Computer Exercise 4 (R)

1. Read the data in **Weedklr.txt**. Three weed killers were compared for their suitability when growing soybeans (one wants a weed killer that does not harm the beans). The weed killers were tested in a complete randomized design, with 24 plots of beans (8 plots per weed killer treatment). In each plot the number of damaged soybean plants was counted. Your job is to determine if the weed killers differed in how much they harmed the beans. Before going on, make sure that the variable `Weedklr` is made into a factor.

A variable that is a count of the number of "events" occurring in an area (or in an interval of time) often follows a Poisson distribution, and in that case a square root transformation can be helpful to stabilize the variances. To get some feeling for the effect of different transformations, you will be trying out three commonly used ones, ranging from "mild" to "severe". Create three new variables, computed as `sqrt(Damage)`, `log(Damage)`, and `-1/Damage` (the minus sign is just to make the transformed variable increase with Damage, which makes it easier to interpret the results).

Investigate variance homogeneity for the original variable (Damage) and for the three transformations. You can use the `bartlett.test` function for this. Note that the variable Weedklr first has to be transformed into a categorical value.

R console

```
bartlett.test(dat$Damage ~ dat$Weedklr)
```

Alternatively, you can use the more robust `fligner.test`. In connection with choosing a suitable transformation, it is very helpful to look at some graphs. In R you can get a nice visual impression by first fitting a linear model, e.g. for the untransformed variable

R console

```
fm1 <- lm(Damage~Weedklr, data=dat)
```

and then going through the sequence of diagnostic plots you get with the command

```
plot(fm1)
```

(the first plot in the sequence is perhaps the most important one in this case). Go through this also for each of the transformations. Which transformations give acceptable homogeneity of variances? Finally, is there a difference between the weed killer treatments?

A convenient way of finding a suitable transformation is to use the `boxcox` function in the package MASS. Load this package and give the command `boxcox(dat$Damage ~ dat$Weedklr)`. The maximum of the curve in the plotted output gives the best power transformation to use (-1 is inverse, 0 is log, ½ is square root, etc).

As an alternative to transformation, do a Kruskal-Wallis test on Damage (or on one of the transformed variables; it does not matter which). Do you get the same conclusion as before?

2. Read the data in **PeaSec.txt**. The variable PSLen gives the length of pea sections grown in tissue culture. There are various treatments given by a 2% concentration of different sugars. The treatments are coded as 1 = control, 2 = glucose, 3 = fructose, 4 = mixture of glucose and fructose, 5 = sucrose. You should make the treatment variable into a factor with suitable levels, for instance with the command

R console

```
dat$Treatm <- factor(dat$Treatm, levels=c(1, 2, 3, 4, 5),
              labels=c("contr", "gluc", "fruc", "gl+fr", "suc"))
```

assuming you read into the data frame `dat`.

First, check the distributions of PSLen in the different treatment groups, to see if the assumptions for analysis of variance are reasonably fulfilled. If the variances seem heterogeneous, try transforming PSLen to fix this problem. For instance: try square root, log and inverse, and why not try one over PSLen raised to three as well (this is a hint; it may be more convenient to use 1000000/PSLen^3 to avoid very small numbers). To help you in selecting a transformation you can use plots like

R console

```
fm1 <- lm(dat$PSLen ~ dat$Treatm)
```

```
plot(fm1)
```

and the Bartlett test, for instance

R console

```
bartlett.test(dat$PSLen ~ dat$Treatm)
```

or, use the `fligner.test`. You can also look at the distribution of residuals of an analysis of variance, for instance using the `norm.fit` function from last week (you need to read it in using `source("NormFit.R")`)

```
norm.fit(resid(fm1))
```

You can also use the function `plot.groups.sd` (that your teacher has written for you). Read it in with the command `source("PlotGroups.R")` and the use it like

```
plot.groups.sd(dat$PSLen,dat$Treatm)
```

to get a plot similar to the one you may have made in exercise 6 this week. The plot shows you the data points and the means and standard deviations of the treatment groups (there is also a function `plot.groups.se` in the file you read, which will plot the standard errors instead of the standard deviations).

Also, you can use the `boxcox` function to select a suitable transformation.

Second, when you have settled on a transformation for one-way anova, use this to test if there is an effect of sugar treatment section length. From here on, it is assumed that your transformed variable is termed `trPSLen`.

Third, perform planned comparisons, by devising appropriate contrasts. (a) Check if the control treatment differs from the four treatments with sugar added. (b) Check if sucrose (a disaccharide) differs from the other sugars (monosaccharides). (c) Check if mixed monosaccharides differs from pure monosaccharides. (d) Check if glucose differs from fructose. The following contrast matrix has columns corresponding to the four contrasts

```
cntr <- cbind(c(-4,1,1,1,1), c(0,1,1,1,-3),
              c(0,1,1,-2,0), c(0,1,-1,0,0))
```

and you would use it like this

R console:

```
fm <- lm(trPSLen ~ Treatm, contrast=list(Treatm=cntr),
data=dat)

summary(fm)
```

What do you conclude about the various effects? Try to see if you can figure out how to repeat the contrasts using the “emmeans” package. Take a look at the code you used in the previous computer exercise (no. 3).

Finally, compare with what you would have concluded from a post-hoc test.

R console:

Load the multcomp package

```
library(multcomp)

fm <- lm(PSLen ~ Treatm, data=dat)

mc <- glht(fm, linfct = mcp(Treatm = "Tukey"))

summary(mc)
```

also try

```
plot(mc)
```

3. Read the data in **Cation.txt**. The data are membrane potentials (Pot, in mV) for four different cation systems (Cation, coded as 1 = Ca-Li, 2 = Ca-Na, 3 = Ca-K, 4 = Sr-Na) at different electrolyte activity ratios (given as logarithm of activity ratio, LogAR). Make Cation into a factor, perhaps like this

R console
```
dat$Cation <- factor(dat$Cation, levels=c(1, 2, 3, 4),
                labels=c("CaLi", "CaNa", "CaK", "SrNa"))
```

First do a one-way anova on Pot with Cation as factor

R console
```
fm1 <- lm(Pot~Cation, data=dat)
```

Do you detect significant differences in membrane potential between the cation systems?

Then introduce `LogAR` as a covariate, and perform an ancova. Begin by checking for parallelism by fitting the model

R console
```
fm2 <- lm(Pot ~ Cation + LogAR + Cation:LogAR, data=dat)
```

The term `Cation:LogAR` on the right-hand side is the way to express a Cation-LogAR interaction term in R. Such a term can be interpreted as differences in slope between the cation systems. You test the null hypothesis of parallel slopes by testing this term, for instance using

R console
```
Anova(fm2)
```

(you may need the command `library(car)` first). If you can accept that there is a common regression slope, test if there is an effect of Cation on Pot. This you can do by fitting a model

R console
```
fm3 <- lm(Pot ~ Cation + LogAR, data=dat)
```

without the interaction term, and then doing `Anova(fm3)`. What do you find? Also, determine from the output if there is a significant effect of LogAR on Pot, and find the estimated common slope of the regression lines (you find the slope the from `summary(fm3)`). You can also produce a plot with the data points and the parallel regression lines if you first read some code that you teacher wrote

```
source("PlotAncova.R")
```

and then make the plot with the command

```
plot.ancova.com(dat$Pot, dat$LogAR, dat$Cation)
```

(there is also a function `plot.ancova.sep` that makes a plot with separate slopes for each group).

In this case the mean log activity ratio is not the same for the different cation systems. In order to get a meaningful comparison between the different systems, one can compute the so-called adjusted means. The adjusted mean Pot for a given Cation is the value indicated by the regression line at the overall average LogAR (usually the unweighted mean over groups); in this way one compares the different cation systems at the same log activity ratio. Find the adjusted means of the four levels of Cation. First compute the group means of LogAR, which you can do with the tapply function. Tapply is a very useful function with many functions that are built in, but you could also define your own function and use that. See `?tapply` for more.

```
group.ave <- tapply(dat$LogAR, dat$Cation, mean)
```

and then get the overall mean LogAR as

```
ave <- mean(group.ave)
```

You can now get the adjusted means as the predicted Pot for the different cation systems, at the goup average LogAR, using the predict function in R

```
cation <- levels(dat$Cation)

logar <- rep(ave, 4)

pred.dat <- data.frame(Cation=cation, LogAR=logar)

predict(fm3, pred.dat)
```

The predict function gives the predicted values of the response of the fitted model fm3 corresponding to each of the "*x*-values" in the data frame pred.dat. To see what it is about, it might help to look at the data in pred.dat (use the command `fix(pred.dat)`).

4. Read the data in **Protand.txt**. The file contains average development times (days until adult eclosion; TAdult) for female and male common blue butterflies (*Polyommatus icarus*) raised in environmental cabinets with different photoperiods (hours daylight: Light).

For many insects in seasonal environments, males tend to develop somewhat faster than females, and this phenomenon is called protandry. First look for protandry in the data by doing a one-way anova on TAdult with Sex as factor.

R console

```
fm1 <- lm(TAdult ~ Sex, data=dat)

Anova(fm1)
```

What do you find?

Photoperiod can affect development time (this occurs for insects in seasonal environments: they are in a "greater hurry" to grow up later in the season). Try using Light as a covariate, fitting the model

R console

```
fm2 <- lm(TAdult ~ Sex + Light, data=dat)
```

if you think this model with a common slope is OK. Do you find any effect of Sex on TAdult? Is the effect an example of protandry? How big is the effect? (Although one should look at the adjusted means to answer the last question, it does not matter in this case, since the mean of Light is the same for females and males). You might also look at a plot of the data and fitted model (perhaps using the function `plot.ancova.com` you used in the last exercise)

5. Read the data in **Allometr.txt**, containing the sex (1=male, 2=female), total adult dry weight, and abdomen dry weight (in mg) of 53 *P. napi* (green-veined white) butterflies. Your job is to determine if females tend to have bigger abdomens than males. While doing this, you will also perform some model criticism.

First, make `Sex` into a factor and then log transform the dry weights (this is not really necessary for the analysis, but do it anyway). You need to read in the car package to use the Anova() function and get type II sums of squares. Check if there is an effect of sex on log abdomen weight, using a one-way analysis of variance (or *t*-test), for instance

R console

```
library(car)

fm1 <- lm(logAbdW ~ Sex, data=dat)

Anova(fm1)
```

Then see if your result could simply be due to sexual size dimorphism (since bigger butterflies tend to have bigger abdomens), by checking for an effect of sex on log total weight, for instance

R console

```
fm2 <- lm(logTotW ~ Sex, data=dat)

Anova(fm2)
```

Next, control for log total body weight, by performing an ancova. Before accepting the conclusion from this analysis, check if the female and male regression slopes are parallel

R console

```
fm3 <- lm(logAbdW ~ Sex + logTotW + Sex:logTotW, data=dat)
```

```
Anova(fm3)
```

Next, extend your model criticism by doing

```
plot(fm3)
```

and looking at the plots. Do you detect any outliers and/or particularly influential points? (Hint: case number 35 in the data set is an outlier and, from the Cook's distances plot, case 35 has a large influence on the model fitting).

Next, try to explain the reason for this outlier, which is a male. Plot males and females in the same graph, but with different symbols for the sexes. For instance, load the code in the file **PlotAncova.R**, with the command `source("PlotAncova.R")` and do (assuming you have added the log-transformed variables to the data frame)

```
plot.ancova.sep(dat$logAbdW, dat$logTotW, dat$Sex)
```

Could it be that the outlier male really is a female?

Next, change the (possibly) wrongly sexed butterfly to a female, (for instance using the command `fix(dat)`, or in the text file directly), and repeat the ancova. Do you get parallel regression lines this time? If so, perform the ancova

R console

```
fm4 <- lm(logAbdW ~ Sex + logTotW, data=dat)
```

and check the results from it. Do females have (relatively) bigger abdomens?