Practical 5
R – linear models, contrasts.

1.

- Open the *R*. If you have, load the Workspace (.RData-fail) saved in last week (*Load Workspace* ...) andrun the package *Rcmdr* (command `library(Rcmdr)`).

  Fix the dataset 'students' as the default dataset by pressing the *Data set*: <*No active dataset*> button: 
;

- If you haven't the workspace, what to load), import the dataset using menus *Data* -> *Import data* -> … (follow the guide from last practical)

  or run the following command in script window:

```
students = read.csv("http://ph.emu.ee/~ktanel/DK_0007/studentsR_eng.csv", header=TRUE,
                                                                  sep=";", dec=",")
```

- As an alternative you may save the students dataset as an *Excel* fail from the course internet page and import it into the *R Commander* (*Data -> Import data -> from Excel, Access or dBase data set…*).

2.

Irrespective to the analyses made in last week's practical try to predict the students' head girth (head circuit?, head line? head circumference? 'peaümbermõõt' in Estonian). And the goal should be to get so good model as possible.

2.1.

As you remember, the head circuit was more strongly correlated with weight than with height (if you don't remember, perform the correlation analysis).

So, the first task should be to predict students' head circuit based on the weight.

a) As the regression equation is also the linear model, the function `lm` (linear model) can be used in the form (more about model building in R look at the next page):

```
peaymb_GLM.1 <- lm(peaymb ~ kaal, data=students)
summary(peaymb_GLM.1)
```

The sign '<-' means assign and can be replaced with '=';

`peaymb_GLM.1` is the model name and command `summary` prints out basic statistics concerning the model.

If you don't want to save the model for further analyses (prediction, residuals' analysis, …), the command without assign the modeling results to some variable can be used:

```
summary(lm(peaymb ~ kaal, data=students))
```

Remarks about model building in R. The general rules and operators used in model construction in *R* are following.

The ~ operator is basic in the formation of models in *R*. An expression of the form `y ~ model` is interpreted as a specification that the response `y` is modelled by a linear predictor specified symbolically by `model`. Such a model consists of a series of terms separated by + operators. The terms themselves consist of variable and factor names separated by `:` operators. Such a term is interpreted as the interaction of all the variables and factors appearing in the term.

In addition to + and `:`, a number of other operators are useful in model formulae.

The `*` operator denotes factor crossing: `a*b` interpreted as `a+b+a:b`.

The `^` operator indicates crossing to the specified degree. For example `(a+b+c)^2` is identical to `(a+b+c)*(a+b+c)` which in turn expands to a formula containing the main effects for `a`, `b` and `c` together with their second-order interactions.

The `%in%` operator indicates that the terms on its left are nested within those on the right. For example `a + b%in%a` expands to the formula `a + a:b`.

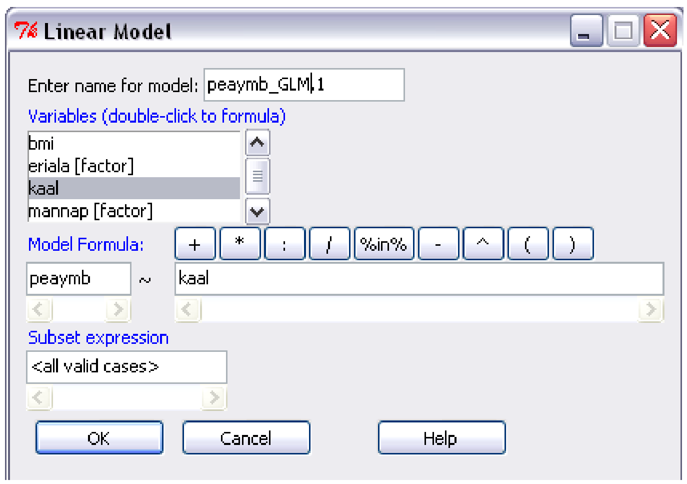
The `-` operator removes the specified terms, so that `(a+b+c)^2 - a:b` is identical to `a + b + c + b:c + a:c`. It can be used also to remove the intercept term: `y ~ x - 1` is a line through the origin. A model with no intercept can be also specified as `y ~ x + 0` or `y ~ 0 + x`.

While formulae usually involve just variable and factor names, they can also involve arithmetic expressions. The formula `log(y) ~ a + log(x)` is quite legal. When such arithmetic expressions involve operators which are also used symbolically in model formulae, there can be confusion between arithmetic and symbolic operator use. To avoid this confusion, the function `I()` can be used to bracket those portions of a model formula where the operators are used in their arithmetic sense. For example, in the formula `y ~ a + I(b+c)`, the term `b+c` is to be interpreted as the sum of `b` and `c`.

----

b) The same analysis with *R Commander*:

Statistics -> Fit models -> Linear model …



Result:

```
> peaymb_GLM.1 <- lm(peaymb ~ kaal, data=students)

> summary(peaymb_GLM.1)

Call:
lm(formula = peaymb ~ kaal, data = students)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4109 -1.4582  0.2312  1.8826  9.4470

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.1738     1.8000   25.10  < 2e-16 ***
kaal          0.1632     0.0277    5.89 6.57e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.907 on 90 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.2782, Adjusted R-squared: 0.2702
F-statistic: 34.69 on 1 and 90 DF,  p-value: 6.573e-08
```

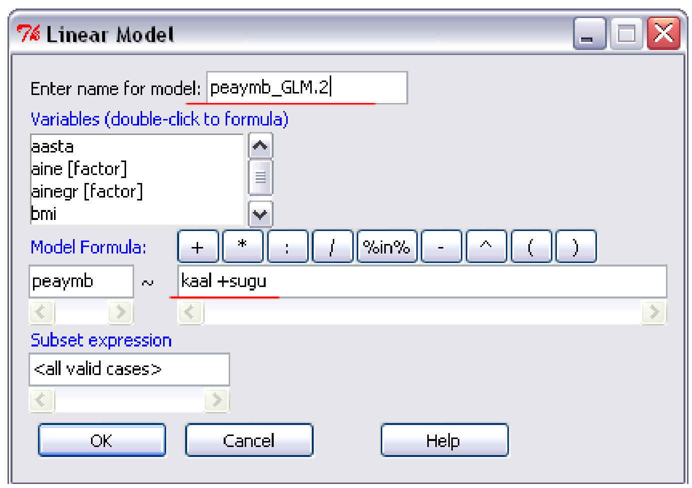
Head circ. = 45,1738 + 0,1632×Weight

The weight effect is statistically significant

2.2.

- Is this possible to get more precise prediction considering also the sex?

In *R Commander*:
Statistics -> Fit models -> Linear model …



Commands:

```
> peaymb_GLM.2 <- lm(peaymb ~ kaal +sugu, data=students)

> summary(peaymb_GLM.2)
```

Result:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 45.55234    2.59115  17.580  < 2e-16 ***
kaal         0.15945    0.03325   4.795 6.49e-06 ***
sugu[T.N]   -0.18006    0.88216  -0.204    0.839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.923 on 89 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.2786, Adjusted R-squared: 0.2624
F-statistic: 17.18 on 2 and 89 DF,  p-value: 4.893e-07
```

Head circ. | Sex=N
= 45,55 – 0,18 + 0,159×Weight
and
Head circ. | Sex=M
= 45,55 + 0 + 0,159×Weight

By default the *R* takes the effect of factor's first level equal to 0 (as a base) – at present the effect of sex 'M' is considered as a base.

The difference between men and women is 0.18 cm: `sugu[T.N]    -0.18006`, but this difference is not statistically significant (p = 0,839). This means that the sex effect is not statistically significant. Also the R^2 was not changed compared with the model without sex.

- In spite of that let's try to include into the model the sex and weight interaction (Why? I don't know. Quite often the modeling is just playing and controlling of different ideas …. Ok, I was calculating the correlation coefficients between weight and head circuit by sex and found these be different – look at the figures in exercise 4.3 of last week's practical … :).

```
peaymb_GLM.3 = lm(peaymb ~ kaal + sugu + kaal:sugu, data=students)
summary(peaymb_GLM.3)
```

The same model is fitted according to the command using the operator `*`:

```
lm(peaymb ~ kaal*sugu, data=students)
```

Result:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     34.96314    6.06303   5.767 1.18e-07 ***
kaal             0.29989    0.07996   3.751 0.000315 ***
sugu[T.N]       12.13460    6.45439   1.880 0.063409 .  
kaal:sugu[T.N]  -0.16877    0.08765  -1.925 0.057398 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.879 on 88 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.3077, Adjusted R-squared: 0.2841
F-statistic: 13.04 on 3 and 88 DF,  p-value: 4.021e-07
```

The sex effect is still not statistically significant (p = 0,063) as also the sex*weight-interaction (p = 0,057), but both these p-values are on the limit and also the R^2 increased by some percent – so I prefer the last model.

The women' and men' head circuits are predictable by formulas:

Head circ. | Sex="N" = 34,96 + 12,13 + (0,300 – 0,169)×Weight = 47,09 + 0,131×Weight

Head circ. | Sex="M" = 34,96 + 0 + (0,300 + 0)×Weight = 34,96 + 0,300×Weight

The p-value in the last row of output ($p = 4{,}02\times10^{-7}$) says, that the constructed model is statistically significant.

- Remark.

As the last model estimates different regression coefficients for men and women, are the same effects estimable also from the model without the weight's main effect:

```
summary(lm(peaymb ~ sugu + kaal:sugu, data=students))
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 34.96314    6.06303   5.767 1.18e-07 ***
sugu[T.N]   12.13460    6.45439   1.880 0.063409 .  
suguM:kaal   0.29989    0.07996   3.751 0.000315 ***
suguN:kaal   0.13112    0.03591   3.651 0.000443 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.879 on 88 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared: 0.3077, Adjusted R-squared: 0.2841
F-statistic: 13.04 on 3 and 88 DF,  p-value: 4.021e-07
```

To be convinced that the equations to predict the men's and women's head circuits are identical with those got before, write down the corresponding equations based on the parameters' estimates from the new analysis.

2.3.

Quite often it is not enough to prefer some model based only on descriptive statistics (like R^2, for example). If the comparable models are hierarchical, it is possible to test the hypothesis about advantage of more complex model. In *R* the corresponding test can be performed with function `anova`.

a) For example, if you have two models

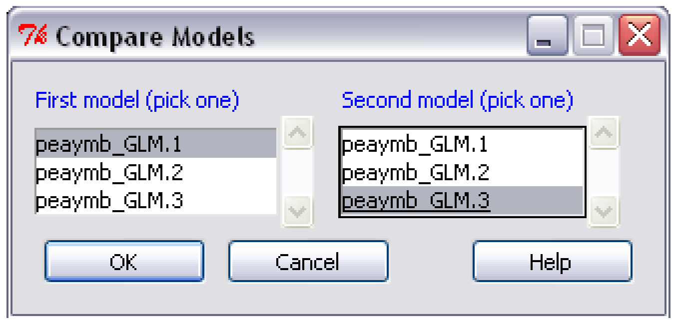
```
peaymb_GLM.1 <- lm(peaymb ~ kaal, data=students)
peaymb_GLM.3 <- lm(peaymb ~ kaal + sugu + sugu:kaal, data=students)
```

you can compare them with command

```
anova(peaymb_GLM.1,peaymb_GLM.3)
```

b) You can also order the same test from *R Commander* menus:

Models -> *Hypothesis tests* -> *Compare two models …*



↓

```
> anova(peaymb_GLM.1,peaymb_GLM.3)
Analysis of Variance Table

Model 1: peaymb ~ kaal
Model 2: peaymb ~ kaal + sugu + kaal * sugu
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     90 760.66
2     88 729.57  2     31.09 1.8752 0.1594
```

Conclusion: more complex model is not statistically significantly better (p = 0,159).
At the same time, suppressing the potentially interesting fact that the relationship between weight and head circuit depends on sex only due to the p-value bigger than 0.05 is in my opinion also not right …

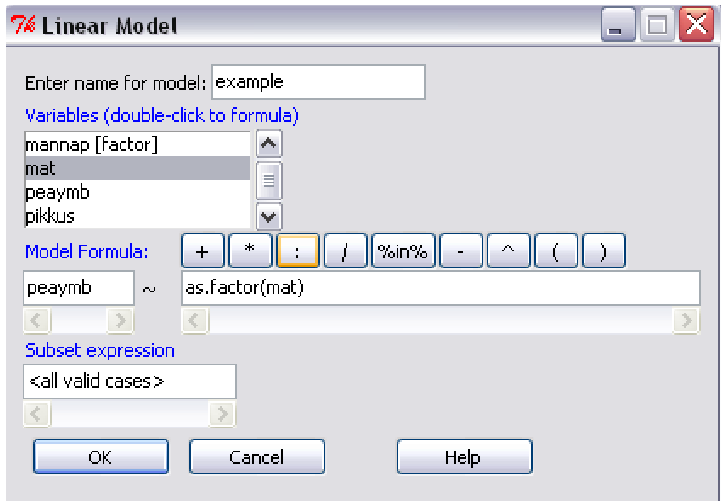
2.4.

Futher you can study, is the head circuit related with the study specification or math grade.

- Analysing the effect of math grade it’s important to ask the *R* to consider the numerical trait ’mat’ as discrete factor and not as the continuous numeric argument of regression analyses (the last is the default option for numerical arguments in *R*). The simpliest variant is to add into the model instead of the term `mat` the function `as.factor(mat)`:

```
summary(lm(peaymb ~ as.factor(mat), data=students))
```

The same in *R Commander*:
Statistics -> Fit models -> Linear model …



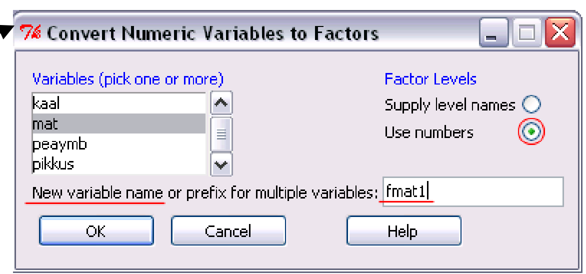
- As an alternative you may add into the dataset ’*students*’ new trait which is already formatted as discrete factor and use in modelling this new variable.

- One possibility to make the new factor variable ’fmat1’ with the same numerical grades is to apply the function `as.factor` in script window:

```
students$fmat1 = as.factor(students$mat)
```

or run the same command in *R Commander* menus

Data -> Manage variables in active data set
-> Convert numeric variables to factors…



The new variable ’fmat1’ has the same numerical values ’3’, ’4’ and ’5’, but it is already considered as factor.

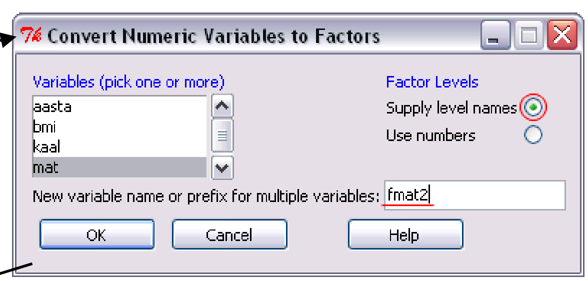
……………………………………………………………….

Applying the following command in script window

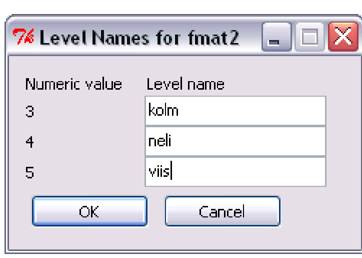
```
students$fmat2 = factor(students$mat, labels=c('kolm','neli','viis'))
```

or ordering from *R Commander* menus

Data -> Manage variables in active data set
-> Convert numeric variables to factors…



it is possible to create the new factor variable with nonnumeric values (the name of the new variable is ’fmat2’ and it’s values are ’three’, ’four’ and ’five’ in Estonian).



3.

Dataset: http://ph.emu.ee/~ktanel/DK_0007/kala.xls

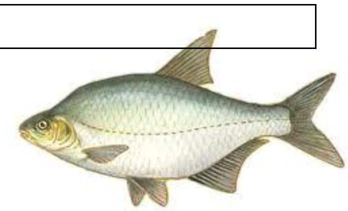
The following data about Estonian fishes [‘fish’ = ‘kala’ in Estonian] are part of Mariann Nõlvak master thesis;

– 5 fishing places (‘Võrtsjärv’, ‘Kärevere’, ‘Kastre’, ‘Praaga’ and ‘Peipsi järv’), years 2004-2006;
– 6 species (in Estonian: haug [= ‘pike’ in English], särg [roach], latikas [bream], luts [burbot], ahven [perch] and koha[pikeperch]);
– the length and weight of fishes is measured, sex (‘e’ – female, ‘i’ – male) and infestation with the larvae of broad tapeworm *Diphyllobothrium latum* is determined;
– also the fishing season (kevad-suvi [spring-summer] and sügis-talv [autumn-winter]) is registered.

Import the dataset into *R Commander* and

fix the imported dataset ‘kala’ as the default dataset: 

3.1. Mudeli parameetrite hindamine



How depends the weight of breams [latikas] from fishing place and sex?

Let’s model the weight of breams with following two-factorial model:

$$y_{ijk} = \mu + K_i + S_j + \varepsilon_{ijk},$$

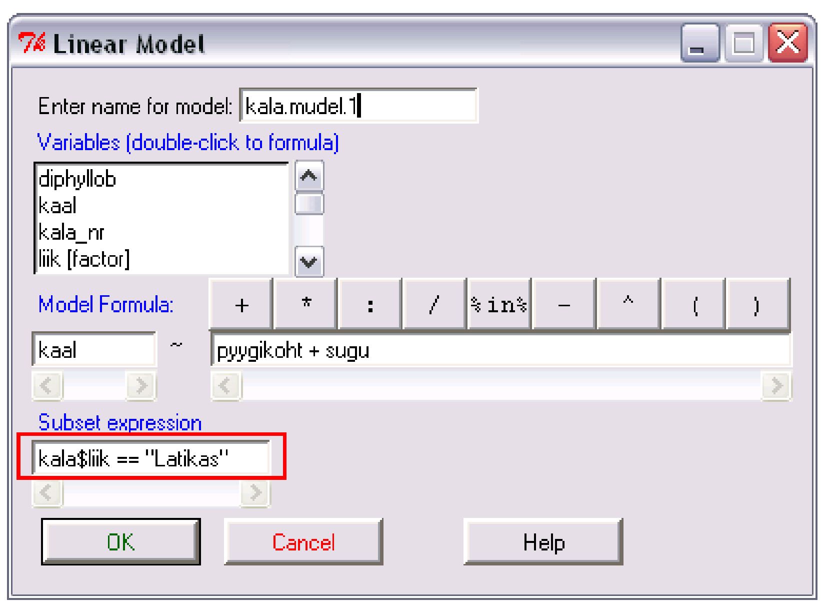
where y_{ijk} is the weight of k^{th} fish cached from place i and having sex j, K_i is the effect of place i (i=1,…,5) and S_j is the effect of sex j (j=1,2).

a) The model is implemented with command

```
kala.mudel.1 = lm(kaal ~ pyygikoht + sugu, data=kala, subset=kala$liik=="Latikas")
summary(kala.mudel.1)
```

b) or in *R Commander*

Statistics -> Fit models -> Linear model …



Excerpt from results:

```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            1020.06      32.34  31.544  < 2e-16 ***
pyygikoht[T.Kärevere]   -95.63      39.88  -2.398   0.0176 *
pyygikoht[T.Peipsi]      76.94      40.22   1.913   0.0575 .
pyygikoht[T.Praaga]     -63.66      42.27  -1.506   0.1340
pyygikoht[T.Võrtsjärv] -372.63      39.01  -9.551  < 2e-16 ***
sugu[T.i]              -113.36      28.14  -4.028 8.52e-05 ***
```

Intercept 1020.1 shows the estimate of average weight of female breams caught in Kastre (by default *R* equates the effects of first levels of all factors with 0, in alphabetic order the first place is 'Kastre' and the first sex is 'e'), the standard deviation of the estimate is 32,3 g.

Other estimates measure the average differences from female breams caught in Kastre and *p*-values are showing the statistical significance of these differences.

For example, the average weight of female breams caught in Kärevere is estimable as 1020,1 – 95,6 = 924,5 g and it differs significantly from the average weight of females caught in Kastre (*p* = 0,0176).

3.2. Kontrastide konstrueerimine

Also we can estimate the average weight of male breams in Võrtsjärv:

1020,1 – 372,6 – 113,4 = 534,1 g.

But to test the difference from female Kastre breams, the **contrast** (= uniquely estimable linear combination of model parameters) should be constructed and the difference from 0 must be tested.

a) This can be done with command

```
linear.hypothesis(kala.mudel.1, c(0,0,0,0,1,1), c(0))
```

*) The first argument of function `linear.hypothesis` determines the model name (`kala.mudel.1`) based on which the contrast is constructed,

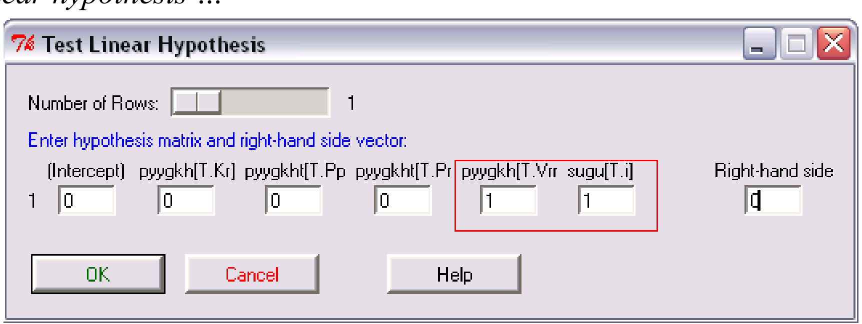
*) the second argument defines the vector of weights assigned to the model non-null parameters (*R* omits the parameters which are equated to zero to guarantee the unique estimates), the factors are ordered as in corresponding modelling command and the factors' levels are in alphabetic order,

*) the third argument defines the contrast value at null hypothesis.

b) The same with *R Commander* menus

(if necessary you should fix the right model for *R Commander* menu commands: Model: kala.mudel.1)

Models -> Hypothesis tests -> Linear hypothesis …



Result:

```
Hypothesis:
pyygikoht[T.Võrtsjärv] + sugu[T.i] = 0

Model 1: kaal ~ pyygikoht + sugu
Model 2: restricted model

  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    167 4600679
2    168 8085728  -1  -3485049 126.50 < 2.2e-16 ***
```

Difference is statistically significant (*p* < 0,001).

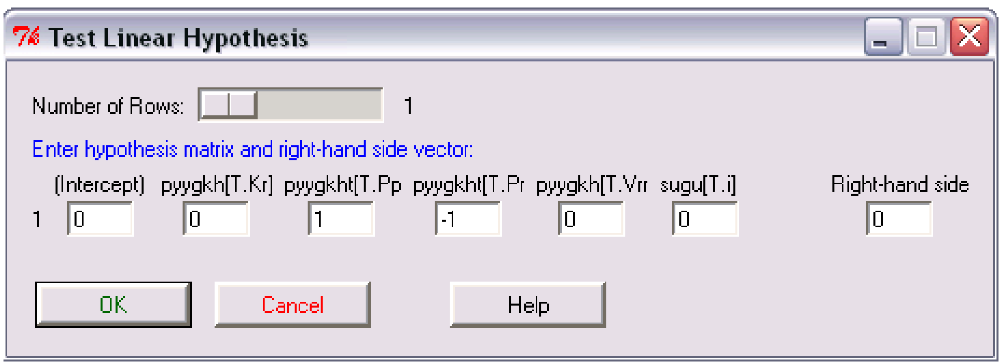
3.3.

But are the Peipsi and Praaga breams significantly different?

Solution:

```
linear.hypothesis(kala.mudel.1, c(0,0,1,-1,0,0), c(0))
```

or



Result:

```
Hypothesis:
pyygikoht[T.Peipsi] - pyygikoht[T.Praaga] = 0

Model 1: kaal ~ pyygikoht + sugu
Model 2: restricted model

  Res.Df     RSS Df Sum of Sq      F   Pr(>F)
1    167 4600679
2    168 4892439  -1   -291760 10.591 0.001376 **
```

Answer: they differ significantly (*p* = 0,0014).

3.4. Estimation of mean values and their confidence intervals

95%-confidence intervals for average weights of Peipsi and Praaga male breams can be found with commands

```
predict(kala.mudel.1, data.frame(pyygikoht="Peipsi", sugu="i"), interval="confidence")
predict(kala.mudel.1, data.frame(pyygikoht="Praaga", sugu="i"), interval="confidence")
```

```
          fit      lwr      upr
[1,] 983.6416 918.6315 1048.652
```

```
          fit     lwr      upr
[1,] 843.0402 779.111 906.9695
```

So, the average weights of Peipsi and Praaga male breams are with 95%-probability in intervals 918,6…1048,7 g and 779,1…907,0 g, correspondingly.

3.5. Testing the statistical significance of factors' effects

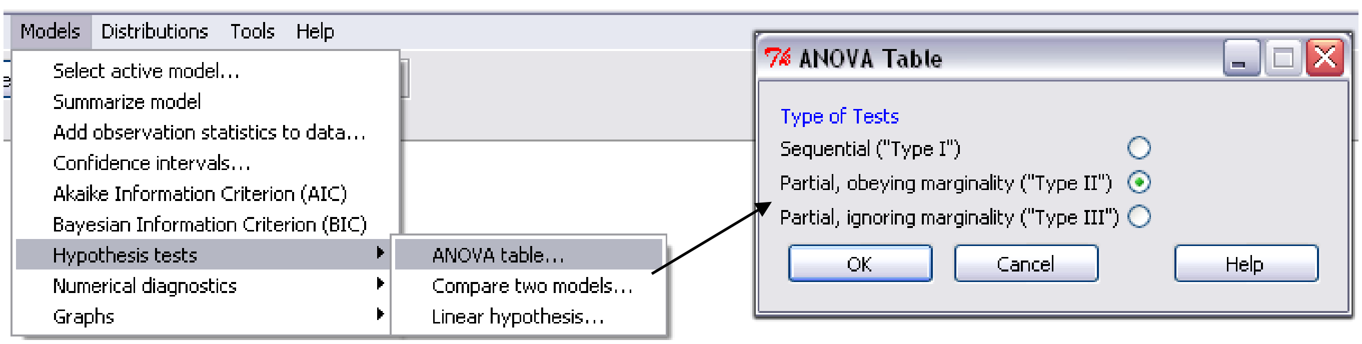
Are the effects of fishing place and sex statistically significant?

Hypothesis about the factors' statistical signifacance can be tested with command

```
Anova(kala.mudel.1)
```

or in *R Commander* menus: *Models -> Hypothesis tests -> ANOVA table*

(if necessary fix the lastly fitted model as the default model for menu commands: 
)



Results:

```
Anova Table (Type II tests)

Response: kaal
           Sum Sq  Df F value    Pr(>F)
pyygikoht 4105949   4  37.260 < 2.2e-16 ***
sugu       447008   1  16.226  8.52e-05 ***
Residuals 4600679 167
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Both the effects of place and sex are statistically significant ($p<0,05$).

4.

4.1. Depends the weight of breams additionally on the season (kevad-suvi [spring-summer] and sügis-talv [autumn-winter])?

Lets add the season effect L_k (k=1,2) into the model:

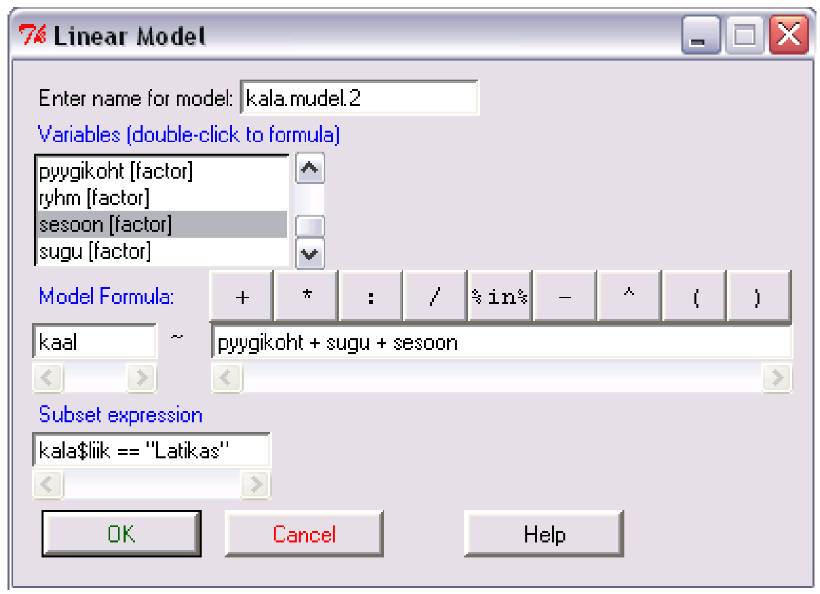
$$y_{ijkl} = \mu + K_i + S_j + L_k + \varepsilon_{ijkl}\,.$$

a) The corresponding command in script window can be used:

```
kala.mudel.2 = lm(kaal~pyygikoht+sugu+sesoon, data=kala, subset=kala$liik=="Latikas")
summary(kala.mudel.2)
```

b) Or with *R Commander*:

Statistics -> Fit models -> Linear model …



Excerpt from results:

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               1028.706     30.808  33.391  < 2e-16 ***
pyygikoht[T.Kärevere]      -52.639     39.190  -1.343   0.1811
pyygikoht[T.Peipsi]        116.895     39.332   2.972   0.0034 **
pyygikoht[T.Praaga]         -4.334     42.459  -0.102   0.9188
pyygikoht[T.Võrtsjärv]    -322.351     38.864  -8.294 3.57e-14 ***
sugu[T.i]                 -117.052     26.768  -4.373 2.16e-05 ***
sesoon[T.sygis-talv]      -114.337     26.388  -4.333 2.54e-05 ***
```

Breams caught on autumn-winter weight on an average 114,3 g less than breams caught on spring-summer period and this difference is statistically significant (p = 2,54e-05 < 0,001).

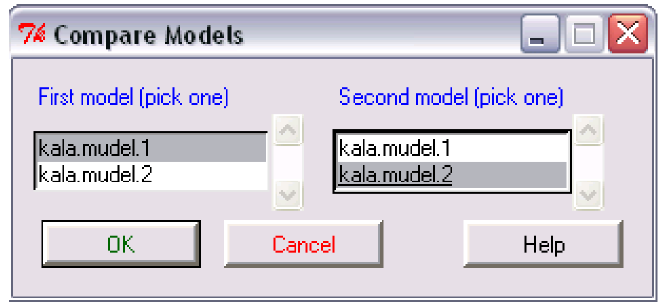
4.2. Will the new model fit the weight of breams better?

Solution:

```
anova(kala.mudel.1,kala.mudel.2)
```

or

Models -> Hypothesis tests -> Compare two models…



Result:

```
Analysis of Variance Table

Model 1: kaal ~ pyygikoht + sugu
Model 2: kaal ~ pyygikoht + sugu + sesoon
  Res.Df     RSS  Df Sum of Sq      F    Pr(>F)
1    167 4600679
2    166 4133230   1    467449 18.774 2.542e-05 ***
```

Yes, the new model is statistically significantly better ($p < 0{,}001$).

4.3.

Lets study additionally, is the average weight of breams caught upstream from Tartu (Kärevere and Võrtsjärv) significantly different from average weight of breams caught downstream from Tartu (Peipsi, Praaga ja Kastre).

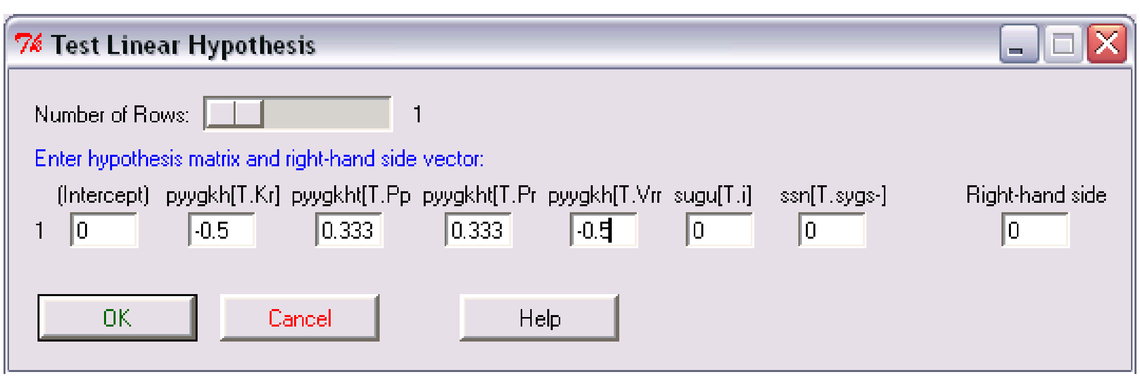
Based on the modeling results presented in last page the average effect of fishing places upstream from Tartu is (–52,6 – 322,4) / 2 = –187,5 g.
And the average effect of fishing places downstream from Tartu is (0 – 4,3 + 116,9) / 3 = 37,5 g.

To test the difference of calculated effects (this is equivalent to testing the difference of average weights) the difference of corresponding contrast from 0 must be tested:

```
linear.hypothesis(kala.mudel.2, c(0,-0.5,0.333,0.333,-0.5,0,0), c(0))
```

or

Models -> Hypothesis tests -> Linear hypothesis…



Result:

```
Hypothesis:
-0.5 pyygikoht[T.Kärevere] + 0.333 pyygikoht[T.Peipsi] + 0.333 pyygikoht[T.Praaga] - 0.5 pyygikoht[T.Võrtsjärv] = 0

Model 1: kaal ~ pyygikoht + sugu + sesoon
Model 2: restricted model

  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1    166 4133230
2    167 6125026 -1  -1991796 79.995 7.113e-16 ***
```

The difference is statistically significant ($p < 0{,}001$).

4.4.

To get the quick overview about the effects of factors included in the model, it is convenient to use the following *R Commander* command:

Models -> Graphs -> Effect plots