R Commander
and
the NMBU plugin

A short introduction

by

Kristian Hovde Liland

Created using the packages:

RcmdrPlugin.NMBU version 1.8.7

and

mixlm version 1.1.5

Contents

1 Introduction

This document is intended to give a brief overview of typical usage of the R Commander for educational purposes. It is not a comprehensive guide, and only covers a minor subset of the available dialogues and functions in the R Commander and the NMBU plugin.

For instructions regarding installation, please refer to the platform specific documents at http://repository.umb.no/R.

Given a complete installation of the R Commander and the NMBU plugin, starting up is done by first starting your preferred R GUI (R x64 / R i386 / R.app / RStudio / …) and the writing the following in the R Console:

```
library(nmbu)
```

Start-up can take a few seconds depending on availability of an internet connection (triggers download of the newest update of the R package RcmdrPlugin.NMBU) and alternatively the speed of the connection. A successful start-up will show a variation of the following window:

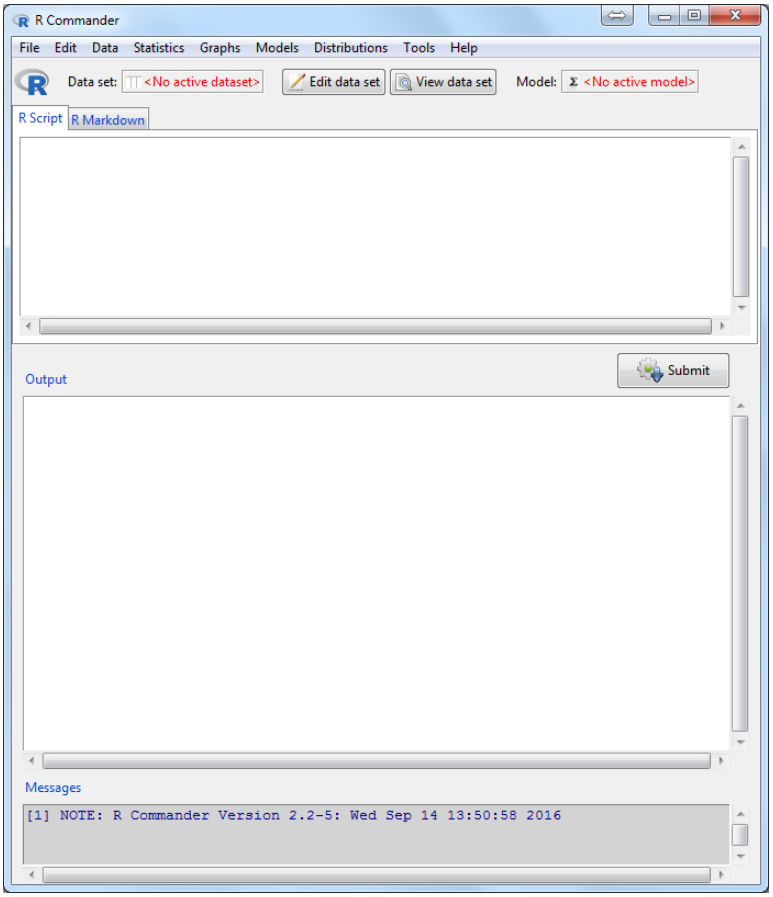


Figure 1: Freshly started R Commander (in Windows 7 colours). RStudio users will not have an Output pane or Messages pane below the Submit button.

1.1 Copying results to a text editor

Use the font `Courier New` (or equivalent monospaced font) and single line spacing on text copied from R or RStudio into Word or other text editors to keep the alignment seen in R/RStudio. The first line that is copied sometimes looses a couple of spaces in the beginning when pasted into Word.

Table copied without changing font:

Anova Table (Type II tests)

Response: height

	Sum Sq	Df	F value	Pr(>F)	
gender	7.2586	1	2.7225	0.15985	
length	19.5869	1	7.3466	0.04225	*
Residuals	13.3306	5			

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

Table with single spaced `Courier New`

```
Anova Table (Type II tests)

Response: height
           Sum Sq Df F value  Pr(>F)
gender     7.2586  1  2.7225 0.15985
length    19.5869  1  7.3466 0.04225 *
Residuals 13.3306  5
---
Signif. codes:  0 '***' 0.001 '**' 0.01
'*' 0.05 '.' 0.1 ' ' 1
```

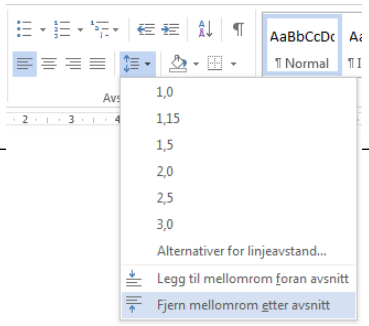


Figure 2: Effect of choice of font and line spacing.
Example from Norwegian Microsoft Word 2013 (Remove space after paragraph)

2 Data handling

2.1 Automatic import

See the Appendix for example data to play with.

1. Copy data with headings/variable names from a document, spreadsheet or other type of table.
2. Select the menu item `Data -> Automatic import from clipboard ….`
3. Write a suitable name for the imported data set (no spaces or special/mathematical signs).
4. A successful import leads to a summary of what was imported (check it for errors!) and the name of the data in blue letters below the menu bar.

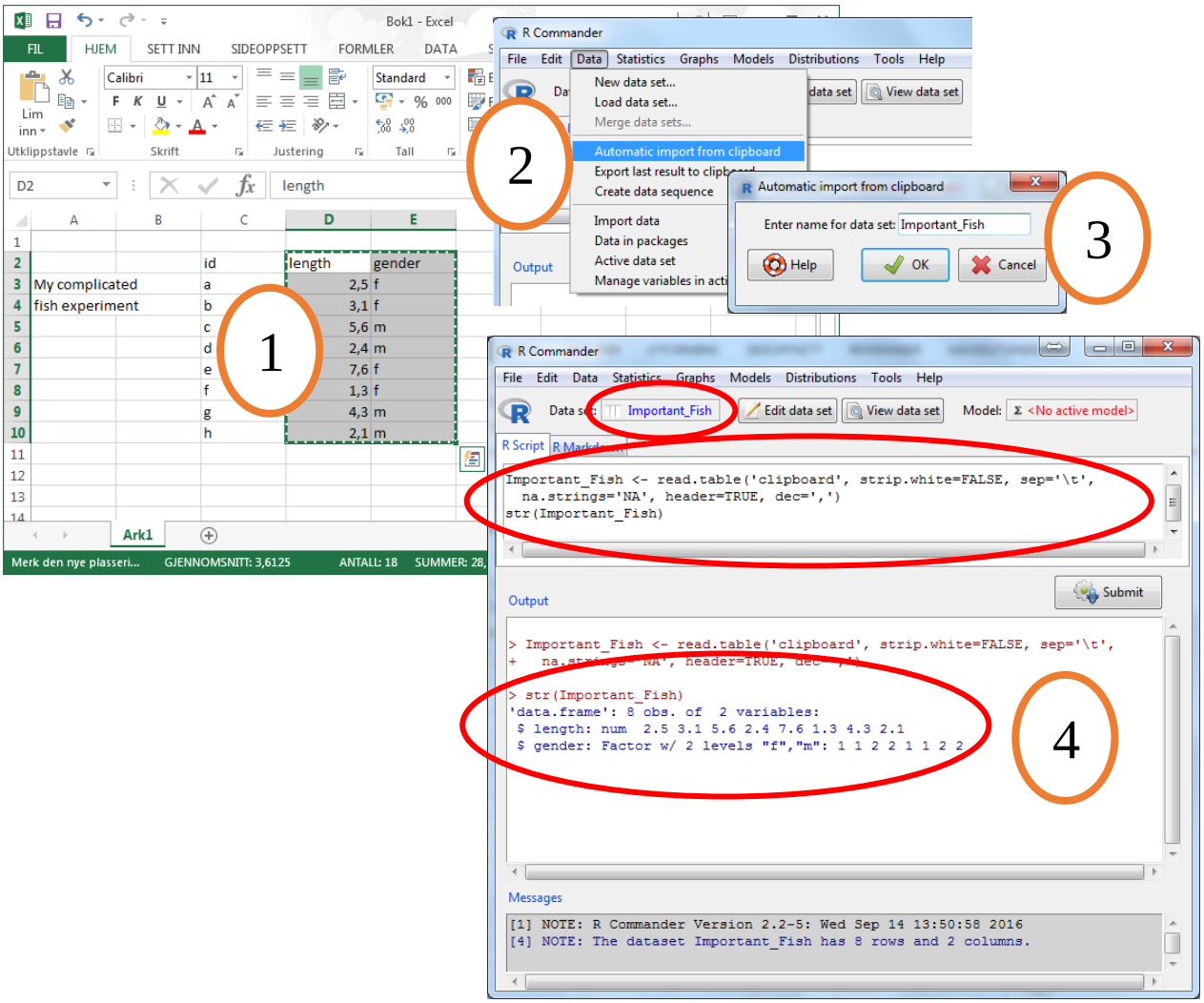


Figure 3: Automatic import from clipboard

The automatic import will attempt to recognize use of commas and dots as decimal marks, use of a header line, and various column separators. If the import failed in some regards, please check if the read.table command in the R Script pane (see Figure 3, part 4) can be adjusted to interpret your data format correctly.

2.2 Import from data files

The R Commander can import from several data formats. These are found in the `Data -> Import data` submenu (see Figure 4).

- `from text file, clipboard, or URL`... is especially suited for "flat" data files, e.g. from instruments, with filenames ending with .txt or .dat.
- `from Excel file`... can handle well-organised sheets where there is a single data table.

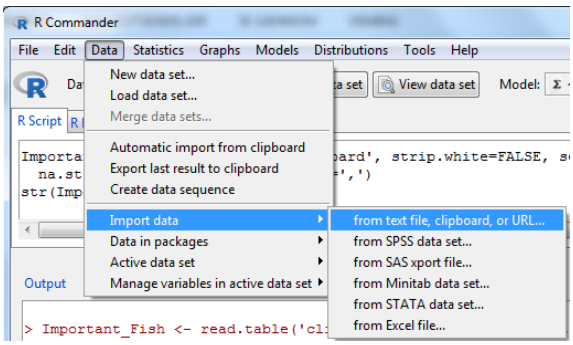


Figure 4: Import data menus

2.3 Loading / saving R data

If the data set has been saved in R's format with a file name ending in .RData or .rda, this is simply loaded through the menu item `Data -> Load data set`....

Save the current data set through the menu item `Data -> Active data set -> Save active data set`... (see Figure 5).

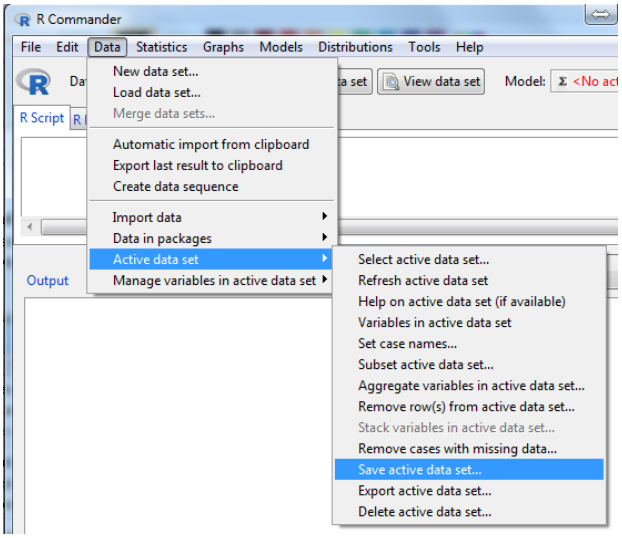


Figure 5: Saving a data set

2.4 Managing variables

From the submenu `Data -> Active data set` one can perform actions involving **all** variables in the active data set:

- `Stack variables in active data set…`
  (prepare data for ANOVA or two-sample testing)
- `Subset active data set…`
- `Save active data set…`
- `Export active data set…`
- `Delete active data set…`

From the submenu `Data -> Manage variables in active data set` one can perform the following actions (and more) for **single** variables in the active data set:

- `Sort…`
- `Mean centre…`
- `Standardize…`
- `Convert numeric values to factor…` (e.g. to use as groupings/levels in two sample tests, ANOVA, group colouring, etc.)
- `Compute new variable…` (make a new variable from an expression based on existing variables)
- `Reorder factor levels…` (e.g. to change the reference level in regression)

3 Graphics

3.1 Scatter plots

When looking for a relationship between two continuous variables, use `Graphs -> Scatterplot...` to get a first impression (see Figure 6). Many extras can be added through the Options pane, e.g. a linear regression line. `Plot by groups...` adds different colours and symbols to groups.

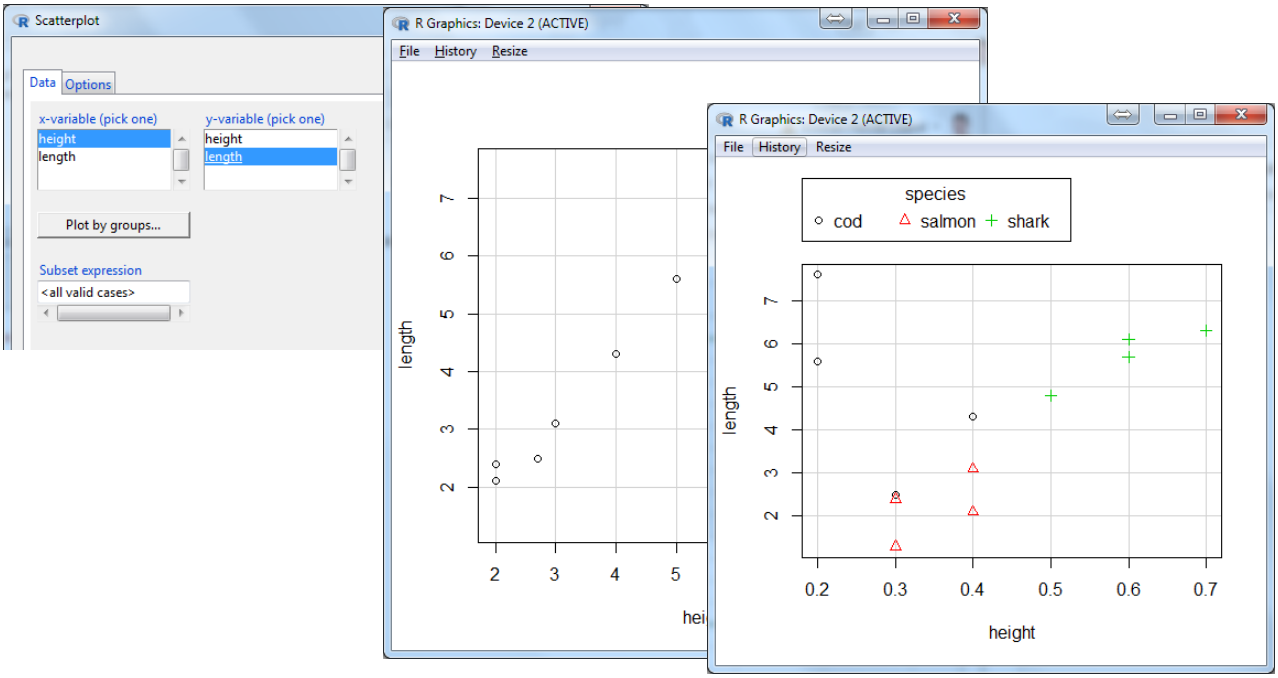


Figure 6: Scatter plot

For more advanced relationships with confidence intervals and prediction intervals, one can use the `Graphs -> Fitted regression plot...` (see Figure 7).

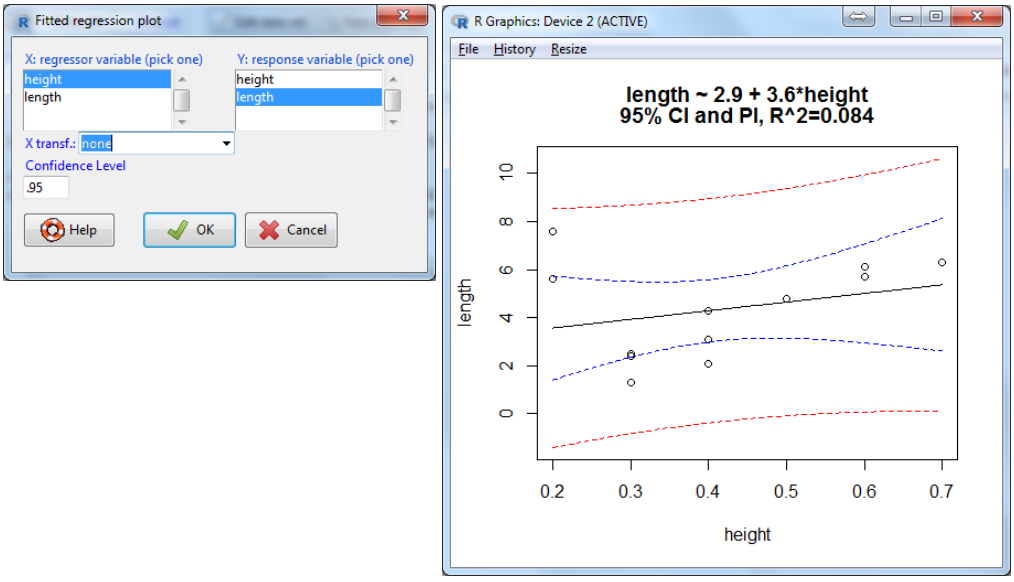


Figure 7: Scatter plot with fitted regression

3.2 Plots of means

Data with a grouping variable can be plotted group-wise with error bars using the Graphs -> Plot of means… (see Figure 8).

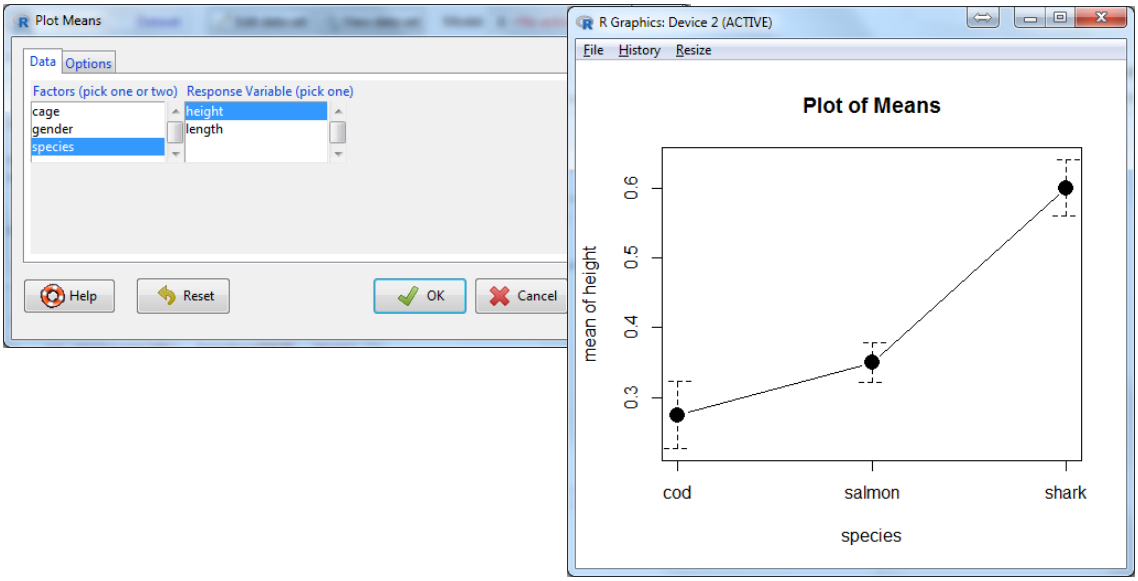


Figure 8: Plot of means

3.3 Line plots

If your data has a natural order and possibly contains a grouping variable, the Graphs -> Line and point plot… is useful (see Figure 9).

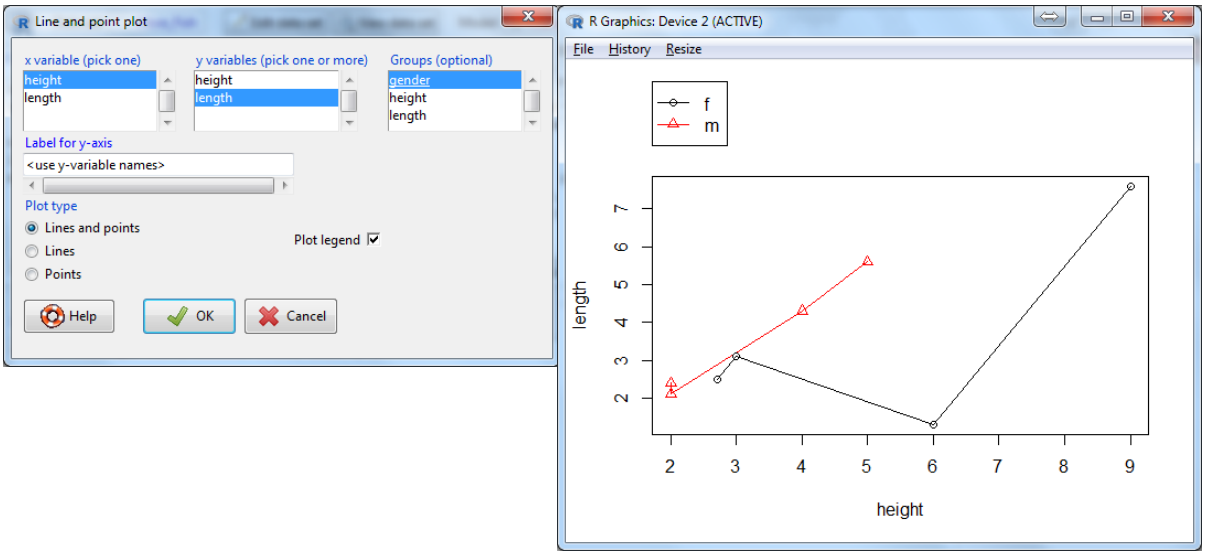


Figure 9: Line and point plot with grouping

3.4 More plots

Some much used plots are histograms and boxplots, also found in the `Graphs` menu (see Figure 10).

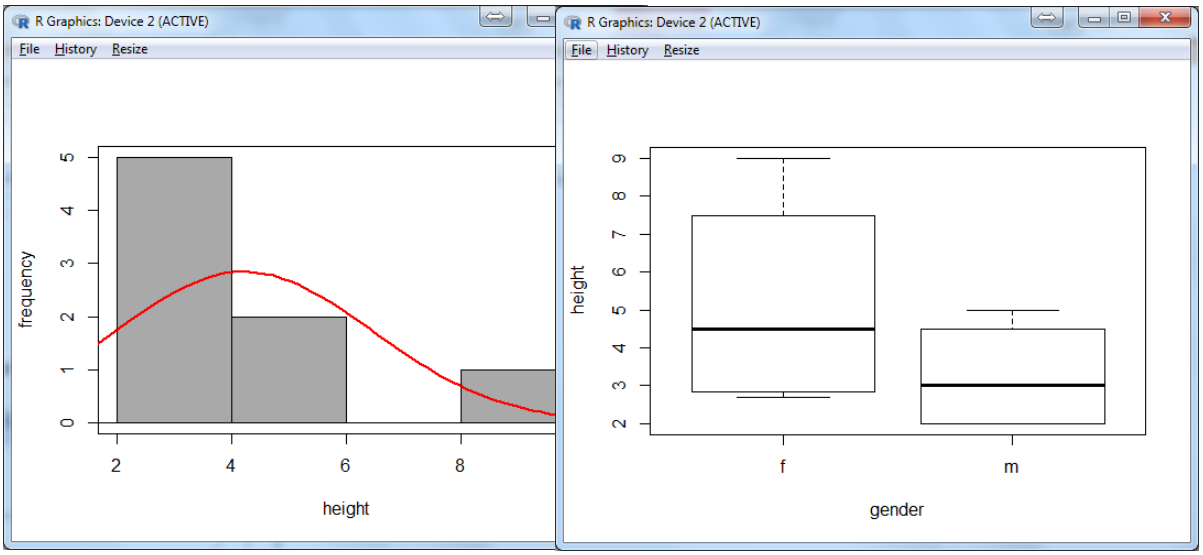


Figure 10: Histogram and boxplot

4 Statistics

The Statistics section only shows some of the available tests in the R Commander and does not explain when to use the different methods or which assumptions need to be made for them to be correct.

4.1 Descriptive statistics

Descriptive statistics are available through the `Statistics -> Summaries -> Numerical summaries...` menu item. The chosen variables are summarized according to the ticked statistics, optionally summarized by groups if selected (see Figure 11).

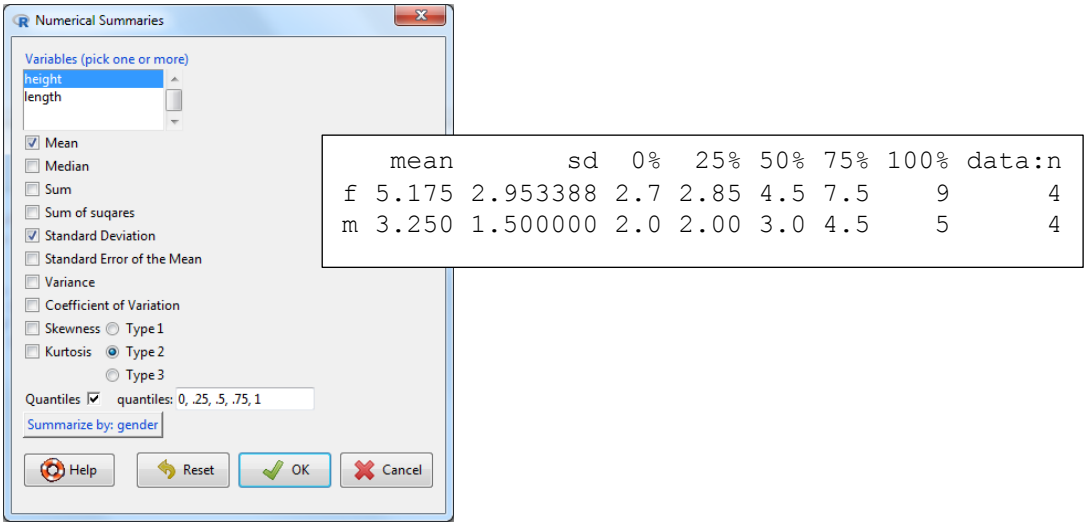


Figure 11: Numerical summaries of variables in active data set

4.2 Comparisons of means

There are several options when testing for differences between mean values in the R Commander. These are mainly found in the t-test (unknown standard deviation(s)) and z-test (known standard deviation(s)) submenus of `Statistics -> Means`.

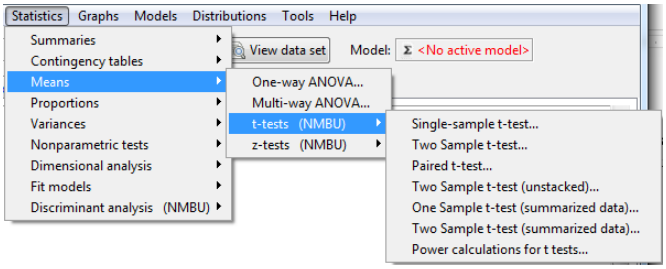


Figure 12: Comparisons of means

A short summary of the possibilities:

Name	Comparison	Variant	Assumption
Single sample / One sample	Mean value against a chosen value (default = 0)	Two-sided or one-sided	
Two sample	Difference between two means against a chosen value	Two-sided or one-sided	Equal/non-equal variance
Paired	Mean of paired differences against a chosen value	Two-sided or one-sided	

The Two Sample tests are based on stacked data (see **2.4 Managing variables**). In addition, there are variants based on summarized data and an unstacked two-sample test. Corresponding tests for proportions are found in the submenu `Statistics -> Proportions`.

4.3 Regression

The menu item `Statistics -> Fit models -> Linear model...` is the most versatile and useful way of performing regression in the R Commander (see Figure 13). A simpler alternative with fewer options is the `Linear regression...` menu item in the same submenu.

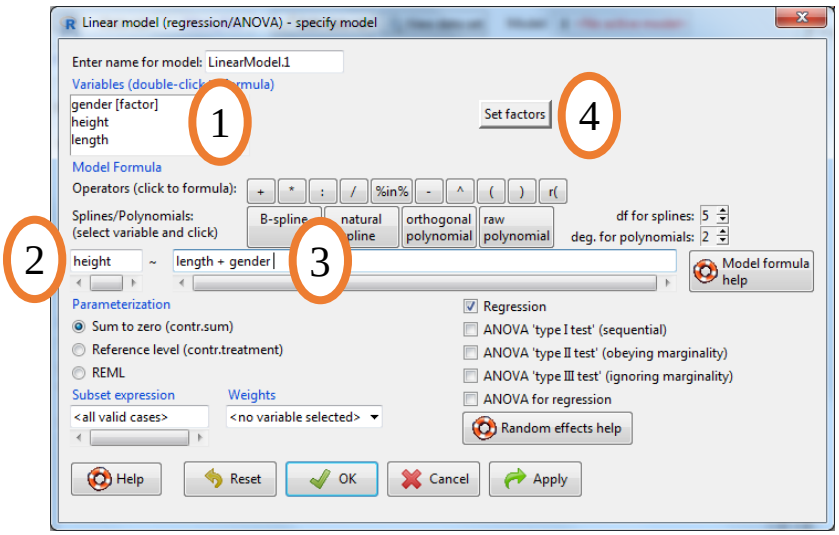


Figure 13: Regression through the `Linear model` dialogue.

Variables are shown in the top left (1). Double-clicking on variables will copy them to the response field (2) (first variable) and to the predictor field (3) (after the first variable). One can also write and edit in the fields by hands. If a predictor is coded as numeric (continuous), but should be included in the regression model as a factor (grouping/categorical), this can be done using the `Set factors` button (4). For model selection, prediction, diagnosis and model graphics see section **4.5 Models**.

```
Call:
lm(formula = height ~ length + gender, data = More_Fish)

Residuals:
       1        2        3        4        5        6        7        8
-1.57971 -1.75720  0.15837 -0.29502  0.66164  2.67526  0.19293 -0.05628

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.3376     1.2076   1.108   0.3184
length        0.7958     0.2936   2.710   0.0423 *
gender(f)     0.9526     0.5773   1.650   0.1599
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

s: 1.633 on 5 degrees of freedom
Multiple R-squared: 0.6695,
Adjusted R-squared: 0.5372
F-statistic: 5.063 on 2 and 5 DF,  p-value: 0.06282
```

Estimated regression coefficients

P-values for two-sided test (coeff. ≠ 0)

Model $\hat{\sigma}$

R^2 and R^2_{adj}

F-test for model

Figure 14: Summary printout from regression.

4.4 Analysis of variance (ANOVA)

When performing analysis of variance, we use the same interface as when doing regression (see previous section, Figure 13 and Figure 15). We need to make sure that there are only factor predictors (1) (categorical/grouping), and we need to specify which ANOVA type (SS) should be displayed (2).

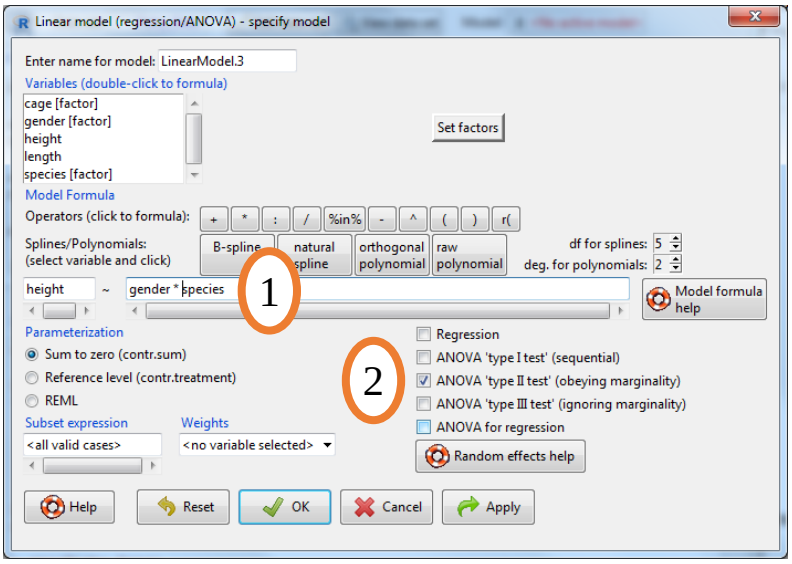


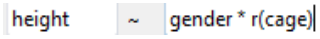
Figure 15: Analysis of variance using the `Linear model` dialogue.

```
Anova Table (Type II tests)

Response: height
                Sum Sq Df F value  Pr(>F)
gender        0.002604  1  0.3255 0.59299
species       0.236771  3  9.8655 0.01532 *
gender:species 0.011563  2  0.7227 0.53005
Residuals     0.040000  5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 16: Printout from ANOVA type II.

To specify a random effect, enclose it in `r()` in the specification of predictors, e.g.



The printout will change accordingly (see Figure 17).

For model selection, prediction, diagnosis and model graphics see section **4.5 Models**.

```
Analysis of variance (unrestricted model)
Response: height
            Mean Sq Sum Sq Df F value Pr(>F)
gender       0.0008 0.0008  1    0.04 0.8743
cage         0.0008 0.0008  1    0.04 0.8743
gender:cage  0.0208 0.0208  1    0.63 0.4520
Residuals    0.0333 0.2667  8       -      -

               Err.term(s) Err.df   VC(SS)
1 gender               (3)      1    fixed
2 cage                 (3)      1 -0.00333
3 gender:cage          (4)      8 -0.00417
4 Residuals              -      -  0.03333
(VC = variance component)

            Expected mean squares
gender       (4) + 3 (3) + 6 Q[1]
cage         (4) + 3 (3) + 6 (2)
gender:cage  (4) + 3 (3)
Residuals    (4)
```

Figure 17: Printout of mixed effect model.

4.5 Models

After a regression or ANOVA model has been fitted, several options become available in the `Models` menu (see Figure 18). The active model is shown in blue in the top right corner of the R Commander, e.g. 

.

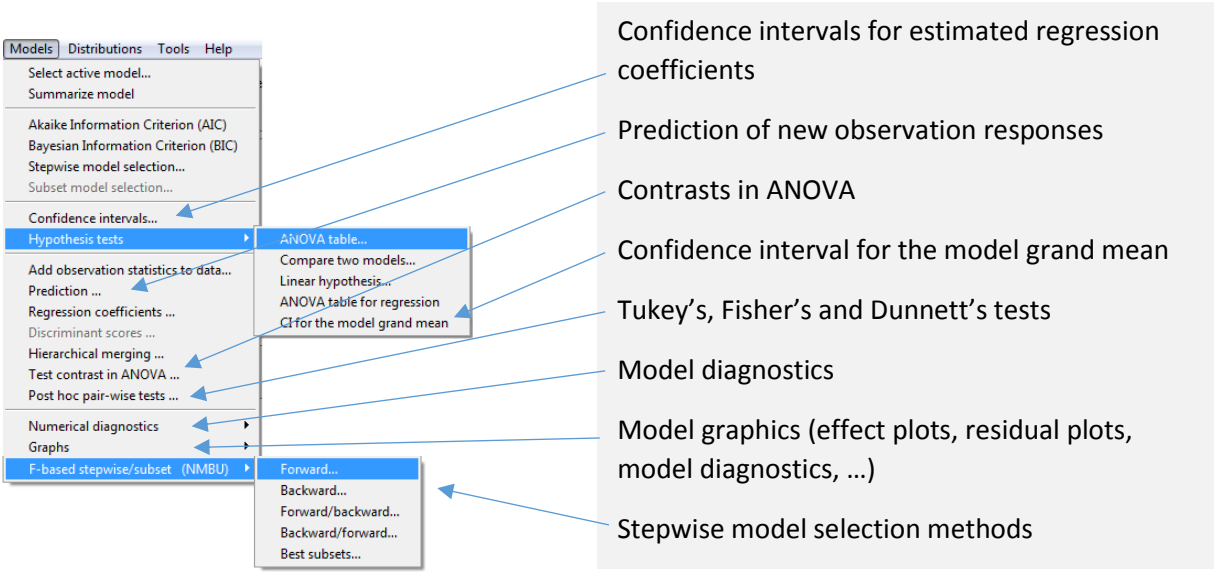


Figure 18: Models menu for working with fitted models.

4.6 Clustering

In the submenu `Statistics -> Dimensional analysis -> Cluster analysis` there are two main types of clustering available, k-means and hierarchical. With the k-means clustering the number of clusters is chosen in advance (see Figure 19) and an iterative procedure is used to search for clusters in the input variables. Redoing the clustering may lead to a different result.

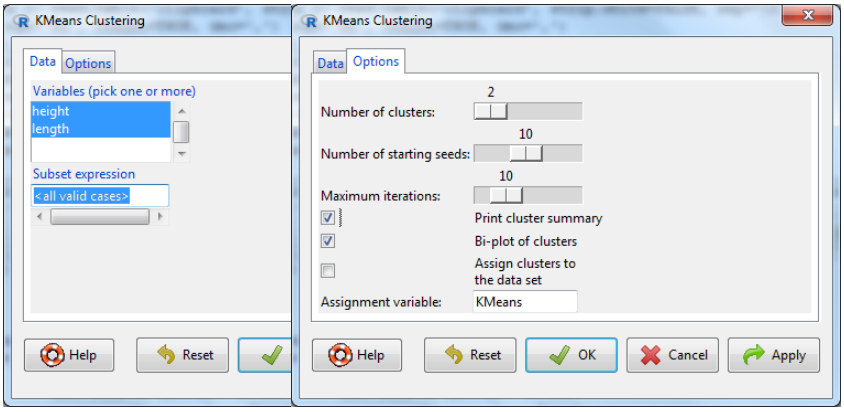


Figure 19: K-means clustering

In hierarchical clustering one has to choose a clustering method and a distance measure (see Figure 20) which will heavily affect the resulting clusters, usually visualized in a dendrogram.

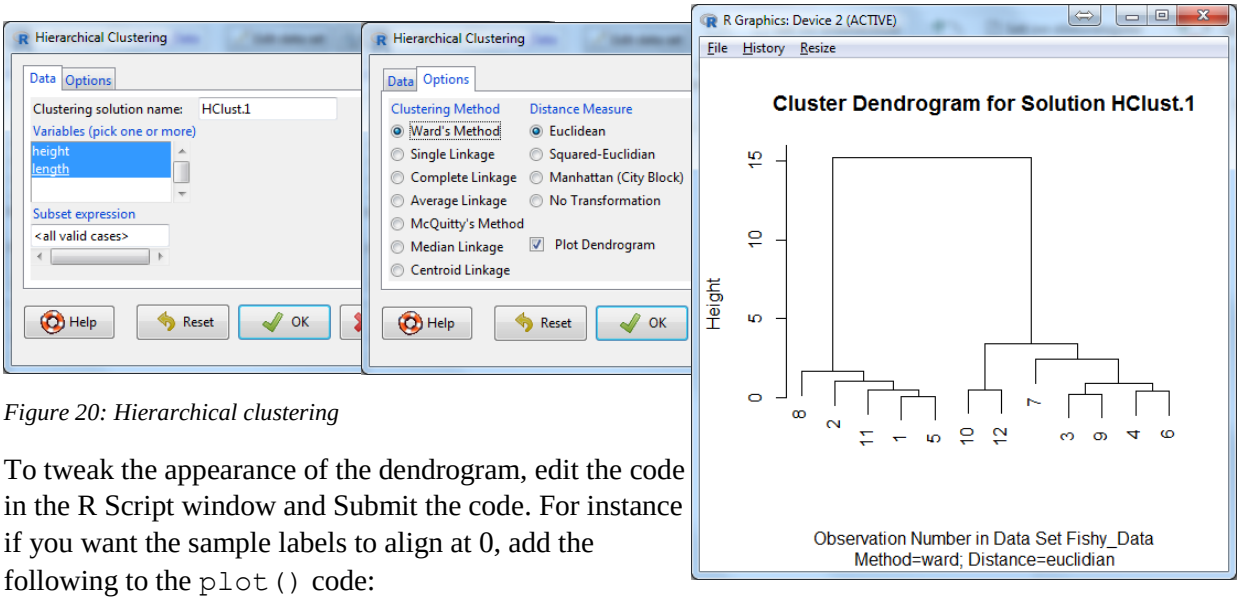
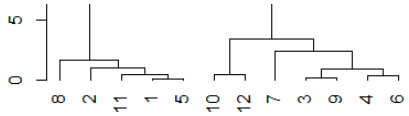


Figure 20: Hierarchical clustering

To tweak the appearance of the dendrogram, edit the code in the R Script window and Submit the code. For instance if you want the sample labels to align at 0, add the following to the `plot()` code:

```
, hang = -1
```

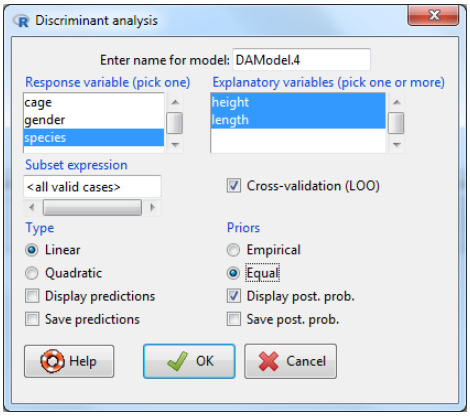
… so it becomes:




```
plot(HClust.1, main= "Cluster Dendrogram for Solution HClust.1", xlab=
  "Observation Number in Data Set Fishy_Data",
  sub="Method=ward; Distance=euclidian", hang = -1)
```

4.7 Classification

One can use the menu item `Statistics -> Discriminant analysis -> LDA/QDA` to perform classification. The response must be a factor (categorical/grouping) while the predictors must be numeric (continuous). Predictors can also be saved scores from principal component analysis/regression or partial least squares.




```
          True
Predicted cod salmon shark
  cod       2      0     0
  salmon    2      4     0
  shark     0      0     4
  Total     4      4     4
  Correct   2      4     4

Proportions correct
   cod salmon  shark
   0.5    1.0    1.0
```

Figure 21: Linear discriminant analysis

If the LDA/QDA contains exactly two predictor variables, one can plot the decision regions using the `Models -> Graphs -> 2D discriminant plot` menu item (see Figure 22).

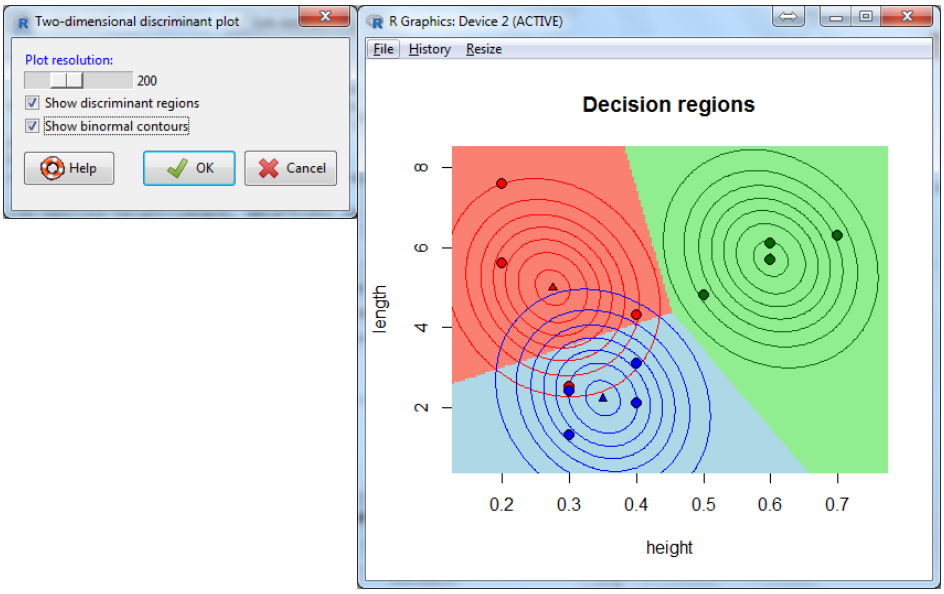


Figure 22: LDA decision regions

4.8 Principal component analysis

A basic tool to reveal structure in multivariate data is found in `Statistics -> Fit models -> Principal component analysis` (see Figure 23).

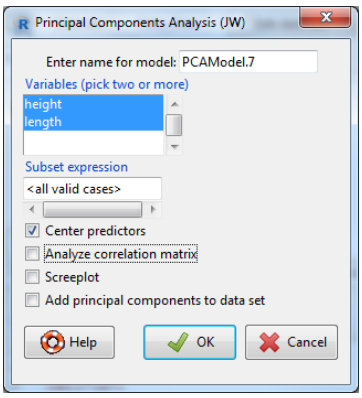


Figure 23: Principal component analysis

```
Component loadings:
              PC1         PC2
height 0.0236499 -0.9997203
length 0.9997203  0.0236499

Component variances:
[1] 4.01647546 0.02405485

Importance of components:
                            PC1      PC2
Standard deviation        2.004 0.15510
Proportion of Variance    0.994 0.00595
Cumulative Proportion     0.994 1.00000
```

When a PCA model has been fitted, one can use the `Models -> Graphs -> PCA/PCR/PLS plots` menu item to plot loadings, scores, biplots and correlation loadings from the model (see Figure 24).

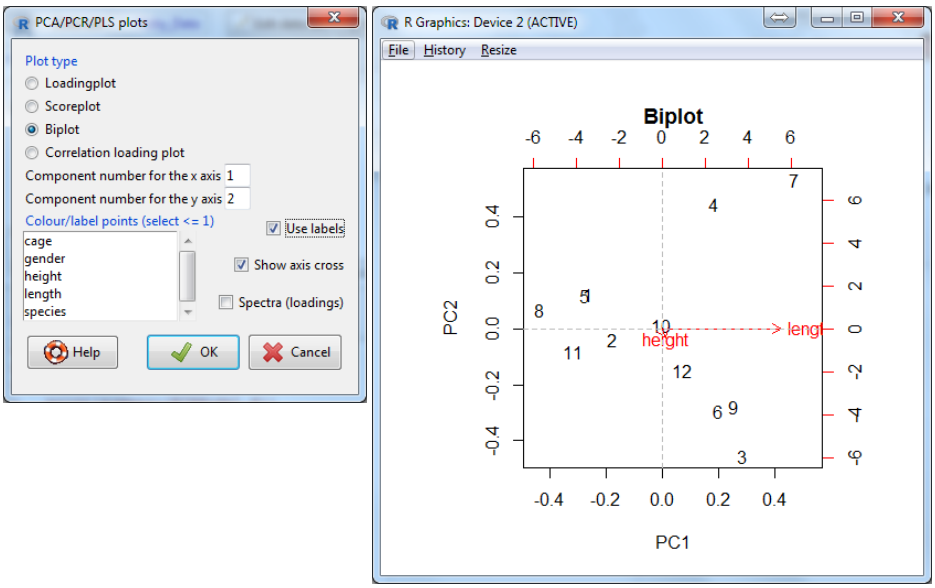
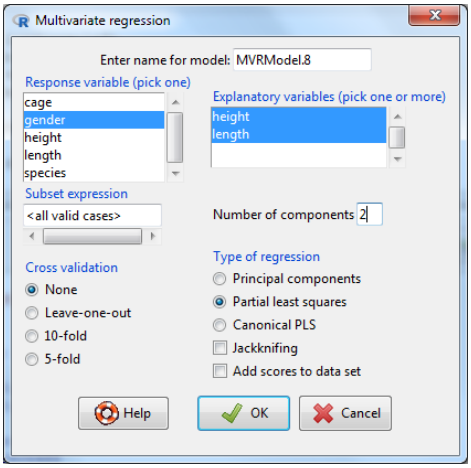


Figure 24: PCA plotting

4.9 Principal component regression and partial least squares

Multiple linear regression can be performed using principal component regression (PCR) or partial least squares (PLS) regression through the menu item `Statistics -> Fit models -> Multivariate regression...` (see Figure 25). This is especially useful when there are more variables than subjects/objects.




```
Data:       X dimension: 12 2
      Y dimension: 12 1
Fit method: kernelpls
Number of components considered: 2
TRAINING: % variance explained
        1 comps   2 comps
X       99.4043  100.0000
gender   0.7569    0.8432
```

Figure 25: Principal component regression and partial least squares

The number of components extracted cannot be higher than the number of objects or variables. When the PCR/PLS model has been fitted, the same plotting tools as in PCA become available (see Figure 24).

5 Appendix

Fictitious data that can be copied to the R Commander to test the described methods

length	gender	height	species	cage
2,5	f	0,3	cod	a
3,1	f	0,4	salmon	a
6,3	f	0,7	shark	a
5,6	m	0,2	cod	a
2,4	m	0,3	salmon	a
5,7	m	0,6	shark	a
7,6	f	0,2	cod	b
1,3	f	0,3	salmon	b
6,1	f	0,6	shark	b
4,3	m	0,4	cod	b
2,1	m	0,4	salmon	b
4,8	m	0,5	shark	b