

## Praktikum 5


### R – lineaarsed mudelid, kontrastid.

Tänane ja ka järgmine praktikumi püüavad anda pisikese ülevaate statistilisest modelleerimisest lineaarsete mudelite abil. Üldised lineaarsed mudelid (*GLM, general linear models*) kujutavad enesest eksperimendipõhises statistikas kasutatavate traditsiooniliste meetodite nagu t-test, regressioon- ja dispersioonanalüüs edasiarendust, mistap tähendab ka nimetatud meetodite rakendamine tegelikult statistilist modelleerimist. Aga erinevalt tavastatistikast, kus konstrueeritud mudelid ja kontrollitavad hüpoteesid on enamasti selgelt paigas, kaasneb üldisemate mudelite rakendamisega sageli palju mängimist ja proovimist ning lõpptulemus võib tegelike andmetes peituvate seoste kõrval sõltuda märksa enam uurija enese ettekujutusest võimalikest seostest ja struktuuridest oma andmeis ning oskusest seda kõike matemaatilise mudelina esitada. Seetõttu ei ole üldiste lineaarsete mudelite rakendamisel tihti esmatähtis mingite mõjude või seoste statistilise olulisuse testimine (kuigi enamasti, aga mitte alati, jõutakse ka selleni, aga siis ka pigem mõttega, et kas on alust mingi faktori tasemeid või ühte mudelit teisest erinevaks lugeda), vaid hoopis mudeli parameetrite (erinevate faktorite mõjude suuruse) hindamine ja olemuse tõlgendamine.

Kuigi selle kursuse raames ei jõua mudelitest väga palju rääkida, võite huvi korral silmad üle lasta järgmisest loengukonspektist: [http://www.eau.ee/~ktanel/VL\\_0192/pt6\\_2010.pdf](http://www.eau.ee/~ktanel/VL_0192/pt6_2010.pdf) (mis on küll osa loomade aretusväärtuste hindamise kursusest ja seeläbi kohati pisut spetsiifiline, hõlmates siiski küllalt suure osa üldiste lineaarsete mudelite olemusest).

#### 1.

- Avage R, seejärel möödunud praktikumi lõpus salvestatud .RData-fail (*Load Workspace ...*) – kui on, mida avada –, ja käivitage lisamoodul *Rcmdr* (näiteks käsuga `library(Rcmdr)`).

Kui teil oli *R-i Workspace*, mida avada ja see sisaldas ka vajalikku tudengite andmefaili, fikseerige nimetatud fail: ;

- või siis importige nimetatud fail

(näiteks käsuga: `students = read.csv("http://www.eau.ee/~ktanel/DK_0007/studentsR.csv", header=TRUE, sep=";", dec=",")`)

ja fikseerige siis.

- Veel ühe alternatiivina võite kursuse kodulehelt salvestada tudengite andmestiku *Excel*'i failina ja importida selle siis *R Commander*'isse (*Data -> Import data -> from Excel, Access or dBase data set...*).

**2.**

Ükskõik, kas möödunud nädalal tegelesite sellega või mitte, püüdke sissejuhatuseks prognoosida tudengite peaümberrõõtu, kusjuures eesmärgiks võiks olla saada nii hea mudel peaümberrõõtu prognoosimiseks, kui võimalik.

**2.1.**

Mäletatavasti oli tudengite peaümberrõõtu tugevamini seotud kaaluga kui pikkusega (kui ei usu või mäleta, leidke uuesti näiteks korrelatsioonikordajad).

Seega võikski esmalt prognoosida peaümberrõõtu kaalu alusel.

a) Et regressioonivõrrand on samuti üldine lineaarne mudel, on regressioonanalüüs teostav ka üldiste lineaarsete mudelite konstrueerimiseks mõeldud funktsiooni `lm` abil:

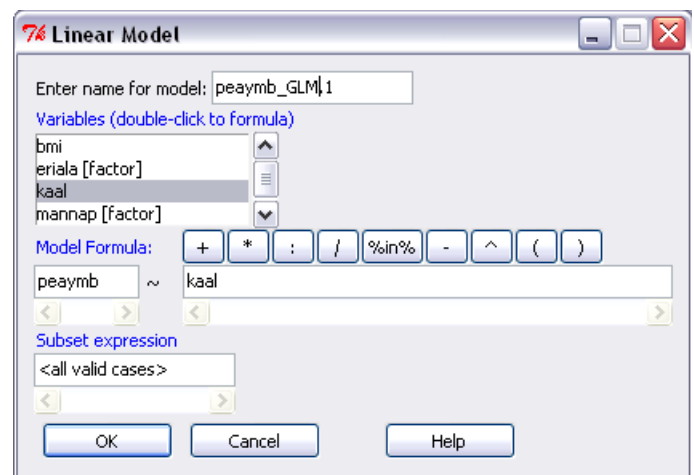
```
peaymb_GLM.1 <- lm(peaymb ~ kaal, data=students) 1
summary(peaymb_GLM.1)
```

Selles käsus `mark <-` tähendab omistamist ja võib olla asendatud ka võrdusmärgiga `=`, `peaymb_GLM.1` on mudelile antud nimi ja käsk `summary` trüüb välja olulisema antud mudelit puudutava statistika. Kui teinekord on soov üksnes kiirelt kontrollida, kas mõni lisatud argument on statistiliselt oluline (vmt) ja ei ole kindlat soovi mudelit hiljem edasi uurida (jääke, prognoose jne), siis ei ole vaja mudelile ka mingit nime omistada (ja seega ka selle nime all salvestada), vaid piisab üksnes käsust kujul

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

b) *R Commander*'is on lineaarsed mudelid tellitavad menüüst

*Statistics -> Fit models -> Linear model ...*



<sup>1</sup> Märkused. Mudelite konstrueerimiseks *R*'is on vajalik tunda kasutatavaid operaatoreid (tehete märke) ja mudelite konstrueerimise eeskirju.

- Oluliseim mudelite esitamisel kasutatav operaator *R*'is on `~`.  
Avaldis `y ~ model` tähistab funktsioontunnuse `y` väärtuste lähendamist lineaarse võrrandiga `model`.  
Taoline võrrand sisaldab märgiga `+` eraldatud muutujaid (näiteks `a + b`).  
Muutujad kujutavad enesest argument- e faktortunnuste nimesid.
- Faktorite `a` ja `b` koosmõju märgitakse kujul `a:b`.
- Esitus `a*b` tähendab `a + b + a:b`.

- Operaator  $\wedge$  võimaldab esitada lühemalt ette antud järku koosmõjusid koos peaeftidega. Näiteks on esitus  $(a+b+c)^2$  sama, mis  $(a+b+c)*(a+b+c)$ , mis omakorda hõlmab faktorite  $a$ ,  $b$  ja  $c$  peamõjusid ja kõiki kahekaupa koosmõjusid:  $a + b + c + a:b + a:c + b:c$ .
- Esitus  $b \%in\% a$  märgib, et faktor  $b$  on allutatud faktorile  $a$ .
- Operaator  $-$  on kasutusel mingi muutuja mudelist välja jätmiseks, näiteks on esitus  $(a+b+c)^2 - a:b$  identne mudeliga  $a + b + c + b:c + a:c$ .
- Ilma vabaliikmeta mudel on esitatav kujul  $y \sim x - 1$ . Alternatiivsed viisid vabaliikmeta mudeli defineerimiseks on  $y \sim x + 0$  ja  $y \sim 0 + x$ .
- Lisaks faktorite nimedele võivad mudelid sisaldada ka matemaatilisi funktsioone.  
Näiteks  $\log(y) \sim a + \log(x)$ .  
Kui rakendatavad funktsioonid sisaldavad mudelite konstrueerimisel kasutatavaid sümboleid, tuleks arusaamatuste vältimiseks need osad mudelist, kus operaatoreid kasutatakse aritmeetiliste tehete mõttes, eraldada kasutades funktsiooni  $I()$ .  
Näiteks mudelis  $y \sim a + I(b+c)$  kasutatakse tunnuse  $y$  modelleerimiseks tunnuste  $b$  ja  $c$  summat.

## Tulemus:

```
> 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
```

Peaümbormõõt =  $45,1738 + 0,1632 \times \text{Kaal}$

Kaalu mõju on statistiliselt oluline

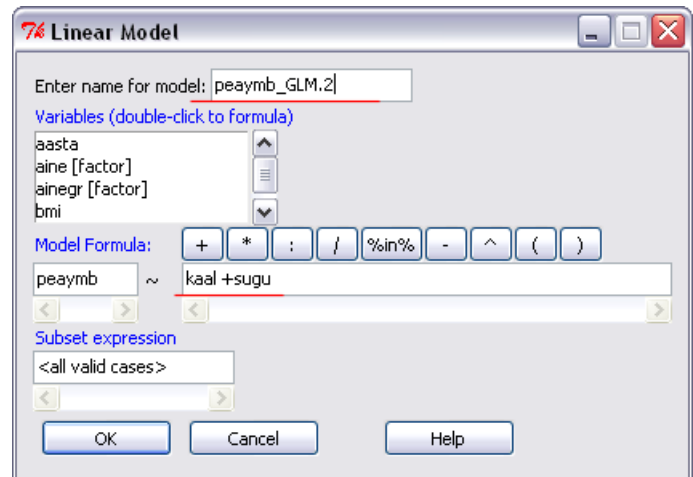
## 2.2.

- Aga kas lisaks kaalule ka tudengite soo arvestamine võimaldab peäumbermõõtu täpsemini prognoosida?

Märkus. Analüüsi, millega ühe korraga analüüsitakse nii pideva argumenttunnuse kui ka diskreetse faktortunnuse mõju, nimetatakse mõnel pool ka **kovariatsioonanalüüsiks** (sisuliselt on tegu kokku pandud regressioon- ja dispersioonanalüüsiga).

*R Commander*'is:

*Statistics -> Fit models -> Linear model ...*



Käsud:

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

```
> summary(peaymb_GLM.2)
```

Tulemus:

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

Kuna diskreetsete faktorite üksikute tasemete mõjud pole üheselt hinnatavad, võtab *R* vaikimisi neist esimese (antud juhul Sugu="M") mõju võrdseks 0-ga (nö baasiks).

Seega tähendab *R*'i väljundi rida `sugu[T.N] -0.18006`, et hinnanguliselt on naisterahvaste keskmine peäumbermõõt meeste keskmisest peäumbermõõdust 0,18 cm võrra väiksem.

Seda, et hinnatud erinevust ei ole erilist põhjust 0-st erinevaks lugeda, näitab näiteks hinnangu standardviga 0,88 cm. Robustselt võttes peaks hinnatud parameetri tegelik väärtus jääma 95%-lise tõenäosusega vahemikku (parameetri hinnang  $\pm 2$  standardviga),

st, et antud juhul peaks meeste ja naiste peäumbermõõtude vaheline erinevus 95%-lise tõenäosusega jääma vahemikku -1,06 kuni 0,70 cm, ehk olema ligikaudu 0.

Olulisuse tõenäosuse  $p = 0,839$  alusel tuleks järeldada, et naisterahvaste peäumbermõõdud ei erine statistiliselt oluliselt meesterahvaste peäumbermõõdudest (sisuliselt tähendab see, et soo mõju peäumbermõõdule ei ole statistiliselt oluline).

Seda, et soo mudelisse lisamine prognoosi täpsust ei parandanud, näitab ka eelmise mudeliga võrreldes peaaegu mitte muutnud  $R^2$  väärtus.

Soovi korral võib siiski kirja panna mudelid nii naiste kui ka meeste peäumbermõõtude tarvis:

$$\text{Peäumbermõõt} \mid \text{Sugu}=\text{"N"} = 45,55 - 0,18 + 0,159 \times \text{Kaal} = 45,37 + 0,159 \times \text{Kaal}$$

$$\text{Peäumbermõõt} \mid \text{Sugu}=\text{"M"} = 45,55 + 0 + 0,159 \times \text{Kaal} = 45,55 + 0,159 \times \text{Kaal}$$

- Aga veel, sobitaks õige mudelisse lisaks soole ka soo ja kaalu koosmõju (Et miks? Ega ma ei teagi, tuli lihtsalt selline mõte, et ehk on peaümberrõõdu ja kaalu vaheline seos meestel ja naistel erinev. Tegelikult tegin vist varem hajuvusdiagrammid või leidsin korrelatsioonikordajad mõlemale soole eraldi (vt joonist eelmise praktikumi ülesandes 4.3), ja tegin seda vaid isiklikule kogemusele tuginedes tekkinud idee kontrollimiseks).

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

Vastavalt  $R^2$ 'i mudelite esitamise süntaksile annab sama tulemuse ka operaatori `*` abil esitatud mudel:

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

Tulemus:

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

Soo mõju pole endiselt statistiliselt oluline ( $p = 0,063$ ), nagu pole seda ka soo ja kaalu koosmõju ( $p = 0,057$ ), aga et mõlemal juhul on  $p$ -väärtus nõu piiri peal ja ka  $R^2$  on üsna mitme protsendi võrra suurem, eelistaksin mina viimast mudelit.

Naiste ja meeste peaümberrõõdud on sellest mudelist prognoositavad järgmiste valemite abil:

$$\text{Peaümberrõõd} \mid \text{Sugu}=\text{"N"} = 34,96 + 12,13 + (0,300 - 0,169) \times \text{Kaal} = 47,09 + 0,131 \times \text{Kaal}$$

$$\text{Peaümberrõõd} \mid \text{Sugu}=\text{"M"} = 34,96 + 0 + (0,300 + 0) \times \text{Kaal} = 34,96 + 0,300 \times \text{Kaal}$$

Väljatrüki viimases reas olev olulisuse tõenäosus ( $p = 4,02 \times 10^{-7}$ ) ütleb, et konstrueeritud mudel tervikuna on statistiliselt oluline.

- Märkus.

Et sisuliselt on tegu meeste ja naiste tarvis eraldi kehakaalu ja peaümberrõõdu vahelise lineaarse seose hindamisega, saame sama hea mudeli ka ilma kehakaalu peamõjuta.

```
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

Veendumaks, et ka võrrandid meeste ja naiste peaümberrõõdude prognoosimiseks tulevad identsed eelnevalt kirja pandutega, kirjutage vastavad valemid välja kõrvaloleva analüüsi tulemustest.

**2.3.**

Sageli ei piisa mudelite võrdlemisel üksnes kirjeldavatel statistikutel (nagu näiteks  $R^2$ ) baseeruvaist võrdlustest. Juhul, kui võrreldavad mudelid on hierarhilised (üks on teisest saadav mingite argumentide lisamise teel), on võimalik testida ka hüpoteesi keerulisema mudeli paremuse statistilisest olulisusest. *R*-s on lihtsaim viis selle teostamiseks kasutada funktsiooni `anova`.

a) Näiteks kui teil on kaks mudelit

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

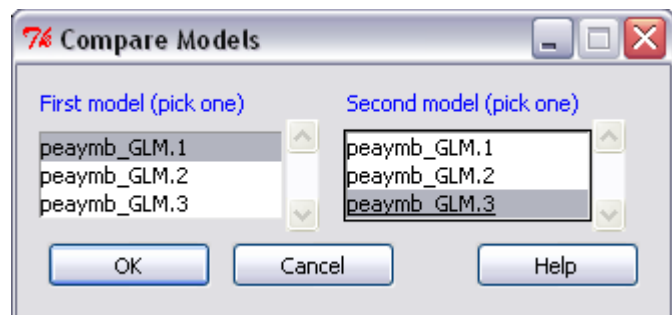
```
peaymb_GLM.3 <- lm(peaymb ~ kaal + sugu + sugu*kaal, data=students)
```

tuleb nende mudelite võrdlemiseks sisestada käsk

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

b) Sama test on valitav ka *R Commander*'i menüüdest:

*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 | <u>0.1594</u> |

Järeldus: komplekssem mudel ei ole statistiliselt oluliselt parem ( $p = 0,159$ ). Samas tähendaks üksnes lihtsama mudeli juurde jäämine põhjendusega, et mis sest keerulisemast ikka rääkida, kui ta ei ole statistiliselt oluliselt parem, seda, et võimalik huvitav peäumbermõõdu ja kaalu vahelise seose sõltuvus soost jääks ära märkimata ...

## 2.4.

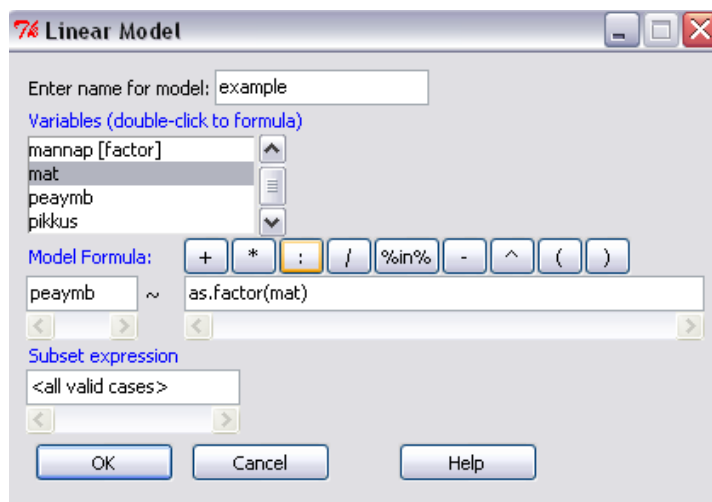
Aga veel – põhimõtteliselt võib uurida ka seda, kas ehk peaümberrõõd hoopis õpitava erialaga või matemaatikahindega seotud pole.

- Viimasel juhul tuleb  $R$ 'ile eraldi öelda, et ta arvulise tunnusega 'mat' mitte regressioonanalüüsi ei püüaks teostada, vaid seda ikka diskreetse faktorina käsitleks. Lihtsaim variant on kirjutada mudelisse liikme mat asemel `as.factor(mat)`:

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

Või sama *R Commander*'is:

*Statistics -> Fit models -> Linear model ...*



- Alternatiivina võib tekitada andmetabelisse 'students' uue,  $R$ 'i poolt juba diskreetse faktor-tunnusena käsitletava tunnuse (kuidas, vt erinevate võimaluste kirjeldusi allpool) ning kasutada mudelis seda.

- Märkus. Diskreetse arvtunnuse baasil faktortunnuse tekitamine.

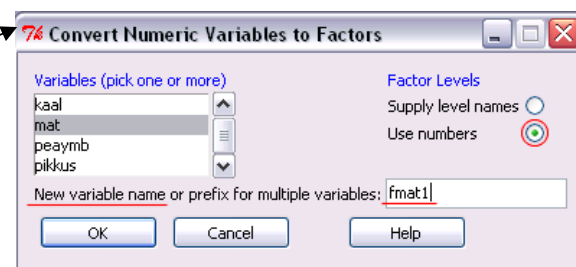
Skriptiaknasse sisestatava funktsiooniga `as.factor`

```
students$fm1 = as.factor(students$mat)
```

või *R Commander*'i menüükäsuga

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

on loodav samuti arvuliste väärtustega '3', '4' ja '5' faktortunnus nimega 'fm1'.



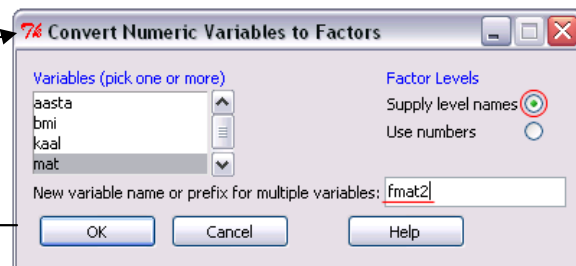
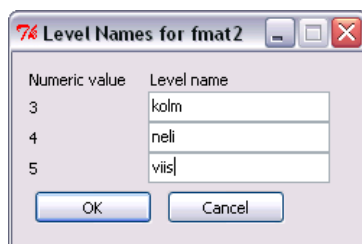
Skriptiakna käsuga

```
students$fm2 = factor(students$mat, labels=c('kolm', 'neli', 'viis'))
```

või *R Commander*'i menüükäsuga

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

on loodav sõnaliste väärtustega faktortunnus (nimega 'fm2').



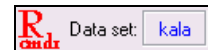
**3.**

Andmestik: [http://www.eau.ee/~ktanel/DK\\_0007/kala.xls](http://www.eau.ee/~ktanel/DK_0007/kala.xls)

Andmestik, mis on osa Mariann Nõlvaku poolt aastail 2004-2006 kogutud Eesti kalade andmebaasist, sisaldab järgmisi andmeid:

- kala number (lihtsalt identifitseerimiseks);
- liik (6 liiki: haug, särg, latikas, luts, ahven ja koha);
- rühm: röövkala või lepiskala;
- 5 püügikohta (Võrtsjärv, Kärevere, Kastre, Praaga ja Peipsi järv);
- püügisesoon (kevad-suvi või sügis-talv);
- kaal ja pikkus;
- sugu;
- laiussiga (*Diphyllbothrium latum*) nakatumine ('diphyl' = 0 või 1);
- laiussi leidude arv kalal ('diph\_arv').

Importige andmestik *R Commander*'isse ja määrake vaikimisi andmestikuks:

**3.1. Mudeli parameetrite hindamine**

Kuidas sõltub latikate kaal elukohast ja soost?

Modelleerime latikate kaalu järgmise lineaarse mudeli abil:

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

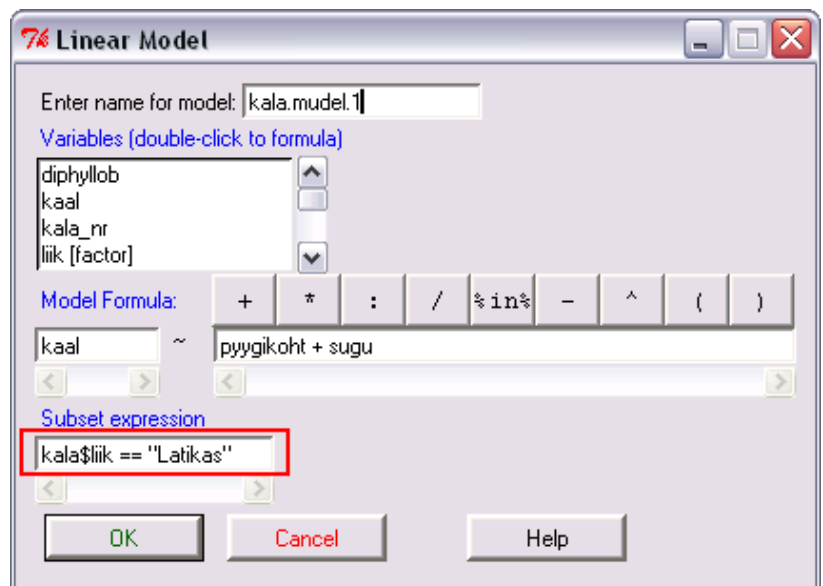
kus  $y_{ijk}$  on  $i$ -st kohast püütud  $j$ . sugu  $k$ . kala kaal,  $K_i$  on  $i$ . püügikoha mõju ( $i=1, \dots, 5$ ) ja  $S_j$  on  $j$ . soo mõju ( $j=1, 2$ ).

**a)** Toodud mudel on rakendatav käsuga

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

**b)** või siis *R Commander*'i menüüdest

*Statistics -> Fit models -> Linear model ...*





Väljavõtt tulemustest:

| 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.Vörtsjärv] | -372.63  | 39.01      | -9.551  | < 2e-16  | *** |
| sugu[T.i]              | -113.36  | 28.14      | -4.028  | 8.52e-05 | *** |

Vabaliige 1020,1 näitab Kastrest püütud emaste kalade keskmise kaalu hinnangut ( $R$  võtab vaikumisi 0-ga võrdseks iga faktori esimese taseme mõju, milleks püügikohtade tähestikulise järjekorra alusel on 'Kastre' ja sugude osas 'e'), saadud hinnangu standardhälve on 32,3 g.

Ülejäänud hinnangud näitavad keskmist erinevust Kastrest püütud emaste kalade keskmisest ja  $p$ -väärtused näitavad vastava erinevuse statistilist olulisust.

Näiteks Kärevere emaste latikate keskmiseks kaaluks saame  $1020,1 - 95,6 = 924,5$  g ja see erineb statistiliselt oluliselt Kastrest püütud emaste latikate keskmisest ( $p = 0,0176$ ).

### 3.2. Kontrastide konstrueerimine

Samuti saame arvutada Vörtsjärve isaste latikate keskmise kaalu:

$$1020,1 - 372,6 - 113,4 = 534,1 \text{ g.}$$

Testimaks viimase erinevust Kastrest püütud emaste latikate keskmisest, tuleb täiendavalt moodustada vastav **kontrast** (= üheselt hinnatav mudeli parameetrite lineaarkombinatsioon) ja testida selle erinevust nullist.

a) Selleks tuleb sisestada käsk

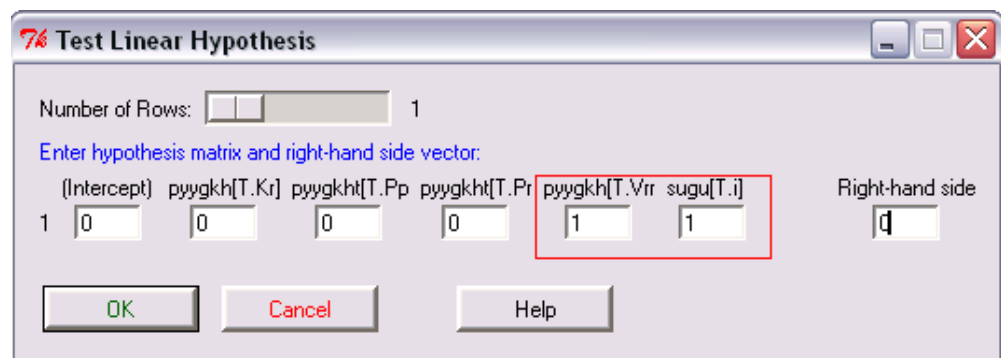
```
linear.hypothesis(kala.mudel.1, c(0,0,0,0,1,1), c(0))
```

Siin esimene argument märgib analüüsi aluseks olevat mudelit, teine omistab mudeli kõigile nullist erinevaks hinnatud parameetritele kaalud (faktorid täpselt sellises järjekorras, nagu mudelis kirjas, ja faktorite tasemed tähestikulises järjekorras), kolmas argument näitab, millega peaks defineeritud kontrast võrduma nullhüpoteesi korral.

b) Sama analüüs *R Commander*'i abil

(vajadusel tuleb määrata *R Commander*'i menüükäskudes kasutatav mudel: Model:  )

*Models -> Hypothesis tests -> Linear hypothesis ...*



Tulemus:

```
Hypothesis:
pyygikoht[T.Vörtsjärv] + 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 *** |

Erinevus on statistiliselt oluline ( $p < 0,001$ ).

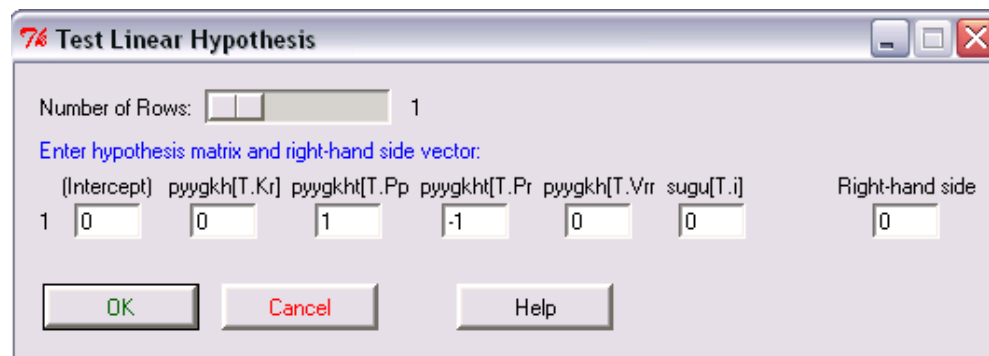
### 3.3.

Aga kas Peipsi ja Praaga latikad erinevad oluliselt?

Lahendus:

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

või



Tulemus:

```
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 ** |

Vastus: erinevad küll ( $p = 0,0014$ ).

### 3.4. Keskmiste väärtuste hindamine

95%-usaldusintervallid Peipsi ja Praaga isaste latikate keskmistele kaaludele on leitavad käskudega

```
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
```

Seega jäävad Peipsi ja Praaga isaste latikate keskmised kaalud 95%-tõenäosusega vahemikesse 918,6...1048,7 g ja 779,1...907,0 g.

### 3.5. Faktorite mõju statistilise olulisuse testimine

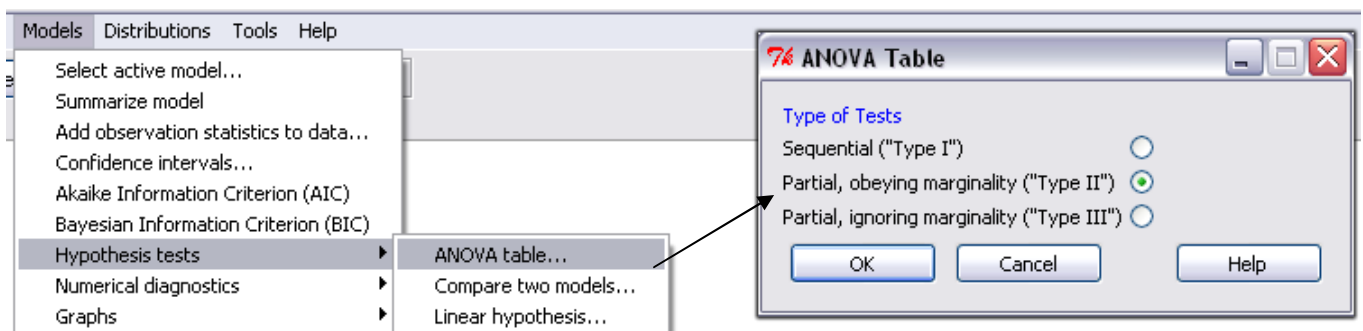
Kas aga püügikoha ja soo mõjud latikate kaalule on üldse statistiliselt olulised?

Hüpoteese faktorite mõjude statistilise olulisuse kohta saab testida käsuga

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

või siis *R Commander*'i menüüdest *Models* -> *Hypothesis tests* -> *ANOVA table*

(vajadusel määrata viimati konstrueeritud mudel vaikimisi mudeliks: Model: `kala.mudel.1`)



Tulemused:

```
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
```

Nii püügikoha kui ka soo mõju latikate kaalule on statistiliselt oluline ( $p < 0,05$ ).

4.

4.1. Aga kas latikate kaal sõltub lisaks ka sesoonist (kevad-suvi, sügis-talv)?

Lisame mudelile sesooni efekti  $L_k$  ( $k=1,2$ ):

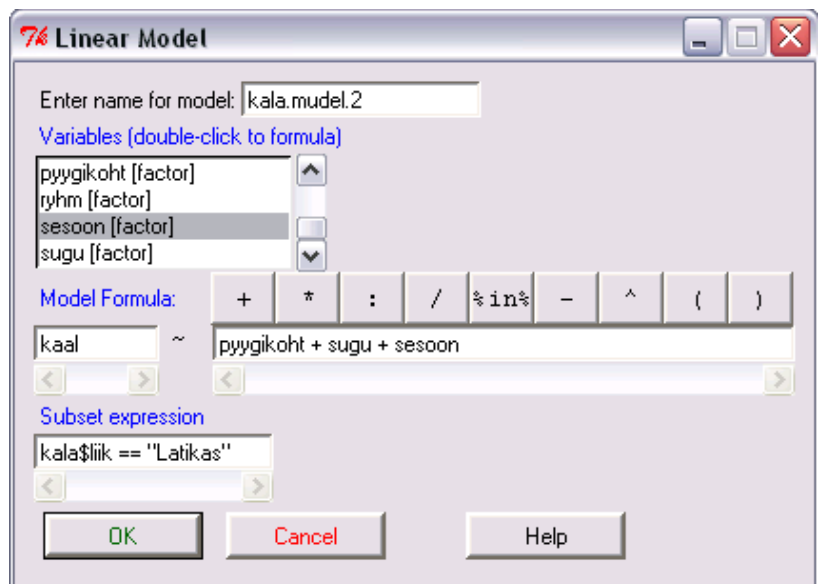
$$y_{ijkl} = \mu + K_i + S_j + L_k + \varepsilon_{ijkl}.$$

a)

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

b) Alternatiivina toodud käskudele võib kasutada ka *R Commander*'it:

Statistics -&gt; Fit models -&gt; Linear model ...



Väljavõtt tulemustest:

| 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 *** |

Sügisel-talvel püütud latikad on keskmiselt 114,3 g kergemad kui kevadel-suvel püütud ja see erinevus on statistiliselt oluline ( $p = 2,54e-05 < 0,001$ ).

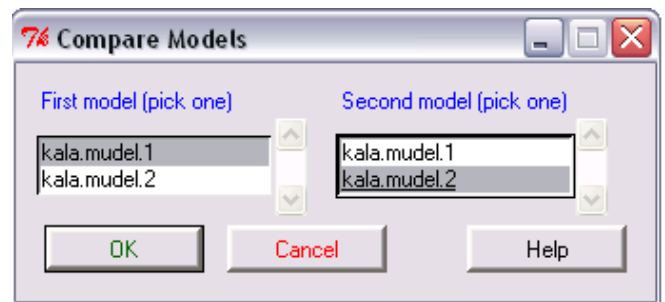
#### 4.2. Kas uus mudel modelleerib latikate kaalu täpsemini?

Lahendus:

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

või

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



Tulemus:

```
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 ***
```

Jah, uus mudel annab statistiliselt oluliselt parema tulemuse ( $p < 0,001$ ).

#### 4.3.

Uurime veel näiteks, kas Tartust ülesvoolu (Kärevere ja Võrtsjärv) püütud latikate keskmine kaal erineb oluliselt Tartust allavoolu (Kastre, Peipsi ja Praaga) püütud latikate keskmisest kaalust.

Vastavalt eelmise lehekülje lõpus toodud tulemuste tabelile on Tartust ülesvoolu jäävate püügi-kohtade (Kärevere ja Võrtsjärv) keskmine mõju  $(-52,6 - 322,4) / 2 = -187,5$  g.

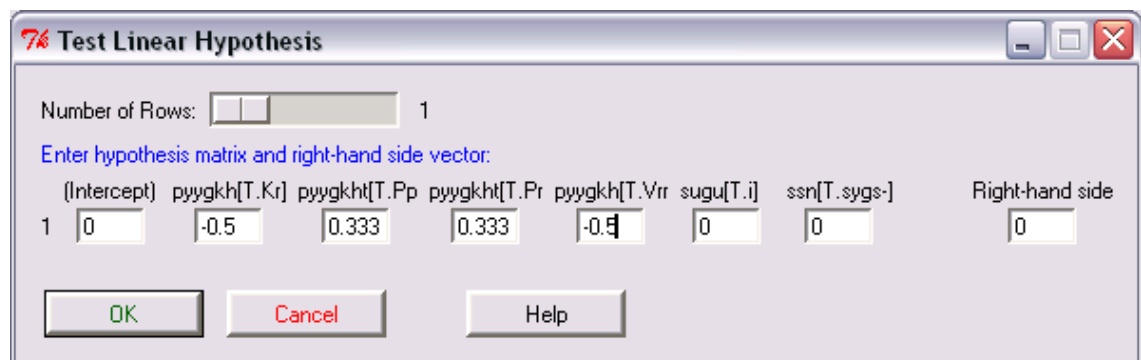
Tartust allavoolu jäävate püügikohtade keskmine mõju on  $(0 - 4,3 + 116,9) / 3 = 37,5$  g.

Leitud keskmiste mõjude (ja seega ka keskmiste kaalude) erinevuse testimiseks tuleb testida vastava kontrasti 0-st erinemist:

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

või

*Models -> Hypothesis tests -> Linear hypothesis...*



Tulemus:

```
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 ***
```

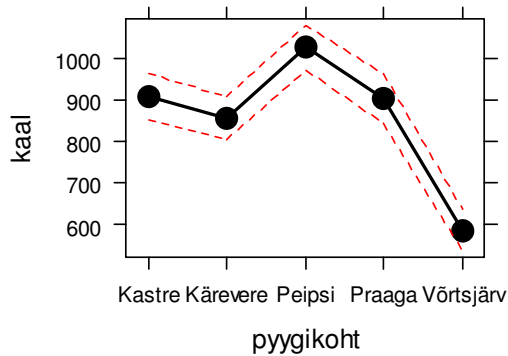
Erinevus on statistiliselt oluline ( $p < 0,001$ ).

#### 4.4.

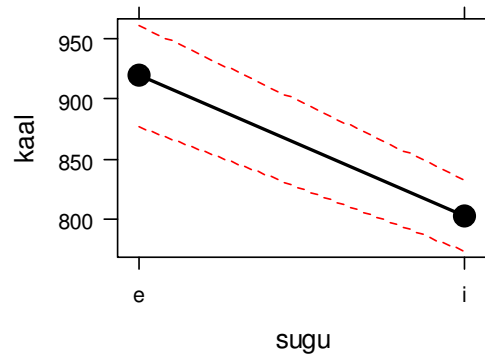
Mudelisse kaasatud faktorite mõjudest ja nende täpsusest kiire ülevaate saamiseks on mugav kasutada *R Commander*'i käsku

*Models -> Graphs -> Effect plots*

**pyygikoht effect plot**



**sugu effect plot**



**sesoon effect plot**

