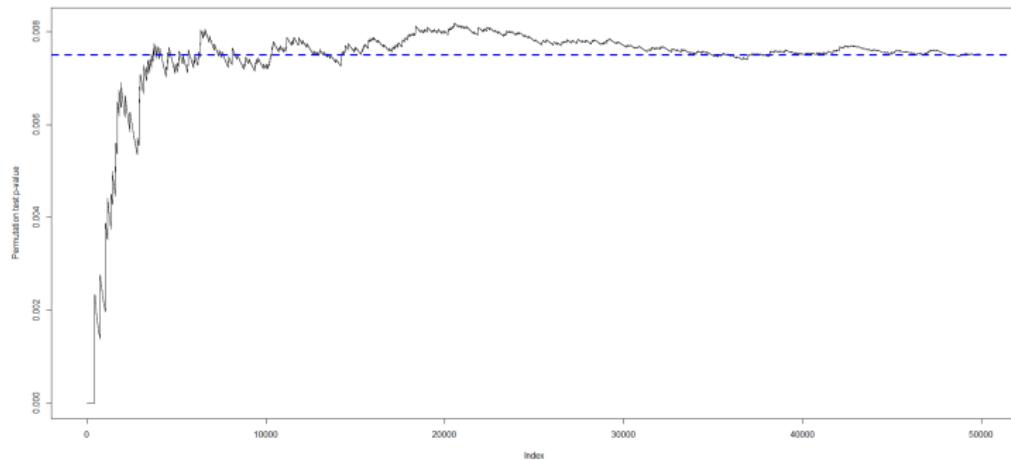
# Sensitivity Analysis - the permutation test

R code:

```
my.reaction.times<-Mobile$Time
my.t.statistics<-numeric(50000)
for(i in 1:50000){
  index<-sample(1:64,32)
  Y1.temp<-my.reaction.times[index]
  Y2.temp<-my.reaction.times[-index]
  my.t.statistics[i]<-t.test(Y1.temp,Y2.temp)$statistic
}
my.p.value<-length(my.t.statistics[abs(my.t.statistics)>T])/50000
my.p.value
[1] 0.0075
```

# Sensitivity Analysis - the permutation test

- The permutation test supports the previous conclusions - but have we performed enough simulations?



## Exercise: MTcars Example

We want to compare miles per gallon for cars with and without manual transmission.

- Access the builtin data set mtcars with the command  
`data(mtcars)`
- Plot the Miles per Gallon for the two groups (`am=0` or `1`).
- Formulate the relevant hypothesis to test, and the alternative.
- Are the underlying assumptions for the t-test fulfilled?
- What is the estimated difference in mpg, and the corresponding 95% confidence interval?
- What can we conclude about  $H_0$ ?

## Paired t-test: The Glucose Data

- The *Glucose12* data set features data from two different methods to measure blood glucose.
- 73 subjects have had their levels of blood glucose measured with both methods.
- We would like to know if the two methods measure the same?

```
head(glucose12)
```

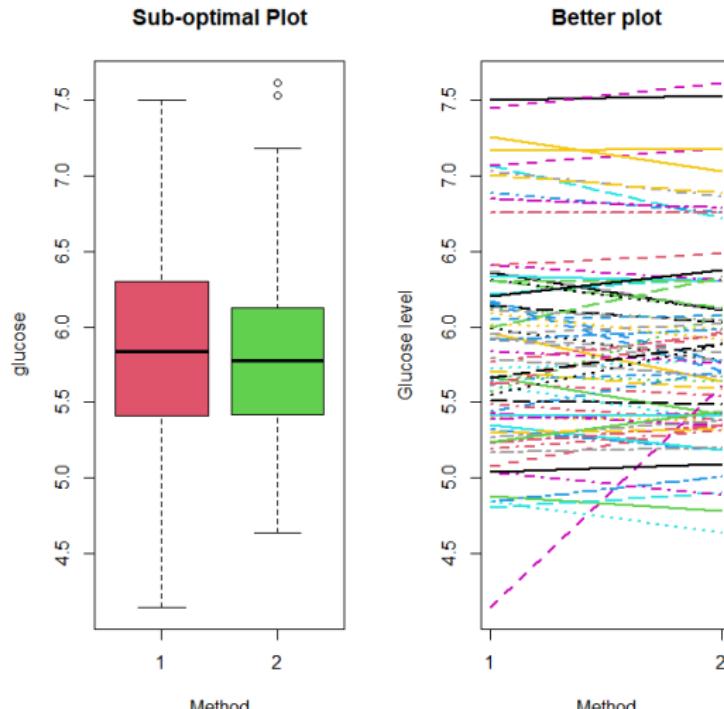
	subject	Glucose1	Glucose2
1	1	6.36	6.11
2	2	5.08	5.35
3	3	6.12	6.04
4	4	5.65	5.69
5	5	7.07	6.72
6	6	5.43	5.34

## Paired t-test: The Glucose Data

- For paired data, each subject act as its own control.
- Greatly reduces person-to-person variation, and may give a much more powerful test.
- We will look for differences between the two types of measurements.
  - Are the differences independent of the size of the measurements?
  - Do we need to look at relative differences (log-transform data)?
- Overall, we wish to investigate if the difference between the two types of measurements is 0.

# Paired t-test: The Glucose Data

Why is the first plot not optimal?



# Model for Paired Data

## Data:

Method 1:  $X_i, i = 1, \dots, 73$

Method 2:  $Y_i, i = 1, \dots, 73$

Difference:  $D_i = X_i - Y_i, i = 1, \dots, 73$

**Model:** Differences are assumed independent and identically distributed with  $D_i \sim N(\mu, \sigma^2)$ -

## Assumptions

- Assumptions on differences as above;
- NO further assumptions on X and Y.

## Hypotheses

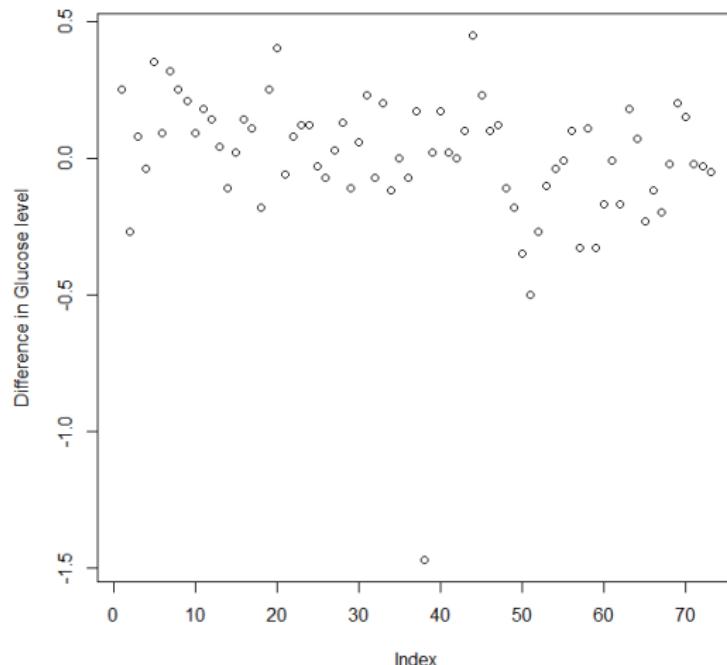
$$H_0 : \mu = 0 \quad H_1 : \mu \neq 0$$

# Assumptions for the Paired t-test

- Independence - consider circumstances.
- Same variances - look for patterns in the scatter plot of differences. IF data are normally distributed: Look at mean of methods vs. differences in methods.
- Normality - consider the qq-plot vs. the normal distribution.

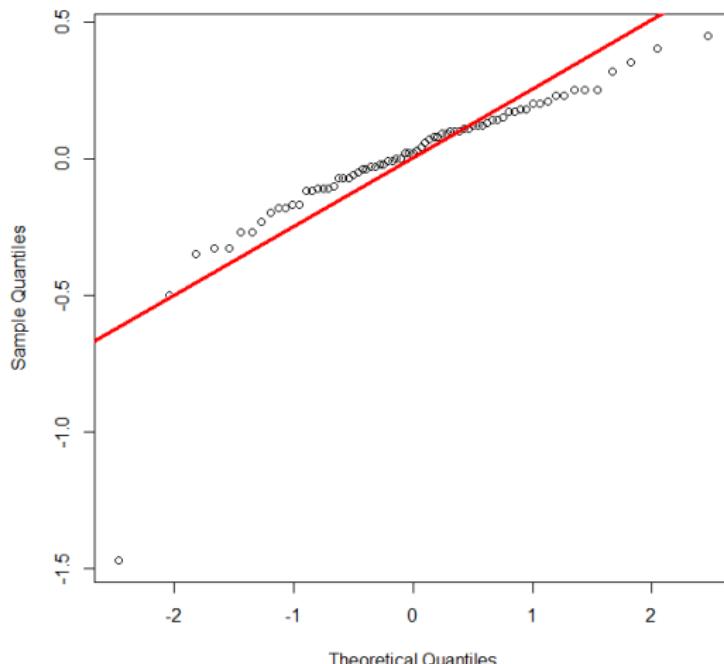
# Assumptions for the Paired t-test

Same variances:



# Assumptions for the Paired t-test

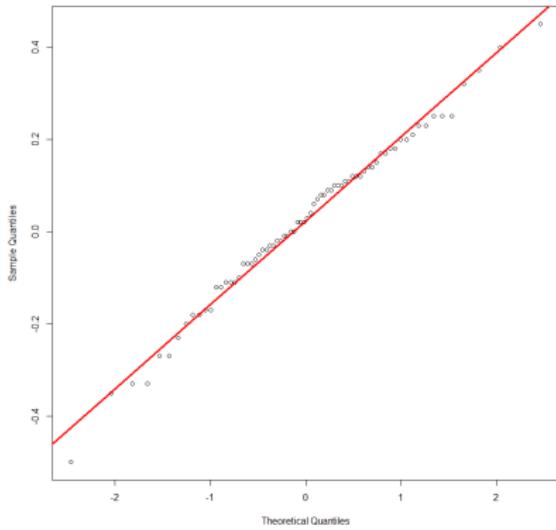
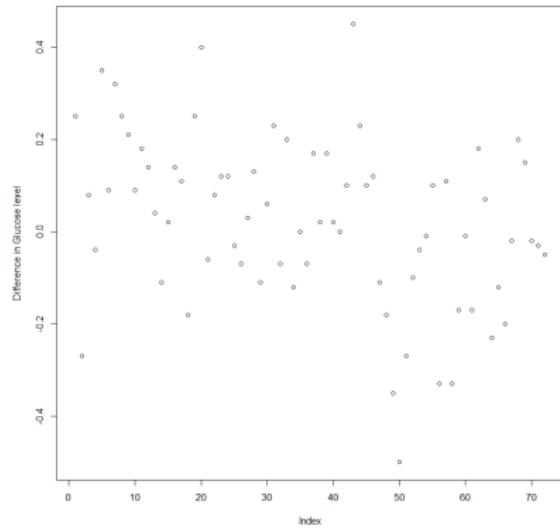
Normality:



# Assumptions for the Paired t-test

Continuing without the outlier:

```
glucose12.new<-glucose12[glucose12$Glucose1-glucose12$Glucose2>-1,]  
glucose12.new$D<-glucose12.new$Glucose1-glucose12.new$Glucose2
```



# Estimation of Method Difference

We wish to estimate the mean difference, ie. the parameter  $\mu$ . This puts us back to a one-sample problem. We have 72 measurements of the same normally distributed variable  $D$ :

- The estimate of the mean difference is  $\hat{\mu} = \bar{D} = 0.0238$ .
- The standard deviation of  $D$ : 0.1822
- The standard error of  $\hat{\mu}$ :  $SEM = \frac{sd(D)}{\sqrt{n}} = \frac{0.1822}{\sqrt{72}} = 0.0215$

# Confidence Interval

The 95% confidence interval for  $\mu$ :

$$\begin{aligned}\bar{D} &\pm t_{97.5\%}(72) \times SEM \\ &= 0.0238 \pm 1.9935 \times 0.0215 \\ &= [-0.0191; 0.0666]\end{aligned}$$

Compare with the standard confidence interval from normal errors:

$$\begin{aligned}\bar{D} &\pm 1.96 \times SEM \\ &= [-0.0183; 0.0658]\end{aligned}$$

# Confidence Interval

We found a confidence interval of

$$[-0.0191; 0.0666]$$

- This interval includes 0; a t-test will show that the data do not support a systematic bias at the 5% test level.
- We will expect the mean of the differences in a similar experiment with 72 subjects to be within this interval, with 95% probability.
- In our study, there could still be a systematic bias less than 0.019, but the t-test from the study will not have enough power to detect it.

## Test of No Bias

Let us test the hypothesis  $H_0 : \mu = 0$  against the alternative  $H_1 : \mu \neq 0$ :

$$T = \frac{\hat{\mu} - 0}{SEM} = \frac{0.0238 - 0}{0.0215} = 1.1059 \sim t(71)$$

The p-value in a t-distribution with 71 degrees of freedom is  $p = 0.27$ , so the hypothesis is accepted; the data do not support a systematic difference between the two methods at the 5% test level.

Note the correspondence between test and confidence interval:

- If the CI contains 0, the t-test will be statistically insignificant;
- If the CI does not contain 0, the t-test will be statistically significant.

## Paired t-test in R

- We don't have to calculate this by hand but we can use **R**:

```
t.test(glucose12.new$D)
```

One Sample t-test

```
data: glucose12.new$D
```

```
t = 1.1059, df = 71, p-value = 0.2725
```

```
alternative hypothesis: true mean is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.01906955 0.06656955
```

```
sample estimates:
```

```
mean of x
```

```
0.02375
```

# Paired t-test in R - Alternative Formulation

```
t.test(glucose12.new$Glucose1,glucose12.new$Glucose2,paired=TRUE)
```

## Paired t-test

```
data: glucose12.new$Glucose1 and glucose12.new$Glucose2
t = 1.1059, df = 71, p-value = 0.2725
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.01906955  0.06656955
sample estimates:
mean of the differences
                0.02375
```

# Exercise - Mobile Phones 1

Recall the study of reaction times when driving. In this exercise we have results from a paired study design, where each subject performs both the ‘Mobile’ and the ‘Control’ experiment.

- Load the data `Mobile_Matched.txt`.
- Make relevant plots of the data, and formulate the hypotheses to test the method difference.
- Evaluate the model control, and perform the test.

## Exercise - Mobile Phones 2

Repeat the analysis from the previous exercise, but this time transform the original reaction times with any log transform.

- See if the check for normality check went better. Comment and compare to the previous exercise.
- Present your results both on the chosen log-scale and back-transformed to the original scale. Is the conclusion altered compared to the non-transformed data?

# Case: Brain

## Story

The dataset `brainweight.txt` contains measurements of the weight of both brain and body for different mammals.

## Data

Variable	Description
art	species of mammal
body	weight of body (kg)
brain	weight of brain (gram)

## Exercise

1. Make a scatterplot of body against brain. Do you see any association between the variables?
2. Make a log transform of both body and brain. Make a scatterplot of the transformed variables. Do you see any association now?
3. Fit a regression model describing `log(brain)` by `log(body)`
4. Write up the mathematical model for `log(brain)`. Then take the anti-log on both sides to obtain an expression for `brain` and simplify the equation.  
What is the structure of the relation between brain and body weight? Is it linear?
5. Estimate confidence intervals for the parameters in the model for `brain`.
6. Test the hypothesis that the brain-to-body ratio is constant (does not depend on body weight).
7. Save the code that performs your analysis in a script and add plenty of comments to your code.

## Extra

1. Are there any outliers in the data? Take a look at the observations that are the least well fitted by your model. Can you explain these observations?
2. Perform model diagnostics on the model in question 3. and argue that each of the model assumptions are fulfilled.
3. Illustrate the fitted model on the data. Also quantify the model uncertainty.
4. What if a new mammal was discovered with a body weight of 5kg. What would the likely brain weight of this mammal be? To quantify the uncertainty in your estimate, provide an appropriate interval.
5. Investigate the quality of the data source and realize that one mammal appears twice. Which mammal is it and which of the entries is the right one (use google or wikipedia to get the facts)? Are there other problems with the data?

# Case: PCB

## Story

The concentration of polychlorinated biphenyl (PCB) residues in a series of lake trout from Cayuga Lake, NY, were reported in Bache et al. (1972). The ages of the fish were accurately known, because the fish were annually stocked as yearlings and distinctly marked as to year class. Each whole fish was mechanically chopped, ground, and thoroughly mixed, and 5-gram samples taken. The samples were treated and PCB residues in parts per million (ppm) were estimated using column chromatography. Data are available in the file `troutpcb.txt`.

The main objective of this case is to obtain a reasonable description (model) of the relation between age of trouts in lake Cayuga and PCB concentration. This includes model fitting, model diagnostics, quantification of uncertainty, communication of the model, and illustration of the model on the data.

## Data

Variable	Description
PCB	PCB concentration in ppm
Age	Age of the trout in years

## Exercise

1. Download `troutpcb.txt` from Campus Net and save it in a new empty folder. Open RStudio or another R editor of your choice and change the *working directory* to the directory where you saved `troutpcb.txt`. In RStudio you go to **Session → Set Working Directory** and navigate to relevant directory.

Now open a new R script (In RStudio go to **File → New File → R Script**). Now save the file in the same place as `troutpcb.txt`.

2. Read the data into R with something like

```
pcb <- read.table("troutpcb.txt", header=TRUE, sep="\t")
```

Now inspect the data and see that they have been read in correctly:

```
head(pcb)
str(pcb)
summary(pcb)
```

3. Plot the data using something like

```
plot(PCB ~ Age, data=pcb)
```

4. Fit the linear regression model and add the fit to the plot:

```
fm <- lm(PCB ~ Age, data=pcb)
abline(fm)
```

Now `fm` holds the *linear model* object and you can for instance extract the coefficients, i.e. parameter estimates with `coef(fm)`.

What is your impression of the model fit — do you think you have found a reasonable model?

5. Perform model diagnostics and plot the residuals:

```
par(mfrow=c(2, 2)) ## split plotting region in 4
plot(fm, which=1:4)
```

Realize that there is a serious problem with this fit: Argue that the variance is not constant and that a log-transform could remedy the problem.

6. Take the logarithm of PCB concentration and save the variable in the data frame:

```
pcb$log10PCB <- log10(pcb$PCB)
## Even better R code:
## pcb <- within(pcb, log10PCB <- log10(PCB))
```

Argue that  $\log_{10}$  is better than  $\log_e$  (Hint: consider interpretation of the axes in a plot of age versus log PCB concentration).

7. Illustrate log PCB versus age, fit the model using log PCB and check the residuals again:

```
par(mfrow=c(1, 1))
plot(log10PCB ~ Age, data=pcb)
fm2 <- lm(log10PCB ~ Age, data=pcb)
abline(fm2)
```

```
par(mfrow=c(2, 2))
plot(fm2, which=1:4)
```

Argue that the residuals are much better behaved. Is there evidence of variance heterogeneity? There is one observation which might be influential; which one?

8. The relation between PCB and age now has the following form:

$$\log_{10} PCB = \hat{\beta}_0 + \hat{\beta}_1 Age$$

Taking the anti-log on both sides, we obtain

$$\begin{aligned} PCB &= 10^{\hat{\beta}_0 + \hat{\beta}_1 Age} \\ &= 10^{\hat{\beta}_0} \cdot 10^{\hat{\beta}_1 Age} \\ &= \hat{\beta}'_0 \cdot \hat{\beta}'_1^{Age} \quad \text{where } \hat{\beta}'_0 = 10^{\hat{\beta}_0}, \hat{\beta}'_1 = 10^{\hat{\beta}_1} \end{aligned}$$

Compute the coefficients and insert the numbers for  $\hat{\beta}'_0$  and  $\hat{\beta}'_1$ . How many decimals do you think are appropriate here?

9. Observe that while the model for  $\log PCB$  is additive, the resulting model for  $PCB$  is *multiplicative*!

The results of a multiplicative model can be interpreted as

- PCB concentration increases with  $XX\%$  per year
- The rate of increase is between  $YY$  and  $ZZ$  percent per year based on a 95% confidence interval.

Compute the right numbers for  $XX$ ,  $YY$  and  $ZZ$ .

Note that if  $L$  and  $U$  are the lower and upper confidence limits for  $\hat{\beta}_j$ , then  $10^L$  and  $10^U$  are appropriate confidence limits for  $10^{\hat{\beta}_j}$ . Observe that while the CI is symmetric for  $\hat{\beta}_j$ , it is *asymmetric* for  $10^{\hat{\beta}_j}$  — uncertainty is not symmetric in the multiplicative model.

10. Suppose you are writing a paper based on these analyses and want to include a table of your findings. You compute the following:

```
B <- coef(summary(fm2))
df <- data.frame(Estimate = round(10^B[, 1], 2),
                  Lower = round(10^confint(fm2)[1, ], 2),
                  Upper = round(10^confint(fm2)[2, ], 2),
                  "p-value" = format.pval(B[, 4], digits=3, eps=1e-3))
df
```

What does the `format.pval` function do?

The  $p$ -value measures evidence against the null hypothesis:  $H_0 : \beta'_1 = X$ . What is the value of  $X$  here?

To communicate your results verbally, you write that “There is xxxx evidence that the rate of increase in PCB different from  $X$ .” Choose xxxx among *no*, *little*, *some* or *strong*.

You may also want to add something like this to your paper:

- The rate of increase is significantly different from  $WW$  ( $p$ -value? $KKK$ ;  $t = JJJ$ ;  $df = NN$ )

Substitute the relevant numbers. Here “?” can be one of “ $<$ ”, “ $=$ ”, or “ $>$ ”.

11. Illustrate the model on the original scale of measurement.

First plot PCB versus age:

```
plot(PCB ~ Age, data=pcb, ylim=c(0, 35),
      xlab="Age [years]", ylab="PCB concentration [ppm]", bty="n",
      las=1)
```

Then compute the model fit and confidence intervals for the fitted model:

```
xval <- seq(1, 12, length=500)
pred <- predict(fm2, newdata=data.frame(Age=xval),
                interval="confidence")
lines(xval, 10^pred[, "fit"], lwd=2)
lines(xval, 10^pred[, "lwr"], col="red", lwd=2, lty=2)
lines(xval, 10^pred[, "upr"], col="red", lwd=2, lty=2)
```

You can also add *prediction* intervals to the figure:

```
pred <- predict(fm2, newdata=data.frame(Age=xval),
                 interval="prediction")
lines(xval, 10^pred[, "lwr"], col="blue", lty=3, lwd=2)
lines(xval, 10^pred[, "upr"], col="blue", lty=3, lwd=2)
```

Finally you may want to add a legend:

```
legend("topleft", legend=c("Fit", "95% Confidence interval",
                           "95% Prediction interval"),
       lwd=2, col=c("black", "red", "blue"), lty=1:3, bty="n")
```

12. Save the code that performs your analysis in a script and add plenty of comments to your code.

### Extra exercises

1. Investigate the sensitivity to the outlier identified previously. Do that by excluding this point from the data, then refit the model and compare the estimates of  $\beta_1'$  and their 95% confidence intervals for the two models. How large is the change in the coefficient relative to the statistical uncertainty? Perform model diagnostics on this model; do the residuals behave better or worse?
2. Another remedy is to transform `Age` using `Age^c`. We could choose  $c$  from the data, but here we will use the cube root transform, i.e.  $Age^{1/3} = \sqrt[3]{Age}$ . Fit the model for  $\log_{10}PCB$  using  $\sqrt[3]{Age}$  instead of just `Age` and inspect the residuals. What is your assessment of the residual diagnostics now?
3. Visualize the model uncertainty by displaying the four different model fits on the same graph. The four models are:
  - (a) The linear fit:  $PCB \propto Age$
  - (b) The log-linear model:  $\log_{10}PCB \propto Age$
  - (c) The log-linear model with the influential observation excluded.
  - (d) The log-linear model using the transformed age:  $\log_{10}PCB \propto \sqrt[3]{Age}$

Which model would you place most trust in?

4. (Difficult) Interpretation of the model with  $\log_{10}PCB \propto \sqrt[3]{Age}$  is hampered by the fact that rate of increase in PCB with Age is not constant. Derive and compute the rate of increase in PCB as a function of age and plot the relation. Hint: we can interpret the rate of increase as the slope of the relation between  $\log_{10}PCB$  and age.

Illustrate the rate of increase in Age as a function of PCB for the log-linear models with `Age` and  $\sqrt[3]{Age}$

# Linear Regression - Part 1

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 7th, 2025

## Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Learning objectives

After this session you should be able to:

- ① Understand what a linear regression model is and be able to fit it to data;

# Learning objectives

After this session you should be able to:

- ① Understand what a linear regression model is and be able to fit it to data;
- ② Understand the assumptions of linear regression models and perform model diagnostics;

# Learning objectives

After this session you should be able to:

- ① Understand what a linear regression model is and be able to fit it to data;
- ② Understand the assumptions of linear regression models and perform model diagnostics;
- ③ Try polynomials or logarithms if the assumptions are not fulfilled;

# Learning objectives

After this session you should be able to:

- ① Understand what a linear regression model is and be able to fit it to data;
- ② Understand the assumptions of linear regression models and perform model diagnostics;
- ③ Try polynomials or logarithms if the assumptions are not fulfilled;
- ④ Illustrate the fitted model on the data.

# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# Simple Linear Regression

- The association between two continuous variables:
- $Y$  response / outcome / dependent variable;
- $X$  explanatory / covariate / independent variable.

# Data

- Observations of pairs  $(x_i, y_i)$  for all  $i = 1, \dots, n$  individuals or units.
- Note that regression is **not symmetrical** in  $X$  and  $Y$ .
- In some studies it is possible to chose  $X$  beforehand, this gives more precise results.

## Example

Is higher education associated with higher murder rates?

- Crime data from 2003 for the US.
- Murder rate: The annual number of murders per 100,000 people in the population.
- Poverty: Percentage of residents with income below the poverty level.
- High school: Percentage of the adult residents who have at least a high school education.
- College: Percentage of the adult residents who have a college education.
- Single parent: Percentage of families headed by a single parent.

The outcome  $Y$  is the murder rate and the explanatory variable  $X$  college.  
**Always start by plotting the data!**

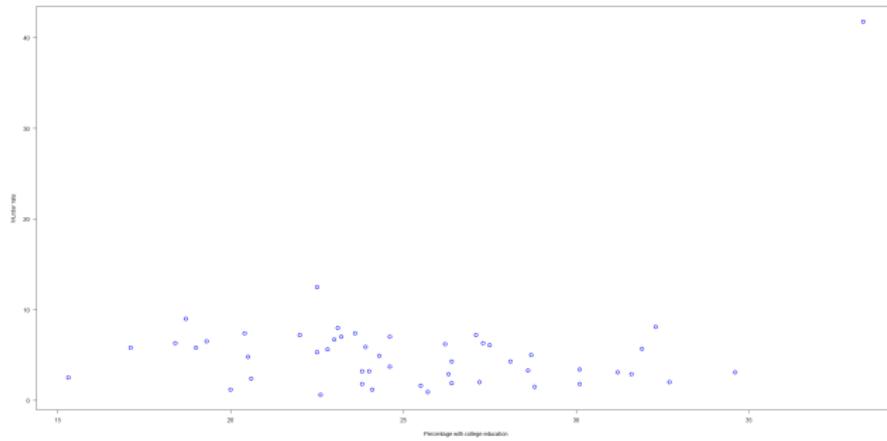
# Read data into R

```
crime <- read.delim("us_statewide_crime.txt")
head(crime)

##           State violent.crime.rate murder.rate poverty high.school college
## 1      Alabama                 486       7.4     14.7     77.5    20.4
## 2      Alaska                  567       4.3      8.4     90.4    28.1
## 3      Arizona                 532       7.0     13.5     85.1    24.6
## 4      Arkansas                 445       6.3     15.8     81.7    18.4
## 5 California                 622       6.1     14.0     81.2    27.5
## 6 Colorado                   334       3.1      8.5     89.7    34.6
##   single.parent unemployed metropolitan
## 1          26.0        4.6       70.2
## 2          23.2        6.6       41.6
## 3          23.5        3.9       87.9
## 4          24.7        4.4       49.0
## 5          21.8        4.9       96.7
## 6          20.8        2.7       84.0
```

# Scatter plot

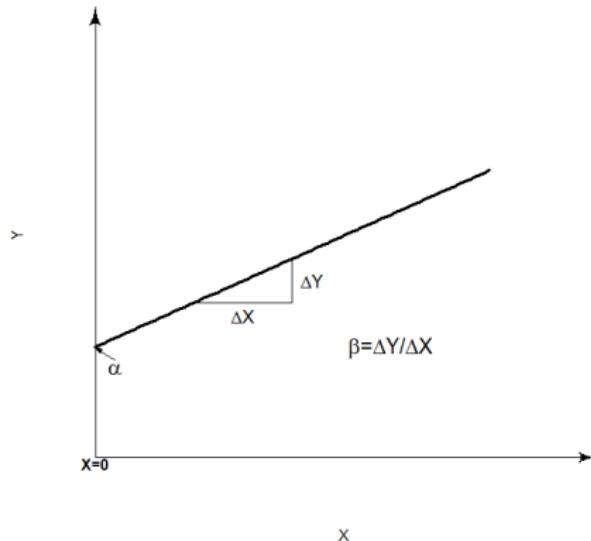
```
plot(crime$college, crime$murder.rate, las=1, cex=1.5, col ="blue",  
lwd=2, ylab="Murder rate",  
xlab="Percentage with college education")
```



# Mathematical Model

The expression for a straight line:

$$Y = \alpha + \beta X$$



# The Parameters

- $\alpha$ : The intercept. The murder rate in a state with 0 percentage college education. **Often a meaningless extrapolation!** The intercept has the same units as the outcome.

# The Parameters

- $\alpha$ : The intercept. The murder rate in a state with 0 percentage college education. **Often a meaningless extrapolation!** The intercept has the same units as the outcome.
- $\beta$ : The slope, the regression coefficient. The difference in murder rate between two states with a difference of 1 percentage in college education. **Often the main parameter of interest.**

# Statistical Model

$$Y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, \dots, n \quad \varepsilon_i \sim N(0, \sigma^2)$$

where

$Y_i$  is the response/dependent variable (random)

$\alpha$  is the intercept (fixed, unknown)

$x_i$  is the covariate/independent variable (fixed)

$\beta$  is the slope (fixed, unknown)

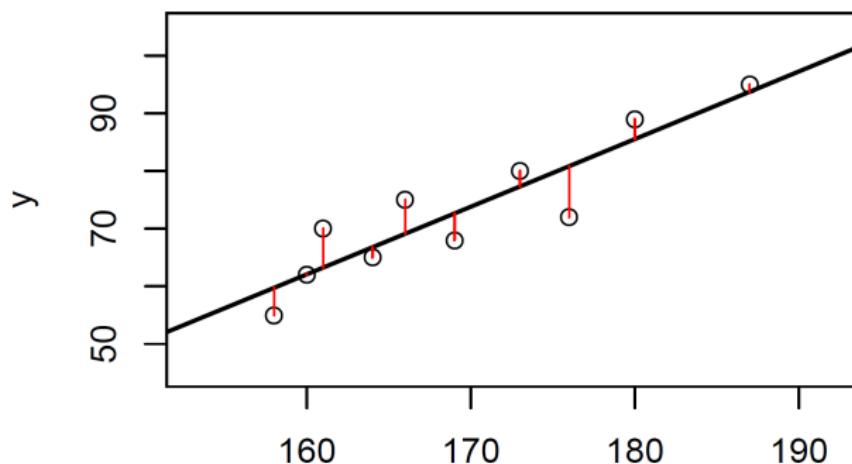
$\varepsilon_i$  is the unobserved random error

$n$  is the number of observations (known)

$\sigma^2$  is the residual variance (fixed, unknown)

## Residuals

The  $\varepsilon_i$  are the difference between the observed  $Y_i$  and the expectation. The best fitting line is found by minimizing the squared residuals (the distances to the line - the estimated random errors). The observations are fixed, while the line varies.



# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# Estimation

Find  $\alpha$  and  $\beta$  so the squared distances to the line becomes as small as possible. I.e. minimize:

$$\sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2$$

The resulting estimates:

$$\hat{\beta} = \frac{SP_{xy}}{SSD_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

where:

$\bar{x}$  is the average of  $x_i$  and  $\bar{y}$  is the average of  $y_i$ .

# Regression in R

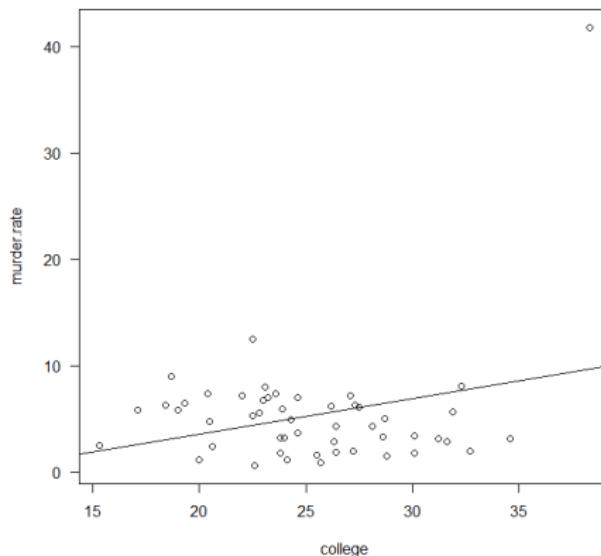
```
reg1 <- lm(murder.rate ~ college, data = crime)
summary(reg1)

##
## Call:
## lm(formula = murder.rate ~ college, data = crime)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.058     4.363   -0.70   0.487
## college      0.333     0.170    1.96   0.056 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.61 on 49 degrees of freedom
## Multiple R-squared:  0.0724, Adjusted R-squared:  0.0534
## F-statistic: 3.82 on 1 and 49 DF,  p-value: 0.0563
```

## Estimated regression line

Estimated line:  $\hat{y} = \hat{\alpha} + \hat{\beta}x$ , murder.rate =  $-3.06 + 0.33 \text{ college}$

```
plot(murder.rate ~ college, data = crime, las=1)  
abline(reg1)
```



### 3 Estimates from the regression analysis

- 2 estimates from the line (**intercept and slope**)
- The **variation** in  $y$  around the regression line ( $\sigma^2$ ). Here the variation in murder rate for states with the same college percentage.

The estimate for  $\sigma^2$  is:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

Nearly the average squared distance but  $n$  (the number of observations) is replaced by  $n - 2$  (the degrees of freedom). An estimator for  $\sigma$  is  $s = \sqrt{s^2}$  which is denoted **residual standard error** in the output ( $s = 5.61$ ).

# Trust, but confirm - estimating parameters

```
# BY HAND
x <- crime$college
y <- crime$murder.rate
(beta <- sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2))

## [1] 0.33307

(alpha <- mean(y)- beta*mean(x))

## [1] -3.0581

Fitted <- alpha + beta*x
Resid <- y - Fitted
sigma2 <- sum(Resid^2)/(length(y)-2)
sqrt(sigma2)

## [1] 5.6146
```

# Uncertainty of the Estimates

- How good are our estimates of the unknown parameters  $\alpha$  and  $\beta$ ?
- If we repeated the experiment how different would the estimates be?

# Uncertainty of the Estimates

- How good are our estimates of the unknown parameters  $\alpha$  and  $\beta$ ?
- If we repeated the experiment how different would the estimates be?
- It may been shown that  $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{SSD_x})$ .
- The estimate of  $\beta$  does not vary so much if:
  - ① If  $\sigma^2$  is small, i.e. the points are close to the line.
  - ② If  $SSD_x$  is large, i.e. the x values are spread out.

## Confidence interval

The **estimated** uncertainty of  $\hat{\beta}$  is called the standard error of  $\hat{\beta}$ :

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SSD_x}}$$

We use the standard error to construct the 95% confidence interval

$$\hat{\beta} \pm t_{0.975}(n - 2) \times SE(\hat{\beta})$$

Where  $t_{0.975}(n - 2)$  is the 97.5% percentile in the t-distribution with  $n - 2$  degrees of freedom.

## Confidence interval

The **estimated** uncertainty of  $\hat{\beta}$  is called the standard error of  $\hat{\beta}$ :

$$SE(\hat{\beta}) = \frac{s}{\sqrt{SSD_x}}$$

We use the standard error to construct the 95% confidence interval

$$\hat{\beta} \pm t_{0.975}(n - 2) \times SE(\hat{\beta})$$

Where  $t_{0.975}(n - 2)$  is the 97.5% percentile in the t-distribution with  $n - 2$  degrees of freedom.

**In our example:**  $0.3331 \pm 2.0096 \times 0.1703 = (-0.009; 0.675)$ .

## Test for $\beta = 0$

We can try to simplify the model, most often by:

$$H_0 : \beta = 0$$

I.e. no effect of the covariate (here college education). We use a t-test:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim t(n - 2)$$

Here  $t = \frac{0.3331}{0.1703} = 1.956$  with a p-value:

$$2 \cdot P(T > |t|) = 2 \cdot (1 - P(T < |t|))$$

```
2*(1-pt(1.956, 49))
```

```
## [1] 0.056179
```

# Tables of results in R

```
confint(reg1)

##              2.5 % 97.5 %
## (Intercept) -11.8259006 5.70976
## college     -0.0092443 0.67538

# Nice table
tab <- cbind(coef(summary(reg1))[, 1:2], "Lower" = confint(reg1)[, 1],
             "Upper" = confint(reg1)[, 2])
tab

##           Estimate Std. Error      Lower      Upper
## (Intercept) -3.05807   4.36303 -11.8259006  5.70976
## college     0.33307   0.17034  -0.0092443  0.67538

# Nice table with p-values
data.frame(round(tab, 2),
            "p-value" = format.pval(coef(summary(reg1))[, 4], digits = 3, eps = 1e-3))

##           Estimate Std. Error      Lower      Upper p.value
## (Intercept) -3.06       4.36 -11.83    5.71  0.4867
## college     0.33       0.17  -0.01    0.68  0.0563
```

# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# Model Check

- Is this a good model?

# Model Check

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.

# Model Check

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.
- We need to check:
  - Are the assumptions behind the model fulfilled?
  - Are there any influential observations that we need to check up on?

# Model Check

- Is this a good model?
- We can only trust our conclusions if the model is appropriate.
- We need to check:
  - Are the assumptions behind the model fulfilled?
  - Are there any influential observations that we need to check up on?
- One would like to check earlier, but we need the estimates to be able to check.

# Residual Analysis

The statistical model was:

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2) \text{ independent}$$

We have to check:

- Normal residuals (observed - fitted)
- Variance homogeneity (one  $\sigma^2$ )
- Linear effect of  $X$ .

Notice that we have no assumption about normal distribution for  $X$ .

# Residual Plots

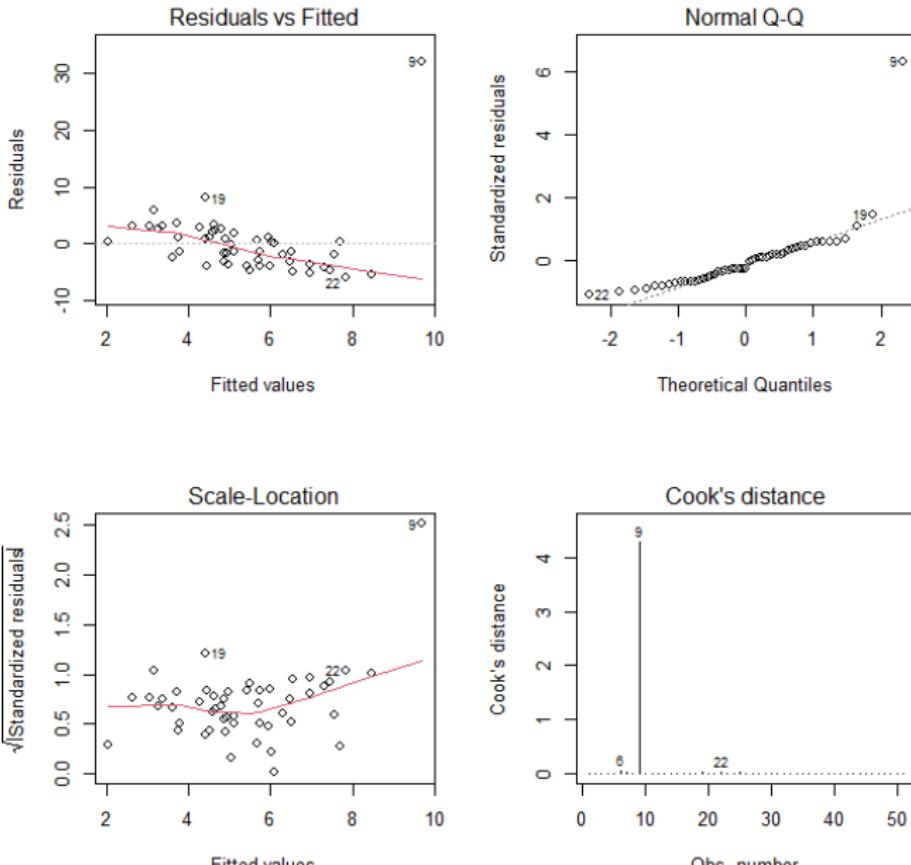
We will do graphical model checks where the residuals are plotted:

- Residuals plotted in qq-plots for normality (are the points close to straight line).
- Residuals plotted vs. fitted values ( $\hat{y}_i$ ) to check variance homogeneity (Look for trumpet shape and other irregularities).
- Residuals plotted vs. explanatory variable ( $x_i$ ) to check for linearity. (Look for curves or S-shape)

# Influential Observations

- May the results change if a few observations are left out, or are they insensitive to minor alterations of the data set?
- Influential observations are not necessarily problematic (although the may be). They impact on the strength of the evidence. If the conclusion rests on for example the presence of a single observation, reservations need to be taken.
- Remove one observation at a time and see how results change.
- Calculate Cook's distance for each observation  $i$ . A measure of how much the results change if observation  $i$  is left out.
- Plot in R

```
par(mfrow = c(2, 2))
plot(reg1, which = 1:4)
```

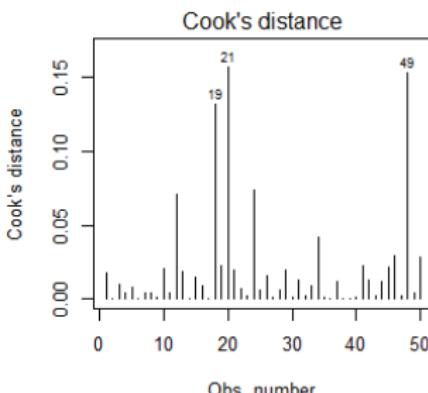
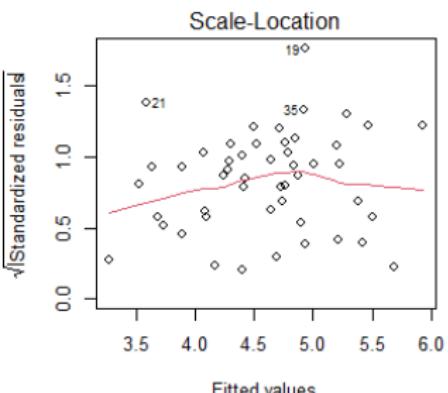
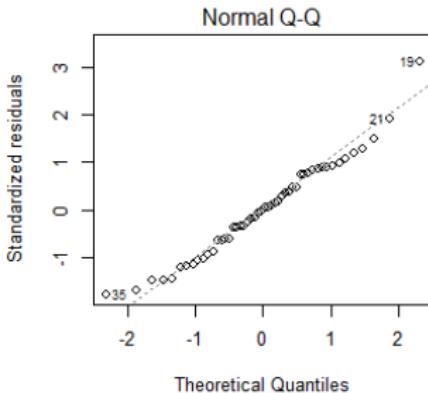
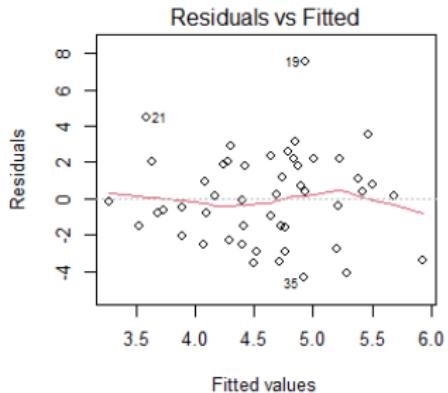


## Model check in example

- Something is different about observation 9!
- It is the observation from Washington DC.
- Washington DC is the capital area of the USA; not a state like the rest.
- I will decide to leave it out, because it is structurally different from the rest, and I will remember that conclusions are not valid for DC.

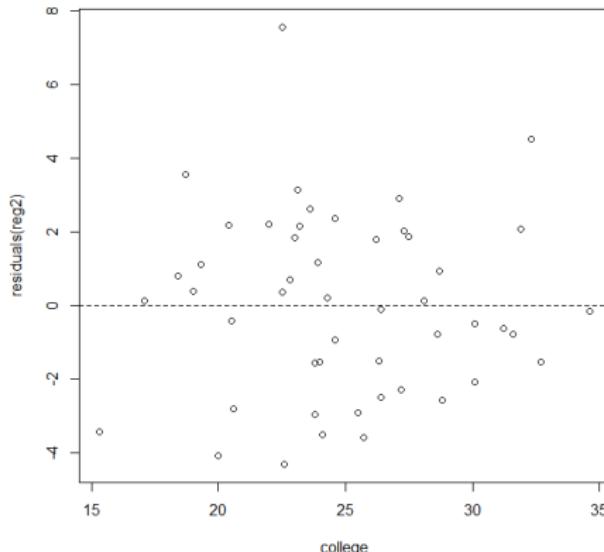
# New Data

```
crime50 <- crime[-9, ]  
head(crime50)  
  
##           State violent.crime.rate murder.rate poverty high.school college  
## 1    Alabama                 486       7.4   14.7     77.5   20.4  
## 2    Alaska                  567       4.3    8.4     90.4   28.1  
## 3    Arizona                 532       7.0   13.5     85.1   24.6  
## 4    Arkansas                 445       6.3   15.8     81.7   18.4  
## 5 California                622       6.1   14.0     81.2   27.5  
## 6 Colorado                  334       3.1    8.5     89.7   34.6  
## single.parent unemployed metropolitan  
## 1          26.0        4.6      70.2  
## 2          23.2        6.6      41.6  
## 3          23.5        3.9      87.9  
## 4          24.7        4.4      49.0  
## 5          21.8        4.9      96.7  
## 6          20.8        2.7      84.0  
  
reg2 <- lm(murder.rate ~ college, data = crime50)
```



# Check linearity

```
plot(residuals(reg2) ~ college, data = crime50,  
      ylab = "Residuals")  
abline(h = 0)
```



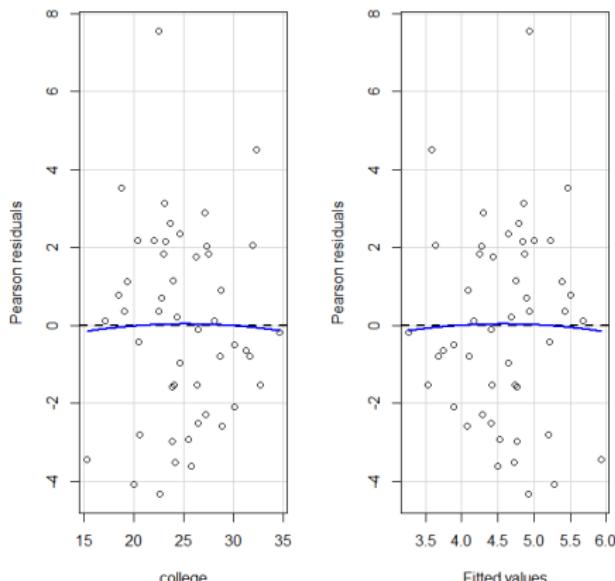
## Check linearity, alternative using the car library

- Instead of making the plot ourselves we can use the package `car` (companion to applied regression).

```
library(car)  
residualPlots(reg2)
```

# Check linearity, alternative using the car library

```
##           Test stat Pr(>|t|)  
## college      -0.128    0.899  
## Tukey test   -0.128    0.898
```



# Estimates for data without DC

```
summary(reg2)

##
## Call:
## lm(formula = murder.rate ~ college, data = crime50)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  8.0416    2.0648   3.89   0.0003 ***
## college     -0.1379    0.0816  -1.69   0.0977 .
##
## Residual standard error: 2.46 on 48 degrees of freedom
## Multiple R-squared:  0.0561, Adjusted R-squared:  0.0364
## F-statistic: 2.85 on 1 and 48 DF,  p-value: 0.0977
```

## Estimates for data without DC, in a table

	Estimate	Std. Error	Lower	Upper	p.value
## (Intercept)	8.04	2.06	3.89	12.19	<0.001
## college	-0.14	0.08	-0.30	0.03	0.0977

- The fitted model has the form:

$$\text{murder.rate}_i = 8.04 - 0.14 \cdot \text{college}_i + \epsilon_i, \quad N(0, 2.46^2)$$

## Estimates for data without DC, in a table

	Estimate	Std. Error	Lower	Upper	p.value
## (Intercept)	8.04	2.06	3.89	12.19	<0.001
## college	-0.14	0.08	-0.30	0.03	0.0977

- The fitted model has the form:

$$\text{murder.rate}_i = 8.04 - 0.14 \cdot \text{college}_i + \epsilon_i, \quad N(0, 2.46^2)$$

- For every percentage increase in college education the murder rate decreases with 0.14 with 95% CI -0.03 to 0.30.

## Estimates for data without DC, in a table

	Estimate	Std. Error	Lower	Upper	p.value
## (Intercept)	8.04	2.06	3.89	12.19	<0.001
## college	-0.14	0.08	-0.30	0.03	0.0977

- The fitted model has the form:

$$\text{murder.rate}_i = 8.04 - 0.14 \cdot \text{college}_i + \epsilon_i, \quad N(0, 2.46^2)$$

- For every percentage increase in college education the murder rate decreases with 0.14 with 95% CI -0.03 to 0.30.
- College education is not statistically significant for the murder rate as 0 is in the confidence interval, and  $p = 0.1$ .

## Estimates for data without DC, in a table

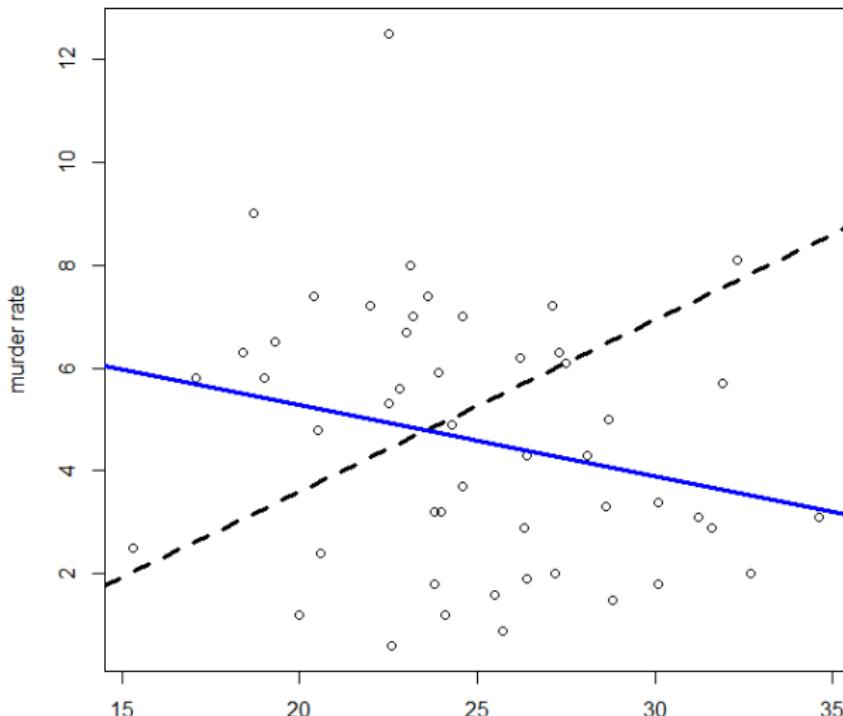
	Estimate	Std. Error	Lower	Upper	p.value
## (Intercept)	8.04	2.06	3.89	12.19	<0.001
## college	-0.14	0.08	-0.30	0.03	0.0977

- The fitted model has the form:

$$\text{murder.rate}_i = 8.04 - 0.14 \cdot \text{college}_i + \epsilon_i, \quad N(0, 2.46^2)$$

- For every percentage increase in college education the murder rate decreases with 0.14 with 95% CI -0.03 to 0.30.
- College education is not statistically significant for the murder rate as 0 is in the confidence interval, and  $p = 0.1$ .
- Washington DC was a very influential observation and changed the negative association to positive. However, DC is not a state and may behave in another fashion than the rest. This observation has therefore been left out of further analyses.

# The Regression Lines



# When is Regression Wrong?

Find a partner and explain to each other what is wrong with the use of regression in these examples.

- Winning times at the Boston marathon have followed a straight line decreasing trend from 160 minutes in 1927 to 130 minutes in 2004. After fitting a regression line you use the equation to predict that the winning time in 2023 will be about 123 minutes.
- A regression between  $x = \text{years of education}$  and  $y = \text{annual income}$  for 100 people shows a modest positive trend; person number 101 dropped out of school after the 10th grade and is now a multimillionaire. Since it is wrong to leave out data, we should report all results including this point resulting in a negative slope.

# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# Confidence limits for regression line

The confidence interval for the fitted line at  $x_0$ :

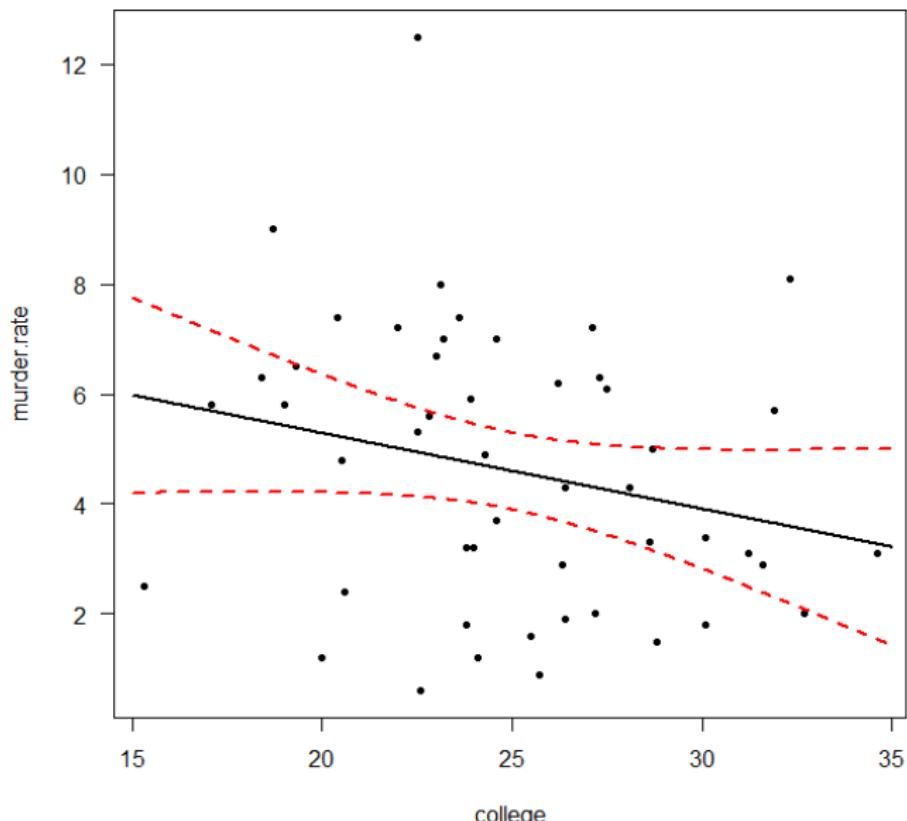
$$\hat{\alpha} + x_0 \hat{\beta} \pm t_{\alpha/2}(n - 2) \cdot s \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SSD_x}}$$

- Expresses uncertainty about the **fitted line/model**.
- These limits become narrower when the number of observations is increased.
- The limits are narrowest close to the mean of  $x$ .

# Confidence intervals with R

```
xval <- seq(from = 15, to = 35, length.out = 500)
newData <- data.frame(college = xval)
Pred.ci <- predict(reg2, newdata = newData,
                    interval = "confidence",
                    level = .95)

## Plot data, model and intervals:
plot(murder.rate ~ college, data = crime50, pch = 20, las = 1)
lines(xval, Pred.ci[, "fit"], lwd = 2)  ## or use: abline(reg2)
lines(xval, Pred.ci[, "lwr"], lty = 2, col = "red", lwd = 2)
lines(xval, Pred.ci[, "upr"], lty = 2, col = "red", lwd = 2)
```



# Prediction limits for regression line

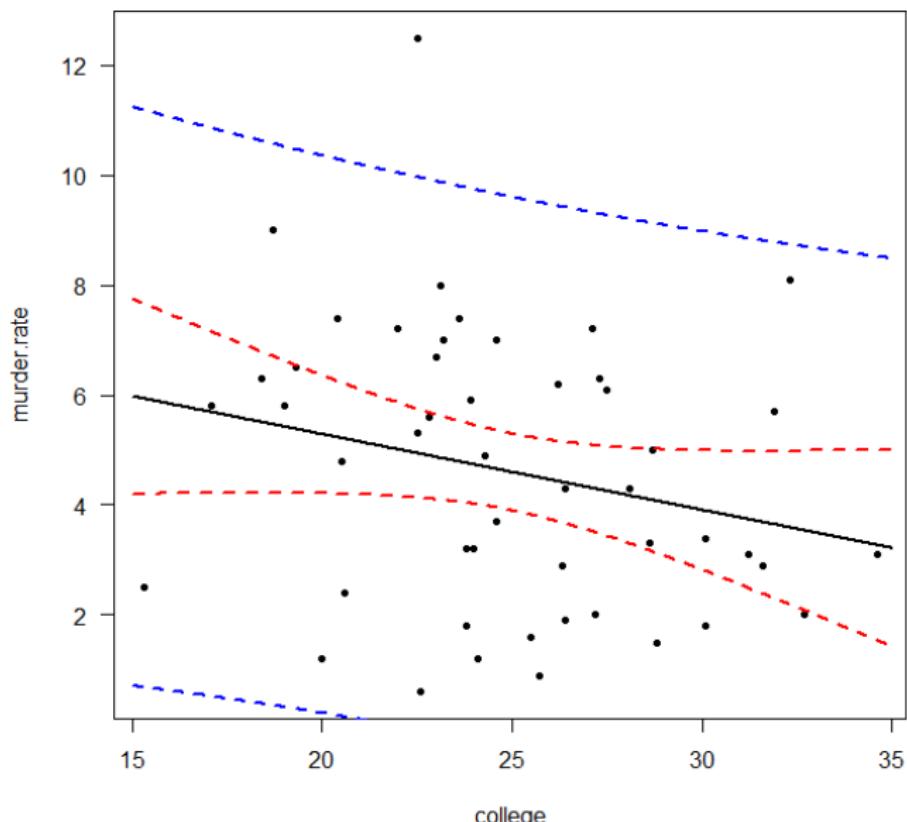
The prediction interval for the fitted line at  $x_0$ :

$$\hat{a} + x_0 \hat{\beta} \pm t_{\alpha/2}(n - 2) \cdot s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{XX}}}$$

- Expresses uncertainty about the a single observation  $(x_0, y_0)$  when  $x_0$  is known.
- Is used to decide whether a new observation is atypical, as the limits will include approximately 95% of future observations.
- The region where we would expect to see 95% of future observations.

# Computing prediction intervals with R

```
## Prediction interval for a new observation:  
Pred.pi <- predict(reg2, newdata = newData,  
                     interval = "prediction")  
## Add prediction intervals to plot:  
lines(xval, Pred.pi[, "lwr"], lty=2,  
      col="blue", lwd=2)  
lines(xval, Pred.pi[, "upr"], lty=2,  
      col="blue", lwd=2)  
## could add legend here.
```



# Prediction and Confidence Limits

- The narrow limits are confidence limits:
  - Corresponds to standard error.
  - Is used to assess the uncertainty of the estimate.
  - Depends a lot on the value of  $x_0$ .

# Prediction and Confidence Limits

- The narrow limits are confidence limits:
  - Corresponds to standard error.
  - Is used to assess the uncertainty of the estimate.
  - Depends a lot on the value of  $x_0$ .
- The broad limits are prediction limits:
  - Corresponds to standard deviation.
  - Also called reference range.
  - Is used to assess individual observations.
  - Approximately calculated as  $\pm 2$  residual standard error.

# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# What do we do if the model does not fit?

- In our example we had an influential observation, but without this observation the model had a good fit.
- We were checking:
  - Normality of residuals.
  - Variance homogeneity.
  - Linear effect of covariate.

## Example about timber hardness

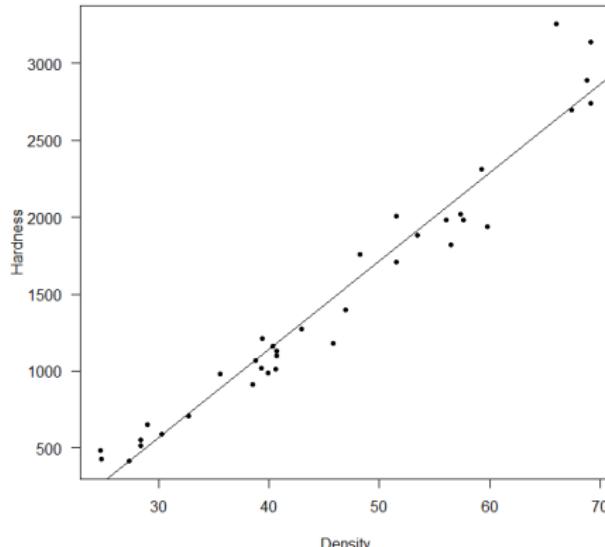
- We have 36 observations of timber hardness and density from some trees in Australia.
- The aim of the study was to estimate parameters that define the relationship between timber hardness and density, so we could predict (unknown) timber hardness in future samples based on the measured density.
- Our first model could be a simple regression:

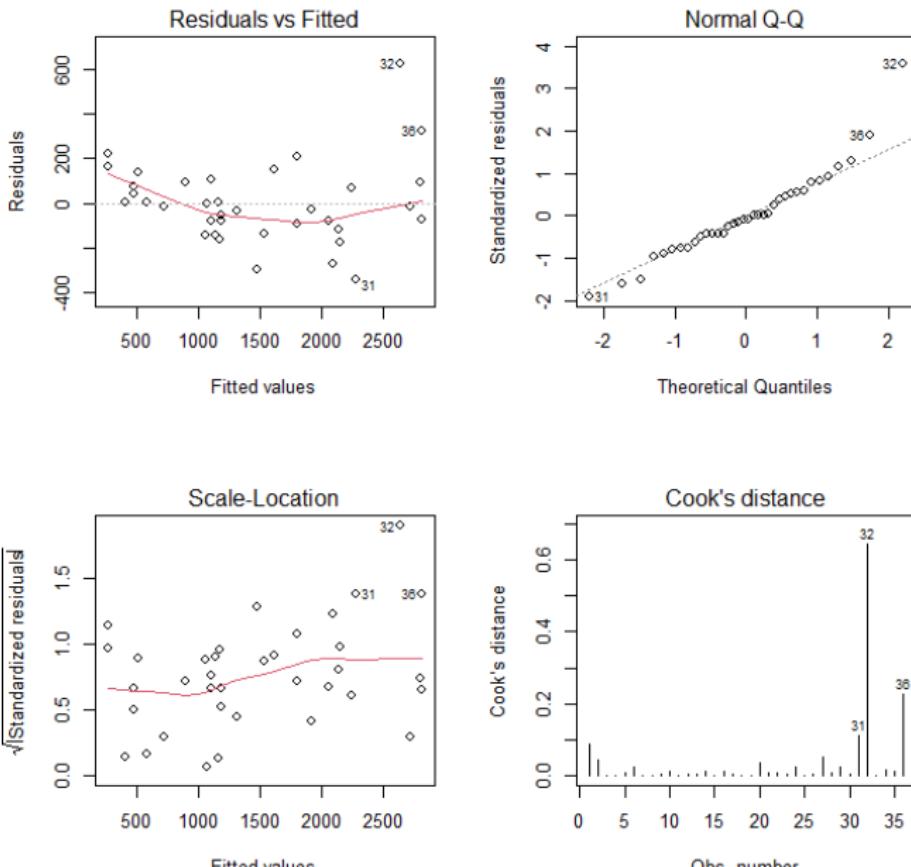
$$\text{Hardness}_i = \alpha + \beta \text{Density}_i + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

- But does this model fit?

# The Janka Example

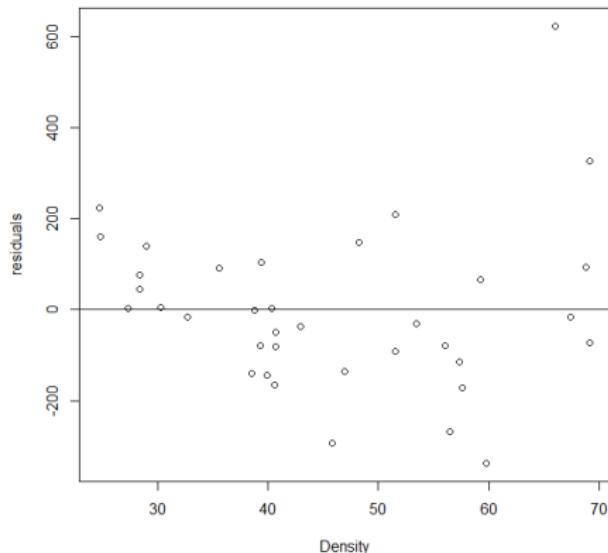
```
reg3 <- lm(Hardness ~ Density, data = janka)
plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
abline(reg3)
```





# Check linearity

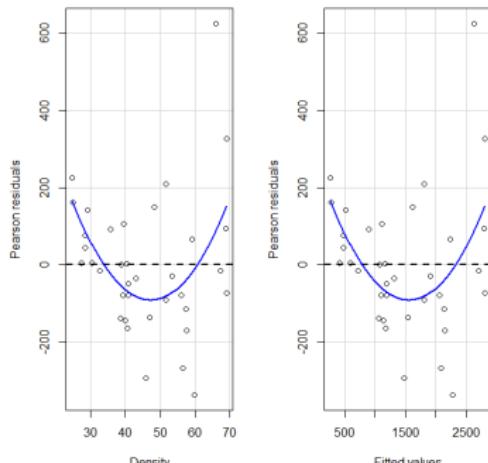
```
plot(residuals(reg3) ~ Density, data = janka,  
      ylab = "residuals")  
abline(h = 0)
```



# Check linearity, using the car library

```
library(car)  
residualPlots(reg3)
```

```
##           Test stat Pr(>|t|)  
## Density      3.248   0.003  
## Tukey test   3.248   0.001
```



# Model check plots, what are we looking for?

- ① **Plot of Residuals vs Fitted** Variance homogeneity. Problem if we see a **trumpet shape**, which we do here. We can also look for non-linear shapes - apart from the trumpet shape a curve is present.
- ② **Normal Q-Q** Residuals have to follow a normal distribution. The observations should follow **a straight line**. They do here (sort of); this kind of model abberation cannot be detected by a qq-plot.
- ③ **Scale-Location** Variance homogeneity. The line should be horizontal (supplements plot 1). We see a slight increase.
- ④ **Cook's distance** Shows potentially influential observations.
- ⑤ **Residual plots for the linear effect** This should look like a random scatter of dots. **Looking for shapes** e.g. polynomial, here we see a curved association.

# Problems with normality

## Normally distributed residuals

If the normal distribution is not valid then we lose power and the prediction limits are not valid.

What can we do?

- If we have a heavy tail to the right, then try transforming the response with the logarithm. We will get back to alternative transformations in an exercise on Thursday.
- Non-parametric methods, simulation

# Problems with Variance Homogeneity

## Variance Homogeneity

If the variance is not constant then we lose power and the prediction limits are not valid.

What can we do?

- If we have a trumpet shape in the residual plot, then try transforming the response with the logarithm.

# Problems with Variance Homogeneity

## Variance Homogeneity

If the variance is not constant then we lose power and the prediction limits are not valid.

What can we do?

- If we have a trumpet shape in the residual plot, then try transforming the response with the logarithm.

## Trumpet shape

- Trumpet shapes are often seen when measuring small positive values, e.g. concentrations.
- Can be regarded as a constant relative uncertainty, constant coefficient of variation.

$$\text{Coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}}$$

# Normality and Variance Homogeneity

- The assumption about normality (and variance homogeneity) are not so crucial for the estimates (NB: It is very crucial for statistical inference in general, when the models are more complicated).
- Here we are using the t-distribution for  $\hat{\beta}$  and this needs normality but the **Central Limit Theorem** helps.
- **Central Limit Theorem:** sums of more and more observations approaches normality.
- **The prediction limits cannot be used** (concern single observations).

# Problems with Linearity

## Linearity

If we have assumed linearity without it being approximately ok then we cannot interpret the model. **What can we do?**

- If we see a curve in the residual plots we can add our covariate squared, cubed...
- Transform variables with log, square root, inverse...

# Polynomial regression models

What if the relation between  $y$  and  $x$  is not a straight line?

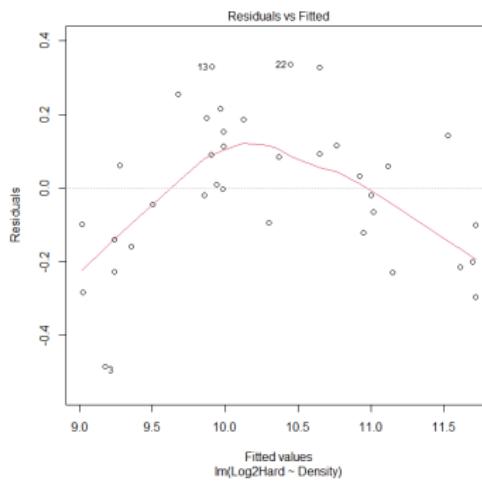
We can use a  $p$ th order polynomial to fit curves:

$$Y_i = \beta_0 + x_i\beta_1 + x_i^2\beta_2 + \dots + x_i^p\beta_p + \varepsilon_i$$

This is still a linear model since it is linear in the parameters,  $\beta$ .

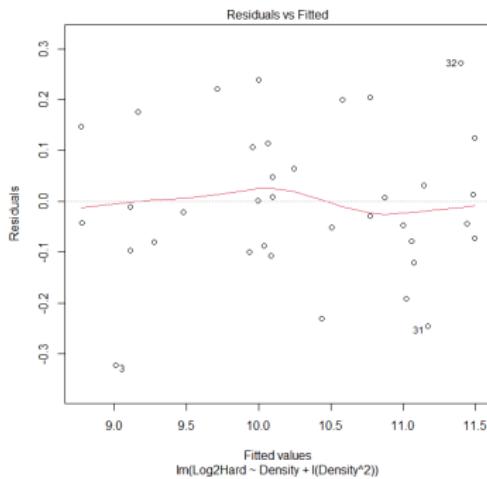
# The Janka Example, a better model

```
janka$Log2Hard <- log2(janka$Hardness)
reg4 <- lm(Log2Hard ~ Density, data = janka)
plot(reg4, which = 1)
```



# The Janka Example, Adding a Squared Term

```
#TRANSFORM THE OUTCOME WITH LOG2 AND ADD SQUARED DENSITY
reg5 <- lm(Log2Hard ~ Density + I(Density^2), data = janka)
plot(reg5, which=1)
```

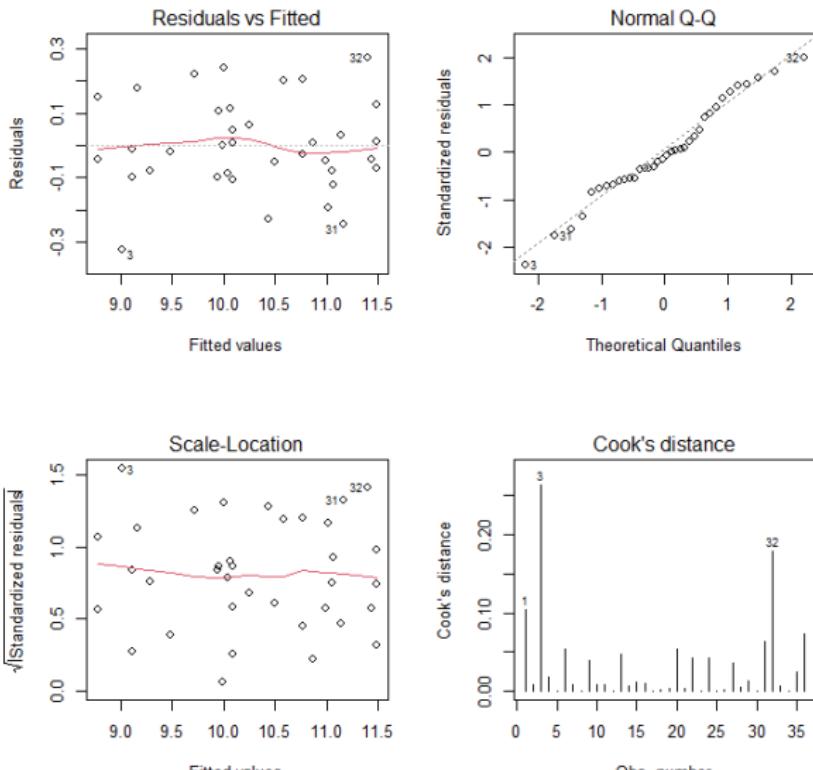


# The Janka Example, Final Model

```
# LOOK AT THE RESULTS
summary(reg5)

##
## Call:
## lm(formula = Log2Hard ~ Density + I(Density^2), data = janka)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.969895  0.301078 19.83   < 2e-16 ***
## Density     0.132029  0.013425  9.83  2.5e-11 ***
## I(Density^2) -0.000754  0.000141  -5.35  6.5e-06 ***
##
## Residual standard error: 0.145 on 33 degrees of freedom
## Multiple R-squared:  0.972, Adjusted R-squared:  0.971
## F-statistic: 579 on 2 and 33 DF,  p-value: <2e-16
```

# The Janka Example, Final Model



# The Janka Example, Final Model

We have estimated

$$\text{Log2Hardness}_i = 5.97 + 0.13\text{Density} - 0.001\text{Density}^2 + \varepsilon_i$$

But this is on the log2 scale. We will want to see the results on the original scale.

# The Janka Example, Final Model

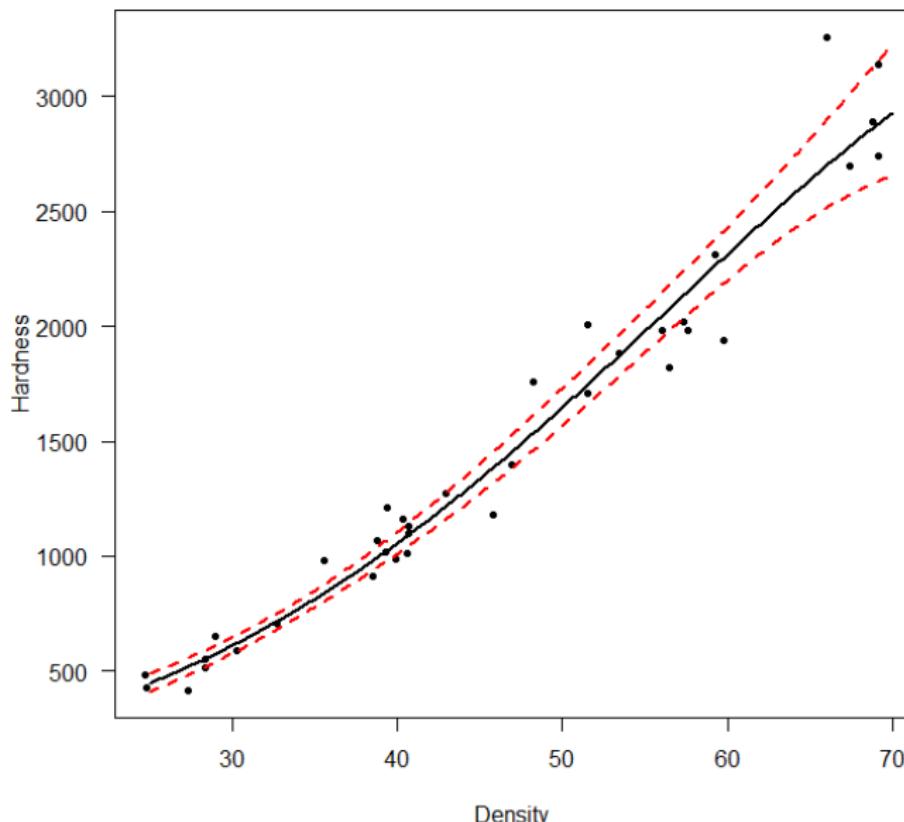
We have estimated

$$\text{Log2Hardness}_i = 5.97 + 0.13\text{Density} - 0.001\text{Density}^2 + \varepsilon_i$$

But this is on the log2 scale. We will want to see the results on the original scale.

```
xval <- seq(from = 25, to = 70, length.out = 500)
newData <- data.frame(Density = xval)
Pred.ci <- predict(reg5, newdata = newData,
                    interval = "confidence", level = .95)

# ON THE ORIGINAL SCALE, WE USED log2
plot(Hardness ~ Density, data = janka, pch = 20, las = 1)
lines(xval, 2^Pred.ci[, "fit"], lwd = 2)
lines(xval, 2^Pred.ci[, "lwr"], lty = 2, col = "red", lwd = 2)
lines(xval, 2^Pred.ci[, "upr"], lty = 2, col = "red", lwd = 2)
```



# Dos and don'ts with polynomial regression

- It is often a good idea to center  $x$ :  $x' = (x - \bar{x})$  before fitting the models
- Retain the simplest order polynomial that fits the data (reasonably) well
- But: retain  $\beta_{p-1}$  in the model if  $\beta_p$  is present

# Overview

- 1 Simple Linear Regression
- 2 Estimation
  - Uncertainty of the Estimates
- 3 Model Check
  - Residual Analysis
  - Influential Observations
- 4 Confidence and Prediction Limits
- 5 Transformation of Data
- 6 Exercises

# Exercises for “simple” linear regression

- ① PCB in trouts  
⇒ Transformations, multiplicative models, residuals, linear regression models in R
- ② Brain weight (If more time)  
⇒ Transformations outliers predictions

# Linear Regression - Part 2

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 7th, 2025

## Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Learning objectives

After this session you should be able to:

- ① Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data

# Learning objectives

After this session you should be able to:

- ① Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- ② Interpret the result from a *multiple linear regression*

# Learning objectives

After this session you should be able to:

- ① Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- ② Interpret the result from a *multiple linear regression*
- ③ Understand and use interactions.

# Learning objectives

After this session you should be able to:

- ① Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- ② Interpret the result from a *multiple linear regression*
- ③ Understand and use interactions.
- ④ Do model reduction as backwards selection.

# Overview

## 1 Multiple Linear Regression

## 2 Estimation

- First MLR in Example

## 3 Building a MLR

- GAM
- Interaction
- Trees

## 4 New model for the ozone data

## 5 Model Check

## 6 Testing

## 7 Exercises

# Multiple Linear Regression

- The association between several continuous variables.
- $Y$  response / outcome / dependent variable
- $X_1, \dots, X_p$  explanatory / covariates / independent variables.

# Data

Observations of sets  $(x_{1i}, \dots, x_{pi}, y_i)$  for all  $i = 1, \dots, n$  individuals or units.

Unit	$x_1$	$x_2$	$\cdots$	$x_p$	$y$
1	$x_{11}$	$x_{12}$	$\cdots$	$x_{1p}$	$y_1$
2	$x_{21}$	$x_{22}$	$\cdots$	$x_{2p}$	$y_2$
2	$x_{31}$	$x_{32}$	$\cdots$	$x_{3p}$	$y_3$
:	:	:	$\cdots$	:	:
$n$	$x_{n1}$	$x_{n2}$	$\cdots$	$x_{np}$	$y_n$

# The Multiple Linear Regression (MLR) model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$$

Aim: Identify one (or several) reasonable model(s) that are:

- As simple as possible
- Captures the relevant structures in the data

General rule: **Keep variables that contribute — drop variables that don't**

# Important Issues to Consider

- Which explanatory variables to include
- Curvature in the response to the explanatory variables
- Interactions between explanatory variables (will return to this)
- Correlation between explanatory variables

# Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:

# Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- Example: Drowning and ice cream sales
  - ① It seems that the higher the sales of ice cream the more drowning accidents.

# Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- Example: Drowning and ice cream sales
  - ① It seems that the higher the sales of ice cream the more drowning accidents.
  - ② Is this because people eat ice cream at the beach, and then cannot swim?

# Use and Abuse of Multiple Linear Regression?

- Multiple Linear Regression may correspond to the scientific question of interest.
- With multiple explanatory variables, the predictions become more precise (more of the variability is explained).
- Assume that one of the explanatory variables is of greatest interest. There may be another variable connected to both the variable of interest and the outcome:
- Example: Drowning and ice cream sales
  - ① It seems that the higher the sales of ice cream the more drowning accidents.
  - ② Is this because people eat ice cream at the beach, and then cannot swim?
  - ③ Or is there a 3rd variable (season) influencing both sale of ice cream and drowning accidents?

## Example: Air pollution studies

How is ozone concentration related to wind speed, air temperature and solar radiation?

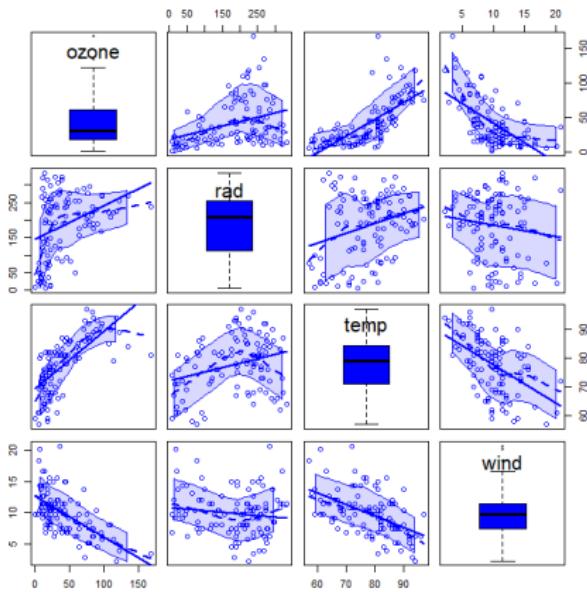
- We have 111 observations of ozone, wind speed, temperature and radiation.

The outcome (response)  $Y$  is the ozone concentration and the explanatory variables are  $X_1$  radiation (rad),  $X_2$  temperature (temp) and  $X_3$  wind speed.

Always start by plotting the data!

# Scatter plot

```
library(car)
scatterplotMatrix(~ ozone + rad + temp + wind,
                  diagonal = list(method="boxplot"), data = oz)
```



# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

# The Regression Model

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi} + \varepsilon_i$$

Traditional assumptions:

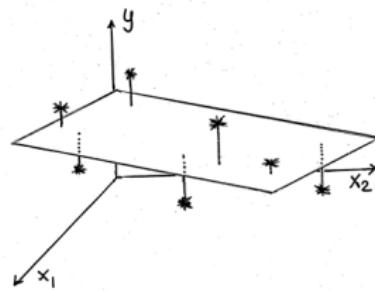
$\varepsilon_i \sim N(0, \sigma^2)$ , independent.

Least squares:

Find the  $\beta_0, \beta_1, \dots, \beta_p$  to minimize  
the sum of the squared distances:

$$SS(\beta_0, \beta_1, \dots, \beta_p) =$$

$$\sum (y_i - (\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}))^2$$



# Matrix Notation

If  $n = 6$  and  $p = 3$  then we can write the model using matrix notation:

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \\ 1 & x_{51} & x_{52} & x_{53} \\ 1 & x_{61} & x_{62} & x_{63} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

Using compact notation we have:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

# Estimation

Using Least Squares method for estimation we get:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

The estimated uncertainty on the estimate (variance):

$$var(\hat{\beta}) = \sigma^2 (X^T X)^{-1}$$

When we have estimates for  $\beta_0, \beta_1 \dots, \beta_p$  then we can calculate the expected values for the outcome:

$$\hat{y} = X \hat{\beta}$$

The value  $\hat{y}_i$  is called the fitted value, or expected value. This corresponds to the value on the regression line.

## Estimation continued

As for simple linear regression we also have the **residuals** (what is left):

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

Using the matrix notation:

$$\hat{\varepsilon} = \mathbf{y} - \hat{\mathbf{y}}$$

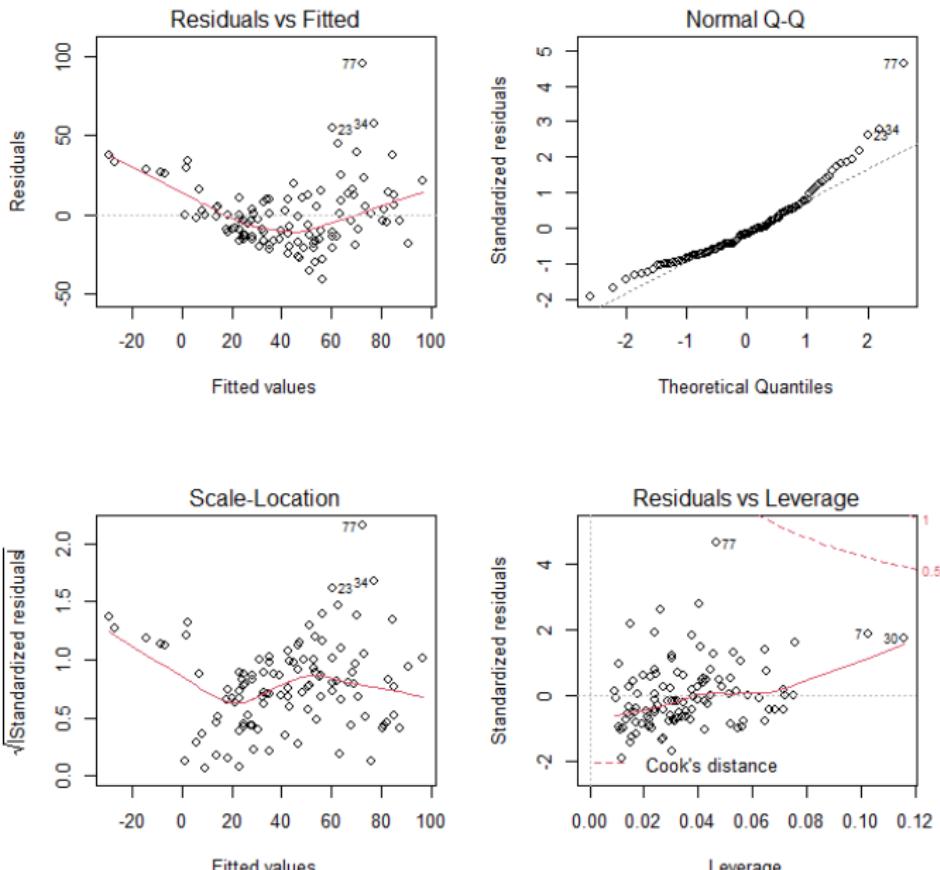
The model variance  $\sigma^2$  is estimated:

$$\hat{\sigma}^2 = s^2 = \frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n - (p + 1)} = MSE$$

# Multiple Linear Regression

```
reg1 <- lm(ozone ~ rad + temp + wind, data = oz)
summary(reg1)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.2321   23.0420  -2.79  0.0063 **
## rad          0.0598    0.0232   2.58  0.0112 *
## temp         1.6512    0.2534   6.52  2.4e-09 ***
## wind        -3.3376    0.6538  -5.10  1.4e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.2 on 107 degrees of freedom
## Multiple R-squared:  0.606, Adjusted R-squared:  0.595
## F-statistic: 54.9 on 3 and 107 DF,  p-value: <2e-16
```



# Interpreting the Estimates

- It is always important to interpret the model parameters
  - give an explanation in words of the results.

## Interpreting the Estimates

- It is always important to interpret the model parameters
  - give an explanation in words of the results.
- $\hat{\beta}_i$  is the effect of variable  $X_i$  on  $Y$ , when all other variables are fixed.

## Interpreting the Estimates

- It is always important to interpret the model parameters
  - give an explanation in words of the results.
- $\hat{\beta}_i$  is the effect of variable  $X_i$  on  $Y$ , when all other variables are fixed.
- $\hat{\beta}_i$  is a slope. The expected change in  $Y$  when  $X_i$  changes one unit and the remaining variables are unchanged.

## Interpreting the Estimates

- It is always important to interpret the model parameters
  - give an explanation in words of the results.
- $\hat{\beta}_i$  is the effect of variable  $X_i$  on  $Y$ , when all other variables are fixed.
- $\hat{\beta}_i$  is a slope. The expected change in  $Y$  when  $X_i$  changes one unit and the remaining variables are unchanged.
- The effect is corrected for the effect of the other explanatory variables.

# The Estimates from the MLR

	Estimate	Std..Error	Lower	Upper	p.value
## (Intercept)	-64.23	23.04	-109.91	-18.55	0.00628
## rad	0.06	0.02	0.01	0.11	0.01124
## temp	1.65	0.25	1.15	2.15	< 0.001
## wind	-3.34	0.65	-4.63	-2.04	< 0.001

And the variance is estimated by:

$$\hat{\sigma}^2 = 21.2^2 = 449.44$$

The Intercept =  $\hat{\beta}_0$  is the expected ozone when wind=0, rad=0 and temp=0, not so interesting.

# The Estimates from the MLR

	Estimate	Std..Error	Lower	Upper	p.value
## (Intercept)	-64.23	23.04	-109.91	-18.55	0.00628
## rad	0.06	0.02	0.01	0.11	0.01124
## temp	1.65	0.25	1.15	2.15	< 0.001
## wind	-3.34	0.65	-4.63	-2.04	< 0.001

And the variance is estimated by:

$$\hat{\sigma}^2 = 21.2^2 = 449.44$$

The Intercept =  $\hat{\beta}_0$  is the expected ozone when wind=0, rad=0 and temp=0, not so interesting.

temp =  $\hat{\beta}_2$  is a slope. The ozone level increases by 1.65 when the temperature increases by 1 **for fixed wind and radiation**.

# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

# Building a MLR

- Perhaps this model was too simple.
- We want to include radiation, temperature and wind but we don't know whether it is reasonable with linear effects.
- Perhaps we need a curve.

# GAM

- This morning we talked about adding squared terms of continuous explanatory variables. After looking at residual plots and seeing a non-random shape.

# GAM

- This morning we talked about adding squared terms of continuous explanatory variables. After looking at residual plots and seeing a non-random shape.
- In many cases, however, we have one or more continuous explanatory variables, but no a priori reason to choose one particular parametric form over another for describing the shape of the relationship between the response variable and the explanatory variable(s).

# GAM

- This morning we talked about adding squared terms of continuous explanatory variables. After looking at residual plots and seeing a non-random shape.
- In many cases, however, we have one or more continuous explanatory variables, but no a priori reason to choose one particular parametric form over another for describing the shape of the relationship between the response variable and the explanatory variable(s).
- Generalized additive models (GAMs) are useful in such cases because they allow us to capture the shape of a relationship between  $y$  and  $x$  without having to chose a particular parametric form beforehand.

# GAM

- We are replacing the linear form in the regression model

$$\sum_j \beta_j X_j$$

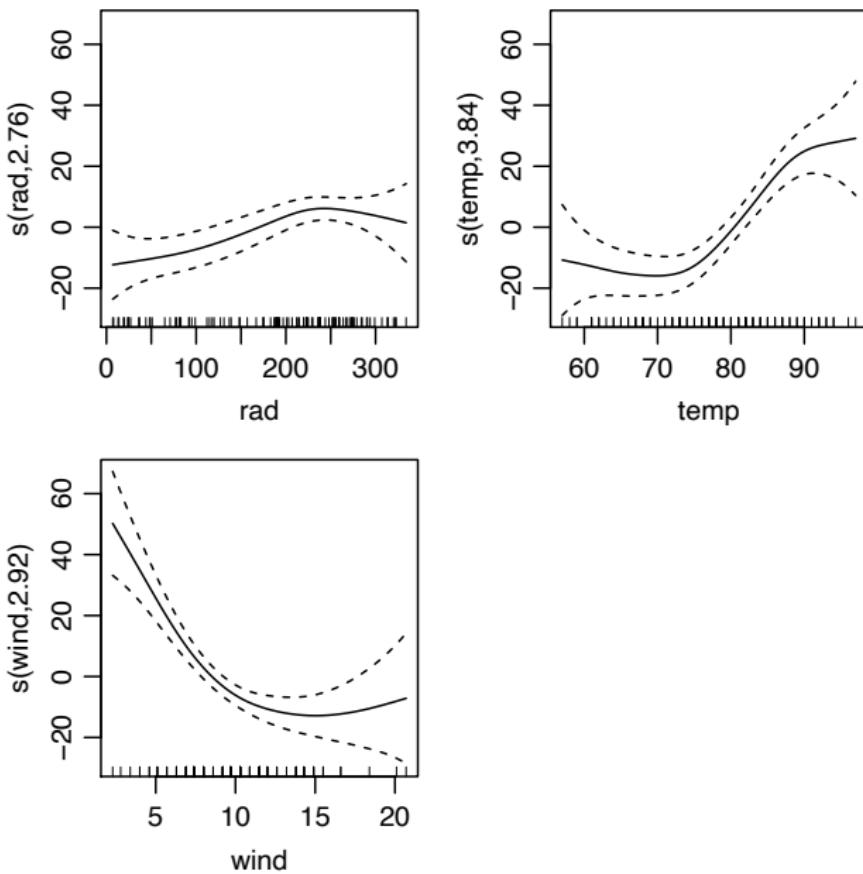
- By the sum of smooth functions

$$\sum_j s_j(X_j)$$

- The functions  $s_j$  are unspecified smooth functions estimated using a non-parametric smoother

# GAM for the ozone example

```
library(mgcv)
par(mfrow = c(2,2), mgp = c(2,0.7,0), mar =
c(3,3,1,1))
model <- gam(ozone ~ s(rad) + s(temp) + s(wind), data = oz)
plot(model)
par(mfrow = c(1,1))
```



# Ideas from GAM

- The confidence intervals are sufficiently narrow to suggest that the curvature in the relationship between ozone and temperature is real
- The curvature of the relationship with wind is questionable
- A linear model may well be all that is required for solar radiation

# What if the effect of temperature depends on wind speed?

We had the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Where  $y_i$  is the observed ozone concentration  $i$ ,  $x_{1i}$  the radiation,  $x_{2i}$  the temperature and  $x_{3i}$  the wind speed.

We are assuming that temperature and wind have an **additive effect** on the ozone concentration.

The effect of temperature is assumed the same for all wind speeds.

# What if the effect of temperature depends on wind speed?

We had the model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

Where  $y_i$  is the observed ozone concentration  $i$ ,  $x_{1i}$  the radiation,  $x_{2i}$  the temperature and  $x_{3i}$  the wind speed.

We are assuming that temperature and wind have an **additive effect** on the ozone concentration.

The effect of temperature is assumed the same for all wind speeds.

Perhaps the additive model is too simple, we can include a **multiplicative term** (also called an interaction).

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4(x_{2i} \cdot x_{3i}) + \varepsilon_i$$

where  $\beta_4$  accounts for the **interaction** between temperature and wind.

# Taylor expansion of real functions

Why is a multiplicative term a relevant idea?

- Power series expansion of a smooth one-dimensional function:

$$f(x) = f(x_0) + \sum_{n=1}^{\infty} a_n (x - x_0)^n, \quad a_n = f^{(n)}(x_0) n!$$

- 1<sup>st</sup> order Taylor expansion:

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0) \cdot (x - x_0) + r_1(x - x_0) \\ &= \underbrace{f(x_0) - f'(x_0) \cdot x_0}_{\alpha} + \underbrace{f'(x_0) \cdot x}_{\beta} + r_1(x - x_0) \\ &= \alpha + \beta x + r_1(x - x_0) \end{aligned}$$

# Taylor expansion of real functions

- If the relation between  $Y$  and  $x$  is really  $f$ :

$$Y = \alpha + \beta x + \varepsilon$$

$\varepsilon = r_1(x - x_0) + \epsilon$  covers both model aberrations and stochasticity.

- If the model aberration is too big to be handled by the general uncertainty  $\varepsilon$ , we may resort to a finer model description, Taylor expansion to the 2<sup>nd</sup> order (here in arbitrary dimensions):

$$\begin{aligned} f(x) = & f(x_0) + \langle f'(x_0), x - x_0 \rangle \\ & + \frac{1}{2}(x - x_0)^T \cdot f''(x_0) \cdot (x - x_0) + r_2(x - x_0) \end{aligned}$$

- The matrix in the second term contains the coefficients to the 2<sup>nd</sup> order multiplicative model terms.

# Trees

- We need to get some ideas about which interactions to include.

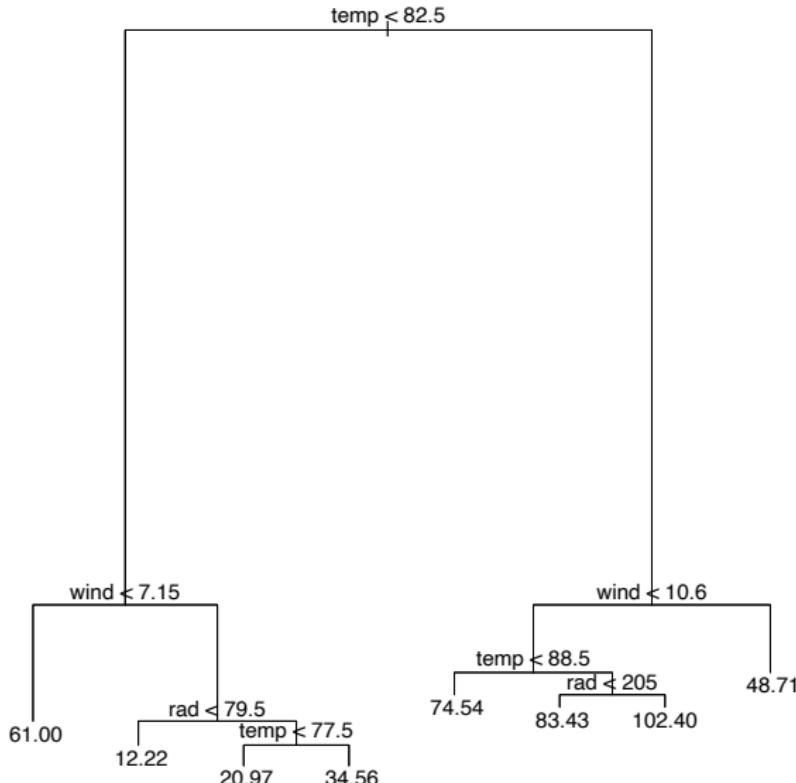
# Trees

- We need to get some ideas about which interactions to include.
- Trees can help identify interactions
- Good for initial data inspection
- The splits that gives the largest reduction in variance for each part.
- At the leaves we have the mean value in that subset of the data

# Tree in example

```
library(tree)
model<-tree(ozone~., data = oz)
plot(model)
text(model)
```

# Tree in example



# Trees

- If the number of covariates is large, the single tree method will no longer work. One will then need to resort to sparse methodology.
- One such bases itself on the Random Forest methodology, selecting subsets of covariates at random and construct corresponding trees.
- Selection of interacting terms, after correcting for correlation/dependence, can be made following *Behr et al 2022*,  
<https://doi.org/10.1073/pnas.2118636119>.
- The procedure by *Behr et al 2022* is, however, outside the scope of this course.

# Ideas from the tree

- Temperature is by far the most important
- Wind speed important at both high and low values. Low wind is associated to higher mean ozone levels.
- Possible interaction between wind and temperature and wind and radiation.

# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

# New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\&\quad + \beta_6(x_{2i} \cdot x_{3i}) + \beta_7(x_{1i} \cdot x_{3i}) + \varepsilon_i\end{aligned}$$

- $y_i = \text{ozone}$ ,  $x_{1i} = \text{radiation}$ ,  $x_{2i} = \text{temp}$ ,  $x_{3i} = \text{wind}$  and  $\varepsilon_i \sim N(0, \sigma^2)$ .

# New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\&\quad + \beta_6(x_{2i} \cdot x_{3i}) + \beta_7(x_{1i} \cdot x_{3i}) + \varepsilon_i\end{aligned}$$

- $y_i$ =ozone,  $x_{1i}$ =radiation,  $x_{2i}$ =temp,  $x_{3i}$ =wind and  $\varepsilon_i \sim N(0, \sigma^2)$ .
- The next step is to simplify the model

# New model for the ozone data

- We are now ready with a more complex model for the ozone data.
- We will include curvature for wind and temperature and interactions between wind and temperature and wind and radiation.

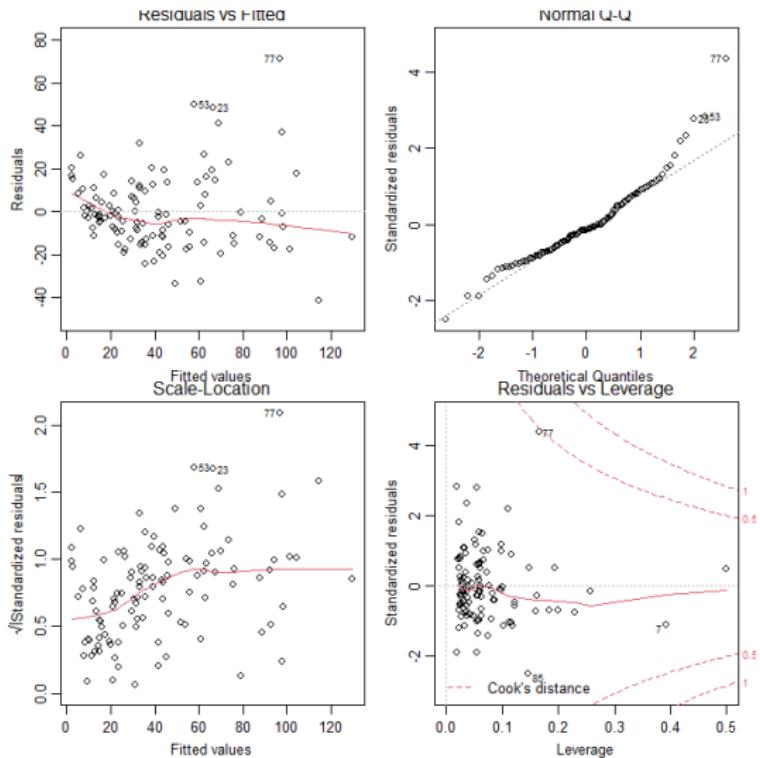
$$\begin{aligned}y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{2i}^2 + \beta_5 x_{3i}^2 \\&\quad + \beta_6(x_{2i} \cdot x_{3i}) + \beta_7(x_{1i} \cdot x_{3i}) + \varepsilon_i\end{aligned}$$

- $y_i$ =ozone,  $x_{1i}$ =radiation,  $x_{2i}$ =temp,  $x_{3i}$ =wind and  $\varepsilon_i \sim N(0, \sigma^2)$ .
- The next step is to simplify the model
- We must not forget to check the underlying assumptions!

# The new model

```
reg2 <- lm(ozone ~ rad + temp + wind + I(temp^2) +
            I(wind^2) + temp:wind + rad:wind, data = oz)
summary(reg2)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 514.40147  193.78358    2.65  0.0092 **
## rad          0.21295   0.06928    3.07  0.0027 **
## temp        -10.65404   4.09489   -2.60  0.0106 *
## wind         -27.39197   9.61700   -2.85  0.0053 **
## I(temp^2)    0.06780   0.02241    3.03  0.0031 **
## I(wind^2)    0.61940   0.14577    4.25  4.7e-05 ***
## temp:wind    0.16967   0.09446    1.80  0.0754 .
## rad:wind    -0.01356   0.00609   -2.23  0.0281 *
##
## Residual standard error: 17.9 on 103 degrees of freedom
## Multiple R-squared:  0.729, Adjusted R-squared:  0.711
## F-statistic: 39.6 on 7 and 103 DF,  p-value: <2e-16
```



## Work with the person next to you

- Use the model with interaction.
- What is the expected ozone concentration:
  - If the level of radiation is 100, temperature is 60 and wind speed is 20.
  - If the level of radiation is 185, temperature is 80 and wind speed is 10.
  - If the level of radiation is 185, temperature is 80 and wind speed is 5.

# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

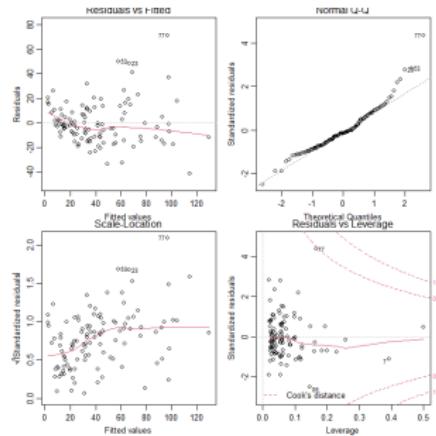
# Model Check in detail

As in the simple linear regression we must check our model assumptions before we interpret our model too much. We have to check:

- Normal residuals (observed - fitted), using qq-plots.
- Variance homogeneity (one  $\sigma^2$ ), residual plots against fitted values.
- Linear effect of  $X_1, \dots, X_p$ , residual plots against each covariate.

# Model Check in example

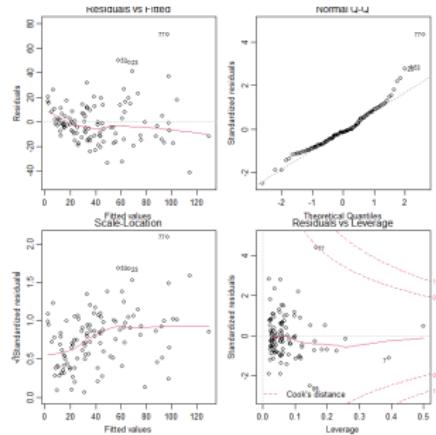
```
par(mfrow=c(2, 2))
plot(reg2, which=1:4)
```



- The model check still did not look good!
- Problem with variance homogeneity.

# Model Check in example

```
par(mfrow=c(2, 2))
plot(reg2, which=1:4)
```

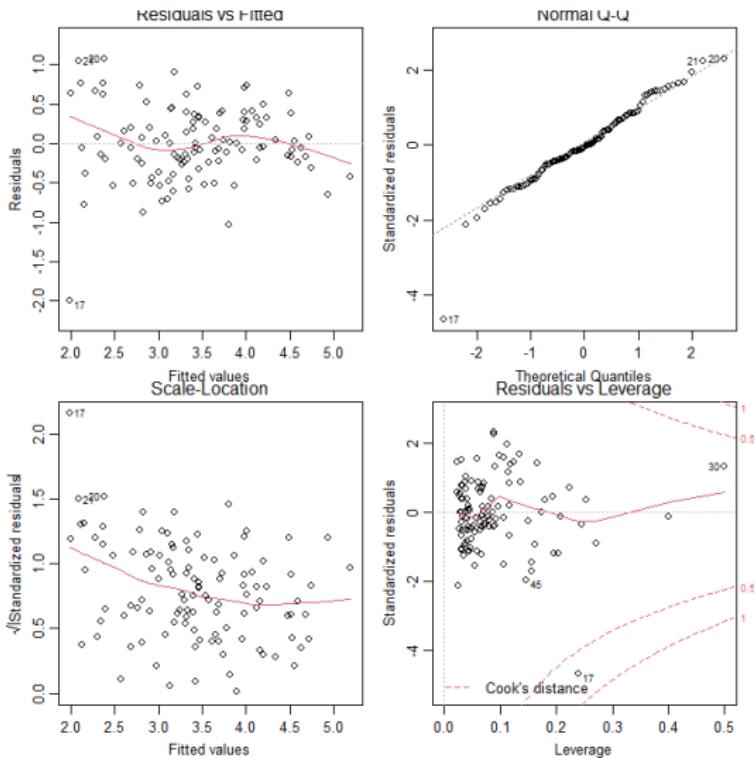


- The model check still did not look good!
- Problem with variance homogeneity.
- What should we do?

# New model for $\log(\text{ozone})$

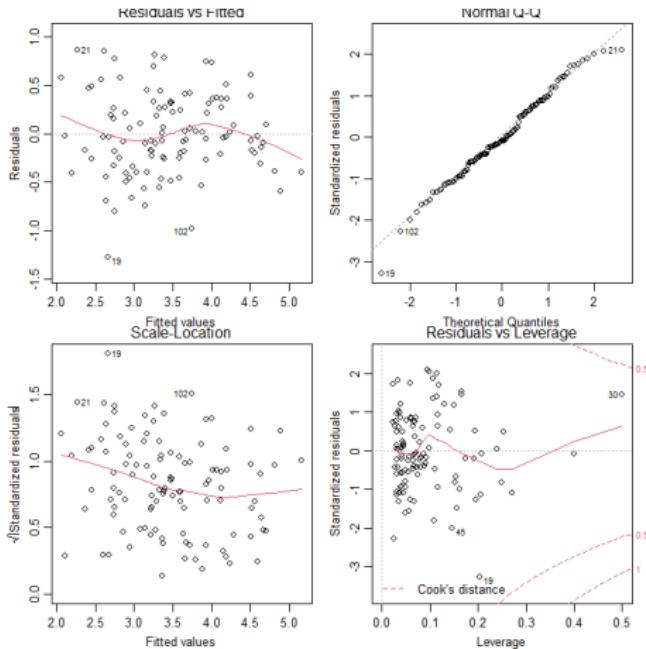
- We need to start from scratch with all the original explanatory variables included.
- We would expect the curvature to have changed.
- We can run a new GAM and do a new Tree;
- With just a few explanatory variables, we can also choose brute force and include all 2nd order effects.
- The new starting model should be:

```
reg3 <- lm(log(ozone) ~ rad + temp + wind + I(temp^2) +  
           I(wind^2) + I(rad^2) +  
           rad:temp + rad:wind +  
           temp:wind, data = oz)
```



# Removing Outlier

```
reg3 <- update(reg3, data=oz[-17,])
```



# Model Estimates

```
summary(reg3)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.378e+01	5.064e+00	2.721	0.00768	**
rad	-1.670e-05	5.470e-03	-0.003	0.99757	
temp	-2.208e-01	1.076e-01	-2.051	0.04287	*
wind	-6.291e-01	2.425e-01	-2.594	0.01092	*
I(temp^2)	1.270e-03	6.014e-04	2.111	0.03725	*
I(wind^2)	1.038e-02	3.608e-03	2.876	0.00493	**
I(rad^2)	-1.353e-05	6.260e-06	-2.161	0.03308	*
rad:temp	9.794e-05	6.208e-05	1.578	0.11780	
rad:wind	-3.984e-05	1.552e-04	-0.257	0.79794	
temp:wind	4.429e-03	2.385e-03	1.857	0.06628	.

# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

# Testing

- When we finally have a satisfactory starting model then we often want to simplify.
- Sometimes we have **specific questions**, i.e.
  - Does the crime rate depend on the level of education?
  - Did the intervention make the children eat more healthily?
- Other times we have a lot of variables and are mainly looking for **structures** in the data.

# Model selection

- It is **not trivial** to chose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.

# Model selection

- It is **not trivial** to chose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we **will** often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.

# Model selection

- It is **not trivial** to chose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we **will** often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.
- Be **cautious when using automated approaches**. If the number of variables is not too large, then **it is** better to think it through and keep an eye on what **is** happening.

# Model selection

- It is **not trivial** to chose a good starting model. Often it is a combination of theoretical knowledge and/or practical experience, and then the hypotheses we want to test.
- Sometimes we have a large amount of data with many variables and not much knowledge. Here we **will** often take a more exploratory approach with some sensible strategies, and an automated approach when looking for a model.
- Be **cautious when using automated approaches**. If the number of variables is not too large, then **it is** better to think it through and keep an eye on what **is** happening.
- A rule of thumb: The number of parameters in the model should be less than (number of observations)/5.

# Backwards and Forwards selection

## Backwards

- Start with the largest model, the most complex, and remove variables which are not significant, **one at a time**. Continue until all variables are significant.

# Backwards and Forwards selection

## Backwards

- Start with the largest model, the most complex, and remove variables which are not significant, **one at a time**. Continue until all variables are significant.

## Forwards

- Start with the model that only includes an intercept. Add a variable one at a time starting with the most significant. Continue until none of the remaining variables are significant.

# Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

# Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

# Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

Never remove a lower order term if it is part of a higher order term

# Tests of main effects and interactions

Never remove a main effect if it is part of an interaction

In R the function `drop1()` will help you not to make mistakes.

Never remove a lower order term if it is part of a higher order term

You don't know if there is (evidence of) an interaction unless you look for it

# Model reduction in example

```
drop1(reg3,test="F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
    rad:temp + rad:wind + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.020	-173.05			
I(temp^2)	1	0.84769	19.867	-170.25	4.4570	0.037253 *
I(wind^2)	1	1.57299	20.593	-166.31	8.2704	0.004925 **
I(rad^2)	1	0.88826	19.908	-170.03	4.6703	0.033077 *
rad:temp	1	0.47342	19.493	-172.35	2.4891	0.117796
rad:wind	1	0.01253	19.032	-174.98	0.0659	0.797937
temp:wind	1	0.65574	19.675	-171.32	3.4477	0.066284 .

---

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

# Model reduction in example

```
drop1(reg3,test="F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
    rad:temp + rad:wind + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.020	-173.05			
I(temp^2)	1	0.84769	19.867	-170.25	4.4570	0.037253 *
I(wind^2)	1	1.57299	20.593	-166.31	8.2704	0.004925 **
I(rad^2)	1	0.88826	19.908	-170.03	4.6703	0.033077 *
rad:temp	1	0.47342	19.493	-172.35	2.4891	0.117796
rad:wind	1	0.01253	19.032	-174.98	0.0659	0.797937
temp:wind	1	0.65574	19.675	-171.32	3.4477	0.066284 .
<hr/>						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

Notice the use of `test="F"`. We will remove `rad:wind`.

# Model reduction in example

```
reg4 <- update(reg3, ~. -rad:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

$\text{log(ozone)} \sim \text{rad} + \text{temp} + \text{wind} + I(\text{temp}^2) + I(\text{wind}^2) + I(\text{rad}^2) + \text{rad:temp} + \text{temp:wind}$

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.032	-174.98			
$I(\text{temp}^2)$	1	0.83641	19.869	-172.25	4.4387	0.037611 *
$I(\text{wind}^2)$	1	1.56240	20.595	-168.30	8.2914	0.004864 **
$I(\text{rad}^2)$	1	0.88847	19.921	-171.96	4.7150	0.032243 *
$\text{rad:temp}$	1	0.56397	19.596	-173.77	2.9929	0.086687 .
$\text{temp:wind}$	1	0.66581	19.698	-173.20	3.5333	0.063029 .

---

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

# Model reduction in example

```
reg4 <- update(reg3, ~. -rad:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
    rad:temp + temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.032	-174.98			
I(temp^2)	1	0.83641	19.869	-172.25	4.4387	0.037611 *
I(wind^2)	1	1.56240	20.595	-168.30	8.2914	0.004864 **
I(rad^2)	1	0.88847	19.921	-171.96	4.7150	0.032243 *
rad:temp	1	0.56397	19.596	-173.77	2.9929	0.086687 .
temp:wind	1	0.66581	19.698	-173.20	3.5333	0.063029 .

---

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

We will remove rad:temp. We will continue like this until all variables

# Model reduction in example

```
reg4 <- update(reg4, ~. -rad:temp)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
    temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.596	-173.77			
rad	1	1.58795	21.184	-167.20	8.2655	0.004919 **
I(temp^2)	1	1.27863	20.875	-168.81	6.6554	0.011310 *
I(wind^2)	1	1.60681	21.203	-167.10	8.3636	0.004679 **
I(rad^2)	1	0.69249	20.289	-171.95	3.6045	0.060450 .
temp:wind	1	0.57900	20.175	-172.56	3.0138	0.085580 .

---

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

# Model reduction in example

```
reg4 <- update(reg4, ~. -rad:temp)
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(temp^2) + I(wind^2) + I(rad^2) +
    temp:wind
```

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		19.596	-173.77			
rad	1	1.58795	21.184	-167.20	8.2655	0.004919 **
I(temp^2)	1	1.27863	20.875	-168.81	6.6554	0.011310 *
I(wind^2)	1	1.60681	21.203	-167.10	8.3636	0.004679 **
I(rad^2)	1	0.69249	20.289	-171.95	3.6045	0.060450 .
temp:wind	1	0.57900	20.175	-172.56	3.0138	0.085580 .

---

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

We will remove temp:wind.

# Model reduction in example

```
reg4 <- update(reg4, ~. -temp:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		20.175	-172.56			
rad	1	1.44844	21.623	-166.94	7.3948	0.0076784 **
temp	1	0.33447	20.509	-172.75	1.7076	0.1942126
wind	1	2.32037	22.495	-162.59	11.8462	0.0008361 ***
I(temp^2)	1	0.69977	20.875	-170.81	3.5725	0.0615534 .
I(wind^2)	1	1.09518	21.270	-168.75	5.5913	0.0199242 *
I(rad^2)	1	0.57877	20.754	-171.45	2.9548	0.0886278 .
<hr/>						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

# Model reduction in example

```
reg4 <- update(reg4, ~. -temp:wind)
drop1(reg4, test = "F")
```

Single term deletions

Model:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)
<none>		20.175	-172.56			
rad	1	1.44844	21.623	-166.94	7.3948	0.0076784 **
temp	1	0.33447	20.509	-172.75	1.7076	0.1942126
wind	1	2.32037	22.495	-162.59	11.8462	0.0008361 ***
I(temp^2)	1	0.69977	20.875	-170.81	3.5725	0.0615534 .
I(wind^2)	1	1.09518	21.270	-168.75	5.5913	0.0199242 *
I(rad^2)	1	0.57877	20.754	-171.45	2.9548	0.0886278 .
<hr/>						
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1						

We will remove the squared effect of rad.

# Model reduction in example

```
reg4 <- update(reg4, ~. - I(rad^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)					
<none>		20.754	-171.45								
rad	1	4.1707	24.924	-153.31	20.8996	1.339e-05 ***					
temp	1	0.2728	21.027	-172.02	1.3669	0.2450244					
wind	1	2.3402	23.094	-161.70	11.7269	0.0008827 ***					
I(temp^2)	1	0.6390	21.393	-170.12	3.2022	0.0764518 .					
I(wind^2)	1	1.0639	21.818	-167.95	5.3312	0.0229249 *					
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

# Model reduction in example

```
reg4 <- update(reg4, ~. - I(rad^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

	Df	Sum of Sq	RSS	AIC	F value	Pr(>F)					
<none>		20.754	-171.45								
rad	1	4.1707	24.924	-153.31	20.8996	1.339e-05 ***					
temp	1	0.2728	21.027	-172.02	1.3669	0.2450244					
wind	1	2.3402	23.094	-161.70	11.7269	0.0008827 ***					
I(temp^2)	1	0.6390	21.393	-170.12	3.2022	0.0764518 .					
I(wind^2)	1	1.0639	21.818	-167.95	5.3312	0.0229249 *					
---											
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

We will remove the squared effect of temp.

# Model reduction in example

```
reg4 <- update(reg4, ~. -I(temp^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(wind^2)
             Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>            21.393 -170.12
rad        1     3.9973 25.390 -153.27 19.6192 2.325e-05 ***
temp       1    11.5647 32.958 -124.58 56.7617 1.807e-11 ***
wind       1     3.3253 24.718 -156.22 16.3212 0.000102 ***
I(wind^2)  1     1.6759 23.069 -163.82  8.2258  0.004993 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Model reduction in example

```
reg4 <- update(reg4, ~. -I(temp^2))
drop1(reg4, test = "F")
```

Single term deletions

Model:

```
log(ozone) ~ rad + temp + wind + I(wind^2)
             Df Sum of Sq    RSS      AIC F value    Pr(>F)
<none>            21.393 -170.12
rad        1     3.9973 25.390 -153.27 19.6192 2.325e-05 ***
temp       1    11.5647 32.958 -124.58 56.7617 1.807e-11 ***
wind       1     3.3253 24.718 -156.22 16.3212 0.000102 ***
I(wind^2)  1     1.6759 23.069 -163.82  8.2258  0.004993 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We will stop the model reduction as all variables are statistically significant.

## Words of caution

- Do not blindly use automatic stepwise variable selection procedures
- Don't confuse and combine model search and selection with *confirmatory hypothesis testing*; we need a fitting and sensible model for the latter
- The model space is often very large — there may be more than one model that may explain the data equally well
- Always consider the use of interactions, polynomials and transformations; even though you may decide against them in the end.

## Words of caution

- Do not blindly use automatic stepwise variable selection procedures
- Don't confuse and combine model search and selection with *confirmatory hypothesis testing*; we need a fitting and sensible model for the latter
- The model space is often very large — there may be more than one model that may explain the data equally well
- Always consider the use of interactions, polynomials and transformations; even though you may decide against them in the end.

Model selection is an art — it takes practice to master

# Final Model

We have subsequently removed  $rad : wind$ ,  $rad : temp$ ,  $temp : wind$ ,  $rad^2$ , and  $temp^2$  through backwards selection. Coefficients in final model:

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	1.1932358	0.5990022	1.992	0.048963	*
rad	0.0022097	0.0004989	4.429	2.33e-05	***
temp	0.0419157	0.0055635	7.534	1.81e-11	***
wind	-0.2208189	0.0546589	-4.040	0.000102	***
I(wind^2)	0.0068982	0.0024052	2.868	0.004993	**
---					
Signif. codes:	0	'***'	0.001	'**'	0.01
	'*'	0.05	'.'	0.1	' '
	1				

# Common issues arising in MLR

- **Differences in the measurement scales** of the explanatory variables, leading to large variation in the sums of squares and hence to an ill-conditioned matrix.
  - Can consider **standardizing**, i.e. subtracting the mean and dividing by the standard deviation.
- **Multicollinearity**, in which there is a near-linear relation between two of the explanatory variables (nearly the same information), leading to unstable parameter estimates.
  - Perhaps choose only one of several colinear variables, or use PCA (Lecture 10)
- **Parameter proliferation** where quadratic and interaction terms soak up more degrees of freedom than our data can afford.
  - Careful selection of interaction and quadratic terms, for example through the methods discussed today, trees and GAM.

# Learning objectives

After this session you should be able to:

- ① Understand what a *multiple linear regression* (MLR) models is and be able to fit it to data
- ② Interpret the result from a *multiple linear regression*
- ③ Understand and use interactions.
- ④ Do backwards selection.

# Overview

1 Multiple Linear Regression

2 Estimation

- First MLR in Example

3 Building a MLR

- GAM
- Interaction
- Trees

4 New model for the ozone data

5 Model Check

6 Testing

7 Exercises

## 2 Exercises for Multiple linear regression

- ① Process: Understand process loss as a function of other continuous variables
- ② Cheese: Describe the taste of matured cheese as a function of chemical descriptors.

# Case: Cheese

## Story

As cheddar cheese matures, a variety of chemical processes take place. The taste of matured cheese is related to the concentration of several chemicals in the final product. In a study of cheddar cheese from the LaTrobe Valley of Victoria, Australia, samples of cheese were analyzed for their chemical composition and subjected to taste tests. Overall taste scores were obtained by combining scores from several tasters. The data are available in `cheese.txt`.

## Data

Variable	Description
case	Sample number
taste	Subjective taste test score, obtained by combining scores of several tasters
acetic	Natural log of concentration of acetic acid
h2s	Natural log of concentration of hydrogen sulfide ( $H_2S$ )
lactic	Concentration of lactic acid

## Exercise

1. Use scatter plots and simple linear regression to examine which variables influence the taste score.
2. Fit a multiple regression model to explain taste as a function of the other variables.
3. Determine a reasonable model for taste considering transformations, interactions and quadratic terms.
4. Perform model diagnostics and revise your model if needed.
5. Predict the taste score of a cheese where log acetic is 5.3, log  $H_2S$  concentration is 8.0 and lactic acid concentration is 3.0.
6. What is the potential problem with this prediction?
7. Save the code that performs your analysis in a script and add plenty of comments to your code.

# Case: Process

## Story

The dataset process.txt contains measurements of air flow, water temperature, and acid concentration of a process loss. The aim of this case is to explain the process loss as a function of the other variables.

## Data

Variable	Description
loss	loss from process
airflow	air flow
watertemp	water temperature
acidconc	acid concentration

## Exercise

1. Plot the variables and make a graphical assessment. Which variables could be helpful in explaining process loss?
2. Using simple linear regression, assess whether air flow, water temperature and acid concentration have an influence on process loss.
3. Now use a multiple linear regression to assess the effects of air flow, water temperature and acid concentration on process loss.

Notice what happens to the significance of the variables: One of the variables was borderline significant in the simple linear regression, but is not significant in the multiple linear regression. How do you explain this?

4. Determine a reasonable model for process loss based on the variables available. Use model diagnostics/residual analyses, transformation of variables, polynomials and interactions to aid your model search.
5. Perform model diagnostics on your final model and subject it to criticism.
6. Write up a mathematical expression summarizing your final model.
7. Summarize the significance of the variables included in the model.
8. Save the code that performs your analysis in a script and add plenty of comments to your code.

# One and Two-way Analysis of Variance

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 8th, 2025

## Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Plan for today

Lecture on one-way analysis of variance

Exercise

Recap of exercise and discussion

Lecture on two-way analysis of variance and the general linear model

Exercise

Recap of exercises and discussion

# Outline, 1 and 2 Way ANOVA

## 1 One-way Analysis of Variance

- Descriptive
- ANOVA Model
- Data Example
- Model Control
- Multiple Testing
- Exercise

## 2 Two-way Analysis of Variance

- Interaction
- Interaction estimates
- Hypotheses
- Exercise

# Overview

1

## One-way Analysis of Variance

- Descriptive
- ANOVA Model
- Data Example
- Model Control
- Multiple Testing
- Exercise

2

## Two-way Analysis of Variance

- Interaction
- Interaction estimates
- Hypotheses
- Exercise

# One-way Analysis of Variance

- Comparing the means in more than two groups.
- One-way ANOVA gives joint test for any differences between the groups.
- **One-way** because only one grouping.
- **Analysis of variance** because we are comparing the variance between groups to the variance within groups.

## Example: Birth Weight (Hosmer & Lemeshow data 2000)

- \* Birth weight of 189 babies.
- \* Race is grouped in three groups: white, black and other.
- \* We want to compare the birth weight for these three groups.
- \* The first step is to describe the data using simple tables and plots.

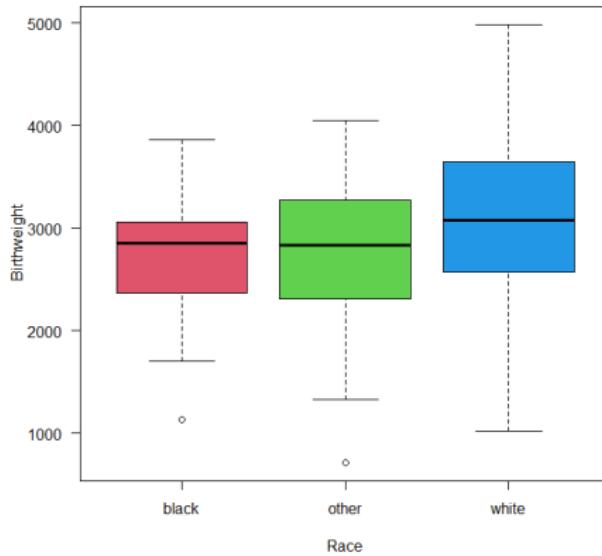
## Example: Birth Weight

```
lbw <- read.delim("lbw.txt")
by(lbw$bwt, lbw$race, summary)

## lbw$race: black
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      1140    2370   2850     2720    3060     3860
## -----
## lbw$race: other
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      709     2310   2840     2800    3270     4050
## -----
## lbw$race: white
##      Min. 1st Qu. Median      Mean 3rd Qu.      Max.
##      1020    2580   3080     3100    3650     4990
```

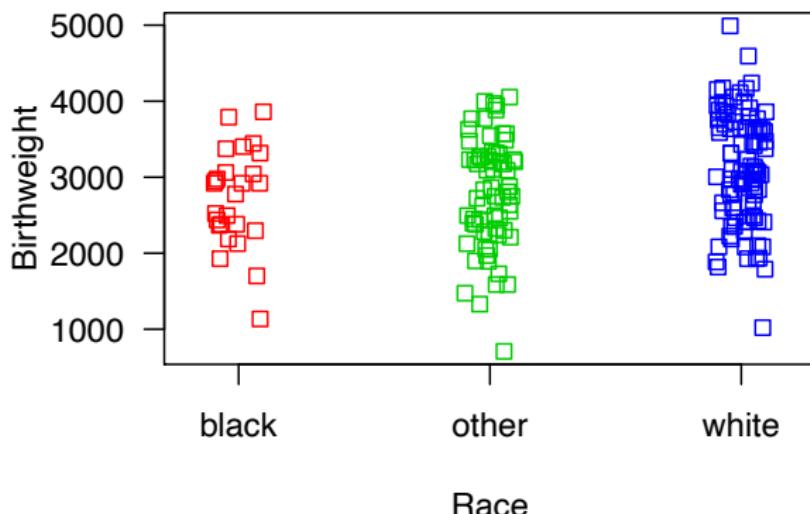
# Example: Birth Weight, Boxplot

```
boxplot(bwt ~ race, data = lbw, xlab = 'Race', ylab = 'Birthweight'  
las = 1, col = 2:4)
```



## Example: Birth Weight, Stripchart

```
stripchart(bwt ~ race, data = lbw, vertical = TRUE, xlab = "Race",  
ylab = "Birthweight", method = "jitter", las = 1, col = 2:4)
```



# One-way Analysis of Variance Model

Let  $Y_{gi}$  be birth weight for child  $i$  in group  $g$ ,  $i \in \{1, \dots, n_g\}$  and  $g \in \{1, \dots, k\}$ .

We assume:

$$\underbrace{Y_{gi}}_{\text{Observation } i \text{ in group } g} = \underbrace{\mu_g}_{\text{Mean in group } g} + \underbrace{\epsilon_{gi}}_{\text{Individual variation}}$$

The observations are assumed to follow a normal distribution within each group, with a common variance  $\sigma^2$

$$\epsilon_{gi} \sim N(0, \sigma^2), \quad Y_{gi} \sim N(\mu_g, \sigma^2)$$

These model assumptions need to be checked before interpreting the results.

# Test of Hypothesis

In a one-way analysis of variance there is only one hypothesis: **Are the group means the same?**

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

If the hypothesis is accepted it means that we can assume the simpler model:

$$Y_{gi} = \mu + \epsilon_{gi}, \quad i \in \{1, \dots, n_g\}, \quad g \in \{1, \dots, k\}$$

# ANOVA math: Sums of Squares

Decomposition of deviation from grand mean:

$$y_{gi} - \bar{y}_{\cdot} = (y_{gi} - \bar{y}_{g\cdot}) + (\bar{y}_{g\cdot} - \bar{y}_{\cdot})$$

Decomposition of variation (sum of squares, SS).

$$\underbrace{\sum_{gi} (y_{gi} - \bar{y}_{\cdot})^2}_{\text{Total variation}} = \underbrace{\sum_{gi} (y_{gi} - \bar{y}_{g\cdot})^2}_{\text{Within groups}} + \underbrace{\sum_{gi} (\bar{y}_{g\cdot} - \bar{y}_{\cdot})^2}_{\text{Between groups}}$$

$y_{gi}$   $i^{th}$  observation in group  $g$ .

$\bar{y}_{g\cdot}$  Average in group  $g$ .

$\bar{y}_{\cdot}$  Overall average or "grand mean".

# Decomposition of Variation

Total = Between + Within

$$SS_{Total} = SS_{Between} + SS_{Within}$$

$$N - 1 = (k - 1) + (N - k)$$

F-test statistic

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{SS_{Between}/(k - 1)}{SS_{Within}/(N - k)}$$

# Decomposition of Variation

Total = Between + Within

$$\begin{aligned}SS_{Total} &= SS_{Between} + SS_{Within} \\N - 1 &= (k - 1) + (N - k)\end{aligned}$$

F-test statistic

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{SS_{Between}/(k - 1)}{SS_{Within}/(N - k)}$$

## Hypothesis test

Reject null hypothesis if F large, i.e. if the variation *between* groups is too large compared to the variation *within* groups.

# Decomposition of Variation

Total = Between + Within

$$\begin{aligned}SS_{Total} &= SS_{Between} + SS_{Within} \\N - 1 &= (k - 1) + (N - k)\end{aligned}$$

F-test statistic

$$F = \frac{MS_{Between}}{MS_{Within}} = \frac{SS_{Between}/(k - 1)}{SS_{Within}/(N - k)}$$

Hypothesis test

Reject null hypothesis if F large, i.e. if the variation *between* groups is too large compared to the variation *within* groups.

F-test statistic

The F-test statistic follows F-dist. with  $(k - 1)$  and  $(N - k)$  df.

Reject if F is large in F-distribution.

## Test of Hypothesis in Example

We want to test whether the birth weight of the children are the same irrespective of race. This corresponds to the hypothesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3$$

If the hypothesis is accepted it means that we can assume the simpler model:

$$Y_{gi} = \mu + \epsilon_{gi}, \quad i \in \{1, \dots, n_g\}, \quad g \in \{1, 2, 3\}$$

# ANOVA IN R (Analysis of Variance Table)

```
model1<-lm(bwt ~ race, data = lbw)
anova(model1)

## Analysis of Variance Table
##
## Response: bwt
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## race        2 5048361 2524181   4.949 0.008052 **
## Residuals 186 94866938  510037
```

# The F test

The test statistic F is calculated as:

$$F = \frac{\text{Mean Sq race}}{\text{Mean Sq residual}} = \frac{2524181}{510037} = 4.949$$

- Nominator (Mean Sq race) is the variation between races.
- Denominator (Mean Sq residual) is the residual variation (or within group).

We reject the null hypothesis of equal means if F is large, i.e. when large variation between groups compared to variation within groups.

# The F test

The test statistic F is calculated as:

$$F = \frac{\text{Mean Sq race}}{\text{Mean Sq residual}} = \frac{2524181}{510037} = 4.949$$

- Nominator (Mean Sq race) is the variation between races.
- Denominator (Mean Sq residual) is the residual variation (or within group).

We reject the null hypothesis of equal means if F is large, i.e. when large variation between groups compared to variation within groups.

Here  $F \sim F(2, 186)$  and the test probability is found to be

$p = P(F \geq 4.495) = 0.008 < 0.05$  so the null hypothesis is rejected

# ANOVA IN R (estimates)

```
summary(model1)

##
## Call:
## lm(formula = bwt ~ race, data = lbw)
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2719.7     140.1   19.42 <2e-16 ***
## raceother    84.3     165.0    0.51   0.610
## racewhite   383.3     157.9    2.43   0.016 *
##
## Residual standard error: 714 on 186 degrees of freedom
## Multiple R-squared:  0.0505, Adjusted R-squared:  0.0403
## F-statistic: 4.95 on 2 and 186 DF,  p-value: 0.00805
```

# ANOVA IN R (estimates)

```
confint(model1)

##           2.5 % 97.5 %
## (Intercept) 2443.38 2996.00
## raceother   -241.22  409.86
## racewhite    71.83  694.81
```

# ANOVA IN R (estimates)

```
confint(model1)
```

```
##             2.5 % 97.5 %
## (Intercept) 2443.38 2996.00
## raceother   -241.22  409.86
## racewhite    71.83  694.81
```

- Intercept is the level of the reference group. R has taken the first group alphabetically or numerically. In our example raceblack.
- racewhite this is the difference in birth weight between white and black babies.
- raceother this is the difference in birth weight between other and black babies.

# Model Assumptions

There are some assumptions connected to the model that should be checked:

# Model Assumptions

There are some assumptions connected to the model that should be checked:

- Independent observations.

# Model Assumptions

There are some assumptions connected to the model that should be checked:

- Independent observations.
- Variance homogeneity.

# Model Assumptions

There are some assumptions connected to the model that should be checked:

- Independent observations.
- Variance homogeneity.
- Normally distributed observations.

# Model Control 1: Independent Observations

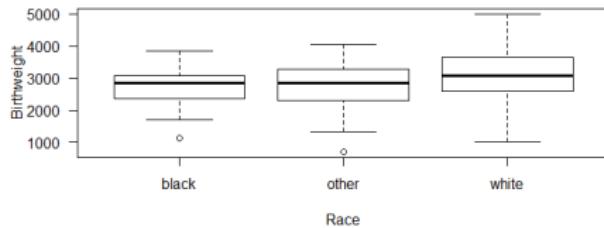
Based on knowledge about the sample.

- \* No twins, siblings,...
- \* Only one observation from each person.

## Model Control 2: Same Variance in Groups

Also called variance homogeneity (just one  $\sigma^2$ ). Can be checked:

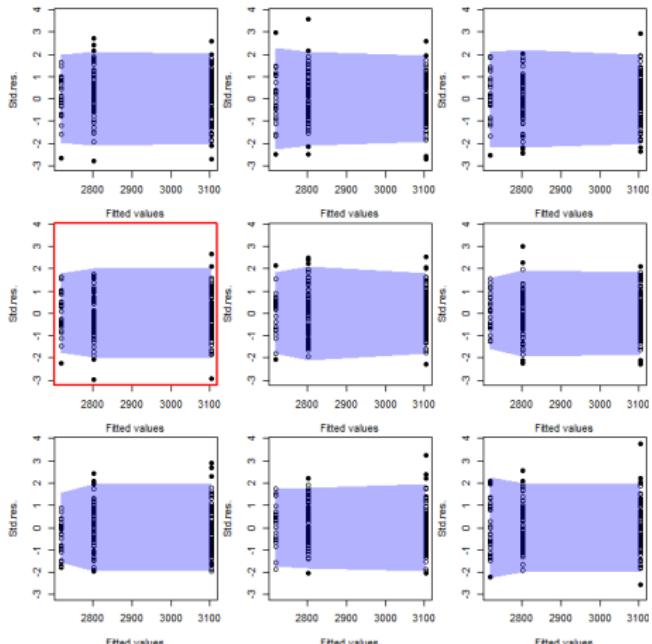
- Boxplots



- Residuals versus fitted values, look for trumpet shape.
- Test the hypothesis of equal variances (Levene's test)

# Model Control 2: Residual plot

```
library(MESS)  
wallyplot(model1)
```



## Model Control 2: Levene's Test

Test the hypothesis that

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

```
library(car)
leveneTest(model1)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df    F value   Pr(>F)
## group     2    0.4666  0.6277
##          186
```

## Model Control 2: Levene's Test

Test the hypothesis that

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2$$

```
library(car)
leveneTest(model1)

## Levene's Test for Homogeneity of Variance (center = median)
##          Df    F value   Pr(>F)
## group     2    0.4666  0.6277
##          186
```

As  $p = 0.6277 > 0.05$  then  $H_0$  is accepted.

## Model Control 3: Normal Distribution

We have assumed that the observations follow a normal distribution within each group. This can be checked:

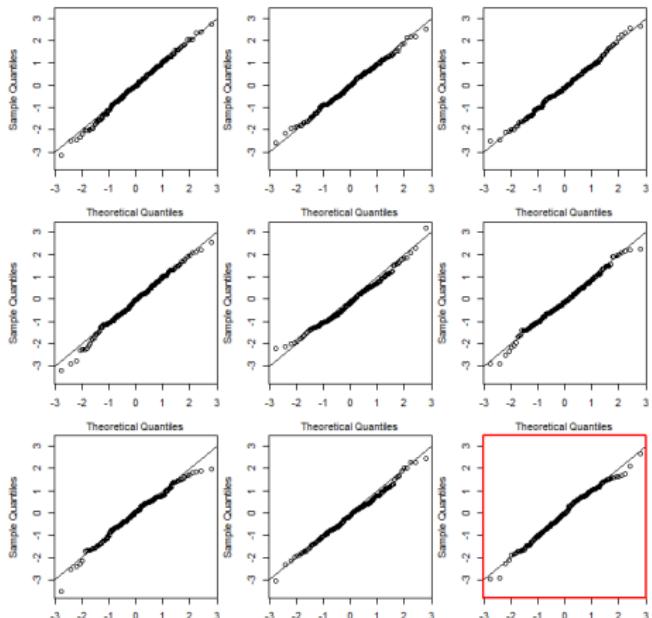
- Draw qq-plots for each group (only if many observations)
- Draw qq-plots for the residuals

$$r_{gi} = Y_{gi} - \hat{\mu}_g = Y_{gi} - \bar{Y}_g$$

- Not advisable to do formal tests for normality
  - If the data set is large then nearly always reject normality.
  - If the data set is small then nearly always accept.

# Model Control 3: QQ-plots

```
qqwrap <- function(x, y, ...) {qqnorm(y,main="",...)  
  abline(a=0, b=1)}  
wallyplot(model11, FUN=qqwrap)
```



# Conclusion for the Birthweight Data

- The model check showed that the assumptions were fulfilled.
- The F-test showed that there is a significant difference in birthweight for different races.
- But not where the differences were to be found.

# Conclusion for the Birthweight Data

- The model check showed that the assumptions were fulfilled.
- The F-test showed that there is a significant difference in birthweight for different races.
- But not where the differences were to be found.

## Pairwise Comparisons

Give problems with multiple testing. We have  $m = k(k - 1)/2$  possible tests so true level of significance (probability of rejecting a true  $H_0$ )  $1 - (1 - \alpha)^m$ . For  $k=3$  we get 0.14.

# Multiple Testing

No completely satisfactory solution to the multiple testing problem:

# Multiple Testing

No completely satisfactory solution to the multiple testing problem:

- Try to avoid the problem (focus the project).

# Multiple Testing

No completely satisfactory solution to the multiple testing problem:

- Try to avoid the problem (focus the project).
- Select a small number of comparisons in the *planning* stage.

# Multiple Testing

No completely satisfactory solution to the multiple testing problem:

- Try to avoid the problem (focus the project).
- Select a small number of comparisons in the *planning* stage.
- Make a graph of the averages  $\pm 2 \times SEM$  and judge visually.

# Multiple Testing

No completely satisfactory solution to the multiple testing problem:

- Try to avoid the problem (focus the project).
- Select a small number of comparisons in the *planning* stage.
- Make a graph of the averages  $\pm 2 \times SEM$  and judge visually.
- Modify the t-test.
  - Bonferroni test at level  $\alpha/m$ .
  - Conservative (accepts in general too often).

# Pairwise Tests in R, no correction

```
pairwise.t.test(lbw$bwt, lbw$race, p.adj = "none")  
  
##  
##  Pairwise comparisons using t tests with pooled SD  
##  
## data:  lbw$bwt and lbw$race  
##  
##      black   other  
## other  0.6100 -  
## white  0.0161  0.0093  
##  
## P value adjustment method: none
```

# Pairwise Tests in R, bonferroni

```
pairwise.t.test(lbw$bwt, lbw$race, p.adj = "bonferroni")
```

Pairwise comparisons using t tests with pooled SD

data: lbw\$bwt and lbw\$race

	black	other
other	1.000	-
white	0.048	0.028

# Exercise: One-way ANOVA

- Exercise 1 Birth Weight and Mother's Weight.

# Overview

## 1 One-way Analysis of Variance

- Descriptive
- ANOVA Model
- Data Example
- Model Control
- Multiple Testing
- Exercise

## 2 Two-way Analysis of Variance

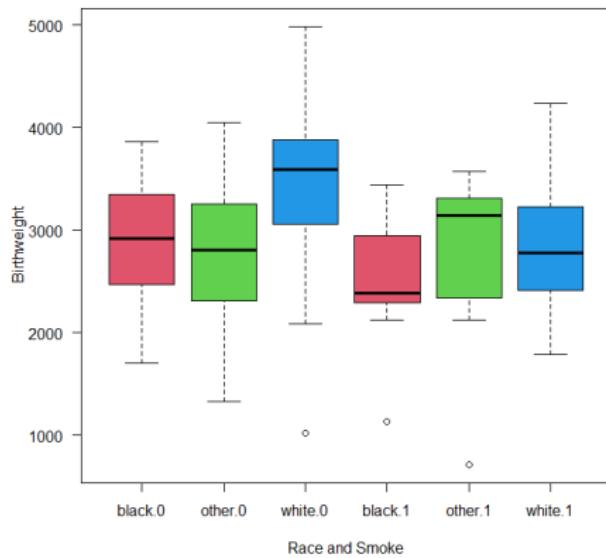
- Interaction
- Interaction estimates
- Hypotheses
- Exercise

# Two-way analysis of variance

- \* Before we had one factor (race), now we have two (race and smoke).
- \* We want to study how race and smoking during pregnancy are associated with birth weight.
- \* Our outcome is still birth weight, factor 1: race (3 groups), factor 2: smoke (2 groups).

# Example: Birth weight, Boxplot

```
boxplot(bwt ~ race*smoke, data=lbw, xlab = 'Race and Smoke',  
ylab = 'Birthweight', las = 1, col = 2:4)
```



## Two-way analysis of variance, additive model

We have an effect of race (r) and smoke (s):

$$Y_{rsi} = \mu + \alpha_r + \beta_s + \epsilon_{rsi}, \quad \epsilon_{rsi} \sim N(0, \sigma^2)$$

and the effects are **additive**.

Here  $Y_{rsi}$  is the birth weight for baby  $i$  of race  $r$  and mother's smoking status  $s$ .

$\alpha_r$  is the effect of race.

$\beta_s$  is the effect of smoking.

# Interaction

- \* Or effect modification.
- \* How do the factors work together?
- \* Maybe the effect of smoking is not the same for all white, black and other.

## Model with Interaction

$$Y_{rsi} = \mu + \alpha_r + \beta_s + \alpha\beta_{rs} + \epsilon_{rsi}, \quad \epsilon_{rsi} \sim N(0, \sigma^2)$$

Here  $Y_{rsi}$  is the birth weight for baby  $i$  of race  $r$  and mother's smoking status  $s$ .

$\alpha$  and  $\beta$  are called main effects and  $\alpha\beta$  the interaction.

The effect of smoking can depend on race and the effect of race can depend on smoking status.

## Model with Interaction

$$Y_{rsi} = \mu + \alpha_r + \beta_s + \alpha\beta_{rs} + \epsilon_{rsi}, \quad \epsilon_{rsi} \sim N(0, \sigma^2)$$

Here  $Y_{rsi}$  is the birth weight for baby  $i$  of race  $r$  and mother's smoking status  $s$ .

$\alpha$  and  $\beta$  are called main effects and  $\alpha\beta$  the interaction.

The effect of smoking can depend on race and the effect of race can depend on smoking status.

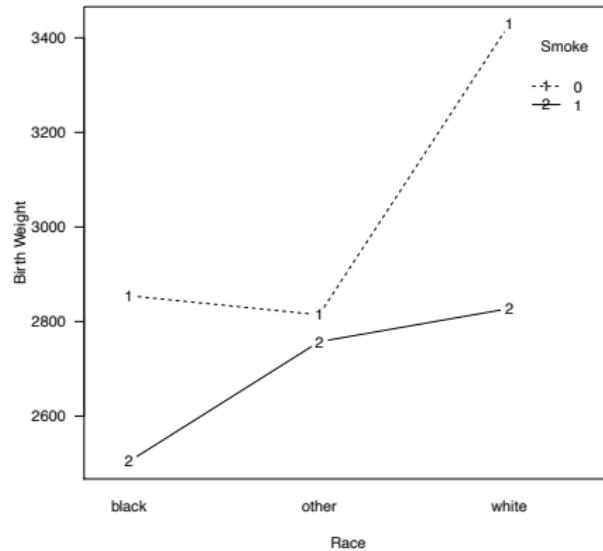
We would like to simplify the model to the additive ( $\alpha\beta_{rs}$ ):

$$Y_{rsi} = \mu + \alpha_r + \beta_s + \epsilon_{rsi}$$

Here both smoking and race have an effect on birth weight. But the effect of smoking does not depend on race and vice versa.

# Birth Weight Example with Interaction

```
interaction.plot(lbw$race, lbw$smoke, lbw$bwt, type=c("b"),
ylab = "Birth Weight", las=1, trace.label = "Smoke", xlab = "Race")
```



# Birth Weight Example with Interaction

```
lbw$smoke <- as.factor(lbw$smoke)
model2 <- lm(bwt ~ race + smoke + race:smoke, data = lbw)
summary(model2)

##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)              2854.50    170.84  16.708 <2e-16 ***
## raceother                -40.26    194.11  -0.207  0.83591
## racewhite                 574.25    199.50   2.878  0.00447 **
## smoke1                  -350.50    275.47  -1.272  0.20486
## raceother:smoke1         293.43    351.13   0.836  0.40443
## racewhite:smoke1        -250.87    309.00  -0.812  0.41792
##
## Residual standard error: 683.4 on 183 degrees of freedom
## Multiple R-squared:  0.1447, Adjusted R-squared:  0.1213
## F-statistic: 6.191 on 5 and 183 DF,  p-value: 2.505e-05
```

# Understanding the Estimates

Interpretation of the estimates from the previous slide.

- Intercept = 2854.50

The estimated birth weight for the reference group. Here race = "black" and smoke = 0.

- raceother = -40.26.

The effect of Other vs. Black for the ref. group (i.e. non-smoker).

- racewhite = 574.25

The effect of White vs. Black for the ref. group (i.e. non-smoker).

- smoke1 = -350.50

The effect of smoker vs non-smoker for the ref. group (i.e. Black).

## Understanding the Estimates (contd)

\* **raceother:smoke1** = 293.43

The extra effect of Other and smoker. The total effect of smoker vs nonsmoker for Other  $-350.50 + 293.43 = -57.07$

\* **racewhite:smoke1** = -250.87

The extra effect of White and smoker. The total effect of smoker vs nonsmoker for White  $-350.50 - 250.87 = -601.37$

## Understanding the Estimates (contd)

The estimated birth weight for combinations of race and smoke.

Smoke	Race		
	Black	Other	White
0	2854.50	2854.50 - 40.26 = 2814.24	2854.50 + 574.25 = 3428.75
1	2854.50 - 350.50 + 293.43 = 2504	2854.50 - 350.50 - 40.26 + 2757.17	2854.50 - 350.50 + 574.25 - 250.87 = 2827.38

# Hypotheses

Model without interaction:

$$M_2 : \mu_{rsi} = \mu + \alpha_r + \beta_s$$

$$H_{20} : (\alpha\beta)_{rs} = 0 \text{ for all } r, s$$

$$H_{2A} : (\alpha\beta)_{rs} \neq 0$$

No effect of race:

$$M_3 : \mu_{rsi} = \mu + \beta_s$$

$$H_{30} : \alpha_r = 0 \text{ for all } r$$

$$H_{3A} : \alpha_r \neq 0$$

No effect of smoking:

$$M_{3*} : \mu_{rsi} = \mu + \alpha_r$$

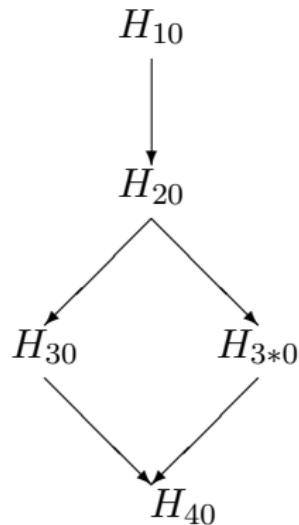
$$H_{3*0} : \beta_s = 0 \text{ for all } s$$

$$H_{3*A} : \beta_s \neq 0$$

No effect of smoking or race:

$$M_{3*} : \mu_{rsi} = \mu$$

# Hypotheses in Two-way Analysis of Variance



- If  $H_{10}$  or  $H_{20}$  is rejected then the analysis is finished.
- $H_{30}$  and  $H_{3*0}$  are on the same level.
- If reject  $H_{30}$  then test  $H_{3*0}$ .

# Hypotheses in Birth Weight Example

```
anova(model2)
```

```
## Analysis of Variance Table
##
## Response: bwt
##              Df  Sum Sq Mean Sq   F value    Pr(>F)
## race          2 5048361 2524181   5.4052 0.005229 ***
## smoke         1 7298537 7298537 15.6288 0.000110 ***
## race:smoke   2 2108643 1054321   2.2577 0.107499
## Residuals 183 85459758 466993
```

Note:

$$\text{Mean Sq} = \frac{\text{Sum Sq}}{Df}$$

## Test for No Interaction, Birth Weight Example

We will try to simplify the model by testing the hypothesis of no interaction:

$$F = \frac{\text{Mean Sq race : smoke}}{\text{Mean Sq Residual}} = \frac{1054321}{466993} = 2.26$$

Under the hypothesis of no interaction we have  $F \sim F(2, 183)$  and the test probability is found as:

$$p = P(F \geq f) = P(F \geq 2.26) = 0.107 > 0.05$$

## Test for No Interaction, Birth Weight Example

We will try to simplify the model by testing the hypothesis of no interaction:

$$F = \frac{\text{Mean Sq race : smoke}}{\text{Mean Sq Residual}} = \frac{1054321}{466993} = 2.26$$

Under the hypothesis of no interaction we have  $F \sim F(2, 183)$  and the test probability is found as:

$$p = P(F \geq f) = P(F \geq 2.26) = 0.107 > 0.05$$

The hypothesis of no interaction is accepted.

# Simplifying the model, Birth Weight Example

```
model3 <- lm(bwt ~ race + smoke, data = lbw)
anova(model3)

## Analysis of Variance Table
##
## Response: bwt
##             Df  Sum Sq Mean Sq   F value    Pr(>F)
## race        2 5048361 2524181    5.3327 0.0056021 **
## smoke       1 7298537 7298537   15.4191 0.0001214 ***
## Residuals 185 87568401 473343
```

# Simplifying the model, Birth Weight Example

```
model3 <- lm(bwt ~ race + smoke, data = lbw)
anova(model3)

## Analysis of Variance Table
##
## Response: bwt
##              Df  Sum Sq Mean Sq   F value    Pr(>F)
## race        2 5048361 2524181    5.3327 0.0056021 **
## smoke       1 7298537 7298537   15.4191 0.0001214 ***
## Residuals 185 87568401 473343
```

Test for no effect of race  $H_{30}$ :

$$F = \frac{\text{Mean Sq race}}{\text{Mean Sq Residual}} = \frac{2524181}{473343} = 5.3327, p = 0.006 < 0.05$$

# Simplifying the model, Birth Weight Example

```

model3 <- lm(bwt ~ race + smoke, data = lbw)
anova(model3)

## Analysis of Variance Table
##
## Response: bwt
##             Df  Sum Sq Mean Sq   F value    Pr(>F)
## race        2 5048361 2524181     5.3327 0.0056021 **
## smoke       1 7298537 7298537    15.4191 0.0001214 ***
## Residuals 185 87568401 473343

```

Test for no effect of race  $H_{30}$ :

$$F = \frac{\text{Mean Sq race}}{\text{Mean Sq Residual}} = \frac{2524181}{473343} = 5.3327, p = 0.006 < 0.05$$

Test for no effect of smoke  $H_{3*0}$ :

$$F = \frac{\text{Mean Sq smoke}}{\text{Mean Sq Residual}} = \frac{7298537}{473343} = 15.4191, p = 0.0001 < 0.05$$

# Final model, Birth Weight Example

```
confint(model3)
```

```
##                  2.5 %    97.5 %
## (Intercept) 2605.5688 3163.0662
## raceother   -320.3599  313.0773
## racewhite    148.5607  752.5194
## smoke1      -643.0746 -212.9761
```

## Interpretation of estimates:

For two mothers of the **same race** where one **is** a smoker and the other nonsmoker, then the birth weight of the smoker's baby will be 428g less than for the nonsmoker. The difference could be as much as 643g or as little as 213g based on the 95% CI.

# Exercise: Two-way ANOVA

- Exercise: Beer Tasting
- Exercise: Fish

# Exercise 2, Beer Tasting

---

## Story

When beer is produced the beer's shelf life can be extended by the addition of a preservative. However, one would not want to change the taste of the beer.

In this study a preservative has been added to beer at four different concentrations at 2, 4, or 6 weeks after the beer has been produced. This is a study with two factors preservative (4 groups) and age of beer (3 groups). Each combination of the two factors is repeated twice (replication). The beer is tasted by a trained panel of experts and rated on a scale from 0, 1, 2, 3 where 3 is the best beer taste. The data available are the average scores from the panel.

We would like to test whether amount of preservative and age of beer is associated with the taste.

## Data

Amount: amount of preservative added (2, 3, 4, 5).

Age: age of beer (2, 4, 6).

Rep: replication

Rating: rating of the beer.

## Exercise

1. Get the data Beer\_data.txt.
2. Illustrate the ratings with a boxplot.
3. Fit a two-way analysis of variance, include the interaction between preservative and age.
4. Test the model assumptions.
5. Reduce the model as much as possible and interpret the results.
6. What combination of preservative and age gives the best rating of the beer?

# Exercise 1, Birth Weight and Mother's Weight

---

## Story

There has been a study of 189 women to find out which factors may influence the birth weight of a baby. We would like to study whether the weight of the mother is associated with the weight of the baby.

## Data

- ID: Id of observation
- LOW: Indicator of low birth weight
- AGE: Age of mother
- LWT: Mother's weight in pounds
- RACE: Race of mother
- SMOKE: Smoking status
- PTL: History of Premature Labor
- HT: Indicator of hypertension
- UI: Presence of Uterine Irritability
- FTV: Number of Physician Visits During the First Trimester
- BWT: Birth weight in grams

## Exercise

1. Get the data lbw.txt. We would like to compare birth weight for different weight of mothers'.
2. Create a new variable Weight\_group grouping maternal weight four equally sized groups.
3. Illustrate the birth weight in the four weight groups with a figure.
4. Is the mean birth weight the same in the four maternal weight groups?
5. Remember to check the model assumptions.
6. Use pairwise tests adjusting for multiplicity to see where there might be differences.

## Hint

Find the quartiles of mother's weight using `summary()` (just read them from the screen)

```
summary( BWTdata$LWT)
```

Weight\_group should be a new column in the data frame with value 1 for the lightest mothers, 2 for the next group, 3 for the third group and 4 for the heaviest mothers.

```
BWTdata$Weight_grp[BWTdata$LWT < p25]<-1
```

```
BWTdata$Weight_grp[BWTdata$LWT >= p25 & BWTdata$LWT < p50]<-2
```

```
BWTdata$Weight_grp[BWTdata$LWT >= p50 & BWTdata$LWT < p75]<-3
```

```
BWTdata$Weight_grp[BWTdata$LWT >= p75]<-4
```

## Case: Fish

### Story

To study the influence of ocean grazers on regeneration rates of seaweed in the intertidal zone, a researcher scraped rock plots free of seaweed and observed the degree of regeneration when certain types of seaweed-grazing animals were denied access. The grazers were limpets (L), small fishes (f) and large fishes (F). A plot was taken to be a square rock surface, 100 cm on each side. Each plot received one of six treatments, named here by which grazers were allowed access.

symbol	description
Lff	All three grazers were allowed access
ff	Limpets were excluded by surrounding the plot with caustic paint
Lf	Large fish were excluded by covering the plot with a coarse net
f	Limpets and large fish were excluded
L	Small and large fish were excluded by covering the plot with a fine net
C	Control: limpets, small fish and large fish were all excluded

Because the intertidal zone is a highly variable environment, the researcher applied the treatments in eight blocks of twelve plots each. Within each block she randomly assigned treatments to plots so that each treatment was applied to two plots. The blocks covered a wide range of tidal conditions

symbol	description
Block 1	just below high tide level, exposed to heavy surf
Block 2	just below high tide level, protected from the surf
Block 3	midtide, exposed
Block 4	midtide, protected
Block 5	just above low tide level, exposed
Block 6	just above low tide level, protected
Block 7	on near-vertical rock wall, midtide level, exposed
Block 8	on near-vertical rock wall, midtide level, protected

### Variables

variable name	description
cover	percentage of regenerated seaweed
block	blocks for different tidal situations
treat	treatment, i.e. what fishes are excluded

### Exercise

1. Make appropriate plots to investigate whether treatments or blocks have any influence on the percentage of regenerated seaweed
2. Fit a two-way analysis of variance model.
3. Check the underlying assumptions of your previous model. If the assumptions were not fulfilled then improve your model with a transformation.
4. Test the model using a 5% significance level
5. Is there evidence of any interaction effects?
6. Try to combine some of the treatment group and test whether your combinations are allowed
7. Split the block in two factors zone (4 levels) and exposed yes/no. Use zone\*exposure instead of block. Is the interaction statistically significant?

Hint:  $y=cover/100$  is a number between 0 and 1, so perhaps a logit transformation might work

$$\text{Logit}(y) = \log(y/(1-y))$$

1

# Linear Models

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 8th, 2025

## Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, statistical report writing,  
logistic Regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Overview

1 The Linear Model  
• ANCOVA

2 Vital Capacity and Cadmium

3 Exercises

# Terminology

For continuous outcomes (e.g. birth weight)

- **Regression:** The covariates are also continuous.
  - Simple (linear) regression: Just one covariate.
  - Multiple (linear) regression: Two or more covariates.
- **Variance analysis:** Covariates are categorical (grouped, factors).
  - One-way analysis of variance: Just one covariate (factor).
  - Two-way analysis of variance: Two covariates (factors).
- **General linear model:** Both types of covariates in the same model.
  - Analysis of covariance: Exactly one continuous and one categorical covariate.

# The General Linear Model (GLM)

$Y_i$  is the outcome for person  $i$  and  $(X_{i1}, \dots, X_{ik})$  are explanatory covariates e.g. age of person  $i$ , or a "dummy" variable:

$$X_{ij} = \begin{cases} 1 & \text{if person } i \text{ is from group } j \\ 0 & \text{if person } i \text{ is not from group } j \end{cases}$$

E.g.  $X_{i1} = 1$  if person  $i$  a boy and  $X_{i1} = 0$  if person  $i$  a girl.  
Model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + \varepsilon_i$$

Where  $\varepsilon_i \sim N(0, \sigma^2)$  and independent.

The predicted values are called  $\hat{Y}_i$ .

# Model Reduction in GLM

- \* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

# Model Reduction in GLM

- \* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

Model Sum of Squares  $SS_{model} = \sum(\hat{Y}_i - \bar{Y})^2$

- Explained variation
- How much do the predicted values vary?
- Large is good

# Model Reduction in GLM

- \* In a general linear model we can split the variation.

$$SS_{total} = SS_{model} + SS_{residual}$$

Model Sum of Squares  $SS_{model} = \sum(\hat{Y}_i - \bar{Y})^2$

- Explained variation
- How much do the predicted values vary?
- Large is good

Residual Sum of Squares  $SS_{residual} = \sum(Y_i - \hat{Y}_i)^2$

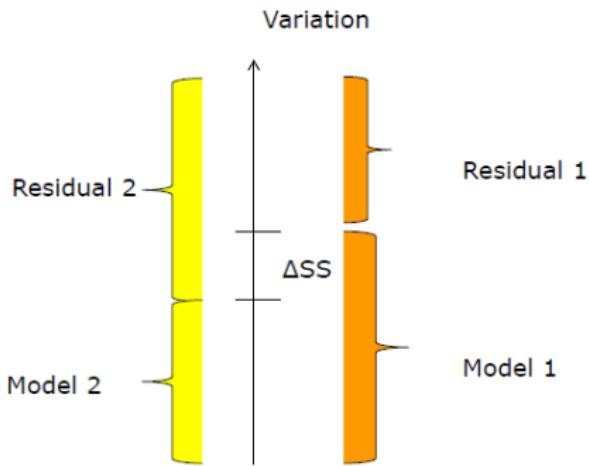
- Variation not explained by model.
- How large are the differences between observed and predicted?
- Small is good.

# Model Reduction - F test

- \* We want to compare two models.  
The original (no. 1) and a simplified (the hypothesis, no. 2).
- \* Is it ok to use the simplified model? Is it good enough?
- \* Note the models must be **nested**, i.e. you get one from the other by setting parameters to zero ("remove effects").
- \* We look at changes in model sum of squares:  
How much less is explained by the simpler model?

$$\Delta SS = SS_{model1} - SS_{model2}$$

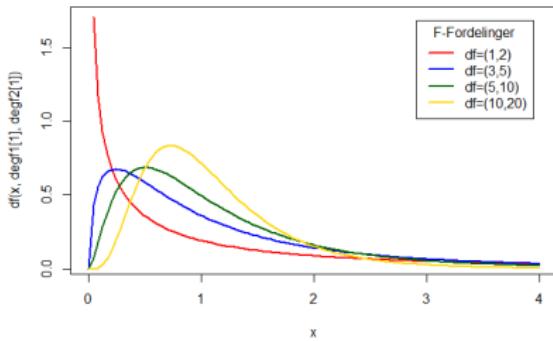
# Model Reduction - contd.



- More parameters can explain (a little) more variation  $\Delta SS > 0$ .
- **How much more?**
- How large  $\Delta SS$  before test significant?

# F-test

- The size of  $\Delta SS$  is seen together with the reduction in parameters  $\Delta DF = Df_1 - Df_2$ .
- $\Delta SS$  is compared to the residual variation from the larger model.



$$F = \frac{\Delta SS / \Delta Df}{SS_{residual} / Df_1} \sim F(\Delta Df, Df_1)$$

# The $R^2$ Statistic

- The  $R^2$  statistic is given as

$$R^2 = \frac{SS_{model}}{SS_{total}}$$

- Often referred to as the *coefficient of determination*.
- Measures how much of the variation that the model explains, large is good. Is found in the summary output from lm.
- A high  $R^2$  gives a model that explains a lot; but says absolutely **nothing** about whether it is a *sensible* explanation.
- Whether the explanations are *sensible* in modelling terms, is decided from the model control.

# The Adjusted $R^2$ Statistic

- The  $R^2$  automatically increases when you add explanatory variables to the model. This is not always sensible.
- To correct for this phenomenon, one often uses the *adjusted*  $R^2$ ,  $\bar{R}^2$  instead:

$$MS_{model} = SS_{model}/df_{model}; MS_{res} = SS_{res}/df_{res};$$

$$MS_{total} = SS_{model}/df_{total}$$

$$\bar{R}^2 = 1 - \frac{MS_{res}}{MS_{total}}$$

- Also found in the summary output of `lm`.

# Analysis of Covariance - The Simplest GLM

- \* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
- \* What could be the aim of such an analysis?

# Analysis of Covariance - The Simplest GLM

- \* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
- \* What could be the aim of such an analysis?
  - To study the two covariates.

# Analysis of Covariance - The Simplest GLM

- \* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
- \* What could be the aim of such an analysis?
  - To study the two covariates.
  - Remove bias, e.g. correct for height differences when comparing lung capacity of smokers and non-smokers.

# Analysis of Covariance - The Simplest GLM

- \* A (historical) term for a model with exactly one categorical covariate (group, factor) and exactly one continuous covariate.
- \* What could be the aim of such an analysis?
  - To study the two covariates.
  - Remove bias, e.g. correct for height differences when comparing lung capacity of smokers and non-smokers.
  - Increase the power in a randomized clinical trial by reducing the unexplained part of the variance, e.g. by including age as a covariate.

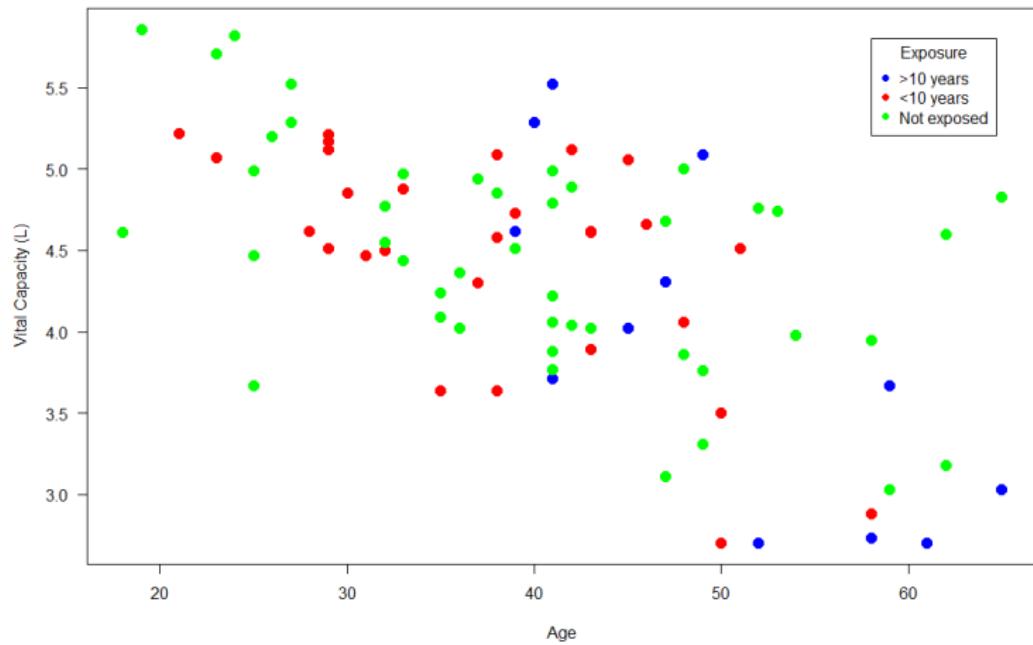
## Example: Vital Capacity and Cadmium

- We have data from a study of the effect of exposure to cadmium on the vital capacity. (From P. Armitage & G. Berry: Statistical methods in medical research. 2nd ed. Blackwell 1987)
- Vital capacity is the maximum amount of air a person can expel from the lungs after a maximum inhalation.
- We have measurements of vital capacity (L), age and exposure to cadmium ( $> 10$  years,  $< 10$  years, not exposed).
- Start by plotting the data!

# Rcode for plots, scatter plot

```
CADdata <- read.csv("cadmium.txt", sep="")  
CADdata$group <- as.factor(CADdata$group)  
  
#PLOT THE DATA WITH DIFFERENT COLOURS IN EXPOSURE GROUPS  
plot(CADdata$age, CADdata$vitcap,  
      col = c("blue", "red", "green")[CADdata$group],  
      xlab = "Age", ylab = "Vital Capacity (L)",  
      las = 1, cex = 1.5, pch = c(16,16,16))  
legend(55,5.8, c(">10 years", "<10 years", "Not exposed"),  
       col = c("blue", "red", "green"),  
       pch = c(16,16,16), title = "Exposure")
```

## Scatter Plot



## Rcode for plots, boxplot

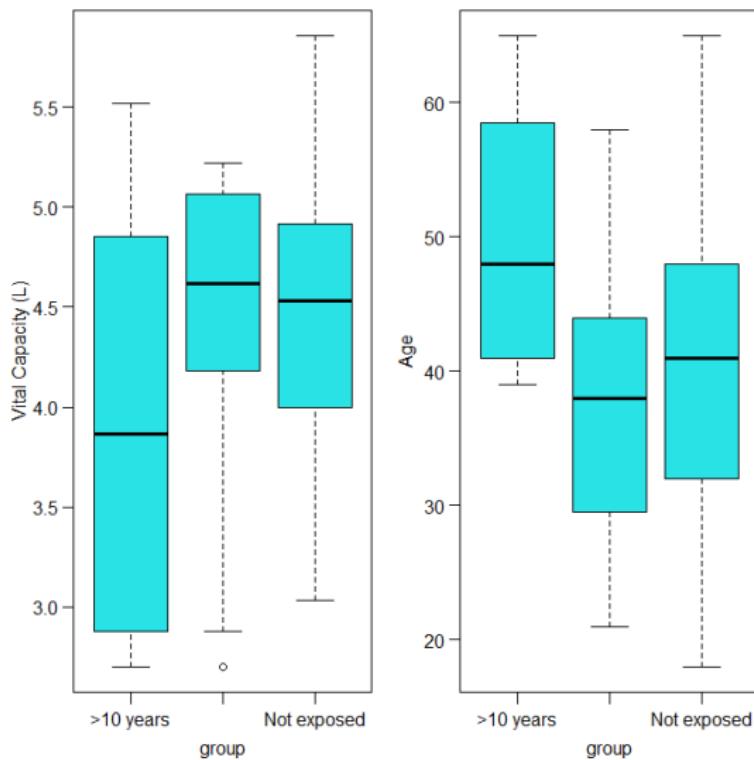
```
#TWO PLOTS NEXT TO EACH OTHER
par(mfrow = c(1,2), mgp = c(2,0.7,0), mar = c(3,3,1,1))

boxplot(vitcap ~ group, data = CADdata, ylab =
    'Vital Capacity (L)', las = 1, xaxt = "n", col = 5)
axis(1, at = c(1,2,3),
    labels = c(">10 years", "<10 years", "Not exposed"))

boxplot(age ~ group, data = CADdata , ylab = 'Age',
    las = 1, xaxt = "n", col = 5)
axis(1, at = c(1,2,3),
    labels = c(">10 years", "<10 years", "Not exposed"))

#BACK TO ONE PLOT
par(mfrow = c(1,1))
```

## Boxplots



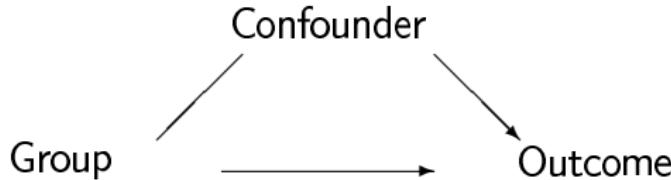
# Comparing Groups

Comparing groups that are not quite comparable (e.g. cadmium exposure).

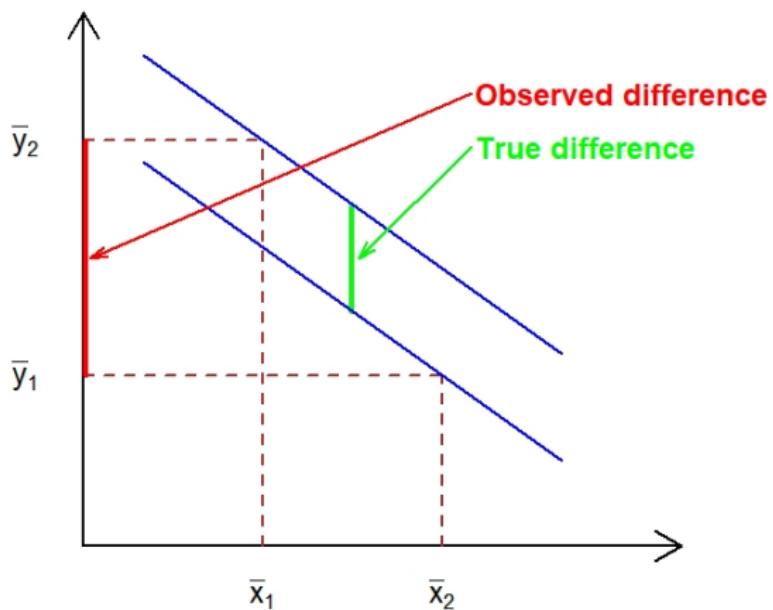
**Confounder:** A variable that

- Has an effect on the outcome.
- Is associated to group (different ages in groups)

This can cause **bias**.

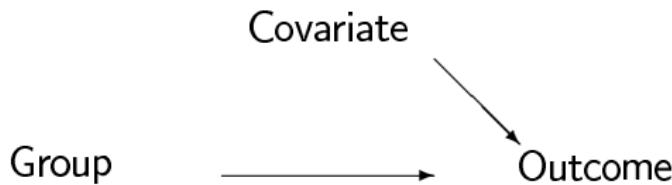


# Illustration of Confounding and ANCOVA



# Adjustment

Even if the distribution of the covariate is the same in the groups, then it can reduce the variation.



- This gives greater power.
- But remember that we are answering a different scientific question (which one?).

# Trying to Avoid Bias

**Matching** Choose individuals so they are similar for important disturbing covariates. Remember to include the matching variables as covariates. Otherwise one can create bias due to unmeasured confounding. Do not interpret the effect

**Randomization** Draw lots between intervention groups.

**Adjust** Include the skew covariate in the model.

# Overview

1 The Linear Model  
• ANCOVA

2 Vital Capacity and Cadmium

3 Exercises

# Vital Capacity and Cadmium

The model for vital capacity

$$Y_i = \beta_0 + \beta_{>10} X_{i,>10} + \beta_{<10} X_{i,<10} + \beta_{age} X_{i,age} + \epsilon_i$$

Here

$X_{i,>10} = 1$  if person  $i$  is exposed  $> 10$  years 0 otherwise.

$X_{i,<10} = 1$  if person  $i$  is exposed  $< 10$  years 0 otherwise.

$X_{i,age}$  = age of person  $i$ .

## Exercise:

- Work in pairs. Online: Work with yourself ☺.
- Draw a sketch of how you envision the above model on a piece of paper.

# Vital Capacity and Cadmium

We have a model with three parallel lines:

 $\beta_{age}$ 

Common slope.

 $\beta_0$ 

Intercept for not exposed

 $\beta_0 + \beta_{<10}$ 

Intercept for exposed < 10 years

 $\beta_0 + \beta_{>10}$ 

Intercept for exposed > 10 years

# Model Check

- Normally distributed residuals ( $y - \hat{y}$ ) (qq-plot).
- Independent observations.
- Variance homogeneity (residual plot).
- Linear effects (residual plots).

## Assumption about Independence

A simple assessment: “Random sample”, “Each individual only sampled once”

# Model Check, using built in plot

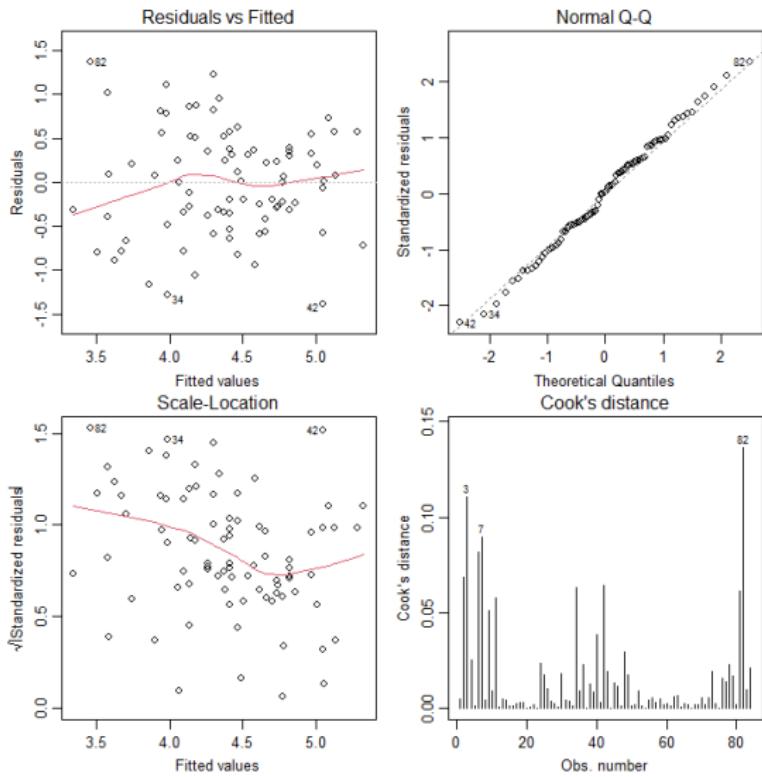
```
#EXPO WHERE NOT EXPOSED 1, <10 IS EXPO==2, >10 is EXPO==3
CADdata$expo[CADdata$group==3] <- 1
CADdata$expo[CADdata$group==2] <- 2
CADdata$expo[CADdata$group==1] <- 3

#DECLARE EXPO AS A FACTOR
CADdata$expo<-as.factor(CADdata$expo)

#Initial model
Model1<-lm(vitcap ~ age + expo, data = CADdata)

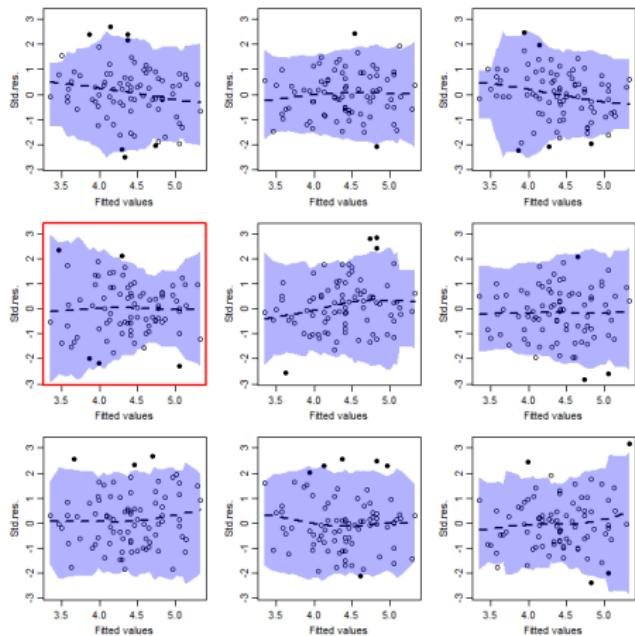
#Model check
par(mfrow = c(2,2), mgp = c(2,0.7,0), mar = c(3,3,1.5,1))
plot(Model1, which = 1:4)
par(mfrow = c(1,1))
```

# Model Check, using built in plot



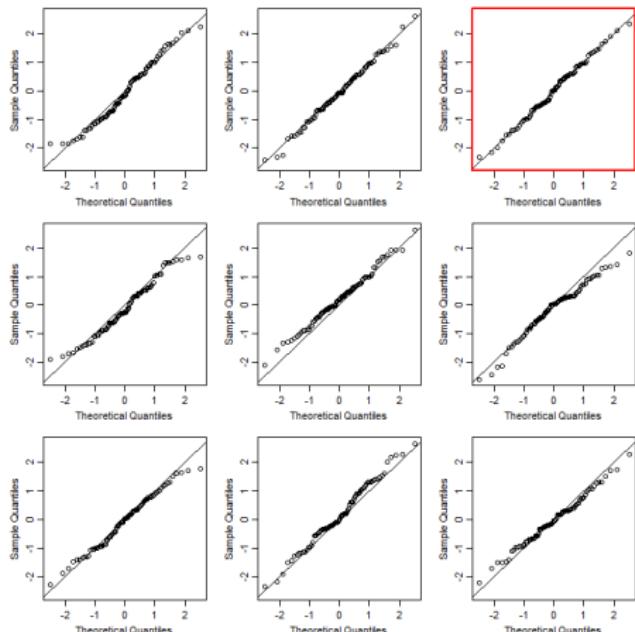
# Extra plots if in doubt: Plot to check variance homogeneity

```
library(MESS)  
wallyplot(Model1)
```



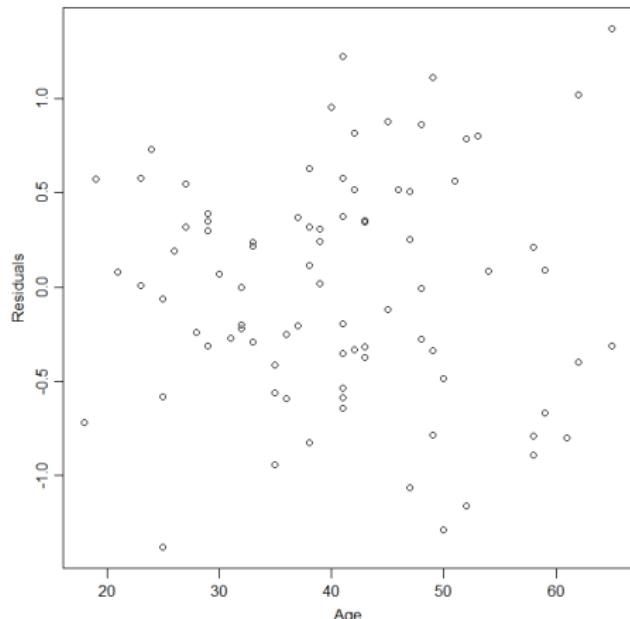
# Plot to check normal residuals

```
qqwrap <- function(x, y, ...) {qqnorm(y,main="",...); abline(a=0, b=1)}  
wallyplot(Model1, FUN=qqwrap)
```



# Plot to check linearity of age

```
plot(CADdata$age, Model1$residuals, xlab = 'Age',  
ylab = 'Residuals')
```

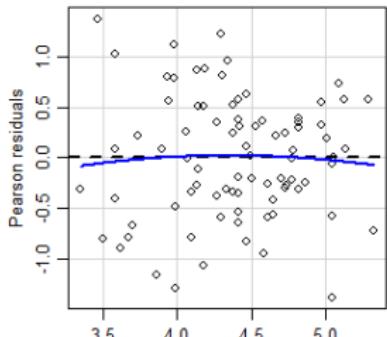
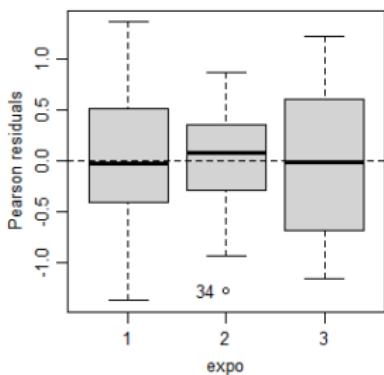
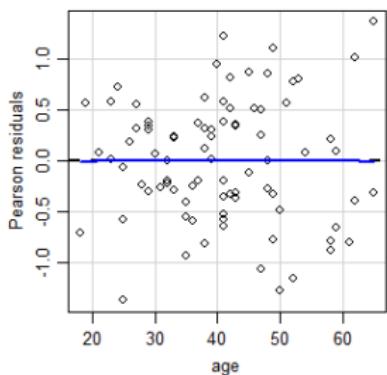


# Plot to check linearity of age using library(car)

```
library(car)
residualPlots(Model1)

##           Test stat  Pr(>|t|)
## age          -0.0535  0.9575
## expo
## Tukey test   -0.3918  0.6952
```

# Plot to check linearity of age using library(car)



# The Model check went well

- Normally distributed residuals ( $y - \hat{y}$ ) (qq-plot) (straight line).
- Independent observations. (Cannot check, have to assume).
- Variance homogeneity (residual plot, no trumpet).
- Linear effects (residual plots, looks random).
- We could also look for influential observations looking at Cook's distance.

# Estimates

```
Model1 <- lm(vitcap ~ age + expo, data = CADdata)
summary(Model1)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.044917  0.268025 22.554  < 2e-16 ***
## age         -0.039775  0.006322 -6.291  1.57e-08 ***
## expo2        -0.070198  0.148669 -0.472    0.638
## expo3        -0.116935  0.209236 -0.559    0.578
##
## Residual standard error: 0.6127 on 80 degrees of freedom
## Multiple R-squared:  0.3696, Adjusted R-squared:  0.3459
## F-statistic: 15.63 on 3 and 80 DF,  p-value: 4.323e-08
```

# Table of results in R

```

confint(Model1)

##                   2.5 %      97.5 %
## (Intercept) 5.51153040 6.57830307
## age         -0.05235723 -0.02719313
## expo2        -0.36605755 0.22566252
## expo3        -0.53332814 0.29945819

# Nice table
tab <- cbind(coef(summary(Model1))[, 1:2], "Lower" = confint(Model1)[, 1],
             "Upper" = confint(Model1)[, 2])

# Nice table with p-values
data.frame(round(tab, 2),
            "p-value" = format.pval(coef(summary(Model1))[, 4], digits = 3, eps = 1e-3))

##           Estimate Std. Error Lower Upper p.value
## (Intercept) 6.04      0.27  5.51  6.58 <0.001
## age        -0.04      0.01 -0.05 -0.03 <0.001
## expo2       -0.07      0.15 -0.37  0.23  0.638
## expo3       -0.12      0.21 -0.53  0.30  0.578

```

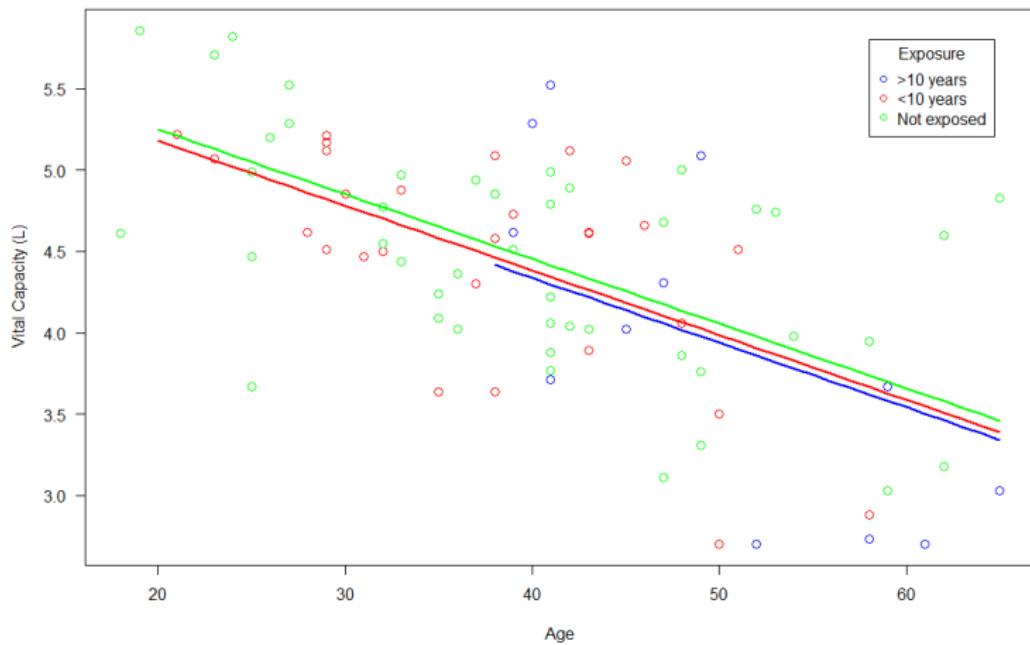
## Estimates from the output

From the R output we got:

- $\hat{\beta}_{age}$  -0.04 (-0.05; -0.03) (Common slope)
- $\hat{\beta}_0$  6.04 (5.51; 6.58) (Intercept for not exposed)
- $\hat{\beta}_{<10}$  -0.07 (-0.37; 0.23) (Extra intercept for exposed < 10 years)
- $\hat{\beta}_{>10}$  -0.12 (-0.53; 0.30) (Extra intercept for exposed > 10 years)

And the variance  $\sigma^2 = 0.613^2 = 0.376$ .

## Fitted Lines



# Interaction

- The vital capacity decreases with -0.04 L per year.
- Is it reasonable that the vital capacity decreases with the same rate in all three exposure groups?
- Allow different slopes in the three groups → Include an interaction between age and group.

# Estimates, from model with interaction

```
Model2 <- lm(vitcap ~ age + expo + age:expo, data = CADdata)
summary(Model2)
```

## Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
## (Intercept)	5.680291	0.313426	18.123	< 2e-16 ***
## age	-0.030613	0.007547	-4.066	0.000117 ***
## expo2	0.549740	0.575884	0.955	0.342728
## expo3	2.503148	1.041842	2.403	0.018655 *
## age:expo2	-0.015919	0.014547	-1.094	0.277170
## age:expo3	-0.054498	0.021070	-2.587	0.011554 *
##				
## Residual standard error:	0.5942	on 78 degrees of freedom		
## Multiple R-squared:	0.422	Adjusted R-squared:	0.385	
## F-statistic:	11.39	on 5 and 78 DF,	p-value: 2.871e-08	

# Test Interaction

```
drop1(Model2, test = "F")  
  
## Single term deletions  
##  
## Model:  
## vitcap ~ age + expo + age:expo  
##          Df   Sum of Sq   RSS      AIC    F value    Pr(>F)  
## <none>           27.535 -81.689  
## age:expo     2   2.4995 30.035 -78.391      3.5402  0.03376 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Test Interaction

```
drop1(Model2, test = "F")  
  
## Single term deletions  
##  
## Model:  
## vitcap ~ age + expo + age:expo  
##          Df   Sum of Sq   RSS      AIC    F value    Pr(>F)  
## <none>           27.535 -81.689  
## age:expo     2   2.4995 30.035 -78.391      3.5402  0.03376 *  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So the interaction is statistically significant  $0.03376 < 0.05$  and we need this more complex model.

# The same model different parametrization

- We want to be able to get the three intercepts and slopes directly from the output.
- Notice the '0', says not to have common intercept.
- This parametrization not for testing the interaction but for understanding.

```
Model2B<-lm(vitcap ~ 0 + expo + age:expo, data = CADdata)
```

# The same model different parametrization

```
summary(Model2B)
```

```
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|) 
## expo1      5.680291  0.313426 18.123   < 2e-16 ***
## expo2      6.230031  0.483122 12.895   < 2e-16 ***
## expo3      8.183438  0.993579  8.2436  3.28e-12 ***
## expo1:age -0.030613  0.007547 -4.056   0.000117 ***
## expo2:age -0.046532  0.012436 -3.742   0.000347 ***
## expo3:age -0.085111  0.019672 -4.327   4.44e-05 ***
## 
## Residual standard error: 0.5942 on 78 degrees of freedom
## Multiple R-squared:  0.9835, Adjusted R-squared:  0.9822 
## F-statistic: 774.5 on 6 and 78 DF,  p-value: <2.2e-16
```

```
#confint(Model2B)
```

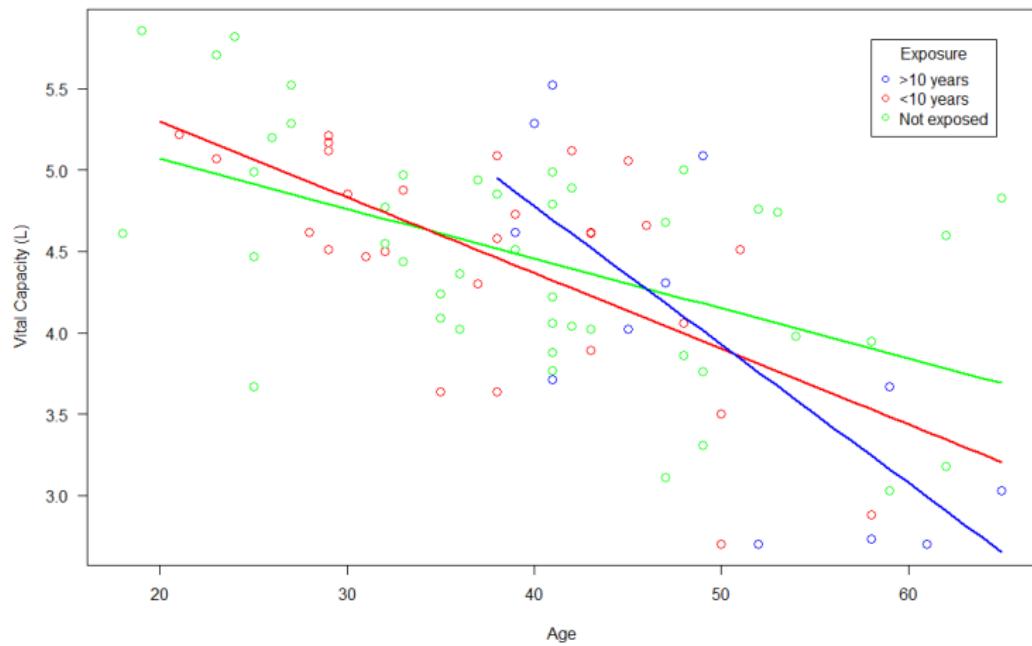
# Estimates from the output with interaction

From the R output we got:

Intercept for not exposed	5.68 (5.06; 6.30)
Slope for not exposed	-0.03 (-0.05; -0.02)
Intercept for exposed < 10 years	6.23 (5.27; 7.19)
Slope for exposed < 10 years	-0.05 (-0.07; -0.02)
Intercept for exposed > 10 years	8.18 (6.21; 10.16)
Slope for exposed > 10 years	-0.09 (-0.12; -0.05)

And the variance  $\sigma^2 = 0.594^2 = 0.353$ .

## Fitted Lines Interaction



# Overview

1 The Linear Model  
• ANCOVA

2 Vital Capacity and Cadmium

3 Exercises

# Exercises

- Exercise 4: Prostate Cancer
- Exercise 5: Birth weight and gestation week

# Exercise 5, ANCOVA

---

## Story

There has been a study of 490 women (an extended version of the previous data) to find out which factors may influence the birth weight of a baby. We would like to study whether the gender and hypertension in the mother are associated with the weight of the baby.

## Data

- id: Id of observation
- bweight: Birth weight in grams
- lowbw: Indicator of low birth weight
- gestwks: Gestation weeks
- preterm: Indicator of preterm baby
- matage: Mother's age
- hyp: Indicator of hypertension
- sex: Sex of baby (1 = male, 2 = female)
- sexalph: Sex of baby ("male", "female")

The data are tab delimited.

## Exercise

1. Get the data births.txt. We would like to compare birth weight for different sex of the child and gestation weeks.
2. Calculate a new variable gest24=gestwks-24
3. Make some plots of the data
4. Fit a linear model to birth weight with gest24 and sex as explanatory variables, with separate slopes for boys and girls.
5. Reduce the model as much as possible and interpret the results.
6. What is the interpretation of the intercept estimate?
7. Plot the estimated regression lines from the final model.

# Exercise 4, Prostate Cancer

---

## Story

The data come from a study conducted by Stamey et al. (1989) where associations between the level of a prostate-specific antigen and a number of potentially explanatory covariates were studied. The aim of the study was to see whether any of the covariates could predict the level of prostate-specific antigen.

## Data

### Outcome

lpsa: level of prostate-specific antigen

### Covariates

lcavol: log cancer volume

lweight: log prostate weight

age: age

lbph: log of the amount of benign prostatic hyperplasia

svi: seminal vesicle invasion (0 = no, 1 = yes)

lcp: log of capsular penetration

gleason: Gleason score (6,7,8,9)

pgg45: percent of Gleason scores

## Exercise

- Read in the data prostate.txt and make descriptive plots of plot the lpsa, lcavol and svi.
- Fit an ANCOVA with lpsa as outcome and lcavol as continuous and svi as categorical covariates, with different slopes for svi=1 and svi=0
- Do a model check.
- Reduce the initial model until there only are significant covariates left. Estimate the parameters with 95% confidence intervals.
- Plot the estimated regression lines from the final model.
- If time look at the rest of the covariates and fit a general linear model, which covariates are statistically significant for the level of prostate-specific antigen?

# Writing Statistical Reports

January 9, 2025

Anders Stockmarr

Section for Statistics and Data Analysis, DTU.

anst@dtu.dk

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$

$\Theta^{\sqrt{17}} + \Omega \int_a^b \delta e^{i\pi} =$   
 $\infty = \{2.71828182845904523536028747135266249775724709369995957497474... \}$   
 $\Sigma \gg \chi^2$

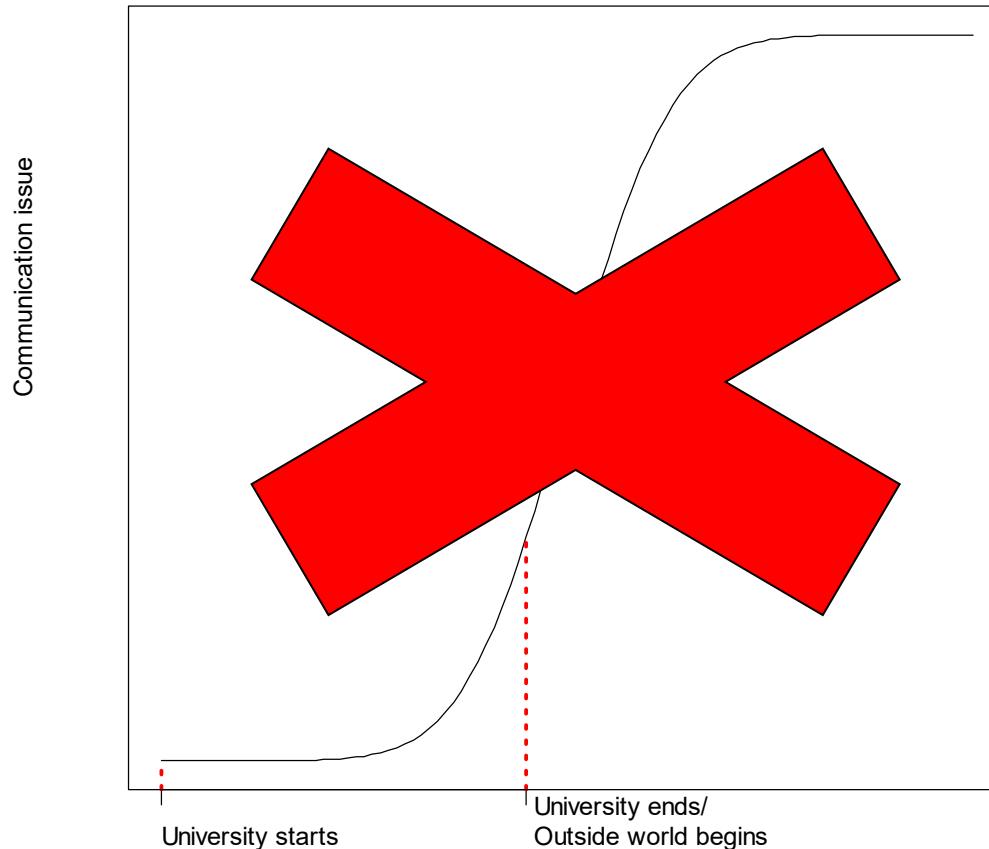
DTU Compute

Department of Applied Mathematics and Computer Science

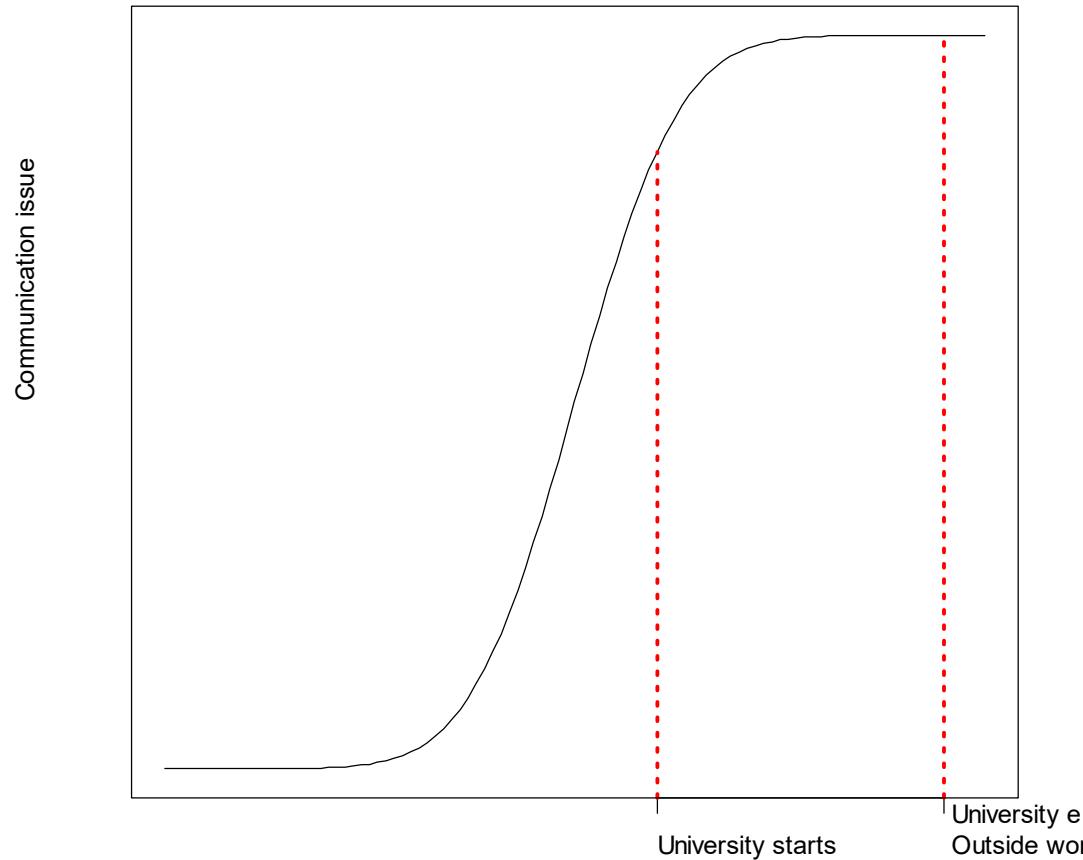
# Why write statistical reports?

- At University: Perhaps because teachers says so.
- Other reasons could be:
  - To document what you have done in a study?
  - Yes, always, but that is not the whole issue when it comes to **statistical** reports.
  - Statistics is a science;
  - applied statistics is the **application of this science within another discipline**;
  - the reader of the report is therefore **not necessarily equipped** to put the results in context;
  - there is therefore, also, an issue of **communication**.

# Statistical reports at university and elsewhere



# Statistical reports at university and elsewhere



# Statistical Reports

- Issues:
  - Document your methods;
    - **matter-of-fact, listing, specifying.**
  - Tell your story;
    - **Why, what, how.**
  - Communicate your story;
    - **Context, examples, discussion.**
- Portrait **the Client** – the expected reader/end-user (this person may or may not exist IRL).

# Statistical Reports – the Client

- Portrait **the Client** – the expected reader/end-user.
- **The Client** could be an external person;
- Or **the Client** could be **yourself**, in 12 months time when you have forgotten most about the analysis.

# Statistical Reports – the Client

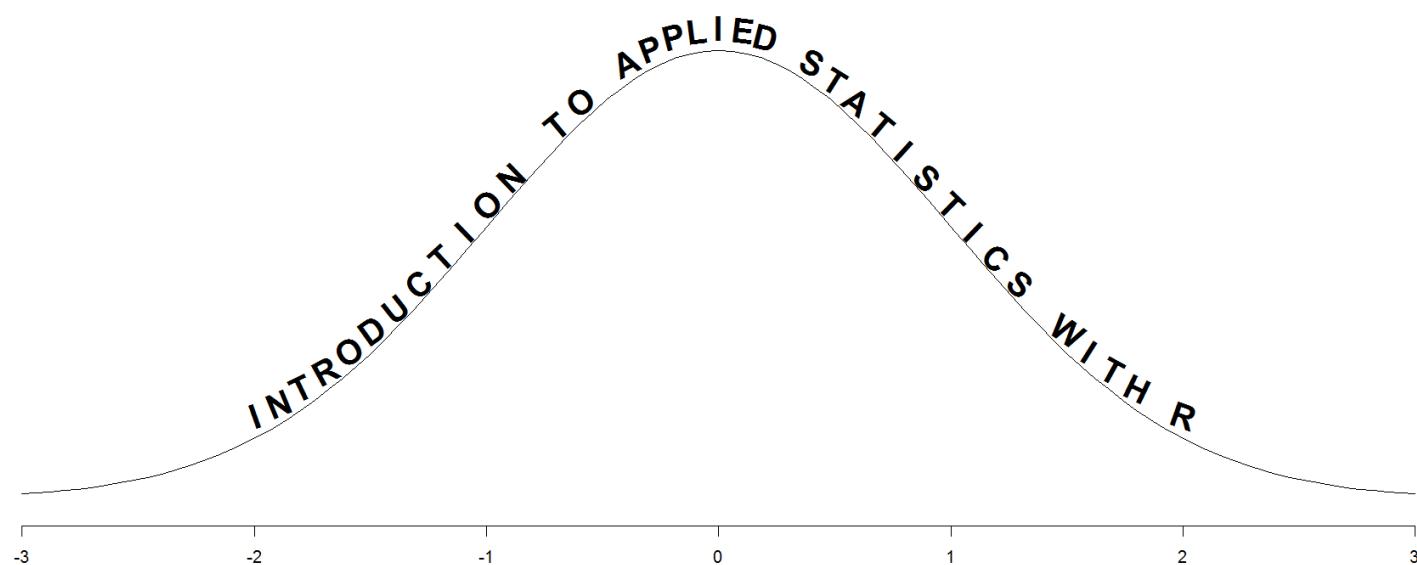
- Portrait **the Client** – the expected reader/end-user.
  - What does **the client** already know? (**basic/advanced science on the subject, statistical methods, project circumstances**)
  - What does **the client** not know? (**basic/advanced science on the subject, statistical methods, project circumstances**)
  - What is the interest of **the Client**? (**research question, p-values, effect parameters, issues with data handling**)
  - What is NOT the interest of **the Client**? (**R code, issues with data handling, intermediate analyses**)
- Adapt the contents and structure (not the results though ☺) to fit the knowledge and interests of **the Client**.

# Report Structure – Contents of the Good Report

- Front page.
- A summary (less than one page).
- A table of contents.
- Introduction.
- Description of data.
- Statistical analyses, results and discussion.
- Conclusion.
- Appendices.

# Front page

- Your **name**.
- Your **Affiliation** (Here: student number)
- The **title** of your report
- Perhaps a nice **picture** to display, to give an idea of the contents:



# Summary

- Should be short.
- Summary of questions posed, the main results and conclusions.
- **The Client will read this section:** Include what you want **the Client** to know about your study; and what **the Client** should look for details about in the report.
- Other readers than the Client may never read more than this! Be sure to include the main findings here.

# Introduction

- Should set the context of the report, give a background description and a formulation of the main reason for the work.
- Should contain one or more specific research question(s) that the work is (was) supposed to answer.
- Pay attention to the **Client** portrait here. What level of information is necessary to set the context properly?

# Description of Data

- Describe the data!
- What are the different data types available? which are outcomes, which are explanatory variables etc.
- Possible means of describing data:
  - verbal means;
  - summary tables; means, medians, quantiles etc.
  - graphical means; scatter plots, histograms etc.

# Statistical Analyses, Results and Discussion

- Three different sections!

## Statistical analyses:

- For each research question, specify the statistical methods used, and why you use exactly these.
- Address if the methods are proper. Are there any assumptions behind the methods (normality, independence etc.) that you needed to verify, and how did that go?

## Results:

- Present results in a matter-of-fact manner.

To communicate results, graphs are nice, but tables are needed as well. Only in special situations should tables not be supplied.

- Example: For investments on the stock market, you need a value that is not read off from a graph.

# Statistical analyses, Results and Discussion

## Discussion:

- **Interpret** the results in the context, in order to communicate the results to the Client.
- Example: Result: " $\beta < 0$ ,  $p=0.0001$ ". Interpretation: "The coefficient is significantly negative, so people who were given treatment A have a higher one-year expected survival rate".
- **Discuss** the results in the context, and relate it to the interests of the Client:
  - Example: "The sample was chosen representatively, so the results applies at the population level. However, the data were self-reported so the estimate may be biased downwards".
  - In particular, determine the level of reliability of the results. Did the data pass the model control? And similar issues.

# Conclusion

- Summarize the analyses.
- Make sure that you address the research question(s) formulated in the introduction.
- Make the conclusions with the weight that your analysis indicate that you can (ie. for example, reservations for things like self-reported data).
- The Conclusion is **the second section that you can count on the Client reading**. Therefore, the interests of **the Client** (ie. questions /context of interest) should be addressed.

# Appendices

- This is for issues of documentation.
- Necessary information, which is not of immediate interest.
- Examples: R code, unproblematic model control charts, graphs which are not of major interest.

# Specifics – Level of Information

- Pay attention to that you **supply enough information** so that the demands for documentation are satisfied.
- **Replication principle:** With access to data, **the Client** should be able to replicate your analysis, based on facts from the report.
- Pay attention to that you do not **supply redundant information**: The most common communication error is oversized reports on relatively simple problems. **Oversized reports** will simply **not be read in full** by **the Client**.
- For each piece of information, consider if it is necessary for either:
  - Documentation/the replication principle;
  - Results;
  - Communication.

If a piece of infomation is not neccessary, **LEAVE IT OUT.**

# Specifics – **the Client**

- The Client is not stupid, avoid patronizing **the Client**.
- On the other hand, the **Client** portrait may reveal lack of knowledge in specific areas: Adress these areas.
- In particular, **the Client**, as the intended reader, may be a non-statistician, and not trained in statistical methodology. If this is the case, consider this when communicating the results.
- The **Client** interests are typically the results and their context, and not how you got there. IE: The R code is typically **NOT** the interest of the Client. Put the R code in an appendix, the justification of the R code is often only for documentational purposes.

# Specifics – Confidence Intervals etc.



- An estimator is usually worthless unless accompanied by an assessment of the level of statistical uncertainty of it. Supply this uncertainty, always.
- Statistical uncertainty can be represented in many forms; but usually a **standard confidence interval** is the right choice.
- Consider this for predicted values as well, also in graphical presentations.
- Include the information in results and discussion.
- Include the full relevant information in summary statements. Example:
  - Simple statement: "Treatment A was significantly better than Treatment B."
  - Detailed statement: "The odds of success were increased with 80% for treatment A when compared to treatment B ( $p=0.002$ ), with a 95% confidence interval of 65% to 87%".
- The better detailed statement both includes **effect size, significance level** and **statistical uncertainty**.

# The Client and Your Report.... Make the Proper Reservations



"'BE CAREFUL'! ALL YOU CAN,  
TELL ME IS 'BE CAREFUL'?"

# Analysis of Categorical Data

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 9th, 2025

# Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, Writing statistical reports,  
Logistic regression

Friday Introduction to repeated measures , Principal  
Component Analysis

# Outline

## ① Categorical Data Introduction

- RR and OR

## ② Confounding

## ③ RxC Tables

## ④ Exercises

# Outline

## 1 Categorical Data Introduction

- RR and OR

## 2 Confounding

## 3 RxC Tables

## 4 Exercises

# Categorical Data

- **Binary data**
  - Yes/No
  - Dead/Alive
- **Nominal** ("label", several groups)
  - Eye colour: Blue/ Brown / Grey / Green
  - Where do you live: Denmark, Germany, Sweden.
- **Ordinal**
  - How do you feel today?: Very unhappy, unhappy, OK, happy, very happy.
  - Do you try to eat healthily?: Never, Sometimes, Always
- **Interval** (does have a numerical distance between values)
  - BMI categories (<25, 25-30, 30+).
  - Annual income groups.

## Example: Colour Blind

We have a study of 270 children where we have registered whether they were colour blind or not.

	Colour blind		
	Yes	No	Total
Girls	1	119	120
Boys	6	144	150
Total	7	263	270

## Example: Colour Blind

We have a study of 270 children where we have registered whether they were colour blind or not.

	Colour blind		
	Yes	No	Total
Girls	1	119	120
Boys	6	144	150
Total	7	263	270

Boy	Colour_blind	Count
1	1	6
1	0	144
0	1	1
0	0	119

Outcome: Colour blind yes/no

Covariate: Sex boy/girl.

## Example: Tables in R

```
colourTab <- xtabs(Count ~ Boy + Colour_blind, data = colour_dat)
```

```
#Print the table
ftable(colourTab)
```

	Colour_blind	0	1
Boy			
0		119	1
1		144	6

```
#Row totals
margin.table(colourTab,1)
```

Boy	0	1
120	150	

```
#Row percentages
prop.table(colourTab,1)
```

	Colour_blind		
Boy		0	1
0	0.991666667	0.008333333	
1	0.960000000	0.040000000	

```
#Column total
margin.table(colourTab,2)
```

Colour_blind	0	1
263	7	

# Risk Ratio

We want to compare the probability that a boy is colour blind ( $p_1$ ) with the probability that a girl is colour blind ( $p_0$ ).

- The probabilities are unknown.
- Variation from random sampling of children for the study.
- Estimate the probabilities

$$\hat{p}_1 = \frac{\text{"number colour blind boys"}}{\text{"number of boys"}} = \frac{6}{150} = 0.04$$

$$\hat{p}_0 = \frac{\text{"number colour blind girls"}}{\text{"number of girls"}} = \frac{1}{120} = 0.0083$$

# Risk Ratio

We want to compare the probability that a boy is colour blind ( $p_1$ ) with the probability that a girl is colour blind ( $p_0$ ).

- The probabilities are unknown.
- Variation from random sampling of children for the study.
- Estimate the probabilities

$$\hat{p}_1 = \frac{\text{"number colour blind boys"}}{\text{"number of boys"}} = \frac{6}{150} = 0.04$$

$$\hat{p}_0 = \frac{\text{"number colour blind girls"}}{\text{"number of girls"}} = \frac{1}{120} = 0.0083$$

- A measure to compare probabilities

- Risk Ratio (RR) =  $\frac{p_1}{p_0}$

## Example: Colour Blind

- \* How many colour blind children would we expect?
- \* Assume that the probability of colour blindness is 0.026 (7 out of 270), independent of gender.

Then we would expect:

  - for 150 boys:  $150 \times 0.026 = 3.9$
  - for 120 girls:  $120 \times 0.026 = 3.1$
- \* We observed 6 and 1; we need statistical methods to decide whether this was a coincidence, or whether colour blindness differs for girls and boys.

# Binomial Distribution

X=Number of events (colour blind children) out of N, with p= the probability of event.

$$P(X = x) = \binom{N}{x} p^x (1 - p)^{N-x}$$

Here  $p$  is the unknown parameter (the probability of colour blind). Our best guess at  $p$  (the estimate) is the observed proportion of colour blind.

$$\hat{p} = \frac{x}{N} = \frac{7}{270}$$

# Binomial Distribution Approximate CI

If N is large then

$$s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

The approximate confidence interval for  $\hat{p}$  using a Normal approximation:

$$\hat{p} \pm 1.96s.e.(\hat{p})$$

# Binomial Distribution Approximate CI

If N is large then

$$s.e.(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

The approximate confidence interval for  $\hat{p}$  using a Normal approximation:

$$\hat{p} \pm 1.96s.e.(\hat{p})$$

For the girls and boys from the example:

```
library(epitools)
binom.approx(colourTab[,2],margin.table(colourTab,1))
```

x	n.Boy	n.Freq	proportion.Freq	lower.Freq	upper.Freq
0	1	0	0.008333333	-0.007931503	0.02459817
1	6	1	0.040000000	0.008640576	0.07135942

# Binomial Distribution 'Exact' CI

Uses the correspondence between test and confidence interval (CI).  
The exact CI includes the p's that would be accepted in an 'exact' test.  
For the girls and boys from the example:

```
binom.exact(colourTab[ , 2], margin.table(colourTab, 1))
```

x	n.Boy	n.Freq	proportion.Freq	lower	upper
0	1	0	120	0.008333333	0.0002109595 0.04555551
1	6	1	150	0.040000000	0.0148185211 0.08502781

# Compare $p_0$ and $p_1$

- ① Risk ratio:  $\frac{p_1}{p_0}$ .
- ② Odds ratio:  $\frac{p_1}{1-p_1} / \frac{p_0}{1-p_0}$ .

But what are odds? - and why do we need them?

# Odds and Probability

- Definition of odds:

$$\begin{aligned}Odds(A) &= \frac{\text{Probability}(A)}{\text{Probability}(A \text{ does not happen})} \\&= \frac{\text{Probability}(A)}{1 - \text{Probability}(A)}\end{aligned}$$

- Back to probabilities

$$\text{Probability}(A) = \frac{Odds(A)}{1 + Odds(A)}$$

# Odds and Probability

- At the bookmaker: The odds for "SønderjyskE" winning against FCK.
- DanskeSpil 2 November 2014 odds=6.1.

$$Odds(SønderjyskE \text{ wins}) = \frac{P(SønderjyskE \text{ wins})}{P(SønderjyskE \text{ does not win})} = \frac{1}{6.1}$$

- The probability that SønderjyskE wins:

$$\frac{1/6.1}{1 + 1/6.1} = 0.14$$

- (Result: FCK-Sønderjydske 1-1)

# Characteristics of Odds

- ✓ Odds are between 0 and infinity.
- ✓ Often log odds (no boundaries).
- ✓ When the probability is 0.5 then odds are 1.
- ✓ Odds are larger than probability.
- ✓ **Note:** When the probability is small ( $\leq 0.1$ ) then probability and odds nearly equal.

# Odds Ratio

The odds ratio is the ratio between the odds in the two groups.

$$OR = \frac{Odds(group_1)}{Odds(group_0)} = \frac{p_1}{1-p_1} / \frac{p_0}{1-p_0}$$

Group	Response	
	No	Yes
0 (ref)	a	b
1	c	d

$$OR = \frac{d/(c+d)/c/(c+d)}{b/(a+b)/a(a+b)} = \frac{d/c}{b/a} = \frac{ad}{bc}$$

# OR in R

```
library(epitools)
oddsratio(colourTab, method = "wald")

$data
    Colour_blind
Boy      0 1 Total
  0     119 1   120
  1     144 6   150
Total  263 7   270

$measure
  odds ratio with 95% C.I.
Boy estimate      lower      upper
  0 1.000000        NA        NA
  1 4.958333 0.5887152 41.76055
```

# OR in R

```
library(epitools)
oddsratio(colourTab, method = "wald")

$data
    Colour_blind
Boy      0 1 Total
  0     119 1   120
  1     144 6   150
Total  263 7   270

$measure
  odds ratio with 95% C.I.
Boy estimate      lower      upper
  0 1.000000        NA        NA
  1 4.958333 0.5887152 41.76055
```

Odds for a boy being colour blind are 4.96 (95% CI 0.6 to 41.8) times larger than for girls.

# RR in R

Remember the  $RR = \frac{p_1}{p_0}$

```
riskratio(colourTab)
```

```
$data
```

	Colour_blind		
Boy	0	1	Total
0	119	1	120
1	144	6	150
Total	263	7	270

```
$measure
```

risk ratio with 95% C.I.

Boy	estimate	lower	upper
0	1.0	NA	NA
1	4.8	0.5858252	39.32914

## RR in R

Remember the  $RR = \frac{p_1}{p_0}$

```
riskratio(colourTab)
```

```
$data
```

	Colour_blind		
Boy	0	1	Total
0	119	1	120
1	144	6	150
Total	263	7	270

```
$measure
```

risk ratio with 95% C.I.

Boy	estimate	lower	upper
0	1.0	NA	NA
1	4.8	0.5858252	39.32914

The risk of a boy being colour blind is 4.8 (95% CI 0.6 to 39.3) times larger than for girls.

# Odds Ratio and Risk Ratio

- OR varies freely from 0 to infinity.
- RR always between 1 and OR.
- OR is symmetric

$$OR(response = 1) = \frac{1}{OR(response = 0)}$$

- RR is not symmetric

$$RR(response = 1) \neq \frac{1}{RR(response = 0)}$$

- For rare events,  $OR \approx RR$ .

# $\chi^2$ (Chisquare) Test

The Hypothesis: OR=1 or equivalently RR=1.

Observed:

Group	Response		Total
	No	Yes	
0 (ref)	a	b	a+b
1	c	d	c+d
Total	a+c	b+d	N

# $\chi^2$ (Chisquare) Test

The Hypothesis:  $OR=1$  or equivalently  $RR=1$ .

Observed:

Group	Response		Total
	No	Yes	
0 (ref)	a	b	$a+b$
1	c	d	$c+d$
Total	$a+c$	$b+d$	N

Expected:

Group	Response		Total
	No	Yes	
0 (ref)	$(a+b)(a+c)/N$	$(a+b)(b+d)/N$	$a+b$
1	$(c+d)(a+c)/N$	$(c+d)(b+d)/N$	$c+d$
Total	$a+c$	$b+d$	N

$$\chi^2 = \sum \frac{(Obs - Expected)^2}{Expected}$$

# Test Colour Blind

Are the odds of being colour blind the same for boys and girls?  
Equivalently is colour blindness independent of sex?

$$H_0 : p_0 = p_1$$

Use:

- $\chi^2$  test.

```
> chisq.test(colourTab, correct=FALSE)
```

Pearson's Chi-square test

```
data: colourTab  
X-squared = 2.6472, df = 1, p-value = 0.1037
```

# Test in Example

The Hypothesis:  $OR=1$  or equivalently  $RR=1$ .

Observed:

Obs Expected	Colour Blind		Total
	No	Yes	
Boy			
0 (ref)	119	1	120
	116.9	3.1	
1	144	6	150
	146.1	3.9	
Total	263	7	270

# OR and Chi2 test in R

```
> epitools::oddsratio(colourTab, method="wald")
$data
  Colour_blind
Boy      0 1 Total
  0     119 1    120
  1     144 6    150
Total  263 7    270

$measure
  odds ratio with 95% C.I.
Boy estimate      lower      upper
  0 1.000000        NA        NA
  1 4.958333 0.5887152 41.76055

$p.value
  two-sided
Boy midp.exact fisher.exact chi.square
  0          NA          NA          NA
  1 0.1200585 0.1363846 0.1037323
```

# OR and Chi2 test in R

```
> epitools::oddsratio(colourTab, method="wald")
$data
  Colour_blind
Boy      0 1 Total
  0     119 1    120
  1     144 6    150
Total  263 7    270

$measure
  odds ratio with 95% C.I.
Boy estimate      lower      upper
  0 1.000000        NA        NA
  1 4.958333 0.5887152 41.76055

$p.value
  two-sided
Boy midp.exact fisher.exact chi.square
  0        NA        NA        NA
  1  0.1200585  0.1363846  0.1037323
```

The hypothesis of  $OR=1$  is accepted  $p=0.14 > 0.05$ , but CI very wide.

# Exercise

Identify each variable as nominal, ordinal or interval.

- ① UK political party preference (Labour, Conservative, Social Democrat).
- ② Depression rating (none, mild, moderate, severe, very severe).
- ③ Patient survival (in number of months).
- ④ University location (Lyngby, Copenhagen, Odense, Aarhus, Aalborg).
- ⑤ Favorite beverage (water, juice, milk, soft drink, beer, wine).
- ⑥ Appraisal of company's inventory level (too low, about right, too high).

# Outline

- 1 Categorical Data Introduction
  - RR and OR

- 2 Confounding

- 3 RxC Tables

- 4 Exercises

# Confounding

Instead of just a risk factor (boy/girl) and an outcome (colour blindness) one might have a third factor.

**Example:** Two treatments (A and B) for kidney stone. The outcome is success or failure of the treatment. We also have registered whether the stone was small or large.

	Treatment	Stone	Success1	Count
1	A	Small	1	81
2	A	Small	0	6
3	A	Large	1	192
4	A	Large	0	71
5	B	Small	1	234
6	B	Small	0	36
7	B	Large	1	55
8	B	Large	0	25

# Tables in R

```
mytable <- xtabs(Count ~ Treatment + Success1 + Stone,  
data = kidney)  
ftable(mytable)
```

		Stone	Large	Small
Treatment	Success1			
A	0		71	6
	1		192	81
B	0		25	36
	1		55	234

The order of the variables in `xtabs` is important. First exposure, second outcome, last extra factors.

# Ignoring the Size of the Stone

```
Treat_Succ <- margin.table(mytable, 1:2)
oddsratio(Treat_Succ, method = "wald")
```

```
$data
```

```
    Success1
```

Treatment	0	1	Total
A	77	273	350
B	61	289	350
Total	138	562	700

```
$measure
```

```
    odds ratio with 95% C.I.
```

Treatment	estimate	lower	upper
A	1.000000	NA	NA
B	1.336276	0.9188954	1.943238

The odds of success for treatment B are 1.34 times the odds for A.

# The Effect of Treatment for Small Stones

```
Small <- mytable[ , , 2]
oddsratio(Small, method = "wald")

$data
    Success1
Treatment  0   1 Total
      A     6  81   87
      B    36 234   270
      Total 42 315   357

$measure
    odds ratio with 95% C.I.
Treatment  estimate      lower      upper
      A 1.00000000      NA      NA
      B 0.4814815 0.1956696 1.184775
```

Treatment A is better for small stones OR=0.48.

# The Effect of Treatment for Large Stones

```
Large <- mytable[ , , 1]
oddsratio(Large, method = "wald")

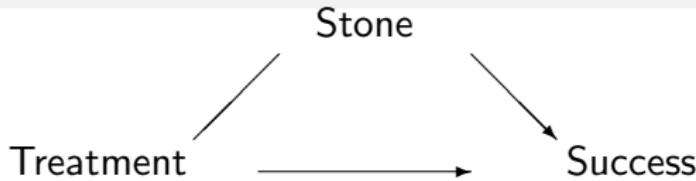
$data
  Success1
Treatment  0   1 Total
  A      71 192 263
  B      25  55  80
  Total 96 247 343

$measure
  odds ratio with 95% C.I.
Treatment  estimate    lower    upper
  A 1.0000000    NA      NA
  B 0.8135417 0.47147 1.403801
```

Treatment A is better for large stones OR=0.81.

How can this happen?

# Confounding



A confounder is:

- Associated with outcome:  
e.g., smaller kidney stones have higher rate of success.
- Associated with the treatment:  
e.g., doctors have chosen treatment A for difficult cases.
- Not a result of treatment, i.e. not an intermediate variable.  
Not a statistical property; cannot be seen from tables; common sense is required.

# Smaller kidney stones have higher rate of success.

```
Stone_Succ <- margin.table(mytable, 3:2)
oddsratio(Stone_Succ, method = "wald")
```

```
$data
```

```
    Success1
```

Stone	0	1	Total
Large	96	247	343
Small	42	315	357
Total	138	562	700

```
$measure
```

```
    odds ratio with 95% C.I.
```

Stone	estimate	lower	upper
Large	1.00000	NA	NA
Small	2.91498	1.955863	4.344429

# Smaller kidney stones have higher rate of success.

```
Stone_Succ <- margin.table(mytable, 3:2)
oddsratio(Stone_Succ, method = "wald")
```

```
$data
```

```
    Success1
```

Stone	0	1	Total
Large	96	247	343
Small	42	315	357
Total	138	562	700

```
$measure
```

```
    odds ratio with 95% C.I.
```

Stone	estimate	lower	upper
Large	1.00000	NA	NA
Small	2.91498	1.955863	4.344429

OR 2.9 (95% CI 1.96 to 4.34) for success with small stone compared to large.

# Doctors have chosen treatment A for difficult cases.

```
Stone_Treat <- margin.table(mytable, c(3,1))  
oddsratio(Stone_Treat, method = "wald")
```

```
$data
```

```
    Treatment
```

Stone	A	B	Total
Large	263	80	343
Small	87	270	357
Total	350	350	700

```
$measure
```

```
    odds ratio with 95% C.I.
```

Stone	estimate	lower	upper
Large	1.00000	NA	NA
Small	10.20259	7.20504	14.44721

# Doctors have chosen treatment A for difficult cases.

```
Stone_Treat <- margin.table(mytable, c(3,1))  
oddsratio(Stone_Treat, method = "wald")
```

```
$data
```

```
    Treatment
```

Stone	A	B	Total
Large	263	80	343
Small	87	270	357
Total	350	350	700

```
$measure
```

```
    odds ratio with 95% C.I.
```

Stone	estimate	lower	upper
Large	1.00000	NA	NA
Small	10.20259	7.20504	14.44721

Small stones have been treated with B.

# Controlling for Confounding

- We could have randomized the treatment.
- We can keep the confounder constant.

# Controlling for Confounding

- We could have randomized the treatment.
- We can keep the confounder constant.

Hold the confounder constant:

- Compare treatments within strata (small stones, and large).
- **If the estimates are similar** we calculate a combined estimate as a suitable average (No more on this today).
- Fit a logistic regression model (more about this in the afternoon).

# Outline

- 1 Categorical Data Introduction
  - RR and OR

- 2 Confounding

- 3 RxC Tables

- 4 Exercises

# RxC tables

Observed Expected	Caffeine Intake				Total
	0	1-150	151-300	300+	
Married	652	1537	598	242	3029
	705.83	1488.01	578.07	257.09	
Prev. Married	36	46	38	21	141
	32.86	69.27	26.91	11.97	
Single	218	327	106	67	718
	167.31	352.72	137.03	60.94	
Total	906	1910	742	330	3888

# Chi-square test in RxC tables

- As for a 2x2 table.
- Hypothesis: Caffeine intake the same irrespective of marital status (independence in table).

$$\chi^2 = \sum \frac{(\text{Observed}-\text{Expected})^2}{\text{Expected}}$$

- Follows a  $\chi^2$  distribution with  $(r-1)(c-1)$  degrees of freedom.

# Chi-square test in RxC tables, contd.

- Test for independence gives a p-value, **but we are not finished yet.**
- **If the test is significant.**
  - Describe the connections. The p-values does not show where the associations are found.
- **If the test is not significant.**
  - There might still be some associations.
- **In both cases describe the table with percentages and plots.**

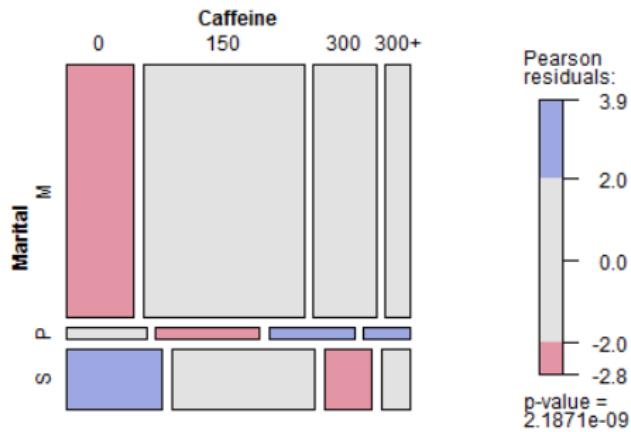
# RxC tables

Observed Row %	Caffeine Intake				Total
	0	1-150	151-300	300+	
Married	652 21.53	1537 50.74	598 19.74	242 7.99	3029
Prev. Married	36 25.53	46 32.62	38 26.95	21 14.89	141
Single	218 30.36	327 45.54	106 14.76	67 9.33	718
Total	906	1910	742	330	3888

Pearsons  $\chi^2(6) = 51.6556$ ,  $p < 0.0001$

# Mosaic Plot

```
library(vcd)
mosaic(mytable, shade = TRUE, legend = TRUE)
```



# Outline

## 1 Categorical Data Introduction

- RR and OR

## 2 Confounding

## 3 RxC Tables

## 4 Exercises

# Exercises

- Exercise 1 Admission to Berkeley
- Exercise 2 Popular

# Exercise 1, Applicants to Berkeley

---

Friday June 7, 2024

## Story

The data are from applicant data from the graduate school at the University of California at Berkeley for autumn 1973. It presents decisions by gender of applicant for the six largest graduate departments. The University of California, Berkeley was sued for bias against women who had applied for admission to graduate schools there. The admission figures for the fall of 1973 showed that men applying were more likely than women to be admitted.

## Data

File: Admission.txt.

Department : A, B, C, D, E and F.

Sex: M or F.

Admitted: No or Yes

Count: The number of applicants for each combination of the three factors.

## Exercise

- Read in the data admission.txt into a three-way table.
- Calculate the OR and RR (with 95% confidence limits) for men being admitted compared to women not taking department into account.
- Is the probability for being admitted the same for men and women?
- Do all departments have the same probability of admitting? Illustrate the results with proportions and a plot.
- Do men and women apply to the same departments? Illustrate the results with proportions and a plot.
- Why is department a potential confounder for the effect of gender on being admitted?
- Calculate the OR and RR for men vs. women admitted for each department separately.

## Hint

If the full three-way table is found in mytable organized with Sex, Admitted, Department then you can take the 2x2 table for department A:

```
DepA <- mytable[, , 1]
```

# Exercise 2: Case: Popular

---

*Friday June 7, 2024*

## Story

Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.

## Data

File: popular.txt

variable name	description
Gender	Boy or girl
Grade	4, 5 or 6
Age	Age in years
Race	White, Other
Urban/Rural	Rural, Suburban, or Urban school district
School	Brentwood Elementary, Brentwood Middle, Ridge, Sand, Eureka, Brown, Main Portage, Westdale Middle
Goals	Student's choice in the personal goals question where options were 1 = Make Good Grades, 2 = Be Popular, 3 = Be Good in Sports
Grades	Rank of "make good grades"(1=most important for popularity, 4=least important)
Sports	Rank of "being good at sports"(1=most important for popularity, 4=least important)
Looks	Rank of "being handsome or pretty"(1=most important for popularity, 4=least important)
Money	Rank of "having lots of money"(1=most important for popularity, 4=least important)

## Exercise

1. Analyze the relationship between gender and goals.
2. Analyze the relationship between age and goals.
3. Could you suggest other analysis that may be interesting?
4. Save the results of your analysis in a text document (e.g. latex, word or star-office).

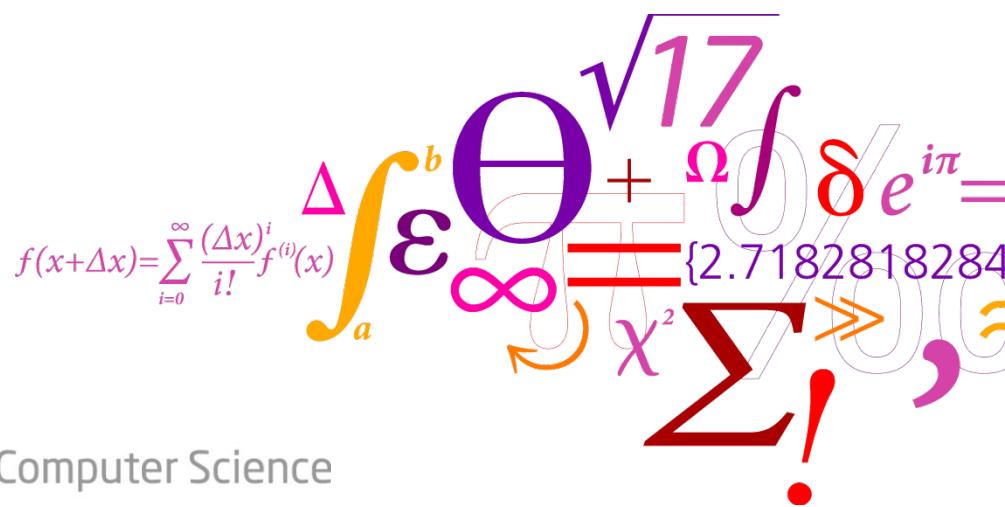
# Logistic Regression

January 9th, 2025

Anders Stockmarr

Section for Statistics and Data Analysis, DTU

anst@dtu.dk

$$f(x+\Delta x) = \sum_{i=0}^{\infty} \frac{(\Delta x)^i}{i!} f^{(i)}(x)$$


DTU Compute

Department of Applied Mathematics and Computer Science

---

# Programme

- Monday : Statistical Inference, the t-test
- Tuesday : Simple and Multiple Regression
- Wednesday : ANOVA, ANCOVA, and Linear Models
- Thursday: Categorical Data, Writing Statistical Reports, **Logistic regression**
- Friday: Repeated Measurements, Principal Component Analysis

# Contents:

1. Introduction.
2. Main example: Sperm competition among horseshoe crabs.
3. Exercise.
3. Logistic regression for frequency data.
5. Exercises.
6. Logistic regression for ordinal data.

# **What you should be able to do after the lecture:**

- a) Identify data suitable for logistic regression.
- b) Carry out simple logistic regression analyses, and estimate the parameters.
- c) Perform standard model control of logistic regression models.

# Introduction

# Logistic Regression

- Applies to:
- Binary data
  - Yes/No
  - Dead/Alive
- Frequency data
  - Percentage of sick people
  - Ratio of bycatch for fishing trawlers
- [Nominal data]
- Ordinal data

# Color Blind Example

X=Number of events (colour blind children) out of N. With p the probability of event, it holds that

$$P(X = x) = \binom{N}{x} p^x (1 - p)^{N-x};$$

X is binomially distributed  $(N, p)$ .

The optimal estimator for p is the observed proportion of colour blind:

$$\hat{p} = X/N$$

In the example from the Categorical Data Session,

$$\hat{p} = 7/270 \approx 0.026$$

# Color Blind Example

$$\hat{p} = X/N$$

Requires the observations to be *repetitions*;

Ie. each person investigated is assumed to have the **same** probability  $p$  of being colorblind.

If this probability *varies from person to person*; depending on t.ex. gender, but perhaps also quantitative genetic information (such as t.ex. the number of alleles at a locus associated with color blindness), a different type of analysis is required.

# Data Example: Sperm Competition in Horseshoe Crabs



<http://people.biology.ufl.edu/dsasson>

# Crab Data

```
setwd("C:/<your data directory>")
crab.data<-read.table("crab.data.txt")
head(crab.data)
```

	color	spine	width	satell	weight	y
1	3	3	28.3	8	3050	1
2	4	3	22.5	0	1550	0
3	2	1	26.0	9	2300	1
4	4	3	24.8	0	2100	0
5	4	3	26.0	4	2600	1
6	3	3	23.8	0	2100	0

Nominal variables;  
categories

# Horseshoe Crab Data Analysis

Central Question for sperm competition analysis:

**What is the probability that a female has a satellite?**

"y" in the crab dataset denotes the presence/absence of satellites.

If satellites attach themselves to females completely at random,  $\mathbb{Y}$  will be either 0 or 1 with the same probability for all individuals:

$$P(Y = 1) = p; \quad P(Y = 0) = 1 - p,$$

Where  $p$  is the probability of having a satellite.

# Horseshoe Crab Data Analysis

- Let  $\mathbb{X}$  be the number of females with a satellite attached; then

$$P(\mathbb{X} = x) = \binom{N}{x} p^x (1 - p)^{N-x}, \quad \hat{p} = \frac{\mathbb{X}}{N}.$$

Finding N and  $\hat{p}$ :

```
N<-length(crab.data$y)
```

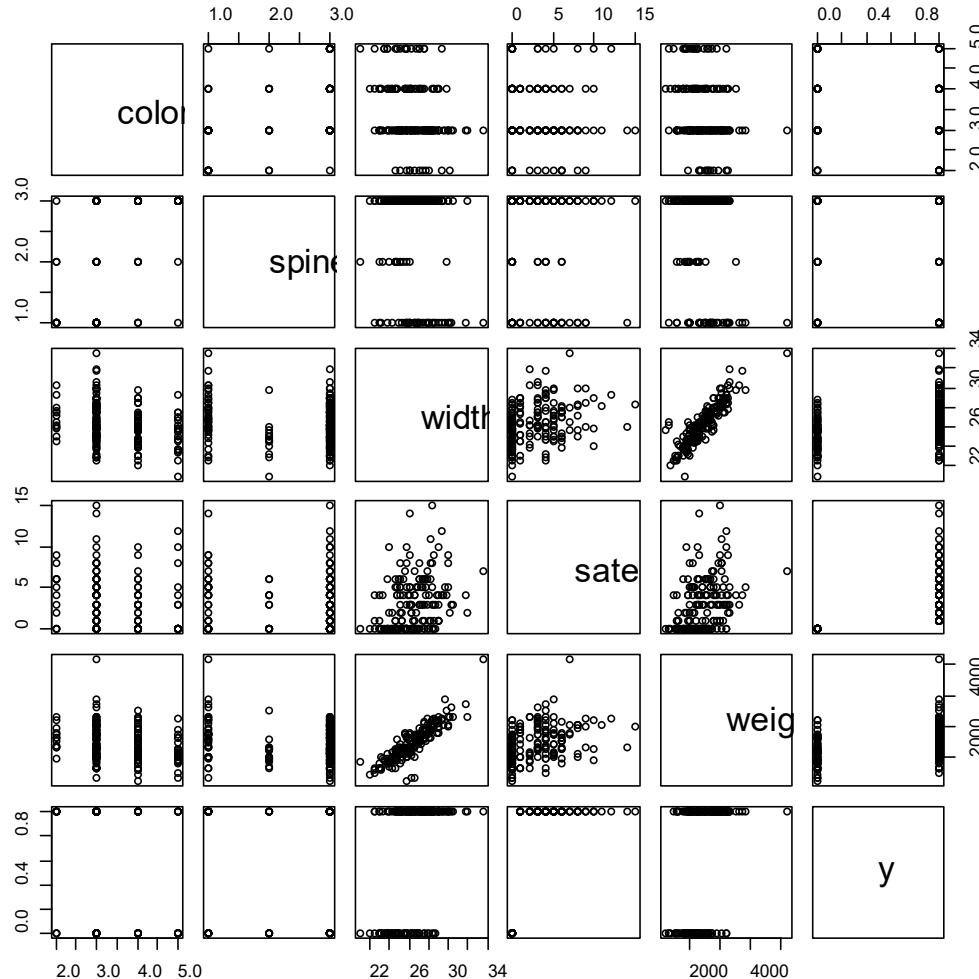
```
N
```

```
[1] 173
```

```
sum(y)/N
```

```
[1] 0.6416185
```

# Horseshoe Crab Data Analysis



# Horseshoe Crab Data Analysis

*H:  $p$  depends on the width of the crab*

*Linear regression is one bid on how to model the effect. But there isn't really much hope, as the data hardly satify the normality assumption.*

*Lets see what happens....*

# Horseshoe Crab Data Analysis

- Linear regression:

```
analysis<-lm(y~width, data=crab.data)
```

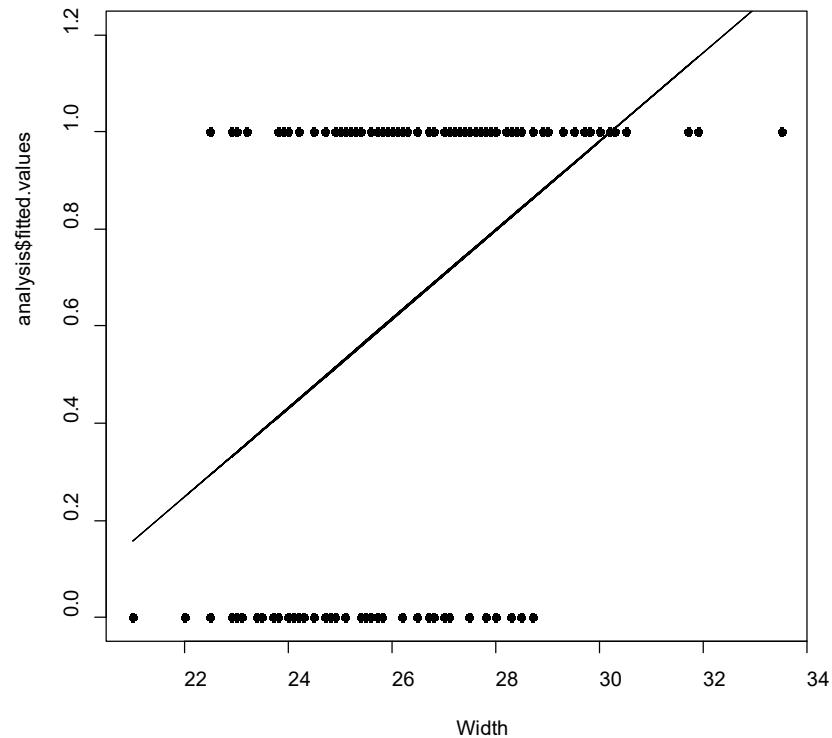
```
analysis
```

```
Call:
```

```
lm(formula = y ~ width,  
   data = crab.data)
```

Coefficients:

(Intercept)	width
-1.76553	0.09153



# Dependency of Width: Logistic Regression

Odds of a satellite (similar to Categorical Data session):

$$\frac{p}{1 - p}$$

Log(odds):

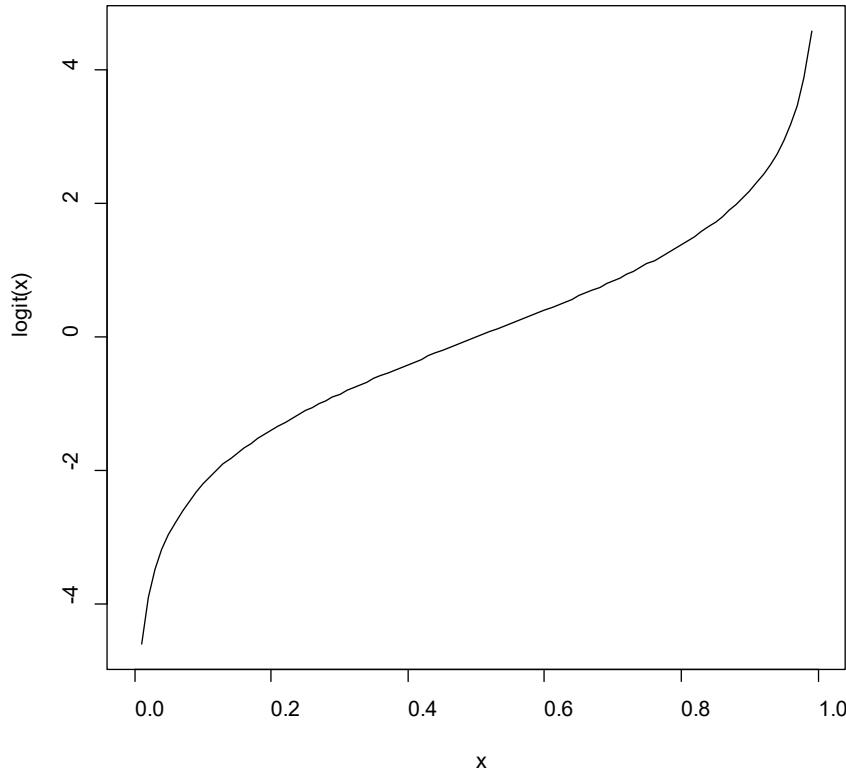
$$\log\left(\frac{p}{1 - p}\right)$$

This is the logit function:

$$\text{logit}(p) = \log\left(\frac{p}{1 - p}\right)$$

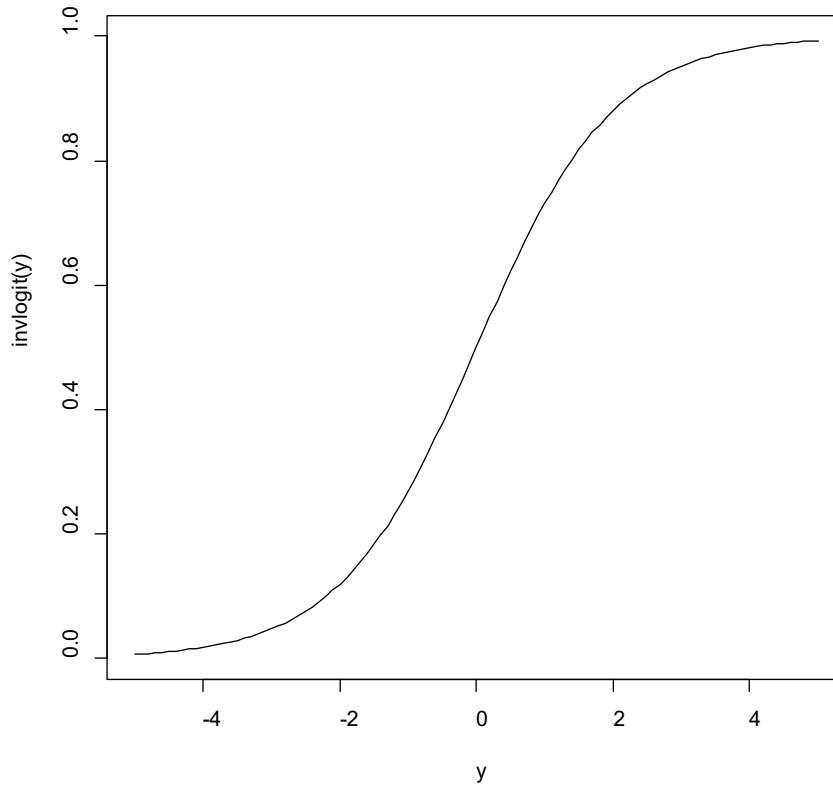
# The Logit Function

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$



# The Inverse Logit Function

$$\text{invlogit}(y) = \frac{e^y}{1 + e^y}$$



# Dependency of Width: Logistic Regression

Model:

$$\text{logit}(p) = \alpha + \beta \cdot \text{width}$$

R: Use the `glm` function with option `family=binomial(link=logit)` :

```
analysis<-glm(y~width,family=binomial(link=logit),data=crab.data)
analysis
```

```
Call: glm(formula = y ~ width, family = binomial(link = logit), data =
crab.data)
```

Coefficients:

(Intercept)	width
-12.3508	0.4972

Degrees of Freedom: 172 Total (i.e. Null); 171 Residual

Null Deviance: 225.8

Residual Deviance: 194.5 AIC: 198.5

# Dependency of Width: Logistic Regression

Model:

$$\text{logit}(p) = \alpha + \beta \cdot \text{width}$$

```
summary(analysis)
```

Call:

```
glm(formula = y ~ width, family = binomial(link = logit), data = crab.data)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.0281	-1.0458	0.5480	0.9066	1.6942

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	<u>-12.3508</u>	2.6287	-4.698	2.62e-06 ***
width	<u>0.4972</u>	0.1017	4.887	1.02e-06 ***
---				

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

(Dispersion parameter for binomial family taken to be 1)

```
Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 194.45 on 171 degrees of freedom
AIC: 198.45
```

Number of Fisher Scoring iterations: 4

# Dependency of Width: Logistic Regression

Model:

$$P(\mathbb{Y}_i = 1) = p_i, \quad \text{logit}(p_i) = \alpha + \beta \cdot \text{width}_i, \quad i = 1, \dots, 173.$$

Test:

```
drop1(analysis, test="Chisq")
```

Single term deletions

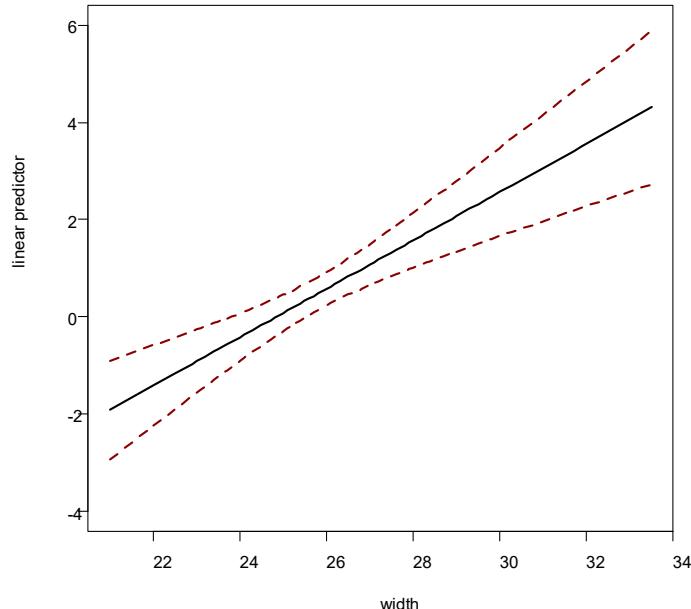
Model:

```
y ~ width
```

	Df	Deviance	AIC	LRT	Pr (>Chi)						
<none>		194.45	198.45								
width	1	225.76	227.76	31.306	2.204e-08 ***						
				---							
Signif. codes:	0	'***'	0.001	'**'	0.01	'*'	0.05	'.'	0.1	' '	1

# Model Control I:

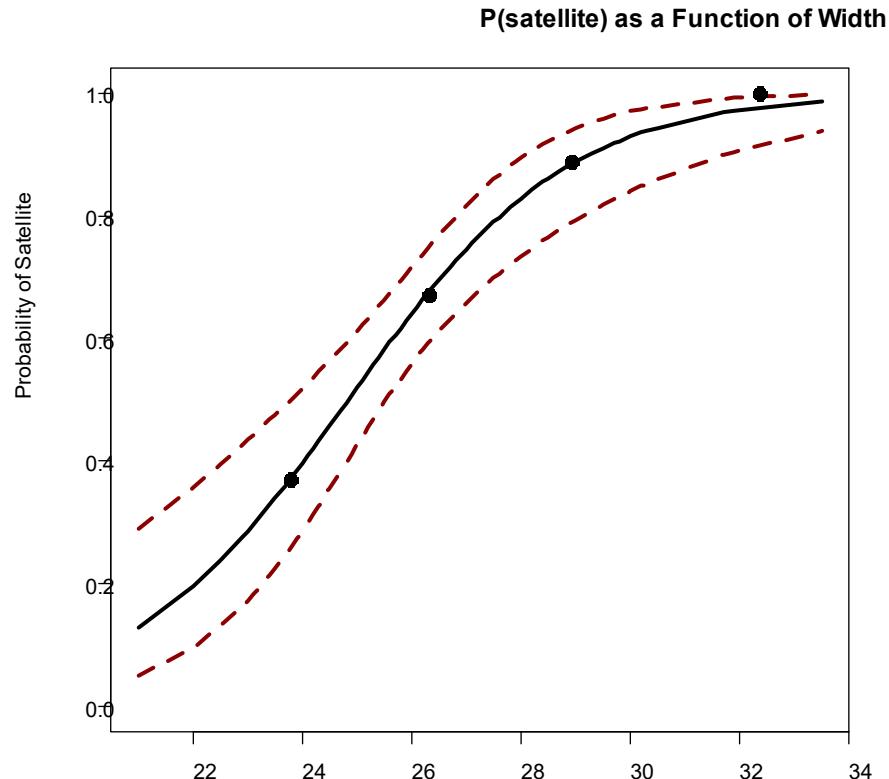
```
prediction.temp<-as.data.frame(predict(analysis,se.fit=T))  
prediction.data<-data.frame(pred=prediction.temp$fit,  
                             upper=prediction.temp$fit+  
                             1.96*prediction.temp$se.fit,  
                             lower=prediction.temp$fit-  
                             1.96*prediction.temp$se.fit)
```



# Model Control II:

```
prediction.data.original<-invlogit(prediction.data)
```

Plot with original data grouped frequencies:



# Model Control III:

Polynomial regression:

$$\text{logit}(p_i) = \alpha + \beta \cdot \text{width}_i + \gamma \cdot \text{width}_i^2$$

```
analysis2<-update(analysis,~.+I(width^2))  
drop1(analysis2,test="Chisq")
```

Single term deletions

Model:

```
y ~ width + I(width^2)
```

	Df	Deviance	AIC	LRT	Pr (>Chi)
<none>		193.63	199.63		
width	1	194.10	198.10	0.47378	0.4913
I(width^2)	1	194.45	198.45	0.82542	0.3636

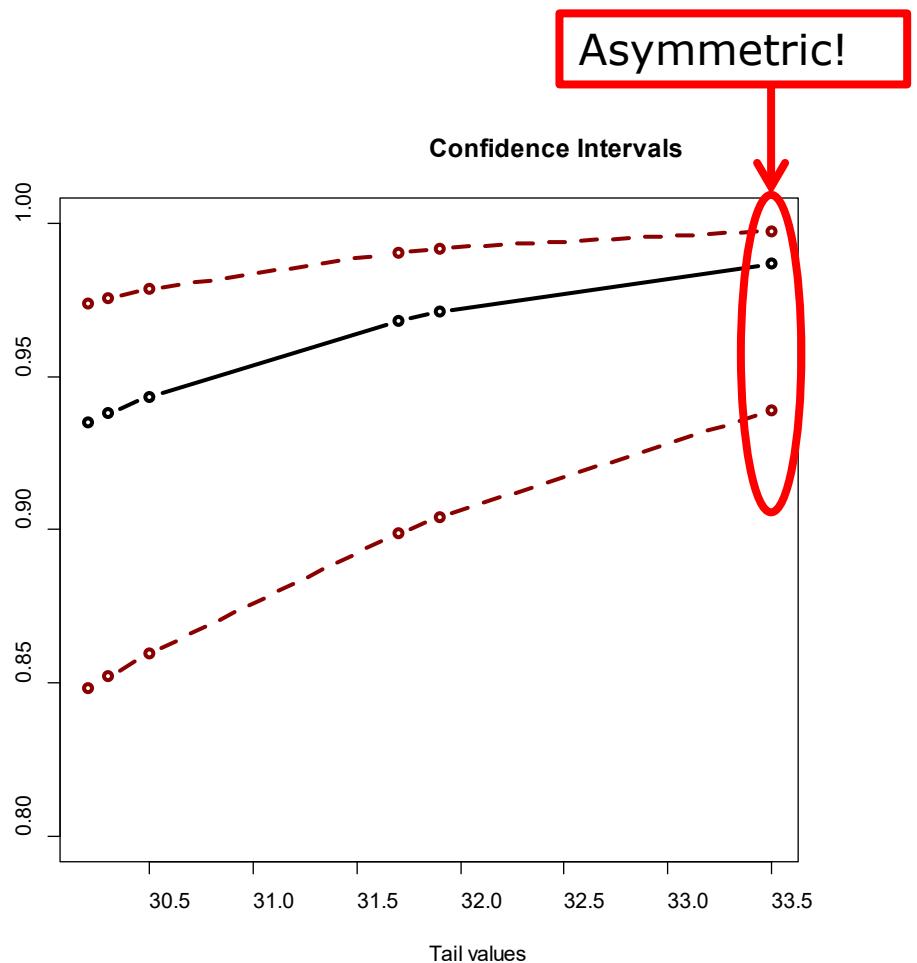
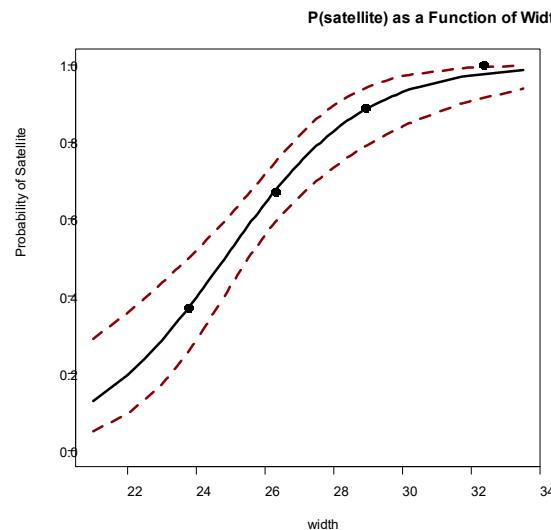
# Model control IV:

- Summary:
- Plot the predictors and the confidence intervals, group the original data, and check if they fall into the confidence area.
- Polynomial regression; if multiple covariates apply consider interaction terms (ie. the product of the covariates).

It is concluded that the model is a fair description of the data.

# Crab Data Analysis

```
tail(prediction.data.original)
pred      upper      lower
168 0.9349627 0.9736655 0.8482453
169 0.9379216 0.9753623 0.8522055
170 0.9434658 0.9784454 0.8598511
171 0.9680587 0.9904320 0.8987182
172 0.9709946 0.9916535 0.9041445
173 0.9866974 0.9972205 0.9387802
```



# Crab Data Analysis Revisited

```
summary(analysis)
```

```
glm(formula = y ~ width, family = binomial(link = logit), data  
= crab.data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-12.3508	2.6287	-4.698	2.62e-06 ***
width	0.4972	0.1017	4.887	1.02e-06 ***
---				

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

$$\hat{\alpha} = -12.3508, \quad v(\hat{\alpha}) = 2.6287^2 = 6.910$$
$$\hat{\beta} = 0.4972, \quad v(\hat{\beta}) = 0.1017^2 = 0.01035$$

# Crab Data Analysis Revisited

Estimates are correlated. Covariance between estimators:

```
summary(analysis)$cov.scaled
```

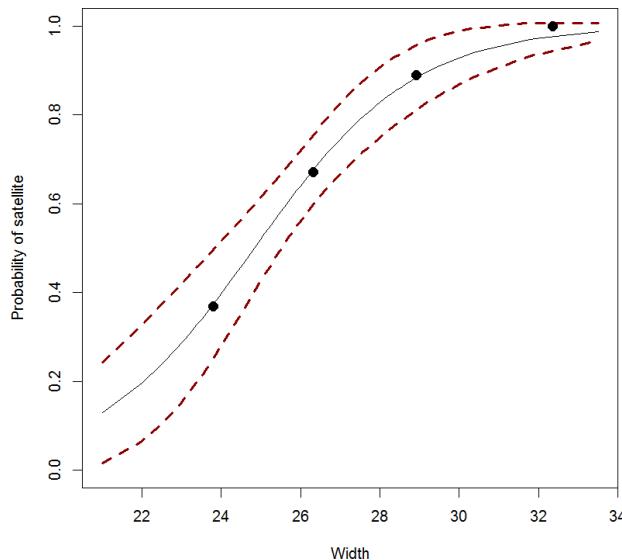
	(Intercept)	width
(Intercept)	6.9101576	-0.26684761
width	-0.2668476	0.01035012

$$\text{cov}(\hat{\alpha}, \hat{\beta})$$

# Prediction Intervals – Brute Force (not recommended)

- You can predict directly on the original scale with predict:

```
predict(analysis, type="response", se.fit=T)
```

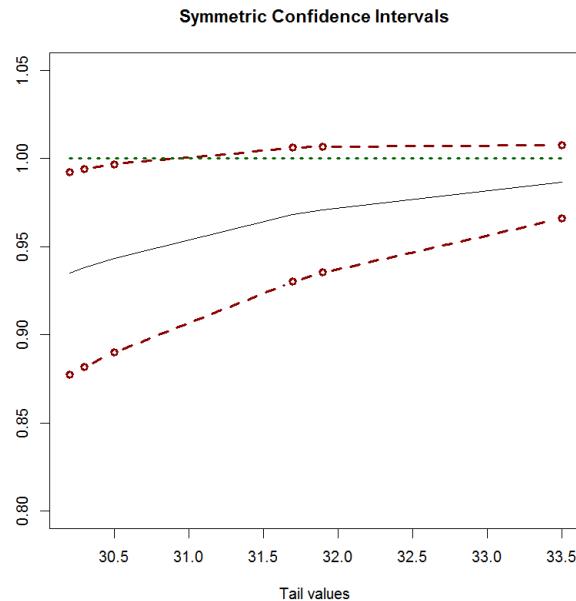


- Symmetric intervals; do not reflect that the link scale is the appropriate for that.

# Prediction Intervals – Brute Force (not recommended)

- You can predict directly on the original scale with predict:

```
predict(analysis, type="response", se.fit=T)
```



Hard to interpret intervals - t. ex. values above 1; not super for a probability.

**Assigning  $\pm 1.96\text{sd}$  should be done on the link scale, not the original/response scale.**

# Prediction Intervals

Suppose we have an additional crab with width w.

What is a 95% confidence interval for this crab to have satellites?

$$\text{logit}(\hat{p}) = \hat{\alpha} + \hat{\beta}w$$

$$sd = sd(\text{logit}(\hat{p}))$$

$$\hat{p} = \frac{e^{\hat{\alpha} + \hat{\beta}w}}{1 + e^{\hat{\alpha} + \hat{\beta}w}} = \text{invlogit}(\hat{\alpha} + \hat{\beta}w)$$
$$\text{upper} = \text{invlogit}(\hat{\alpha} + \hat{\beta}w + 1.96 * sd)$$
$$\text{lower} = \text{invlogit}(\hat{\alpha} + \hat{\beta}w - 1.96 * sd)$$

# Prediction Intervals

- Assume w=15. In R, use the newdata option in predict():

```
new.data<-data.frame(width=15)
new.prediction<-predict(analysis,newdata=new.data,se.fit=TRUE)
new.prediction.2<-data.frame(fit=new.prediction$fit,
                               upper=new.prediction$fit+1.96*
                                   new.prediction$se.fit,
                               lower=new.prediction$fit-1.96*
                                   new.prediction$se.fit)

invlogit(new.prediction.2)
      fit        upper       lower
1 0.007447815 0.06206388 0.00085019
```

Syntax: ?predict.glm

# Exercise:

recall that we have defined the main model as

```
analysis<-glm(y~width, family=binomial(link=logit), data=crab.data)
```

- 1) Plot the crab data again: `plot(crab.data)`
- 2) Deduce from the graph that another possible predictor for a satellite is the crab weight. Use the `update()` function to add weight to the model as on slide 24. How does that alter the model? If you should choose between width and weight, which one would you choose?
- 3) A third possible predictor for satellites is the color of the female. The color is a nominal covariate where higher value indicates darker skin, so it is added to the model as a factor:

```
analysis2<-update(analysis, ~.+ as.factor(color))
```

Check that color does not add significantly to the model. Which color label stands out the most?

- 4) Create a new dataset that included an indicator for darkskinned females:

```
crab.data.2<-data.frame(crab.data, dark=1*(crab.data$color==5))
```

Add 'dark' to the model with the command

```
analysis2<-update(analysis, ~.+dark, data=crab.data.2)
```

Do satellite males prefer light-skinned or dark-skinned females, or are they indifferent?

# Real Interest:

## Does Horseshoe Crabs recognize high fertility?

- Light Skin of horseshoe female crabs is associated with increased fertility!
- To investigate this, we are not really interested in the effect of the width;
- But we have to model the effect of the width, as it is a **confounder** for the color preference; an attachment could be either because the female is wide, or because it has a light colored skin.
- To model the width effect, the logistic regression model is obvious.

# What if we just used the t-test?

```
t.test(crab.data.3$y[crab.data.3$dark==0],  
        crab.data.3$y[crab.data.3$dark==1])
```

yields p=0.002.

BUT:

```
mean(crab.data.3$width[crab.data.3$dark==1])
```

```
[1] 25.28182
```

```
mean(crab.data.3$width[crab.data.3$dark==0])
```

```
[1] 26.44702
```

```
t.test(crab.data.3$width[crab.data.3$dark==0],  
        crab.data.3$width[crab.data.3$dark==1])
```

yields p=0.01: In this dataset, light.skinned crabs are significantly wider than dark-skinned crabs.

**We cannot know if the conclusion from the t-test is because of the color or the width.**

# Logistic Regression For Frequency Data

# Smoking, Obesity, Snoring (SOS)

Effect of Smoking, Obesity and Snoring on **Hypertension** (Altman (1991, page 353)):

```
sosdata<-read.table("sosdata.txt")
```

```
sosdata
```

	smoking	obesity	snoring	n.tot	n.hyp
1	No	No	No	60	5
2	Yes	No	No	17	2
3	No	Yes	No	8	1
4	Yes	Yes	No	2	0
5	No	No	Yes	187	35
6	Yes	No	Yes	85	13
7	No	Yes	Yes	51	15
8	Yes	Yes	Yes	23	8

# Smoking, Obesity, Snoring (SOS)

Model: Let  $p$  be the probability of hypertension. Then

$$\text{logit}(p) = \alpha + \beta_{\text{smoking}} + \beta_{\text{obese}} + \beta_{\text{snoring}}$$

Thus: The odds ratio of hypertension for a smoker vs. a non-smoker, with the same snoring and obesity status, is given by

$$\exp(\beta_{\text{smoking}})$$

Coding in R:

```
analysis.sos<-glm(n.hyp/n.tot~smoking + obesity + snoring,  
family=binomial(link=logit), weights = n.tot)
```

Compactified table  
on slide 37!  
Requires weights.

# Smoking, Obesity, Snoring (SOS)

```
analysis.sos<-glm(n.hyp/n.tot~smoking + obesity + snoring,  
family=binomial(link=logit), weights = n.tot)  
  
summary(analysis.sos)  
Call:  
glm(formula = n.hyp/n.tot ~ smoking + obesity + snoring, family  
= binomial(link = logit),  
      data = sosdata, weights = n.tot)  
  
Coefficients:  
              Estimate Std. Error z value Pr(>|z|)  
(Intercept) -2.37766   0.38018  -6.254   4e-10 ***  
smokingYes   -0.06777   0.27812  -0.244   0.8075  
obesityYes    0.69531   0.28509   2.439   0.0147 *  
snoringYes    0.87194   0.39757   2.193   0.0283 *  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Smoking, Obesity, Snoring (SOS)

Odds ratios for smoking, obesity and snoring:

```
exp(cbind (OR = coef(analysis.sos), confint(analysis.sos)))
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %
(Intercept)	0.09276726	0.04063914	0.183823
smokingYes	0.93447081	0.53379700	1.594628
obesityYes	2.00432951	1.13345994	3.478922
snoringYes	2.39154432	1.15660384	5.605594

Note that the interval for smoking contains 1; smoking is insignificant.

# Exercises:

## 1) Load the data set surgery as

```
surgery<-read.table("surgery.txt",header=T)
```

The dataset shows the results of a study about Y= whether a patient having surgery with general anesthesia experienced a sore throat on waking up (0=no, 1=yes), as a function of the D= duration of the surgery in minutes; and the T= type of device used to secure the airway (0= laryngeal mask airway, 1= tracheal tube). Fit a logistic regression model using these predictors, interpret parameter estimates, and conduct inference about the effects.

Source: D. Collett, in Encyclopedia of Biostatistics (Wiley, New York 1998), pp.350-358.

## 2) Alternative formulation for frequency data: Access the internal R dataset menarche (proportion of female children that have reached menarche/first menstruation), by typing

```
library(MASS); data(menarche); attach(menarche)
```

Model the matrix `cbind(Menarche, Total-Menarche)` as a function of Age, and make a plot with the data and the fitted logistic regression curve.

# ORDINAL REGRESSION

# Ordinal Regression

- In the lecture on logistic regression, we modeled the probability of a satellite for a Female Horseshoe Crab as

$$\text{logit}(p_i) = \alpha + \beta w_i$$

where  $w$  was the width of the crab.

The logit function is the log of the odds:

$$\text{logit}(p_i) = \log \left( \frac{P(Y_i = 1)}{P(Y_i = 0)} \right)$$

However, this relies on if data are binary. What if there were more response groups than two?

# Ordinal Regression

Let's activate the `ordinal` package, and look at the wine data:

```
library(ordinal)
summary(wine)

  response      rating      temp      contact      bottle      judge
Min.    :12.00  1: 5      cold:36   no  :36     1       : 9    1       : 8
1st Qu.:32.00  2:22      warm:36  yes:36     2       : 9    2       : 8
Median  :46.00  3:26          3       : 9    3       : 9    3       : 8
Mean    :47.22  4:12          4       : 9    4       : 9    4       : 8
3rd Qu.:60.00  5: 7          5       : 9    5       : 9    5       : 8
Max.    :90.00          6       : 9    (Other):18  (Other):24
```

`temp` is the temperature when crushing the grapes, while `contact` indicates contact between juice and skin during the crushing process. The `rating` of the wines have 5 ordinal levels.

# Ordinal Regression

- With the response  $Y=\text{rating}$ , the response is no longer binary, but ordinal. Instead of the success/failure events (satellite/no satellite), we organise the data into the 4 consecutive success events:

$$\begin{aligned} Y \leq 1, Y \leq 2, Y \leq 3, Y \leq 4; \\ Y \leq j, j = 1:4 \end{aligned}$$

None of these events happen if  $Y = 5$ ; so  $Y = 5$  is the 'failure' here. We thus have 4 versions of 'success'; lets model the odds of success:

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \text{temp}_i + \text{contact}_i + \text{temp}_i: \text{contact}_i, j = 1, \dots, 4.$$

Remember that the logit function is the log of the odds. The odds of the event  $\{Y_i \leq j\}$  are

$$\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)} = \frac{P(Y_i \leq j)}{P(Y_i > j)}, j = 1, \dots, 4$$

# Ordinal Regression

In the ordinal package, we can use the `clm` function to do ordinal regression:

```
analysis <- clm(rating ~ temp * contact, data = wine)
drop1(analysis, test="Chisq")
```

Single term deletions

Model:

```
rating ~ temp * contact
          Df      AIC      LRT Pr(>Chi)
<none>           186.83
temp:contact     1 184.98 0.15145    0.6972
```

# Ordinal Regression

No interaction:

```
analysis2 <- update(analysis, ~.-temp:contact)
drop1(analysis2, test="Chisq")
```

Single term deletions

Model:

```
rating ~ temp + contact
```

	Df	AIC	LRT	Pr (>Chi)
<none>		184.98		
temp	1	209.91	26.928	2.112e-07 ***
contact	1	194.03	11.043	0.0008902 ***
---				
Signif. codes:	0	'***'	0.001 '**'	0.01 '*' 0.05 '.' 0.1 ' ' 1

# Ordinal Regression

- Parameter estimates:

```
summary(analysis2)$coefficients
```

	Estimate	Std. Error	z value	Pr (>  z   )
1   2	-1.344383	0.5171020	-2.599842	9.326680e-03
2   3	1.250809	0.4378802	2.856509	4.283277e-03
3   4	3.466887	0.5977604	5.799793	6.639670e-09
4   5	5.006404	0.7309063	6.849584	7.406519e-12
tempwarm	2.503102	0.5286801	4.734625	2.194605e-06
contactyes	1.527798	0.4766226	3.205466	1.348440e-03

- We need to get back to the probabilities of the response groups. We know how to get from the (log) odds to probabilities from the categorical data lecture.

# Ordinal Regression

- Consider the reference group, the situation where `temp="cold"` and `contact="no"`. In this case, the linear predictor for the probabilities  $P(Y \leq j), j = 1, \dots, 4$  is exactly the first four parameter estimates in the table below.

```
summary(analysis2)$coefficients
```

	Estimate	Std. Error	z value	Pr (>  z  )
1 2	-1.344383	0.5171020	-2.599842	9.326680e-03
2 3	1.250809	0.4378802	2.856509	4.283277e-03
3 4	3.466887	0.5977604	5.799793	6.639670e-09
4 5	5.006404	0.7309063	6.849584	7.406519e-12
tempwarm	2.503102	0.5286801	4.734625	2.194605e-06
contactyes	1.527798	0.4766226	3.205466	1.348440e-03

- Lets extract them:

```
my.linear.predictor<- analysis2$alpha
```

# Ordinal Regression

- From (log) odds to probabilities:

```
temp<-exp(my.linear.predictor) / (1+exp(my.linear.predictor))
```

From cumulated probabilities to category probabilities:

```
my.probabilities<-c(temp[1],diff(temp),1-temp[4])  
my.probabilities  
1 | 2            2 | 3            3 | 4            4 | 5            4 | 5  
0.20679013 0.57064970 0.19229094 0.02361882 0.00665041
```

When the temperature is cold and there is no contact, the wine is most often rated 2.

# Ordinal Regression

Cold and no contact:

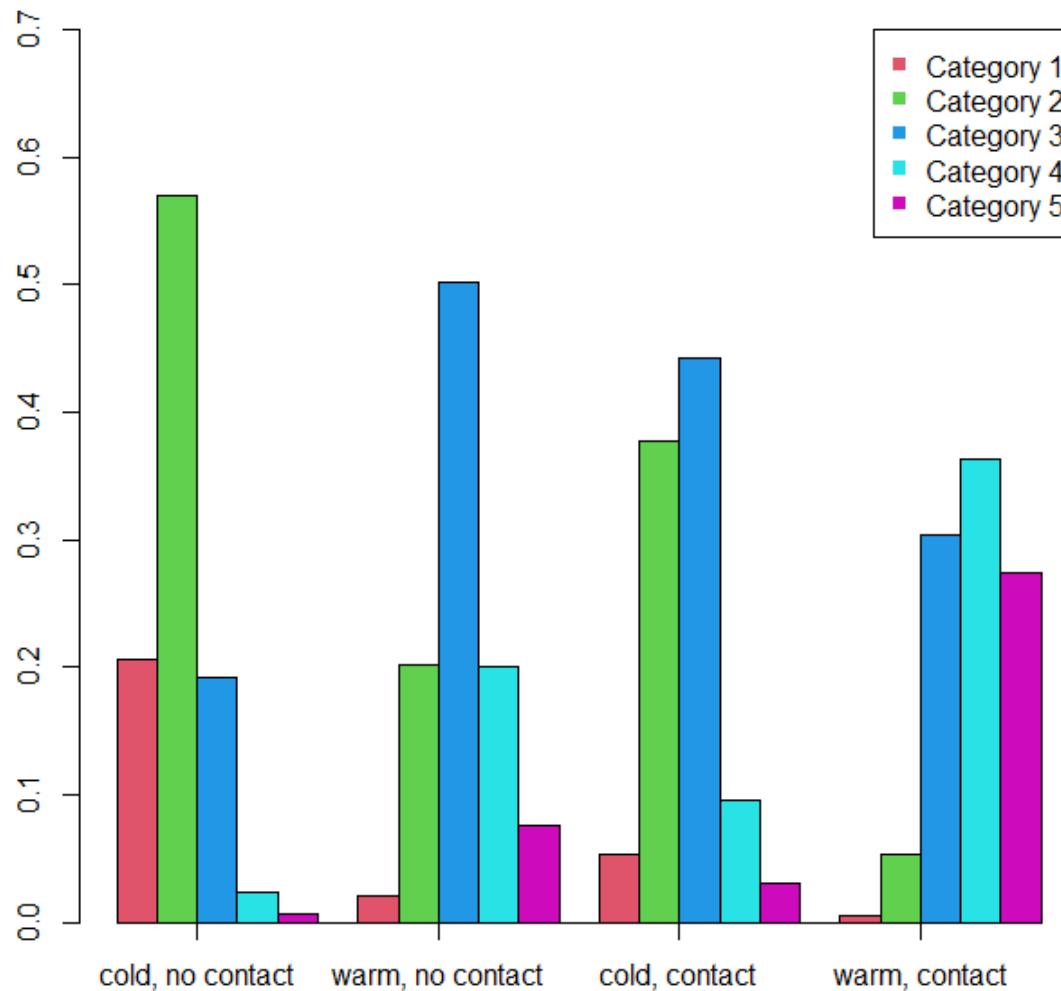
0.20679013 0.57064970 0.19229094 0.02361882 0.00665041

- All estimates in one go:

```
my.predictdata<-expand.grid(temp=c("cold","warm"),contact=c("no","yes"))  
  
cbind(my.predictdata,predict(analysis2,newdata=my.predictdata)$fit)  
temp contact 1 2 3 4 5  
1 cold no 0.206790132 0.57064970 0.1922909 0.02361882 0.00665041  
2 warm no 0.020887709 0.20141572 0.5015755 0.20049402 0.07562701  
3 cold yes 0.053546010 0.37764614 0.4430599 0.09582084 0.02992711  
4 warm yes 0.004608274 0.05380128 0.3042099 0.36359581 0.27378469
```

We recognize the first line as the values we calculated. The wines are generally rated highest when the temperature is warm and there is contact between the juice and skin.

# Ordinal Regression



# Ordinal Regression

## the Proportional Odds assumption and more

- In the formula

$$\text{logit}(P(Y_i \leq j)) = \alpha_j + \text{temp}_i + \text{contact}_i + \text{temp}_i : \text{contact}_i, j = 1, \dots, 4$$

the impact of `temp` and `contact` does not depend on the response level  $j$  at the logit scale.

This assumption is standard, but one may wish to be able to model a changing effect of temperature and contact when the category level changes.

This can be done in R using the `vglm` function from the `VGAM` library. This is referred to as self-study.

General nominal regression can be done in R with the `multinom` function from the `nnet` library. This too is referred to as self-study.

**Confidence intervals on the original scale:** One can use similar methods as in the lecture on logistic regression, to obtain uncertainties about probabilities.

# Solution to exercise about histamine in dogs, course 02935

Elisabeth Wreford Andersen

7 June 2019

## 1 Introduction

In an experiment with 16 dogs the blood histamine concentration was measured at 0, 1, 3, and 5 minutes after injection of morphine or trimethaphane. Before injection the dogs were classified into two groups according to their level of histamine (intact or depleted).

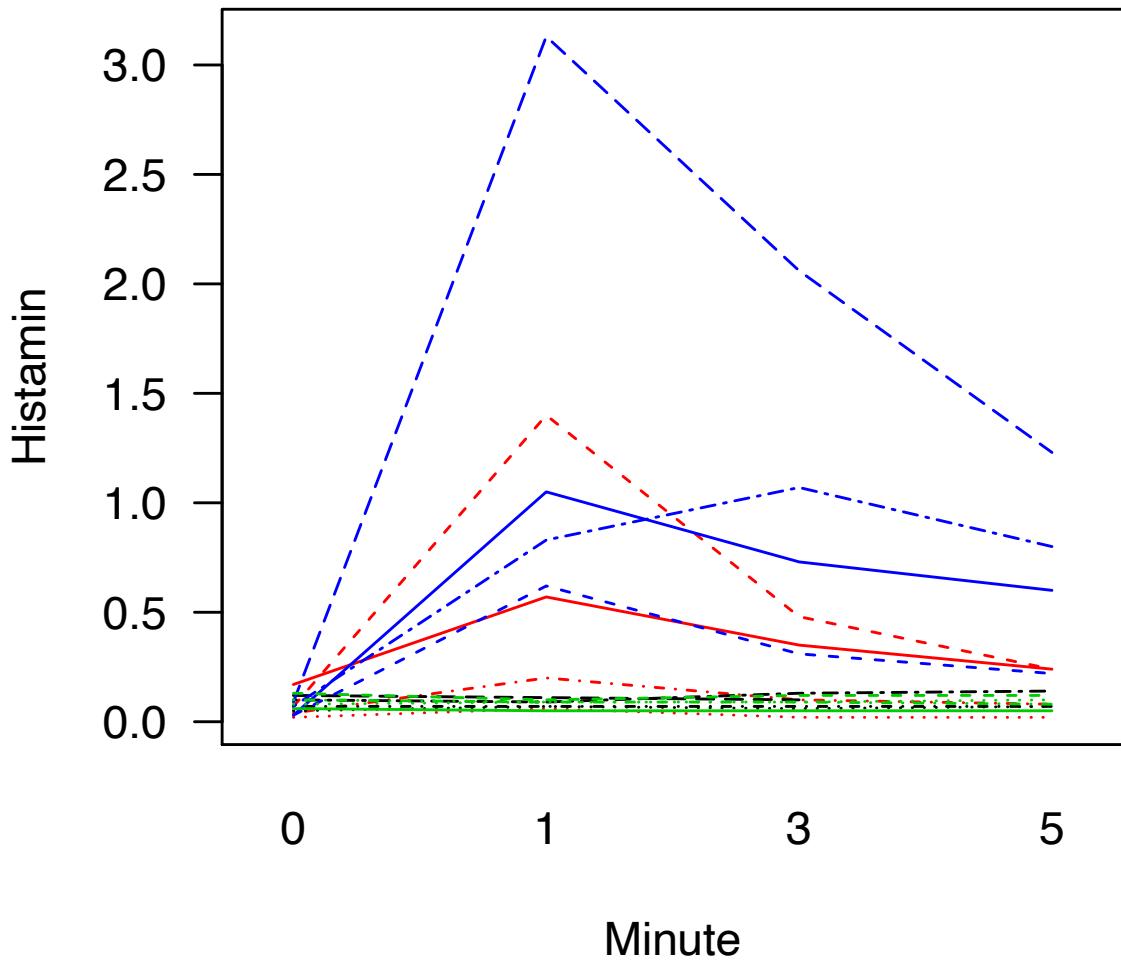
In the code below we will read in the data and set some of the variables to factors. We will define a treatment variable “TRT” with four levels, combining the two drugs and the two levels of histamin before treatment.

```
histamin <- read.table("histamin.txt", header=T, sep=",", dec=".")  
histamin$dog <- factor(histamin$dog)  
histamin$minQ <- histamin$min  
histamin$min <- as.factor(histamin$min)  
histamin$TRT <- as.factor(histamin$treatm:histamin$level)  
summary(histamin)  
  
##      treatm      level      dog      min       hist  
##  morphine:32  deplet:32   1      : 4  0:16   Min.    :0.020  
##  trimetha:32  intact:32   2      : 4  1:16   1st Qu.:0.070  
##                      3      : 4  3:16   Median  :0.100  
##                      4      : 4  5:16   Mean    :0.312  
##                      5      : 4  7:16   3rd Qu.:0.240  
##                      6      : 4  9:16   Max.    :3.130  
##          (Other):40                 NA's    :1  
##      minQ             TRT  
##  Min.    :0.00  morphine:deplet:16  
##  1st Qu.:0.75  morphine:intact:16
```

```
## Median :2.00  trimetha:deplet:16
## Mean    :2.25  trimetha:intact:16
## 3rd Qu.:3.50
## Max.   :5.00
##
```

In Figure 1 we will plot a line for the level of histamin in each dog during the study.

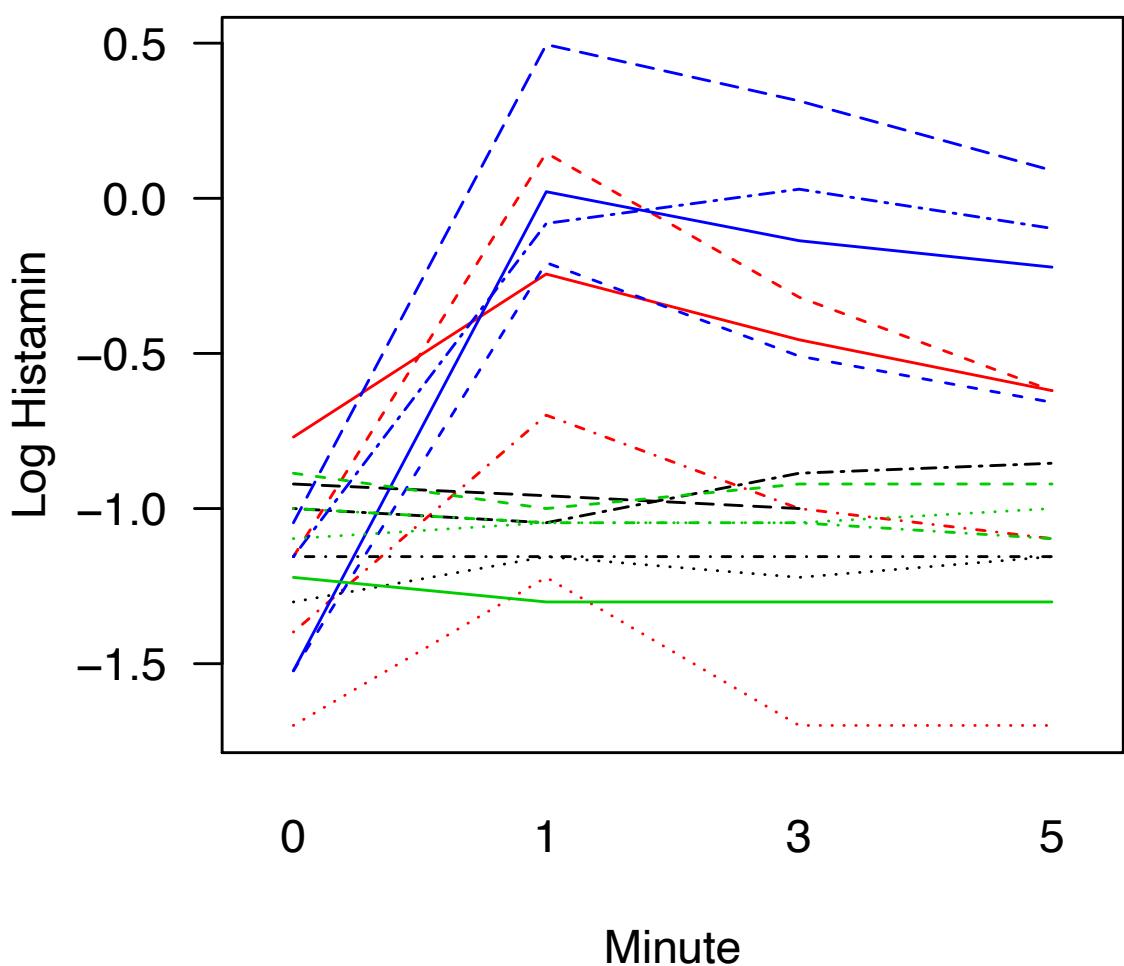
```
interaction.plot(histamin$minQ,
                 histamin$dog, histamin$hist, xlab="Minute",
                 ylab="Histamin", legend=F,
                 col= histamin$TRT, las=1)
```



Figur 1: Individual levels of histamin

Looking at Figure 1 we see that the combination trimetha:intact seems to be different. Some of the lines are very close to zero. Perhaps it is easier to look at log transformed data (Figure 2).

```
histamin$loghist <- log10(histamin$hist)
#A line for each dog
interaction.plot(histamin$minQ,
                 histamin$dog, histamin$loghist, xlab="Minute",
                 ylab="Log Histamin", legend=F,
                 col= histamin$TRT, las=1)
```



Figur 2: Individual levels of log-histamin

The patterns in Figure 2 look similar so we could perhaps look at the average in each treatment group (Figure 3).

```

Mean_data <- aggregate(histamin$loghist,
                       by=list(histamin$min, histamin$TRT),
                       mean)

#Have a look
Mean_data

##      Group.1      Group.2      x
## 1      0 morphine:deplet -1.094188
## 2      1 morphine:deplet -1.078542
## 3      3 morphine:deplet -1.065702
## 4      5 morphine:deplet      NA
## 5      0 morphine:intact -1.255341
## 6      1 morphine:intact -0.504704
## 7      3 morphine:intact -0.868415
## 8      5 morphine:intact -1.008864
## 9      0 trimetha:deplet -1.051204
## 10     1 trimetha:deplet -1.098136
## 11     3 trimetha:deplet -1.078341
## 12     5 trimetha:deplet -1.079690
## 13     0 trimetha:intact -1.311604
## 14     1 trimetha:intact  0.057051
## 15     3 trimetha:intact -0.075516
## 16     5 trimetha:intact -0.221608

#assign names
names(Mean_data) <- c("minute", "TRT", "MeanLogHist")
#Have a look
Mean_data

##      minute      TRT MeanLogHist
## 1      0 morphine:deplet -1.094188
## 2      1 morphine:deplet -1.078542
## 3      3 morphine:deplet -1.065702
## 4      5 morphine:deplet      NA
## 5      0 morphine:intact -1.255341
## 6      1 morphine:intact -0.504704
## 7      3 morphine:intact -0.868415
## 8      5 morphine:intact -1.008864
## 9      0 trimetha:deplet -1.051204
## 10     1 trimetha:deplet -1.098136

```

```

## 11      3 trimetha:deplet -1.078341
## 12      5 trimetha:deplet -1.079690
## 13      0 trimetha:intact -1.311604
## 14      1 trimetha:intact  0.057051
## 15      3 trimetha:intact -0.075516
## 16      5 trimetha:intact -0.221608

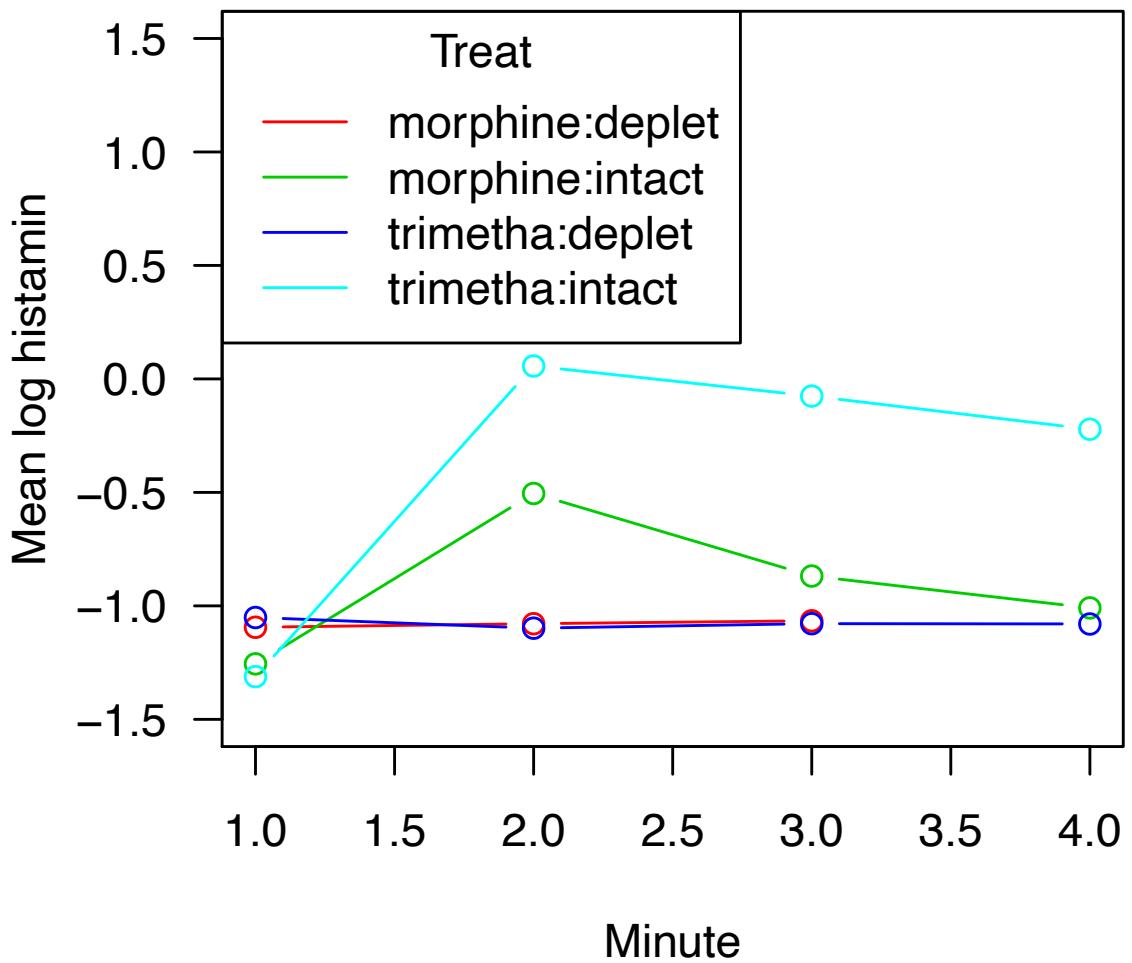
#Plot the means

Grp1<-subset(Mean_data,TRT=="morphine:deplet")
Grp2<-subset(Mean_data,TRT=="morphine:intact")
Grp3<-subset(Mean_data,TRT=="trimetha:deplet")
Grp4<-subset(Mean_data,TRT=="trimetha:intact")

plot(as.numeric(Grp1$minute), Grp1$MeanLogHist, type = "b",
     lty=1, xlab = "Minute", las = 1, col= 2,
     ylab = "Mean log histamin", ylim=c(-1.5,1.5))

lines(as.numeric(Grp2$minute), Grp2$MeanLogHist, type = "b", col = 3, lty = 1)
lines(as.numeric(Grp3$minute), Grp3$MeanLogHist, type = "b", col = 4, lty = 1)
lines(as.numeric(Grp4$minute), Grp4$MeanLogHist, type = "b", col = 5, lty = 1)
legend("topleft",
       legend = c("morphine:deplet",
                  "morphine:intact",
                  "trimetha:deplet",
                  "trimetha:intact"),
       title = "Treat",
       lty = c(1,1,1,1),
       col= 2:5)

```



Figur 3: Average log-histamin in treatment groups

From this plot we can see that the histamin levels are unchanged in the groups with depleted histamine levels. For both groups with intact levels of histamine the histamine concentration increases in both groups after injection. There may be a tendency of a higher rise in the trimethaphane group than in the morphine group.

## 2 Analyse these data using one or more of the “simple” methods

As we have seen in the plots we have an example of repeated measurements, as we have four measurements from each dog (one missing however). One way of analyzing these data could be to chose a summary measure representing the treatment effect for each dog. This means that we again just have one observation per dog and we can use our standard methods to analyse the data.

### 2.1 Analysis 1

In the first analysis we could look at the change from 0 to 1 minute (the change from observation 1 to 2).

```
myfun <- function(x){  
  x[2]-x[1]  
}  
  
Data1 <- aggregate(histamin$loghist,  
                    by=list(histamin$dog, histamin$TRT), myfun)  
  
names(Data1) <- c("dog", "TRT", "diff10")  
  
#have a look  
Data1  
  
##      dog          TRT    diff10  
## 1      5 morphine:deplet -0.045757  
## 2      6 morphine:deplet -0.037789  
## 3      7 morphine:deplet  0.000000  
## 4      8 morphine:deplet  0.146128  
## 5      1 morphine:intact  0.698970  
## 6      2 morphine:intact  0.477121  
## 7      3 morphine:intact  1.301030  
## 8      4 morphine:intact  0.525426  
## 9     13 trimetha:deplet -0.045757  
## 10    14 trimetha:deplet  0.051153  
## 11    15 trimetha:deplet -0.113943  
## 12    16 trimetha:deplet -0.079181  
## 13     9 trimetha:intact  1.315270  
## 14    10 trimetha:intact  1.544068
```

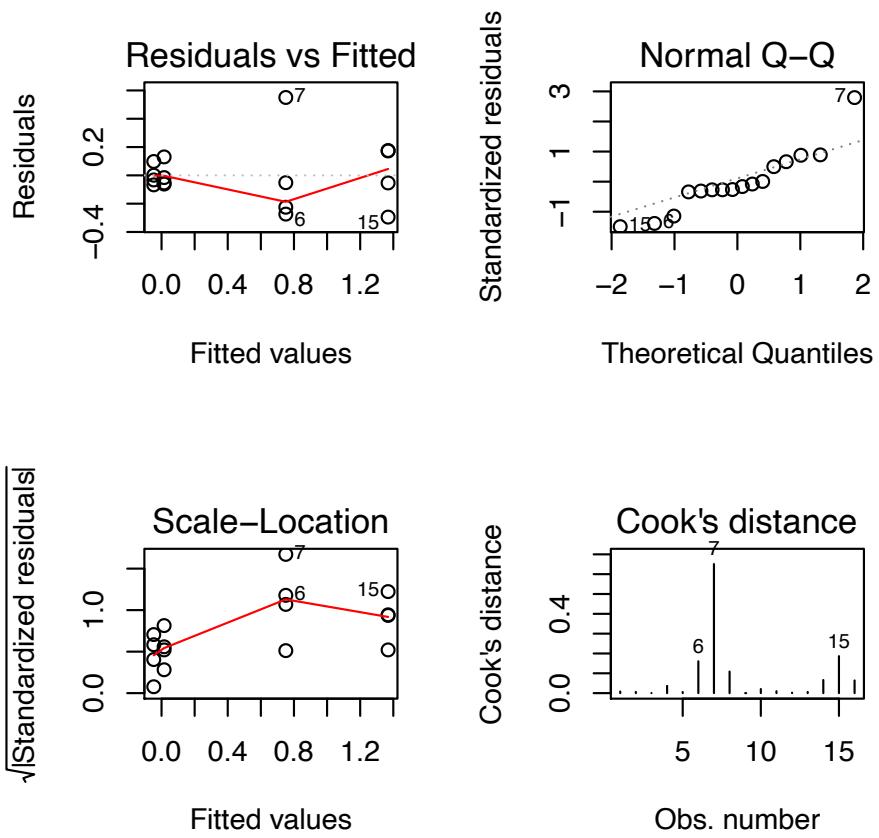
```

## 15 11 trimetha:intact 1.073980
## 16 12 trimetha:intact 1.541302

# Analyse the change from 0 to 1
reg1 <- lm(diff10 ~ TRT, data = Data1)

#Model check
par(mfrow=c(2,2))
plot(reg1, which=1:4)
par(mfrow=c(1,1))

```



Figur 4: Plots for checking simple model 1

In this analysis we are using the change from observation 1 to 2 as our outcome. This results in a one-way analysis of variance.

The model check could perhaps have been better (mainly due to observation 7) but we only have 16 observations so it is difficult to be too sure. The results of the analysis are seen below.

```

summary(reg1)

##
## Call:
## lm(formula = diff10 ~ TRT, data = Data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -0.2947 -0.0628 -0.0420  0.1062  0.5504
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)              0.0156    0.1137   0.14  0.89283
## TRTmorphine:intact      0.7350    0.1608   4.57  0.00064 ***
## TRTtrimetha:deplet     -0.0626    0.1608  -0.39  0.70394
## TRTtrimetha:intact      1.3530    0.1608   8.42  2.2e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.227 on 12 degrees of freedom
## Multiple R-squared:  0.897, Adjusted R-squared:  0.871
## F-statistic: 34.8 on 3 and 12 DF,  p-value: 3.37e-06

confint(reg1)

##
##                               2.5 % 97.5 %
## (Intercept)             -0.23207 0.26336
## TRTmorphine:intact      0.38467 1.08531
## TRTtrimetha:deplet     -0.41290 0.28774
## TRTtrimetha:intact      1.00269 1.70333

anova(reg1)

## Analysis of Variance Table
##
## Response: diff10
##           Df Sum Sq Mean Sq F value Pr(>F)
## TRT        3   5.40   1.799   34.8 3.4e-06 ***
## Residuals 12   0.62   0.052
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

From this analysis (the anova table) we see a significant treatment effect ( $p < 0.0001$ ). The two groups with intact levels have significantly higher increases compared to the reference group (morphine:deplet). Perhaps we would also like to compare the two intact groups directly. We can follow our one-way analysis up by pairwise testing.

```
pairwise.t.test(Data1$diff10, Data1$TRT, p.adj = "none")

##
##  Pairwise comparisons using t tests with pooled SD
##
##  data:  Data1$diff10 and Data1$TRT
##
##          morphine:deplet morphine:intact trimetha:deplet
## morphine:intact  0.00064      -         -
## trimetha:deplet 0.70394      0.00033     -
## trimetha:intact 2.2e-06      0.00234    1.4e-06
##
## P value adjustment method: none
```

Looking at the pairwise tests we can see also a significant difference between trimetha:intact and morphine:intact  $p=0.00234$

## 2.2 Analysis 2

Another possibility for a summary statistic could be to look at the maximum value.

```
Data2 <- aggregate(histamin$loghist,
                   by=list(histamin$dog, histamin$TRT), max)

names(Data2) <- c("dog", "TRT", "max")

#have a look
Data2

##   dog           TRT      max
## 1  5 morphine:deplet -0.853872
## 2  6 morphine:deplet      NA
## 3  7 morphine:deplet -1.154902
## 4  8 morphine:deplet -1.154902
## 5  1 morphine:intact -0.698970
## 6  2 morphine:intact -1.221849
```

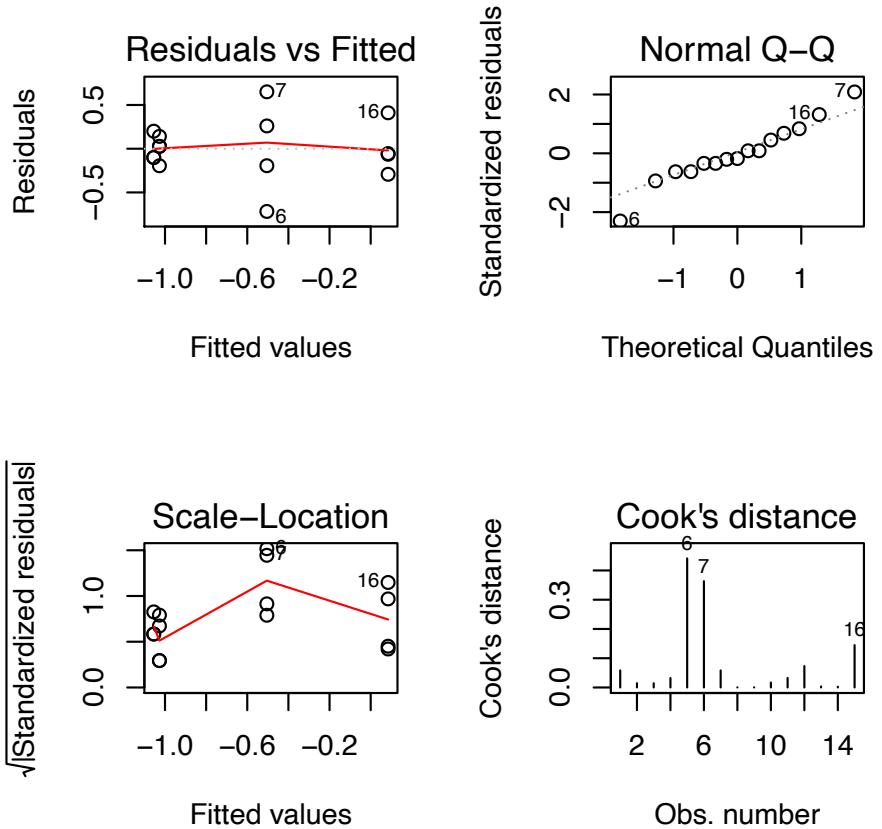
```

## 7    3 morphine:intact  0.146128
## 8    4 morphine:intact -0.244125
## 9   13 trimetha:deplet -1.000000
## 10   14 trimetha:deplet -1.000000
## 11   15 trimetha:deplet -0.886057
## 12   16 trimetha:deplet -1.221849
## 13    9 trimetha:intact -0.207608
## 14   10 trimetha:intact  0.021189
## 15   11 trimetha:intact  0.029384
## 16   12 trimetha:intact  0.495544

# Analyse the max
reg2 <- lm(max ~ TRT, data = Data2)

#Model check
par(mfrow=c(2,2))
plot(reg2, which=1:4)
par(mfrow=c(1,1))

```



Figur 5: Plots for checking simple model 2

In the following we have the results of the analysis looking at the maximum value for each dog.

```
summary(reg2)

##
## Call:
## lm(formula = max ~ TRT, data = Data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -0.7171 -0.1473 -0.0552  0.1708  0.6508 
##
## Coefficients:
```

```

##                               Estimate Std. Error t value Pr(>|t|) 
## (Intercept)           -1.0546    0.2078   -5.08  0.00036 ***
## TRTmorphine:intact   0.5499    0.2749    2.00  0.07074 .
## TRTtrimetha:deplet   0.0276    0.2749    0.10  0.92187 
## TRTtrimetha:intact   1.1392    0.2749    4.14  0.00163 ** 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
## 
## Residual standard error: 0.36 on 11 degrees of freedom 
## (1 observation deleted due to missingness) 
## Multiple R-squared:  0.696, Adjusted R-squared:  0.613 
## F-statistic: 8.39 on 3 and 11 DF,  p-value: 0.00349 

confint(reg2)

##                  2.5 %    97.5 % 
## (Intercept)     -1.511857 -0.59726 
## TRTmorphine:intact -0.055094  1.15480 
## TRTtrimetha:deplet -0.577366  0.63253 
## TRTtrimetha:intact  0.534238  1.74413 

anova(reg2)

## Analysis of Variance Table 
## 
## Response: max 
##              Df Sum Sq Mean Sq F value Pr(>F) 
## TRT          3   3.26   1.09    8.39 0.0035 ** 
## Residuals 11   1.42   0.13 
## --- 
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

```

From this analysis (the anova table) we see a significant treatment effect ( $p = 0.0035$ ). The trimetha:intact group has a significantly higher maximum compared to the reference group (morphine:deplet). Again we can follow our one-way analysis up by pairwise testing.

```

pairwise.t.test(Data2$max, Data2$TRT, p.adj = "none") 

## 
## Pairwise comparisons using t tests with pooled SD 
## 
## data: Data2$max and Data2$TRT 

```

```

##          morphine:deplet morphine:intact trimetha:deplet
## morphine:intact 0.0707           -           -
## trimetha:deplet 0.9219           0.0647      -
## trimetha:intact 0.0016           0.0409     0.0011
##
## P value adjustment method: none

```

Looking at the pairwise tests we can see also a significant difference between trimetha:intact and morphine:intact p=0.0409

## 2.3 Analysis 3

We could also look at the change from 0 to 5 minutes (observations 1 and 4), the change from baseline.

```

myfun2 <- function(x){
  x[4]-x[1]
}

Data3 <- aggregate(histamin$loghist,
  by=list(histamin$dog, histamin$TRT), myfun2)

names(Data3) <- c("dog", "TRT", "diff50")

#have a look
Data3

##   dog          TRT    diff50
## 1  5 morphine:deplet  0.146128
## 2  6 morphine:deplet       NA
## 3  7 morphine:deplet  0.000000
## 4  8 morphine:deplet  0.146128
## 5  1 morphine:intact  0.301030
## 6  2 morphine:intact  0.000000
## 7  3 morphine:intact  0.535113
## 8  4 morphine:intact  0.149762
## 9 13 trimetha:deplet -0.096910
## 10 14 trimetha:deplet  0.096910
## 11 15 trimetha:deplet -0.034762
## 12 16 trimetha:deplet -0.079181
## 13  9 trimetha:intact  0.865301
## 14 10 trimetha:intact  1.301030

```

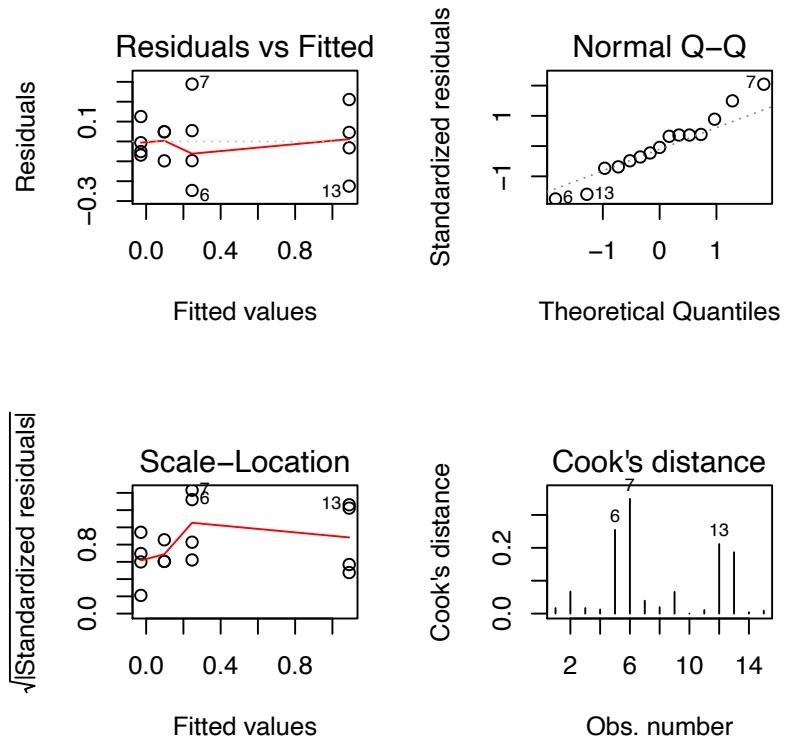
```

## 15 11 trimetha:intact 1.057992
## 16 12 trimetha:intact 1.135663

# Analyse the change from 0 to 5
reg3 <- lm(diff50 ~ TRT, data = Data3)

#Model check
par(mfrow=c(2,2))
plot(reg3, which=1:4)
par(mfrow=c(1,1))

```



Figur 6: Plots for checking initial model

In the following we have the results of the analysis looking at the change from baseline for each dog.

```

summary(reg3)

##
## Call:
## lm(formula = diff50 ~ TRT, data = Data3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.24648 -0.08257 -0.00628  0.05163  0.28864
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.0974    0.0940   1.04    0.32
## TRTmorphine:intact  0.1491    0.1244   1.20    0.26
## TRTtrimetha:deplet -0.1259    0.1244  -1.01    0.33
## TRTtrimetha:intact  0.9926    0.1244   7.98  6.7e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.163 on 11 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.911, Adjusted R-squared:  0.887
## F-statistic: 37.6 on 3 and 11 DF,  p-value: 4.48e-06

confint(reg3)

##                   2.5 % 97.5 %
## (Intercept) -0.10954 0.30438
## TRTmorphine:intact -0.12472 0.42284
## TRTtrimetha:deplet -0.39969 0.14788
## TRTtrimetha:intact  0.71880 1.26636

anova(reg3)

## Analysis of Variance Table
##
## Response: diff50
##             Df Sum Sq Mean Sq F value Pr(>F)
## TRT          3  2.993  0.998   37.6 4.5e-06 ***
## Residuals 11  0.292  0.027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

From this analysis (the anova table) we see a significant treatment effect ( $p < 0.0001$ ). The trimetha:intact group has a significantly higher change from baseline compared to the reference group (morphine:deplet). Again we can follow our one-way analysis up by pairwise testing.

```
pairwise.t.test(Data3$diff50, Data3$TRT, p.adj = "none")

##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Data3$diff50 and Data3$TRT
##
##          morphine:deplet morphine:intact trimetha:deplet
## morphine:intact  0.256          -          -
## trimetha:deplet 0.333          0.036          -
## trimetha:intact 6.7e-06        1.5e-05      9.9e-07
##
## P value adjustment method: none
```

Looking at the pairwise tests we can see that it is mainly trimetha:intact group, which is different from the other three. There is also a significant difference between trimetha:intact and morphine:intact  $p < 0.0001$ .

All in all we get more or less the same conclusions which ever analysis we have chosen. We see that the two depleted groups are very similar and that the trimetha:intact group changes most and significantly more than the morphine:intact group. It depends on the summary measure whether we can see differences between morphine:intact and the depleted groups.

### 3 Using a Random Effects Model

Instead of using the summary measures we could also apply a random effects model making use of all the observations at once. Below we have a simple random effects model with a random effect of dog and an interaction between minute and treatment allowing the differences between the treatment groups to vary at each minute (Often the groups will be close at the beginning).

```
library(nlme)
model1 <- lme(loghist ~ min + TRT + min:TRT,
               random = ~1 | dog, data = histamin, na.action = na.omit)

anova(model1)
```

```

##          numDF denDF F-value p-value
## (Intercept)     1     35 121.793 <.0001
## min            3     35  59.857 <.0001
## TRT            3     12   4.315  0.0278
## min:TRT        9     35  28.724 <.0001

```

From the results we can see that there is a significant interaction between min:TRT ( $p < .0001$ ). To find out where the differences lie one can look at a table of the fixed effects estimates.

```

summary(model1)$tTable

##                               Value Std.Error DF t-value p-value
## (Intercept)      -1.094188  0.163753 35 -6.68196 9.8693e-08
## min1             0.015645  0.082070 35  0.19064 8.4991e-01
## min3             0.028486  0.082070 35  0.34709 7.3060e-01
## min5             0.077696  0.090638 35  0.85721 3.9716e-01
## TRTmorphine:intact -0.161153  0.231581 12 -0.69588 4.9976e-01
## TRTtrimetha:deplet  0.042984  0.231581 12  0.18561 8.5585e-01
## TRTtrimetha:intact -0.217417  0.231581 12 -0.93884 3.6633e-01
## min1:TRTmorphine:intact  0.734991  0.116065 35  6.33259 2.8225e-07
## min3:TRTmorphine:intact  0.358440  0.116065 35  3.08827 3.9273e-03
## min5:TRTmorphine:intact  0.168780  0.122273 35  1.38035 1.7623e-01
## min1:TRTtrimetha:deplet -0.062578  0.116065 35 -0.53916 5.9319e-01
## min3:TRTtrimetha:deplet -0.055623  0.116065 35 -0.47924 6.3475e-01
## min5:TRTtrimetha:deplet -0.106182  0.122273 35 -0.86840 3.9109e-01
## min1:TRTtrimetha:intact  1.353010  0.116065 35 11.65735 1.3246e-13
## min3:TRTtrimetha:intact  1.207602  0.116065 35 10.40454 2.9731e-12
## min5:TRTtrimetha:intact  1.012300  0.122273 35  8.27900 9.2733e-10

```

The intercept is the expected log-histamine level for a dog in morphine:deplet at minute 0. At minute 1 the expected log-histamine for trimetha:intact would be :  $-1.094188 + 0.015645 - 0.217417 + 1.353010 = 0.05705$ .

We also had a random effect and the estimated variances are seen below.

```

#The estimates random effects
VarCorr(model1)

## dog = pdLogChol(1)
##           Variance StdDev
## (Intercept) 0.093789 0.30625
## Residual    0.013471 0.11606

```

```
#intra class correlation  
0.09378856/(0.09378856+0.01347106)  
  
## [1] 0.87441
```

The intraclass correlation describes how strongly measurements from the same dog resemble each other. Here we have a quite high value and we could also see in Figure 2 that even though the dogs have similar patterns then some dogs generally have high levels and some low.

### 3.1 Model check

As for the linear model we still have assumptions underlying the model that should be checked. First we will check the assumptions concerning the residual error  $\epsilon_i \sim N(0, \sigma^2)$ . This is like what we did for the linear models but we have to construct more of the residual plots ourselves.

```
#residual plot  
plot(model1)  
  
#qqplot for residuals  
qqnorm(residuals(model1))  
qqline(residuals(model1))
```

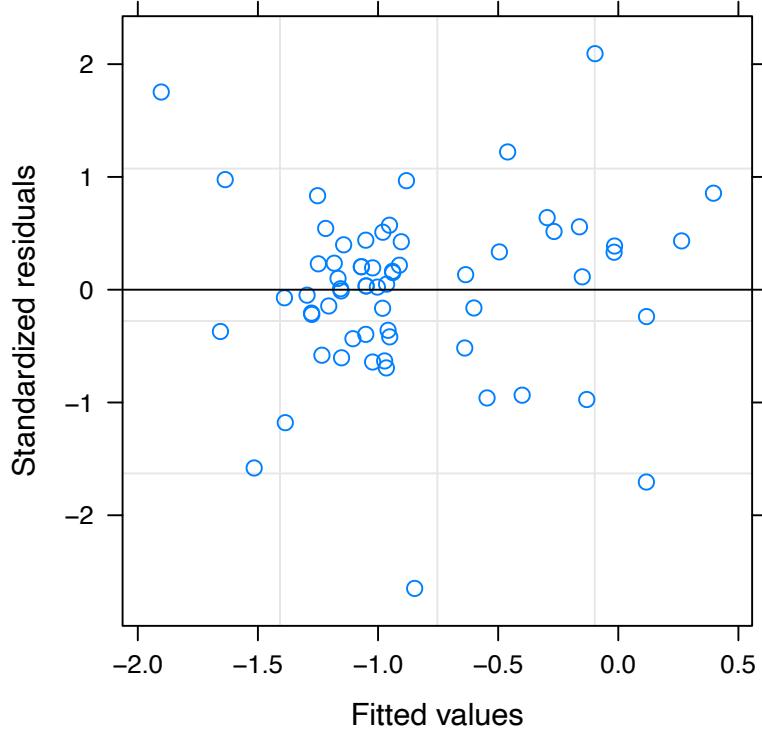
From Figure 7 we see in the top plot that it looks like a random scatter of dots, so the variance homogeneity looks ok and the qq plot looks like a straight line so the normality of residuals is ok.

In Figure 8 we are checking the normal distribution of the random effect.

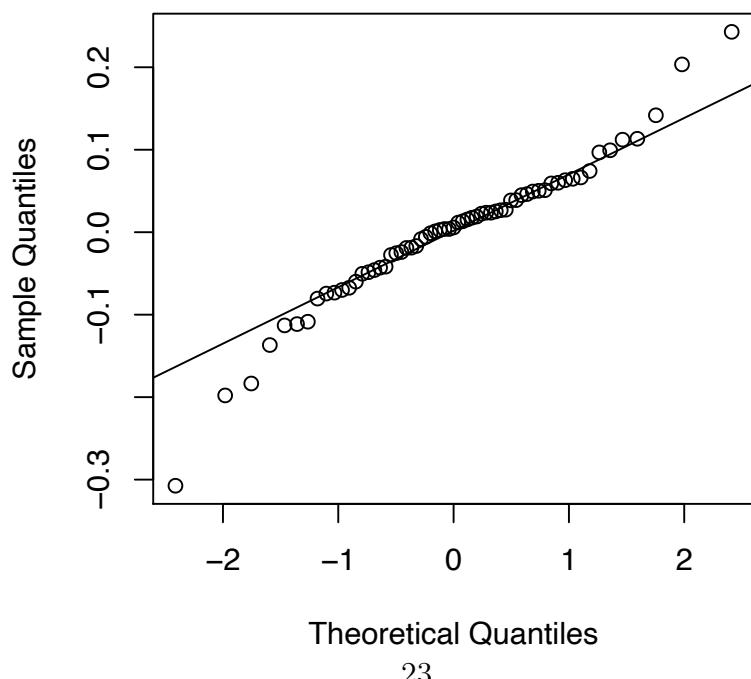
```
#qqplot for random effect
qqnorm(ranef(model1)$"(Intercept)")
qqline(ranef(model1)$"(Intercept)")
```

This approach will only make sense if the number of levels for a factor is not too small (here we have 16 dogs).

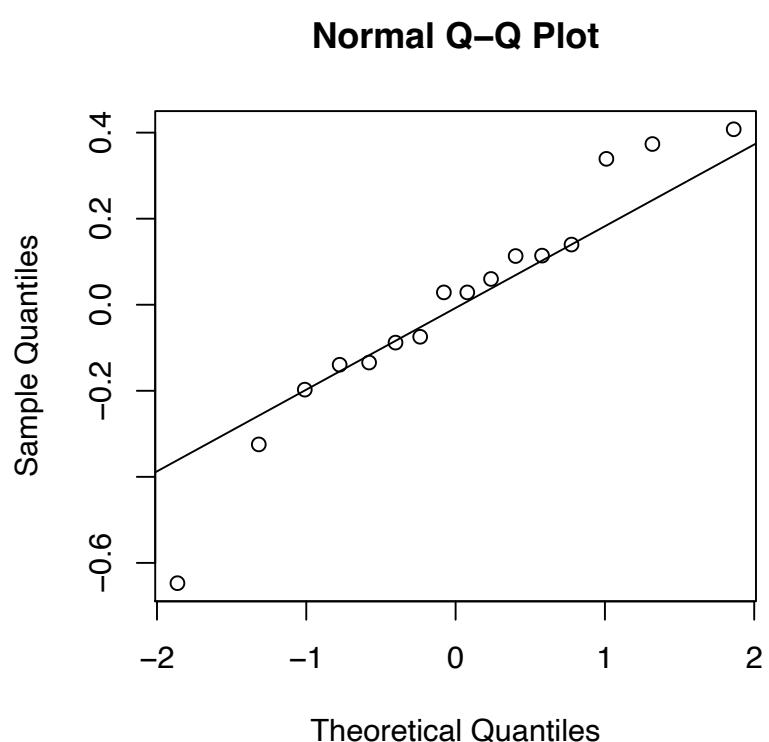
In this example the assumptions underlying the model seem ok.



**Normal Q–Q Plot**



Figur 7: Residual plots for checking residual error



Figur 8: QQ plot for random effect

# Repeated measures, simple methods

Anders Stockmarr

Course developers: Anders Stockmarr, Elisabeth Wreford Andersen,  
Per Bruun Brockhoff

DTU Department of Applied Mathematics and Computer Science  
Section for Statistics and Data Analysis  
Technical University of Denmark  
[anst@dtu.dk](mailto:anst@dtu.dk)

January 10th 2025

# Plan for this week

Monday Statistical inference, and the t-test

Tuesday Simple and Multiple regression

Wednesday ANOVA, ANCOVA, and linear models

Thursday Categorical data, Writing statistical reports,  
Logistic regression

Friday Repeated measurements, Principal Component  
Analysis

# Overview of this module

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a summary statistic
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

# Aim of this module

- Present simple methods for dealing with repeated measurements - and a few more advanced applications
- Easy to use, and a lot better than pretending to have independent observations
- Useful even after more advanced models are presented
- See how to specify these models in R

# Overview

## 1 The repeated measurements setup

- Aspirin Example
- Activity of rats

## 2 Separate analysis for each time-point

- Example: rats data

## 3 Analysis of a summary statistic

- Example: rats data

## 4 Random effects model - simple version

- Example: rats data

## 5 Random effects model - advanced version

## 6 Pros and cons of simple approaches

# The repeated measurements setup

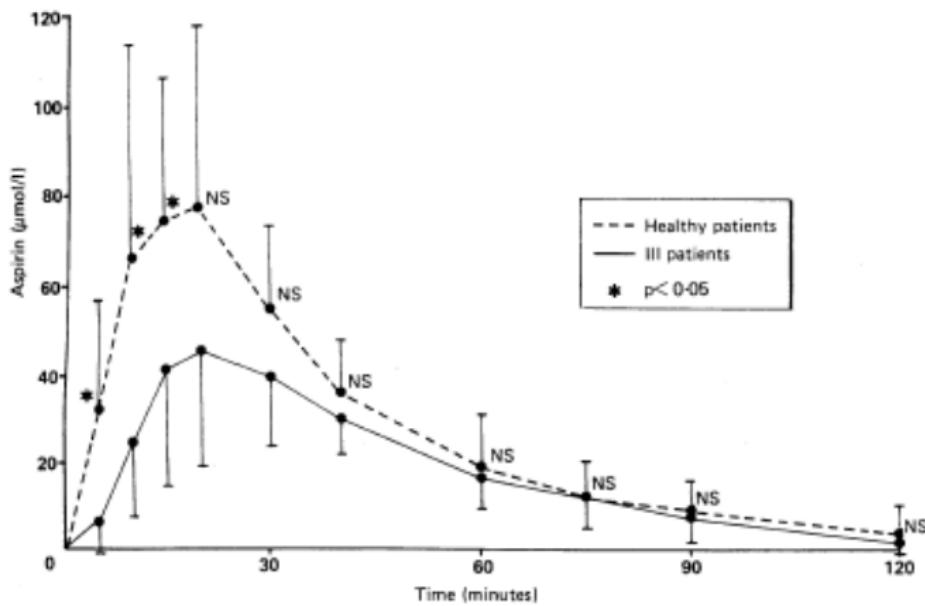
- Several "individuals"
- Several measurements on each individual
- Two measurements on the same individual might be **correlated**
- Might even be highly correlated if "**close**" and less correlated if "**far apart**"

# The repeated measurements setup

- Several "individuals"
- Several measurements on each individual
- Two measurements on the same individual might be **correlated**
- Might even be highly correlated if "**close**" and less correlated if "**far apart**"
- **Typical example** (Matthews et al 1990):
  - Two groups of patients (**ill/healthy**)
  - Treated with aspirin
  - The aspirin concentration is measured in the blood 10 times during 2 hours

*To pretend all observations are independent can lead to wrong conclusions*

# Traditional figure of means at each time point



# What is the aim of the study?

It is always important to consider the aim of the study!

① To describe pattern over time

- Do we have a change over time?
- Linear or curve?
- Is the pattern the same for all groups?

② Can we see a difference between groups?

- Is the difference the same at all time points?
- Is the difference in the levels?
- Is the difference in the trend?

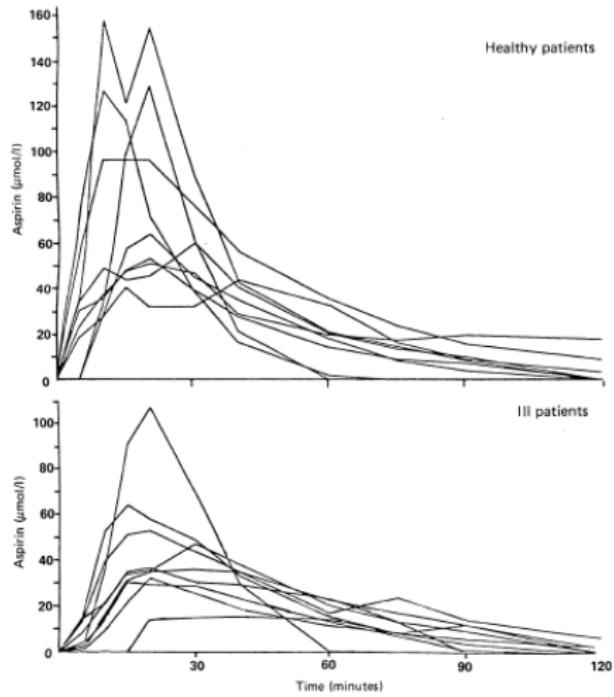
# Problems using the traditional presentation of the data

- Comparing the groups at each time point.
  - Not efficient
  - Tests are not independent, carried out on the same individuals
  - Difficult to interpret
- The pattern over time
  - This can only be studied if we use that each individual has several observations

## Take care when using average curves

- Important structures can be overlooked if you start with the average curves.
- You cannot see the variation over time.
- Always make a plot of the individual observations.
- Only use the average if the patterns are similar, i.e. we only see shifts up and down.

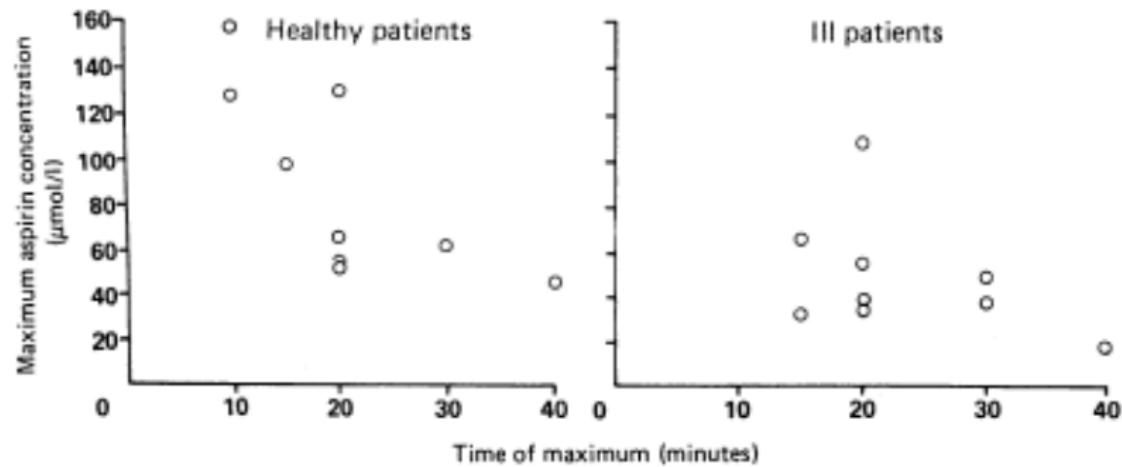
# Individual curves



# Analysis of summary statistics

- Choose a single measure to **summarize** the individual curves
- This again reduces the data set to **independent** observations
- Popular choices of summary measures - "relevant" feature extraction:
  - Total increase (last point minus first point)
  - Area under curve (AUC)
  - Maximum or minimum point
  - Average over time
  - Slope in regression with time
  - (or higher order polynomial coefficients)
- Good method with few and easily checked assumptions
- Information may be lost
- Important to choose a **good summary measure**

# Summary Statistics



# Analysis of aspirin study

TABLE I—*Analysis of data from aspirin study*

	Area under curve	Maximum concentration
<i>Healthy patients (n = 9)</i>		
Arithmetic mean (SD) ( $\mu\text{mol/l}$ )	26·5 (8·8)	86·0 (41·5)
Geometric mean ( $\mu\text{mol/l}$ )	25·4	77·8
<i>Ill patients (n = 9)</i>		
Arithmetic mean (SD) ( $\mu\text{mol/l}$ )	17·5 (5·0)	46·7 (26·3)
Geometric mean ( $\mu\text{mol/l}$ )	16·8	41·2
Ratio of geometric means	1·52	1·89
95% Confidence interval	1·11 to 2·08	1·14 to 3·13
t Test	2·83 (df = 16)	2·66 (df = 16)
p Value	0·01	0·02

## Example: Activity of rats

Summary of experiment:

- 3 treatments: 1, 2, 3 (concentrations)
- 10 cages per treatment
- 10 months
- The response is activity ( $\log(\text{count})$ ) of intersections of light beam during 57 hours)

# The rats data

```
rats <- read.csv("rats.txt")  
  
# make treatment and cage factors  
rats$treatm <- factor(rats$treatm)  
rats$cage <- factor(rats$cage)  
  
# make two versions of the time variable  
# - one quantitative and one factor  
rats$monthQ <- rats$month  
rats$month <- factor(rats$month)  
  
summary(rats)  
str(rats)
```

In this example we have repeated measurements from each cage.

# Data in two formats, long

This is the format we want. One response (here lnc) in each line

```
##   treatm cage month    lnc
## 1      1     1     1 9.9323
## 2      1     1     2 9.6447
## 3      1     1     3 9.7628
## 4      1     1     4 9.6014
## 5      1     1     5 9.3227
## 6      1     1     6 9.2463
## 7      1     1     7 9.0739
## 8      1     1     8 9.2077
## 9      1     1     9 9.1670
## 10     1     1    10 8.8319
```

# Data in two formats, wide

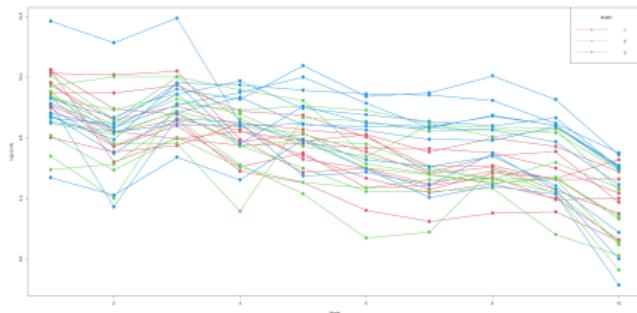
One line for each cage.

```
##      treatm cage   lnc.1   lnc.2   lnc.3   lnc.4   lnc.5   lnc.6   lnc.7   lnc.8
## 1          1     1 9.9323 9.6447 9.7628 9.6014 9.3227 9.2463 9.0739 9.2077
## 11         1     3 10.0547 9.7384 9.6928 9.4676 9.5297 9.3837 9.2052 9.2582
## 21         1     5 9.7448 9.4278 9.5995 9.2614 9.3744 9.0857 9.1068 9.1732
## 31         1     7 9.8660 9.8687 9.9329 9.7183 9.6861 9.5047 9.2622 9.2699
## 41         1     9 9.9552 9.5536 9.4930 9.4369 9.4520 9.4147 9.4107 9.3751
##      lnc.9 lnc.10
## 1  9.1670 8.8319
## 11 9.0681 8.8753
## 21 8.9907 8.9995
## 31 9.1555 9.1584
## 41 9.2475 9.0435
```

Not usually the format wanted for analysis but adapted to plots and group averages.

# The rats data - plotting the individual patterns

```
plot(rats$monthQ,rats$lnc,xlab="Months",
      ylab="log(count)",type ="n")
for(i in 1:dim(rats2)[1]){
  lines(1:10,rats2[i,-(1:2)],type="b",col=rats2$treatm[i]+1,
        pch=16,lwd=2,cex=2)
}
legend("topright",legend=1:3,col=2:4,pch=16,lty=2,title="treatm")
```

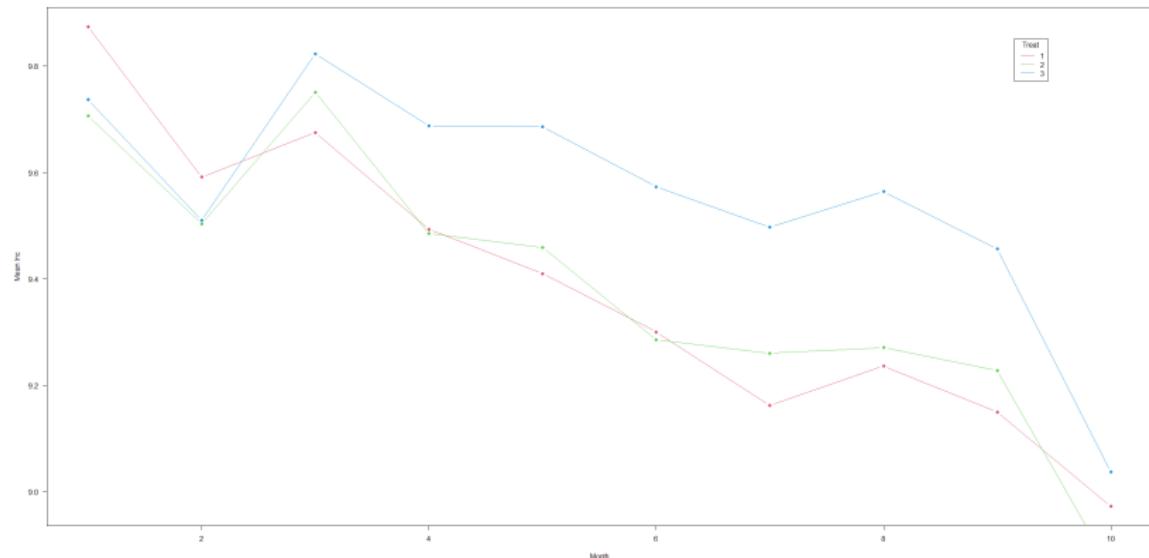


# The rats data - plotting the average patterns

```
Mean_data <- aggregate(rats$lnc, by = list(rats$month, rats$treatm)
                         mean)
names(Mean_data) <- c("month", "treatm", "Meanlnc" )
#Plot the means
Grp1<-subset(Mean_data,treatm==1)
Grp2<-subset(Mean_data,treatm==2)
Grp3<-subset(Mean_data,treatm==3)

plot(as.numeric(Grp1$month), Grp1$Meanlnc, type = "b", pch=16,
     xlab = "Month", ylab = "Mean lnc", las = 1, col= 2)
lines(as.numeric(Grp2$month), Grp2$Meanlnc, type = "b",
      col = 3, pch=16)
lines(as.numeric(Grp3$month), Grp3$Meanlnc, type = "b",
      col = 4, pch=16)
legend(locator(1),           # we will place it with a mouse click
       legend = c("1","2","3"),
       title = "Treat", lty = c(2,2,2), col= 2:4)
```

# The rats data - plotting the average patterns



# Overview

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a summary statistic
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

## Separate analysis for each time-point (continued)

- Select a fixed time point
- The observations at that time (one from each individual) are independent
- Do a **separate** analysis for the observations at that time
- This is not wrong, but (possibly) a lot of **information is wasted**
- This can be done for several time-points, but
  - Difficult to reach a **coherent** conclusion
  - Sub-tests are not **independent**
  - Tempting to select time-points supporting our preference
  - Mass significance: If many tests are carried out at 5% level some might be significant by chance. (Bonferroni correction: Use significance level  $0.05/n$  instead of 0.05)

# Separate analysis of rats data

```
#use the by function to make 10 tests
byMonth <- by(rats, rats$monthQ,
               function(x) anova(lm(lnc ~ treatm, data = x)))
#The largest effect at month 8
byMonth[[8]]

## Analysis of Variance Table
##
## Response: lnc
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatm     2  0.649   0.324    7.29  0.0029 **
## Residuals 27  1.201   0.044
```

## Separate analysis of rats data

- The model at each month is:

$$\ln c_i = \mu + \alpha(\text{treatm}_i) + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \quad i = 1 \dots 30$$

- The result of the ten tests for no treatment effect:

Month	1	2	3	4	5	6	7	8	9	10
F-value	1.22	0.27	1.02	2.30	3.87	4.10	4.70	7.29	4.09	0.88

Compare with  $F_{95\%;2,27} = 3.35$  or  $F_{99.5\%;2,27} = 6.49$  if Bonferroni correction is used

# Overview

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a **summary statistic**
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

## Rats data analyzed via summary measure

- The log of the total activity in each cage is chosen as summary measure  $\ln\text{Tot} = \log(\text{Total count})$
- The one way ANOVA model becomes:

$$\ln\text{Tot}_i = \mu + \alpha(\text{treatm}_i) + \varepsilon_i, \quad \varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2), \quad i = 1 \dots 30$$

- Notice the simplicity of the model and the relative few assumptions

# Rats data analyzed via a summary measure – log total activity

```
rats$count <- exp(rats$lnc)
DataTot <- aggregate(rats$count, by = list(rats$cage,
                                             rats$treatm), sum)
names(DataTot) <- c("cage", "treatm", "Tot")
head(DataTot)

##    cage treatm     Tot
## 1     1      1 124864
## 2     3      1 131277
## 3     5      1 110166
## 4     7      1 145418
## 5     9      1 128819
## 6    11      1 136214
```

## log total activity (contd.)

```
reg1 <- lm(log(Tot) ~ treatm, data = DataTot); summary (reg1)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.73354   0.05409 216.91  <2e-16 ***
## treatm2     -0.00567   0.07650  -0.07   0.941
## treatm3      0.16728   0.07650    2.19   0.038 *
##
## Residual standard error: 0.171 on 27 degrees of freedom
## Multiple R-squared:  0.196, Adjusted R-squared:  0.137
## F-statistic:  3.3 on 2 and 27 DF,  p-value: 0.0522

anova(reg1)

## Analysis of Variance Table
##
## Response: log(Tot)
##             Df Sum Sq Mean Sq F value Pr(>F)
## treatm      2  0.193  0.0965    3.3  0.052 .
## Residuals  27  0.790  0.0293
```

## Rats data analyzed via a summary measure - slopes

```
# byCage is a regression with different slope and intercept
# for each cage. Saves the coefficients
byCage <- coef(lm(lnc ~ -1 + cage + monthQ:cage, data = rats))
slope1 <- data.frame(matrix(byCage, nrow=30, byrow=F))

names(slope1) <- c("Intercept", "Slope")

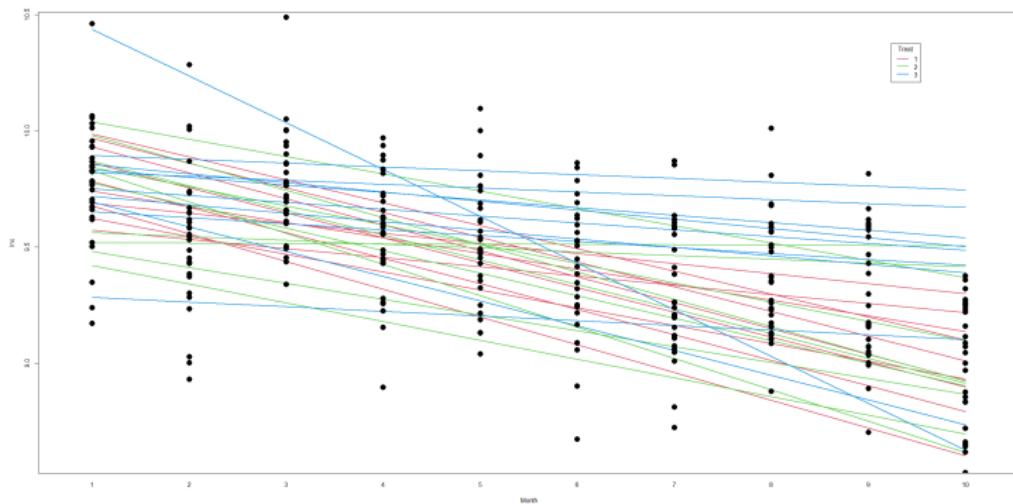
#We also need info about cage and treat
SlopeData <- cbind(DataMean[ , 1:2], slope1)
head(SlopeData)

##   cage treatm Intercept      Slope
## 1     1       1    9.9685 -0.107163
## 2     3       1   10.0411 -0.111593
## 3     5       1    9.6981 -0.076675
## 4     7       1   10.0831 -0.098336
## 5     9       1    9.8067 -0.066990
## 6    11       1    9.7310 -0.043147
```

# Rats data analyzed via a summary measure - plotting slopes

```
# Individual slopes:  
fit <- unlist(by(rats, rats$cage,  
                  function(x) fitted.values(lm(lnc ~ monthQ, data=x))))  
names(fit) <- NULL  
  
#plotting the linear fit by cage  
interaction.plot(rats$monthQ, rats$cage, fit,  
                 xlab="Month", ylab="lnc", legend=F,  
                 col=as.numeric(SlopeData$treatm)+1, lty=1, lwd=2)  
  
lines(rats$monthQ,rats$lnc, type="p", pch=16, cex=2)  
legend(locator(1), # we will place it with a mouse click  
       legend = c("1","2","3"), title = "Treat",  
       lty = c(1,1,1), lwd=2, col= 2:4)
```

# Rats data analyzed via a summary measure - plotting slopes



## Rats data analyzed via a summary measure - slopes (contd)

```
reg3 <- lm(Slope ~ treatm, data = SlopeData)
summary(reg3)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0879    0.0142   -6.20  1.2e-06 ***
## treatm2      0.0100    0.0200    0.50   0.620
## treatm3      0.0355    0.0200    1.77   0.088 .
##
## Residual standard error: 0.0448 on 27 degrees of freedom
## Multiple R-squared:  0.11, Adjusted R-squared:  0.0439
## F-statistic: 1.67 on 2 and 27 DF,  p-value: 0.208
```

## Rats data analyzed via a summary measure - slopes (contd)

```
reg3 <- lm(Slope ~ treatm, data = SlopeData)
summary(reg3)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.0879    0.0142   -6.20  1.2e-06 ***
## treatm2      0.0100    0.0200    0.50   0.620
## treatm3      0.0355    0.0200    1.77   0.088 .
##
## Residual standard error: 0.0448 on 27 degrees of freedom
## Multiple R-squared:  0.11, Adjusted R-squared:  0.0439
## F-statistic: 1.67 on 2 and 27 DF,  p-value: 0.208
```

- Treat 1 has been chosen as reference with estimated slope -0.09.
- Treat 2 is very similar to Treat 1 with slope  $-0.09 + 0.01 = -0.08$
- Treat 3 has slope closer to 0:  $-0.09 + 0.04 = -0.05$

## Rats data analyzed via a summary measure - slopes (contd)

```
anova(reg3)

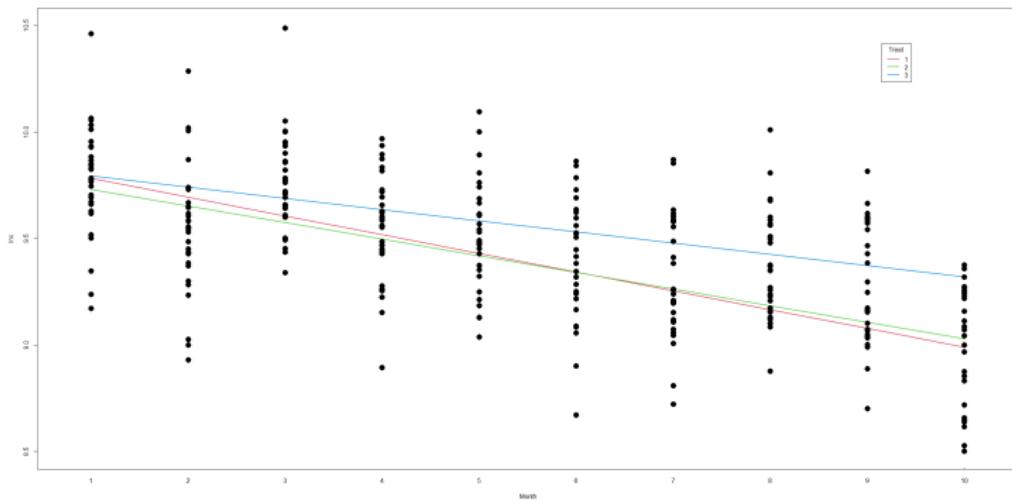
## Analysis of Variance Table
##
## Response: Slope
##           Df Sum Sq Mean Sq F value Pr(>F)
## treatm     2 0.0067 0.00334   1.67   0.21
## Residuals 27 0.0542 0.00201
```

From the one-way analysis of variance we do not see a significant difference in the slopes between the three groups.

# Rats data analyzed via a summary measure - plotting slopes

```
# Treatment average slopes:  
fit2 <- unlist(by(rats, rats$treatm,  
                     function(x) fitted.values(lm(lnc ~ monthQ, data=x))))  
names(fit2) <- NULL  
  
#plotting the linear fit by treatment  
interaction.plot(rats$monthQ, rats$treatm, fit2,  
                  xlab="Month", ylab="lnc", legend=F,  
                  col=2:4, lty=1, lwd=2, ylim=c(8.5,10.5))  
  
lines(rats$monthQ,rats$lnc, type="p", pch=16, cex=2)  
legend(locator(1),  
       legend = c("1","2","3"), title = "Treat",  
       lty = c(1,1,1), lwd=2, col= 2:4)
```

# Rats data analyzed via a summary measure - plotting slopes



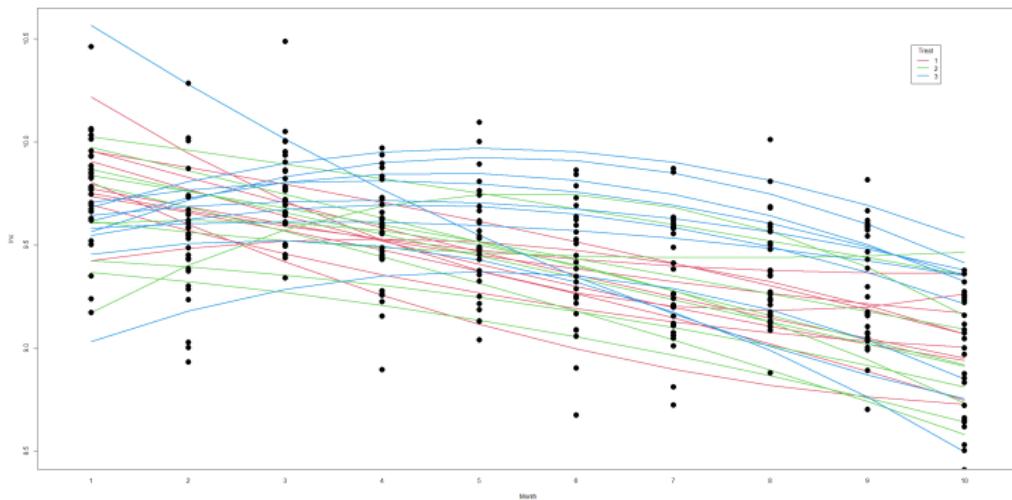
# Rats data analyzed via a summary measure - plotting curves

```
# Individual curves: Look similar apart from one of the blue
rats$monthQ2 <- rats$monthQ^2
fit3 <- unlist(by(rats, rats$cage,
  function(x) fitted.values(lm(lnc ~ monthQ+monthQ2, data=x))))
names(fit3) <- NULL

#plotting the fit by cage
interaction.plot(rats$monthQ, rats$cage, fit3,
  xlab="Month", ylab="lnc", legend=F,
  col=as.numeric(SlopeData$treatm)+1, lty=1, lwd=2,)

lines(rats$monthQ,rats$lnc, type="p", pch=16, cex=2)
legend(locator(1),
  legend = c("1","2","3"), title = "Treat",
  lty = c(1,1,1), lwd=2, col= 2:4)
```

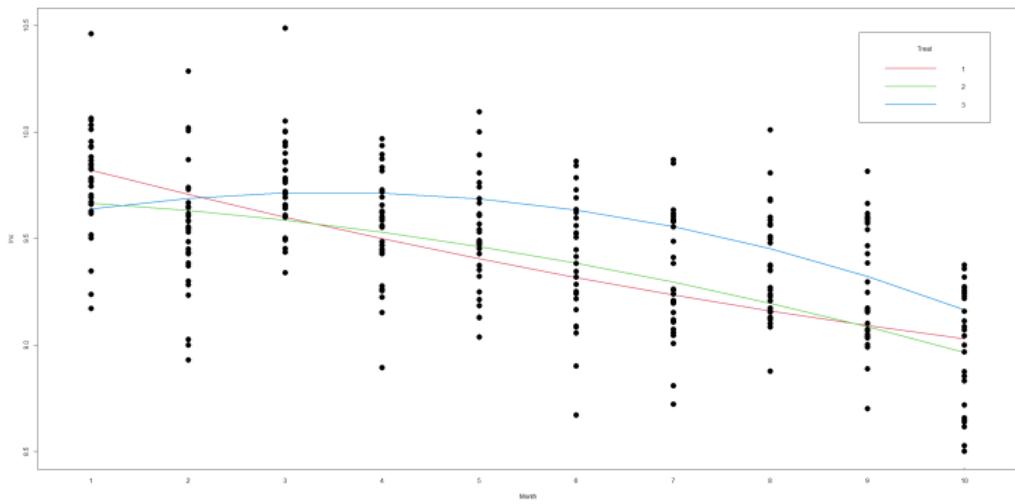
# Rats data analyzed via a summary measure - plotting curves



## Rats data analyzed via a summary measure - plotting curves

```
# Treatment average curves:  
fit4 <- unlist(by(rats, rats$treatm,  
  function(x) fitted.values(lm(lnc ~ monthQ+monthQ2, data=x))))  
names(fit4) <- NULL  
  
#plotting the fit by treatment  
interaction.plot(rats$monthQ, rats$treatm, fit4,  
  xlab="Month", ylab="lnc", legend=F,  
  col=2:4,lty=1,lwd=2,ylim=c(8.5,10.5))  
  
lines(rats$monthQ,rats$lnc, type="p",pch=16,cex=2)  
legend(locator(1),  
  legend = c("1","2","3"), title = "Treat",  
  lty = c(1,1,1),lwd=2,col= 2:4)
```

# Rats data analyzed via a summary measure - plotting curves



# Rats data analyzed via a summary measure - curvatures

```
byCage2 <- coef(lm(lnc ~ -1 + cage + monthQ:cage +
                     monthQ2:cage, data = rats))

# I want a data frame, one row for each cage
curve <- data.frame(matrix(byCage2, nrow=30, byrow=F))
names(curve) <- c("Intercept", "monthQ", "monthQ2")
# We also need info about cage and treat
CurveData <- cbind(DataMean[, 1:2], curve)
head(CurveData)

##   cage treatm Intercept    monthQ    monthQ2
## 1     1      1  10.0465 -0.146171  0.0035462
## 2     3      1  10.0886 -0.135343  0.0021591
## 3     5      1   9.8302 -0.142721  0.0060042
## 4     7      1  10.0237 -0.068641 -0.0026996
## 5     9      1   9.8664 -0.096853  0.0027148
## 6    11      1   9.8445 -0.099881  0.0051576
```

## Rats data analyzed via a summary measure - curvatures (contd)

Each estimate of the curve is used as the response

```
reg4 <- lm(monthQ2 ~ treatm, data = CurveData)
summary(reg4)

## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.00320   0.00298   1.08  0.29185
## treatm2     -0.00854   0.00421  -2.03  0.05237 .
## treatm3     -0.01608   0.00421  -3.82  0.00071 ***
##
## Residual standard error: 0.00941 on 27 degrees of freedom
## Multiple R-squared:  0.351, Adjusted R-squared:  0.303
## F-statistic: 7.3 on 2 and 27 DF,  p-value: 0.00291
```

- For Treat 1 the estimated curve is positive (nearly zero 0.003)
- For Treat 2 the estimated curve is negative (0.003-0.009=-0.006)
- For Treat 2 the estimated curve is also negative (0.003-0.016=-0.013)

# Rats data analyzed via a summary measure - curvatures (contd)

```
anova(reg4)
```

```
## Analysis of Variance Table
##
## Response: monthQ2
##           Df  Sum Sq  Mean Sq F value Pr(>F)
## treatm     2 0.00129 0.000647    7.3 0.0029 **
## Residuals 27 0.00239 0.000089
```

When using the curvature as the response we get a significant difference between the three groups ( $p = 0.003 < 0.05$ )

# Overview

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a summary statistic
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

# Random effects model - simple version

- This model uses **all observations** instead of reducing to one observation per individual
- We can test the **time-by-treatment interaction**
- Add “individual” as a random effect
- Makes measurements on same individual correlated
- Unfortunately **equally correlated** no matter if they are “close” or “far apart”
- Can be considered first step in modeling the actual covariance structure
- Usually only good for short series

## Rats data analyzed via random effects approach

- The model can now be enhanced to:

$$\ln c_i = \mu + \alpha(\text{treatm}_i) + \beta(\text{month}_i) + \gamma(\text{treatm}_i, \text{month}_i) + d(\text{cage}_i) + \varepsilon_i,$$

- The covariance structure of this model is:

$$\text{cov}(y_{i_1}, y_{i_2}) = \begin{cases} 0 & , \text{ if } \text{cage}_{i_1} \neq \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_d^2 & , \text{ if } \text{cage}_{i_1} = \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_d^2 + \sigma^2 & , \text{ if } i_1 = i_2 \end{cases}$$

# Rats data analyzed via random effects approach

- The model can now be enhanced to:

$$\ln c_i = \mu + \alpha(\text{treatm}_i) + \beta(\text{month}_i) + \gamma(\text{treatm}_i, \text{month}_i) + d(\text{cage}_i) + \varepsilon_i,$$

- The covariance structure of this model is:

$$\text{cov}(y_{i_1}, y_{i_2}) = \begin{cases} 0 & , \text{ if } \text{cage}_{i_1} \neq \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_d^2 & , \text{ if } \text{cage}_{i_1} = \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \sigma_d^2 + \sigma^2 & , \text{ if } i_1 = i_2 \end{cases}$$

One can think of the model as:

$\log(\text{count}) = \text{"mean" + "between cage variation" + "within cage variation"}$

# Multilevel model structure

- This model is also called a two-level model.

Level	1	2
Unit	Each separate obs.	Cages
Variation	Within cage $\sigma^2$	Between cage $\sigma_d^2$
Covariates	month month:treatm	treatm

# Model synonyms

- Two-level model
- Mixed model with random subject level
- Mixed model with random intercept
- Model with compound symmetry correlation structure
- Model with exchangeable correlation structure

## Compound symmetry

The model implies that all observations from the same cage are correlated with the intra-class correlation:

$$\text{Corr}(Y_{i_1}, Y_{i_2}) = \frac{\text{cov}(Y_{i_1}, Y_{i_2})}{\sqrt{\text{var}(Y_{i_1})}\sqrt{\text{var}(Y_{i_2})}} = \frac{\sigma_d^2}{\sigma_d^2 + \sigma^2}$$

Here  $\text{cage}_{i_1} = \text{cage}_{i_2}$  and  $i_1 \neq i_2$ .

We are not taking the distance between observations into account.

Maybe too simple? Perhaps observations close in time are more similar than observations far apart.

This correlation structure is called **exchangeable** or compound symmetry.

# Potential mistakes when leaving the random effect out

Level	Unit	Covariates
1	Each separate obs.	month
		month:treatm
2	Cages	treatm

If a random effect is present:

- Potential bias in the mean (ignoring that observations go together)
- Estimates on Level 1 may have too much variation. Ignoring pairs → too high p-values. Effects can be overlooked
- Estimates on Level 2 may have too little variation → too small p-values. “Noise” can become an effect

# Rats data analyzed via random effects approach in R

```
library(nlme)
model1 <- lme(lnc ~ month + treatm + month:treatm,
               random = ~1 | cage, data = rats)
anova(model1)

##                               numDF denDF F-value p-value
## (Intercept)              1     243   85525  <.0001
## month                  9     243      46  <.0001
## treatm                  2      27      3  0.0557
## month:treatm            18    243      2  0.0059
```

# Rats data analyzed via random effects approach in R

```
library(nlme)
model1 <- lme(lnc ~ month + treatm + month:treatm,
               random = ~1 | cage, data = rats)
anova(model1)

##                               numDF denDF F-value p-value
## (Intercept)              1     243   85525  <.0001
## month                  9     243      46  <.0001
## treatm                  2      27      3  0.0557
## month:treatm            18    243      2  0.0059
```

So we have a significant interaction between treatment and month.

## Table of fixed effects

#This time summary gives a lot of output!

```
#summary(model1)
```

#A table of fixed effects estimates

```
summary(model1)$tTable
```

##	Value	Std.Error	DF	t-value	p-value
## (Intercept)	9.874280	0.080856	243	122.12192	4.1653e-220
## month2	-0.282870	0.087062	243	-3.24908	1.3213e-03
## month3	-0.199010	0.087062	243	-2.28585	2.3124e-02
## month4	-0.381285	0.087062	243	-4.37948	1.7692e-05
## month5	-0.464289	0.087062	243	-5.33288	2.2101e-07
.....					
## month7:treatm3	0.472180	0.123124	243	3.83500	1.6005e-04
## month8:treatm3	0.465124	0.123124	243	3.77770	1.9913e-04
## month9:treatm3	0.443817	0.123124	243	3.60464	3.7919e-04
## month10:treatm3	0.202159	0.123124	243	1.64192	1.0190e-01

# Table of random effects

```
#The estimates random effects  
VarCorr(model1)  
  
## cage = pdLogChol(1)  
##           Variance StdDev  
## (Intercept) 0.027478 0.16577  
## Residual    0.037899 0.19468  
  
#intra class correlation  
0.027478/(0.027478+0.037899)  
  
## [1] 0.4203
```

$$\hat{\sigma}_d^2 = 0.03 \text{ and } \hat{\sigma}^2 = 0.04$$

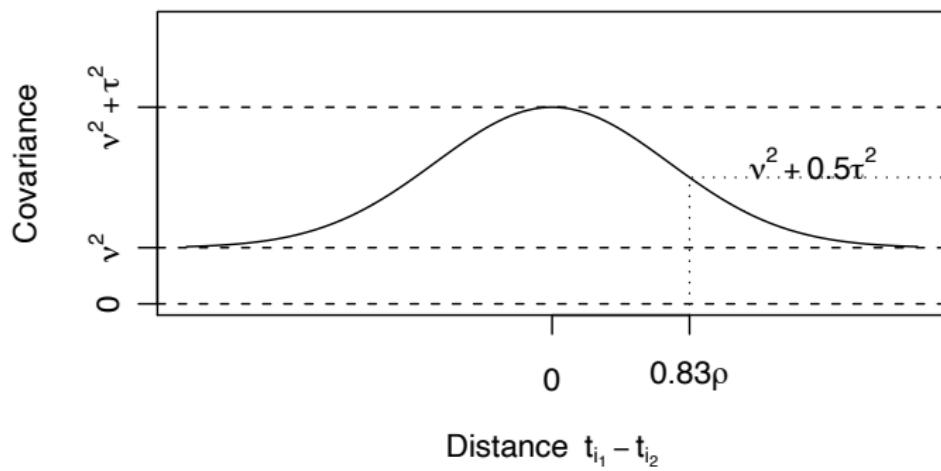
# Overview

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a summary statistic
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

# Advanced Covariance Structures: Gaussian spatial correlation

- *Spatial* covariance structures, depending on “how far” observations are apart.
- Gaussian decline of dependency:

$$Cov(Y_{i_1}, Y_{i_2}) = \begin{cases} 0 & , \text{ if } \text{indiv}_{i_1} \neq \text{indiv}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 \exp\left\{\frac{-(t_{i_1} - t_{i_2})^2}{\rho^2}\right\} & , \text{ if } \text{indiv}_{i_1} = \text{indiv}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 + \sigma^2 & , \text{ if } i_1 = i_2 \end{cases}$$



# Rats data via spatial Gaussian correlation model

- Model for the entire observational vector:

$$\begin{aligned} \mathbf{Y} &\sim N(\mu, \mathbf{V}), \quad \text{where} \\ \mu_i &= \mu + \alpha(\text{treatm}_i) + \beta(\text{month}_i) + \gamma(\text{treatm}_i, \text{month}_i), \quad \text{and} \\ V_{i_1, i_2} &= \begin{cases} 0 & , \text{ if } \text{cage}_{i_1} \neq \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 \exp \left\{ \frac{-(\text{month}_{i_1} - \text{month}_{i_2})^2}{\rho^2} \right\} & , \text{ if } \text{cage}_{i_1} = \text{cage}_{i_2} \text{ and } i_1 \neq i_2 \\ \nu^2 + \tau^2 + \sigma^2 & , \text{ if } i_1 = i_2 \end{cases} \end{aligned}$$

- R code for model implementation:

```
analysis<-lme(lnc~month+treatm+month:treatm,
  random=~1|cage,
  correlation=corGaus(form=~as.numeric(month)|cage,nugget=T),
  data=rats)
```

- Partial R output:

Random effects:

Formula: ~1 | cage

(Intercept) Residual

StdDev: 0.1404056 ( $= \hat{\nu}$ ) 0.2171559 ( $= \sqrt{\hat{\sigma}^2 + \hat{\tau}^2}$ )

Correlation Structure: Gaussian spatial correlation

Formula: ~as.numeric(month) | cage

Parameter estimate(s):

range nugget

2.3863954 ( $= \hat{\rho}^2$ ) 0.2186744 ( $= \hat{\sigma}^2 / (\hat{\sigma}^2 + \hat{\tau}^2)$ )

Number of Observations: 300

Number of Groups: 30

- Notice the R parametrization of the variance parameters

# Other spatial correlation structures

- R has a number of build-in correlation structures. A few examples:

<b>correlation=</b>	<b>Name</b>	<b>Correlation term</b>
sp(gau)(t)	Gaussian	$\tau^2 \exp\left\{-\frac{(t_{i1}-t_{i2})^2}{\rho^2}\right\}$
sp(exp)(t)	exponential	$\tau^2 \exp\left\{-\frac{ t_{i1}-t_{i2} }{\rho}\right\}$
ar(1)	autoregressive(1)	$\tau^2 \rho^{ i_1-i_2 }$
un	unstructured	$\tau_{i_1,i_2}^2$

## How to select the correlation structure -Stationarity

- For all the listed covariance structures, the value of  $Cov(Y_t, Y_{t+u})$  does not depend on the time  $t$ ; only on the time difference  $u$ .
- We say that the error process  $\varepsilon_t$  is weakly stationary of order 2;
- weakly because the stationarity is defined from moments, and not distributions (for Gaussian processes this is the same though);
- of order 2 because the stationarity is defined through 2nd order moments (variance, covariance).

# The semi-variogram

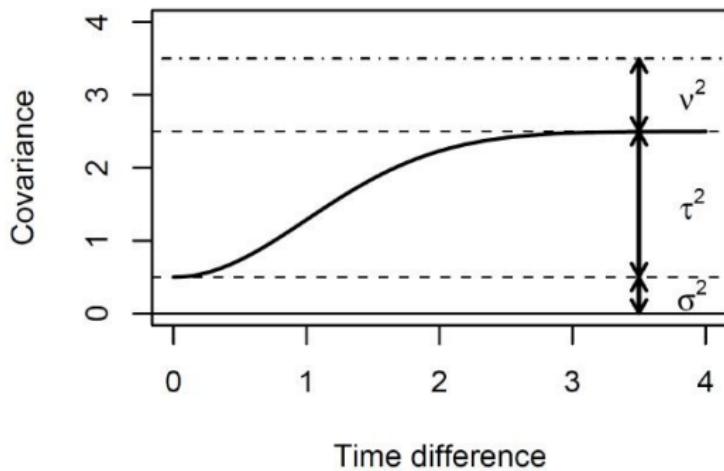
- The semi-variogram plots  $\gamma(u) = \frac{1}{2}V(Y_t - Y_{t+u})$ ,  $u > 0$ :

$$\begin{aligned}\gamma(u) &= \frac{1}{2}V(Y_t - Y_{t+u}) \\ &= \frac{1}{2}(V(Y_t) + V(Y_{t+u}) - 2Cov(Y_t, Y_{t+u})) \\ &= V(Y_t) - Cov(Y_t, Y_{t+u}) \\ &= \nu^2 + \tau^2 + \sigma^2 - \nu^2 - \tau^2\lambda(u) \\ &= \sigma^2 + \tau^2(1 - \lambda(u)),\end{aligned}$$

where  $\lambda(u) = exp(-u^2/\rho^2)$ ,  $exp(-u/\rho)$ ,  $\rho^u$  for spatial Gaussian, exponential and AR1 correlation structures, respectively.

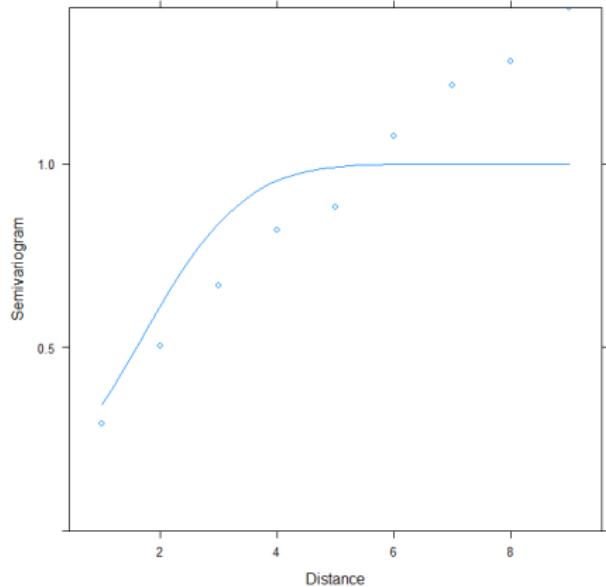
# The semi-variogram

- Theoretical semi-variogram, Gaussian correlation structure:
- Plot of  $\sigma^2 + \tau^2(1 - \lambda(u))$ , where  $\lambda(u) = \exp(-u^2/\rho^2)$ :



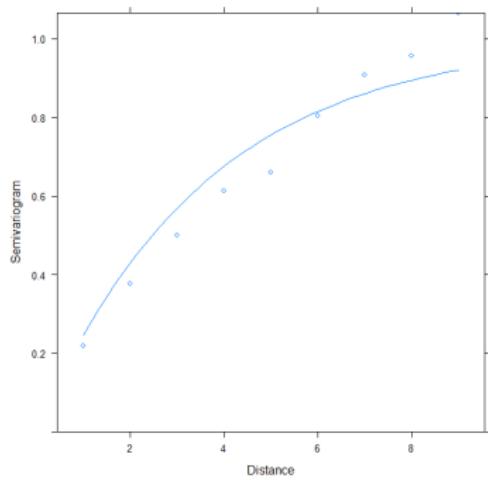
# The semi-variogram

```
plot(Variogram(analysis,form=~monthQ|cage,data=rats))
```



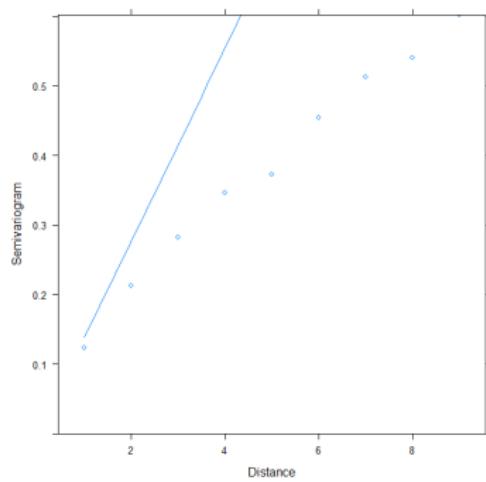
# The semi-variogram, exponentially decreasing correlation

```
analysis2<-lme(lnc~month+treatm+month:treatm, random=~1|cage,  
                  correlation=corExp(form=~monthQ|cage,nugget=T),  
                  data=rats)  
plot(Variogram(analysis2,form=~monthQ|cage,data=rats))
```



# The semi-variogram, linearly decreasing correlation

```
analysis3<-lme(lnc~month+treatm+month:treatm, random=~1|cage,  
correlation=corLin(form=~monthQ|cage),  
data=rats)  
plot(Variogram(analysis3,form=~monthQ|cage,data=rats))
```



# The semi-variogram, summing up

- The semi-variogram **compares a theoretical and empirical correlation structure**. Similarity indicates a good model.
- For the *rats data*, the semi-variogram indicates that the optimal correlation structure seems to be an exponentially decreasing spatial correlation structure, rather than the standard spatial Gaussian structure.

# The rats data, partial final model output

- Random effects:

Formula: ~1 | cage

(Intercept) Residual

StdDev: 0.07828533 ( $= \hat{\nu}$ ) 0.2510369 ( $= \sqrt{\hat{\sigma}^2 + \hat{\tau}^2}$ )

Correlation Structure: Exponential spatial correlation

Formula: ~as.numeric(month) | cage

Parameter estimate(s):

range nugget

3.556503e+00 ( $= \hat{\rho}$ ) 3.341370e-08 ( $= \hat{\sigma}^2 / (\hat{\sigma}^2 + \hat{\tau}^2)$ )

Number of Observations: 300

Number of Groups: 30

# Overview

- 1 The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- 2 Separate analysis for each time-point
  - Example: rats data
- 3 Analysis of a summary statistic
  - Example: rats data
- 4 Random effects model - simple version
  - Example: rats data
- 5 Random effects model - advanced version
- 6 Pros and cons of simple approaches

# Pros and cons of approaches

## Separate analysis for each time-point

- + Not wrong
- Can be confusing
- Difficult to reach coherent conclusion
- In general not very informative

## Analysis of summary statistic

- + Good method with few and easily checked assumptions
- Important to choose good summary measure(s)

## Random effects approach - simple version

- + Good method for short series
- + Uses all observations
- Usually not good for long series

## Random effects approach - qadvanced version

- + Works for short and long series
- + Uses all observations
- Requires appropriate choice of covariance decay function

# Overview of this module

- ① The repeated measurements setup
  - Aspirin Example
  - Activity of rats
- ② Separate analysis for each time-point
  - Example: rats data
- ③ Analysis of a summary statistic
  - Example: rats data
- ④ Random effects model - simple version
  - Example: rats data
- ⑤ Random effects model - advanced version
- ⑥ Pros and cons of simple approaches

## Exercise: Histamin in dogs

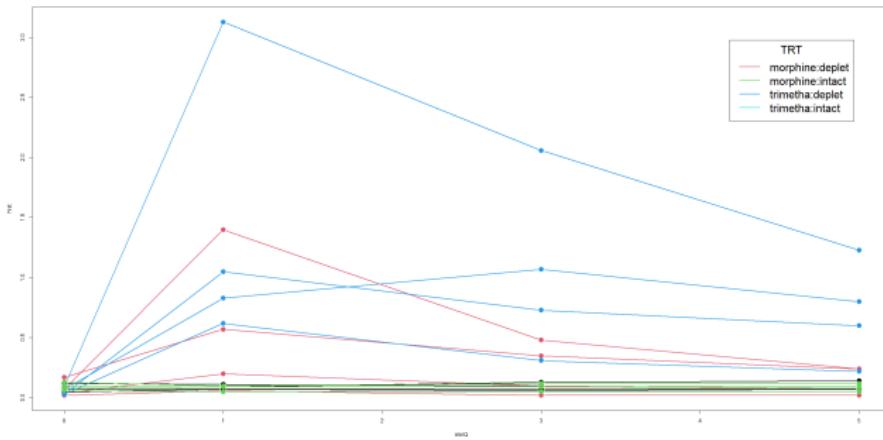
In an experiment with 16 dogs the blood histamine concentration was measured 0, 1, 3, and 5 minutes after injection of morphine or trimethaphane. Before injection the dogs were classified into two groups according to their level of histamine (intact or depleted).

```
histamin <- read.table("histamin.txt", header=T, sep=",", dec=".")  
histamin$dog <- factor(histamin$dog)  
histamin$minQ <- histamin$min  
histamin$min <- factor(histamin$min)  
histamin$TRT <- factor(paste(histamin$treatm,":",histamin$level,  
                         sep=""))  
#summary(histamin)
```

# Histamin in dogs

```
plot(histamin$minQ,histamin$hist,xlab="minQ",ylab="hist",pch="")  
for(i in 1:64){  
  temp<-histamin[histamin$dog==i,]  
  lines(temp$minQ,temp$hist,col=(1:4)[temp$TRT[1]],lwd=2,type="b",  
    pch=16,cex=2)  
}  
legend(locator(1),  
  legend = levels(histamin$TRT), title = "TRT",  
  lty = c(1,1,1,1),lwd=2,col= 2:5,cex=2)
```

# Histamin in dogs



## Histamin in dogs - questions

First of all use the TRT factor for the analysis (defining 4 groups of dogs).

- ① Make some plots of the data, for instance one line for each dog (maybe colored differently in each TRT group).
- ② Analyze these data using one or more of “the simple methods”.
- ③ How would you approach making a conclusion for the “real” treatment: morphine vs. trimetha?