# Laboratory exercises

The laboratory exercises outlined in the following pages are designed to allow you to develop some expertise in using statistical software (R) to analyze data. R is powerful statistical software but like all software, it has its limitations. In particular, it is dumb: it cannot think for you, it cannot tell you whether the analysis you are attempting to do is appropriate or even makes any sense, and it cannot interpret your results.

## General points to keep in mind

- Before attempting any statistical procedure, you must familiarize yourself with what the procedure is actually doing. This does not mean you actually have to know the underlying mathematics (although this certainly helps!), but you should at least understand the principles involved in the analysis. Therefore, before doing a laboratory exercise, read the appropriate section(s) in the lecture notes. Otherwise, the output from your analyses - even if done correctly - will seem like drivel.
- The laboratories are designed to complement the lectures, and *vice versa*. Owing to scheduling constraints, it may not be possible to synchronize the two perfectly. But feel free to bring questions about the laboratories to class, or questions about the lectures to the labs.
- Work on the laboratories at your own speed: some can be done much more quickly than others, and one laboratory need not correspond to one laboratory session. In fact, for several laboratories we have allotted two laboratory sessions. Although you will not be "graded" on the laboratories per se, be aware that completing the labs is essential. If you do not complete the labs, it is very unlikely that you will be able to complete the assignments and the final exam. So take these laboratories seriously!
- The objective of the first lab is to allow you to acquire or review the minimum knowledge required to complete the following laboratory exercises with R. There are always several methods to accomplish something in R, but you will only find simple ways in this manual. Those amongst you that want to go further will easily find many examples of more detailed and sophisticated methods. In particular, I point you to the following resources:

  **R for beginners**
  http://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf

**Using R for psychological research: A simple guide to an elegant package**
http://www.personality-project.org/r/

**An introduction to R**
http://cran.r-project.org/doc/manuals/R-intro.html

If you prefer paper books, the CRAN web site has a commented list at:
http://www.r-project.org/doc/bib/R-books.html

Finally, I also recommend this short reference card:

**R reference card by Tom Short**
http://cran.r-project.org/doc/contrib/Short-refcard.pdf

## What is R and why use it in this course

R is multiplatform free software forming a system for statistical computation and graphics. R is also a programming language specially designed for statistical data analysis. It is a dialect of the S language. S-Plus is another dialect of the S language, very similar to R, incorporated into a commercial package. S-Plus has a built-in graphical design intreface that some find convivial.

R has 2 major advantages for this course. Initially, you will find that it also has one inconvenience. However, this "inconvenience" will rapidly force you to acquire very good working habits. So, I see it as a third advantage.

The first advantage is that you can install it freely on you personal computer(s). This is important because it is by doing analyses that you will learn and eventually master biostatistics. This implies that you have easy and unlimited access to a statistical software package.

The second advantage is that R can do everything in statistics. R was conceived to be extensible and has become the preferred tool for statisticians around the world. The question is not "Can R do this?" but rather "How can I do this in R?". And Google is your friend.

No other software package offers you these two advantages.

The inconvenience of R is that one has to type commands (or copy and paste code) rather than use a menu and select options. If you do not know what command to use, nothing will happen. It is therefore not that easy when you start. However, it is possible to rapidly learn to make basic operations (open a data file, plot data, and run a simple

analysis). And once you understand the operating principle, you can easily find examples on the Web for more complex analyses and graphs for which you can adapt the code.

This is exactly what you will do in the first lab to familiarize yourself with R.

Why is this inconvenience really an advantage in my mind? Because this way of doing things is more efficient and will save you time on the long run. I guarantee it. Believe me, you will never do an analysis only once. As you'll proceed through analyses, you will find data entry errors, discover that the analysis must be run separately for subgroups, find extra data, have to rerun the analysis on transformed data, or you will make some analytical error along the way. If you use a graphical interface with menus, redoing an analysis implies that you reclick here, enter values there, select some options, etc. Each of these steps is a potential source of error. If, instead, you use lines of codes, you only have to fix the code and submit to repeat instantaneously the entire analysis. And you can perfectly document what you did, leaving an audit trail for the future. This is how pros work and can document the quality of the results of their analyses.

## Installation

To install R on a computer, go to `http://cran.r-project.org/`. You will find compiled versions (binairies) for your preferred operating system (Windows, MacOS, Linux).

The version I installed when I started these notes is 2.9.1 for Windows (36 Mbytes). I installed it with the default options, although along the way I installed quite a number of extra packages.

Note : R has already been installed on the lab computers (the version may be slightly different, but this should not matter).

## General lab instructions

- Bring a USB key or equivalent so you can save your work. Alternatively, email your results to yourself.
- Read the lab exercise before coming to the lab. Read the R code and come with questions about the code.
- During pre-labs, listen to the special instructions
- Do the laboratory exercises at your own rhythm, in teams. Then, I recommend that you start (complete?) the lab assignment so that you can benefit from the presence of the TA or prof.

- During your analyses, copy and paste results in a separate document, for example in your preferred word processing program. Annotate abundantly
- Each time you shut down R, save the history of your commands (ex: labo1.1 rHistory, labo1.2.rHistory, etc). You will be able to redo the lab rapidly, get code fragments, or more easily identify errors.
- Create your own "library" of code fragments (snippets). Annotate it abundantly. You will thank yourself later.

### And for the dilettantes or those completely refractory to programming...

If you find the absence of a graphical user interface too spartan or if you have recurring problems finding your way that make you waste too much time, then install and load the graphical interface distributed by CRAN and that works under Windows, Mac and Linux (so they say, although I only tested it under Windows).

The library (module) is called Rcmdr. It is not part of the base distribution, and you will need to install it first into R on your own computer (it has been installed already in the lab, but is not loaded automatically when you start R). To install this package, you need an internet connection. In the command window, enter the command:

```
install.packages("Rcmdr", dependencies=TRUE)
```

A window will open to let you choose one of the web sites that distribute Rcmdr. The download and installation will take several minutes because several other libraries are needed and will be automatically downloaded. Once the Rcmdr library will have been installed, you will need to start it to get to the graphical interface. To do so, simply enter the command:

```
library(Rcmdr)
```

A new window will open, with menus to access most of the common commands. One of the panels of the Rcmdr window is a log of the R commands created by Rcmdr as you make choices in the graphical user interface. Study this code to learn, maybe more easily, the R language.

# Lab- R tutorial

After completing this laboratory exercise, you should be able to :

- Open R data files
- Import rectangular data sets
- Export R data to text files
- Verify that data were imported correctly
- Examine the distribution of a variable
- Examine visually and test for normality of a variable
- Calculate descriptive statistics for a variable
- Transform data

## Importing and exporting data

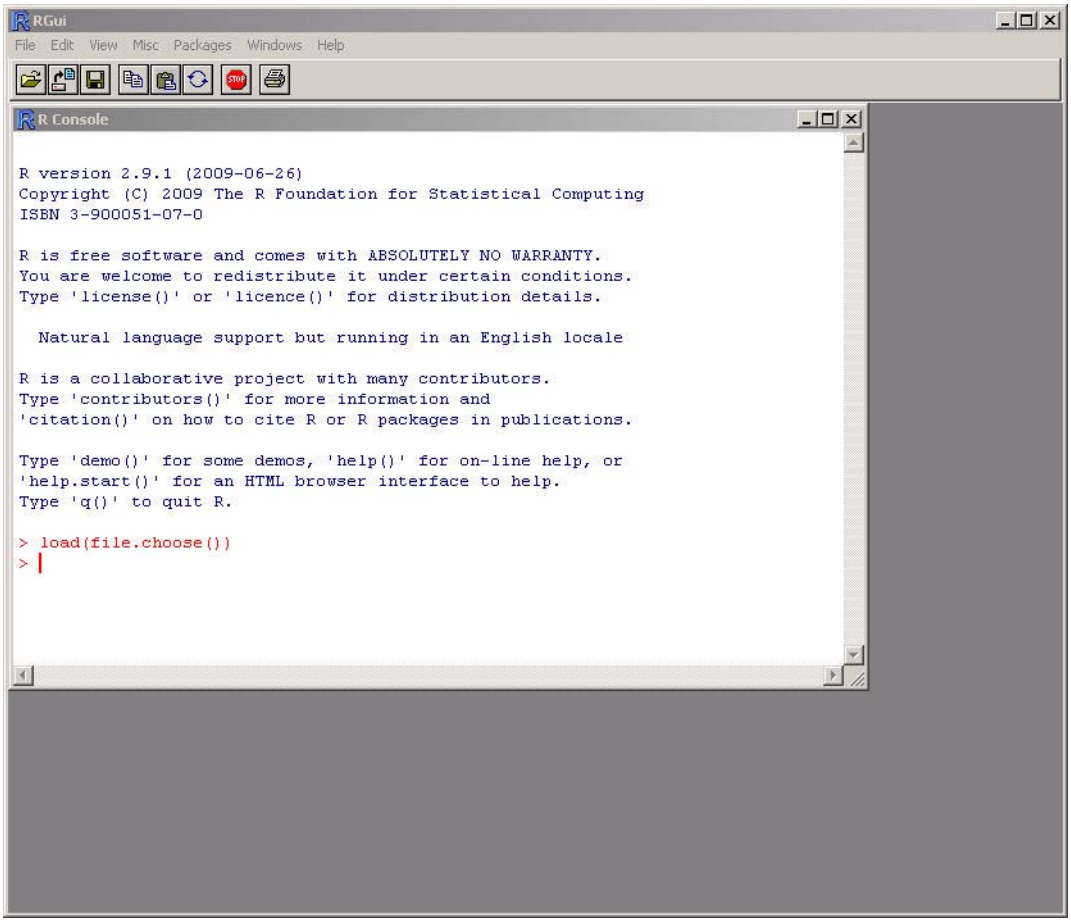
### Open and save an R data file

Data for the lab exercises and assignments are provided to you in R format (files with extension .Rdata). To open these files, if your operating system allows it, you can simply click on the file to start a new R session. Alternatively, once R is started, you can type the following at the R console window:

```
load(file.choose())
```

to open a dialog box that will allow you to navigate to a folder on your computer to choose the file you want to load. After having made your selection, you will be returned to the R console window with no apparent change. :

**Figure 1.**




To verify that the data were read and loaded properly, you can list all objects in memory with the ls() function, or get a more detailed description with ls.str():

```
> load(file.choose())
(#Choose the file ErablesGatineau.Rdata)
> ls()
[1] "ErablesGatineau"
> ls.str()
ErablesGatineau : 'data.frame': 100 obs. of  3 variables:
 $ station: Factor w/ 2 levels "A          ",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ diam   : num  22.4 36.1 44.4 24.6 17.7 ...
 $ biom   : num  732 1171 673 1552 504 ...
```

R confirms that the object `ErablesGatineau` is in memory. `ErableGatineau` is a data.frame that contains 100 observations (lines) of 3 variables (columns) : `station`, a variable of type Factor with 2 levels, and `diam` and `biom` that are 2 numeric variables.

### Entering data in R

R is not the ideal environment to input data. It is possible, but the syntax is heavy and makes me pull my hair. Use your preferred worksheet or database program instead. It will be more efficient and less frustrating.

### Cleaning up / correcting data

Another operation that can be frustrating in R. My advice: unless you want to keep track of all corrections made (so that you can go back to the original data), do not change data in R. Return to the original data file (in a worksheet or database), correct the data there, and then reimport into R. It is simple to resubmit the few lines of code to reimport data. Doing things this way will leave you with a single version of your data file that has all corrections, and the code that allows you to repeat the analysis exactly.

### Importing data from the clipboard

To import a rectangular block of data into R, select the data block and copy it to the clipboard. (**Ctrl-C** in Windows) then, in the R console window, enter the command:

```
myDF<-read.delim("clipboard")
```

to create the myDF data frame from the data contained in the clipboard.

### Example for clipboard data importation

1. Open the text file `simulies.orford.libby.txt`. Select all of its contents and copy it to the clipboard.
2. Enter the following commands in the R console window (ignore the comment lines starting with #)

```
# Import clipboard data into a data.frame called simulies.
simulies<-read.delim("clipboard")

# Check what has been imported
ls.str()

# List the contents of the simulies data.frame
simulies
```

You should get :

```
simulies : 'data.frame':        615 obs. of  4 variables:
 $ Site  : Factor w/ 2 levels "Lovering","Orford": 2 2 2 2 2 2 2 2 2 2 ...
 $ Sp    : Factor w/ 1 level "mixtum/fuscum": 1 1 1 1 1 1 1 1 1 1 ...
 $ Days  : int  78 78 78 78 78 78 78 78 78 78 ...
```

```
 $ Length: num  4.1 2.8 2.2 3.2 3.2 3.1 3.1 4.1 4.1 3.2 ...

        Site            Sp Days Length
1     Orford mixtum/fuscum   78    4.1
2     Orford mixtum/fuscum   78    2.8
3     Orford mixtum/fuscum   78    2.2
4     Orford mixtum/fuscum   78    3.2
5     Orford mixtum/fuscum …
```

(... hundreds of lines of data ...)

```
…   Lovering mixtum/fuscum   78    0.7
613 Lovering mixtum/fuscum   78    0.7
614 Lovering mixtum/fuscum   78    0.7
615 Lovering mixtum/fuscum   78    0.8
```

### Import data from Excel.

Save the data from Excel in a .csv format (**File>Save as**).

Import these data into R with the command read.csv.

For example, to create a data frame named age from a csv file on disk, you can type the command: :

```
>
age<-read.csv(file.choose())
```

(If you do this from the file called age.csv that I exported from the Excel file age.xls, and then type age to list the data frame contents, you will get:

```
> age
   ageclass    sex count
1       0-9 female 17619
2       0-9   male 17538
3     10-19 female 17947
4     10-19   male 18207
5     20-29 female 21344
6     20-29   male 21401
7     30-39 female 19138
8     30-39   male 18837
9     40-49 female 13135
10    40-49   male 12568
11    50-59 female 11617
12    50-59   male 10661
13    60-69 female 11053
14    60-69   male  9374
15    70-79 female  7712
16    70-79   male  5348
17      80+ female  4114
18      80+   male  1926
```

**Trap** : Be careful if you use a language where the comma is used as a decimal point. By default, R uses decimal points and you will not get the result you expect. There is a modified version of read.csv() called `read.csv2()` that fixes this problem.

### Exporting data from R.

```
write.csv(mydata, file = "outfilename.csv", row.names =
FALSE)
```

where `mydata` is the data frame name to be exported and `outfilename.csv` is the name of the file to be exported. Note that this file will be created in the working directory (which can be changed using the menu **File>Change dir**, or by the command setwd())

## Preliminary examination of data

The first step of data analysis is to examine the data at had. This examination will tell you if the data were correctly imported, whether the numbers are credible, whether all data came in, etc. This initial data examination often will allow you to detect unlikely observations, possibly due to errors at the data entry stage. Finally, the initial plotting of the data will allow you to visualize the major trends that will be confirmed later by your statistical analysis.

The file `sturgeon.Rdata` contains data on sturgeons from the Saskatchewan River. These data were collected to examine how sturgeon size varies among sexes (`sex`), sites (`location`), and years (`year`).

⌨ Make sure you start from a blank slate by removing all objects in memory by typing the command

```
rm(list=ls())
```

⌨ Open the file `sturgeon.Rdata`.

⌨ To get a list of the content of the data file loaded in memory, type the command

ls.str() :

```
sturgeon : 'data.frame':        186 obs. of  9 variables:
 $ fklngth : num  37 50.2 28.9 50.2 45.6 ...
 $ totlngth: num  40.7 54.1 31.3 53.1 49.5 ...
 $ drlngth : num  23.6 31.5 17.3 32.3 32.1 ...
 $ rdwght  : num  15.95 NA 6.49 NA 29.92 ...
 $ age     : num  11 24 7 23 20 23 20 7 23 19 ...
 $ girth   : num  40.5 53.5 31 52.5 50 54.2 48 28.5 44 39 ...
 $ sex     : Factor w/ 2 levels "FEMALE","MALE": 2 1 2 1 2 1 1 2 2 2 ...
 $ location: Factor w/ 2 levels "CUMBERLAND","THE_PAS": 2 2 2 2 2 2 2 2 2 2
...
 $ year    : Factor w/ 3 levels "1978","1979",..: 1 1 1 1 1 1 1 1 1 1 ...
```

### Summary statistics

⌨ To get summary statistics on the contents of the data frame sturgeon, type the command:

```
summary(sturgeon)
```

```
    fklngth         totlngth        drlngth
 Min.   :24.96   Min.   :28.15   Min.   :14.33
 1st Qu.:41.00   1st Qu.:43.66   1st Qu.:25.00
```

```
 Median :44.06   Median :47.32   Median :27.00
 Mean   :44.15   Mean   :47.45   Mean   :27.29
 3rd Qu.:48.00   3rd Qu.:51.97   3rd Qu.:29.72
 Max.   :66.85   Max.   :72.05   Max.   :41.93
                 NA's   :85.00   NA's   :13.00
     rdwght           age           girth
 Min.   : 4.73   Min.   : 7.00   Min.   :11.50
 1st Qu.:18.09   1st Qu.:17.00   1st Qu.:40.00
 Median :23.10   Median :20.00   Median :44.00
 Mean   :24.87   Mean   :20.24   Mean   :44.33
 3rd Qu.:30.27   3rd Qu.:23.50   3rd Qu.:48.80
 Max.   :93.72   Max.   :55.00   Max.   :73.70
 NA's   : 4.00   NA's   :11.00   NA's   :85.00
     sex            location     year
 FEMALE:106   CUMBERLAND: 85   1978:45
 MALE  : 80   THE_PAS   :101   1979:68
                                1980:73
```

For each variable, R lists the minimum, the maximum, the median that is the 50$^{th}$ percentile, here the 93$^{rd}$ value of the 186 observations ordered in ascending order, values at the first (25%) and third quartile (75%), and the number of missing values in the column.

Note that several variables have missing values (NA). Only the variables `fklngth` (fork length), `sex`, `location`, and `year` have 186 observations.

**Trap** : Beware of missing values. Several R functions are sensitive to missing values and you will frequently have to do your analyses on data subsets without missing data, or by using optional parameters in various commands. We will get back to this, but you should always pay attention and take note of missing data when you do analyses.

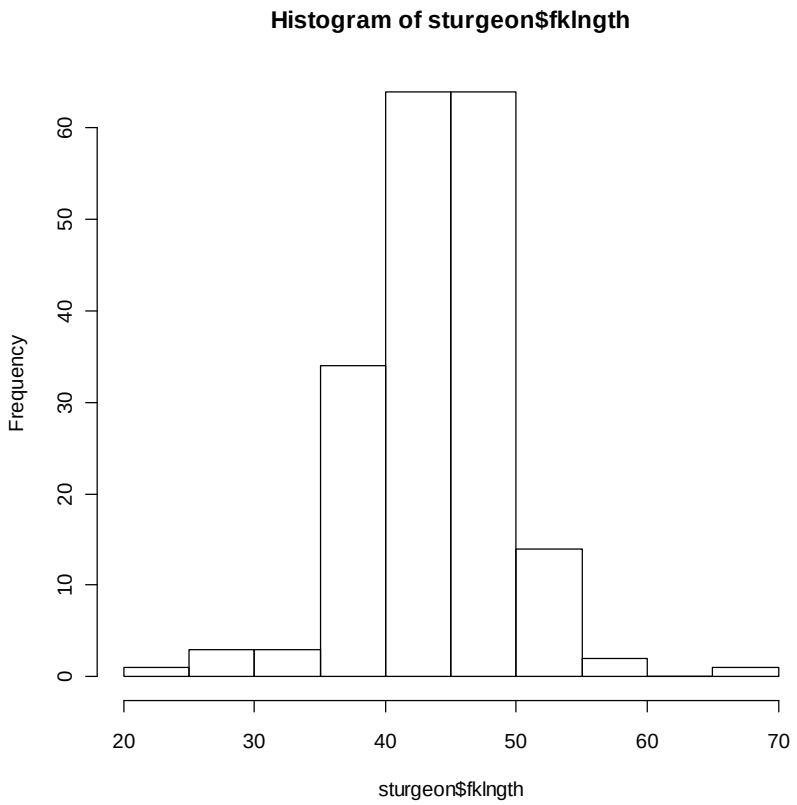
### Histogram, empirical probability density, boxplot, and visual assessment of normality.

Let's look more closely at the distribution of `fklngth`.

The command `hist()` will create a histogram. For the histogram of fklngth in the sturgeon data frame, type the command:

```
hist(sturgeon$fklngth)
```

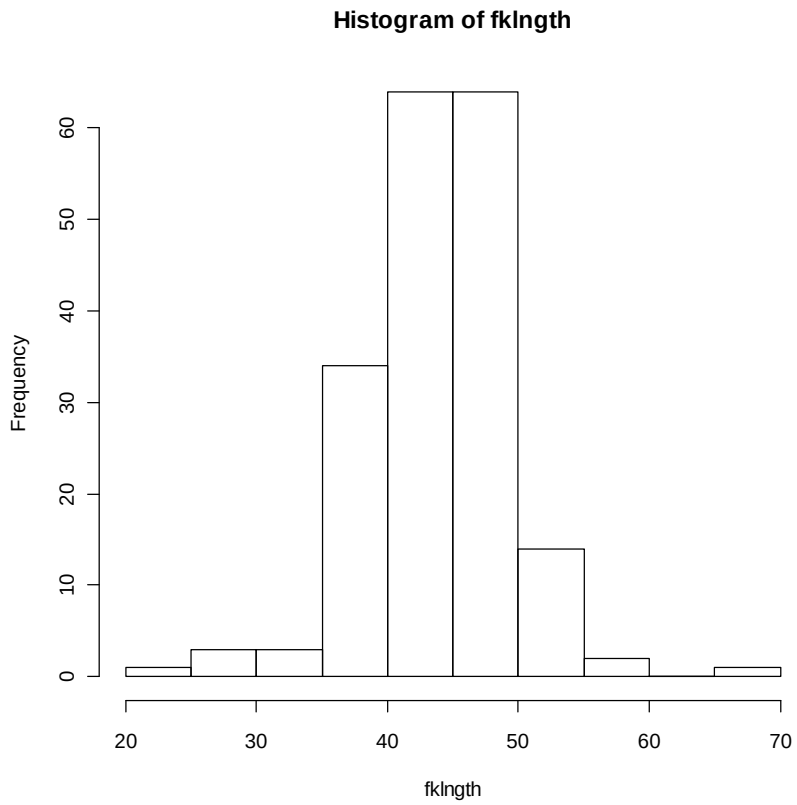
**Figure 2.**




The data appear to be approximately normal. This is good to know. Note that this syntax is a bit heavy as you need to prefix variable names by the data frame name. You can lighten the syntax by making the variables directly accessible by commands by typing the command `attach(sturgeon)`

```
attach(sturgeon)
hist(fklngth)
```

**Figure 3.**



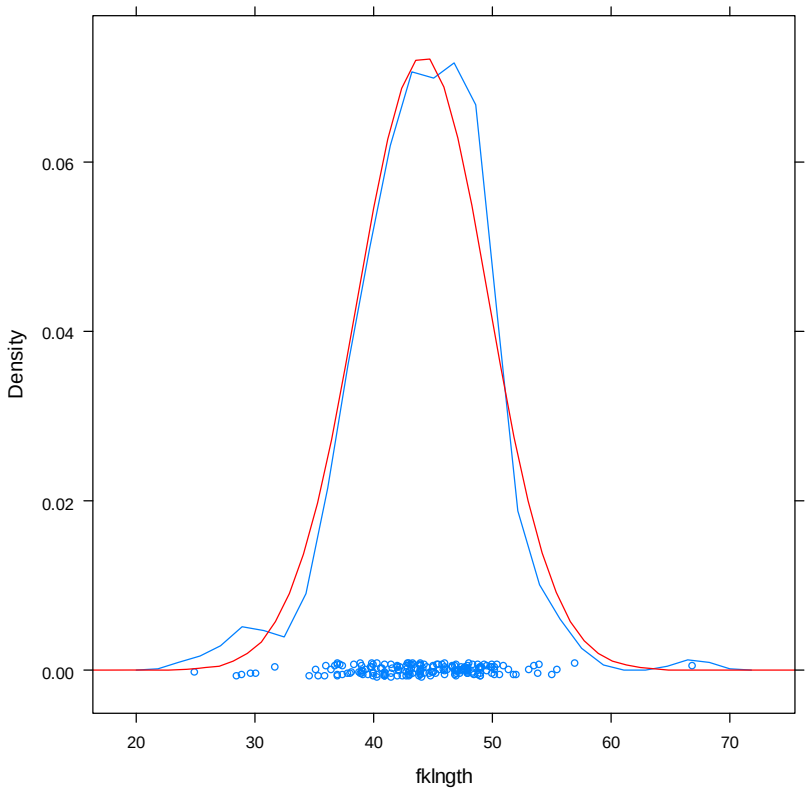

This histogram (Fig. 3) is a very classical representation of the distribution. Histograms are not perfect however because their shape partly depends on the number of bins used, more so for small samples. One can do better, especially if you want to visually compare the observed distribution to a normal distribution. But you need to come up with a bit of extra R code (or be adept at cut and paste...)

I suggest you cut and paste the following code in a new script window (**File->New script**, or **Ctrl-n** in Windows)and then execute it:
:

```
# Distribution des données (stripchart) avec densité de
probabilité empirique et distribution normale (en rouge)
superposée

require(lattice)
densityplot(fklngth,
      panel=function(x,...){
            panel.densityplot(na.omit(x))
            panel.mathdensity(dmath=dnorm,
args=list(mean=mean(na.omit(x)), sd=sd(na.omit(x))),
            col="red")
      }
)
```

You should get :

**Figure 4.**

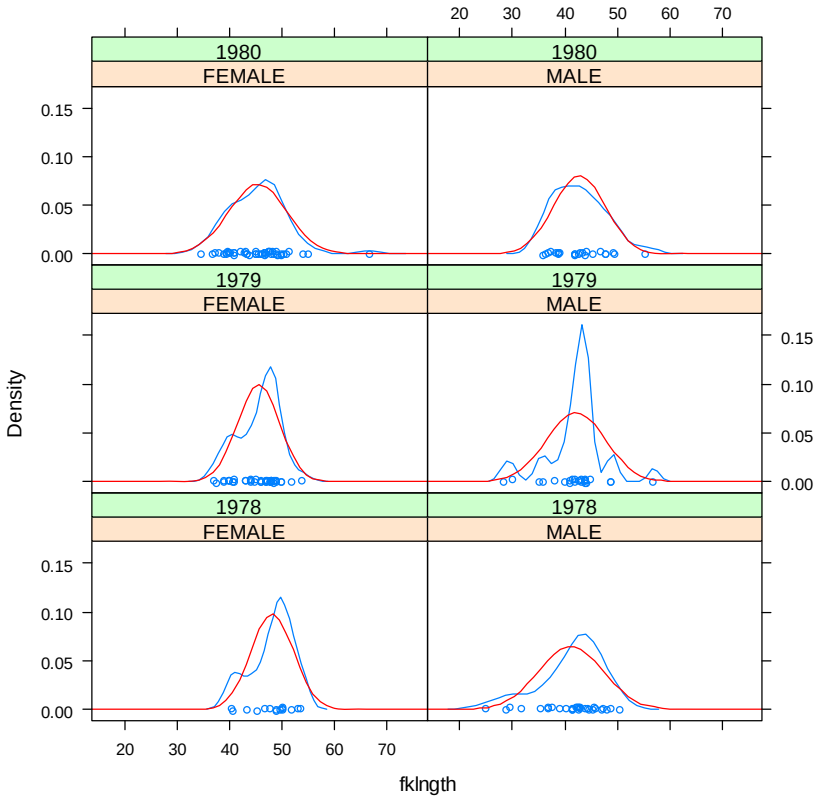



Each observation is represented by an open circle. To better see superimposed points, a small vertical jitter is applied. The red line is the normal distribution with the same mean and standard deviation as the data. The other line, in blue, is the empirical distribution, smoothed from the observations.

If you feel a bit more adventurous, you can plot the distribution of fklngth per sex and year groups with:
:

```
# Distribution des données de fklngth par sex et year avec
densité de probabilité empirique et distribution normale (en
rouge) superposée

densityplot(~fklngth|sex+year,
      panel=function(x,...){
            panel.densityplot(na.omit(x))
            panel.mathdensity(dmath=dnorm,
args=list(mean=mean(na.omit(x)), sd=sd(na.omit(x))),
            col="red")
      }
)
```

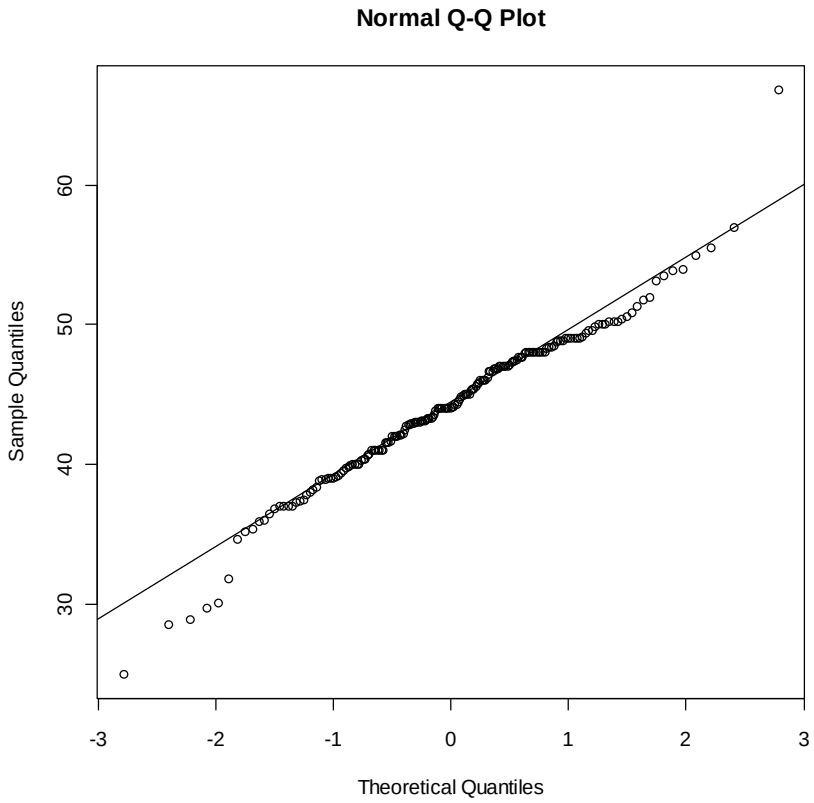
**Figure 5.**




Another way to visually assess normality of data is the QQ plot that is obtained by the pair of commands `qqnorm()` and `qqline()`,

```
qqnorm(fklngth)
qqline(fklngth)
```

**Figure 6.**




Perfectly normal data would follow the straight diagonal line. Here there are deviations in the tails of the distribution and a bit to the right of the center. Compare this representation to the two preceding graphs. You will probably agree that it is easier to visualize how data deviate from normality by looking at a histogram of an empirical probability density than by looking at the QQ plots. However, QQ plots are often automatically produced by various statistical routines and you should be able to interpret them. In addition, one can easily run a formal test of normality in R with the command shapiro.test() that computes a statistic (W) that measures how tightly data fall around the straight diagonal line of the QQ plot. If data fall perfectly on the line, then W=1. If W is much less than 1, then data are not normal.

For the `fklngth` data :

```
shapiro.test(fklngth)
```

```
        Shapiro-Wilk normality test

data:  fklngth
W = 0.9722, p-value = 0.0009284
```

W is close to 1, but far enough to indicate a statistically significant deviation from normality).

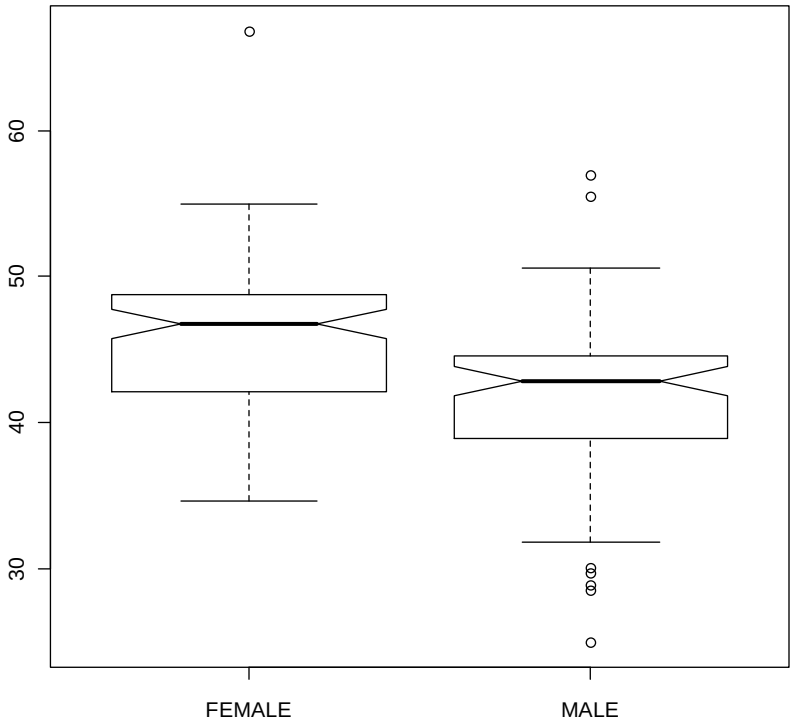
Visual examination of very large data sets is often made difficult by the superposition of data points. Boxplots are an interesting alternative.

The command

```
boxplot(fklngth~sex, notch=TRUE)
```

produces a boxplot of fklngth for each sex , and adds whiskers..

**Figure 7.**




The slightly thicker line inside the box of Fig. 7 indicates the median. The width of the notch is proportional to the uncertainty around the median estimate. One can visually assess the approximate statistical significance of differences among medians by looking at the overlap of the whiskers (here there is no overlap and one could tentatively conclude that the median female size is larger than the median male size). Boxes extend from the first to third quartile (the 25th to 75th percentile if you prefer). Bars (whiskers) extend above and below the

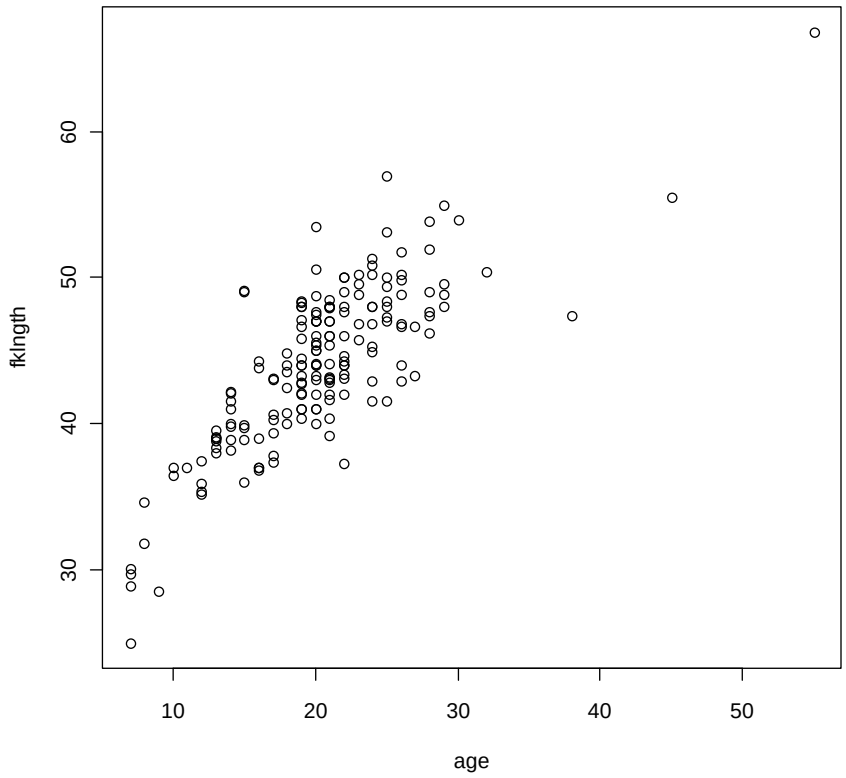
boxes from the minimum to the maximum observed value or, if there are extreme values, from the smallest to the largest observed value within 1.5x the interquartile range from the median. Observations exceeding the limits of the whiskers (hence further away from the median than 1.5x the interquartile range, the range between the 25th and 75th percentile) are plotted as circles. These are outliers, possibly aberrant data.

## Scatterplots

In addition to histograms and other univariate plots, it is often informative to examine scatter plots. The command `plot(y~x)` produces a scatter plot of y on the vertical axis (the ordinate) vs x on the horizontal axis (abscissa).

⌨ Create a scatterplot of fklngth vs age using the plot() command. You should obtain:

**Figure 8.**



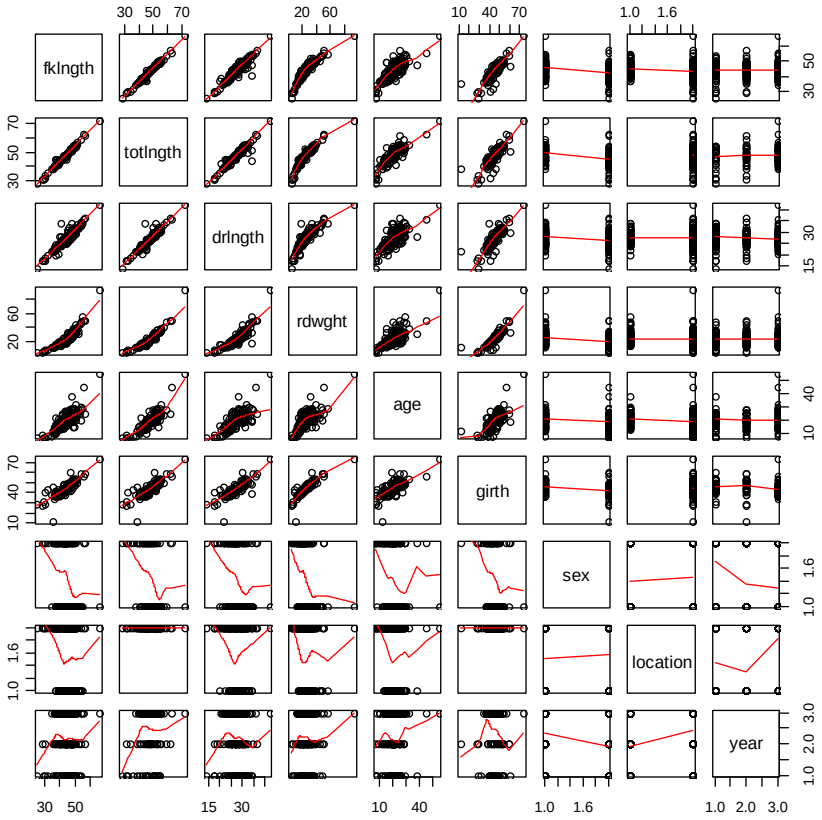
R has a function to create all pairwise scatterplots rapidly called `pairs()`. One of `pairs()` options is the addition of a lowess trace on each plot to that is a smoothed trend in the data.

To get the plot matrix with the lowess smooth for all variable sin the sturgeon data frame, enter the command

```
pairs(sturgeon, panel=panel.smooth)
```

and you should get

**Figure 9.**




## Creating data subsets

You will frequently want to do analyses on some subset of your data.. The command subset() is what you need to isolate cases meeting some criteria. For example, to create a subset of the sturgeon data frame that contains only females caught in 1978, you could write: :

```
> sturgeon.female.1978<-subset(sturgeon,sex=="FEMALE" &
year=="1978")
> sturgeon.female.1978
```

```
     fklngth totlngth  drlngth rdwght age girth
2   50.19685 54.13386 31.49606     NA  24  53.5
4   50.19685 53.14961 32.28346     NA  23  52.5
```

```
6   49.60630 53.93701 31.10236  35.86  23  54.2
7   47.71654 51.37795 33.97638  33.88  20  48.0
15  48.89764 53.93701 29.92126  35.86  23  52.5
105 46.85039       NA 28.34646  23.90  24    NA
106 40.74803       NA 24.80315  17.50  18    NA
107 40.35433       NA 25.59055  20.90  21    NA
109 43.30709       NA 27.95276  24.10  19    NA
113 53.54331       NA 33.85827  48.90  20    NA
114 51.77165       NA 31.49606  35.30  26    NA
116 45.27559       NA 26.57480  23.70  24    NA
118 53.14961       NA 32.67717  45.30  25    NA
119 50.19685       NA 32.08661  33.90  26    NA
123 49.01575       NA 29.13386  37.50  22    NA
       sex   location year
2   FEMALE    THE_PAS 1978
4   FEMALE    THE_PAS 1978
6   FEMALE    THE_PAS 1978
7   FEMALE    THE_PAS 1978
15  FEMALE    THE_PAS 1978
105 FEMALE CUMBERLAND 1978
106 FEMALE CUMBERLAND 1978
107 FEMALE CUMBERLAND 1978
109 FEMALE CUMBERLAND 1978
113 FEMALE CUMBERLAND 1978
114 FEMALE CUMBERLAND 1978
116 FEMALE CUMBERLAND 1978
118 FEMALE CUMBERLAND 1978
119 FEMALE CUMBERLAND 1978
123 FEMALE CUMBERLAND 1978
```

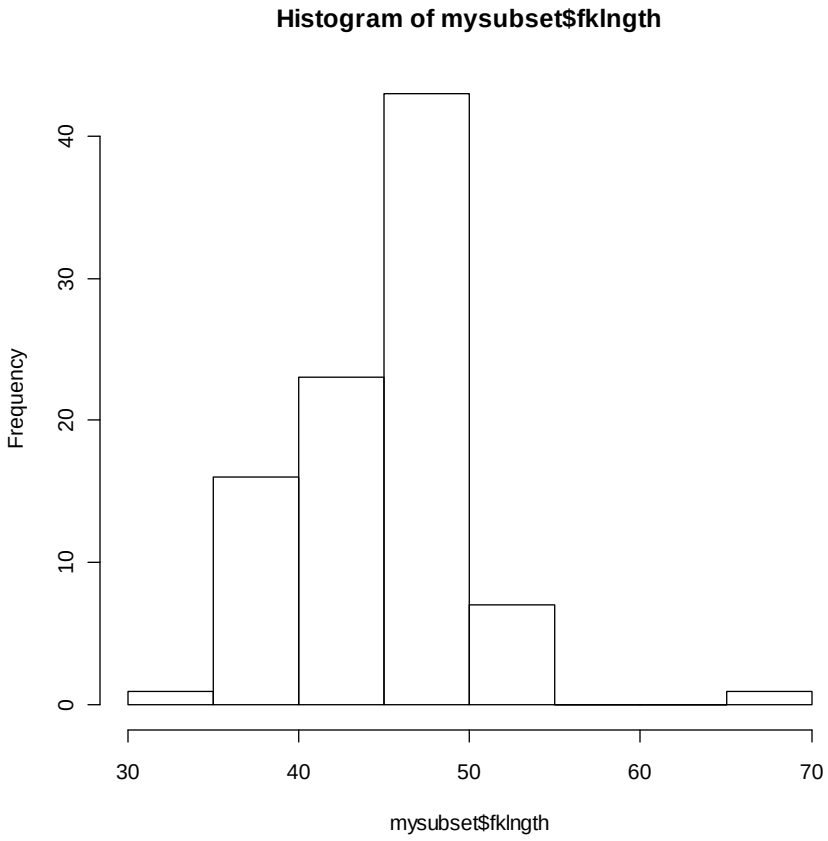
**Trap**: When using criteria to select cases, be careful of the == syntax to mean equal to. In this context, if you use a single =, you will not get what you want.

The following table lists the most common criteria to create expressions and their R syntax.

| Operator | Meaning | Operator | Meaning |
|---|---|---|---|
| == | Equal to | != | Not equal to |
| > | Larger than | < | Less than |
| >= | Larger than or equal | <= | Less than or equal to |
| & | And | \| | Or |

⌨ Using the commands `subset()` and `hist()`, create a histogram for females caught in 1979 and 1980 (hint: sex=="FEMALE" & (year =="1979" | year=="1980"))

**Figure 10.**




## Data transformation

You will frequently transform raw data to better satisfy assumptions of statistical tests. R will allow you to do that easily..

The most used functions are probably:

log10()

sqrt()

ifelse()

You can use these functions directly within commands, create vector variables, or add columns in data frames.

To do a plot of the decimal log of fklngth vs age, you can simply use the log10 function within the plot command:

```
plot(log10(fklngth)~age)
```

To create a vector variable, an orphan variable if you wish, one that is not part of a data frame, called `lfklngth` and corresponding too the decimal log of `fklngth`, simply enter:

```
lfklngth<-log10(fklngth)
```

If you want this new variable to be added to a data frame, then you must prefix the variable name by the data frame name and the $ symbol. For example to add the variable `lfkl` containing the decimal log of `fklngth` to the sturgeon data frame, enter::

```
sturgeon$lfkl<-log10(fklngth)
```

`lfkl` will be added to the data frame sturgeon for the R session. Do not forget to save the modified data frame if you want to keep the modified version.

For conditional transformations, you can use the function `ifelse()`. For example, to create a new variable called dummy with a value of 1 for males and 0 for females, you can use:

:

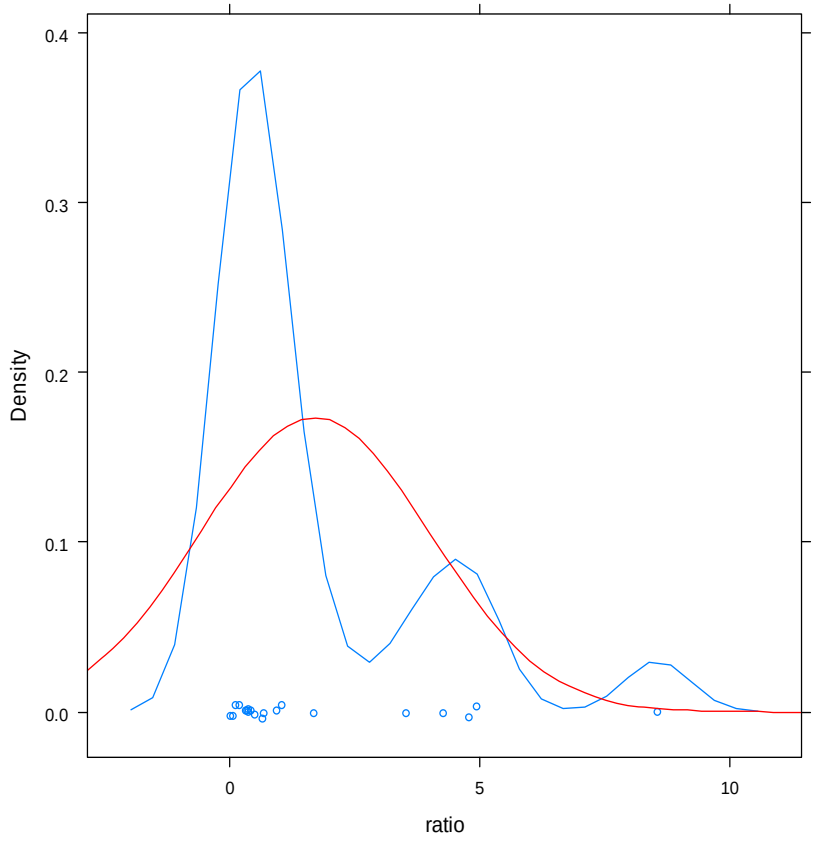
```
dummy<-ifelse(sex=="MALE",1,0)
```

## Exercise

The file salmonella.Rdata contains numerical values for the variable called ratio for two environments (milieux: IN VITRO or IN VIVO) and for 3 strains (souches).

Examine the ratio variable and make a graph to visually assess normality for the SAUVAGE strain.

**Figure 11.**




# Commented R code for this laboratory exercise

```
# Exercise - Introduction to R for BIO4158 Applied Biostatistics
# Exercice - Introduction à R pour BIO4558 Biostatistiques Appliquées
# Auteur: Antoine Morin (amorin@uottawa.ca)
#  Dependencies: lattice


#################################################################
# Commands for importing and exporting data
# Commandes pour importer et exporter des données

# Read R data file ErablesGatineau.Rdata
# Lire le fichier de données R appelé ErablesGatineau.Rdata

#  load(file.choose()) # local version
load(url("http://www.antoinemorin.com/biostats/2010/Erables-
Gatineau.Rdata")) # online version

# Lister objets en mémoire
# List memory objects

ls()

# Décrire objets en mémoire
# Describe memory objects

ls.str()
```

```
# Lire données dans le presse-papier
# Read data from the clipboard

myDF <- read.delim("clipboard")

# Lire données exportées d'Excel dans un fichier de format csv
# Read data exported from Excel into a text file in csv format

age <- read.csv(file.choose())

# Save dataframe to a R data file
# Sauvegarder un dataframe dans un fichier de données R

save(age, file = "age.Rdata")

############################################################
# Examen des données du fichier sturgeon.Rdata
# Examination of the data of the file sturgeon.Rdata

# Effacer toutes les données en mémoire pour partir à neuf
# Erase all objects in memory to start with a clean slate

rm(list = ls())


# Charger le fichier en utilisant une boîte de dialogue, choisir stur-
geon.Rdata
# Load a datafile using the dialog box, choose sturgeon.Rdata

load(file.choose()) # local version
#  load(url("http://www.antoinemorin.com/biostats/2010/sturgeon.Rdata")) #
online version


# Obtenir une liste décrivant ce qui est en mémoire après le chargement du
fichier
# Get a list of what has been loaded to memory after opening the file

ls.str()

# Obtenir un sommaire des variables dans le data.frame sturgeon
# Get a summary of the contents of the sturgeon data.frame

summary(sturgeon)

# Faire un histogramme de la variable fklngth dans le data.frame sturgeon
# Make a histogram of the variable fklngth in the sturgeon data.frame in mem-
ory

hist(sturgeon$fklngth)

# Rendre les variables contenues dans le data.frame sturgeon directement
disponibles
# Make variables in the sturgeon data.frame directly callable

attach(sturgeon)

# Refaire le même histogramme que précédemment
# Redo the same histogram as before
ls()
hist(fklngth)

# Distribution des données (stripchart) avec densité de probabilité empirique
# et distribution normale (en rouge) superposée
# Data distribution (stripchart) with empirical probability density
# and superimposed normal distribution (in red)

require(lattice)
densityplot(fklngth, panel = function(x, ...) {
    panel.densityplot(na.omit(x))
    panel.mathdensity(dmath = dnorm, args = list(mean = mean(na.omit(x)),
        sd = sd(na.omit(x))), col = "red")
})

# Distribution des données sur fklngth par sex et year
# Data distribution fklngth by subsets of sex and year
```

```
densityplot(~fklngth | sex + year, panel = function(x,
    ...) {
    panel.densityplot(na.omit(x))
    panel.mathdensity(dmath = dnorm, args = list(mean = mean(na.omit(x)),
        sd = sd(na.omit(x))), col = "red")
})

# QQ plot de fklngth

qqnorm(fklngth)
qqline(fklngth)

# Wilks-Shapiro test of normality on fklngth

shapiro.test(fklngth)

# Boxplot of fklngth by sex, with whiskers

boxplot(fklngth ~ sex, notch = TRUE)


# Diagramme de dispersion de fklngth en fonction de l'age
# Scatterplot of fklngth as a function of age

plot(fklngth ~ age)


# Matrice de diagramme de dispersion bivariés avec lowess
# Matrix of scatterplots of all pairs of variables in dataframe sturgeon, with
lowess trace

pairs(sturgeon, panel = panel.smooth)

# Créer un sous-ensemble des données pour les femelles capturées en 1978
# Create a subset with only females captured in 1978
sturgeon.female.1978 <- subset(sturgeon, sex == "FEMALE" &
    year == "1978")

# Créer un histogramme de fklngth pour le sous ensemble des données  pour les
femelles en 1979 et 1980
# Create a histogram of fklngth from a subset for females in 1979 and 1980

mysubset = subset(sturgeon, sex == "FEMALE" & (year ==
    "1979" | year == "1980"))
hist(mysubset$fklngth)
```

# Lab - Power analysis with G*Power

After completing this laboratory, you should

- be able to compute the power of a t-test with G*Power
- be able to calculate the required sample size to achieve a desired power level with a t-test
- be able to calculate the detectable effect size by a t-test given the sample size, the power and $\alpha$
- understand how power changes when sample size increases, the effect size changes, or when $\alpha$ decreases
- understand how power is affected when you change from a two-tailed to a one-tailed test.

## The theory

### What is power?

*Power is the probability of rejecting the null hypothesis when it is false.*

### Why do power analyses?

**Assess strength of the evidence.** Power analysis, performed after accepting a null hypothesis, can help assess the probability of rejecting the null if it were false, and if the magnitude of the effect was equal to that observed (or to any other given magnitude). This type of *a posteriori* analysis is very common.

**Design better experiments.** Power analysis, performed prior to conduct an experiment (but most often *after* a preliminary experiment), can be used to determine the number of observations required to detect an effect of a given magnitude with some probability (the power). This type of *a priori* experiment should be more common.

**Estimate minimum detectable effect.** Sampling effort is often pre-determined (when you are handed data of an experiment already completed), or extremely constrained (when logistics dictates what can be done). Whether it is *a priori* or *a posteriori*, power analysis can help you estimate, for a fixed sample size and a given power, what is the minimum effect size that can be detected.

### Factors affecting power

For a given statistical test, there are 3 factors that affect power:

**Decision criteria.** Power is related to $\alpha$, the probability level at which one rejects the null hypothesis. If this decision criteria is made very strict (i.e. if critical $\alpha$ is set to a very low value, like 0.1% or $p$=0.001), then power will be lower than if the critical $\alpha$ was less strict.

**Sample size.** The larger the sample size, the larger the power. As sample size increases, one's ability to detect small effect sizes as being statistically significant gets better.

**Effect size.** The larger the effect size, the larger the power. For a given sample size, the ability to detect an effect as being significant is higher for large effects than for small ones. Effect size measures how false the null hypothesis is.

## What is G*Power?

G*Power is free software developed by quantitative psychologists from the University of Dusseldorf in Germany that is available at: http://www.psycho.uni-duesseldorf.de/abteilungen/aap/gpower3/. It is available in MacOS and Windows versions.

G*Power will allow you to do power analyses for the majority of statistical tests we will cover during the term without making lengthy calculations and looking up long tables and figures of power curves. It is a really useful tool that you need to master.

⌨ Download the software and install it on your computer and your workstation (if it is not there already).

## How to use G*Power

### General principle

Using G*Power generally involves 3 steps:

1. Choosing the appropriate test
2. Choosing one of the 5 types of available power analyses
3. Enter parameter values and press the **Calculate** button

## Types of power analyses

**A priori.** Computes the sample size required given α, β, and the effect size. This type of analysis is useful when planning experiments.

**Compromise.** Computes α and β for a given β/α ratio, sample size, and effect size. Less commonly used (I have never used it myself) although it can be useful when theβ/α ratio has meaning, for example when the cost of type I and type II errors can be quantified.

**Criterion.** Computes α for a givenβ, sample size, and effect size. In practice, I see little interest in this. Let me know if you see something I don't!

**Post-hoc.** Computes the power for a given α, effect size, and sample size. Used frequently to help in the interpretation of a test that is not statistically significant, but only if an effect size that is biologically significant is used (and not the observed effect size). Not relevant when the test is significant.

**Sensitivity.** Computes the detectable effect size for a given α, β, and sample size. Very useful at the planning stage of an experiment.

## How to calculate effect size

G*Power can perform power analyses for several statistical tests. The metric for effect size depends on the test. Note that other software packages often use different effect size metrics and that it is important to use the correct one for each package. G*Power has an effect size calculator for many tests that only requires you to enter the relevant values. The following table lists the effect size metrics used by G*Power for the various tests.

.

| Test | Index of effect size | Calculation |
|---|---|---|
| t-test on Means | d | $d = \frac{\lvert\mu_1 - \mu_2\rvert}{\sigma}$ |
| t-test on Correlations | r | |
| Other t-tests | f | $f=\mu_1/\sigma$ |

| | | |
|---|---|---|
| F-test (ANOVA) | f | $f = \frac{\sqrt{\frac{\sum_{i=1}^{k}(\mu_i - \mu)^2}{k}}}{\sigma}$ |
| Other F-tests | $f^2$ | $f^2 = \frac{R_p^2}{1 - R_p^2}$ where $R_p^2$ is the squared partial correlation coefficient |
| Chi-Square test | w | $w = \sqrt{\sum_{i=1}^{m}\frac{(p_{0i} - p_{1i})^2}{p_{0i}}}$ where $p_{0i}$ is the proportion in category i predicted by the null (0) and alternative (1) hypothesis |

## Power analysis for a t-test on two independent means

The objective of this lab is to learn to use G*Power and understand how the 4 parameters of power analyses (α, β, sample size and effect size) are related to each other. For this, you will only use the standard t-test to compare two independent means. This is the test most used by biologists, you have all used it, and it will serve admirably for this lab. What you will learn today will be applicable to all other power analyses.

Jaynie Stephenson studied the productivity of streams in the Ottawa region. She has measured fish biomass in 36 streams, 18 on the Shield and 18 in the Ottawa Valley. She found that fish biomass was lower in streams from the valley (2.64 g/m$^2$, standard deviation=3.28) than from the Shield (3.31 g/m$^2$, standard deviation=2.79.). When she tested the null hypothesis that fish biomass is the same in the two regions by a t-test, she obtained:

```
Pooled-Variance Two-Sample t-Test

t = -0.5746, df = 34, p-value = 0.5693
```

She therefore accepted the null hypothesis (since p is much larger than 0.05) and concluded that fish biomass is the same in the two regions.

### Post-hoc analysis

Using the observed means and standard deviations, we can use G*Power to calculate the power of the two-tailed t-test for two independent means, using the observed effect size (the difference between the two means, weighted by the standard deviations) for $\alpha$=0.05.

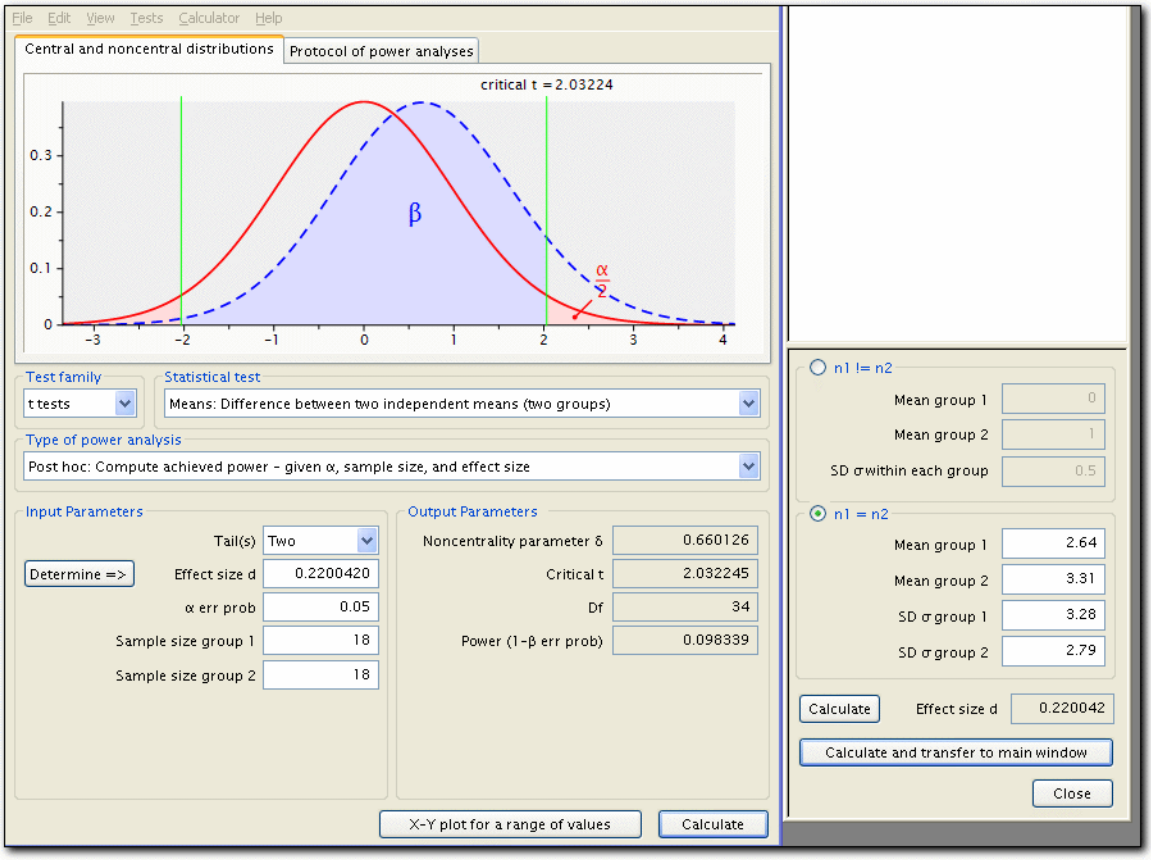
⌨Start G*Power.

1. In **Test family**, choose: t tests
2. For **Statistical test**, choose: Means: Difference between two independent means (two groups)
3. For **Type of power analysis**, choose: Post hoc: Compute achieved power - given $\alpha$, sample size, and effect size
4. At **Input Parameters**,
   in the box**Tail(s)**, chose: Two,
   check that $\alpha$ **err prob** is equal to 0.05
   Enter 18 for **Sample size group** 1 et 2

   then, to calculate effect size (d), click on **Determine =>**
5. In the window that opens,

   select **n1 = n2**, then
   enter the two means(**Mean group** 1 et 2)
   the two standard deviations(**SD s group** 1 et 2)
   and click on **Calculate and transfer to main window**
6. After you click on the Calculate button in the main window, you should get the following:

**Figure 12.**




Let's examine this plot.

- The curve on the left, in red, corresponds to the expected distribution of the t-statistics when $H_0$ is true (i.e. when the two means are equal) given the sample size (18 per region) and the observed standard deviations.
- The vertical green lines correspond to the critical values of t for $\alpha$=0.05 and a total sample size of 36 (2x18).
- The shaded pink regions correspond to the rejection zones for $H_0$. If Jaynie had obtained a t-value outside the interval delimited by the critical values ranging from -2.03224 to 2.03224, she would then have rejected $H_0$, the null hypothesis of equal means. In fact, she obtained a t-value of -0.5746 and concluded that the biomass is equal in the two regions.
- The curve on the right, in blue, corresponds to the expected distribution of the t-statistics if $H_1$ is true (here $H_1$ is that there is a difference in biomass between the two regions equal to 3.33-2.64=0.69g/$m^2$, given the observed standard deviations). This distribution is what we should observe if $H_1$ was true and that we repeated a large number of times the experiment using random

samples of 18 streams in each of the two regions and calculated a t-statistic for each sample. On average, we would obtain a t-statistic of about 0.6.

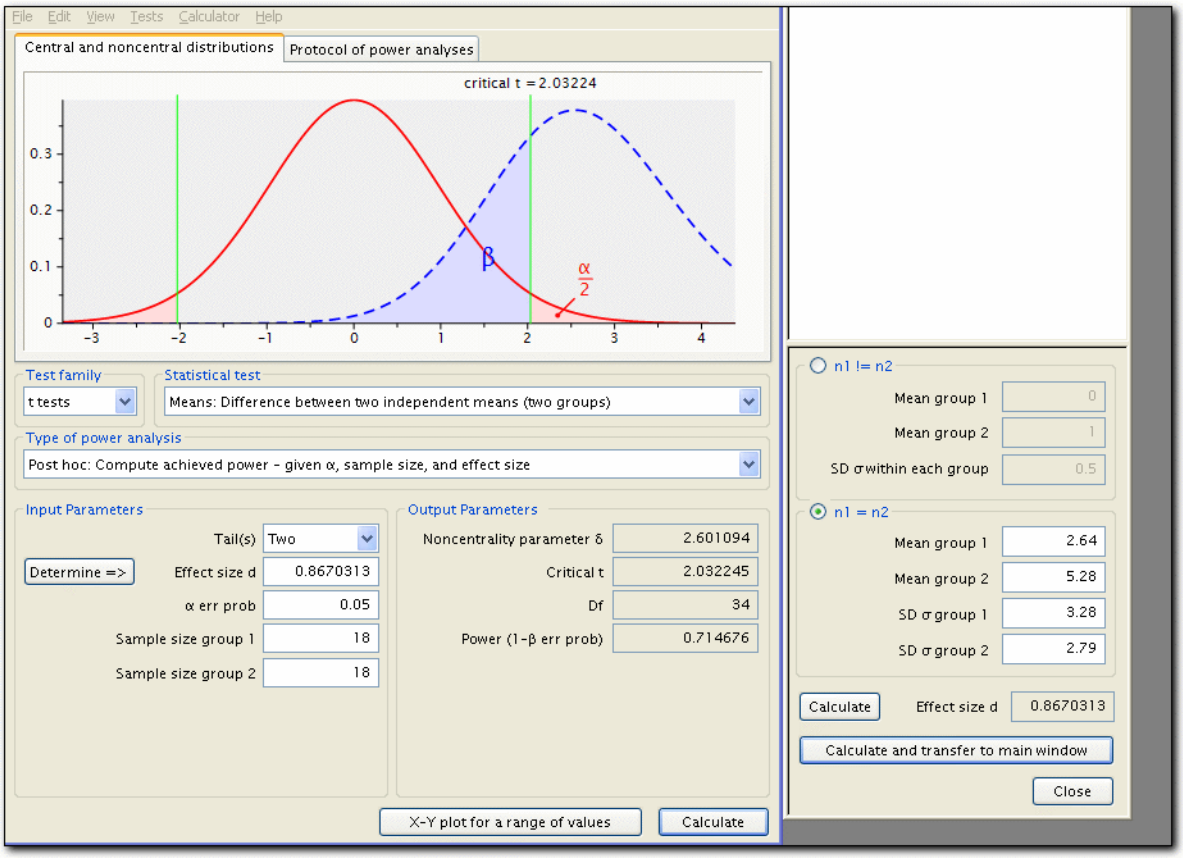
- Note that there is considerable overlap of the two distributions and that a large fraction of the surface under the right curve is within the interval where $H_0$ is accepted between the two vertical green lines at -2.03224 and 2.03224. This proportion, shaded in blue under the distribution on the right is labeled $\beta$ and corresponds to the risk of type II error (accept $H_0$ when $H_1$ is true).
- Power is simply 1-$\beta$, and is here 0.098339. Therefore, if the mean biomass differed by 0.69g/$m^2$ between the two regions, Jaynie had only 9.8% chance of being able to detect it as a statistically significant difference at $\alpha$=5% with a sample size of 18 streams in each region.

Let's recapitulate: The difference in biomass between regions is not statistically significant according to the t-test. It is because the difference is relatively small relative to the precision of the measurements. It is therefore not surprising that that power, i.e. the probability of detecting a statistically significant difference, is small. Therefore, this analysis is not very informative.

**Indeed, a post hoc power analysis using the observed effect size is not useful.** It is much more informative to conduct a post hoc power analysis for an effect size that is **different** from the observed effect size. But what effect size to use? It is the biology of the system under study that will guide you. For example, with respect to fish biomass in streams, one could argue that a two fold change in biomass (say from 2.64 to 5.28 g/$m^2$) has ecologically significant repercussions. We would therefore want to know if Jaynie had a good chance of detecting a difference as large as this before accepting her conclusion that the biomass is the same in the two regions. So, what were the odds that Jaynie could detect a difference of 2.64 g/$m^2$ between the two regions? G*Power can tell you if you cajole it the right way.

Change the mean of group 2 to 5.28, recalculate effect size, and click on Calculate to obtain:

**Figure 13.**




The power is 0.71, therefore Jaynie had a reasonable chance (71%) of detecting a doubling of biomass with 18 streams in each region.

Note that this post hoc power analysis, done for an effect size considered biologically meaningful, is much more informative than the preceeding one done with the observed effect size (which is what too many students do because it is the default of so many power calculation programs). Jaynie did not detect a difference between the two regions. There are two possibilities: 1) there is really no difference between the regions, or ") the precision of measurements is so low (because the sample size is small and/or there is large variability within a region) that it is very unlikely to be able to detect even large differences. The second power analysis can eliminate this second possibility because Jaynie had 71% chances of detecting a doubling of biomass.

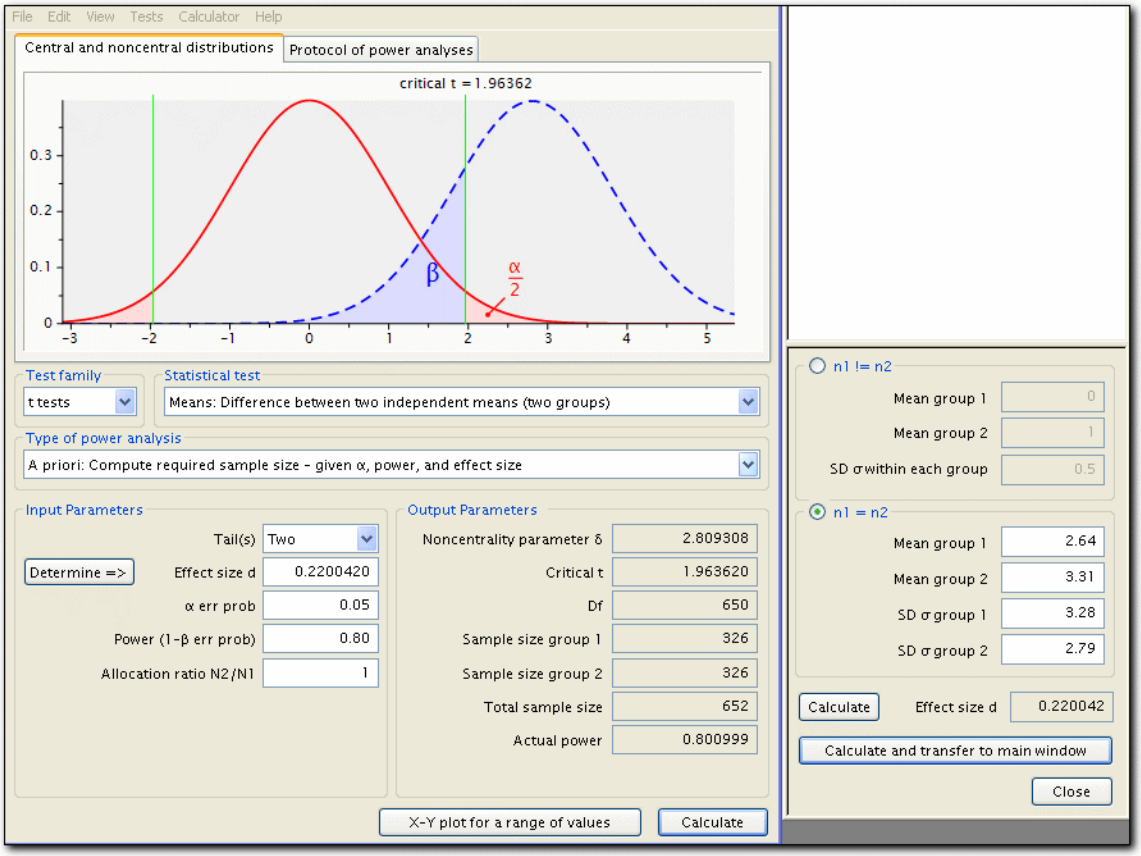
## A priori analysis

Suppose that a difference in biomass of 3.33-2.64=0.69g/$m^2$ can be ecologically significant. The next field season should be planned so that Jaynie would have a good chance of detecting such a difference in

fish biomass between regions. How many streams should Jaynie sample in each region to have 80% of detecting such a difference (given the observed variability)?

⌨ Change the type of power analysis in G*Power to A priori: **Compute sample size - given α, power, and effect size**. Ensure that the values for means and standard deviations are those obtained by Jaynie. Recalculate the effect size metric and enter 0.8 for power and you will obtain:

**Figure 14.**




Ouch! The required sample would be of 326 streams in each region! It would cost a fortune and require several field teams otherwise only a few dozen streams could be sampled over the summer and it would be very unlikely that such a small difference in biomass could be detected. Sampling fewer streams would probably be in vain and could be considered as a waste of effort and time: why do the work on several dozens of streams if the odds of success are that low?

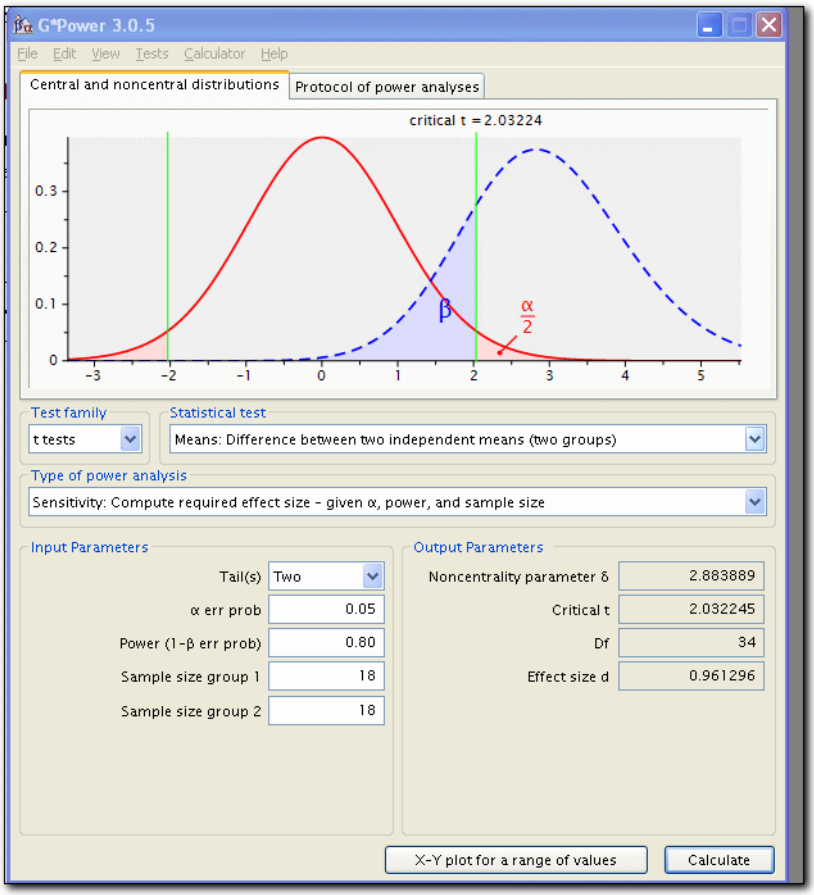
⌨If we recalculate for a power of 95%, we find that 538 streams would be required from each region. Increasing power means more work!

### Sensitivity analysis - Calculate the detectable effect size

Given the observed variability, a sampling effort of 18 streams per region, and with $\alpha$=0.05, what effect size could Jaynie detect with 80% probability ($\beta$=0.2)?

⌨Change analysis type in G*Power to Sensitivity: **Compute required effect size - given α, power, and sample size** and check that sample size is 18 in each region. You should get:

**Figure 15.**




The detectable effect size for this sample size, $\alpha$=0.05 and $\beta$=0.2 (or power of 80%) is 0.961296. **Attention**, this effect size is the metric d and is dependent on sampling variability. Here, d is approximately equal to

$$d = \frac{\left|\bar{X}_1 - \bar{X}_2\right|}{\sqrt{s_1^2 + s_2^2}}$$

To convert this d value without units into a value for the detectable difference in biomass between the two regions, you need to multiply d by the denominator of the equation:

$$\text{detectable difference} = \left|\bar{X}_1 - \bar{X}_2\right| = d\sqrt{s_1^2 + s_2^2}$$

$$\text{detectable difference} = 0.961296\sqrt{3.28^2 + 2.79^2}$$

$$\text{detectable difference} = 4.139437$$

Therefore, with 18 streams per region, $\alpha$=0.05 and $\beta$=0.2 (so power of 80%), Jaynie could detect a difference of 4.14g/m$^2$ between regions, a bit more than a doubling of biomass.

## Important points to remember

- Post hoc power analyses are relevant only when the null hypothesis is accepted because it is impossible to make a type II error when accepting $H_0$.
- With very large samples, power is very high and minute differences can be statistically detected, even if they are not biologically significant.
- When using a stricter significance criteria ($\alpha$<0.05) power is reduced.
- Maximizing power implies more sampling effort, unless you use a more liberal statistical criteria ($\alpha$>0.05)
- The choice of $\beta$ is somewhat arbitrary. $\beta$=0.2 (power of 80%) is considered relatively high by most.

## Assignment

Black fly larvae (Diptera: Simuliidae) were sampled in February at the outlet of two lakes of the Eastern Townships (Lakes Orford and Lovering). The body length of each larva was measured. The data are in the file simulies.orford.libby.txt

The relationship between body length (L, in mm) and dry mass (M, in $\mu$g) for the dominant species (*P. mixtum/fuscum*) is:

$$M = 1.36L^{3.05}$$

1. Calculate the mean mass and its standard deviation at each site.

2. Using the mean mass of *P. mixtum/fuscum* at Lovering as a reference and the standard deviations observed at each site, calculate the power of a two-tailed t test for independent means

   a) if the mass difference was 5 μg, $\alpha$=0.05, and that 100 larvae were sampled at each site
   b) if the mass difference was 20 μg, $\alpha$=0.05, and that 100 larvae were sampled at each site
   c) if the mass difference was 50 μg, $\alpha$=0.05, and that 100 larvae were sampled at each site

3. Calculate the required sample size to detect, with a two-tailed t-test for independent means, a difference of 50 μg between the means

   a) with a power of 80% and $\alpha$=0.05
   b) with a power of 80% and $\alpha$=0.001
   c) with a power of 95% and $\alpha$=0.05

4. Calculate the detectable effect size (d) with a two-tailed t-test on independent means, given the observed standard deviations

   a) with a power of 80%, $\alpha$=0.05, and measurements on 10 larvae at each site
   b) with a power of 80%, $\alpha$=0.05, and measurements on 200 larvae at each site
   c) with a power of 80%, $\alpha$=0.05, and measurements on 20 larvae at one site and 380 larvae at the second site

5. Calculate the difference in average mass, in μg, that can be detected given your estimate of the minimum detectable effect size (d) at 4a, b, and c.

# Lab- Linear correlation and simple linear regression

After completing this laboratory exercise, you should be able to:

- Use R to produce a scatter plot of the relationship between two variables.
- Use R to carry out some simple data transformations.
- Use R to compute the Pearson product-moment correlation between two variables and assess its statistical significance.
- Use R to compute the correlation between pairs of ranked variables using the Spearman rank correlation and Kendall's tau.
- Use R to assess the significance of pairwise comparisons from a generalized correlation matrix using Bonferroni-adjusted probabilities.
- Use R do a simple linear regression
- Use R to test the validity of the assumptions underlying simple linear regression
- Use R to assess significance of a regression by the bootstrap method
- Quantify effect size in simple regression and perform a power analysis using G*Power

## Scatter plots

Correlation and regression analysis should always begin with an examination of the data: this is a critical first step in determining whether such analyses are even appropriate for your data.

Suppose we are interested in the extent to which length of male sturgeon in the vicinity of The Pas and Cumberland House covaries with weight. To address this question, we look at the correlation between `fklngth` and `rdwght`.

Recall that one of the assumptions in correlation analysis is that the relationship between the two variables is linear. To evaluate this assumption, a good first step is to produce a scatterplot:


Using the file `sturgeon.Rdata`, load the sturgeon data frame and attach() it to make variable names directly available in functions. Make

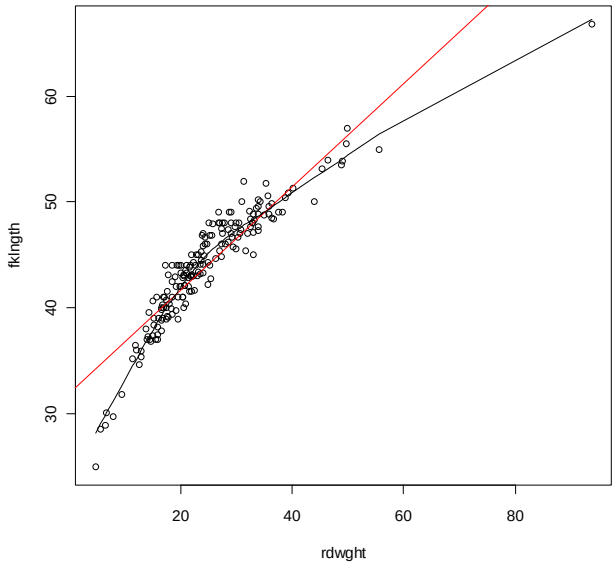
a scatter plot of `fklngth` vs `rdwght` fit with a locally weighted regression (Loess) smoother and a linear regression

```
plot(fklngth~rdwght)
abline(lm(fklngth~rdwght), col="red")
goodcases<-!(is.na(fklngth)|is.na(rdwght))
lines(lowess(fklngth[goodcases]~rdwght[goodcases])).
```

(note here the fancy coding required to eliminate missing values that interfere with the creation of the lowess trace)

⌨ Does this curve suggest a good correlation between the two? Based on visual inspection, does the relationship between these two variables appear linear?

**Figure 16.** f




There is some evidence of nonlinearity, as the curve appears to have a negative second derivative (concave down). This notwithstanding, it does appear the two variables are highly correlated.

⌨ The car library contains an improved version of the `plot()` command, called `scatterplot()`, that allows the creation of similar scatterplots more easily without having to do as much gymnastics around missing data. To use it, you must first download the library from CRAN (it is not part of the base installation). To get it from one of the mirror sites of CRAN, enter the following command in the console window:

```
install.packages("car", dependencies=TRUE)
```

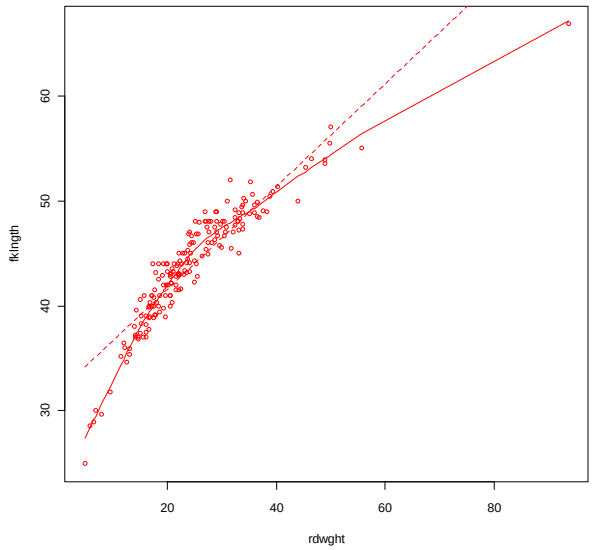
Alternatively, you can use the R menu **Packages ->Install packages** and select the car package from the (long) list. Once the library is install, you still must load it in memory before you can use its functions.

library(car)

Finally, you can use the `scatterplot()` function.

```
scatterplot(fklngth~rdwght, reg.line=lm, smooth=TRUE,
labels=FALSE,   boxplots=FALSE, span=0.5, data=sturgeon)
```

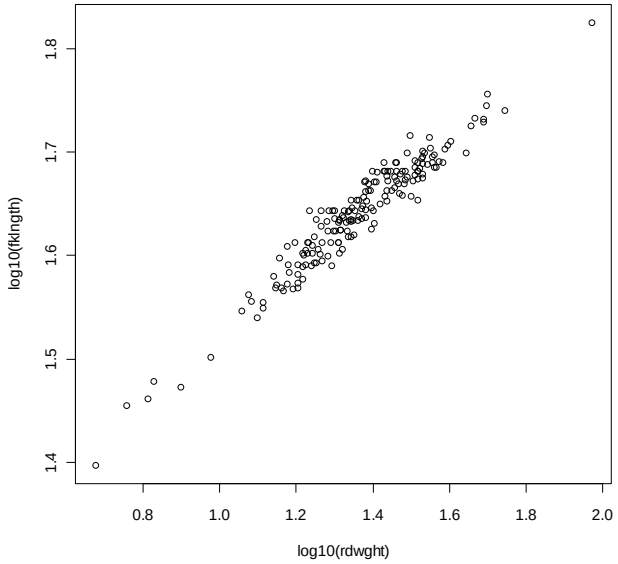
**Figure 17.**




⌨ Refaites le diagramme de dispersion, sans les lignes, avec des axes logarithmiques:

```
plot(log10(fklngth)~log10(rdwght))
```

**Figure 18.**




Compare the diagrams before and after the transformation(Figs 17 et 18).Since the relationship is more linear after transformation, correlation analysis should be done on the transformed data.

## Data transformations and the product-moment correlation

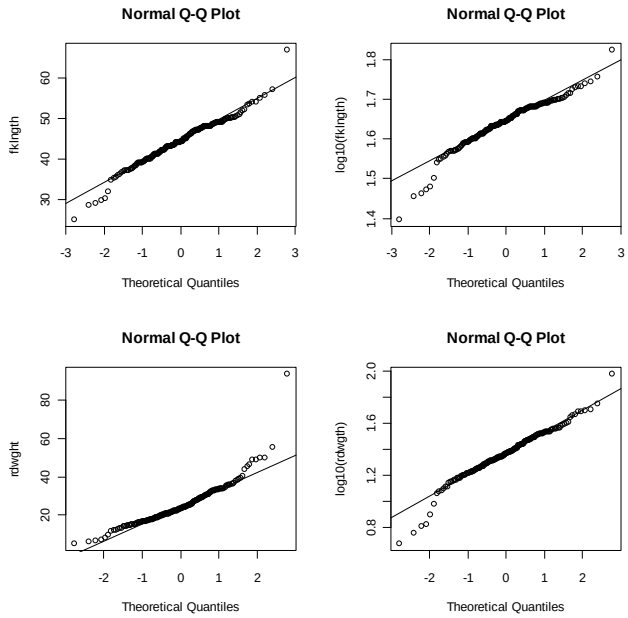
Recall that another assumption underlying significance testing of the product-moment correlation is that the distribution of the two variables in question is bivariate normal. Testing multivariate normality is a very dicey proposition, and S-PLUS has no way of doing it easily. However, we can test to see whether each of the two variables are normally distributed using the same procedures outlined in the exercise on two-sample comparisons. If the two variables are each normally distributed, then one is usually (relatively) safe in assuming the joint distribution is normal, although this needn't necessarily be true.

⌨ Examine the distribution of the 4 variables (the two original variables and the log-transformed variables). What do you conclude from visual inspection of these plots?

The following graph contains the 4 QQ plots. It was produced by the code below that starts with the par() command to ensure that all 4 plots would appear together on the same page in 2 rows and 2 columns: :

```
par(mfrow = c(2, 2))
qqnorm(fklngth,ylab="fklngth")
qqline(fklngth)
qqnorm(log10(fklngth),ylab="log10(fklngth)")
qqline(log10(fklngth))
qqnorm(rdwght,ylab="rdwght")
qqline(rdwght)
qqnorm(log10(rdwght),ylab="log10(rdwgth)")
qqline(log10(rdwght))
par(mfrow = c(1, 1))
```

**Figure 19.**



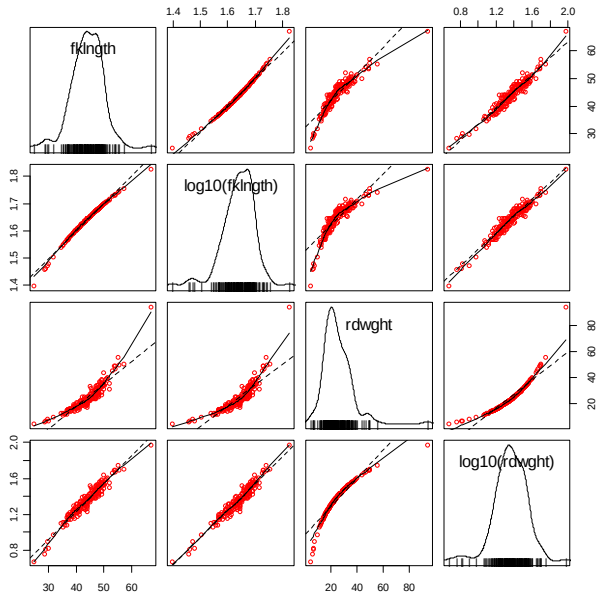

None of these distributions are perfectly normal, but deviations are mostly minor..

To generate a scatterplot matrix of all pairs of variables, with linear regression and lowess traces, you can use `scatterplot.matrix()` (from the `car` library):

```
scatter-
plot.matrix(~fklngth+log10(fklngth)+rdwght+log10(rdwght),
reg.line=lm,   smooth=TRUE, span=0.5, diagonal = 'density')
```

to get

**Figure 20.**




Next, calculate the Pearson product-moment correlation between each pair (untransformed and log transformed) using the cor() command. However, to do this, it will be easier if you first add your transformed data as columns in the sturgeon data frame:

```
sturgeon$lfklngth <- with(sturgeon, log10(fklngth))
sturgeon$lrdwght <- with(sturgeon, log10(rdwght))
```

Then you can get the correlation matrix by:

```
cor(sturgeon[,c("fklngth","lfklngth","lrdwght","rdwght")],
  use="complete.obs")
```

Note the **use="complete.obs"** parameter. It tells R to keep only lines of the data frame where all variables were measured. If there are missing data, some lines will be removed, but correlations will be calculated for the same subset of cases for all pairs of variables. One could use, instead, **use="pairwise.complete.obs"**, to tell R to only eliminate observations when values are missing for this particular pair of variables. In this situation, if there are missing values in the data frame, the sample size for pairwise correlations will vary. In general, I recommend you use the option **use="complete.obs",** unless you have so many missing values that it eliminates the majority of your data.

Why is the correlation between the untransformed variables the smallest in the set?

```
          fklngth  lfklngth   lrdwght    rdwght
fklngth  1.0000000 0.9921435 0.9645108 0.9175435
lfklngth 0.9921435 1.0000000 0.9670139 0.8756203
lrdwght  0.9645108 0.9670139 1.0000000 0.9265513
rdwght   0.9175435 0.8756203 0.9265513 1.0000000
```

Several things should be noted here. First, the correlation between fork length and round weight is high, regardless of which variables are used: so as might be expected, heavier fish are also longer, and vice versa. Second, the correlation is greater for the transformed variables than the untransformed variables. Why? Because the correlation coefficient is inversely proportional to the amount of scatter around a straight line. If the relationship is curvilinear (as it is for the untransformed data), the scatter around a straight line will be greater than if the relationship is linear. Hence, the correlation coefficient will be smaller.

## Testing the significance of correlations and Bonferroni probabilities

It's possible to test the significance of individual correlations using the commands window. As an example, let's try testing the significance of the correlation between `lfklngth` and `rdwght` (the smallest correlation in the above table).

From R console window, enter the following to test the correlation between `lkfl` and `rdwght`:

```
>cor.test(sturgeon$lfklngth, sturgeon$rdwght, alterna-
tive="two.sided",   method="pearson")

Pearson's product-moment correlation

data:  sturgeon$lfklngth and sturgeon$rdwght
t = 24.3223, df = 180, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8367345 0.9057199
sample estimates:
      cor
```

0.8756203

We see here that the correlation is highly significant ($p < 2.2e-16$), which is no surprise given how high the correlation coefficient is (0.8756).

It's important to bear in mind that when you are testing the probability of correlations, the probability of finding any one correlation that is "significant" by pure chance increases with the

number of pairwise correlations examined. Suppose, for example, that you have five variables; there are then a total of 10 possible pairwise correlations, and from this set, you would probably not be surprised to find at least one that is "significant" purely by chance. One way of avoiding the problem is to adjust individual $\alpha$ levels for pairwise correlations by dividing by the number of comparisons, k, such that: $\alpha' = \alpha/k$ (Bonferroni probabilities), i.e. if initially, $\alpha = 0.05$ and there are a total of 10 comparisons, then $\alpha' = 0.005$.

In the above example where we examined correlations between `FKLNGTH` and `RDWGHT` and their log and square-root transformations, it would be appropriate to adjust the $\alpha$ at which significance is tested by the total number of correlations in the matrix (in this case, 15, so $\alpha'=0.0033$). Does your decision about the significance of the correlation between `lkfl` and `rdwght` change?

An interesting and important question in conservation biology is whether there are biodiversity hotspots, i.e. areas where the biodiversity of different taxonomic groups, e.g. birds, mammals and plants, are all high. If so, we would predict that a sample of different areas would show positive correlations between species richness (number of different species) of different groups.

## Non-parametric correlations: Spearman's rank and Kendall's tau

The analysis done with the sturgeon data in the section above suggests that one of the assumptions of correlation, namely, bivariate normality, may not be valid for `fklngth` and `rdwght` nor for the log transforms of these variables. Finding an appropriate transformation is sometimes like looking for a needle in a haystack; indeed, it can be much worse simply because for some distributions, there is no transformation that will normalize the data. In such cases, the best option may be to go to a non-parametric analysis that does not assume bivariate normality or linearity. All such correlations are based on the ranks rather than the data themselves: two options available in R are Spearman and Kendall's tau.

From the R console window, test the correlation between `fklngth` and `rdwght` using both the Spearman and Kendall's tau. The following commands will produce the correlations:

```
>cor.test(sturgeon$fklngth, sturgeon$rdwght, alterna-
tive="two.sided",   method="spearman")

        Spearman's rank correlation rho
```

```
data:  sturgeon$fklngth and sturgeon$rdwght
S = 47971.33, p-value < 2.2e-16
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.9522546
```

```
>cor.test(sturgeon$fklngth, sturgeon$rdwght, alterna-
tive="two.sided", method="kendall")Kendall's rank correla-
tion tau
```

```
 Kendall's rank correlation tau

data:  sturgeon$fklngth and sturgeon$rdwght
z = 16.3578, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8208065
```

Contrast these results with those obtained using the Pearson product-moment correlation. Why the difference?

Test the non-parametric correlations on pairs of the transformed variables. You should immediately note that the non-parametric correlations are identical for untransformed and transformed variables. This is because we are using the ranks, rather than the raw data, and the rank ordering of the data does not change when a transformation is applied to the raw values.

Note that the correlations for Kendall's tau (0.820) are lower than for the Spearman rank (0.952) correlation. This is because Kendall's gives more weight to ranks that are far apart, whereas Spearman's weights each rank equally. Generally, Kendalls's is more appropriate when there is more uncertainty about the reliability of close ranks.

The sturgeons in this sample were collected using nets and baited hooks of a certain size. What impact do you think this method of collection had on the shapes of the distributions of `fklngth` and `rdwght`? Under these circumstances, do you think correlation analysis is appropriate at all?

Note that correlation analysis assumes that *each variable is randomly sampled*. In the case of sturgeon, this is not the case: baited hooks and nets will only catch sturgeon above a certain minimum size. Note that in the sample, there are no small sturgeons, since the fishing gear targets only larger fish. Thus, we should be very wary of the correlation coefficients associated with our analysis, as the inclusion of smaller fish may well change our estimate of these correlations.

## Simple linear regression

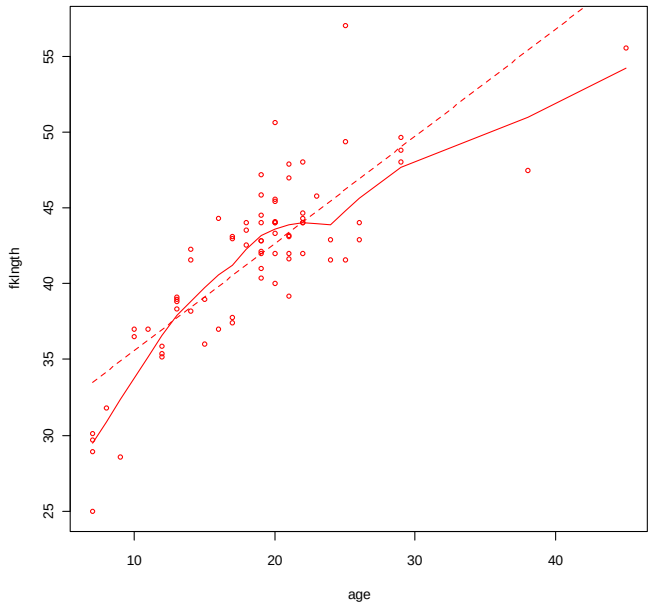
In correlation analysis we are interested in how pairs of variables covary: However, in regression analysis, we are attempting to estimate a model thus predicts a variable (the dependent variable) from another variable (the independent variable)..

As with any statistical analysis, the best way to begin is by looking at your data. If you are interested in the relationship between two variables, say, Y and X, produce a plot of Y versus X just to get a "feel" for the relationship.

⌨ The data file `sturgeon.Rdata` contains data for sturgeons collected from 1978-1980 at Cumberland House, Saskatchewan and The Pas, Manitoba. Make a scatterplot of `fklngth` (the dependent variable) versus `age` (the independent variable) for males and add a linear regression and a loess smoother. What do you conclude from this plot?

```
sturgeon.male <- subset(sturgeon, subset=sex=="MALE")
library(car)
scatterplot(fklngth~age, reg.line=lm, smooth=TRUE, labels=FALSE,
  boxplots=FALSE, span=0.5, data=sturgeon.male)
```

**Figure 21.**




This suggests that the relationship between age and fork length is not linear.

Suppose that we want to know the growth rate of male sturgeon. One estimate (perhaps not a very good one) of the growth rate is given by the slope of the fork length - age regression.

First, let's run the regression with the lm() command, and save its results in an object called RegModel.1

```
RegModel.1 <- lm(fklngth~age, data=sturgeon.male)
```

Nothing appears on the screen, but don't worry, it all got saved in memory.

To see the statistical results, type:

```
summary(RegModel.1)
```

```
Call:
lm(formula = fklngth ~ age, data = sturgeon.male) ❶

Residuals: ❷
    Min      1Q  Median      3Q     Max
-8.4936 -2.2263  0.1849  1.7526 10.8234

Coefficients: ❸
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.50359    1.16873   24.39   <2e-16 ***
age          0.70724    0.05888   12.01   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: ❹ 3.307 on 73 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.664 ❺,  Adjusted R-squared: 0.6594 ❻
F-statistic: 144.3 on 1 and 73 DF,  p-value: < 2.2e-16 ❼
```

❶ A friendly reminder of the model fitted and the data used.

❷ General statistics about the residuals around the fitted model

❸ Fitted model parameter estimates, standard errors, t values and associated probabilities

❹ Square root of the residual variance

❺ Coefficient of determination. It corresponds to the proportion of the total variance of the dependent variable that is accounted for the regression.

❻ The adjusted R-squared accounts for the number of parameters in the model. If you want to compare the performance of several models with different numbers of parameters, this is the one to use.

❼ This is the test of the overall significance of the model. In the simple regression case, this is the same as the test on the slope of the regression.

The estimated regression equation is therefore:

Fklngth = 28.50359+ 0.70724*age

Given the highly significant F-value of the model, as well as the highly significant t-value for the slope of the line, we reject the null hypothesis that there is no relationship between fork length and age.

## Testing regression assumptions

Simple model I regression makes four assumptions:

1. the X variable is measured without error;
2. the relationship between Y and X is linear;
3. that for any value of X, the Y's are independently and normally distributed;
4. the variance of Y for fixed X is independent of X.

Having done the regression, we can now test the assumptions. For most biological data, the first assumption is almost never valid; usually there is error in both Y and X. This means that in general, slope estimates are biased, but predicted values are unbiased. However, so long as the error in X is small relative to the range of X in your data, the fact that X has an associated error is not likely to influence the outcome dramatically. On the other hand, if there is significant error in X, regression results based on a model I regression may give poor estimates of the functional relationship between Y and X. In this case, more sophisticated regression procedures must be employed which are, unfortunately, beyond the scope of this course.

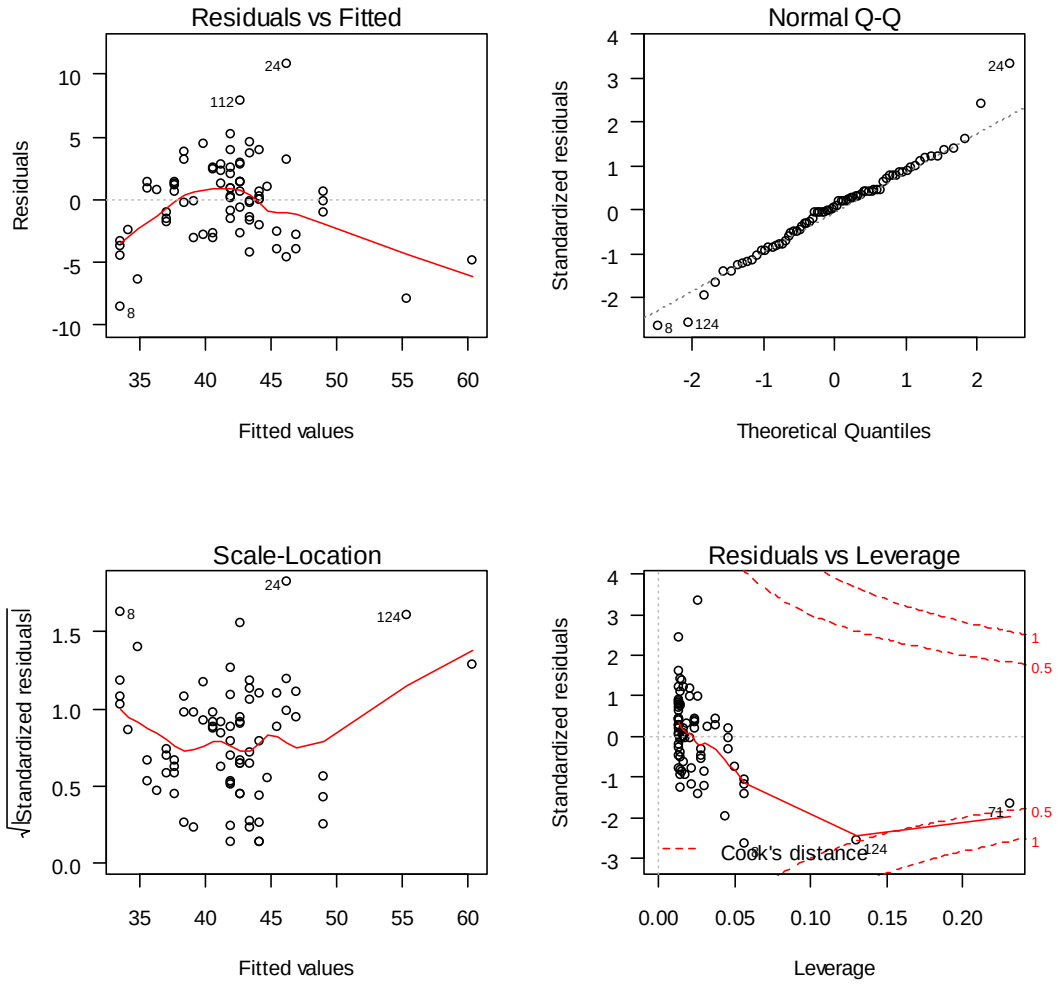
The other assumptions of a model I regression can, however, be tested, or at least evaluated visually. The plot() command can display diagnostics for linear models.

```
par(mfrow = c(2, 2), las=1)
plot(RegModel.1)
```

The par() command is used here to tell R to display 2 rows and 2 columns of graphs per page (there are 4 diagnostic graphs for linear models generated automatically), and the las statement is to tell R to rotate the labels of the Y axis so that they are perpendicular to the Y axis. (Yes, I know, this is not at all obvious.)

You will get:

**Figure 22.**




The first graph (upper left) can tell you about linearity, normality, and homoscedasticity of the residuals. It shows the deviations around the regression vs the predicted values. Remember that the scatterplot (`fklngth` vs `age`) suggested that the relationship between fork length and age is not linear. Very young and very old sturgeons tended to fall under the line, and fish of average age tended to be a bit above the line. This is exactly what the residual vs fitted plot shows. The red line is a lowess trace through these data. If the relationship was linear, it would be approximately flat and close to 0. The scatter of residuals tells you a bit about their normality and homoscedasticity, although this graph is not the best way to look at these properties. The next two are better.

The second graph is to assess the normality of the residuals. It is a QQ plot of the residuals . If the residuals were normally distributed, they would fall very close to the diagonal line. Here, we see it is mostly the case, except in the tails.

The third graph, bottom left and titled Scale-Location, helps with assessing homoscedasticity. It plots the square root of the absolute value of the standardized residual (residual divided by the standard error of the residuals, this scales the residuals so that their variance is 1 ) as a function of the fitted value. This graph can help you visualize whether the spread of the residuals is constant or not. If residuals are homoscedastic, then the average will not with increasing fitted values. Here, there is slight variability, but it is not monotonous (i.e. it does not increase or decrease systematically) and there is no strong evidence against the assumption of homoscedasticity.

The fourth graph, bottom right, plots the residuals as a function of leverage and can help detecting the presence of outliers or points that have a very strong influence on the regression results. The leverage of a point measures how far it is from the other points, but only with respect to the independent variable. In the case of simple linear regression, it is a function of the difference between the observation and the mean of the independent variable. You should look more closely at any observation with a leverage value that is greater than: $2(k+1)/n$, where k is the number of independent variables (here 1), and n is the number of observations. In this case there is 1 independent variable, 75 observations, and points with a leverage higher than 0.053 may warrant particular scrutiny. The plot also gives you information about how the removal of a point from the data set would change the predictions. This is measured by the Cook's distance, illustrated by the red lines on the plot. A data point with a Cook distance larger than 1 has a large influence.

**Trap:** Note that R automatically labels the most extreme cases on each of these 4 plots. It does not mean that these cases are outliers, or that you necessarily need be concerned with them. In any data set, there will always be a minimum and a maximum residual.

So, what is the verdict about the linear regression between `fklngth` and `age`? It fails the linearity, possibly fails the normality, passes homoscedasticity, and this does not seem to be too strongly affected by some bizarre points.

## Formal tests of regression assumptions

In my practice, I seldom use formal tests of regression assumptions and mostly rely on graphs of the residuals to guide my decisions. To my knowledge, this is what most biologists and data analysts do. However, in my early analyst life I was not always confident that I was interpreting these graphs correctly and wished that I had a formal test or a statistic quantifying the degree of deviation from the regression assumptions.

The `lmtest` R package, not part of the base R installation, but available from CRAN, contains a number of tests for linearity and homoscedasticity. And one can test for normality using the Shapiro-Wilk test seen previously.

⌨Install the lmtest package from CRAN.

⌨Run the following commands:

```
library(lmtest)
bptest(fklngth ~ age, varformula = ~
  fitted.values(RegModel.1), studentize=TRUE, data=stur-
geon.male)
dwtest(fklngth ~ age, alternative="greater",
  data=sturgeon.male)
resettest(fklngth ~ age, power=2:3, type="regressor",
  data=sturgeon.male)
shapiro.test(residuals(RegModel.1))
```

to get:

```
> bptest(fklngth ~ age, varformula = ~
+   fitted.values(RegModel.1), studentize=TRUE, data=sturgeon.male)

        studentized Breusch-Pagan test ❶

data:  fklngth ~ age
BP = 0.0035, df = 1, p-value = 0.9531

> dwtest(fklngth ~ age, alternative="greater",
+   data=sturgeon.male)

        Durbin-Watson test ❷

data:  fklngth ~ age
DW = 2.242, p-value = 0.849
alternative hypothesis: true autocorrelation is greater than 0

> resettest(fklngth ~ age, power=2:3, type="regressor",
+   data=sturgeon.male)

        RESET test ❸

data:  fklngth ~ age
RESET = 14.5442, df1 = 2, df2 = 71, p-value = 5.082e-06

> shapiro.test(residuals(RegModel.1))
```

```
        Shapiro-Wilk normality test❹

data:  residuals(RegModel.1)
W = 0.9804, p-value = 0.2961
```

❶The Breusch-Pagan test examines whether the variability of the residuals is constant with respect to increasing fitted values. A low p value is indicative of heteroscedasticity. Here, the p value is high, and supports my visual assessment that the homoscedasticity assumption is met by these data.

❷The Durbin-Watson test can detect serial autocorrelation in the residuals. Under the assumption of no autocorrelation, the D statistic is 2. This test can detect violation of independence of observations (residuals), although it is not foolproof. Here there is no problem identified.

❸The RESET test is a test of the assumption of linearity. If the linearity assumption is met, the RESET statistic will be close to 1. Here, the statistic is much larger (14.54), and very highly significant. This confirms our visual assessment that the relationship is not linear.

❹And the Shapiro-Wilk normality test on the residual confirms that the deviation from normality of the residuals is not large.

## Data transformations in regression

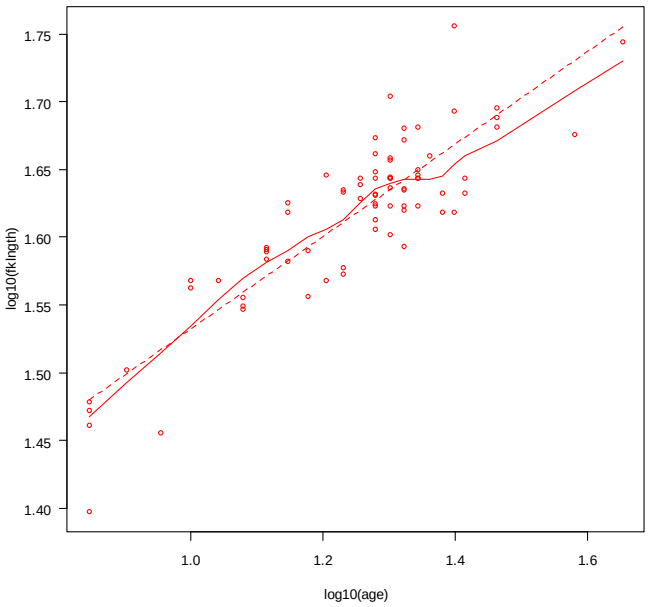
The analysis above revealed that the linearity assumption underlying regression analysis is not met by the `fklngth-age` data. If we want to use regression analysis, data transformations are required:

⌨Let's plot the log-transformed data,

```
scatterplot(log10(fklngth)~log10(age), reg.line=lm,
smooth=TRUE, labels=FALSE,
  boxplots=FALSE, span=0.5, data=sturgeon.male)
```

We get:

**Figure 23.**




and fit the linear regression model on the log-transformed variables:

```
RegModel.2 <- lm(log10(fklngth)~log10(age), data=stur-
geon.male)
summary(RegModel.2)
```

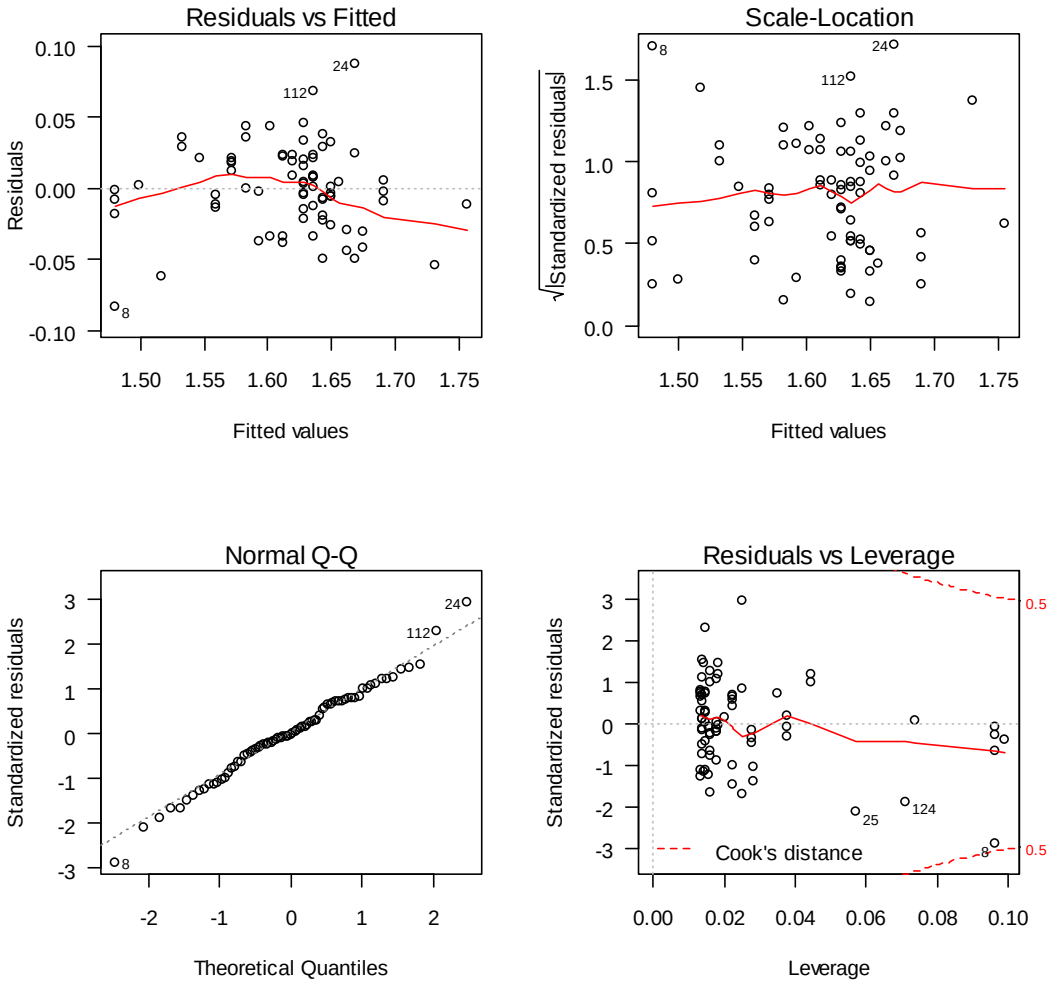
```
Call:
lm(formula = log10(fklngth) ~ log10(age), data = sturgeon.male)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0827935 -0.0168373 -0.0007188  0.0211016  0.0874465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.19199    0.02723   43.77   <2e-16 ***
log10(age)   0.34086    0.02168   15.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03015 on 73 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.772,      Adjusted R-squared: 0.7688
F-statistic: 247.1 on 1 and 73 DF,  p-value: < 2.2e-16
```

Note that by using the log transformed data, the proportion of variation explained by the regression has increased by 10% (from 0.664 to 0.772), a substantial increase.

So the relationship has become more linear. Good. Let's look at the residual diagnostic plots:

plot(RegModel.2)

**Figure 24.**




So things appear a little better than before, although still not ideal. For example, the Residual vs fitted plot still suggests a potential nonlinearity. The QQ plot is nicer than before, indicating that residuals are more normally distributed after the log-log transformation. There is no indication of heteroscedasticity. And, although there are still a few points with somewhat high leverage, none have a Cook's distance above 0.5. It thus seems that transforming data improved things: more linear, more normal, less dependence on extreme data.

Do the formal tests support this visual assessment?

```
bptest(log10(fklngth) ~ log10(age), varformula = ~
  fitted.values(RegModel.2), studentize=TRUE, data=stur-
geon.male)
dwtest(log10(fklngth) ~ log10(age), alternative="greater",
  data=sturgeon.male)
resettest(log10(fklngth) ~ log10(age), power=2:3,
type="regressor",
  data=sturgeon.male)
shapiro.test(residuals(RegModel.2))
```

```
> bptest(log10(fklngth) ~ log10(age), varformula = ~
+   fitted.values(RegModel.2), studentize=TRUE, data=sturgeon.male)

        studentized Breusch-Pagan test

data:  log10(fklngth) ~ log10(age)
BP = 0.2244, df = 1, p-value = 0.6357

> dwtest(log10(fklngth) ~ log10(age), alternative="greater",
+   data=sturgeon.male)

        Durbin-Watson test

data:  log10(fklngth) ~ log10(age)
DW = 2.1777, p-value = 0.6134
alternative hypothesis: true autocorrelation is greater than 0

> resettest(log10(fklngth) ~ log10(age), power=2:3, type="regressor",
+   data=sturgeon.male)

        RESET test

data:  log10(fklngth) ~ log10(age)
RESET = 4.4413, df1 = 2, df2 = 71, p-value = 0.01523

> shapiro.test(residuals(RegModel.2))

        Shapiro-Wilk normality test

data:  residuals(RegModel.2)
W = 0.99, p-value = 0.8246
.
```

Indeed, they do: residuals are still homoscedastic (Breusch-Pagan test), show not autocorrelation (Durbin-Watson test), are normal (Shapiro-Wilk test), and they are more linear (p value of the RESET test is now 0.015, instead of 0.000005). Linearity has improved, but is still violated somewhat.

## Dealing with outliers

In this case, there are no real clear outliers. Yes, observations 8, 24, and 112 are labeled as the most extreme in the last set of residual diagnostic plots. But they are still within what I consider the "reasonable" range. But how does one define a limit to the reasonable? When is an extreme value a real outlier we have to deal with? Opinions vary about the issue, but I favor conservatism.

My rule is that, unless the value is clearly impossible or an error in data entry, I do not delete "outliers" and analyze all my data. Why? Because, I want my data to reflect natural or real variability. Indeed, variability is often what interests biologists the most.

Keeping extreme values is the fairest way to proceed, but it often creates other issues. These values will often be the main reason why the data fail to meet the assumptions of the statistical analysis. The solution is to run the analysis with and without the outliers, and compare the results. In many cases, the two analyses will be qualitatively similar: the same conclusions will be reached, and the effect size will not be very different. Sometimes, however, this comparison will reveal that the presence of the outliers changes the story. The logical conclusion then is that the results depend on the outliers and that the data at hand are not very conclusive.

As an example, let's rerun the analysis after eliminating observations labeled 8, 24, and 112:

```
RegModel.3 <- lm(log10(fklngth)~log10(age), data=stur-
geon.male, subset=!(rownames(sturgeon.male) %in%
c('8','24','112')))
summary(RegModel.3)
```

to obtain:

```
Call:
lm(formula = log10(fklngth) ~ log10(age), data = sturgeon.male,
    subset = !(rownames(sturgeon.male) %in% c("8", "24", "112")))

Residuals:
       Min         1Q     Median         3Q        Max
-0.0691634 -0.0173903  0.0009862  0.0185900  0.0476466

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.22676    0.02431   50.46   <2e-16 ***
log10(age)   0.31219    0.01932   16.16   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02554 on 70 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.7885,     Adjusted R-squared: 0.7855
F-statistic:   261 on 1 and 70 DF,  p-value: < 2.2e-16
```

The intercept, slope, and R squared are about the same, and the significance of the slope is still astronomical. Removing the "outliers" has little effect in this case.

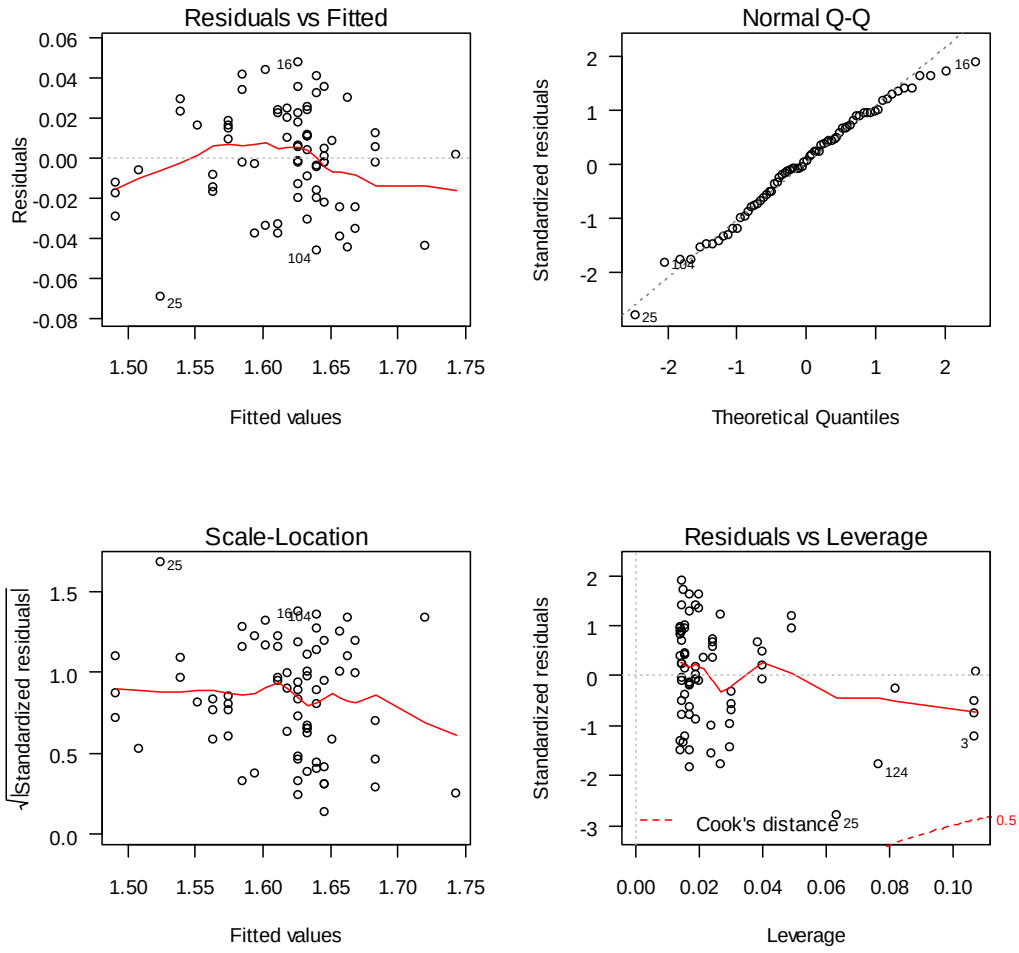
As for the diagnostic residual plots and the formal tests of assumptions:

plot(RegModel.3)

```
library(lmtest)
```

```
sturgeon.male.subset <- subset(sturgeon, sub-
set=!(rownames(sturgeon.male) %in% c('8','24','112')))
bptest(log10(fklngth) ~ log10(age), varformula = ~
  fitted.values(RegModel.3), studentize=TRUE, data=stur-
geon.male.subset)
dwtest(log10(fklngth) ~ log10(age), alternative="greater",
data=sturgeon.male.subset)
resettest(log10(fklngth) ~ log10(age), power=2:3,
type="regressor", data=sturgeon.male.subset)

shapiro.test(residuals(RegModel.3))
```

**Figure 25.**




```
> bptest(log10(fklngth) ~ log10(age), varformula = ~
+   fitted.values(RegModel.3), studentize=TRUE, data=sturgeon.male.subset)

        studentized Breusch-Pagan test

data:  log10(fklngth) ~ log10(age)
BP = 4.8698, df = 1, p-value = 0.02733
```

```
> dwtest(log10(fklngth) ~ log10(age), alternative="greater", data=stur-
geon.male.subset)

        Durbin-Watson test

data:  log10(fklngth) ~ log10(age)
DW = 2.1894, p-value = 0.8896
alternative hypothesis: true autocorrelation is greater than 0

> resettest(log10(fklngth) ~ log10(age), power=2:3, type="regressor",
data=sturgeon.male.subset)

        RESET test

data:  log10(fklngth) ~ log10(age)
RESET = 4.4312, df1 = 2, df2 = 163, p-value = 0.01337

>
> shapiro.test(residuals(RegModel.3))

        Shapiro-Wilk normality test

data:  residuals(RegModel.3)
W = 0.9832, p-value = 0.4502
```

No real difference either. Overall, this suggests that the most extreme values do not have undue influence on the results.

## Quantifying effect size in regression and power analysis

Biological interpretation differs from statistical interpretation. Statistically, we conclude that size increase with age (i.e. the slope is positive and different from 0). But this conclusion alone does not tell if the difference between young and old fish is large. The slope and the scatterplot are more informative than the p-value here. The slope (in log-log space) is 0.34. This means that for each unit increase of X (log10(age)), there is an increase of 0.34 units of log10(fklngth). In other words, when age is multiplied by 10, fork length is multiplied by about 2. Humm, length increases more slowly than age. This slope value (0.34) is an estimate of effect size. It measure how much length changes with age

### Power to detect a given slope

For G*Power calculation, a different metric of effect size is used to facilitate calculations. The effect size (d) in regression can be computed from the slope as:

$$d = \frac{b}{s_b\sqrt{n-k-1}}$$

where b is the estimate of the slope, $s_b$ is the standard error of the slope, n is the number of observations, and k is the number of predictors (1 in simple linear regression).

You can compute an approximate power with G*Power for some slope value that you deem of sufficient magnitude to warrant detection. Go to **Tests: Means: One group: difference from constant**, you would replace b in the equation by the slope you want to detect, but use the estimate of standard error from your data.

For example, suppose that sturgeon biologists deem that a slope of 0.1 for the relationship between log10(fklngth) and log10(age) is meaningful and you wanted to estimate the power to detect such a slope with a sample of 20 sturgeons. Results from the log-log regression contain what we need:

```
summary(RegModel.2)
```

```
Call:
lm(formula = log10(fklngth) ~ log10(age), data = sturgeon.male)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0827935 -0.0168373 -0.0007188  0.0211016  0.0874465

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.19199    0.02723   43.77   <2e-16 ***
log10(age)   0.34086    0.02168   15.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.03015 on 73 degrees of freedom
  (5 observations deleted due to missingness)
Multiple R-squared: 0.772,      Adjusted R-squared: 0.7688
```

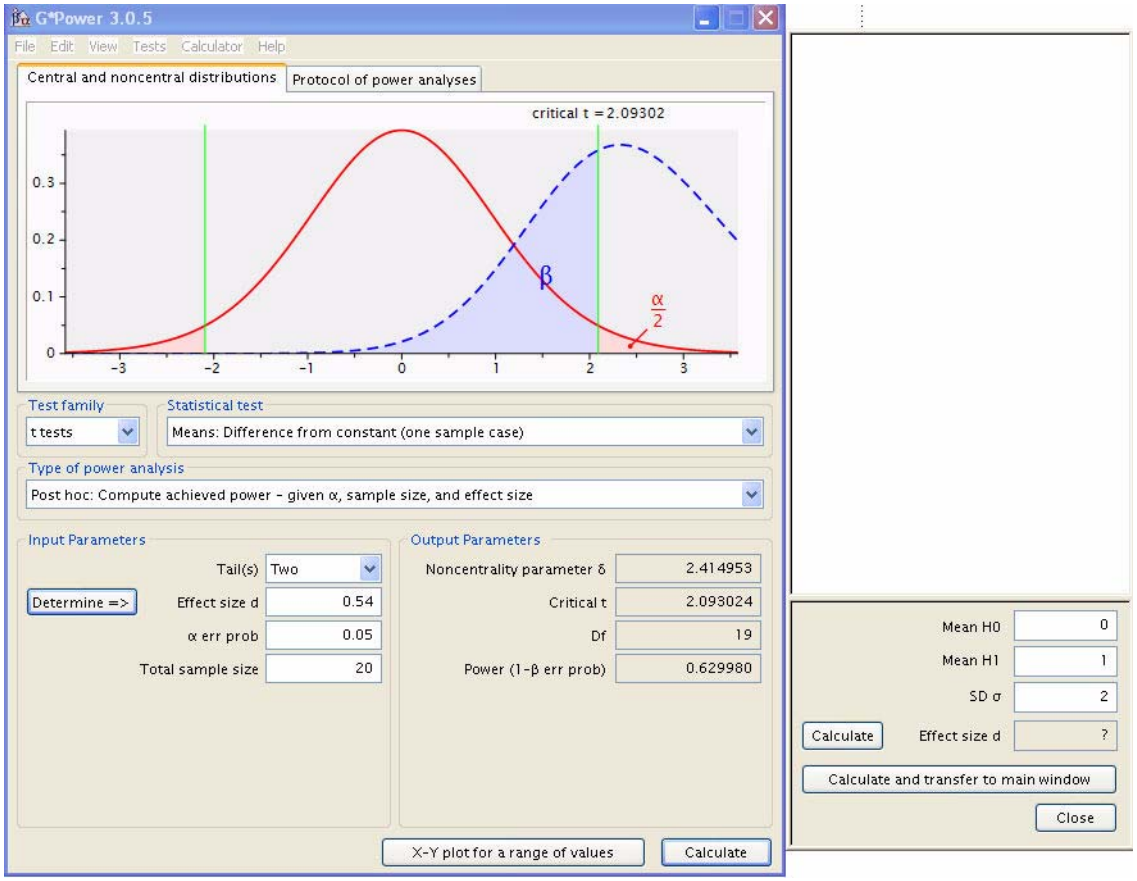
F-statistic: 247.1 on 1 and 73 DF, p-value: < 2.2e-16

The standard error of the slope is 0.02168. There were 75 fish (n=75) in that sample). These are what we need to compute the effect size for G*Power.

$$d = \frac{b}{s_b\sqrt{n-k-1}} = \frac{0.1}{0.02168\sqrt{75-1-1}} = 0.54$$

Armed with this effect size (an assumed slope of 0.1, given the variability around regression of fklngth vs age), you can then use G*Power. Go to **Tests: Means: One group: difference from constant**, and enter the calculated d value, alpha, and the sample size to compute power:

**Figure 26.**

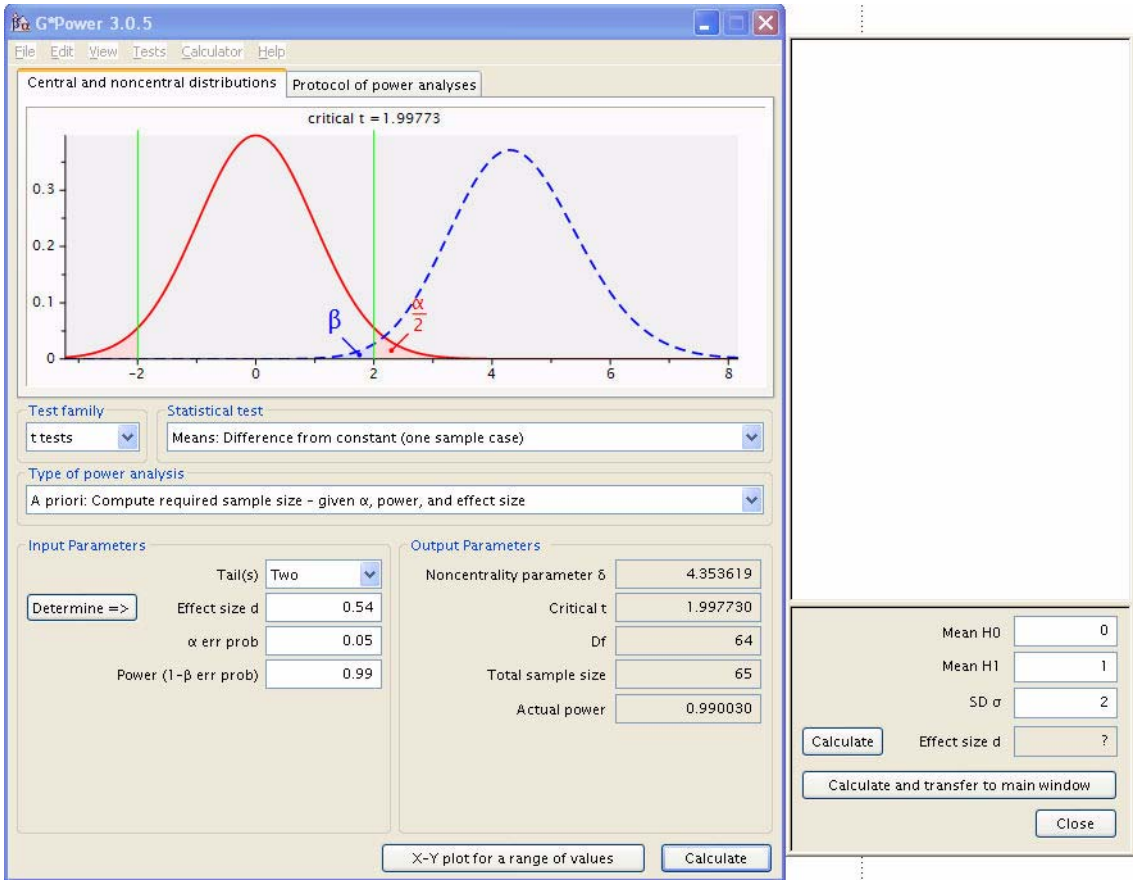



Power to detect a significant slope, if the slope is 0.1, variability of data points around the regression is like in our sample (making that slope having an effect size of 0.54, for a sample of 20 sturgeons, with alpha=0.05 is 0.629. Only about 2/3 of samples of that size would detect a significant effect of age on fklngth.

### Sample size required to achieve desired power

To estimate the sample size required to achieve 99% power to detect a slope of 0.1 (in log-log space), with alpha=0.05, you use the same value of d previously calculated for G*Power (0.54):

**Figure 27.**




By increasing sample size to 65, with the same assumptions as before, power increases to 99%.

## Bootstrapping the simple linear regression with R

A non-parametric test for the intercept and slope of a linear regression can be obtained by bootstrapping:

```
# Bootstrap 95% CI for regression coefficients
library(boot)
# function to obtain regression weights
bs <- function(formula, data, indices) {
  d <- data[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(coef(fit))
}
# bootstrapping with 1000 replications
results <- boot(data=sturgeon.male, statistic=bs,
   R=1000, formula=log10(fklngth)~log10(age))
```

```
# view results
results
> results

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = sturgeon.male, statistic = bs, R = 1000, formula = log10(fklngth)
~
    log10(age))


Bootstrap Statistics :
     original❶     bias❷     std. error❸
t1* 1.1919926 -8.936048e-05  0.03532614
t2* 0.3408557  1.116266e-04  0.02789170
```

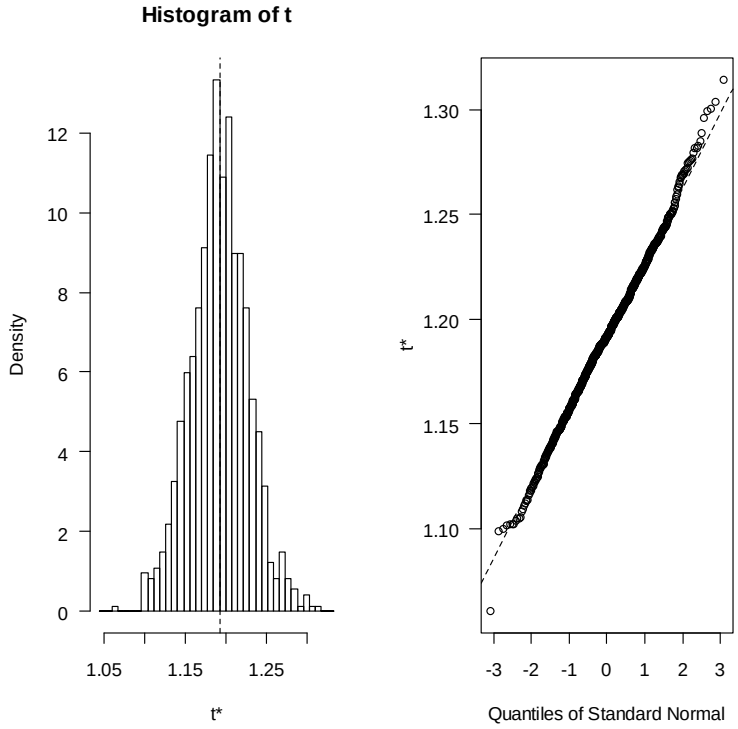
❶ For each parameter in the model (here the intercept is labeled t1* and the slope of the regression line is labeled t2*) , you obtain the original parameter estimate (on all non-bootstrapped data)

❷ bias is the difference between the mean value of all bootstrap estimates and the original value

❸ the standard error of the bootstrap estimate

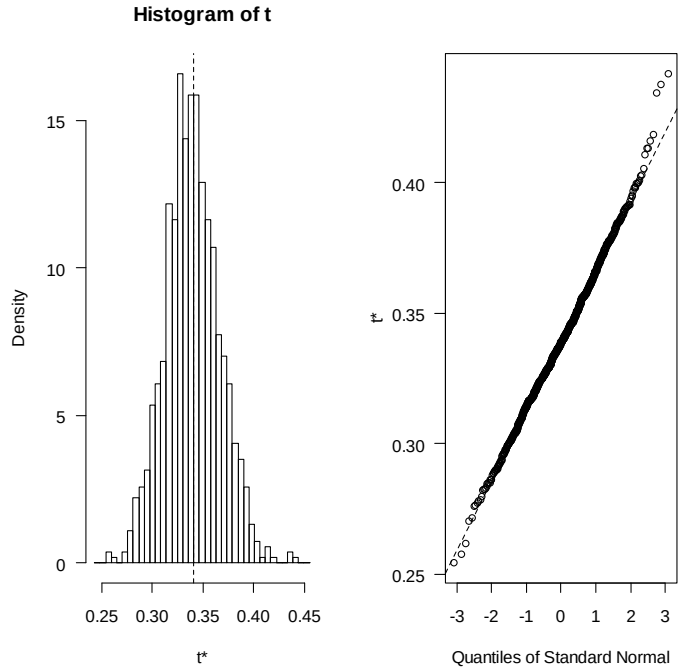
```
plot(results, index=1) # intercept
```

**Figure 28.**




```
plot(results, index=2) # log10(age)
```

**Figure 29.**




The distribution of the bootstrapped estimates is rather Gaussian, with only small deviations in the tails (where it counts for confidence intervals...). One could use the standard error of the bootstrap estimates to calculate a symmetrical confidence interval as mean +- t SE. But, given that R can as easily calculate a bias-corrected adjusted (BCa) confidence interval, or one based on the actual distribution, (Percentile) why not have it do it all:

```
# get 95% confidence intervals
boot.ci(results, type = "all", index = 1)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "all", index = 1)

Intervals :
Level      Normal              Basic
95%   ( 1.124,  1.256 )   ( 1.123,  1.255 )

Level     Percentile            BCa
95%   ( 1.129,  1.261 )   ( 1.104,  1.247 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
Warning message:
In boot.ci(results, type = "all", index = 1) :
  bootstrap variances needed for studentized intervals
```

```
boot.ci(results, type = "all", index = 2)
```

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
```

```
boot.ci(boot.out = results, type = "all", index = 2)

Intervals :
Level      Normal              Basic
95%   ( 0.2898,  0.3946 )   ( 0.2902,  0.3931 )

Level     Percentile            BCa
95%   ( 0.2886,  0.3915 )   ( 0.2966,  0.4044 )
Calculations and Intervals on Original Scale
Some BCa intervals may be unstable
Warning message:
In boot.ci(results, type = "all", index = 2) :
  bootstrap variances needed for studentized intervals
```

Here the 4 types of CI that R managed to calculate are essentially the same. Had data been violating more strongly the standard assumptions (normality, homoscedasticity), then the different methods (Normal, Basic, Percentile, and BCa) would have diverged more. In that case, which one is best? BCa has the favor of most, currently.

## R program

```
#  Exercise - Correlation and regression for BIO4158 Applied Biostatistics
#  Exercice - Corrélation et régression pour BIO4558 Biostatistiques Appli-
quées
#  Auteur: Antoine Morin (amorin@uottawa.ca)
#  Dependencies: boot, car, lmtest

#  Examine data (Sturgeon.Rdata)
#  load(file.choose()) # local version
load(url("http://www.antoinemorin.com/biostats/2010/sturgeon.Rdata")) #
online version

##############################################
# Scatter plots

attach(sturgeon)
plot(fklngth ~ rdwght)
abline(lm(fklngth ~ rdwght), col = "red")
goodcases <- !(is.na(fklngth) | is.na(rdwght))
lines(lowess(fklngth[goodcases] ~ rdwght[goodcases]))

library(car)
scatterplot(fklngth ~ rdwght, reg.line = lm, smooth = TRUE,
  labels = FALSE, boxplots = FALSE, span = 0.5, data = sturgeon)

plot(log10(fklngth) ~ log10(rdwght))

##############################################
# Data transformations and  the product-moment correlation

# Make probability plots of raw and transformed data
par(mfrow = c(2, 2))
qqnorm(fklngth, ylab = "fklngth")
qqline(fklngth)
qqnorm(log10(fklngth), ylab = "log10(fklngth)")
qqline(log10(fklngth))
qqnorm(rdwght, ylab = "rdwght")
qqline(rdwght)
qqnorm(log10(rdwght), ylab = "log10(rdwgth)")
qqline(log10(rdwght))
par(mfrow = c(1, 1))

# Scatterplot matrix
sturgeon$lfklngth <- with(sturgeon, log10(fklngth))
```

```
sturgeon$lrdwght <- with(sturgeon, log10(rdwght))

scatterplot.matrix(~fklngth + lfklngth + lrdwght +
  rdwght, reg.line = lm, smooth = TRUE, span = 0.5, diagonal = "density",
  data = sturgeon)

scatterplot.matrix(~fklngth + log10(fklngth) + rdwght +
    log10(rdwght), reg.line = lm, smooth = TRUE, span = 0.5,
    diagonal = "density")


# Correlation matrix

cor(sturgeon[, c("fklngth", "lfklngth", "lrdwght",
    "rdwght")], use = "complete.obs")


##############################################
#  Testing the significance of correlations and Bonferroni probabilities

# Correlation tests

cor.test(sturgeon$lfklngth, sturgeon$rdwght, alternative = "two.sided",
    method = "pearson")

##############################################
# Non-parametric correlations: Spearman's rank and Kendall's tau

cor.test(sturgeon$lfklngth, sturgeon$rdwght, alternative = "two.sided",
    method = "spearman")
cor.test(sturgeon$fklngth, sturgeon$rdwght, alternative = "two.sided",
    method = "kendall")

##########################################
#  Simple linear regression

sturgeon.male <- subset(sturgeon, subset = sex ==
    "MALE")

# Examine data
library(car)
scatterplot(fklngth ~ age, reg.line = lm, smooth = TRUE,
    labels = FALSE, boxplots = FALSE, span = 0.5, data = sturgeon.male)

# Fit linear regression and plot regression diagnostics
RegModel.1 <- lm(fklngth ~ age, data = sturgeon.male)
summary(RegModel.1)
par(mfrow = c(2, 2), las = 1)
plot(RegModel.1)



#####################################################
# Formal tests of linear models assumptions on residuals

library(lmtest)
#Homoscedasticity
bptest(fklngth ~ age, varformula = ~fitted.values(RegModel.1),
    studentize = TRUE, data = sturgeon.male)
#Serial autocorrelation
dwtest(fklngth ~ age, alternative = "greater", data = sturgeon.male)
#Linearity
resettest(fklngth ~ age, power = 2:3, type = "regressor",
    data = sturgeon.male)
#Normality
shapiro.test(residuals(RegModel.1))

#####################################################
#  Data transformations in regression
```

```
# Redo the same on log-transformed data
scatterplot(log10(fklngth) ~ log10(age), reg.line = lm,
    smooth = TRUE, labels = FALSE, boxplots = FALSE, span = 0.5,
    data = sturgeon.male)
RegModel.2 <- lm(log10(fklngth) ~ log10(age), data = sturgeon.male)
summary(RegModel.2)
plot(RegModel.2)

library(lmtest)
bptest(log10(fklngth) ~ log10(age), varformula = ~fitted.values(RegModel.2),
    studentize = TRUE, data = sturgeon.male)
dwtest(log10(fklngth) ~ log10(age), alternative = "greater",
    data = sturgeon.male)
resettest(log10(fklngth) ~ log10(age), power = 2:3,
    type = "regressor", data = sturgeon.male)
shapiro.test(residuals(RegModel.2))
outlier.test(lm(log10(fklngth) ~ log10(age), data = sturgeon.male))

############################################################
#  Dealing with outliers

# Same analysis, but eliminating data points that deviate from
# the general pattern or have high Cook's distance
# (observations labeled 8, 24, and 112)

RegModel.3 <- lm(log10(fklngth) ~ log10(age), data = sturgeon.male,
    subset = !(rownames(sturgeon.male) %in% c("8", "24", "112")))
summary(RegModel.3)
plot(RegModel.3)
library(lmtest)
sturgeon.male.subset <- subset(sturgeon, subset = !(rownames(sturgeon.male)
%in%
    c("8", "24", "112")))
bptest(log10(fklngth) ~ log10(age), varformula = ~fitted.values(RegModel.3),
    studentize = TRUE, data = sturgeon.male.subset)
dwtest(log10(fklngth) ~ log10(age), alternative = "greater",
    data = sturgeon.male.subset)
resettest(log10(fklngth) ~ log10(age), power = 2:3,
    type = "regressor", data = sturgeon.male.subset)

shapiro.test(residuals(RegModel.3))




############################################################
#  Bootstrapping the simple linear regression with R


# Bootstrap 95% CI for regression coefficients

library(boot)
# function to obtain regression weights
bs <- function(formula, data, indices) {
    d <- data[indices, ]
    fit <- lm(formula, data = d)
    return(coef(fit))
}
# bootstrapping with 1000 replications
results <- boot(data = sturgeon.male, statistic = bs,
    R = 1000, formula = log10(fklngth) ~ log10(age))

# view results
results
plot(results, index = 1) # for intercept
plot(results, index = 2) # for slope
```

```
# get 95% confidence intervals
boot.ci(results, type = "all", index = 1)
boot.ci(results, type = "all", index = 2)
```

# Lab- Two - sample comparisons

After completing this laboratory exercise, you should be able to:

- Use R to visually examine data.
- Use R to compare the means of two normally distributed samples.
- Use R to compare the means of two non-normally distributed samples.
- Use R to compare the means of two paired samples

## Visual examination of sample data

One of the first steps in any type of data analysis is to visualize your data with plots and summary statistics, to get an idea of underlying distributions, possible outliers, and trends in your data. This often begins with plots of the data, such as histograms, probability plots, and box plots, that allow you to get a feel for whether your data are normally distributed, whether they are correlated one to the other, or whether there are any suspicious looking points that may lead you to go back to the original data file to check for errors.

Suppose we want to test the null hypothesis that the size, as indexed by fork length (`fklngth` - the length, in cm, from the tip of the nose to the base of the fork in the caudal fin), of sturgeon at The Pas and Cumberland House is the same. To begin, we have a look at the underlying distributions of the sample data to get a feel for whether the data are normally distributed in each sample. We will not actually test for normality at this point; the assumption of normality in parametric analyses refers always to the residuals and not the raw data themselves. However, if the raw data are non-normally distributed, then you usually have good reason to suspect that the residuals also will be non-normally distributed.

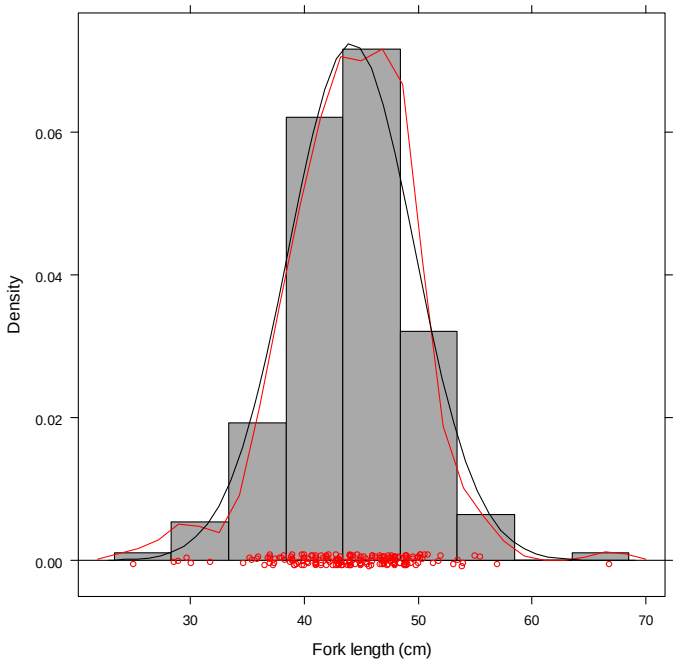
An excellent way to visually compare a data distribution to a normal distribution is to superimpose a histogram of the data and a normal curve. To do so, we must proceed in two steps: 1) tell R that we want to make a histogram with a density curve superimposed, 2) tell R that we want this to be done for both locations.

Using the data file `sturgeon.Rdata`, generate histograms and associated normal densities for `fklngth` data at The Pas and Cumberland House.

```
library(lattice)
```

```
histogram(~fklngth, xlab = "Fork length (cm)", type = "den-
sity",
    panel = function(x, ...) {
        panel.histogram(x, col = "darkgrey", ...)
        panel.densityplot(na.omit(x), col = "red")
       panel.mathdensity(dmath = dnorm, col = "black", args
= list(mean = mean(x),
            sd = sd(x)))
    })
```

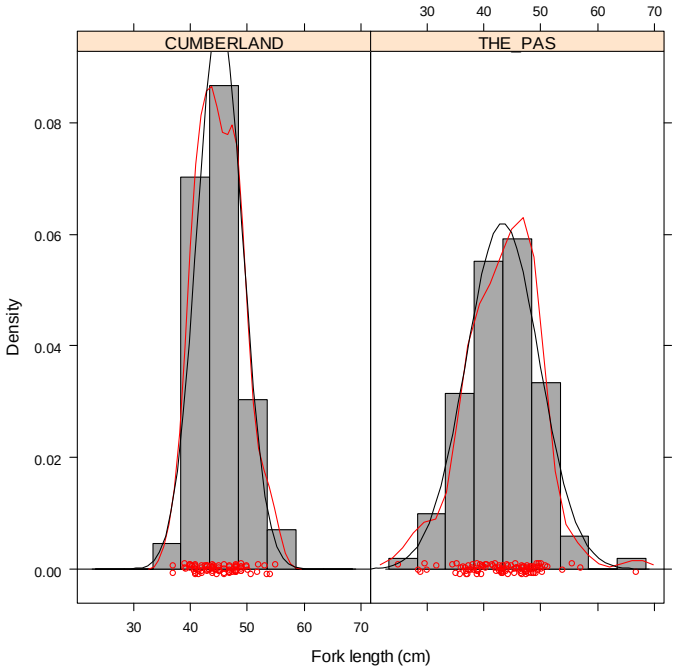
**Figure 30.**




The data shown in this plot are for all cases, but we are interested in examining the histograms by `location`.

```
histogram(~fklngth | location, xlab = "Fork length (cm)",
    type = "density", panel = function(x, ...) {
        panel.histogram(x, col = "darkgrey", ...)
        panel.densityplot(na.omit(x), col = "red")
       panel.mathdensity(dmath = dnorm, col = "black", args
= list(mean = mean(x),
            sd = sd(x)))
    })
```

**Figure 31.**




Based on your visual inspection, are the two samples normally distributed?

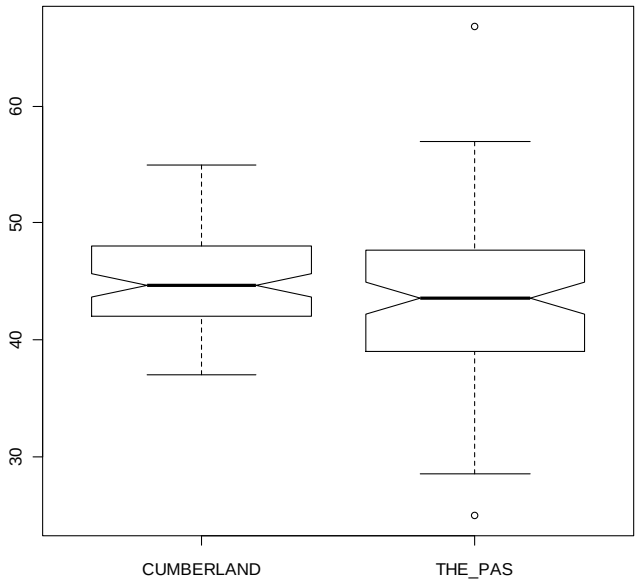
Visual inspection of these plots suggests that this variable is approximately normally distributed in each sample.

Since we are interested in finding out if mean fish size differs among the two locations, it is probably also a good idea to generate a graph that compares the two groups of data. A box plot works well for this.

Generate a box plot of `fklngth` grouped by `location`. What do you conclude about differences in size among the two locations?

```
boxplot(fklngth~location, notch=TRUE)
```

**Figure 32.**




It would appear as though there are not big differences in fish size among the two locations, although fish size at The Pas looks to be a more variable, with a bigger range in size and outliers (defined as values > 1.5* inter-quartile range) at both ends of the distribution.

R program for examining data

```
# load(file.choose()) # local version
load(url("http://www.antoinemorin.com/biostats/2010/stur-
geon.Rdata")) # online version


attach(sturgeon)

library(lattice)

histogram(~fklngth, xlab = "Fork length (cm)", type = "density",
    panel = function(x, ...) {
        panel.histogram(x, col = "darkgrey", ...)
        panel.densityplot(na.omit(x), col = "red")
        panel.mathdensity(dmath = dnorm, col = "black", args =
list(mean = mean(x),
            sd = sd(x)))
    })


histogram(~fklngth | location, xlab = "Fork length (cm)",
    type = "density", panel = function(x, ...) {
        panel.histogram(x, col = "darkgrey", ...)
        panel.densityplot(na.omit(x), col = "red")
        panel.mathdensity(dmath = dnorm, col = "black", args =
list(mean = mean(x),
            sd = sd(x)))
    })


boxplot(fklngth ~ location, notch = TRUE)
)
```

## Comparing means of two independent samples: parametric and non-parametric comparisons

Test the null hypothesis that the mean `fklngth` of The Pas and Cumberland House samples are the same. What do you conclude?

```
> #  t-test assuming equal variances
> t.test(fklngth~location, alternative='two.sided',
conf.level=.95,
+    var.equal=TRUE, data=sturgeon)

        Two Sample t-test

data:  fklngth by location
t = 2.1359, df = 184, p-value = 0.03401
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1308307 3.2982615
sample estimates:
mean in group CUMBERLAND    mean in group THE_PAS
                45.08439                 43.36984


>
> #  t-test assuming unequal variances
```

```
> t.test(fklngth~location, alternative='two.sided',
conf.level=.95,
+   var.equal=FALSE, data=sturgeon)

        Welch Two Sample t-test

data:  fklngth by location
t = 2.2201, df = 169.804, p-value = 0.02774
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1900117 3.2390804
sample estimates:
mean in group CUMBERLAND    mean in group THE_PAS
                45.08439                 43.36984

>
> #  non-parametric wilcoxon test
> wilcox.test(fklngth ~ location, alternative="two.sided",
data=sturgeon)

        Wilcoxon rank sum test with continuity correction

data:  fklngth by location
W = 4973, p-value = 0.06296
alternative hypothesis: true location shift is not equal to 0

>
```

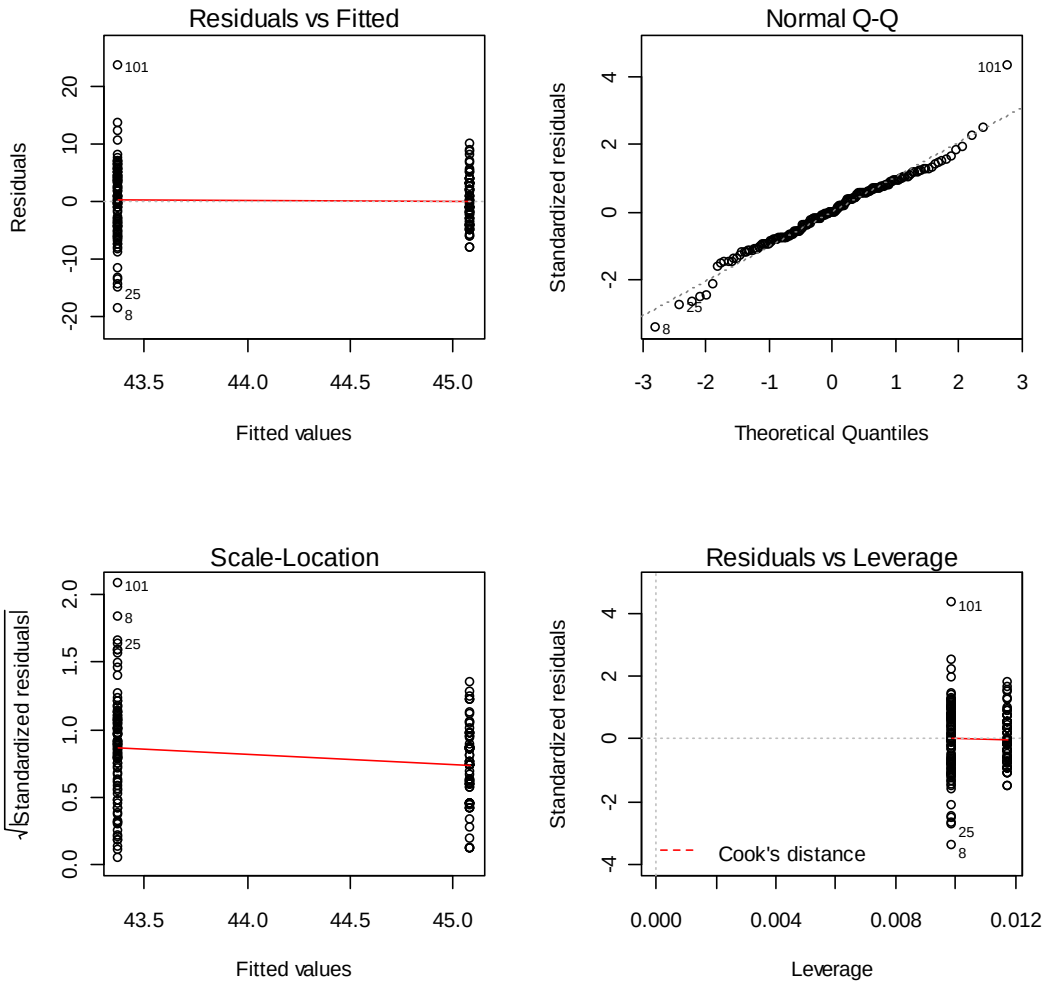
Based on the *t*-test, we would reject the null hypothesis, i.e. there is a significant (but not highly significant) difference in mean fork length between the two populations.

Note that using the Wilcoxon rank sum test, we do not reject the null hypothesis. The two different tests therefore give us two different results. The significant difference obtained using the *t*-test may, at least in part, be due to deviations from normality or homoscedasticity; on the other hand, the non-significant difference obtained using the *U* -statistic may be due to the fact that for fixed sample size, the power of a non-parametric test is lower than the corresponding parametric test. Given the *p* values obtained from both tests, and the fact that for samples of this size (84 and 101), the *t*-test is comparatively robust with respect to non-normality, I would be inclined to reject the null hypothesis.

Before accepting the results of the parametric *t*-test and rejecting the null hypothesis that there is no difference in size between the two locations, it is important to do some formal tests to determine if the model fits the assumption of normally distributed residuals and equal variances. Preliminary examination of the raw data suggested the data appeared roughly normal but there might be problems with variances (since the spread of data for The_Pas was much greater than for Cumberland). We can examine this more closely by looking at the residuals. An easy way to do so, is to fit a linear model and use the residual diagnostic plots:

```
AnovaModel.1 <- lm(fklngth ~ location, data=sturgeon)
par(mfrow = c(2, 2))
plot(AnovaModel.1)
```

**Figure 33.**




The first plot above shows the spread of the residuals around the estimated values for the two groups and allows us to get a feel for whether there are problems with the assumption of homogeneity of variances. If the variances were equal, the vertical spread of the two clusters of points should be about the same. The above plot shows that the vertical spread of the group with the smaller mean is greater than it is for the larger mean, suggesting again that there are problems with the variances. We can test this formally by examining the mean differences in the absolute value of the residuals.

The second graph above is a normal QQ plot (or probability plot) of the residuals of the model. Note that these generally fall on a straight line, suggesting there is no real problem with normality. We can do a formal test for normality on the residuals using the Shapiro-Wilk test.

```
shapiro.test(residuals(AnovaModel.1))


        Shapiro-Wilk normality test

data:  residuals(AnovaModel.1)
W = 0.9747, p-value = 0.001858
```

Hummm. The test indicates that the residuals are not normal. But, given that (a) the distribution is not very far (at least visually) from normal, and that (b) the number of observations in each location is reasonably large (i.e. >30), we do not need to be overly concerned with this violation of the normality assumption.:

How about equality of variance?

```
> require(car)
> levene.test(fklngth ~ location)

Levene's Test for Homogeneity of Variance
       Df F value    Pr(>F)
group   1  11.514 0.0008456 ***
      184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> fligner.test(fklngth ~ location)

        Fligner-Killeen test of homogeneity of variances

data:  fklngth by location
Fligner-Killeen:med chi-squared = 11.537, df = 1, p-value = 0.0006823

> require(lmtest)
> bptest(fklngth ~ location, varformula = ~
+   fitted.values(AnovaModel.1), studentize=TRUE, data=stur-
geon)

        studentized Breusch-Pagan test

data:  fklngth ~ location
BP = 8.8015, df = 1, p-value = 0.00301

>
```

The above are the results of three tests implemented in R (in the `car` and `lmtest` packages) that can be used to test for equal variances in t-tests or linear models involving only discontinuous or categorical independent variables. **Doing the three of them is overkill.** There is not much to prefer one test over another. Levene test is possibly the better known. It tests whether the mean of absolute values of the residuals differs among groups. The Fligner test is considered more robust to normality violations than Levene's and is preferred by some purists for this reason. It compares ranks of the absolute values of the residuals to test whether variances are equal. Finally, the Breusch-Pagan test has the advantage of being applicable to more linear models (it can deal with regression-type continuous independent variables, at least to some extent). It tests whether the studentized (i.e.

scaled by their sd estimate) squared residuals vary with the independent variables in a linear model. In this case, all 3 indicate that variances are unequal.

On the basis of these results, we conclude that there is evidence (albeit weak) to reject the null hypothesis of no difference in `fklngth` by `location`. We have modified the *t*-test to accommodate unequal variances, and are satisfied that the assumption of normally distributed residuals is sufficiently met. Thus, it appears that `fklngth` at Cumberland is greater than `fklngth` at The Pas.

R program for comparing two independent samples

```
#t-test assuming equal variances
t.test(fklngth~location, alternative='two.sided', conf.level=.95,
   var.equal=TRUE, data=sturgeon)

#t-test assuming unequal variances
t.test(fklngth~location, alternative='two.sided', conf.level=.95,
  var.equal=FALSE, data=sturgeon)

#non-parametric wilcoxon test
wilcox.test(fklngth ~ location, alternative="two.sided",
data=sturgeon)

# Diagnostic plots of residuals and tests of normality/homoscedas-
ticity

AnovaModel.1 <- lm(fklngth ~ location, data=sturgeon)
par(mfrow = c(2, 2))
plot(AnovaModel.1)
shapiro.test(residuals(AnovaModel.1))
require(car)
levene.test(fklngth ~ location)
fligner.test(fklngth ~ location)
require(lmtest)
bptest(fklngth ~ location, varformula = ~
  fitted.values(AnovaModel.1), studentize=TRUE, data=sturgeon)
```

## Bootstrap and permutation tests to compare 2 means

Bootstrap and permutation tests can be used to compare means (or other statistics) between pairs of samples. The general idea is simple, and it can be implemented in more ways than I can count.. Here, I use existing tools and the fact that a comparison of means can be construed as a test of a linear model. We will be able to use similar code later on when we fit more complex (but fun!) models.

```
library(boot)
```

The first section defines the function that I called bs that simply extracts coefficients from a fitted model:

```
# function to obtain model coefficients for each iteration
bs <- function(formula, data, indices) {
    d <- data[indices, ]
    fit <- lm(formula, data = d)
    return(coef(fit))
}
```

The second section with the boot() command is where the real work is done: take data in sturgeon, boosttrap R=1000 times, each time fit the model fklngth vs location, and keep the values calculated by the bs function.

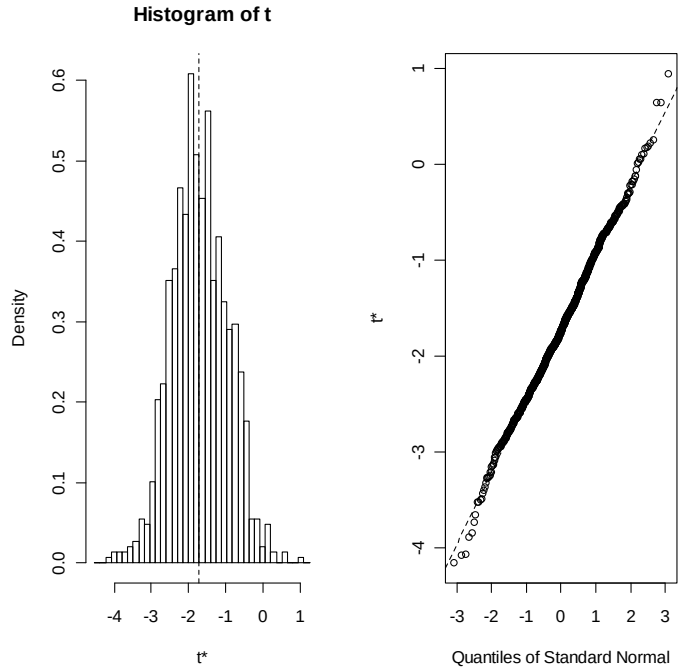
```
# bootstrapping with 1000 replications
results <- boot(data = sturgeon, statistic = bs, R = 1000,
    formula = fklngth ~ location)

# view results
results
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = sturgeon, statistic = bs, R = 1000, formula = fklngth ~
    location)


Bootstrap Statistics :
     original       bias    std. error
t1* 45.084391 -0.005634742   0.4164833
t2* -1.714546  0.012011519   0.7510652
```

So we get the original estimates for the two coefficients in this model : the mean at the first (alphabetical) location, Cumberland, and the difference in means between Cumberland and The Pas ). It is the second parameter, the difference between means, which is of interest here.

```
plot(results, index = 2)
```

**Figure 34.**




```
# get 95% confidence intervals

boot.ci(results, type = "bca", index = 2)

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "bca", index = 2)

Intervals :
Level       BCa
95%   (-3.052, -0.211 )
Calculations and Intervals on Original Scale
```

The 95% CI for the slope does not include 0.

Permutation tests for linear models can easily be done using the lmPerm library:

```
require(lmPerm)

mymodelProb <- lmp(fklngth ~ location, data = sturgeon,
    perm = "Prob")
```

The lmp() function does all the work for us. Here it is run with the option perm to control the stopping rule used. Option Prob stops the sampling when the estimated standard deviation of the p-value falls below some fraction of the estimated. It is one of many stopping rules

that one could use to do permutations on a subset of all the possibilities (because it would take foreeeever to do them all,, even on your fast machine).

```
summary(mymodelProb)

Call:
lmp(formula = fklngth ~ location, data = sturgeon, perm = "Prob")

Residuals:
      Min        1Q    Median        3Q       Max
-18.40921  -3.75370  -0.08439   3.76598  23.48055

Coefficients:
          Estimate Iter❶ Pr(Prob)❷
location1   0.8573 1117   0.0824 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 5.454 on 184 degrees of freedom
Multiple R-Squared: 0.02419,    Adjusted R-squared: 0.01889
F-statistic: 4.562 on 1 and 184 DF,  p-value: 0.03401 ❸
```

❶ In this case, the SPR stopping rule stopped after 1117 iterations. Note that this number will vary each time you run this snippet of code. These are random permutation results, so expect variability.

❷ The estimated probability associated to H0 is `0.0824`. The observed difference in fklngth between the two locations was larger than the permuted differences in about (1-0.0824=about 92%) of the 1117 cases. Mind you, 1117 permutations is not a large number, so small p values can't be expected to be very precise. If it is critical that you get more precise p values, more permutations would be needed. Unfortunately, there is no simple way to control that with lmPerm. Maybe in the next version?

❸ The rest is the standard output for the model fitted to the data, with the standard parametric test. Here the p-value, assuming all assumptions are met, is 0.034.

## Comparing the means of paired samples

In some experimental designs, individuals are measured twice: common examples are the measurement of the same individual at two different times during development, or of the same individual subjected to two different experimental treatments. In these cases, the two samples are not independent (they include the same individuals), and a paired comparison must be made.

The file `skulldat.Rdata` shows measurements of lower face width of 15 North American girls measured at age 5 and again at age 6 years (data from Newman and Meredith, 1956).

⌨Let's first run a standard t-test comparing the face width at age 5 and 6, not taking into account that the data are not independent and that they are consecutive measurements on the same individuals.

```
> t.test(width~age, alternative='two.sided',
+   conf.level=.95, var.equal=TRUE )

        Two Sample t-test

data:  width by age
t = -1.7812, df = 28, p-value = 0.08573
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.43000032  0.03000032
sample estimates:
mean in group 5 mean in group 6
       7.461333        7.661333

>
```

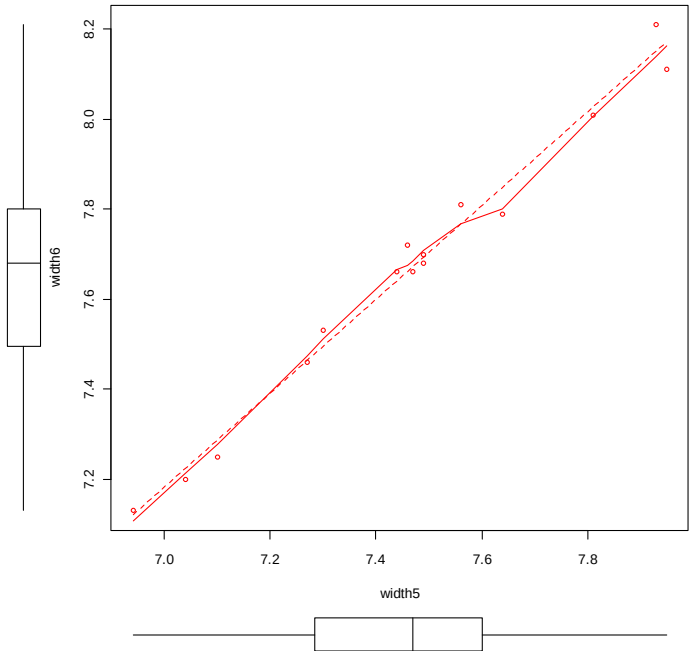
⌨Now, let's run the appropriate paired t-test. What do you conclude? Compare this with the previous result and explain any differences.

```
> t.test(skulldat$width5, skulldat$width6, alterna-
tive='two.sided',
+   conf.level=.95, paired=TRUE)

        Paired t-test

data:  skulldat$width5 and skulldat$width6
t = -19.5411, df = 14, p-value = 1.473e-11
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.2241710 -0.1798290
sample estimates:
mean of the differences
                 -0.202
```

The first analysis above assumes that the two samples of girls at age 5 and 6 are independent samples, whereas the second analysis assumes that the same girl is measured twice, once at age 5 and once at age 6 years.

Note that in the former case, we accept the null based on $p$=.05, but in the latter we reject the null. In other words, the appropriate (paired sample) test shows a very significant effect of age, whereas the inappropriate one does not. The reason is because there is a strong correlation between face width at age 5 and face width at age 6:

**Figure 35.**




with $r = .993$. In the presence of correlation, the standard error of the pairwise difference in face width at age 5 and 6 is much smaller than the standard error of the difference between the mean face width at age 5 and 6. Thus, the associated $t$-statistic will be much larger for a paired sample test, i.e. the power of the test is much greater, and the $p$ values are smaller.

Repeat the above procedure with the nonparametric alternative, the Wilcoxon signed-rank test. What do you conclude?

```
> wilcox.test(skulldat$width5, skulldat$width6, alterna-
tive='two.sided',
+   paired=TRUE)
Warning in wilcox.test.default(skulldat$width5, skulldat$width6, alternative
= "two.sided",  :
  cannot compute exact p-value with ties

        Wilcoxon signed rank test with continuity correction

data:  skulldat$width5 and skulldat$width6
V = 0, p-value = 0.0007211
alternative hypothesis: true location shift is not equal to 0

>
```

So, we reach the same conclusion as we did using the paired sample $t$-test and conclude there are significant differences in skull sizes of girls aged 5 and 6 (what a surprise!).

But, wait a minute. We have used two-tailed tests here. But, given what we know about how children grow, a one-tail hypothesis would be preferable. This can be done by changing the alternative option. One uses the alternative hypothesis to decide if it is "less" or greater". Here, we expect that if there is an effect (i.e the alternative hypothesis), width5 will be less than width6.

```
> t.test(skulldat$width5, skulldat$width6, alterna-
tive='less',
+   conf.level=.95, paired=TRUE)

        Paired t-test

data:  skulldat$width5 and skulldat$width6
t = -19.5411, df = 14, p-value = 7.363e-12
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
       -Inf -0.1837930
sample estimates:
mean of the differences
                -0.202

>
> wilcox.test(skulldat$width5, skulldat$width6, alterna-
tive='less',
+    exact=TRUE, paired=TRUE)
Warning in wilcox.test.default(skulldat$width5, skulldat$width6, alternative
= "less",  :
  cannot compute exact p-value with ties

        Wilcoxon signed rank test with continuity correction

data:  skulldat$width5 and skulldat$width6
V = 0, p-value = 0.0003606
alternative hypothesis: true location shift is less than 0
```

R program for paired samples

```
load(file.choose())
attach(skulldat)

t.test(width~age, alternative='two.sided',
  conf.level=.95, var.equal=TRUE )


t.test(skulldat$width5, skulldat$width6, alternative='two.sided',
  conf.level=.95, paired=TRUE)

scatterplot(width6~width5)

wilcox.test(skulldat$width5, skulldat$width6, alterna-
tive='two.sided',
  paired=TRUE)

t.test(skulldat$width5, skulldat$width6, alternative='less',
  conf.level=.95, paired=TRUE)

wilcox.test(skulldat$width5, skulldat$width6, alternative='less',
   exact=TRUE, paired=TRUE)
```

## References

Bumpus, H.C. (1898) The elimination of the unfit as illustrated by the introduced sparrow, *Passer domesticus*. Biological Lectures, Woods Hole Biology Laboratory, Woods Hole, 11 th Lecture: 209 - 226.

Newman, K.J. and H.V. Meredith. (1956) Individual growth in skeletal bigonial diameter during the childhood period from 5 to 11 years of age. Amer. J. Anat. 99: 157 - 187.

# Lab- Single classification (one-way) ANOVA

After completing this laboratory exercise, you should be able to:

- Use R to do a one-way parametric ANOVA with multiple comparisons
- Use R to test the validity of the parametric ANOVA assumptions
- Use R to perform a one-way non-parametric ANOVA
- Use R to transform your data so that the assumptions of parametric ANOVA are met.

## One-way ANOVA with multiple comparisons

The one-way ANOVA is the multi-sample analog of the *t*-test, which is used to compare two samples. It makes essentially the same assumptions, and in the case of two samples, is in fact mathematically equivalent to the *t*-test.

In 1960 - 1962, the Grand Rapids Dam was built on the Saskatchewan River upstream of Cumberland House. There are anecdotal reports that during dam construction, a number of large sturgeon were stranded and died in shallow pools. Surveys of sturgeon were carried out in 1954, 1958, 1965 and 1966 with fork length (`fklngth`) and round weight (`rdwght`) being recorded (not necessarily both measurements for each individual). These data are in the data file `dam10dat.Rdata`.

### Visualizing the data

Using `dam10dat.Rdata`, you must first change the data type of the numerical variable `year`, so that R recognizes that we wish to treat this variable as a factor variable and not a continuous variable.

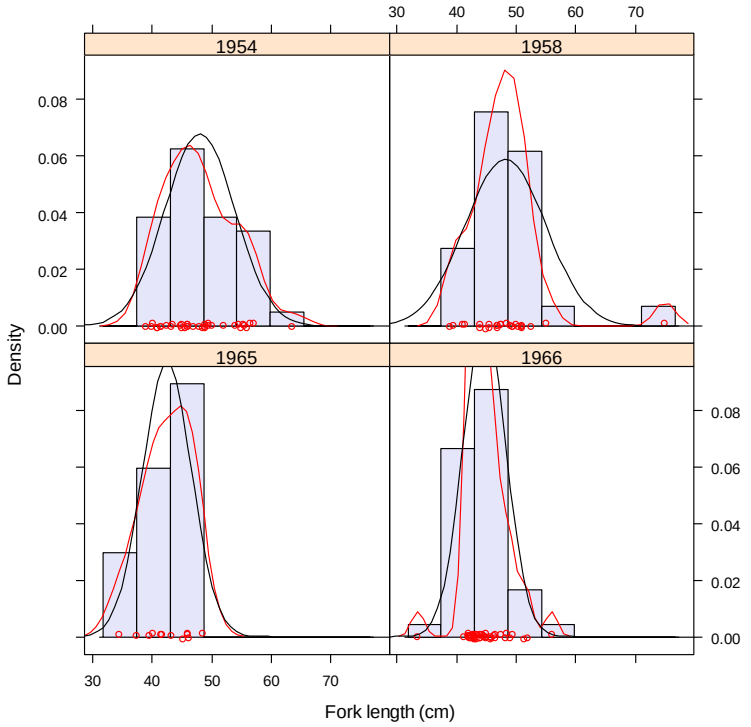
```
Dam10dat$year<-as.factor(Dam10dat$year)
ls.str()
attach(Dam10dat)
```

Next, have a look at the `fklngth` data, just as we did in the last lab for *t*-tests. Create a histogram with density line grouped by year to get a feel for what's happening with your data. What can you say about these data?

```
require(lattice)
histogram( ~ fklngth | year, as.table=TRUE,
```

```
xlab = "Fork length (cm)", type = "density",
panel = function(x, ...) {
      panel.histogram(x,  col="lavender",...)
     panel.densityplot(na.omit(x), col="red")
    panel.mathdensity(dmath = dnorm, col = "black",
            args = list(mean=mean(x),sd=sd(x)))
} )
```

**Figure 36.**




It appears as though there may have been a small drop in FKLNGTH after the construction of the dam, but the data are variable and the effects are not clear. There might also be some problems with normality in the 1954 and 1966 samples, and it looks as though there are outliers in the 1958 and 1966 samples. Let's proceed with testing the assumptions of the ANOVA by running the analysis and looking at the residuals.

R program for visually examining data

```
#load datafile dam10dat.rdata
load(file.choose())
#Check contents
ls.str()
#convert year from numeric to factor
Dam10dat$year<-as.factor(Dam10dat$year)
#verify that change was made
ls.str()
attach(Dam10dat)

#Plot data
require(lattice)
histogram( ~ fklngth | year, as.table=TRUE,
          xlab = "Fork length (cm)", type = "density",
          panel = function(x, ...) {
                   panel.histogram(x,  col="lavender",...)
              panel.densityplot(na.omit(x), col="red")
              panel.mathdensity(dmath = dnorm, col = "black",
                                args = list(mean=mean(x),sd=sd(x)))
          } )
```

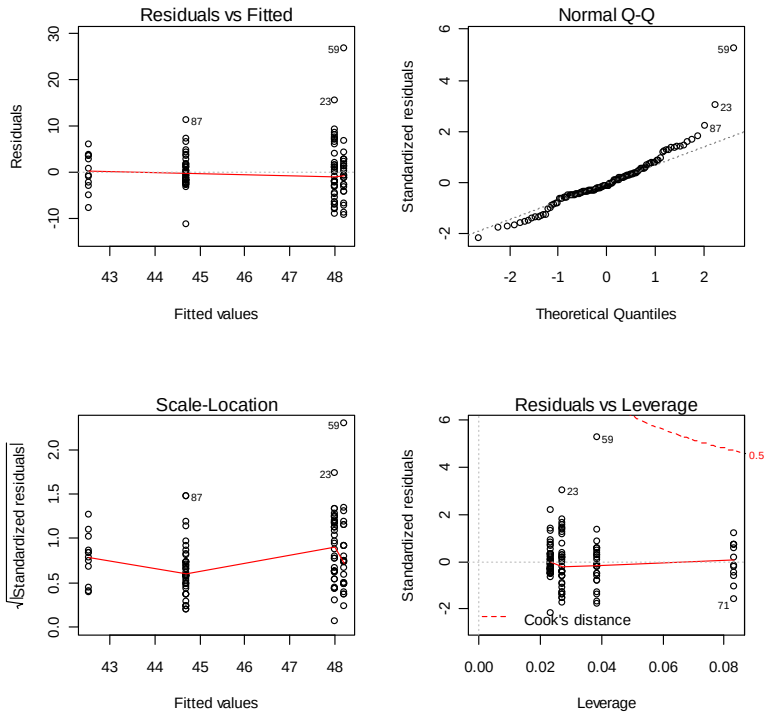
### Testing the assumptions of a parametric ANOVA

Parametric one-way ANOVAs have three major assumptions: (1) the residuals are normally distributed; (2) the error variance is the same for all groups (homoscedasticity); and (3) the residuals are independent. These assumptions must be tested before we can accept the results of any parametric ANOVA.

Carry out a one-way ANOVA on `fklngth` by `year` and produce the residual diagnostic plots

```
opar<-par(mfrow=c(2,2))
anova.model1<-lm(fklngth~year)
plot(anova.model1)
par(opar)
```

**Figure 37.**




Judging from the plots, it looks as though there are problems with both normality and variance heterogeneity. Note that there is one point (case 59) with large expected values and a large residual that appear to lie well off the line: this is the outlier we detected earlier. This point might be expected to inflate the variance for the group it belongs to.

Formal tests will tell us for certain if we should be concerned about normality and variance heterogeneity.

Perform a normality test on the residuals from the ANOVA.

```
shapiro.test(residuals(anova.model1))
```

```
        Shapiro-Wilk normality test

data:  residuals(anova.model1)
W = 0.9157, p-value = 1.63e-06
```

This test confirms our suspicions from the probability plot: the residuals are not normally distributed. Recall, however, that the power here is high, so only small deviations from normality are required to reject the null.

⌨Next, test for homoscedasticity:

```
require(car)
levene.test(fklngth ~ year)
```

```
Levene's Test for Homogeneity of Variance
       Df F value  Pr(>F)
group   3  2.8159 0.04234 *
      114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The probability value tells you that you can reject the null hypothesis that there is no difference in variances among years. Thus, we conclude there is evidence that the variances in the groups are not equal.

## Performing the ANOVA

Let's look at the results of the ANOVA, assuming for the moment that assumptions are met well enough :

```
> summary(anova.model1)

Call:
lm(formula = fklngth ~ year)

Residuals:
     Min       1Q   Median       3Q      Max
-11.2116  -2.6866  -0.7116   2.2103  26.7885

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  48.0243❶     0.8566  56.061  < 2e-16 ***
year1958      0.1872     1.3335   0.140  0.88859
year1965     -5.5077     1.7310  -3.182  0.00189 **
year1966     -3.3127     1.1684  -2.835  0.00542 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.211❷ on 114 degrees of freedom

Multiple R-squared: 0.1355,❸     Adjusted R-squared: 0.1128

F-statistic: 5.957 on 3 and 114 DF,  p-value: 0.0008246❹
```

❶Note the 4 coefficients printed. They can be used to obtain the predicted values for the model (i.e. the group means). The mean fklngth for the first year (1954) is 48.0243. The coefficients for the 3 other years are the difference between the mean for that year and for 1954. So, the mean for 1965 is (48.0243-5.5077=42.5166). For each estimated coefficient, there is a standard error, a t-value and associated probability (for H0 that the coefficient is 0). Note here that coefficients for 1965 and 1966 are both negative and significantly less than 0. Fish were smaller after the construction of the dam than in

1954. Take these p-values with a grain of salt: these are not corrected for multiple comparisons, and they constitute only a subset of the possible comparisons. In general, I pay little attention to this part of the output and look more at what comes next.

❷The square root of the variance of the residuals (observed minus fitted values) corresponds to the amount of variability that is unexplained by the models (here an estimate of how much size varied among fish, once corrected for differences among years)

❸The R-squared is the proportion of the variance of the dependent variable that can be explained by the model. Here the model explains only 13.5% of the variability. Size differences among year are relatively small compared to the ranges of sizes that can occur within years. This corresponds well to the visual impression left by the histograms of fklngth per year.

❹This is the p-value for the "omnibus" test, the test that all means are equal. Here it is much smaller than 0.05 and hence we would reject H0 and conclude that fklngth varies among the years.

The anova() command produces the standard ANOVA table that contains most of the same information.:

```
> anova(anova.model1)
Analysis of Variance Table

Response: fklngth
           Df  Sum Sq Mean Sq F value    Pr(>F)
year        3  485.26  161.75  5.9574 0.0008246 ***
Residuals 114 3095.30   27.15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

The total variability in fklngth sums of square is partitioned into what can be accounted for by year (485.26) and what is left unexplained as residual variability (3095.30). Year indeed explains (485.26/(3095.30+485.26)=.1355 or 13.55% of the variability). The mean square of the residuals is their variance..

### R program for One-way ANOVA

```
#One way ANOVA with multiple comparisons
#ANOVA à un critère de classification et comparaisons multiples

#load datafile Dam10dat.Rdata
load(file.choose())
#Check contents
ls.str()
#convert year from numeric to factor
Dam10dat$year<-as.factor(Dam10dat$year)
#verify that change was made
ls.str()
attach(Dam10dat)

#Plot data
require(lattice)
histogram( ~ fklngth | year, as.table=TRUE,
          xlab = "Fork length (cm)", type = "density",
          panel = function(x, ...) {
                   panel.histogram(x,  col="lavender",...)
              panel.densityplot(na.omit(x), col="red")
              panel.mathdensity(dmath = dnorm, col = "black",
                                args = list(mean=mean(x),sd=sd(x)))
          } )

#Fit anova model and plot residual diagnostics
opar<-par(mfrow=c(2,2)) #saves current par and sets graphic page
to hold 4 graphs
anova.model1<-lm(fklngth~year)
plot(anova.model1)
par(opar) #revert to old par

#Check normality of residuals
shapiro.test(residuals(anova.model1))

#Check homoscedasticity
require(car)
levene.test(fklngth ~ year)

summary(anova.model1)
anova(anova.model1)
```

## Performing multiple comparisons of means test

The pairwise.t.test() function can be used to compare means and adjust (or not) probabilities for multiple comparisons by choosing one of the options for p.adj:

```
pairwise.t.test(fklngth, year, p.adj = "none")
pairwise.t.test(fklngth, year, p.adj = "bonf")
pairwise.t.test(fklngth, year, p.adj = "holm")
pairwise.t.test(fklngth, year, p.adj = "fdr")
```

```
> pairwise.t.test(fklngth, year, p.adj = "none")
```

This compares means of all pairs, not adjusting probabilities.

```
        Pairwise comparisons using t tests with pooled SD
```

```
data:  fklngth and year

     1954   1958   1965
1958 0.8886 -      -
1965 0.0019 0.0022 -
1966 0.0054 0.0079 0.1996

P value adjustment method: none
```

Option "bonf" adjusts the p-values according to the Bonferroni correction. In this case, since there are 6 p-values calculated, it amounts to simply multiplying the uncorrected p-values by 6 (unless the result is above 1, in that case, let adj p = 1).

```
> pairwise.t.test(fklngth, year, p.adj = "bonf")
        Pairwise comparisons using t tests with pooled SD

data:  fklngth and year

     1954  1958  1965
1958 1.000 -     -
1965 0.011 0.013 -
1966 0.033 0.047 1.000

P value adjustment method: bonferroni
```

Option "holm" is the sequential Bonferroni correction, where the p-values are ranked from (i=1) smallest to (N) largest. The correction factor for p-values is then (N-i+1). Here, for example, we have N=6 pairs that are compared. The lowest uncorrected p-value is 0.0019 for 1954 vs 1965. The corrected p-value becomes 0.0019*(6-1+1)=0.011. The second lowest p-value is 0.0022. The corrected p/value is therefore 0.0022*(6-2+1)=0.011. For the highest p-value, the correction is (N-N+1)=1, hence it is equal to the uncorrected probability.

```
> pairwise.t.test(fklngth, year, p.adj = "holm")

        Pairwise comparisons using t tests with pooled SD

data:  fklngth and year

     1954  1958  1965
1958 0.889 -     -
1965 0.011 0.011 -
1966 0.022 0.024 0.399

P value adjustment method: holm
```

The "fdr" option is for controlling the false discovery rate. .

```
> pairwise.t.test(fklngth, year, p.adj = "fdr")

        Pairwise comparisons using t tests with pooled SD

data:  fklngth and year

     1954   1958   1965
1958 0.8886 -      -
1965 0.0066 0.0066 -
1966 0.0108 0.0119 0.2395

P value adjustment method: fdr
```

The four post-hoc tests here tell us the same thing: differences are all between two groups of years: 1954/58 and 1965/66, since all comparisons show differences between the 50's and 60's but no differences within the 50's or 60's. So, in this particular case, the conclusion is not affected by the choice of adjustment method.

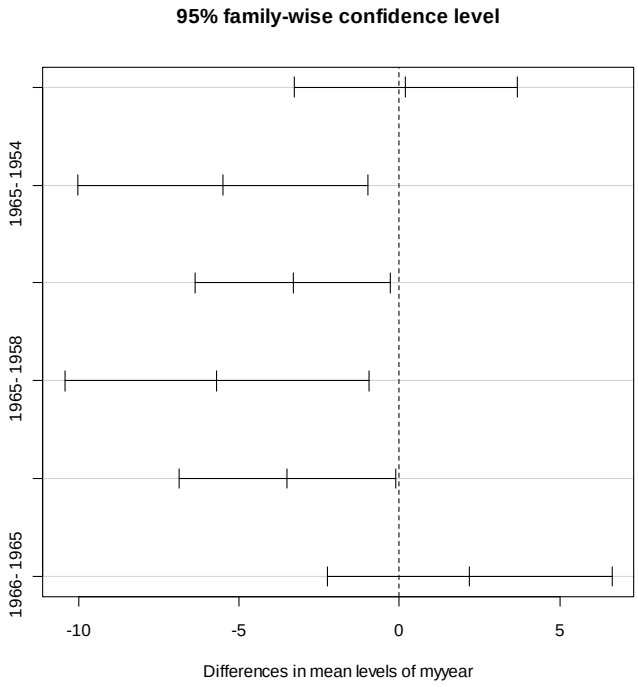
But in other situations, you will observe contradictory results. Which one to choose? Unadjusted p-values are certainly suspect when there are multiple tests. On the other hand, the traditional Bonferroni correction is very conservative, and becomes even more so when there are a large number of comparisons. Recent work suggest that the fdr approach may be a good compromise when there are a lot of comparisons.

The Tukey method of multiple comparisons is one of the most popular and is easily performed with R (note, however, that there is a pesky bug that manifests itself when the independent variable can look like a number rather than a factor, hence the little pirouette with paste to begin):

```
myyear <- as.factor(paste("", year))
TukeyHSD(aov(fklngth ~ myyear))

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = fklngth ~ myyear)

$myyear
                  diff        lwr        upr     p adj
1958- 1954  0.1872141  -3.289570  3.6639986 0.9990071
1965- 1954 -5.5076577 -10.021034 -0.9942809 0.0100528
1966- 1954 -3.3126964  -6.359223 -0.2661701 0.0274077
1965- 1958 -5.6948718 -10.436304 -0.9534397 0.0116943
1966- 1958 -3.4999106  -6.875104 -0.1247171 0.0390011
1966- 1965  2.1949612  -2.240630  6.6305526 0.5710111
```

```
plot(TukeyHSD(aov(fklngth ~ myyear)))
```

**Figure 38.**




The confidence intervals, corrected for multiple tests by the Tukey method, are plotted for differences among years. Unfortunately, the labels are not all printed because they would overlap, but the order is the same as in the preceeding table.

The library multcomp can produce a better plot version, but requires a bit more code:

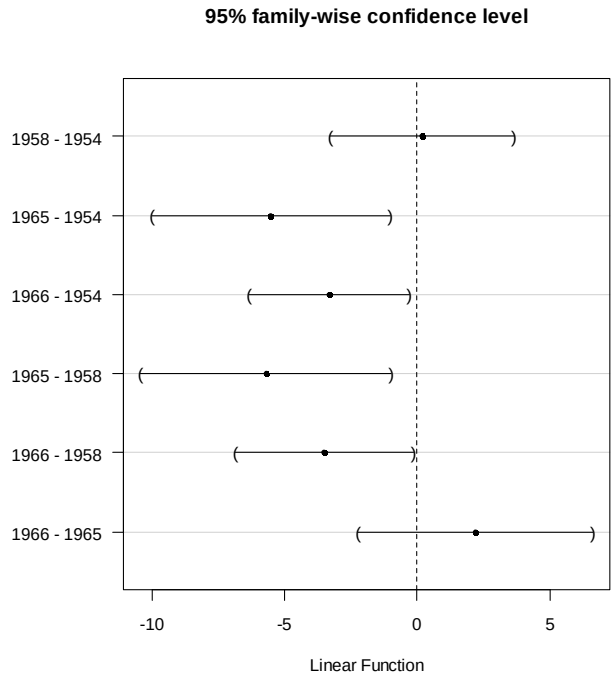
```
# Alternative way to compute Tukey multiple comparisons
# set up a one-way ANOVA

anova.fkl.vs.year <- aov(aov(fklngth ~ year))

# set up all-pair comparisons for factor `year'
library(multcomp)
meandiff <- glht(anova.fkl.vs.year, linfct = mcp(year =
"Tukey"))

confint(meandiff)
```

```
 Simultaneous Confidence Intervals

Multiple Comparisons of Means: Tukey Contrasts


Fit: aov(formula = aov(fklngth ~ year))

Estimated Quantile = 2.5936
95% family-wise confidence level


Linear Hypotheses:
                  Estimate lwr      upr
1958 - 1954 == 0   0.1872  -3.2713   3.6457
1965 - 1954 == 0  -5.5077  -9.9973  -1.0180
```

```
1966 - 1954 == 0  -3.3127  -6.3432  -0.2822
1965 - 1958 == 0  -5.6949 -10.4114  -0.9783
1966 - 1958 == 0  -3.4999  -6.8574  -0.1424
1966 - 1965 == 0   2.1950  -2.2173   6.6073
```

```
old.par <- par(mai = c(1, 1.25, 1, 1))
plot(meandiff)
par(old.par)
```

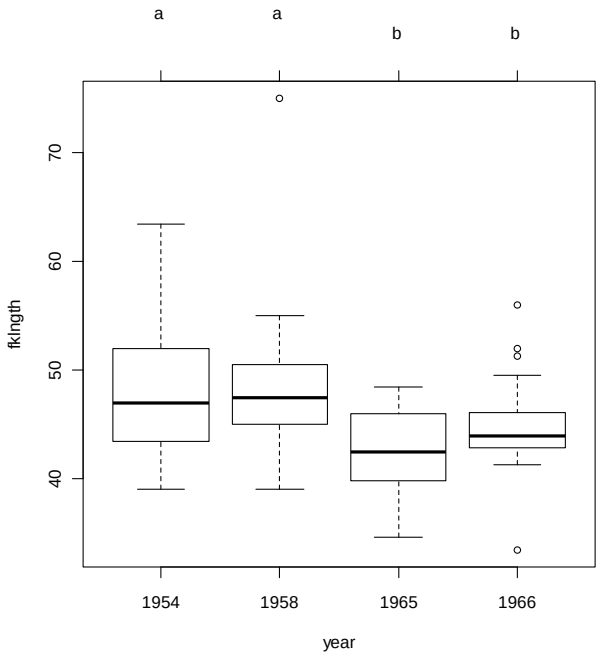
**Figure 39.**




This is slightly better, but it would be even better to plot the means and their confidence intervals, adjusting their width for multiple comparisons:

```
# Compute and plot means and Tukey CI
means <- glht(anova.fkl.vs.year, linfct = mcp(year =
"Tukey"))
cimeans <- cld(means)
# use sufficiently large upper margin
old.par <- par(mai = c(1, 1, 1.25, 1))
# plot
plot(cimeans)
par(old.par)
```

**Figure 40.**




Note the letters appearing on top. Years labelld with the same letter do not differ significantly.

## R commands for One-way ANOVA with post-hoc tests

```
#################################################################
# Performing multiple comparisons of means test

# Multiple comparisons

# These require only the standard stats library automatically
loaded
pairwise.t.test(fklngth, year, p.adj = "none")
pairwise.t.test(fklngth, year, p.adj = "bonf")
pairwise.t.test(fklngth, year, p.adj = "holm")
pairwise.t.test(fklngth, year, p.adj = "fdr")

#  Tukey HSD: note that
#  you need an aov() model and that
#  it does not handle factors that look like numbers, here I add a
space before the year
#  Note also that package HH interferes with this function, use
detach(package:HH)

myyear <- as.factor(paste("", year))
TukeyHSD(aov(fklngth ~ myyear))
plot(TukeyHSD(aov(fklngth ~ myyear)))


# Alternative way to compute Tukey multiple comparisons
# set up a one-way ANOVA
anova.fkl.vs.year <- aov(aov(fklngth ~ year))

# Compute and plot CI for differences between means (Tukey)
library(multcomp)
meandiff <- glht(anova.fkl.vs.year, linfct = mcp(year = "Tukey"))
confint(meandiff)
old.par <- par(mai = c(1, 1.25, 1, 1))
plot(meandiff)
par(old.par)

# Compute and plot means and Tukey CI
means <- glht(anova.fkl.vs.year, linfct = mcp(year = "Tukey"))
cimeans <- cld(means)
# use sufficiently large upper margin
old.par <- par(mai = c(1, 1, 1.25, 1))
# plot
plot(cimeans)
par(old.par)
```

⌨ Repeat the above ANOVA and multiple comparisons of means tests (use the Bonferroni option) using `rdwght` instead of `fklngth`. What do you conclude? Do your analyses support the hypothesis that dam construction resulted in the loss of a significant number of older, larger sturgeon?

```
            Df  Sum Sq Mean Sq F value  Pr(>F)
year         3  1733.2   577.7  4.3302 0.00701 **
Residuals   80 10673.4   133.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
34 observations deleted due to missingness

Pairwise comparisons using t tests with pooled SD

data:  rdwght and year

     1954  1958  1965
1958 1.000 -     -
1965 0.028 1.000 -
1966 0.034 1.000 1.000

P value adjustment method: bonferroni
```

Once again we see a significant difference in means among years, but this time we see only a significant difference between data from 1954 and data from 1965 and 1966. Data from 1958 do not differ from the other means. Try repeating the post-hoc contrasts using the Tukey tests and confirm to yourself that, in this case, they all yield the same results.

So, for this variable, there is no evidence that sturgeon in 1958 differed from those caught in either 1965 or 1966. Part of the problem is the small sample size and high variance (for `rdwght`) of the 1958 and 1965 samples, so the power of the comparisons are lower than in the case of `fklngth` (especially the 1958-65 comparison), where the within-group variances are smaller.

Note that for the purpose of saving space we did not check whether the `RDWGHT` data met the assumptions of the ANOVA. However, this should *always* be done before the results of the anova are accepted conclusively!

On the whole, our analyses supports the hypothesis that construction of the E.B. Campbell Dam resulted in the loss of a number of old, large, sturgeon.

### R commands for One-way ANOVA on RDWGHT with post-hoc tests

```
# Run anova on rdwght values
aov.rdwght.year <- aov(rdwght ~ year, Dam10dat, na.action =
na.exclude)
summary(aov.rdwght.year)
opar <- par(mfrow = c(2, 2))
plot(aov.rdwght.year)
par(opar)
pairwise.t.test(rdwght, year, p.adj = "bonf")
```

## Data transformations and non-parametric ANOVA

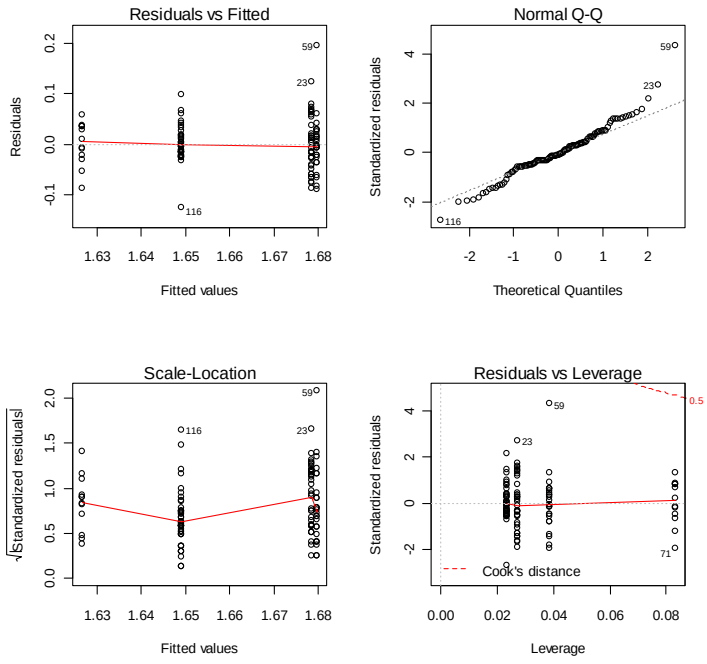
In the above example to examine differences in `fklngth` among `years`, we detected evidence of non-normality and variance heterogeneity. If the assumptions underlying a parametric ANOVA are not valid, there are several options: (1) if sample sizes in each group are reasonably large, parametric ANOVA is reasonably robust with respect to the normality assumption, for the same reason that the t-test is, so the results are probably not too bad; (2) we can transform the data; (3) we can go the non-parametric route.

⌨ Repeat the one-way ANOVA in the section above, but this time run the analysis on the $\log_{10}$ `fklngth`. With this transformation, do some of the problems encountered previously disappear?

```
# Fit anova model on log10 of fklngth and plot residual diagnostics
opar <- par(mfrow = c(2, 2))
anova.model2 <- lm(log10(fklngth) ~ year)
plot(anova.model2)
par(opar)
```

Looking at the residuals, we have:

**Figure 41.**



Things look barely better than before. Running the Wilks-Shapiro test for normality on the residuals, we get:

```
Shapiro-Wilk normality test

data:  residuals(anova.model2)
W = 0.962, p-value = 0.002048
```

Running Levene's test on the absolute value of the residuals we get:

```
Levene's Test for Homogeneity of Variance
       Df F value  Pr(>F)
group   3  2.6611 0.05148 .
      114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
1
```

So, it would appear that we still have some problems with the assumption of normality and are just on the border line of meeting the assumption of homogeneity of variances. You have several choices here: 1) try to find a different transformation to satisfy the assumptions, 2) assume the data are close enough to meeting the assumptions, or 3) perform a non-parametric ANOVA.

R commands for ANOVA on Log10 transformed data

```
# Fit anova model on log10 of fklngth and plot residual diagnos-
tics
opar <- par(mfrow = c(2, 2))
anova.model2 <- lm(log10(fklngth) ~ year)
plot(anova.model2)
par(opar)

# Check normality of residuals
shapiro.test(residuals(anova.model2))

# Check homoscedasticity
require(car)
levene.test(log10(fklngth) ~ year)

summary(anova.model2)
anova(anova.model2)
```

The most commonly used non-parametric analog of the parametric one-way ANOVA is the Kruskall-Wallis one-way ANOVA. Perform a Kruskall-Wallis one-way ANOVA of `fklngth`, and compare these results to the parametric analysis above. What do you conclude?

```
kruskal.test(fklngth, year)

        Kruskal-Wallis rank sum test

data:  fklngth and year
Kruskal-Wallis chi-squared = 15.7309, df = 3, p-value = 0.001288
```

So, the conclusion is the same as with the parametric ANOVA: we reject the null that the mean rank is the same for each year. Thus, despite violation of one or more assumptions, the parametric analysis is telling us the same thing as the non-parametric analysis: the conclusion is, therefore, quite robust.

## Dealing with outliers

Our preliminary analysis of the relationship between `FLNGTH` and `Year` suggested there might be some outliers in the data. These were evident in the box plots of `FKLNGTH` by `YEAR` and flagged as cases 59, 23 and 87 in the residual probability plot and residual-fit plot. In general, you have to have very good reasons for removing outliers from a data set (e.g., you know there was a mistake made in the data collection/entry). However, it is often useful to know how the analysis changes if you remove the outliers from the data set.

Repeat the original ANOVA of `FKLNGTH` by `YEAR` but work with a subset of the data without the outliers.

Have any of the conclusions changed?

```
aov.Damsubset <- aov(fklngth ~ as.factor(year), Dam10dat,
    subset = c(-23, -59, -87)) # removes obs 23, 59 and 87

summary(aov.Damsubset)
```

```
                Df  Sum Sq Mean Sq F value   Pr(>F)
as.factor(year)   3  367.51  122.50  6.8942 0.000267 ***
Residuals       111 1972.36   17.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
shapiro.test(residuals(aov.Damsubset))
```

```
        Shapiro-Wilk normality test

data:  residuals(aov.Damsubset)
W = 0.9853, p-value = 0.2448
```

```
levene.test(log10(fklngth) ~ year, Dam10dat,
     subset = c(-23, -59, -87))
```

```
Levene's Test for Homogeneity of Variance
      Df F value  Pr(>F)
group   3  3.8144 0.01206 *
      111
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Elimination of three outliers, in this case, makes things better in terms of the normality assumption, but does not improve the variances. Moreover, the fact that the conclusion drawn from the original ANOVA with outliers retained does not change upon their removal reinforces the fact that there is no good reason to remove the points..

R commands to rerun ANOVA and assumption tests on subsetted data set

```
aov.Damsubset <- aov(fklngth ~ as.factor(year), Dam10dat,
    subset = c(-23, -59, -87)) # removes obs 23, 59 and 87
summary(aov.Damsubset)
shapiro.test(residuals(aov.Damsubset))
levene.test(log10(fklngth) ~ year, Dam10dat,
    subset = c(-23, -59, -87))
```

## Permutation test

R commands for permutation test in one-way ANOVA

```
#################################################################
#  Permutation Test for one-way ANOVA
#  modified from code written by David C. Howell
#  http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Permuta-
tion%20Anova/PermTestsAnova.html

mod1 <- lm(fklngth ~ year) # Standard Anova
ANOVA <- summary(aov(mod1)) # Save summary to variable
F.factor1 <- ANOVA[[1]]$"F value"[1] # Save observed F value

# Print standard ANOVA results
cat(" The standard ANOVA for these data follows ",
    "\n")
print(ANOVA, "\n")
cat("\n")
cat("\n")

print("Resampling as in Manly with unrestricted sampling of obser-
vations. ")
# Now start resampling
nreps <- 500
Fboot <- numeric(nreps) # initalize vector to receive bootstrap
values
Fboot[1] <- F.factor1
for (i in 2:nreps) {
    newdependent <- sample(fklngth, length(fklngth)) # randomize
dep var
    mod2 <- lm(newdependent ~ year) # refit model
    b <- summary(aov(mod2))
    Fboot[i] <- b[[1]]$"F value"[1] # store F stats

}
probyear <- length(Fboot[Fboot >= F.factor1])/nreps
cat(" The probability value for year is ", probyear,
    "\n")
```

# Lab- Multiway ANOVA: factorial and nested designs

After completing this laboratory exercise, you should be able to:

- Use R to do parametric ANOVAs for 2-way factorial designs with replication.
- Use R to do 2-way factorial design ANOVA without replication
- Use R to do parametric ANOVAs for nested designs with replication.
- Use R to do non-parametric 2-way ANOVAs
- Use R to do multiway pairwise comparisons
-

Be aware that there are a large number of possible ANOVA designs, many of which can be handled by S-PLUS: this laboratory is designed only to introduce you to some of the more common ones.

## Two-way factorial design with replication

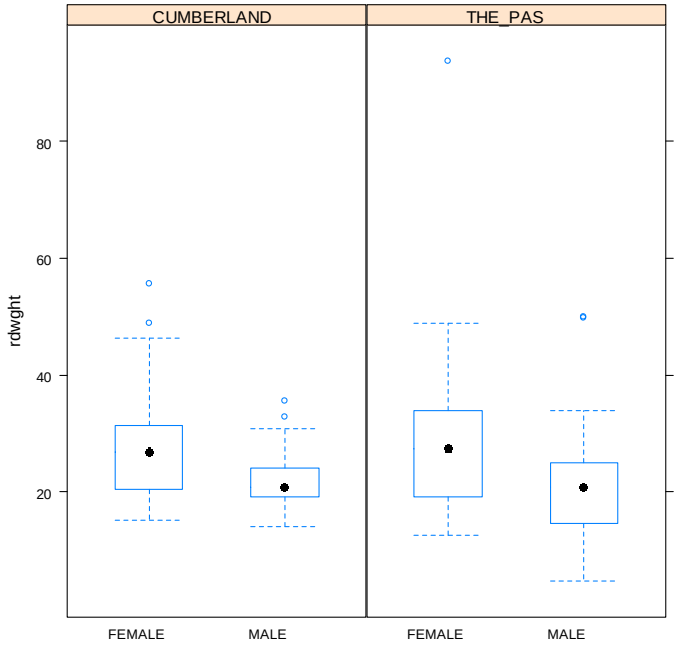
Many experiments are designed to investigate the joint effects of several different factors: in a two-way ANOVA, we examine the effect of two factors, but in principle the analysis can be extended to three, four or even five factors, although interpreting the results from 4- and 5-way ANOVAs can be very difficult.

Suppose that we are interested in the effects of two factors: `location` (Cumberland House and The Pas) and `sex` (male or female) on sturgeon size (data can be found in `Stu2wdat.Rdata`). Note that because the sample sizes are not the same for each group, this is an unbalanced design. Note also that there are missing data for some of the variables, meaning that not every measurement was made on every fish.

### Fixed effects ANOVA (Model I)

⌨ Begin by having a look at the data by generating box plots of `rdwght` for `sex` and `location` from the file `Stu2wdat.Rdd`.

**Figure 42.**




From this, it appears as though females might be larger at both locations. It's difficult to get an idea of whether fish differ in size among the two locations. The presence of outliers on these plots suggests there might be problems meeting normality assumptions for the residuals.

Generate summary statistics for `rdwght` by `sex` and `location`.

```
: FEMALE
: CUMBERLAND
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  15.10   20.40   26.80   27.37   31.40   55.60
------------------------------------------------------------
: MALE
: CUMBERLAND
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  14.00   19.22   20.85   22.14   23.90   35.60
------------------------------------------------------------
: FEMALE
: THE_PAS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.54   19.14   27.39   27.98   33.88   93.72
------------------------------------------------------------
: MALE
: THE_PAS
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.73   14.63   20.79   20.65   24.94   49.94
```

The summary statistics confirm our interpretation of the box plots: females appear to be larger than males, and differences in fish size among locations are small.

⌨Using the file `Stu2wdat.Rdata`, do a two-way factorial ANOVA:

```
# Fit anova model and plot residual diagnostics
# but first, save current par and set graphic page to hold 4
graphs
opar <- par(mfrow = c(2, 2))

anova.model1 <- lm(rdwght ~ sex + location + sex:location,
    contrasts = list(sex = contr.sum, location = contr.sum),
    data = Stu2wdat)
anova(anova.model1)
```

```
Analysis of Variance Table

Response: rdwght
             Df  Sum Sq Mean Sq F value    Pr(>F)
sex           1  1839.6  1839.6 18.6785 2.569e-05 ***
location      1     4.3     4.3  0.0433    0.8355
sex:location  1    48.7    48.7  0.4944    0.4829
Residuals   178 17530.4    98.5
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Trap.** Be careful here. R gives you the sequential sums of squares (Type I) and associated Mean squares and probabilities. These are **not** to be trusted unless the design is perfectly balanced. In this case, there are varying numbers of observations across sex and location combinations and therefore the design is not balanced.

What you want are the partial sums of squares (type III). The easiest way to get them is to use the Anova() function in the car package (note the subtle difference, Anova() is not the same as anova(), remember case matters in R.). However, this is not enough by itself. To get the proper values for the type III sums of square, one also needs to specify contrasts, hence the cryptic contrasts = list(sex = contr.sum, location = contr.sum) above...

```
require(car)
Anova(anova.model1, type = 3)
```

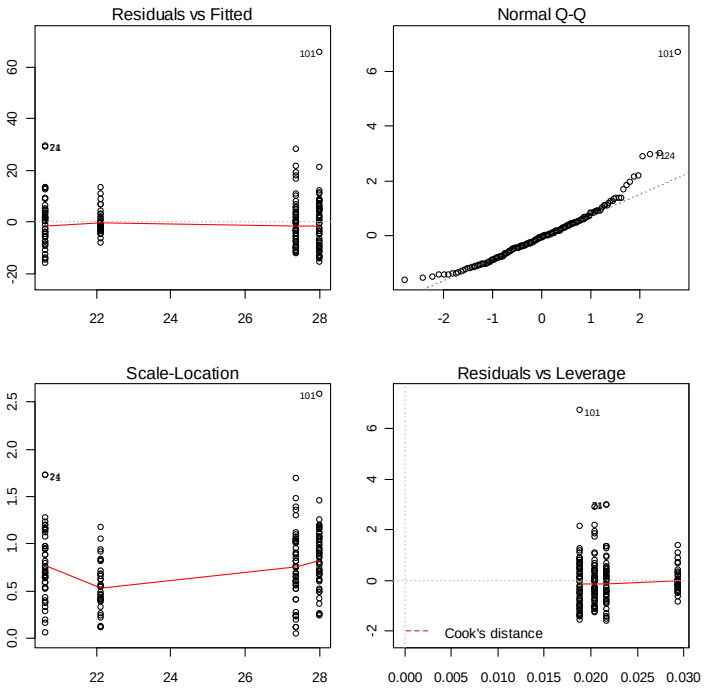
```
Anova Table (Type III tests)

Response: rdwght
             Sum Sq  Df  F value  Pr(>F)
(Intercept)   36716   1 372.8078 < 2e-16 ***
sex             550   1   5.5797 0.01925 *
location          9   1   0.0942 0.75924
sex:location     49   1   0.4944 0.48288
Residuals     17530 178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

On the basis of the ANOVA, we accept 2 null hypotheses: (1) that the effect of `sex` (if any) does not depend on `location` (no interaction), and (2) that there is no difference in the size of sturgeon (pooled over `sex`) between the two `locations`. On the other hand, we reject the

null hypothesis that there is no difference in size between male and female sturgeon (pooled over `location`), precisely as expected from the graphs.

```
plot(anova.model1)
```

**Figure 43.**




As usual, we cannot accept the above results without first ensuring that the assumptions of ANOVA are met. Examination of the residuals plots above shows that the residuals are reasonably normally distributed, with the exception of three outliers flagged on the QQ plot (cases 101, 24, & 71). The residuals vs fit plot shows that the spread of residuals is about equal over the range of the fitted values, again with the exception of a few cases. When we test for normality of residuals we get:

```
shapiro.test(residuals(anova.model1))
```

```
        Shapiro-Wilk normality test

data:  residuals(anova.model1)
W = 0.8721, p-value = 2.619e-11
```

So, there is evidence of non-normality in the residuals.

We will use the Levene's test to examine the assumption of homogeneity of variances, just as we did with the 1-way anova.

```
require(car)
levene.test(rdwght ~ sex * location, data = Stu2wdat)
```

```
Levene's Test for Homogeneity of Variance
       Df F value  Pr(>F)
group   3  3.8526 0.01055 *
      178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

If the assumption of homogeneity of variances was valid, we would be accepting the null that the mean of the absolute values of residuals does not vary among levels of `sex` and `location` (i.e., `group`). The above table shows that the hypothesis is rejected and we conclude there is evidence of heteroscedascticity. All in all, there is good evidence that several important assumptions have been violated. However, whether these violations are sufficiently large to invalidate our conclusions remains to be seen.

## R commands for a Two-way factorial ANOVA

```
# load datafile Stu2wdat.Rdata
# load(file.choose()) # local version
load(url("http://www.antoinemorin.com/biostats/2010/
Stu2wdat.Rdata"))


# Check contents
ls.str()

#########################################################
# Two-way factorial design with replication

# Make boxplot of rdwght vs sex and location
require(lattice)
bwplot(rdwght ~ sex | location, data = Stu2wdat)

# Compute summary statistics by sex and location, but first remove
NA

mydf <- data.frame(rdwght = Stu2wdat$rdwght, sex = Stu2wdat$sex,
    location = Stu2wdat$location)
mydf.nona <- mydf[complete.cases(mydf), ]
attach(mydf.nona)
by(rdwght, list(sex, location), summary)


# Fit anova model and plot residual diagnostics
# but first, save current par and set graphic page to hold 4
graphs
opar <- par(mfrow = c(2, 2))

anova.model1 <- lm(rdwght ~ sex + location + sex:location,
    contrasts = list(sex = contr.sum, location = contr.sum),
    data = Stu2wdat)
anova(anova.model1)

require(car)
Anova(anova.model1, type = 3)
plot(anova.model1)

# Check normality of residuals
shapiro.test(residuals(anova.model1))

# Check homoscedasticity
require(car)
levene.test(rdwght ~ sex * location, data = Stu2wdat)

par(opar)
```

⌨ Repeat this procedure using the data file `Stu2mdat.Rdata`. Now what do you conclude? Suppose you wanted to compare the sizes of males and females: in what way would these comparisons differ between `Stu2wdat.Rdata` and `Stu2mdat.Rdata`?

Note that in this case, we see that at Cumberland House, females are larger than males, whereas the opposite is true in The Pas (you can confirm this observation by generating summary statistics). What happens with the ANOVA (remember, you want Type III sum of squares)?

```
Anova Table (Type III tests)

Response: rdwght
             Sum Sq  Df   F value    Pr(>F)
(Intercept)  106507   1 1081.4552 < 2.2e-16 ***
sex             49   1    0.4944    0.4829
location         9   1    0.0891    0.7656
sex:location   1745   1   17.7220 4.051e-05 ***
Residuals     17530 178
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

In this case, the interaction term `sex:location` is significant but the main effects are not significant.

**R commands for a 2-way ANOVA with Stu2mdat**

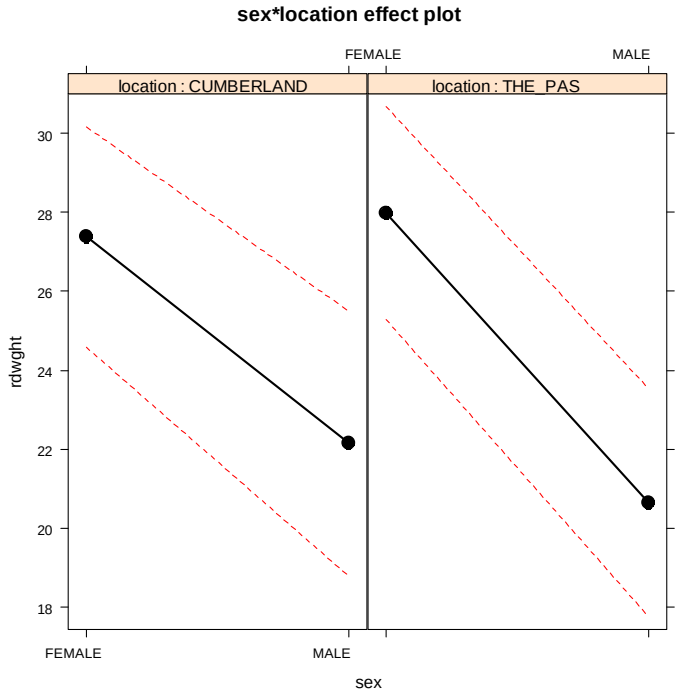
```
anova.model2 <- lm(rdwght ~ sex + location + sex:location,
    contrasts = list(sex = contr.sum, location = contr.sum),
    data = Stu2mdat)
Anova(anova.model2, type = "III")
```

⌨ You might find it useful here to generate plots for the two data files to compare the interactions between `sex` and `location`. The effect plot shows the relationship between means for each combination of factors (also called cell means). Generate an effect plot for the two models using the allEffects command from the effects library:

```
library(effects)
allEffects(anova.model1)
 model: rdwght ~ sex + location + sex:location

 sex*location effect
              location
sex            CUMBERLAND   THE_PAS
  FEMALE           27.37347     27.97717
  MALE             22.14118     20.64652
```

```
plot(allEffects(anova.model1), 'sex:location')
```

**Figure 44.**

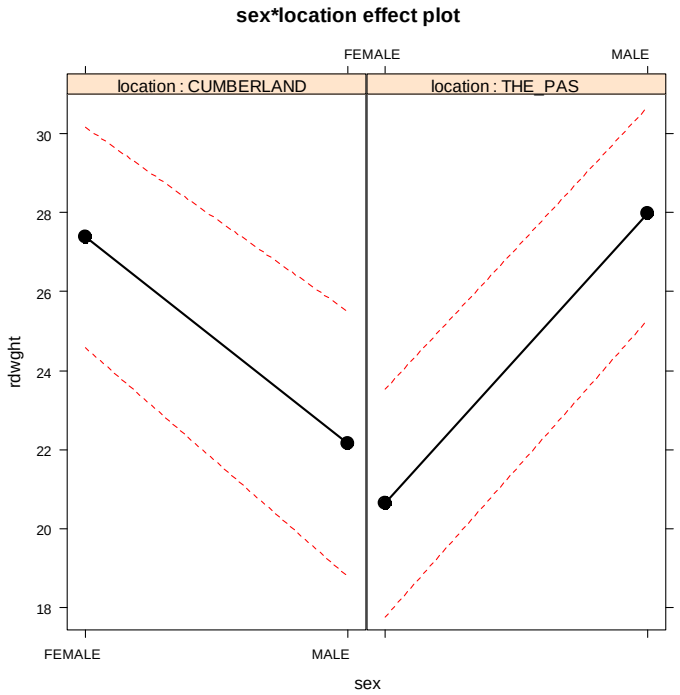



```
allEffects(anova.model2)
```

```
 model: rdwght ~ sex + location + sex:location

 sex*location effect
            location
sex          CUMBERLAND   THE_PAS
  FEMALE        27.37347     20.64652
  MALE          22.14118     27.97717
```

```
plot(allEffects(anova.model2), 'sex:location')
```

**Figure 45.**




There is a very large difference between the results from `Stu2wdat` and `Stu2mdat`. In the former case, because there is no significant interaction, we can essentially pool over the levels of factor 1 (`sex`, say) to test for the effects of `location`, or over the levels of factor 2 (`location`) to test for the effects of `sex`. In fact, if we do so and simply run a one-way ANOVA on the `Stu2wdat` data with `sex` as the grouping variable, we get:

```
Anova(aov(rdwght ~ sex, data = Stu2wdat), type = 3)
Anova Table (Type III tests)

Response: rdwght
             Sum Sq  Df F value    Pr(>F)
(Intercept)  78191   1 800.440 < 2.2e-16 ***
sex           1840   1  18.831 2.377e-05 ***
Residuals    17583 180
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that here the residual sum of squares (17583) is only slightly higher than for the 2-way model (17530), simply because, in the 2-way model, only a small fraction of the explained sums of squares is due to the `location` main effect or the `sex:LOCATION` interaction.

On the other hand, if you try the same trick with `stu2mdat`, you get:

```
Anova(aov(rdwght ~ sex, data = Stu2mdat), type = 3)
```

```
Anova Table (Type III tests)

Response: rdwght
            Sum Sq  Df  F value Pr(>F)
(Intercept)  55251   1 515.0435 <2e-16 ***
sex            113   1   1.0571 0.3053
Residuals    19309 180
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, the residuals sum of squares (19309) is much larger than in the 2-way model (17530), because most of the explained sums of squares is due to the interaction. Note that if we did this, we would conclude that male and female sturgeons don't differ in size. But in fact they do: it's just that the difference is in different directions, depending on location. This is why it is always dangerous to try and make too much of main effects in the presence of interactions!

### Mixed effects ANOVA (Model III)

We have neglected an important component in the above analyses, and that is related to the type of ANOVA model we wish to run. In this example, `Location` is a random effect, whereas `sex` is a fixed effect (because it is "fixed" biologically), and so this model should be treated as a mixed model (Model III) ANOVA.

Note that in these analyses, R treats analyses by default as Model I ANOVA, so that the main effects and the interaction are tested over the residuals mean square. Recall, however, that in a Model III ANOVA, main effects are tested over the interaction mean square or the pooled interaction mean square and residual mean square (depending on which statistician you consult!)

Working with the `Stu2wdat` data, rebuild the ANOVA table for `rdwght` based on the knowledge that `location` is a random factor and `sex` is a fixed factor. To do this, you need to recalculate the F-ratios for the `sex` and `location` terms using the `sex:location` interaction mean square instead of the residual mean square. This is most easily accomplished by hand, making sure you are working with the Type III Sums of squares ANOVA table.

```
Anova Table (Type III tests)

Response: rdwght
              Sum Sq  Df   F value    Pr(>F)
(Intercept)  106507   1 1081.4552 < 2.2e-16 ***
sex            1745   1   17.7220 4.051e-05 ***
location          9   1    0.0891    0.7656
sex:location     49   1    0.4944    0.4829
Residuals     17530 178
---
```

For sex, the new ratio of mean squares is (1745/1)/(49/1)=35.6

To assign a probability to the new F-value, enter the following in the commands window: `1-pf(F, df1, df2)`, where `F` is the newly calculated F-value, and `df1` and `df2` are the degrees of freedom of the numerator (`sex`) and denominator (`SEX:location`), respectively.

```
 1-pf(35.6,1,1)
[1] 0.1057152
```

Note that the *p* value for `sex` is now non-significant. This is because the error MS of the initial ANOVA is smaller than the interaction MS, but mostly because the number of degrees of freedom of the numerator of the F test has dropped from 178 to 1. In general, a drop in the denominator degrees of freedom makes it much more difficult to reach significance.

## 2-way factorial ANOVA without replication

In some experimental designs, there are no replicates within data cells: perhaps it is simply too expensive to obtain more than one datum per cell. A 2-way ANOVA is still possible under these circumstances, but there is an important limitation.

Because there is no replication within cells, there is no error variance: we have simply a row sum of squares, a column sum of squares, and a remainder sum of squares. This has important implications: if there is an interaction, only a model II ANOVA can be entirely tested, while in a Model III ANOVA, only the fixed effect can be tested (over the remainder MS); for Model I ANOVAs, or for random effects in Model III ANOVAs, it is not appropriate to test main effects over the remainder unless we are sure there is no interaction.

A limnologist studying Round Lake in Algonquin Park takes a single temperature (`temp`) reading at 10 different depths (`depth`, in m) at four times (`date`) over the course of the summer. Her data are shown in `Nr2wdat.Rdata`.

Do a two-way unreplicated ANOVA using `temp` as the dependent variable, `date` and `depth` as the factor variables (you will need to recode `depth` to tell R to treat this variable as a factor). Note that there is no interaction term included in this model.

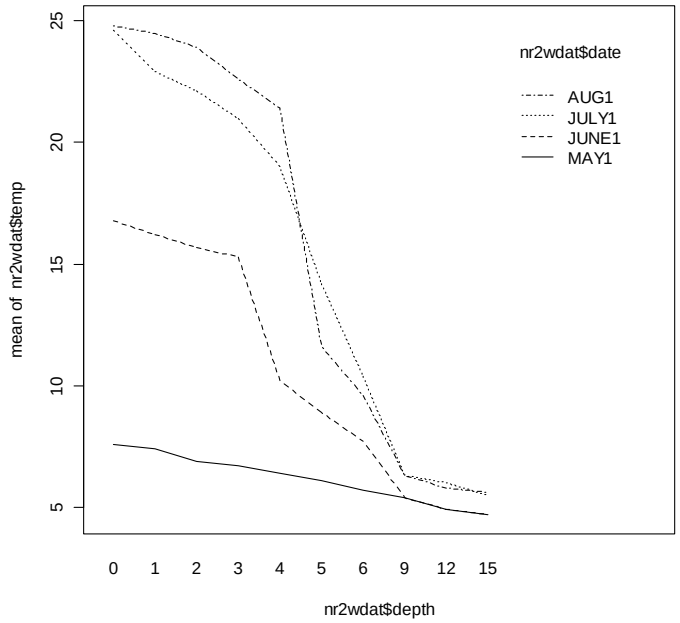
```
anova.model4<-lm(temp ~ date + as.factor(depth), data =
nr2wdat)
Anova(anova.model4, type = 3)

Anova Table (Type III tests)

Response: temp
                 Sum Sq Df  F value    Pr(>F)
```

```
(Intercept)      1511.99  1 125.5652 1.170e-11 ***
date              591.15  3  16.3641 2.935e-06 ***
as.factor(depth) 1082.82  9   9.9916 1.450e-06 ***
Residuals         325.12 27
```

Assuming that this is a Model III ANOVA (`date` random, `depth` fixed), what do you conclude? (Hint: you may want to generate an interaction plot of temp versus depth and month, just to see what's going on.)

```
interaction.plot(nr2wdat$depth, nr2wdat$date, nr2wdat$temp)
```

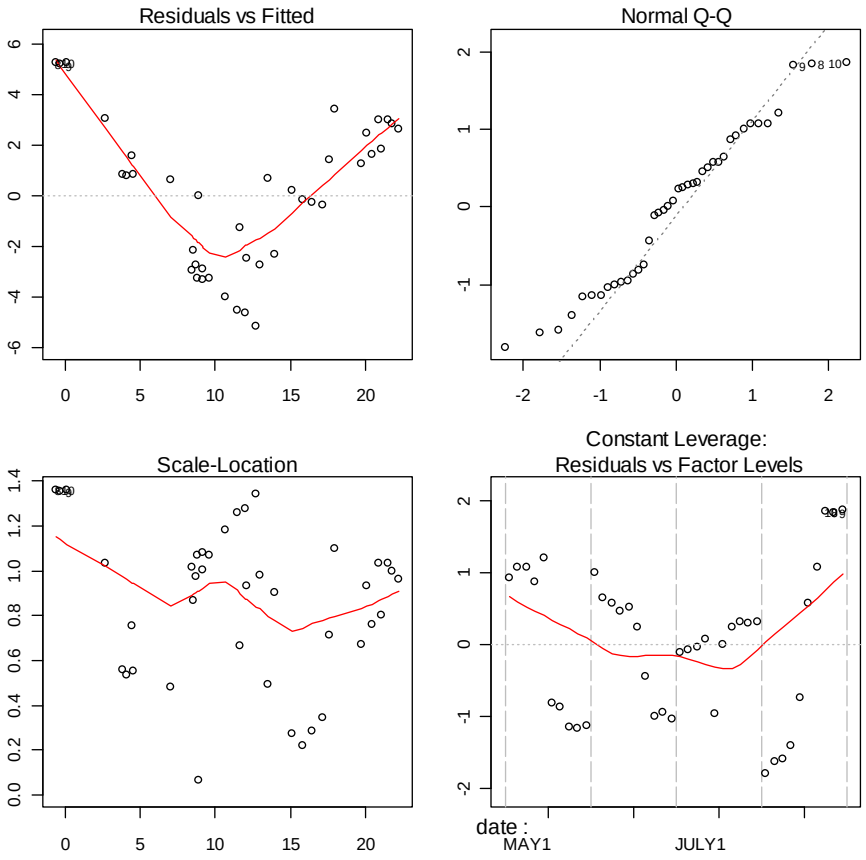
**Figure 46.**




There is a highly significant decrease in temperature as depth increases. To test the effect of month (the (assumed) random factor), we must assume that there is no interaction between depth and month, i.e. that the change in temperature with depth is the same for each month. This is a dubious assumption: if you plot temperature against depth for each month, you should see that the temperature profile becomes increasingly non-linear as the summer progresses (i.e. the thermocline develops), from almost a linear decline in early spring to what amounts to a step decline in August. In other words, the relationship between temperature and depth does change with month, so that if you were to use the above fitted model to estimate, say, the temperature at a depth of 5 m in July, you would not get a particularly good estimate.

In terms of residual diagnostics, have a look at the residuals probability plot and residuals vs fitted values plot.

**Figure 47.**




Testing the residuals for normality, we get p=0.16, so that the normality assumption seems to be O.K. In terms of heteroscedasticity, we can only test among months, using depths as replicates (or among depths using months as replicates). Using depths as replicates within months, we find

```
levene.test(temp ~ date , data = nr2wdat)
Levene's Test for Homogeneity of Variance
      Df F value    Pr(>F)
group  3  17.979 2.679e-07 ***
      36       3
```

So there seems to be some problem here, as can be plainly seen in the above plot of residuals vs fit. All in all, this analysis is not very satisfactory: there appears to be some problems with the assumptions, and the assumption of no interaction between `depth` and `date` would appear to be invalid.

R commands for 2-way ANOVA, unreplicated

```
#load datafile Nr2wdat.Rdata
# load(file.choose()) # local version
load(url("http://www.antoinemorin.com/biostats/2010/
Nr2wdat.Rdata"))


#Check contents
ls.str()

# Run anova
nr2wdat$depth <- as.factor(nr2wdat$depth)
anova.model4 <- lm(temp ~ date + depth, data = nr2wdat)
Anova(anova.model4, type = 3)

interaction.plot(nr2wdat$depth, nr2wdat$date, nr2wdat$temp)
par(mfrow = c(2, 2))
plot(anova.model4)

# Check normality of residuals
shapiro.test(residuals(anova.model4))

# Check homoscedasticity
require(car)
levene.test(temp ~ date, data = nr2wdat)
```

## Nested designs

A common experimental design occurs when each major group (or treatment) is divided into randomly chosen subgroups. For example, a geneticist interested in the effects of genotype on desiccation resistance in fruit flies might conduct an experiment with larvae of three different genotypes. For each genotype (major group), she sets up three environmental chambers (sub-groups, replicates within groups) with a fixed temperature humidity regime, and in each chamber, she has five larvae for which she records the number of hours each larvae survived.

The file `Nestdat.Rdata` contains the results of just such an experiment. The file lists three variables: `genotype`, `chamber` and `survival`. Run a nested ANOVA with `survival` as the dependent variable, `genotype/chamber` as the independent variables (this is the shorthand notation for a chamber effect nested under genotype).

```
nestdat$chamber <- as.factor(nestdat$chamber)
anova.nested <- lm(survival ~ genotype/chamber, data = nest-
dat)
```

What do you conclude from this analysis? What analysis would (should) you do next? (Hint: if there is a non-significant effect of chambers within genotypes, then you can increase the power of between-genotype comparisons by pooling over chambers within genotypes.) Do it! Make sure you check your assumptions!

R program for the nested two-way ANOVA

```
nestdat$chamber <- as.factor(nestdat$chamber)
anova.nested <- aov(survival ~ genotype/chamber, data = nestdat)
anova(anova.nested)
```

```
Analysis of Variance Table

Response: survival
                Df  Sum Sq Mean Sq  F value Pr(>F)
genotype         2 2952.22 1476.11 292.6081 <2e-16 ***
genotype:chamber  6   40.65    6.78   1.3432 0.2639
Residuals        36  181.61    5.04
```

We conclude from this analysis that there is no (significant) variation among chambers within genotypes, but that the null hypothesis that all genotypes have the same dessiccation resistance (as measured by survival) is rejected. In other words, genotypes differ in their survival.

Since the chambers within genotypes effect is non-significant, we may want to pool over chambers to increase our degrees of freedom:

```
anova.simple <- lm(survival ~ genotype, data = nestdat)
anova(anova.simple)
```

```
Analysis of Variance Table

Response: survival
          Df  Sum Sq Mean Sq F value    Pr(>F)
genotype   2 2952.22 1476.11  278.93 < 2.2e-16 ***
Residuals 42  222.26    5.29
```

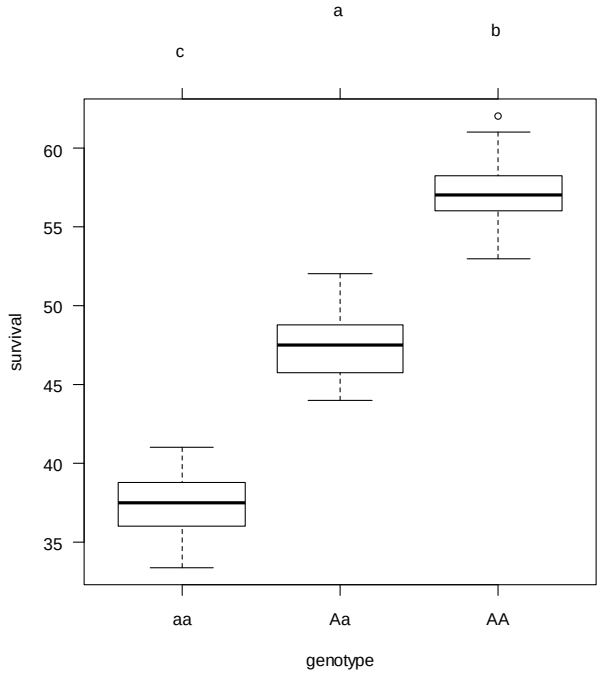
Thus, we conclude that there is significant variation among the three genotypes in dessiccation resistance.

A box plot of survival across genotypes shows clearly that there is significant variation among the three genotypes in dessiccation resistance. This can be combined with a formal Tukey multiple comparison test:

```
par(mfrow=c(1,1))
# Compute and plot means and Tukey CI
means <- glht(anova.simple, linfct = mcp(genotype =
"Tukey"))
cimeans <- cld(means)
# use sufficiently large upper margin
```

```
old.par <- par(mai = c(1, 1, 1.25, 1))
# plot
plot(cimeans, las=1) # las option to put y-axis labels as God
intended them
```

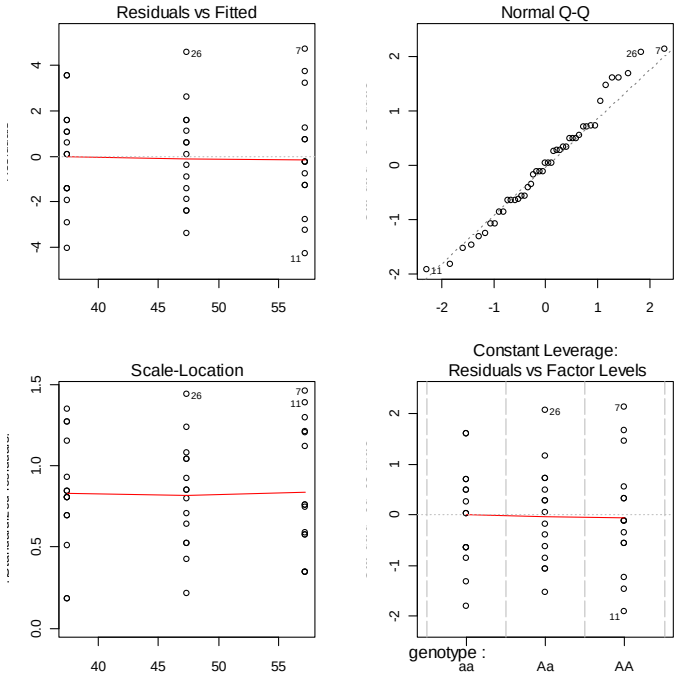
**Figure 48.**




So, we conclude from the Tukey analysis and plot that dessiccation resistance (R) , as measured by larval survival under hot, dry conditions, varies significantly among all three genotypes with R(AA) > R(Aa) > R(aa).

Before concluding this, however, we must test the assumptions. Here are the residual plots and diagnostics for the one-way (unnested) design:

**Figure 49.**




So, all the assumptions appear to be valid, and the conclusion reached above still holds. Note that if you compare the residual mean squares of the nested and one-way ANOVAs (5.04 vs 5.29), they are almost identical. This is not surprising, given the small contribution of the `chamber %in% genotype` effect to the explained sum of squares.

## Two-way non-parametric ANOVA

Two-way non-parametric ANOVA is an extension of the non-parametric one-way methods discussed previously. The basic procedure is to rank all the data in the sample from smallest to largest, then carry out a 2-way ANOVA on the ranks. This can be done either for replicated or unreplicated data:

⌨ Using the data file `Stu2wdat.Rdata`, do a two-factor ANOVA to examine the effects of `sex` and `location` on `rank(rdwght)`.

```
aov.rank <- aov(rank(rdwght) ~ sex * location,
    contrasts = list(sex = contr.sum,
    location = contr.sum), data = Stu2wdat)
Anova(aov.rank, type = 3)
```

```
Anova Table (Type III tests)
```

```
Response: rank(rdwght)
              Sum Sq  Df  F value    Pr(>F)
(Intercept)  1499862   1 577.8673 < 2.2e-16 ***
sex            58394   1  22.4979 4.237e-06 ***
location        1128   1   0.4347    0.5105
sex:location    1230   1   0.4738    0.4921
Residuals     472383 182
```

The Scheirer-Ray-Hare extension of the Kruskall-Wallis test is done by computing a statistic H given by the effect sums of squares (SS) divided by the total MS. The latter can be calculated as the variance of the ranks.. We compute an H statistic for each term. The H-statistics are then compared to a theoretical $\chi^2$ distribution using the command line: `1-pchisq(H, df)`, where `H` and `df` are the calculated H-statistics and associated degrees of freedom, respectively.

⌨ Use the ANOVA table based on ranks to test the effects of `sex` and `location` on `rdwght`. What do you conclude? How does this result compare with the result obtained with the parametric 2-way ANOVA done before?

To calculate the Scheirer-Ray-Hare extension to the Kruskall-Wallis test, you must first calculate the total mean square (MS), i.e. the variance of the ranked data. In this case, there are 186 observations, their ranks are therefore the series 1, 2, 3, ..., 186. The variance can be calculated simply as var(1:186) (Isn't R neat? Cryptic maybe, but neat.)

So we can compute the H statistic for each term:

$H_{sex}$ = 58394/var(1:186) = 20.14

$H_{location}$ = 1128/var(1:186) = 0.39

$H_{sex:location}$ = 1230/var(1:186) = 0.42

And convert these statistics into p-values:

for sex: 1-pchisq(20.14,1) = 7.197556e-06

location 1-pchisq(0.39, 1) = 0.5322994

sex:location 1-pchisq(0.42, 1) = 0.516937

Note that these results are the same as those obtained in our original two-way parametric ANOVA. Despite the reduced power, we still find significant differences between the sexes, but still no interaction and no effect due to location.

There is, however, an important difference. Recall that in the original parametric ANOVA, there was a significant effect of `sex` when we considered the problem as a Model I ANOVA. However, if we consider it as Model III, the significant `sex` effect could in principle disappear, because the df associated with the interaction MS are much smaller than the df associated with the Model I error MS. In this case, however, the interaction MS is about half that of the error MS. So, the significant `sex` effect becomes even more significant if we analyze the problem (correctly) as a Model III ANOVA. Once again, we see the importance of specifying the correct ANOVA design.

## Multiple comparisons

Further hypothesis testing in multiway ANOVAs depends critically on the outcome of the initial ANOVA. If you are interested in comparing groups of marginal means (that is, means of treatments for one factor pooled over levels of the other factor, e.g., between male and female sturgeon pooled over location), this can be done exactly as outlined for multiple comparisons for one-way ANOVAs. For comparison of individual cell means, you must specify the interaction as the group variable.

The file `wmcdat2.Rdata` shows measured oxygen consumption (`O2cons`) of two species (`species` = A, B)) of limpets at three different concentrations of seawater (`conc` = 100, 75, 50%) taken from Sokal and Rohlf, 1995, p. 332.

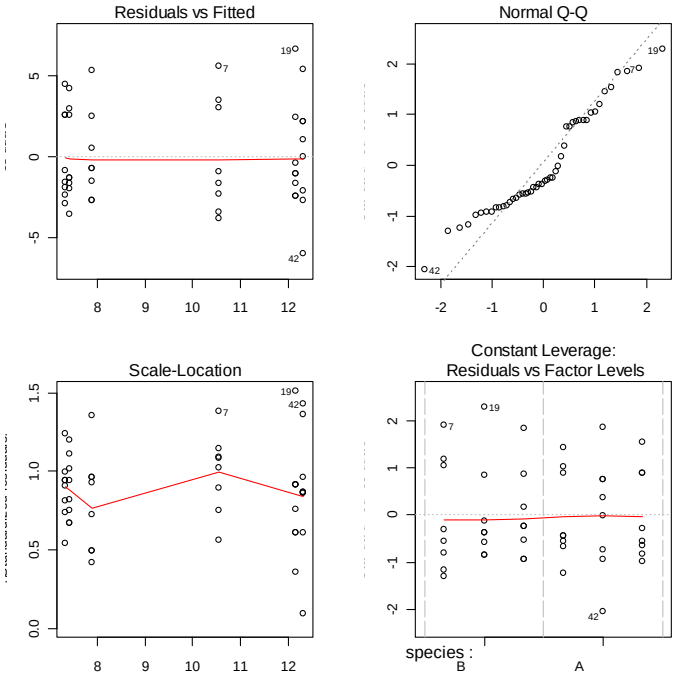
Run a 2-way factorial ANOVA on `wmcdat2` data, using `O2cons` as the dependent variable and `species` and `conc` as the factors. What do you conclude?

The ANOVA table is shown below. Technically, because the sample sizes in individual cells are rather small, this analysis should be repeated using a non-parametric ANOVA. For the moment, let's stick with the parametric analysis.

```
Anova Table (Type III tests)

Response: o2cons
             Sum Sq Df  F value    Pr(>F)
(Intercept)  4441.7  1 464.6163 < 2.2e-16 ***
species        16.6  1   1.7404 0.1942381
conc          181.3  2   9.4833 0.0003993 ***
species:conc   23.9  2   1.2514 0.2965616
Residuals     401.5 42
```

Look at the diagnostic plots:

**Figure 50.**




Homoscedasticity looks ok, but normality less so.. Testing for normality, we get:

```
Shapiro-Wilk normality test

data:  residuals(anova.model5)
W = 0.9369, p-value = 0.01238
```

So there is evidence of non-normality, but otherwise everything looks O.K. Since the ANOVA is relatively robust with respect to non-normality, we proceed, but if we wanted to reassure ourselves, we could run a non-parametric ANOVA, and get the same answer.

On the basis of the ANOVA results obtained above, which means would you proceed to compare? Why?

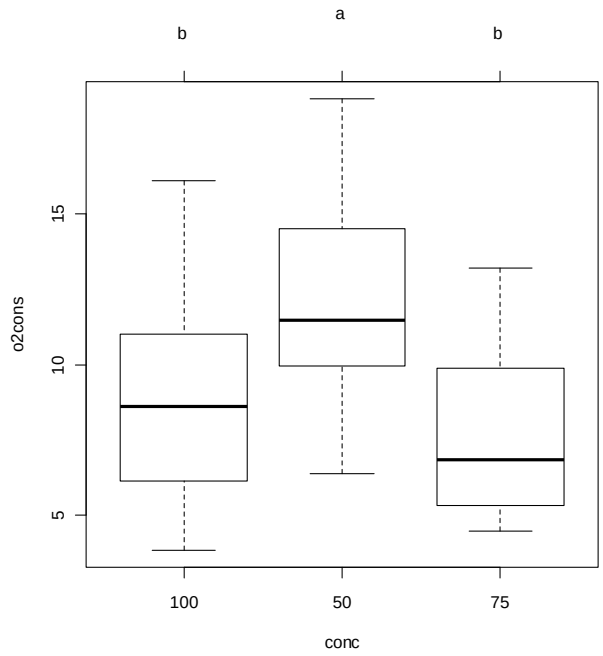
Overall, we conclude that there are no differences among species, and that the effect of concentration does not depend on species (no interaction). Since there is no interaction and no main effect due to species, the only justified comparisons are among salinity concentrations:

```
# fit simplified model
anova.model6 <- lm (o2cons ~ conc, data = wmcdat2)
# Make Tukey multiple comparisons
TukeyHSD(anova.model6)
```

```
Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = o2cons ~ conc, data = wmcdat2)

$conc
           diff        lwr       upr     p adj
50-100  3.25500  0.5692518  5.940748 0.0141313
75-100 -1.38125 -4.0669982  1.304498 0.4325855
75-50  -4.63625 -7.3219982 -1.950502 0.0003793
```

```
par(mfrow=c(1,1))
# Graph of all comparisons for conc
tuk <- glht(anova.model6, linfct = mcp(conc = "Tukey"))
# extract information
tuk.cld <- cld(tuk)
# use sufficiently large upper margin
old.par <- par(mai = c(1, 1, 1.25, 1))
# plot
plot(tuk.cld)
par(old.par)
```

**Figure 51.**




So there is evidence of a significant difference in oxygen consumption at a reduction in salinity to 50% of regular seawater, but not at a reduction of only 25%.

⌨ Repeat the analysis described above using `wmc2dat2.Rdata`. How do your results compare with those obtained for `wmc2dat2.Rdata`?

Using `wmc2dat2.Rdata`, we get:

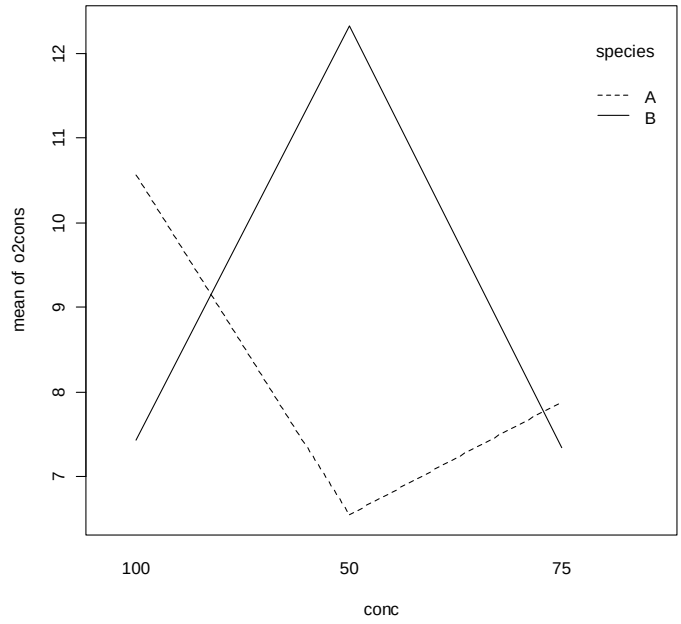
```
Anova Table (Type III tests)
```

```
Response: o2cons
             Sum Sq Df  F value    Pr(>F)
(Intercept)  3618.2  1 381.8998 < 2.2e-16 ***
species         5.8  1   0.6162 0.4368642
conc           29.0  2   1.5283 0.2287237
species:conc  168.2  2   8.8742 0.0006101 ***
Residuals     397.9 42
```

Here there is a large interaction effect, and consequently, there is no point in comparing marginal means. This is made clear by examining an interaction plot:

```
with(wmc2dat2, interaction.plot(conc, species, o2cons))
```

**Figure 52.**




Working still with the `wmc2dat2.Rdata` data set, compare individual cell means (6 in all), with the **Bonferonni** adjustment. To do this, it is helpful to create a new variable to indicate all the combinations of `species` and `conc`:

```
wmc2dat2$species.conc <- paste(wmc2dat2$species,
    wmc2dat2$conc)
```

Then we can conduct pairwise bonferroni comparisons:

```
with(wmcdat2, pairwise.t.test(o2cons, species.conc,
    p.adj = "bonf"))
```

```
Pairwise comparisons using t tests with pooled SD
```

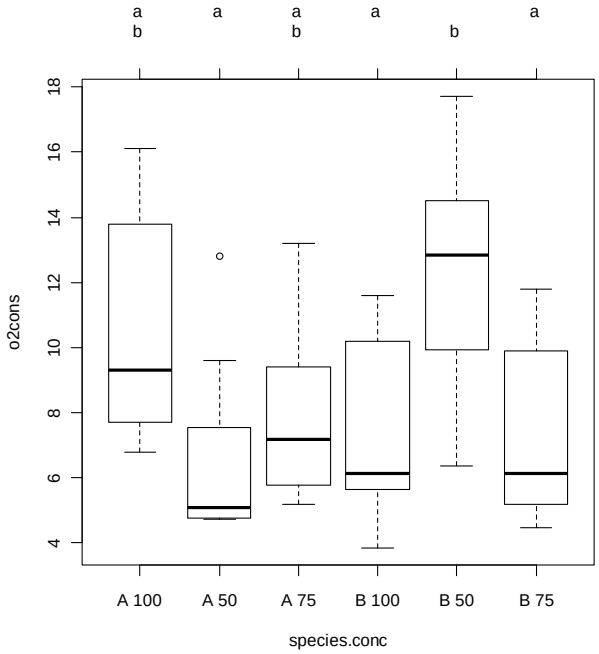
```
data:  o2cons and species.conc

      A 100 A 50  A 75  B 100 B 50
A 50  1.000 -     -     -     -
A 75  1.000 0.124 -     -     -
B 100 0.737 0.056 1.000 -     -
B 50  1.000 1.000 0.096 0.043 -
B 75  0.647 0.048 1.000 1.000 0.036

P value adjustment method: bonferroni
```

These comparisons are a little more difficult to interpret, but the analysis essentially examines for differences among seawater concentrations within species A and for differences among concentrations within species B. We see here that the `o2Cons` at 50% seawater for species B is significantly different from that of 75% and 100% seawater for species B, whereas there are no significant differences in `o2cons` for species A across all seawater concentrations.

I find these outputs rather unsatisfying because they show only p-values, but no indication of effect size. One can get both the conclusion from the multiple comparison procedure and an indication of effect size from the graph produced with the following code:

```
3   fit one-way anova comparing all combinations of spe-
cies.conc combinations
anova.modelx <- aov(o2cons ~ species.conc, data = wmc2dat2)


tuk2 <- glht(anova.modelx, linfct = mcp(species.conc =
"Tukey"))
# extract information
tuk2.cld <- cld(tuk2)
# use sufficiently large upper margin
old.par <- par(mai = c(1, 1, 1.25, 1))
# plot
plot(tuk2.cld)
par(old.par)
```

**Figure 53.**




Note that in this analysis, we have used the error MS = 9.474 from the original model to contrast cell means. Recall, however, that this assumes that in fact we are dealing with a Model I ANOVA, which may or may not be the case (`conc` is certainly a fixed factor, but `species` might be either fixed or random).

## Permutation test for two-way ANOVA

When data do not meet the assumptions of the parametric analysis in two- and multiway ANOVA, as an alternative to the non-parametric ANOVA, , it is possible to run permutation tests to calculate p-values. The following code does it for a two-way Model I ANOVA. I'll leave it to you to adapt it to other situations. (Bonus points for an elegant solution to multi-way models...)

```
################################################################################
#####
# Permutation test for two way ANOVA
# Ter Braak creates residuals from cell means and then permutes across
# all cells
#  This can be accomplished by taking residuals from the full model
#  modified from code written by David C. Howell
#  http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Permutation%20Anova/Per-
mTestsAnova.html


nreps = 500
dependent <- Stu2wdat$rdwght
factor1 <- as.factor(Stu2wdat$sex)
factor2 <- as.factor(Stu2wdat$location)

my.dataframe <- data.frame(dependent, factor1, factor2)
my.dataframe.noNA <- my.dataframe[complete.cases(my.dataframe),
```

```
    ]

mod <- lm(dependent ~ factor1 + factor2 + factor1:factor2,
    data = my.dataframe.noNA)
res <- mod$residuals
TBint <- numeric(nreps)
TB1 <- numeric(nreps)
TB2 <- numeric(nreps)

ANOVA <- summary(aov(mod))
cat(" The standard ANOVA for these data follows ",
    "\n")
F1 <- ANOVA[[1]]$"F value"[1]
F2 <- ANOVA[[1]]$"F value"[2]
Finteract <- ANOVA[[1]]$"F value"[3]
print(ANOVA, "\n")
cat("\n")
cat("\n")


TBint[1] <- Finteract
for (i in 2:nreps) {
    newdat <- sample(res, length(res), replace = FALSE)
    modb <- summary(aov(newdat ~ factor1 + factor2 + factor1:factor2,
        data = my.dataframe.noNA))
    TBint[i] <- modb[[1]]$"F value"[3]
    TB1[i] <- modb[[1]]$"F value"[1]
    TB2[i] <- modb[[1]]$"F value"[2]
}
probInt <- length(TBint[TBint >= Finteract])/nreps
prob1 <- length(TB1[TB1 >= F1])/nreps
prob2 <- length(TB2[TB1 >= F2])/nreps
cat("\n")
cat("\n")
print("Resampling as in ter Braak with unrestricted sampling of cell residu-
als. ")
cat("The probability for the effect of Interaction is ",
    probInt, "\n")
cat("The probability for the effect of Factor 1 is ",
    prob1, "\n")
cat("The probability for the effect of Factor 2 is ",
    prob2, "\n")
```

## Bootstrap for two-way ANOVA

In most cases, permutation tests will be more appropriate than bootstrap in ANOVA designs. However, for the sake of completedness, I have a snippet of code to do bootstrap for you:

```
########################################################################


#  Bootstrap for two-way ANOVA
#  You possibly want to edit bootfunction.mod1 to return other values
#  Here it returns the standard coefficients of the fitted model
#  Requires boot library
#


nreps = 5000
dependent <- Stu2wdat$rdwght
factor1 <- as.factor(Stu2wdat$sex)
factor2 <- as.factor(Stu2wdat$location)

my.dataframe <- data.frame(dependent, factor1, factor2)
my.dataframe.noNA <- my.dataframe[complete.cases(my.dataframe), ]



require(boot)
#  Fit model on observed data
mod1 <- aov(dependent ~ factor1 + factor2 + factor1:factor2,
    data = my.dataframe.noNA)
```

```
#  Bootstrap 1000 time using the residuals bootstraping methods  to
#  keep the same unequal number of observations for each level of the indep.
var.


fit <- fitted(mod1)
e <- residuals(mod1)
X <- model.matrix(mod1)

bootfunction.mod1 <- function(data, indices) {
    y <- fit + e[indices]
    bootmod <- lm(y ~ X)
    coefficients(bootmod)
}

bootresults <- boot(my.dataframe.noNA, bootfunction.mod1,
    R = 1000)


bootresults

#  Calculate 90% CI and plot bootstrap estimates separately
#  for each model parameter

boot.ci(bootresults, conf = 0.9, index = 1)
plot(bootresults, index = 1)

boot.ci(bootresults, conf = 0.9, index = 3)
plot(bootresults, index = 3)

boot.ci(bootresults, conf = 0.9, index = 4)
plot(bootresults, index = 4)

boot.ci(bootresults, conf = 0.9, index = 5)
plot(bootresults, index = 5)
```

# Lab- Multiple regression

After completing this laboratory exercise, you should be able to:

- Use R to fit a multiple regression model, and compare the adequacy of several models using inferential and information theoretic criteria
- Use R to test hypotheses about the effects of different independent variables on the dependent variable of interest.
- Use R to evaluate multicollinearity among (supposedly) independent variables and its effects.
- Use R to do curvilinear (polynomial) regression.

## Multiple regression: points to keep in mind

Multiple regression models are used in cases where there are one dependent variable and several independent, continuous variables. In many biological systems, the variable of interest may be influenced by several different factors, so that accurate description or prediction requires that several independent variables be included in the regression model.

Before beginning, be aware that multiple regression takes time to learn well. Beginners should keep in mind several important points:

1. An overall regression model may be statistically significant even if none of the individual regression coefficients in the model are (caused by multicollinearity);
2. A multiple regression model may be "nonsignificant" even though some of the individual coefficients are (caused by overfitting);
3. Unless "independent" variables are uncorrelated in the sample, different model selection procedures may yield different results.

## The data: let's first have a look

The file `Mregdat.Rdata` contains data of the richness (number of species) of four groups: birds (`bird`, and its log transform `logbird`), plants (`plant, logpl`), mammals (`mammal, logmam`), herptiles (`herptile, logherp`), and the total species richness of all four groups combined (`totsp, logtot`) in 30 wetlands in the Ottawa-Cornwall-Kingston area. Also shown are the latitude and longitude of the wetland (`lat`, `long`), its area (`logarea`), the percentage of the wetland covered by water at all times during the year (`swamp`), the percentage of

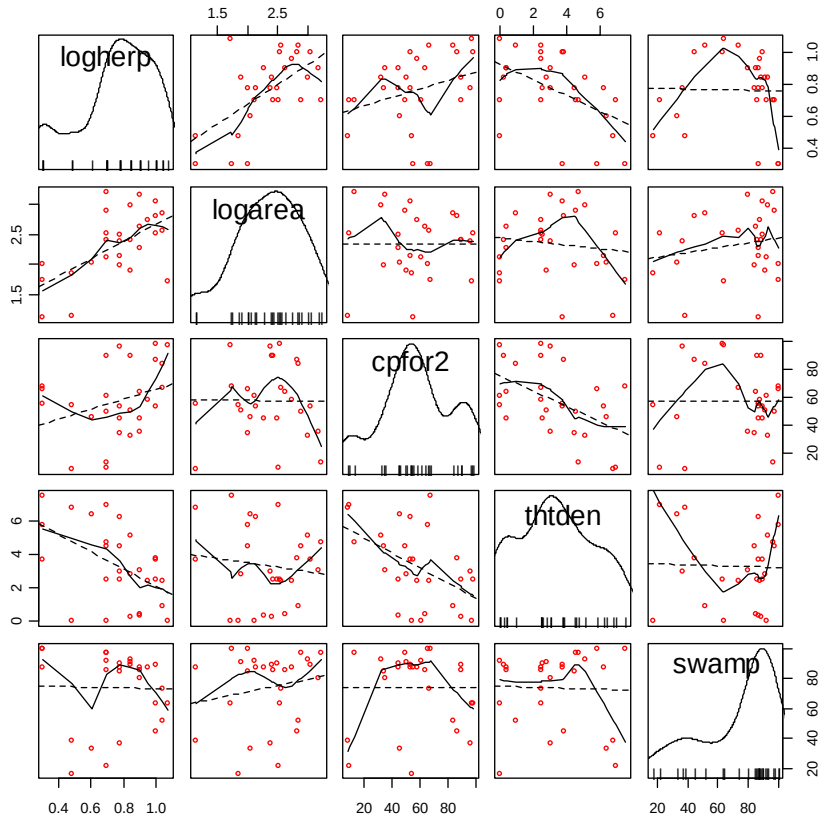
forested land within 1 km of the wetland (`cpfor2`), and the density (in m/hectare) of hard-surface roads within 1 km of the wetland (`thtden`).

We will focus on herptiles for this exercise, so we better first have a look at how this variable is distributed and correlated to the potential independent variables:

```
require(car)
scatterplot.matrix(~logherp + logarea + cpfor2 + thtden +
swamp,
  reg.line = lm, smooth = TRUE, span = 0.5, diagonal = "den-
sity",
  data = mydata)
```

**Figure 54.**

f



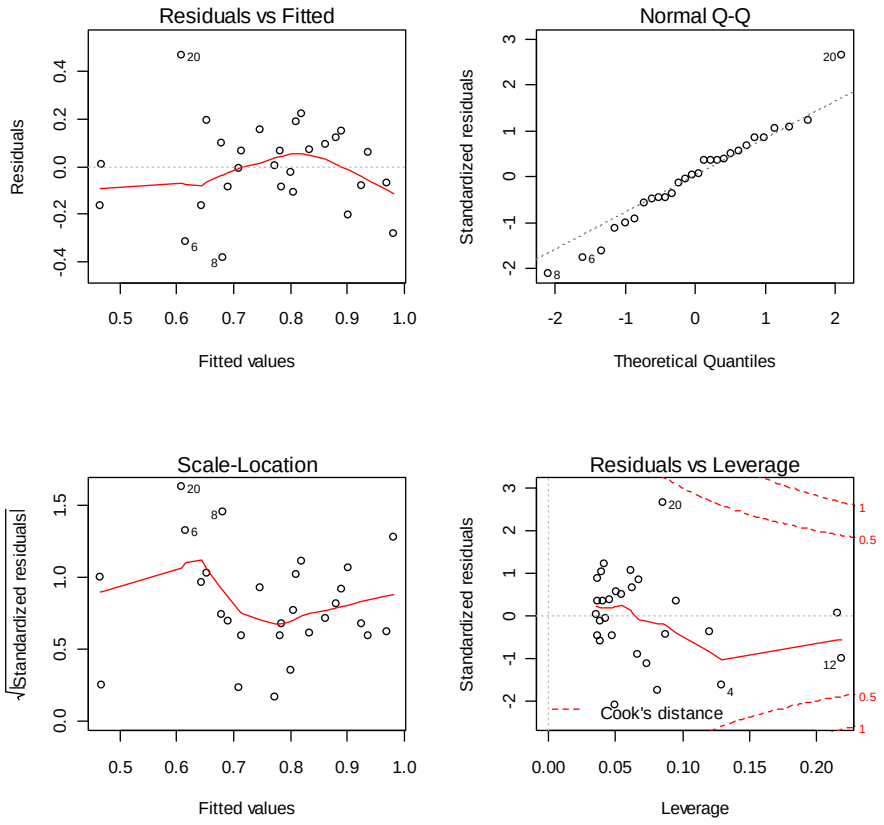
## Multiple regression models from scratch

We begin the multiple regression exercise by considering a situation with one dependent variable and three possibly independent variables. First, we will start from scratch and build a multiple regression model based on what we know from building simple regression models. Next, we will look at automated methods of building multiple regressions models using simultaneous, forward, and backward stepwise procedures.

⌨ Using the `Mregdat.Rdata` data file, regress `logherp` on `logarea` On the basis of the regression, what do you conclude?

```
model.loga<-lm(logherp ~ logarea, data = mydata)
summary(model.loga)

Call:
lm(formula = logherp ~ logarea, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.38082 -0.09265  0.00763  0.10409  0.46977

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.18503    0.15725   1.177 0.249996
logarea      0.24736    0.06536   3.784 0.000818 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1856 on 26 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.3552,     Adjusted R-squared: 0.3304
F-statistic: 14.32 on 1 and 26 DF,  p-value: 0.0008185

opar <- par(mfrow(2,2))
plot(model.loga)
opar
```

**Figure 55.**




It looks like there is a positive relationship between herptile species richness and wetland area: the larger the wetland, the greater the number of species. Note, however, that about 2/3 of the observed variability in species richness among wetlands is not "explained" by wetland area ($R^2$ = .355). Residual analysis shows no major problems with normality, heteroscedasticity or independence of residuals.

⌨ Rerun the above regression, this time replacing `logarea` with `cpfor2` as the independent variable, such that the expression in the formula field reads: `Logherp~cpfor2`. What do you conclude?

```
Call:
lm(formula = logherp ~ cpfor2, data = mydata)

Residuals:
    Min      1Q  Median      3Q     Max
-0.4909 -0.1027  0.0588  0.1603  0.2516

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.609197   0.104233   5.845 3.68e-06 ***
cpfor2      0.002706   0.001658   1.632    0.115
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.2202 on 26 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.09289,    Adjusted R-squared: 0.058
F-statistic: 2.662 on 1 and 26 DF,  p-value: 0.1148
```

According to this result, we would accept the null hypothesis, and conclude that there is no relationship between herptile density and the proportion of forest on adjacent lands. But what happens when we enter both variables into the regression simultaneously?

⌨ Rerun the above regression one more time, this time adding both independent variables into the model at once, such that `logherp~logarea+cpfor2`. What do you conclude?

```
Call:
lm(formula = logherp ~ logarea + cpfor2, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40438 -0.11512  0.01774  0.08187  0.36179

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.027058   0.166749   0.162 0.872398
logarea     0.247789   0.061603   4.022 0.000468 ***
cpfor2      0.002724   0.001318   2.067 0.049232 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.175 on 25 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.4493,     Adjusted R-squared: 0.4052
F-statistic:  10.2 on 2 and 25 DF,  p-value: 0.0005774
```

Now we reject both null hypotheses that the slope of the regression of `logherp` on `logarea` is zero and that the slope of the regression of `logherp` on `cpfor2` is zero.

Why is `cpfor2` a significant predictor of `logherp` in the combined model when it was not significant in the simple linear model? The answer lies in the fact that it is sometimes necessary to control for one variable in order to detect the effect of another variable. In this case, there is a significant relationship between `logherp` and `logarea` that masks the relationship between `logherp` and `cpfor2`. When both variables are entered into the model at once, the effect of `logherp` is controlled for, making it possible to detect a `cpfor2` effect (and vice versa).

⌨ Run another multiple regression, this time substituting `THTDEN` for `cpfor2` as an independent variable (`Logherp~Logarea+thtden`).

```
Call:
lm(formula = logherp ~ logarea + thtden, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-0.31583 -0.12326  0.02095  0.13201  0.31674

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.37634    0.14926   2.521 0.018437 *
logarea      0.22504    0.05701   3.947 0.000567 ***
thtden      -0.04196    0.01345  -3.118 0.004535 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1606 on 25 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.5358,     Adjusted R-squared: 0.4986
F-statistic: 14.43 on 2 and 25 DF,  p-value: 6.829e-05
```

In this case we reject the null hypotheses that there are no effects of wetland area (`logarea`) and road density (`thtden`) on herptile richness (`logherp`). Note here that road density has a negative effect on richness, whereas wetland area and forested area (`cpfor2`; results from previous regression) both have positive effects on herptile richness.

The $R^2$ of this model is even higher than the previous multiple regression model, reflecting a higher correlation between `logherp` and `thtden` than between `logherp` and `cpfor2` (if you run a simple regression between `logherp` and `thtden` and compare it to the `cpfor2` regression you should be able to detect this).

Thus far, it appears that herptile richness is related to wetland area (`logarea`), road density (`thtden`), and possibly forest cover on adjacent lands (`cpfor2`). But, does it necessarily follow that if we build a regression model with all three independent variables, that all three will show significant relationships? No, because we have not yet examined the relationship between `Logarea`, `cpfor2` and `thtden`. Suppose, for example, two of the variables (say, `cpfor2` and `thtden`) are perfectly correlated. Then the `thtden` effect is nothing more than the `cpfor2` effect (and vice versa), so that once we include one or the other in the regression model, none of the remaining variability would be explained by the third variable.

Fit a regression model with `logherp` as the dependent variable and `logarea`, `cpfor2` and `thtden` as the independent variables. What do you conclude?

```
Call:
lm(formula = logherp ~ logarea + cpfor2 + thtden, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.30729 -0.13779  0.02627  0.11441  0.29582

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.284765   0.191420   1.488 0.149867
logarea      0.228490   0.057647   3.964 0.000578 ***
cpfor2       0.001095   0.001414   0.774 0.446516 (1)
thtden      -0.035794   0.015726  -2.276 0.032055 *
---
```

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

Residual standard error: 0.1619 on 24 degrees of freedom
  (2 observations deleted due to missingness)

Multiple R-squared: 0.5471 (2),     Adjusted R-squared: 0.4904
F-statistic: 9.662 on 3 and 24 DF,  p-value: 0.0002291
```

Several things to note here:

(1) As predicted, the regression coefficient for `cpfor2` has become non-significant: once the variability explained by `logarea` and `thtden` is removed, a non-significant part of the remaining variability is explained by `cpfor2`.

(2) The $R^2$ for this model (.547) is only marginally larger than the $R^2$ for the model with only `logarea` and `thtden` (.536), which is again consistent with the non-significant coefficient for `cpfor2`.

Note also that although the regression coefficient for `thtden` has not changed significantly from that obtained when just `thtden` and `logarea` were included in the fitted model (-.036 vs -.042), the standard error for the regression coefficient for `thtden` has increased slightly, meaning the is more precise. If the correlation between `thtden` and `cpfor2` was greater, the change in precision would also be greater.

We can compare the fit of the last two models (i.e., the model with all 3 variables and the model with only `logarea` and `thtden`) to decide which model is best to include.

```
anova(model.loga.cpfor2.thtden, model.loga.thtden)
```

```
Analysis of Variance Table

Model 1: logherp ~ logarea + cpfor2 + thtden
Model 2: logherp ~ logarea + thtden
  Res.Df      RSS Df Sum of Sq     F Pr(>F)
1     24  0.62937
2     25  0.64508 -1  -0.01571 0.599 0.4465
```

From this analysis, we would conclude that the full model with all three variables included does not offer a significant improvement in fit over the model with only `Logarea` and `thtden`. This isn't surprising given that we already know that we cannot reject the null hypothesis of no effect of `cpfor2` in the full model.

Overall, we would conclude, on the basis of these analyses, that:

1. Given the three variables `thtden`, `logarea` and `cpfor2`, the best model is one that includes the first two variables.

2. There is evidence of a negative relationship between herptile richness and the density of roads on adjacent lands.
3. There is evidence that the larger the wetland area, the greater the herptile species richness.

Note that by "best", I don't mean the best possible model, I mean the best one given the three predictor variables we started with. It seems pretty clear that there are other factors controlling richness in wetlands, since even with the "best" model, almost half of the variability in richness is unexplained.

## Stepwise multiple regression procedures

There are a number of techniques available for selecting the multiple regression model that best suits your data. When working with only three independent variables it is often sufficient to work through the different combinations of possible variables yourself, until you are satisfied you have fit the best model. This is, essentially, what we did in the first section of this lab. However, the process can become tedious when dealing with numerous independent variables, and you may find an automatic procedure for fitting models to be easier to work with.

Stepwise regression in R relies on the Akaike Information Criterion, as a measure of goodness of fit (AIC=2k + 2ln(L)) where k is the number of regressors, and L is the maximized value of the likelihood function for the model). This is a statistic that rewards accuracy while penalizing model complexity. If a new model has an AIC lower than that of the current model, the new model is a better fit to the data.

Still working with the `Mregdat` data, run a stepwise multiple regression on the same set of variables:

```
model.loga.cpfor2.thtden<-lm(logherp ~ logarea + cpfor2 +
thtden, data = mydata)
#  Stepwise Regression
library(MASS)
step <- stepAIC(model.loga.cpfor2.thtden, direction="both")
step$anova # display results

Start:  AIC=-98.27 (1)
logherp ~ logarea + cpfor2 + thtden

          Df Sum of Sq     RSS     AIC (2)
- cpfor2   1     0.016   0.645 -99.576
<none>                   0.629 -98.267
- thtden   1     0.136   0.765 -94.794
- logarea  1     0.412   1.041 -86.167

Step:  AIC=-99.58
```

```
logherp ~ logarea + thtden


          Df Sum of Sq     RSS     AIC ❸
<none>                   0.645 -99.576
+ cpfor2   1     0.016   0.629 -98.267
- thtden   1     0.251   0.896 -92.376
- logarea  1     0.402   1.047 -88.013

> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
logherp ~ logarea + cpfor2 + thtden

Final Model:
logherp ~ logarea + thtden


      Step Df   Deviance Resid. Df Resid. Dev       AIC
1                               24  0.6293717 -98.26666
2 - cpfor2  1 0.01570813        25  0.6450798 -99.57640
>
```

Examining the output, we find:

❶ R calculated the AIC for the starting model (here the full model with the 3 independent variables).

❷ The AIC for models where terms are deleted.. Note here that the only way to reduce the AIC is to drop cpfor2.

❸ The AIC for models where terms are added or deleted from the model selected in the first step (i.e. logherp ~ logarea + thtden). Note that none of these models are better.

Instead of starting from the full (saturated) model and removing terms, one can start from a null model and add terms:

```
#  Forward selection approach
model.null<-lm(logherp ~ 1, data = mydata)
step <- stepAIC(model.null, scope=~. + logarea + cpfor2 +
thtden, direction="forward")
step$anova # display results
```

How does it differ from the first procedure?

```
Start:  AIC=-82.09
logherp ~ 1

          Df Sum of Sq     RSS     AIC
+ logarea  1     0.494   0.896 -92.376
+ thtden   1     0.342   1.047 -88.013
+ cpfor2   1     0.129   1.260 -82.820
<none>                   1.390 -82.091

Step:  AIC=-92.38
logherp ~ logarea

         Df Sum of Sq     RSS     AIC
+ thtden  1     0.251   0.645 -99.576
+ cpfor2  1     0.131   0.765 -94.794
```

```
<none>                  0.896 -92.376

Step:  AIC=-99.58
logherp ~ logarea + thtden

         Df Sum of Sq     RSS     AIC
<none>                  0.645 -99.576
+ cpfor2  1     0.016   0.629 -98.267
```

```
> step$anova # display results
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
logherp ~ 1

Final Model:
logherp ~ logarea + thtden


       Step Df  Deviance Resid. Df Resid. Dev       AIC
1                              27  1.3895281 -82.09073
2 + logarea  1 0.4935233        26  0.8960048 -92.37639
3  + thtden  1 0.2509250        25  0.6450798 -99.57640
>
```

You should first notice that the final result is the same as the default stepwise regression and as what we got building the model from scratch. In forward selection, R first fits the least complex model (i.e, with only an intercept), and then adds variables, one by one, according to AIC statistics. Thus, in the above example, the model was first fit with only an intercept. Next, `Logarea` was added, followed by `thtden`. `cpfor2` was not added because it would make AIC increase to above that of the model fit with the first two variables.

Generally speaking, when doing multiple regressions, it is good practice to try several different methods (e.g. all regressions, stepwise, and backward elimination, etc.) and see whether you get the same results. If you don't, then the "best" model may not be so obvious, and you will have to think very carefully about the inferences you draw. In this case, regardless of whether we use automatic, or forward/backward stepwise regression, we arrive at the same model.

When doing multiple regression, always bear in mind the following:

1. Different procedures may produce different "best" models, i.e. the "best" model obtained using forward stepwise regression needn't necessarily be the same as that obtained using backward stepwise. It is good practice to try several different methods and see whether you end up with the same result. If you don't, it is almost invariably due to multicollinearity among the independent variables.

2. Be wary of stepwise regression. As the authors of SYSTAT, another commonly used statistical package, note, "Stepwise regression is probably the most abused computerized statistical technique ever devised. If you think you need automated stepwise

regression to solve a particular problem, you probably don't. Professional statisticians rarely use automated stepwise regression because it does not necessarily find the "best" fitting model, the "real" model, or alternative "plausible" models. Furthermore, the order in which variables enter or leave a stepwise program is usually of no theoretical significance. You are always better off thinking about why a model could generate your data and then testing that model."

3. Remember that just because there is a significant regression of Y on X doesn't mean that X causes Y: correlation does not imply causation!

## Detecting multicollinearity

Multicollinearity is the presence of correlations among independent variables. In extreme cases (perfect collinearity) it will prevent you from fitting some models. When collinearity is not perfect, it reduces your ability to test for the effect of individual variables, **but does not affect the ability of the model to predict**.

The help file for the HH package contains this clear passage about one of the indices of multicollinearity, the variance inflation factors:

"*A simple diagnostic of collinearity is the variance inflation factor, VIF one for each regression coefficient (other than the intercept). Since the condition of collinearity involves the predictors but not the response, this measure is a function of the X's but not of Y. The VIF for predictor i is 1/(1-R_i^2), where R_i^2 is the R^2 from a regression of predictor i against the remaining predictors. If R_i^2 is close to 1, this means that predictor i is well explained by a linear function of the remaining predictors, and, therefore, the presence of predictor i in the model is redundant. Values of VIF exceeding 5 are considered evidence of collinearity: The information carried by a predictor having such a VIF is contained in a subset of the remaining predictors. If, however, all of a model's regression coefficients differ significantly from 0 (p-value < .05), a somewhat larger VIF may be tolerable.*"

VIF indicate by how much the variance of each regression coefficient is increased by the presence of collinearity.

**Trap.** There are several vif() functions (I know of at least three in the packages car, HH, ,and DAAG), and I do not know how they differ.

To quantify multicollinarity, one can simply call the vif() function from the package car:

```
require(car)
vif(model.loga.cpfor2.thtden)
```

```
 logarea   cpfor2   thtden
1.022127 1.344455 1.365970
```

Here there is no evidence that multicollinearity is a problem since all vif are close to 1.

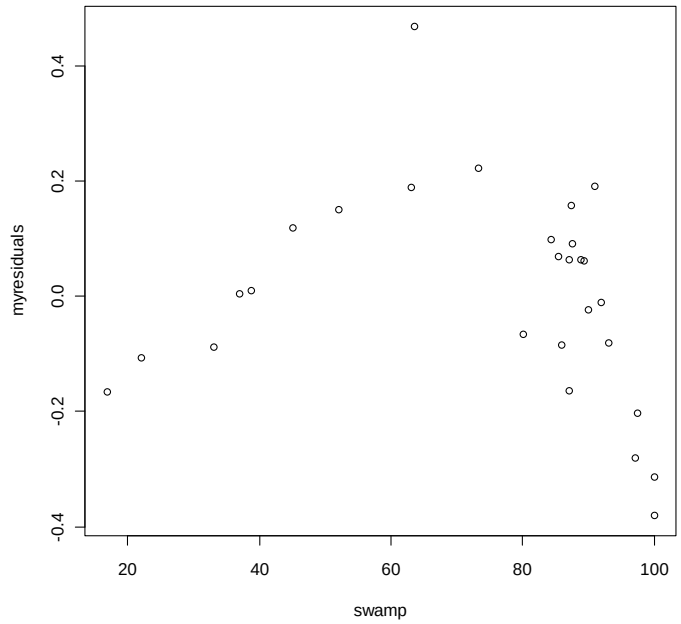
## Polynomial regression

In the regression models considered so far, we have assumed that the relationship between the dependent and independent variables is linear. If not, in some cases it can be made linear by transforming one or both variables. On the other hand, for many biological relationships no transformation in the world will help, and we are forced to go with some sort of non-linear regression method.

The simplest type of nonlinear regression method is polynomial regression, in which you fit regression models that include independent variables raised to some power greater than one, e.g. $X^2$, $X^3$, etc.

⌨ Plot the relationship between the `residuals` of the `logherp-logarea` regression and `swamp`.

**Figure 56.**




⌨ Visual inspection of this graph suggests that there is a strong, but highly nonlinear, relationship between these two variables. Try

regressing the `residuals` of the `logherp-logarea` regression on `swamp`. What do you conclude?

```
Call:
lm(formula = myresiduals ~ swamp, data = mydata.noNA)

Residuals:
      Min        1Q    Median        3Q       Max
-0.350880 -0.138189  0.003131  0.108485  0.458021

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.084571   0.109265   0.774    0.446
swamp       -0.001145   0.001403  -0.816    0.422

Residual standard error: 0.1833 on 26 degrees of freedom
Multiple R-squared: 0.02498,    Adjusted R-squared: -0.01252
F-statistic: 0.666 on 1 and 26 DF,  p-value: 0.4219
```

In other words, the fit is terrible, even though you can see from the graph that there is in fact quite a strong relationship between the two - it's just that it is a non-linear relationship. (If you look at your residuals from this model, you will see strong evidence of nonlinearity, as expected.)

⌨ Rerun the above regression but add a second term in the **Formula** field to represent $(\texttt{swamp})^2$. The expression should appear as: `residuals~SWAMP+SWAMP^2`. What do you conclude? What does examination of the residuals from this multiple regression tell you?

```
Call:
lm(formula = myresiduals ~ swamp + I(swamp^2), data = mydata.noNA)

Residuals:
      Min        1Q    Median        3Q       Max
-0.181185 -0.085350  0.007377  0.067327  0.242455

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.804e-01  1.569e-01  -4.975 3.97e-05 ***
swamp        3.398e-02  5.767e-03   5.892 3.79e-06 ***
I(swamp^2)  -2.852e-04  4.624e-05  -6.166 1.90e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1177 on 25 degrees of freedom
Multiple R-squared: 0.6132,     Adjusted R-squared: 0.5823
F-statistic: 19.82 on 2 and 25 DF,  p-value: 6.972e-06
```

It is clear that once the effects of area are controlled for, a considerable amount of the remaining variability in herptile richness is explained by `swamp`, in a nonlinear fashion. If you examine the residuals, you will see that compared to the linear model, the fit is much better.

⌨ Based on the results from the above analyses, how would you modify the regression model arrived at above? What, in your view, is the "best" overall model? Why? How would you rank the various factors in terms of their effects on herptile species richness?

In light of these results, we might want to try and fit a model which includes `logarea, thtden, cpfor2, swamp` and `swamp^2`:

```
Call:
lm(formula = logherp ~ logarea + cpfor2 + thtden + swamp + I(swamp^2),
    data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.201797 -0.056170 -0.002072  0.051814  0.205626

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.203e-01  1.813e-01  -1.766   0.0912 .
logarea      2.202e-01  3.893e-02   5.656 1.09e-05 ***
cpfor2      -7.864e-04  9.955e-04  -0.790   0.4380
thtden      -2.929e-02  1.048e-02  -2.795   0.0106 *
swamp        3.113e-02  5.898e-03   5.277 2.70e-05 ***
I(swamp^2)  -2.618e-04  4.727e-05  -5.538 1.45e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1072 on 22 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8181,     Adjusted R-squared: 0.7767
F-statistic: 19.78 on 5 and 22 DF,  p-value: 1.774e-07
```

Note that on the basis of this analysis, we should probably drop `cpfor2` and refit using the remaining variables:

```
Call:
lm(formula = logherp ~ logarea + thtden + swamp + I(swamp^2),
    data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-0.19621 -0.05444 -0.01202  0.07116  0.21295

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.461e-01  1.769e-01  -1.957   0.0626 .
logarea      2.232e-01  3.842e-02   5.810 6.40e-06 ***
thtden      -2.570e-02  9.364e-03  -2.744   0.0116 *
swamp        2.956e-02  5.510e-03   5.365 1.89e-05 ***
I(swamp^2)  -2.491e-04  4.409e-05  -5.649 9.46e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1063 on 23 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8129,     Adjusted R-squared: 0.7804
F-statistic: 24.98 on 4 and 23 DF,  p-value: 4.405e-08
```

How about multicollinearity in this model?

```
> vif(model.polynomial.reduced)
   logarea     thtden      swamp I(swamp^2)
  1.053193   1.123491  45.845845  45.656453
```

VIF for the two swamp terms are much higher than the standard threshold of 5. However, this is expected for polynomial terms, and not really a concern given that both terms are highly significant in the

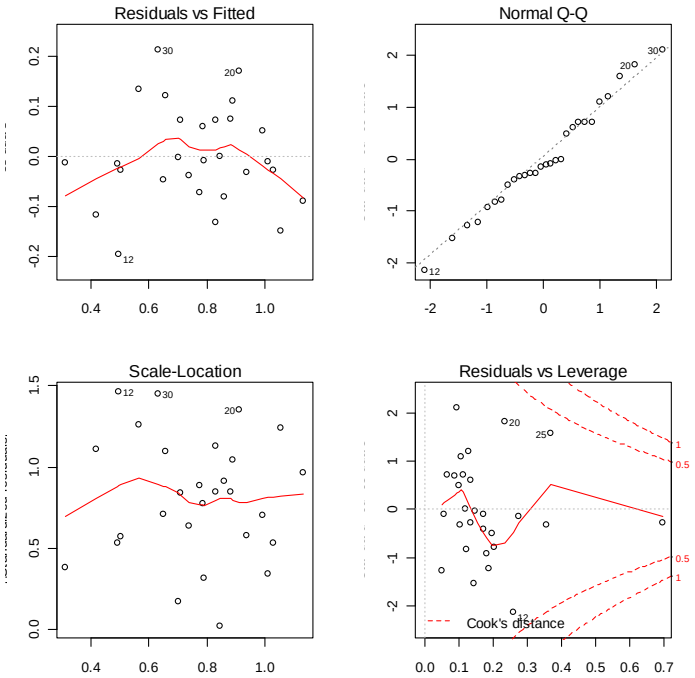
model. The high VIF means that these two coefficients are not estimated precisely, but using both in the model still allows to make a good prediction (i.e. account for the response to swamp).

## Checking assumptions of a multiple regression model

All the model selection techniques or the manual model crafting assumes that the standard assumptions (independence, normality, homoscedasticity, linearity) are met. Given that a large number of models can be fitted, it may seem that testing the assumptions at each step would be an herculean task. However, it is generally sufficient to examine the residuals of the full (saturated) model or of the final model. Terms not contributing significantly to the fit do not affect residuals much, and therefore, the residuals to the full model, or the residuals to the final model, are generally sufficient.

Let's have a look at the diagnostic plots for the final model:

**Figure 57.**




Everything looks about right here. For the skeptic, let's run the formal tests:

```
shapiro.test(residuals(model.polynomial.reduced))
```

```
        Shapiro-Wilk normality test

data:  residuals(model.polynomial.reduced)
W = 0.9837, p-value = 0.9278
```

The residuals do not deviate from normality. Good.

```
>
> #Homoscedasticity
> bptest(logherp ~ logarea + thtden + swamp + I(swamp^2),
varformula = ~fitted.values(model.polynomial.reduced),
+     studentize = TRUE, data = mydata)

        studentized Breusch-Pagan test

data:  logherp ~ logarea + thtden + swamp + I(swamp^2)
BP = 2.1938, df = 1, p-value = 0.1386
```

No deviation from homoscedasticity either. Good.

```
> #Serial autocorrelation
> dwtest(logherp ~ logarea + thtden + swamp + I(swamp^2), alternative =
"greater",
+   data = mydata)

        Durbin-Watson test

data:  logherp ~ logarea + thtden + swamp + I(swamp^2)
DW = 1.725, p-value = 0.2095
alternative hypothesis: true autocorrelation is greater than 0
```

No serial correlation in the residuals, so no evidence of non-independence.

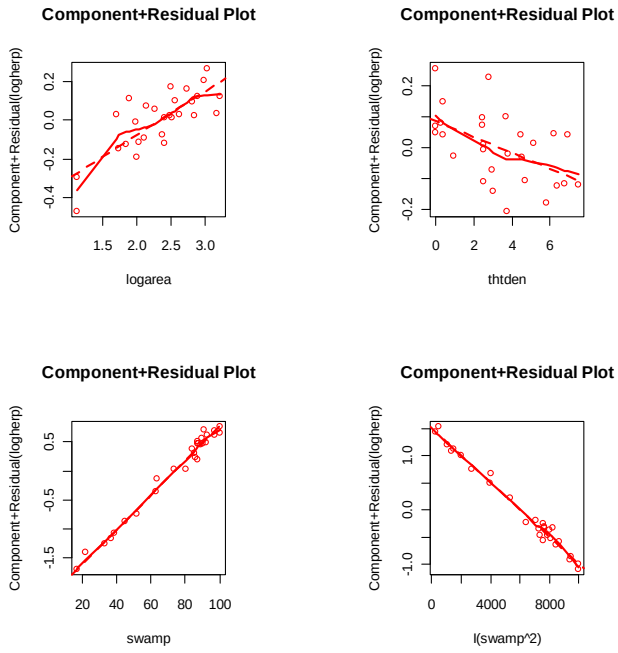
```
>
> #Linearity
> resettest(logherp ~ logarea + thtden + swamp + I(swamp^2), power = 2:3, type
= "regressor",
+     data = mydata)

        RESET test

data:  logherp ~ logarea + thtden + swamp + I(swamp^2)
RESET = 0.9823, df1 = 8, df2 = 15, p-value = 0.4859
```

And no significant deviation from linearity. So it seems that all is fine.

But, still something is missing. How about effect size? How is that measured or viewed? The regression coefficients can be used to measure effect size, although it may be better to standardize them so that they become independent of measurements. But a graph is often better. In this context, some of the most useful graphs are called partial residual plots (or component + residual plots). These plots show how the dependent variable, corrected for other variables in the model, varies with each individual variable. Let's have a look:

```
# Evaluate visually linearity and effect size
# component + residual plot
cr.plots(model.polynomial.reduced, one.page=TRUE, ask=FALSE)
```

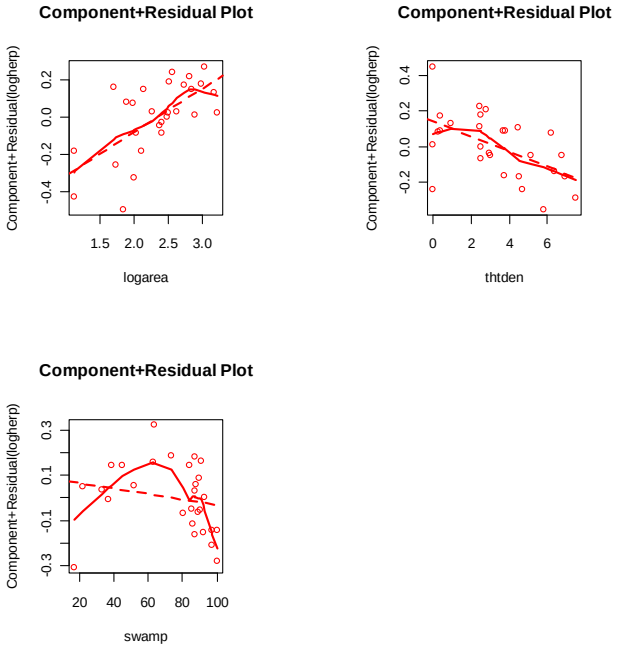
**Figure 58.**




Note that the vertical scale varies among plots. For thtden, the dependent variable (log10(herptile richness)) varies by about 0.4 units over the range of thtden in the sample. For logarea, the variation is about 0.6 log units. For swamp, it is a bit tricky since there are two terms and the have opposite effect (leading to a peaked relationship), so the plots are less informative. However, there is no deviation from linearity to be seen.

To illustrate what these graphs would look like if there was deviation from linearity, let's drop the squared swamp term and produce the graphs and run the resettest:

**Figure 59.**




The lack of linearity along the gradient of swamp becomes obvious. And the resettest also detects a violation from linearity:

```
 RESET test

data:  g
RESET = 6.7588, df1 = 6, df2 = 18, p-value = 0.0007066
```

## Testing for interactions

When there are multiple independent variables one should always be ready to assess interactions. In most multiple regression contexts this is somewhat difficult because adding interaction terms increases overall multicollinearity and because in many cases there are not enough observations to test all interactions, or the observations are not well balanced to make powerful tests for interactions.

Going back to our final model, see what happens if one tries to fit the fully saturated model with all interactions:

```
fullmodel.withinteractions<-lm(logherp ~ logarea * cpfor2 *
thtden * swamp * I(swamp^2), data= Mregdat)

summary(fullmodel.withinteractions)

Call:
lm(formula = logherp ~ logarea * cpfor2 * thtden * swamp * I(swamp^2),
    data = Mregdat)

Residuals:
```

```
ALL 28 residuals are 0: no residual degrees of freedom!

Coefficients: (4 not defined because of singularities)
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                         -5.948e+03         NA      NA       NA
logarea                              3.293e+03         NA      NA       NA
cpfor2                               7.080e+01         NA      NA       NA
thtden                               9.223e+02         NA      NA       NA
swamp                                1.176e+02         NA      NA       NA
I(swamp^2)                          -3.517e-01         NA      NA       NA
logarea:cpfor2                      -3.771e+01         NA      NA       NA
logarea:thtden                      -4.781e+02         NA      NA       NA
cpfor2:thtden                       -1.115e+01         NA      NA       NA
logarea:swamp                       -7.876e+01         NA      NA       NA
cpfor2:swamp                        -1.401e+00         NA      NA       NA
thtden:swamp                        -1.920e+01         NA      NA       NA
logarea:I(swamp^2)                   5.105e-01         NA      NA       NA
cpfor2:I(swamp^2)                    3.825e-03         NA      NA       NA
thtden:I(swamp^2)                    7.826e-02         NA      NA       NA
swamp:I(swamp^2)                    -2.455e-03         NA      NA       NA
logarea:cpfor2:thtden                5.359e+00         NA      NA       NA
logarea:cpfor2:swamp                 8.743e-01         NA      NA       NA
logarea:thtden:swamp                 1.080e+01         NA      NA       NA
cpfor2:thtden:swamp                  2.620e-01         NA      NA       NA
logarea:cpfor2:I(swamp^2)           -5.065e-03         NA      NA       NA
logarea:thtden:I(swamp^2)           -6.125e-02         NA      NA       NA
cpfor2:thtden:I(swamp^2)            -1.551e-03         NA      NA       NA
logarea:swamp:I(swamp^2)            -4.640e-04         NA      NA       NA
cpfor2:swamp:I(swamp^2)              3.352e-05         NA      NA       NA
thtden:swamp:I(swamp^2)              2.439e-04         NA      NA       NA
logarea:cpfor2:thtden:swamp         -1.235e-01         NA      NA       NA
logarea:cpfor2:thtden:I(swamp^2)     7.166e-04         NA      NA       NA
logarea:cpfor2:swamp:I(swamp^2)             NA         NA      NA       NA
logarea:thtden:swamp:I(swamp^2)             NA         NA      NA       NA
cpfor2:thtden:swamp:I(swamp^2)              NA         NA      NA       NA
logarea:cpfor2:thtden:swamp:I(swamp^2)      NA         NA      NA       NA

Residual standard error: NaN on 0 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:     1,      Adjusted R-squared:   NaN
F-statistic:   NaN on 27 and 0 DF,  p-value: NA
```

Indeed, it is not possible to include all 32 terms with only 28 observations. There are not enough data points, R square is one, and the model perfectly overfits the data.

If you try to use an automated routine to "pick" the best model out of this soup, R becomes humorous:

```
step(fullmodel.withinteractions)
```

```
Warning message:
attempting model selection on an essentially perfect fit is nonsense
```

Does this mean you can forget about potential interactions and simply accept the final model without a thought? No. You simply do not have enough data to test for all interactions. But there is a compromise worth attempting, comparing the final model to a model with a subset of the interactions, say all second order interactions, to check whether the inclusion of these interactions improves substantially the fit:

```
full.model.2ndinteractions<- lm(logherp ~ logarea + cpfor2 +
thtden + swamp + I(swamp^2)
  + logarea:cpfor2 + logarea:thtden + logarea:swamp
  + cpfor2:thtden + cpfor2:swamp
```

```
  + thtden:swamp
  , data= mydata)

summary(full.model.2ndinteractions)

Call:
lm(formula = logherp ~ logarea + cpfor2 + thtden + swamp + I(swamp^2) +
    logarea:cpfor2 + logarea:thtden + logarea:swamp + cpfor2:thtden +
    cpfor2:swamp + thtden:swamp, data = mydata)

Residuals:
      Min        1Q    Median        3Q       Max
-0.216880 -0.036534  0.003506  0.042990  0.175490

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.339e-01  6.325e-01   0.686 0.502581
logarea        -1.254e-01  2.684e-01  -0.467 0.646654
cpfor2         -9.344e-03  7.205e-03  -1.297 0.213032
thtden         -1.833e-01  9.035e-02  -2.028 0.059504 .
swamp           3.569e-02  7.861e-03   4.540 0.000334 ***
I(swamp^2)     -3.090e-04  7.109e-05  -4.347 0.000500 ***
logarea:cpfor2  2.582e-03  2.577e-03   1.002 0.331132
logarea:thtden  7.017e-02  3.359e-02   2.089 0.053036 .
logarea:swamp  -5.290e-04  2.249e-03  -0.235 0.816981
cpfor2:thtden  -2.095e-04  6.120e-04  -0.342 0.736544
cpfor2:swamp    4.651e-05  5.431e-05   0.856 0.404390
thtden:swamp    2.248e-04  4.764e-04   0.472 0.643336
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.108 on 16 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared: 0.8658,     Adjusted R-squared: 0.7735
F-statistic: 9.382 on 11 and 16 DF,  p-value: 4.829e-05
```

This model fits the data slightly better than the full model (it explains 86.6% of the variance in logherp, compared to 81.2% for the full model without interactions), but has twice as many parameters. If you look at the individual coefficients, some weird things happen: for example, the sign for logarea has changed. This is one of the symptoms of multicollinearity. Let's look at the variance inflation factors:

```
vif(full.model.2ndinteractions)
       logarea         cpfor2         thtden          swamp     I(swamp^2)
      49.86060       78.49622      101.42437       90.47389      115.08457
logarea:cpfor2 logarea:thtden  logarea:swamp  cpfor2:thtden   cpfor2:swamp
      66.97792       71.69894       67.27034       14.66814       29.41422
  thtden:swamp
      20.04410
```

Ouch. All VIF are above 5, not only the ones involving the swamp terms. This model is not very satisfying it seems. Indeed the AIC for the two models indicate that the model with interactions has less information than the full model (remember, models with the lowest AIC value are to be preferred):

```
> AIC(full.model)
[1] -38.3433
> AIC(full.model.2ndinteractions)
[1] -34.86123
```

The anova() command can be used to test whether the addition of all interaction terms improves the fit significantly:

```
anova(full.model, full.model.2ndinteractions)
```
```
Analysis of Variance Table

Model 1: logherp ~ logarea + cpfor2 + thtden + swamp + I(swamp^2)
Model 2: logherp ~ logarea + cpfor2 + thtden + swamp + I(swamp^2) + loga-
rea:cpfor2 +
    logarea:thtden + logarea:swamp + cpfor2:thtden + cpfor2:swamp +
    thtden:swamp
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     22 0.252820
2     16 0.186507  6  0.066314 0.9481  0.489
```

This test indicates that the addition of interaction terms did not reduce significantly the residual variance around the full model. How about a comparison with the final model without cpfor2?

```
anova(model.polynomial.reduced, full.model.2ndinteractions)
```
```
Analysis of Variance Table

Model 1: logherp ~ logarea + thtden + swamp + I(swamp^2)
Model 2: logherp ~ logarea + cpfor2 + thtden + swamp + I(swamp^2) + loga-
rea:cpfor2 +
    logarea:thtden + logarea:swamp + cpfor2:thtden + cpfor2:swamp +
    thtden:swamp
  Res.Df      RSS Df Sum of Sq      F Pr(>F)
1     23 0.259992
2     16 0.186507  7  0.073486 0.9006 0.5294
```

And this comparison suggests that our final model does not make worse predictions than the full model with interactions.

## Dredging and the information theoretical approach

One of the main critiques of stepwise methods is that the p-values are not strictly correct because of the large number of tests that are actually done. This is the multiple testing problem. In building linear models (multiple regression for example) from a large number of independent variables, and possibly their interactions, there are so many possible combinations that if one were to use Bonferroni type corrections, it would make tests very conservative.

An alternative, very elegantly advocated by Burnham and Anderson (2002, Model selection and multimodel inference: a practical information-theoretic approach. 2nd ed), is to use AIC (or better the AICc that is more appropriate for samples where the number of observations is less that about 40 times the number of variables) to rank potential models, and identify the set of models that are best ones. One can then average the parameters across models, weighing using the probability that it is the best model to obtain coefficients that are more robust and less likely to be unduly affected by multicollinearity.

This approach has been implemented in the package MuMIn

```
library(MuMIn)
```

```
dd <- dredge(full.model.2ndinteractions)
```

Object dd will contain all possible models using the terms of our full model with 2nd order interactions.

Then, we can have a look at the subset of models that have an AICc within 4 units from the lowest AICc model. (Burnham and Anderson suggest that models that deviate by more than 4 AICc units have very little empirical support):

```
> top.models.1 <- get.models(dd, subset = delta < 4) # get
subset of best models
> model.avg(top.models.1) # get averaged coefficients

Model summary: ❶

           Deviance   AICc Delta Weight ❷
2+3+4+5+7      0.23 -35.95  0.00   0.34
2+3+4+5        0.26 -35.56  0.39   0.28
1+2+3+4+5+7    0.22 -33.02  2.93   0.08
2+3+4+5+7+8    0.22 -32.95  3.00   0.08
1+2+3+4+5      0.25 -32.74  3.21   0.07
2+3+4+5+8      0.25 -32.51  3.44   0.06
2+3+4+5+6+7    0.22 -32.17  3.78   0.05
2+3+4+5+6      0.26 -31.99  3.97   0.05

Variables:
             1              2             3             4             5
        cpfor2     I(swamp^2)       logarea         swamp       thtden
             6              7             8
 logarea:swamp logarea:thtden   swamp:thtden

Averaged model parameters: ❸
               Coefficient Variance       SE Unconditional SE  Lower CI
(Intercept)      -2.08e-01 3.81e-03 2.45e-01         2.56e-01 -0.709000
cpfor2           -1.20e-04 4.92e-14 2.78e-04         2.85e-04 -0.000678
I(swamp^2)       -2.68e-04 5.78e-18 4.90e-05         5.15e-05 -0.000369
logarea           1.31e-01 1.97e-04 1.16e-01         1.19e-01 -0.103000
swamp             3.19e-02 1.41e-09 6.12e-03         6.43e-03  0.019300
thtden           -6.84e-02 8.04e-06 5.27e-02         5.39e-02 -0.174000
logarea:swamp     4.38e-05 8.37e-14 2.07e-04         2.17e-04 -0.000381
logarea:thtden    2.14e-02 3.94e-07 2.49e-02         2.54e-02 -0.028400
swamp:thtden     -3.28e-05 4.06e-16 7.95e-05         8.16e-05 -0.000193
                Upper CI
(Intercept)     0.294000
cpfor2          0.000438
I(swamp^2)     -0.000167
logarea         0.365000
swamp           0.044500
thtden          0.037200
logarea:swamp   0.000469
logarea:thtden  0.071100
swamp:thtden    0.000127

Relative variable importance: ❹
    I(swamp^2)        logarea          swamp         thtden logarea:thtden
          1.00           1.00           1.00           1.00           0.55
        cpfor2   swamp:thtden  logarea:swamp
          0.15           0.14           0.10
:
```

❶ You first get the list of the models with an AICc within the desired 4 units of the best model. The variables that are included in the model are coded with the key just below.

❷ For each model, in addition to the AICc, the Akaike weights are calculated. These estimate the probability that this model is the best one. Here we can see that the first (best) model has only 34% chances of being really the best one.

❸ For the subset of models, weighted averages (using Akaike weights) for model parameters are calculated, with 95% CI. Note that, by default, terms missing from a model are assumed to have a coefficient of 0.

❹ I am not sure how this is calculated. But clearly it is related to the influence of each variable in the model.

## Bootstrapping multiple regression

When data do not meet the assumptions of normality and homoscedasticity and that it is not possible to transform the data to meet the assumptions, bootstrap can be used to compute confidence intervals for coefficients. If the distribution of the bootstrapped coefficients is symmetrical and approximately Gaussian, ,then empirical percentiles can be used to estimate the confidence limits.

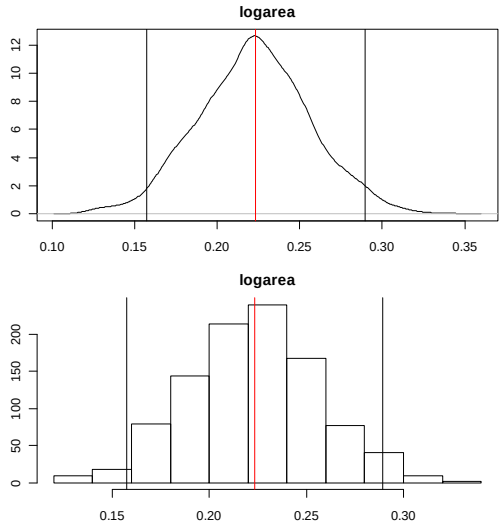
The following code, using library simpleboot has been designed to be easily modifiable and will compute CI using empirical percentiles

```
#############################################################
#######
#  Bootstrap analysis the simple way with library simpleboot

# Define model to be bootstrapped and the data source used
mymodel<-lm(logherp ~ logarea + thtden + swamp + I(swamp^2),
data = mydata)

#  Set the number of bootstrap iterations
nboot<-1000

require(simpleboot)

#  R is the number of bootstrap iterations
#  Setting rows to FALSE indicates resampling of residuals
mysimpleboot<-lm.boot(mymodel, R = nboot, rows = FALSE)

#  Extract bootstrap coefficients
myresults<-sapply(mysimpleboot$boot.list, function (x)
x$coef)

#  Transpose matrix so that lines are bootstrap iterations
and columns are coefficients
tmyresults<-t(myresults)

#  Plot histograms of bootstrapped coefficients
```

```
ncoefs<-length(data.frame(tmyresults))


par(mfrow=c(2,1), mai=c(0.5, 0.5, 0.5, 0.5), ask = TRUE)

for (i in 1 : ncoefs) {
  lab<-colnames(tmyresults)[i]
  x<-tmyresults[,i]

  plot(density(x),main=lab, xlab="")
  abline(v=mymodel$coef[i], col = "red")
  abline(v=quantile(x, c(0.025, 0.975)))
  hist(x, main=lab, xlab="")
  abline(v=quantile(x, c(0.025, 0.975)))
  abline(v=mymodel$coef[i], col = "red")
  }

par(mfrow=c(1,1), ask = FALSE)
```

When run, it will pause to let you have a look at the distribution for each coefficient in the model by producing plots like:

**Figure 60.**




The top plot is the kernel density and the bottom one is the histogram of the bootstrap estimates for the coefficient. On these plots, the red line indicate the value of the parameter in the ordinary analysis, and the two vertical black lines mark the limits of the 95% confidence interval. Here the CI does not include 0 and one can conclude that the effect of logarea on logherp is significantly positive.

Precise values for the limits can be obtained by:

```
> #  Display empirical bootstrap quantiles (not corrected
for bias)
> p <- c(0.005,0.01, 0.025, 0.05, 0.95, 0.975, 0.99, 0.995)
> apply(tmyresults, 2, quantile, p)
      (Intercept)   logarea      thtden      swamp    I(swamp^2)
```

```
0.5%  -0.735605927 0.1321902 -0.048153235 0.01794482 -0.0003529401
1%    -0.706786582 0.1444830 -0.045895772 0.01886999 -0.0003384226
2.5%  -0.644649399 0.1575567 -0.043016668 0.02015046 -0.0003263934
5%    -0.600986378 0.1679613 -0.039140810 0.02166009 -0.0003152855
95%   -0.075274798 0.2801058 -0.012076119 0.03777323 -0.0001859188
97.5% -0.035536091 0.2894439 -0.009660800 0.03929003 -0.0001715579
99%   -0.004244384 0.3014986 -0.007446500 0.04091320 -0.0001612558
99.5%  0.031621494 0.3095612 -0.006314632 0.04301388 -0.0001566147
```

These confidence limits are not reliable when the distribution of the bootstrap estimates deviate from Gaussian. If they do,, then it is preferable to compute so-called bias-corrected accelerated (BCa) confidence limits. The following code does just that:

```
################################################
#  Bootstrap analysis in multiple regression with BCa confi-
dence intervals
#  Preferable when parameter distribution is far from normal

# Bootstrap 95% BCa CI for regression coefficients
library(boot)


# function to obtain regression coefficients for each itera-
tion
bs <- function(formula, data, indices) {
  d <- data[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(coef(fit))
}

# bootstrapping with 1000 replications
results <- boot(data=mydata, statistic=bs,
   R=1000, formula=logherp ~ logarea + thtden + swamp +
I(swamp^2))

# view results
results
plot(results, index=1) # intercept
plot(results, index=2) # logarea
plot(results, index=3) # thtden
plot(results, index=4) # swamp
plot(results, index=5) # swamp^2



# get 95% confidence intervals
boot.ci(results, type="bca", index=1) # intercept
boot.ci(results, type="bca", index=2) # logarea
boot.ci(results, type="bca", index=3) # thtden
boot.ci(results, type="bca", index=4) # swamp
boot.ci(results, type="bca", index=5) # swamp^2
```

```
results

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = mydata, statistic = bs, R = 1000, formula = logherp ~
```

```
    logarea + thtden + swamp + I(swamp^2))


Bootstrap Statistics :
         original        bias     std. error
t1* -0.3460988043  2.613804e-04 2.346177e-01
t2*  0.2232323394 -5.971629e-03 5.222778e-02
t3* -0.0256957141 -3.946859e-04 9.411536e-03
t4*  0.0295628938  4.057821e-04 7.148861e-03
t5* -0.0002490924 -2.438465e-06 5.341945e-05
```

It will also display the standard graphs for each coefficient,, and the resulting BCa CI. For example, for loga:

```
> boot.ci(results, type="bca", index=2) # logarea
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1000 bootstrap replicates

CALL :
boot.ci(boot.out = results, type = "bca", index = 2)

Intervals :
Level       BCa
95%   ( 0.1212,  0.3225 )
Calculations and Intervals on Original Scale
```

Note that the BCa interval is from 0.12 to 0.32, whereas the simpler percentile interval is 0.16 to 0.29. BCa interval here is longer on the low side, and shorter on the high side, which it should given the distribution of bootstrap estimates.

## Permutation test

Permutation tests are more rarely performed in multiple regression contexts than bootstrap. But here is code to do it (**beware. I sometimes get strange results with the lmPerm library and I wonder if it is buggy).**

```
############################################################
###########
#  Permutation in multiple regression
#
#  using lmperm library

require(lmPerm)

#  Fit desired model on the desired dataframe
mymodel<-lm(logherp ~ logarea + thtden + swamp + I(swamp^2),
data = mydata)

mymodelProb<-lmp(logherp ~ logarea + thtden + swamp +
I(swamp^2), data = mydata, perm="Prob")


summary(mymodel)
summary(mymodelProb)
```