Solutions to Computer Exercise 3 (R)

1. The estimated regression of logC3 on cLAT and cLONG is

```
logC3 <- -0.5456 + 0.0424*cLAT - 0.0037*cLONG
```

The regression coefficients are the same as in the lecture, but the intercept has changed to a biologically meaningful value (because the zero-points of the x-values are well within the set of observed data points).

We get $t = 7.833$, $p < 0.001$ for cLAT and $t = -0.837$, $p = 0.405$ for cLONG. The sums of squares are 3.6112 for cLAT, 0.0412 for cLONG, and 4.1200 for the residual. In the regression of logC3 on cLONG, the residual sum of squares is 7.7311, which is greater than the residual sum of squares for the multiple regression. In fact 7.7311 - 4.1200 = 3.6111, so the sum of squares for cLAT in the multiple regression is the improvement in fit, measured as residual sum of squares, when cLAT is added to the model with only cLONG (note that there is slight rounding error in the calculation above). Similarly, in the regression of logC3 on cLAT, the residual sum of squares is 4.1612, and the difference between the two residual sums of squares is 4.1612 - 4.1200 = 0.0412.

The two simple regressions are

```
logC3 <- -0.5456 + 0.0420*cLAT
```

```
logC3 <- -0.5456 - 0.0004*cLONG
```

The regression coefficients are not the same as in the multiple regression, but they are rather similar. This is because there is some covariation between cLAT and cLONG in the data set, but this covariation is not terribly strong (check this by looking at a cLAT vs. cLONG scatter plot). Note finally that the intercepts are the same for all three regressions above: this is a consequence of using centered x-variables (the intercept is in fact equal to the mean of logC3).

2. The estimated regression of volume on diameter is

Volume = -36.94 + 5.066*Diam

with a *p*-value for Diam of less than 0.000001. The proportion of variation explained by the regression is $R^2 = 0.935$, which seems pretty good. Regressing volume on both diameter and height gives

Volume = -57.99 + 4.708*Diam + 0.3393*Height

with a *p*-value for Diam of less than 0.000001 and a *p*-value for Height of 0.0145. Thus, there is a significant additional effect of Height, but since $R^2 = 0.948$ is only slightly bigger than before, one actually gains rather little predictive power by also measuring height. The reason that the regression coefficient for Diam changes when Height is included is that Diam and Height are (positively) correlated with each other. This is also seen in the tolerance values for Diam and Height, which are both equal to 0.730, clearly smaller than one. However, the tolerance is not close to zero, so there is substantial independent variation in Diam and Height.

Log-transforming all the measurements, one gets instead

$$\text{logVolume} = -2.353 + 2.200*\text{logDiam}$$

with $p < 0.001$ for logDiam, $R^2 = 0.954$ and

$$\text{logVolume} = -6.632 + 1.983*\text{logDiam} + 1.117*\text{logHeight}$$

with $p < 0.001$ for logDiam, $p = 0.000008$ for logHeight and $R^2 = 0.978$. There is thus a highly significant additional effect of logHeight, but again the increase in the proportion of explained variation is rather small. Anyway, it is slightly better to use the regression for the log-transformed variables.

3. The anova gives F=6.0995 and p=0.015 for the comparison between sexes and the t-test from last week gave t=2.4697 with the same p-value (note that you must set variances as equal).

4. WingLen1: F=1.321 and p=0.281. WingLen2: F=5.231 and p=0.001. Works fine!

5. For males we get F=6.050 and p=0.0009 and for females F=30.98 and p<0.000001. Thus, there is an effect of host plant and it seems stronger for females. From the planned high vs. low quality comparisons for males, there is a significant difference only for *A. petiolata* (this is the first contrast, called Foodpl1 by R) but not for *B. incana* (the Foodpl2 contrast), whereas for females both plants produce a significant response to quality. The response to host-plant is however significant for both sexes.

6. The anova gives F=23.00 and p=0.00002, so there are differences between melon varieties. From the Tukey test, it seems that the second variety differs from the others and that this is the only (post hoc) significant difference. The pairwise comparisons without any adjustment give additional significant differences, but such a test should only be used for planned comparisons, i.e. comparisons decided on in advance of the test. The sequential Bonferroni adjustment gives the same conclusions as the Tukey test, i.e. that variety 2 differs from each of the others, but we cannot reject that varieties 1, 3 and 4 have the same yield.