Sheet 6

Hand in solution by May 16 in the lecture room

1. **Multivariate regression.** Load the vegetation data from the alp in Grisons.

```
t.d <-
read.table("http://stat.ethz.ch/~stahel/courses/multivariate/datasets/vegenv.dat",
           header=TRUE)
```

We will investigate the relationship between the soil chemistry variables `pH10, P10, N10, C10` (input variables) and the abundance of the plant species `Nardstri, Caluvulg, Festrubr` (target variables).

To perform your analysis conveniently, you will need some procedures that can be found on our website. Make them available by typing in

```
source("http://stat.ethz.ch/education/semesters/ss2011/ams/regr.R")
```

The documentation for the function `regr` from this package can be found online:
`http://stat.ethz.ch/~stahel/regression/regr-description.pdf`

a) Fit a multivariate regression model using the function `regr`. Which input variables influence which of the three target variables in a significant way?
Hint: Use `t.r <- regr(cbind(Nardstri,Caluvulg,Festrubr) ~ pH10+P10+N10+C10, data = t.d)` to perform multivariate regression, and `summary(t.r)` to display the summary of the fitted model.

b) Check the model assumptions using the diagnostics produced by
`plot.regr(t.r,plotselect=c(ta=1,qq=1,resmatrix=1,qqmult=1))`.

c) Remove the most prominent outlier, transform the target variables using the square root, and repeat the analysis. Which coefficients now differ significantly from 0? Compute the partial correlation between `Nardstri` and `Caluvulg`, and explain the meaning of this number in the context of this dataset.
Hint: The residuals can be extracted via `residuals(t.r)`.

d) Now repeat the regression using the principal components of the most common plant species as target variables. To this end, restrict the data to the species whose average number of occurrence is more than once per plot (like in exercise 3 of Sheet 3). Apply the square root transform to these data, and perform regression using the first 3 principal components of these variables as target variables and (as in question **a)**) the input variables `pH10`, `P10`, `N10` and `C10`. Display the residual plots and comment on the outcome. Compute the matrix of partial correlations and explain your findings.
Hint: Use `pca.res <- prcomp(...,scale=TRUE)` and `pc <- pca.res$x` to obtain the principal components. Save the first 3 principal components in a data frame,
`pc <- as.data.frame(pc[,1:3])`.
Fit the model using
`regr(cbind(PC1,PC2,PC3)~pH10+P10+N10+C10, data = cbind(pc,t.d))`.

2. Linear mixing. The dataset `voc.dat` contains hourly measurements of volatile organic compounds and other quantities in Wallisellen from January 1996 till February 1997:

tag, monat, jahr, std, min	Date, Time of the measurement
prpa	Propan (ppbC)
nbu	Butan (ppbC)
X2mprpa	2-Methylpropan (ppbC)
npe	Pentan (ppbC)
X2mbu	2-Methylbutan (ppbC)
nhx	Hexan (ppbC)
X2mpe.3mpe	2-Methylpentan+3-Methylpentan (ppbC)
X224mpe	2,2,4-Trimethylpentan (ppbC)
ethe	Ethen (ppbC)
prpe	Propen (ppbC)
ethi	Ethin (ppbC)
be	Benzol (ppbC)
to	Toluol (ppbC)
and other variables	

We restrict the data to July 1996 (`jahr=1996`, `monat=7`), and we like to do linear unmixing via principle components. Read in the data, retain only the volatile organic compound variables and delete all observations with missing values:

```
d.voc <- read.table("ftp://stat.ethz.ch/Teaching/Datasets/WBL/voc.dat",header=TRUE)
m.voc <- na.omit(d.voc[d.voc$monat==7 & d.voc$jahr==96,6:18])
```

We use `prcomp()` to determine the principle components, since `prcomp()` does not center the data automatically (there is the argument `center=TRUE/FALSE` for it). The transformed data can be obtained by using `prcomp(...)$x`.

a) Determine the principle components (using `center=FALSE`) and give biplots of the first few principal components. Which variables have a big influence on the first component?
Hint: Use `biplot(...,choice=c(1,3))` to plot, say, the first and third principal component.

b) Standardize the dataset `m.voc` in the sense that you divide all variables by their respective mean, and repeat the above (again, do not center the data for the PCA). Also provide pairwise scatter plots of the first few principal components.
Hint: To standardize, use `m.vocs <- sweep(m.voc,2,STATS=mean(m.voc),FUN="/")`.

c) In addition, we can also standardize as described in Section 7.3k of the script, dividing each row of `m.vocs` (the standardized data from **b)**) by its row sum. Do this, and provide pairwise scatter plots of the first few principal components (using `center=TRUE` for the PCA). For a linear mixing model with p components, the scores should lie in a p-simplex (with $p+1$ vertices). Is this apparent from the scatter plots?
Hint: To standardize, use `sweep(m.vocs,1,STATS=rowSums(m.vocs),FUN="/")`.