Praktikum 4

R ja selle lisamoodul *Rcmdr*: sagedustabelid, χ2- ja Fisheri täpne test; korrelatsioon- ja regressioonanalüüs

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://ph.emu.ee/~ktanel/DK_0007/studentsR.csv", header=TRUE, sep=";", dec=",")`

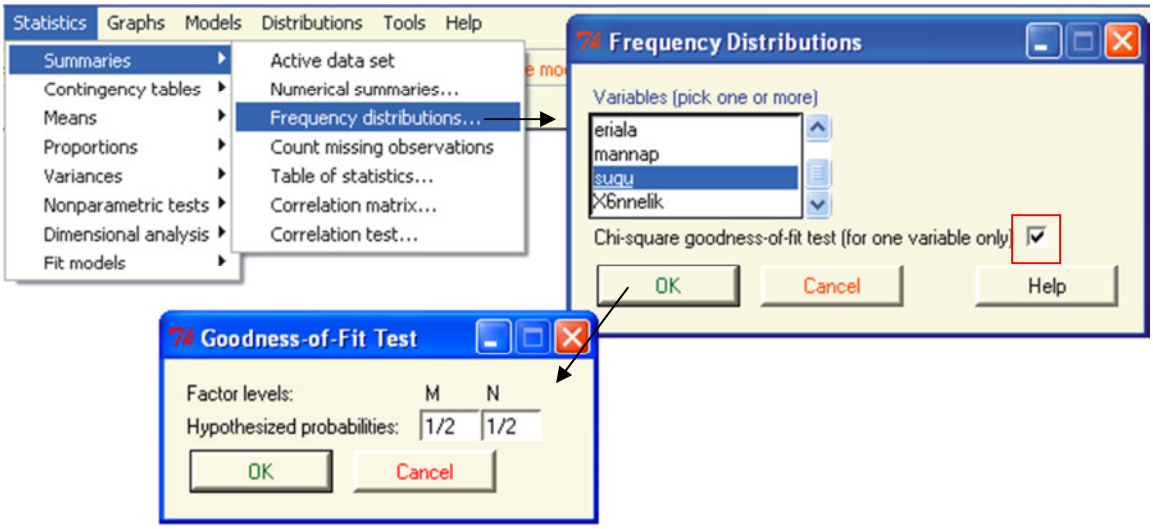
  ja fikseerige siis.

- 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. Sagedustabelid, χ2- ja Fisheri täpne test

2.1. *R Commander*'i menüüdest leitav käsk *Statistics -> Summaries -> Freguency distributions…* võimaldab konstrueerida andmestikus sisalduvale mittearvulisele tunnusele sagedustabeli (st lugeda kokku, kui palju mingeid väärtuseid esineb) ning soovi korral ka testida χ2-testiga väärtuste esinemissageduste erinevust ette antud jaotusest.

- Näiteks võib lasta *R*'il leida andmestikku kuuluvate tudengite soolise jaotuse ja testida selle erinevust 50-50 suhtest:



```
> .Table <- table(students$sugu)

> .Table  # counts for sugu

 M  N
21 79

> 100*.Table/sum(.Table)  # percentages for sugu

 M  N
21 79

> .Probs <- c(0.5,0.5)

> chisq.test(.Table, p=.Probs)

        Chi-squared test for given probabilities

data:  .Table
X-squared = 33.64, df = 1, p-value = 6.631e-09
```

Meeste ja naiste arv

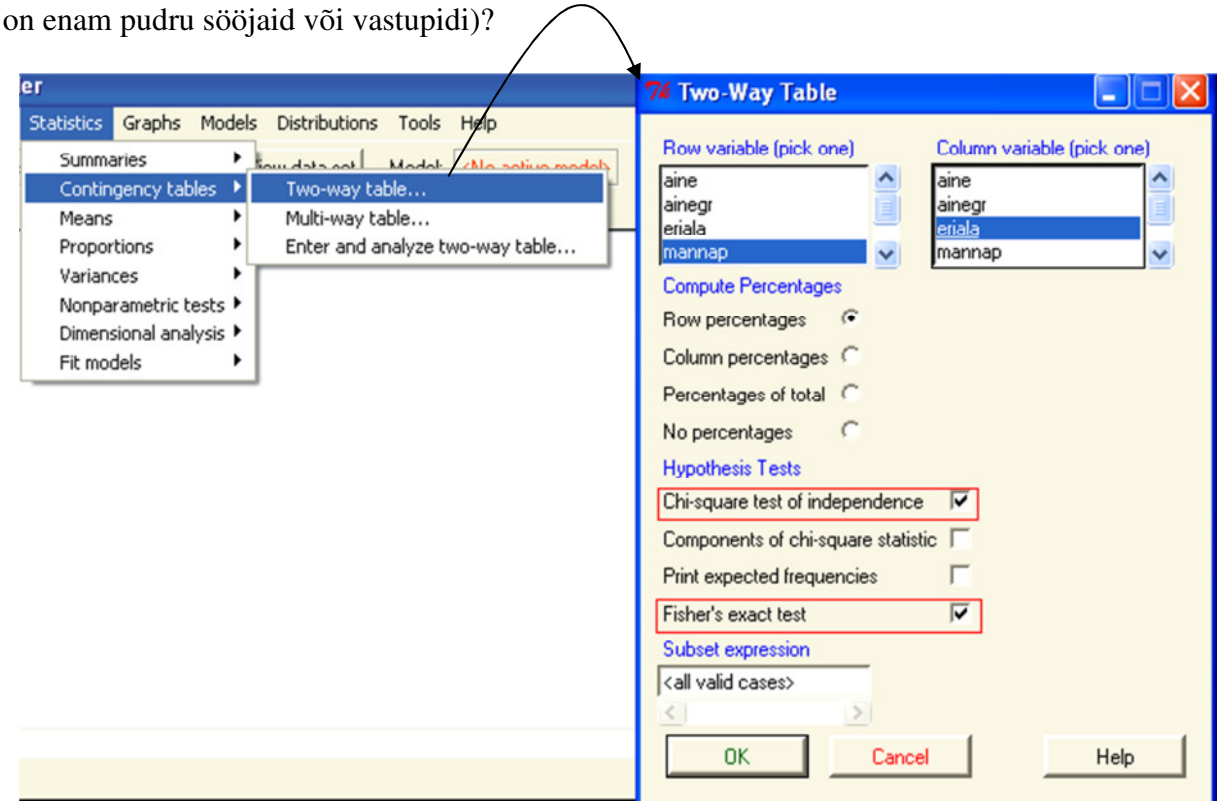
Meeste ja naiste arvu suhtelised sagedused (%), antud juhul võrdsed meeste ja naiste arvuga üksnes põhjusel, et andmestikku kuulub täpselt 100 tudengit

χ^2-testi tulemus ($p = 6.631\times10^{-9}$) näitab, et tudengite soolise jaotuse võib lugeda erinevaks 50-50 suhtest.

- Püüdke rakendada samu käske ka mõne teise tunnuse jaotumise uurimiseks ja testimiseks.
- Saite te kõigist *R Commander*'i poolt skripti aknasse trükitud käskudest aru?

2.2. Kahemõõtmelised sagedustabelid

- Kas eriala ja mannapudru söömine on omavahel seotud (et kas näiteks veterinaaria tudengite seas on enam pudru sööjaid või vastupidi)?



```
> .Table <- xtabs(~mannap+eriala, data=students)

> .Table
           eriala
mannap      LAT LKI
  Ei         23  23
  Jah        22  26
  Nii ja naa  2   4

> rowPercents(.Table) # Row Percentages
           eriala
mannap       LAT  LKI Total Count
  Ei        50.0 50.0   100    46
  Jah       45.8 54.2   100    48
  Nii ja naa 33.3 66.7   100     6

> .Test <- chisq.test(.Table, correct=FALSE)

> .Test

        Pearson's Chi-squared test

data:  .Table
X-squared = 0.6423, df = 2, p-value = 0.7253


> remove(.Test)

> fisher.test(.Table)

        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.7681
alternative hypothesis: two.sided


> remove(.Table)
```

R Commander'i poolt kasutatav käsk

`xtabs(~mannap+eriala,data=students)`

on alternatiiv käsule

`table(students$mannap,students$eriala)`

Käsu

`rowPercents(xtabs(~mannap+eriala, data=students))`

tulemus on peaaegu samaväärne käsu

`100*prop.table(table(students$mannap, students$eriala),1)`

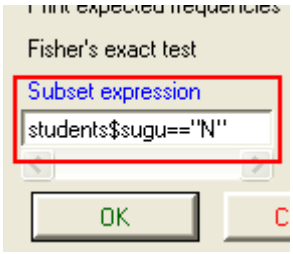
tulemusega. Funktsiooni `prop.table` teine argument väärtusega 1 tähendabki reaprotsentide leidmist.

Proovi järgi.

Nii χ^2- kui ka Fisheri täpse testi (funktsioonid `chisq.test` ja `fisher.test`) tulemusena on korrektne jääda nullhüpoteesi juurde – tudengi eriala ja mannapudru söömine ei ole omavahel seotud (vastavalt $p = 0{,}73$ ja $p = 0{,}77$).

- Paljude analüüside korral sisaldab *R Commander*'i tellimisaken lahtrit <*Subset expression*> analüüsitava andmestiku kitsamaks piiritlemiseks.

Näiteks soovides eelnevalt kirjeldatud analüüse teostada üksnes naisterahvaste baasil, tuleb lahtrisse <*Subset expression*> trükkida lisakitsendus analüüsil kasutatavate andmetabeli '*students*' ridade kohta kujul `students$sugu=="N"`.



Tulemuseks on *R Commander*'i käsk

```
xtabs(~eriala+mannap, data=students, subset=students$sugu=="N")
```

Muidugi on sellele sagedustabeli tegemise käsule ja ka vastavatele χ^2- ning Fisheri täpse testi käskudele olemas samaväärsed alternatiivid:

```
table(students$mannap[students$sugu=="N"],students$eriala[students$sugu=="N"])
chisq.test(table(students$mannap[students$sugu=="N"],students$eriala[students$sugu=="N"]))
fisher.test(table(students$mannap[students$sugu=="N"],students$eriala[students$sugu=="N"]))
```

- Muide, kas te panite tähele χ^2-testi kohta käivat rohelise kirjas hoiatust *R Commander*'i teadete aknas? Mida see hoiatus tähendab?

Aga seda, et üksnes naisterahvaid analüüsides on andmestik χ^2-testi tarvis pisut liiga väike. Küsimusele mannapudru söömisest 'Nii ja naa' vastanute oodatavad sagedused on <5, mistap ei pruugi χ^2-testil arvutatav teststatistik ka nullhüpoteesi kehtides enam χ^2-jaotusele vastavalt jaotuda ning mistõttu ka teststatistiku väärtuse alusel leitav p-väärtus ei pruugi enam õige olla.

```
> .Test$expected # Expected Counts
      mannap
eriala       Ei      Jah Nii ja naa
   LAT 19.49367 21.72152    2.78481
   LKI 15.50633 17.27848    2.21519
```

Mida teha?

Üks variant on kasutada traditsioonilise χ^2-testi asemel sellel baseeruvat Monte-Carlo-testi. Viimase korral ei eeldata andmete alusel leitud teststatistiku χ^2-jaotuse järgset jaotumist (nagu tehakse standardse χ^2-testi korral) vaid hoopis paigutatakse tudengeid suur arv kordi juhuslikult erinevatesse „mannapudru-gruppidesse" ja erialadesse ning arvutatakse iga kord välja traditsioonilise χ^2-statistiku väärtus. Nullhüpoteesile vastavaks jaotuseks, mille alusel p-väärtus leitakse, ei võeta mitte teoreetilist χ^2-jaotust, vaid hoopis kõigi juhuslikult moodustatud sagedustabelite baasil arvutatud teststatistikute jaotus.
Vastava testi *R*-s teostamiseks tuleb funktsioonile `chisq.test` lisada argument `simulate.p.value=TRUE`, soovides täpsustada korduste arvu (vaikimisi 2000), tuleb lisada ka argument `B=5000` (siis korratakse juhuslike sagedustabelite moodustamist 5000 korda).

Et sagedustabeleid moodustatakse juhuslikult, võib ka testi tulemus erinevatel kordadel olla pisut erinev. Näiteks:

```
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  table(students$mannap[students$sugu == "N"], students$eriala[students$sugu ==      "N"])
X-squared = 0.816, df = NA, p-value = 0.6747
```

või

```
        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  xtabs(~eriala + mannap, data = students, subset = students$sugu ==      "N")
X-squared = 0.816, df = NA, p-value = 0.6702
```

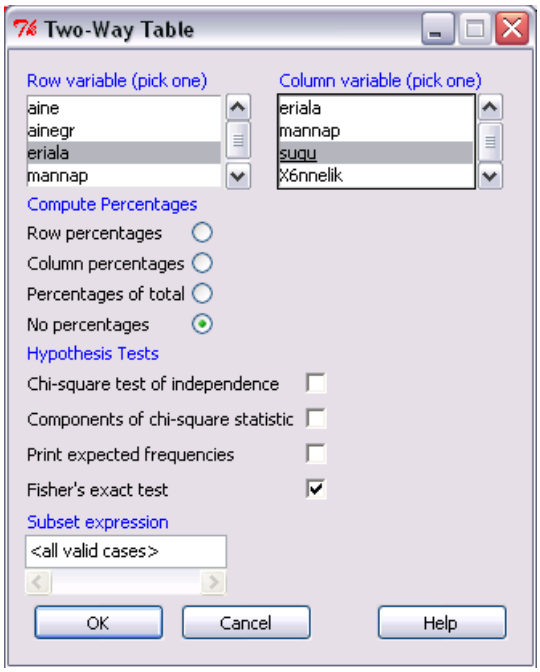
Proovige, kas saate ka erinevad tulemused.

Alternatiivina võib teostada ka Fisheri täpse testi, mis rehkendab välja kõik võimalikud sagedustabelid ja leiab täpse p-väärtuse nende alusel. NB! Suurte andmestike ja/või sagedustabelite puhul teostab *R* ka Fisheri täpse testi Monte-Carlo meetodi kohaselt … (muidu lihtsalt jääkski arvutama).

Ja muidugi võib mannapudru söömisele 'Nii ja naa' vastanud analüüsist üldse välja või siis panna nad näiteks mannaputru söövate tudengitega kokku.

- Juhul, kui analüüsitava sagedustabeli näol on tegu 2×2-tabeliga, leiab Fisheri täpset testi teostav funktsioon `fisher.test` automaatselt ka šansside suhte (*OR*, *odds ratio*) ja selle 95%-usaldus-intevalli.

Näiteks soovides testida, kas meeste-naiste suhe on veterinaarmeditsiini ja loomakasvatuse eriala õppivate tudengite hulgas erinev, võib sooritada Fisheri täpse testi.



Ükskõik, kas *R Commander*'i menüüdest või siis skripti aknasse trükitud vastava käsu abil:

```
fisher.test(students$eriala,students$sugu)
```

Tulemus:

```
        Fisher's Exact Test for Count Data

data:  .Table
p-value = 0.001031
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.02360176 0.51641371
sample estimates:
odds ratio
 0.1350761
```

Olulisuse tõenäosuse (p = 0,001) alusel võib lugeda tõestatuks alternatiivse hüpoteesi: erinevatel erialadel on tudengite sooline jaotus erinev.

Šansside suhe *OR* = 0,135, mis on ligikaudu hinnatav sagedustabelist suhtena (3/18) / (44/35) ≈ 0,133, näitab, et veterinaarmeditsiini eriala tudengite hulgas on meesterahva kohtamise šanss 0,135 korda väiksem kui loomakasvatuse eriala tudengite hulgas.

```
      sugu
eriala  M  N
   LAT  3 44
   LKI 18 35
```

Täpselt sama šansside suhte saame ka pööratud tabeli korral.

```
table(students$sugu,students$eriala) ->
```

```
   LAT LKI
M    3  18
N   44  35
```

OR = (3/44) / (18/35) ≈ 0,133 (*R* annab hinnanguks jällegi 0,135) – meeste hulgas on veterinaar-meditsiini eriala tudengi peale sattumise šanss 0,135 korda väiksem kui naiste hulgas.

Šansside suhte 95%-usaldusintervall (0,024; 0,516) ei sisalda arvu 1, mistõttu võime ka vaid usaldusintervallile tuginedes lugeda tõestatuks alternatiivse hüpoteesi.

3. Korrelatsioonanalüüs

3.1.

Leidke kõigi andmestikku kuuluvate arvetunnuste (va. aasta) vahelised korrelatsioonikordajad, kasutades

a) kas skriptiaknasse sisestatud käske *R*'is või *R* Commander'is:

```
cor(students[,c("bmi","kaal","mat","peaymb","pikkus")], use="pair")
```

Märkused:

¤) puuduv esimene argument (enne koma) käsus
`[,c("bmi","kaal","mat","peaymb","pikkus")]`
ütleb *R*'ile, et analüüsil tuleb kasutada andmestiku kõiki ridu, teine argument ütleb, milliseid veerge analüüsil kasutada;

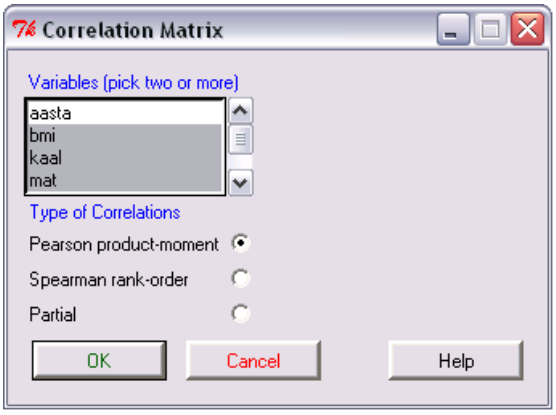
¤) käsk `cor(students)` leiab kõigi andmestikku '*students*' kuuluvate arvtunnuste vahelised lineaarsed korrelatsioonikordajad;

¤) käsk `cor(students$bmi, students$kaal)` leiab üksnes kehamassiindeksi ja kehamassi vahelise lineaarse korrelatsioonikordaja;

¤) lisaargument `use="pair"` on vajalik, kui andmestik sisaldab puuduvaid väärtuseid, selle valiku tagajärjel kasutab R iga kahe tunnuse vahelise seose analüüsil kõiki andmetabeli ridu, kus need **kaks tunnust on mõõdetud**;
alternatiivne lisavalik `use="complete.obs"` kasutab korrelatsioonide arvutamisel aga vaid neid ridu, kus **on mõõdetud kõik analüüsi kaasatud tunnused**;

¤) vaikimisi leiab *R* Pearson'i e lineaarsed korrelatsioonikordajad,
Spearman'i ja Kendall'i korrelatsioonikordajate leidmiseks tuleb kasutada lisavalikuid
`method="spearman"` ja `method="kendall"`
(ka Pearson'i korrelatsioonikordaja arvutamise võib määrata lisavalikuga `method="pearson"`).

b) või samaväärset käsku *R Commander*'i menüüst *Statistics* -> *Summaries* -> *Correlation matrix* …



3.2. Leidke ka peaümbermõõdu ning kaalu, pikkuse ja kehamassiindeksi vahelised astak- ehk Spearman'i korrelatsioonikordajad. Kas on põhjust arvata, et nimetatud tunnuste vahelised seosed ei ole lineaarsed?

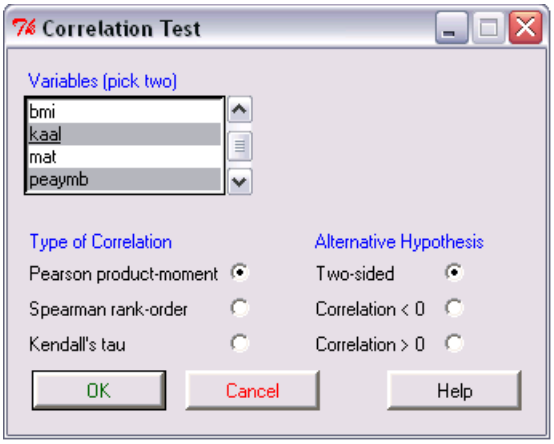
Kui ikka Spearman'i korrelatsioonikordaja Pearson'i korrelatsioonikordajast eriti ei erine, on mõistlik kasutada viimast (on lihtsam ja traditsioonilisem).

3.3. Kas peaümbermõõdu ning kaalu, pikkuse ja kehamassiindeksi vahelised seosed on statistiliselt olulised?

a) Üks võimalus küsimusele vastamiseks on trükkida skripti aknasse vastavate argumentidega funktsioon `cor.test`:

```
cor.test(students$peaymb, students$kaal, method="pearson")
```

b) Alternatiivina on sama analüüs tellitav ka *R Commander*'i menüüst *Statistics -> Summaries -> Correlation test …*



Kui $p < 0.05$, on vastav seos statistiliselt oluline.

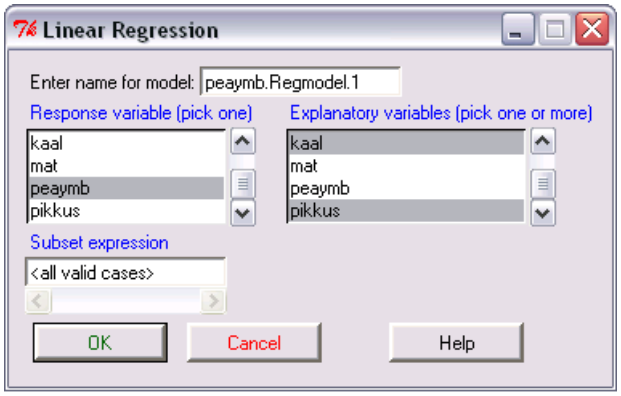
4. Regressioonanalüüs

4.1.

- Inimesed sageli ei tea oma peaümbermõõtu. Kui edukalt on võimalik peaümbermõõtu prognoosida kaalu, pikkuse ja kehamassiindeksi alusel lineaarse regressioonanalüüsi abil? Kas kõik need kolm argumenttunnust on leitud prognoosivõrrandis vajalikud?

Statistics -> Fit models -> Linear regression …

NB! Oluline on anda sobitatavale mudelile nimi, sest selle alusel on edaspidi võimalik tellida sama mudeli kohta täiendavaid analüüse!!



Vastav *R Commander*'i poolt produtseeritav käsk on kujul:

```
peaymb.Regmodel.1 <- lm(peaymb~bmi+kaal+pikkus, data=students)
```

Märkused:

¤) *R*'ile vaikimisi omase omistamise märgi <– asemel võib kasutada ka märki = ;

¤) funktsioonile `lm` võinuks sama mudeli (skripti aknasse trükkides) ette anda ka kujul

```
lm(students$peaymb ~ students$bmi + students$kaal + students$pikkus).
```

- Mudeli `peaymb.Regmodel.1` tulemuste väljastamiseks tuleb kasutada käsku

`summary(peaymb.Regmodel.1)`

Tulemus:

```
Call:
lm(formula = peaymb ~ bmi + kaal + pikkus, data = students)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4769 -1.4241  0.2410  1.8189  9.1708

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  76.7956    48.0530   1.598    0.114
bmi          -0.7785     1.0830  -0.719    0.474
kaal          0.4170     0.3690   1.130    0.261
pikkus       -0.1807     0.2817  -0.642    0.523

Residual standard error: 2.929 on 88 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-Squared: 0.2836,     Adjusted R-squared: 0.2592
F-statistic: 11.61 on 3 and 88 DF,  p-value: 1.747e-06
```

Residuals description

Estimated regression coeficients and corresponding p-values

R-Square, describing the goodness of fit of the model

Model's statistical significance

- *R Commander*'i menüüdest

Models -> Numerical diagnostics and *Models -> Graphs*.

on tellitav hulk mudeli diagnostikaks vajalikke parameetreid ja jooniseid.

Näiteks *Models -> Numerical diagnostics -> Variance-inflation factors* võimaldab analüüsida mudeli argumentide multikollineaarsust (so mudeli argumentide omavahelist seotust – kui mudeli argumendid on omavahel tugevalt seotud, on ka nende mõjud omavahel seotud ja mudelist saadud mõjude hinnangud ei pruugi olla korrektsed).
Enamasti loetakse multikollineaarsus suureks kui VIF > 10 (ja siis tuleks mõni argument mudelist välja jätta).

Antud juhul:

```
> vif(peaymb.Regmodel.1)
     bmi     kaal   pikkus
 94.4982 174.8210  62.0392
```

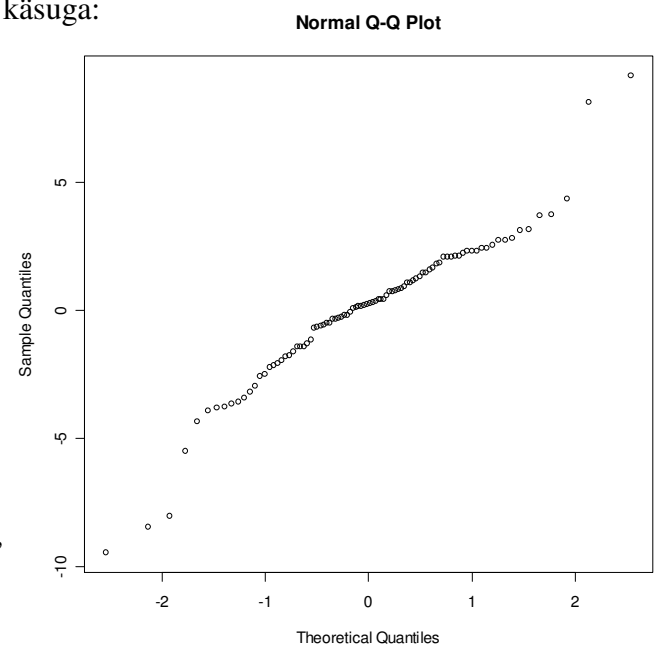
Seega on multikollineaarsus selgelt liiga suur ja vähemalt üks argumentidest oleks mõttekas mudelist välja jätta. Milline? Vaadake korrelatsioone – argument, mis on uuritava tunnusega kõige tugevamini seotud, oleks mõistlik mudelisse alles jätta …

Proovige lihtsamaid vähema arvu argumentidega mudeleid (andes kõigile oma nimed).

- Valik *Models -> Graphs -> Basic diagnostic plots* võimaldab tellida mitmeid mudeli sobivuse ja eelduste täidetuse kontrollimiseks mõeldud jooniseid, muuhulgas saab lasta *R*'il välja joonistada jääkide ja uuritava tunnuse väärtuste vahelise hajuvusdiagrammi (otsustamaks, kas konstrueeritud mudel prognoosib kõiki uuritava tunnuse väärtusi võrdse täpsusega) ja nn jääkide tõenäosuspaberi jääkide normaaljaotusega võrdlemiseks (*normal Q-Q plot*).

Viimane joonis lihtsalt tellitav ka skripti aknast käsuga:

```
qqnorm(resid(peaymb.Regmodel.1))
```



Päris täpselt mudeli jäägid normaaljaotuse järgi ei jaotu (punktid graafikul ei paikne päris täpselt diagonaalsel sirgel), leiduvad üksikud suured jäägid, ja seda nii positiivsete kui ka negatiivsete väärtuste poolt.
Kui võimalik, tuleks püüda mudelit parandada, võttes arvesse täiendavaid argumente või muutes funktsiooni.
Samas, reaalses elus ei õnnestugi sageli eriti täpseid prognoosivõrrandeid konstrueerida, mistõttu võib parema puudumisel jääda ka taoliste jääkidega mudeli juurde.

4.2.

- Tundub, et optimaalseim mudel peaks sisaldama vaid üht argumenti (kas jõudsite ka sellisel järeldusele?), ja selleks argumendiks peaks olema kehamass (tunnus 'kaal').

Vastava analüüsi tulemus on järgmine:

```
> peaymb.Regmodel.4 <- lm(peaymb~kaal, data=students)

> summary(peaymb.Regmodel.4)

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
```

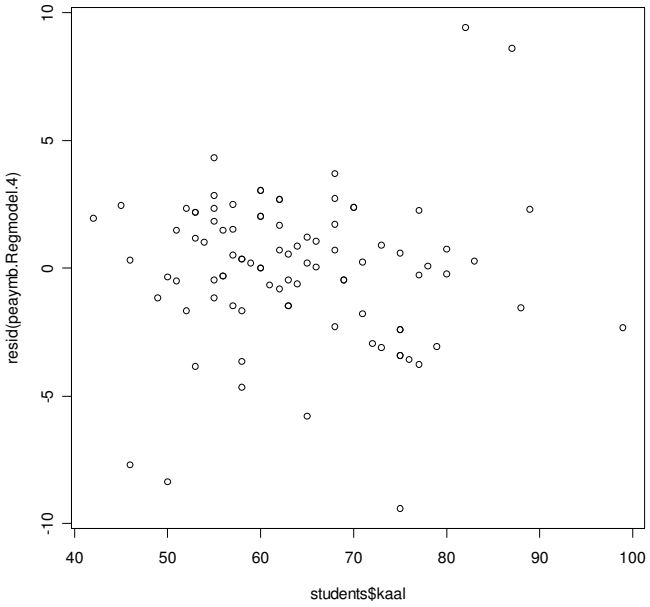
Kas see mudel on statistiliselt oluline?

Oskate kirja panna ka regressioonivõrrandi?

- Taoline lihtne mudel on illustreeritav mitmete skriptiaknasse sisestatud käskude abil tellitavate 2-mõõtmeliste joonistega.

Näiteks on mudeli jääkide graafik tellitav käsuga

```
plot(students$kaal,resid(peaymb.Regmodel.4))
```



Mingit selget tendentsi kõrval olevalt jääkide hajuvusdiagrammilt silma ei hakka, sestap võib mudeli lugeda sobivaks.

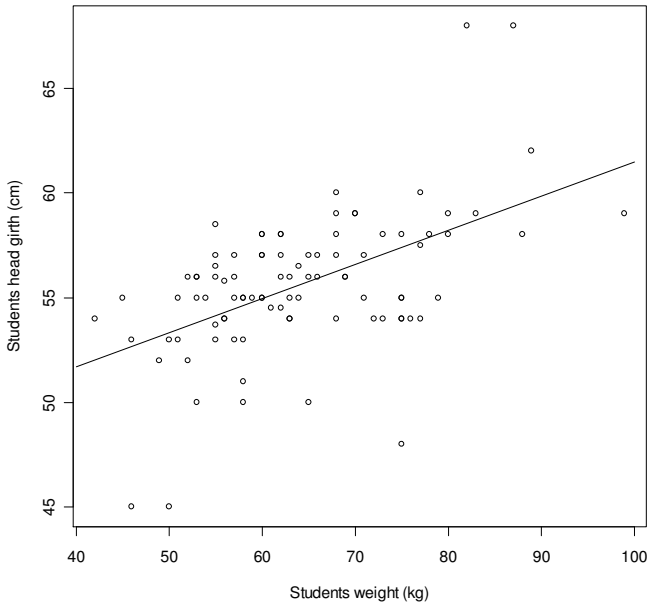
- Peaümbermõõdu prognoos mingite kindlate kehamasside korral (näiteks kehamassidele 40 ja 100 kg) on arvutatav käsuga

```
prediction=predict(peaymb.Regmodel.4,data.frame(kaal=c(40,100)))
```

(prognoositud peaümbermõõtude vektorit ei trükita väljundisse vaid omistatakse muutuja `prediction` väärtuseks, soovides prognoositud väärtusi väljundi aknas näha, tuleks skripti aknasse trükkida neid sisaldava muutuja nimi).

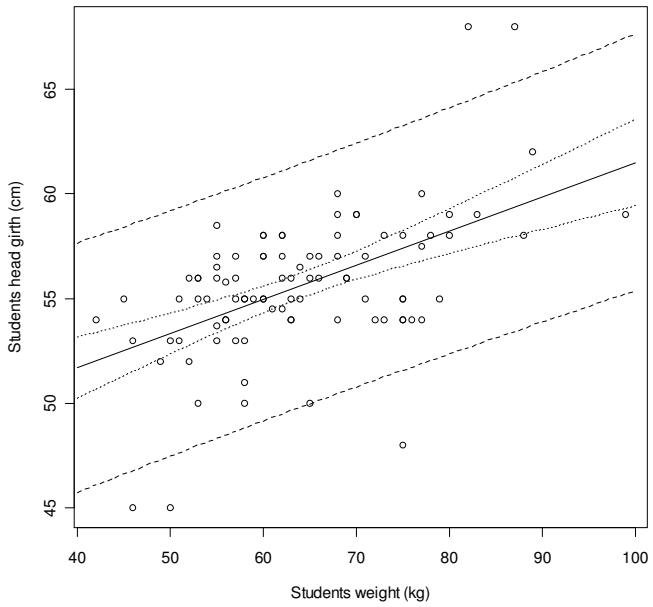
Kehamasside ja peaümbermõõtude hajuvusdiagramm koos regressioonisirgega on tellitav skripti aknasse trükitava käsuga

```
plot(students$kaal,students$peaymb,
   xlab="Students weight (kg)",
   ylab="Students head girth (cm)")
lines(c(40,100),prediction)
   # see käsk joonistab regressioonisirge
```

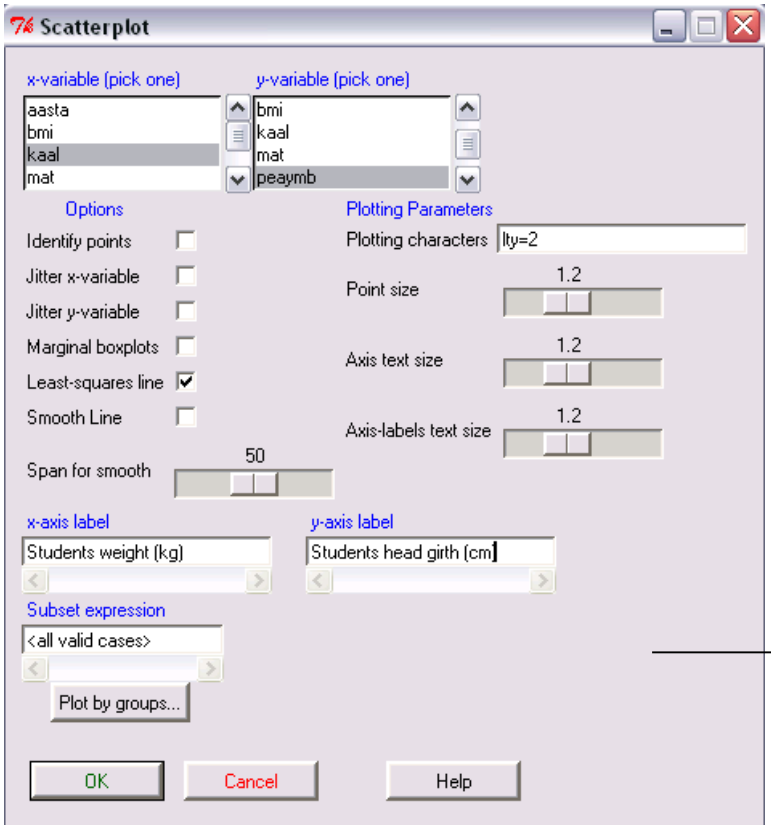


Samale joonisele on lisatavad ka tolerantsi- ja usaldusintervalli laiust näitavad jooned (vaikimisi leitav 95%-tolerantsiintervall näitab piirkonda, kuhu peaks jääma 95% tudengitest oma kehamasside ja peaümbermõõtudega; 95%-usaldusintervall aga näitab, kus paikneb 95%-lise tõenäosusega tegelik kehamassi ja peaümbermõõdu vahelist seost kirjeldav regressioonisirge).

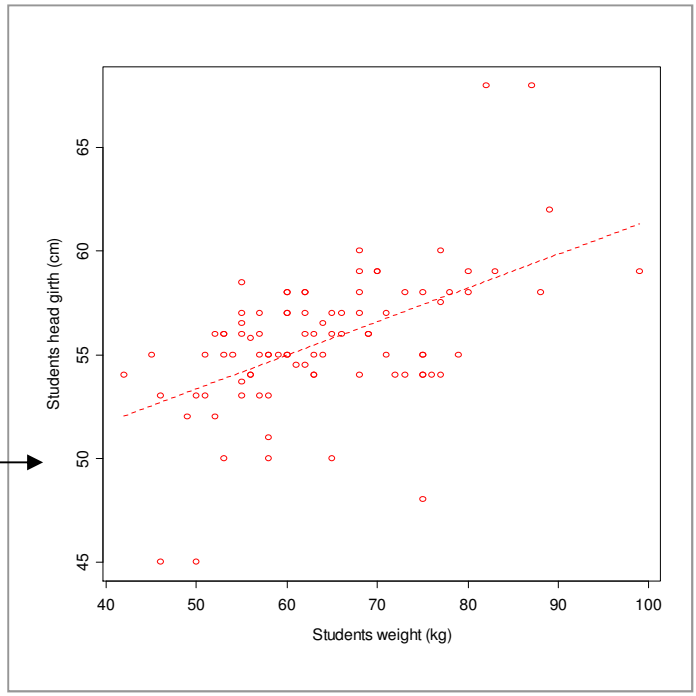
```
prediction=predict(peaymb.Regmodel.4,
 data.frame(kaal=c(40:100)),interval="prediction")
lines(40:100,prediction[,2],lty=2)
lines(40:100,prediction [,3],lty=2)

prediction=predict(peaymb.Regmodel.4,
 data.frame(kaal=c(40:100)),
 interval="confidence")
lines(40:100,prediction[,2],lty=3)
lines(40:100,prediction[,3],lty=3)
```



- Analoogne regressioonisirgega hajuvusdiagramm on tellitav ka *R Commander*'i menüüst *Graphs -> Scatterplot …*



Menüüdest mitte valitavad usaldus- ja/või tolerantsiintervallid võib lisada joonisele skripti aknasse sisestatud käskude `predict` ja `lines` abil.

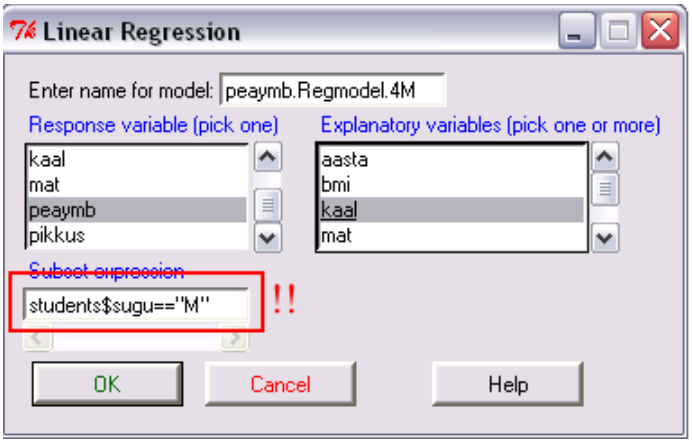


4.3. Kas peaümbermõõdu ja kehamassi vaheline seos on ühesugune nii meeste kui ka naiste korral?

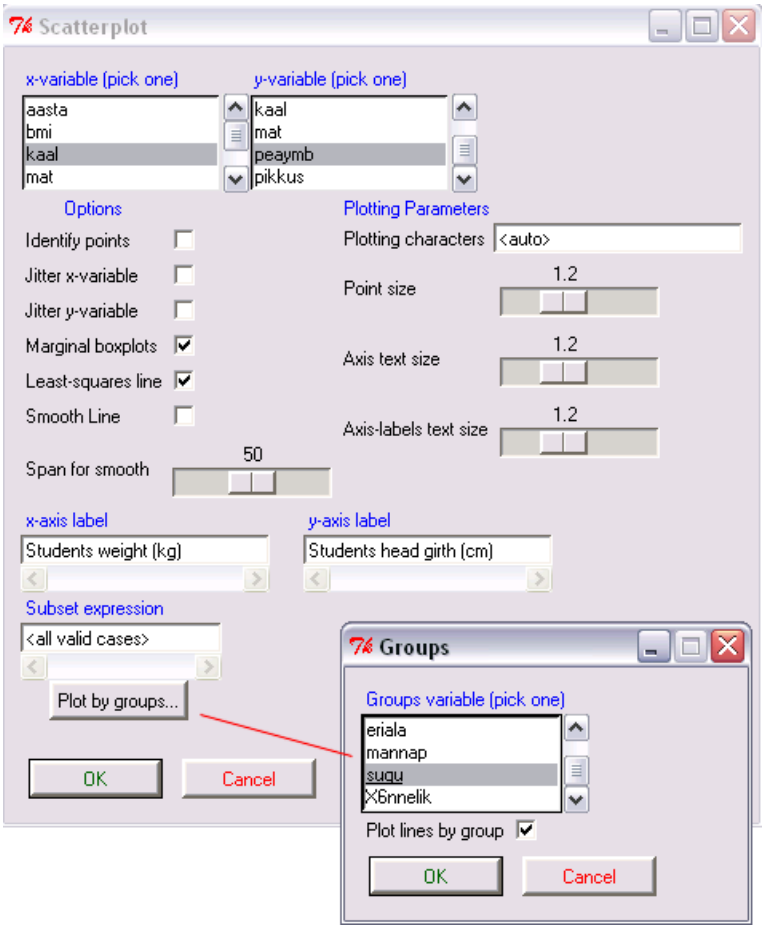
- Üks võimalus on leida regressioonivõrrand eraldi mõlema soo tarvis.
  Näiteks skripti aknasse sisestatud käskude abil:

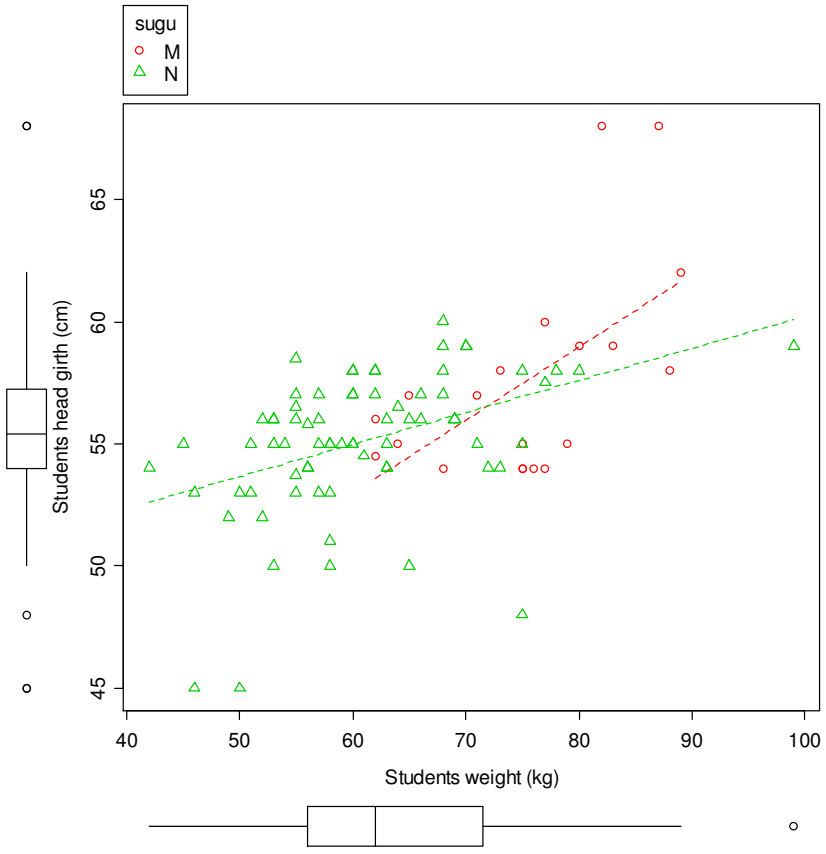
```
peaymb.Regmodel.4M <- lm(peaymb~kaal, data=students[students$sugu=="M",])
summary(peaymb.Regmodel.4M)
peaymb.Regmodel.4N <- lm(peaymb~kaal, data=students[students$sugu=="N",])
summary(peaymb.Regmodel.4N)
```

Või siis *R Commander*'i menüüdest *Statistics -> Fit models -> Linear regression …* märkides vajaliku lisakitsenduse lahtrisse <*Subset expression*>.



- Visuaalselt on selgeim joonistada hajuvusdiagrammile eraldi regressioonisirged nii meeste kui ka naiste jaoks.
  R Commander'i menüükäsu *Graphs -> Scatterplot …* abil on see suhteliselt lihtne (ja samas palju lisavõimalusi pakkuv).





4.4.

Ja lõpetuseks, püüdke aru saada, milliseid erinevaid jooniseid on võimalik konstrueerida *R Commander*'i menüüst *Graphs -> Scatterplot matrix*….