Biomeetria bioloogidele
5. loeng

osa I

# Mitmene regressioonanalüüs

(kovariatsioonanalüüs, 2-way ANOVA)

Interpretatsioon, multikollineaarsus, koosmõjud, …

## Miks?

Miks kasutada funktsioontunnuse $Y$ väärtuste prognoosimisel rohkem kui ühte tunnust?

- Täiendava informatsiooni abil võime jõuda täpsemate prognoosideni;
- Isegi, kui meid huvitab vaid ühe tunnuse ($X$) mõju sõltuvale tunnusele ($Y$), võimaldab segavate faktorite arvesse võtmine kirjeldada meid huvitavat seost täpsemalt;
- Kui argumenttunnuse $X$ mõju funktsioontunnusele ($Y$) muutub sõltuvalt mingi kolmanda tunnuse $Z$ väärtustest (eksisteerib koosmõju tunnuste $X$ ja $Z$ vahel), siis on tunnuse $X$ mõju $Y$-le võimalik kirjeldada vaid mitmese regresioonanalüüsi abil;
- On olukordi, kus peale täiendavate tunnuste lisamist regressioonmudelisse kaovad regressioonanalüüsi eeldustega seotud probleemid (näiteks mudeli jääkide jaotus võib peale täiendavate tunnuste lisamist muutuda normaaljaotuseks).

## Miks mitte kasutada?

Iga regressioonmudelisse lisanduva parameetri hindamisel võime veidi eksida. Paljude hinnatavate parameetritega mudelis võivad ka pisikesed hindamisel tehtavad eksimused kumuleeruda ning tulemuseks saame üsna kehvasti prognoosiva mudeli (mida keerukam on aparaat, seda rohkem on seal ka osi, mis katki minna võivad). Seega ära lisa mudelisse rohkem hinnatavaid parameetreid kui hädapärast vaja (vali endale kõigist kasutuskõlblikest mudelitest lihtsaim)!

(Tuleviku) prognoosimiseks kasutavasse mudelisse pole tark lisada tunnuseid, mille hilisem mõõtmine on kas kallis, tülikas või võimatu – kui teeme prognoosimise liiga kalliks ja vaevaliseks, ei hakata meie ilusat ja häid prognoose andvat mudelit nagunii tarvitama;

Kui eesmärgiks on funktsioontunnuse *Y* väärtuseid mõjutavate tunnuste väljaselgitamine (põhjus-tagajärg seoste leidmine), siis ei tohiks mudelisse lisada ka mõningaid selliseid tunnuseid, mille väärtuste teadmine aitaks *Y*-i väärtuseid täpsemalt prognoosida (täpsemalt hilisemates loengutes, probleemiks vaid vaatlusandmete analüüsimisel, korrektselt korraldatud katsete korral probleemi ei teki).

## Mitmene regressioon – interpretatsioon – pidev tunnus + faktortunnus

```
> summary(lm(kaal~pikkus+factor(sugu)))
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -90.47341    8.60270 -10.517  < 2e-16 ***
pikkus          0.89125    0.05117  17.418  < 2e-16 ***
factor(sugu)2   4.81850    1.02693   4.692 3.31e-06 ***

Residual standard error: 7.845 on 640 degrees of freedom
Multiple R-Squared: 0.5715,     Adjusted R-squared: 0.5701
F-statistic: 426.7 on 2 and 640 DF,  p-value: < 2.2e-16
```

1 cm pikem mees kaalub lühemast mehest keskmiselt 0,89 kg rohkem.
1 cm pikem naine kaalub lühemast naisest keskmiselt 0,89 kg rohkem.

meestudeng (sugu=2) kaalub **samapikast** naistudengist keskmiselt 4,8kg rohkem:

Naise kaal = -90,47 + 0,89125 *pikkus* + ε
Mehe kaal = -90,47 + 4,818 + 0,89125 *pikkus* + ε

Mitmene regressioon – interpretatsioon – hüpoteeside testimine

```
> summary(lm(DVR~pikkus+kaal))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.36094   11.25398   4.120 4.56e-05 ***
pikkus       0.11727    0.07986   1.468   0.1427
kaal         0.13520    0.05744   2.354   0.0190 *
Multiple R-Squared: 0.06167,    Adjusted R-squared: 0.05731
```

Inimese kaalu ja pikkuse pealt tehtud süstoolse vererõhu prognoos on (tõestatavalt) täpsem kui inimese pikkuse pealt tehtav prognoos.

Inimese kaalu ja pikkuse pealt tehtud süstoolse vererõhu prognoos pole (tõestatavalt) täpsem kui inimese kaalu pealt tehtav prognoos.

Arvatavasti tasub mitmese regressiooni asemel teha tavalist regressioonanalüüsi ja argumenttunnusena on targem kasutada kaalu.

```
> summary(lm(DVR~pikkus))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 31.12448    9.24727   3.366 0.000831 ***
pikkus       0.25646    0.05391   4.758 2.67e-06 ***
Multiple R-Squared: 0.04968,    Adjusted R-squared: 0.04748

> summary(lm(DVR~kaal))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.46929    2.51592  24.830  < 2e-16 ***
kaal         0.19758    0.03872   5.103 5.03e-07 ***
Multiple R-Squared: 0.05697,    Adjusted R-squared: 0.05478
```

Vaid kaalu kasutades saame prognoosida vaadeldud tudengite diastoolset vererõhku peaaegu sama täpselt kui kaalu ja pikkust kasutades (uute tudengite vererõhku prognoosides annaks vaid kaalu kasutav mudel ehk isegi täpsema tulemuse...

Võimalik on olukord, kus ükski tunnustest ei lisa täiendavat informatsiooni, kuid kumbki tunnustest omaette võttes on oluline:

```
> summary(lm(DVR~pikkus+kaal, subset=(vanus==19)))
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.1533    19.4720   1.703   0.0904 .
pikkus        0.2080     0.1403   1.482   0.1402
kaal          0.1096     0.1137   0.964   0.3365

> summary(lm(DVR~pikkus, subset=(vanus==19)))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.54114   16.45784   1.430  0.15437
pikkus       0.30348    0.09669   3.139  0.00199 **

> summary(lm(DVR~kaal, subset=(vanus==19)))
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 61.10244    4.85242  12.592  < 2e-16 ***
kaal         0.23136    0.07878   2.937  0.00376 **
```

Võime kasutada diastoolse vererõhu prognoosimisel kas kaalu või pikkust. Sisuliselt on tegemist võrdväärsete alternatiividega – kumba eelistada, jääb suuresti kasutaja otsustada.

Miks antud tunnused on samaväärsed?

Kaalu ja pikkuse vaheline korrelatsioon on väga kõrge – r=0,75. Seega pikk tudeng ka kaalub enamasti rohkem kui lühike tudeng ning arvutil on raske otsustada, kas inimesel on vererõhk kõrge seepärast, et ta kaalub palju, või seepärast, et ta on pikk (või hoopistükkis suure ruumala tõttu).

Ekstremaalne näide:
Kui kaks tunnust käituvad valimis täpselt samamoodi ($X_1 = X_2$), siis järgnevad mudelid prognoosivad kõik võrdväärselt hästi $Y$-tunnuse väärtuseid:

$Y = 2X_1 - 3X_2$
$Y = -X_2$
$Y = -X_1$
$Y = -4X_1 + 3X_2$

Järeldus: lisades mudelile teiste tunnustega korreleeritud tunnuseid võivad varem mudelis olnud tunnuste kordajad kergesti ka märki vahetada.

## Multikollineaarsus

Multikollineaarsus (*multicollinearity*) – argumenttunnuste kõrge korreleeritus (Kui argumenttunnuste korrelatsioon on suurem kui 0,7 siis tavaliselt öeldakse, et tegemist on multikollineaarsusega).

Multikollineaarsus pole probleemiks, kui eesmärgiks on leida sõltuvat tunnust hästi (täpselt) prognoosiv mudel.

Multikollineaarsus muutub probleemiks, kui üritame saadud mudeli parameetreid (põhjuslikult) tõlgendada:

- regressioonikordajate hinnangud sageli äärmiselt ebatäpsed;
- statistiliselt oluliste regressioonikordajate märk võib muutuda ka tehes väikeseid parandusi või muudatusi mudelis;
- statistiliste meetodite võimetus jagada „mõju“ funktsioontunnusele kahe või enama argumenttunnuse vahel;

Näiteid tunnustest, mis võivad tänu multikollineaarsusele segadust põhjustada:

Noortel ja lastel pikkus ≈ kaal ≈ vanus

Rääkides pikkuse mõjust ühele või teisele näitajale peame silmas ehk hoopis vanuse mõju?

Haridus ja sissetulek on sageli tugevalt seotud, seega võib hariduse lisamine mudelile sissetuleku mõju hoopistükkis tagurpidiseks pöörata (ja/või ebaoluliseks muuta);

Vanus ja sünniaasta käivad sageli käsikäes. Seega võib vanemate patsientide tervis olla kehvem sellepärast, et nende lapsepõlv möödus rasketes tingimustes, aga ka sellepärast, et nad on lihtsalt vanad.

Kui haigusega *X* peaaegu alati kaasneb ka haigus *Y*, siis pole sageli statistilise analüüsi abil võimalik öelda, kas mingi tervisenäitaja halvenes haiguse *X* või haiguse *Y* pärast.

...

# Koosmõju

Koosmõjudest - ilma koosmõjudeta mudel

```
> summary(lm(sdp~ravim+sugu))

Residuals:
     Min       1Q   Median       3Q      Max
-29.0498  -6.8232  -0.4295   6.5545  32.1179

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   180.8660     0.7795 232.026   <2e-16 ***
ravimaktiivne  -0.3685     0.7744  -0.476    0.634
suguN         -20.7197     0.7787 -26.607   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 9.973 on 997 degrees of freedom
Multiple R-Squared: 0.5111,     Adjusted R-squared: 0.5101
F-statistic: 521.2 on 2 and 997 DF,  p-value: < 2.2e-16
```

**Koosmõju** – kui argumenttunnuse $X_1$ mõju funktsioontunnusele $Y$ sõltub mingi kolmanda tunnuse $X_2$ väärtusest, siis öeldakse, et eksisteerib tunnuste $X_1$ ja $X_2$ koosmõju tunnusele $Y$.

**Näide**

Uuritakse, kas ja kuidas ravim W mõjutab patsiendi vererõhku. Selgub, et ravimi W toimel naiste vererõhk langeb ja meeste vererõhk ei muutu. Järelikult eksisteerib koosmõju tunnuste *ravim* (saab ravimit W/ ei saa ravimit W) ja *sugu* vahel – ravimi toime sõltub inimese soost.

## Koosmõju uurimine R'is:

```
> m1=lm(sdp~ravim+sugu+ravim*sugu)
> summary(m1)

Call:
lm(formula = sdp ~ ravim + sugu + ravim:sugu)

Residuals:
     Min       1Q   Median       3Q      Max
-36.9906  -6.4563   0.2498   7.0918  28.6907

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)         180.0365     0.7008 256.905  < 2e-16 ***
ravimaktiivne        -0.4365     0.9642  -0.453    0.651
suguN                 5.6875     0.9141   6.222 7.21e-10 ***
ravimaktiivne:suguN  -9.5548     1.2925  -7.393 3.05e-13 ***

Residual standard error: 10.13 on 996 degrees of freedom
Multiple R-Squared: 0.1224,     Adjusted R-squared: 0.1197
F-statistic: 46.29 on 3 and 996 DF,  p-value: < 2.2e-16
```

Parameetrite interpreteerimine

```
Coefficients:
                      Estimate
(Intercept)           180.0365
ravimaktiivne          -0.4365
suguN                   5.6875
ravimaktiivne:suguN    -9.5548
```

$\mu$=180,0365

$\alpha_{platseebo} = 0$ $\alpha_{aktiivne} = -0{,}4365$

$\beta_{mees}=0$ $\beta_{naine} = 5{,}6875$

$\alpha\beta_{platseebo,\ mees} = 0$ $\alpha\beta_{platseebo,\ naine} = 0$

$\alpha\beta_{aktiivne,\ mees} = 0$ $\alpha\beta_{aktiivne,\ mees} = -9{,}5548$

Ravita meeste keskmine: 180.04+0+0+0 = 180,04
Ravitud meeste keskmine: 180.04-0,4365+0+0 = 179,6
Ravita naiste keskmine: 180,0365+0+5,6875+0 = 185,72
Ravitud naiste keskmine: 180,0365-0,4365+5,6875-9,5548 = 175.7327

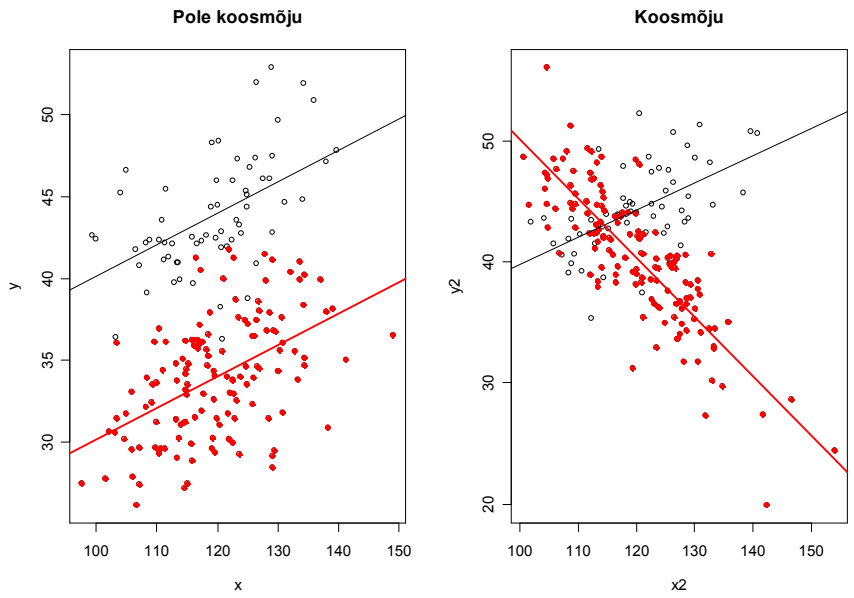
**Faktortunnuse koosmõjud pideva tunnusega**

Näide:

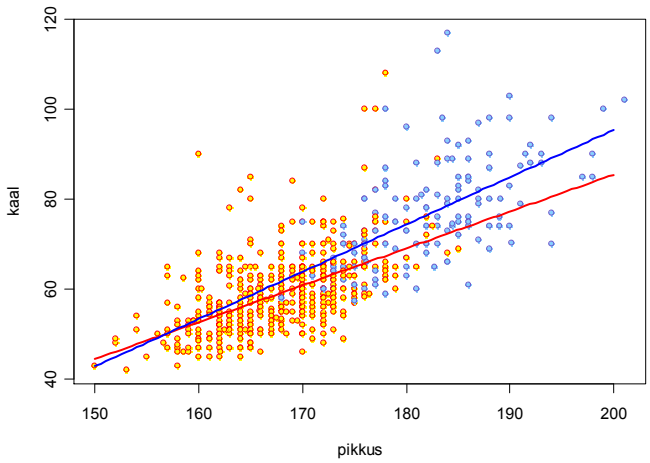
vanuse ja soo koosmõju

tõlgendus 1:
vanus mõjub naistele ja meestele erinevalt;

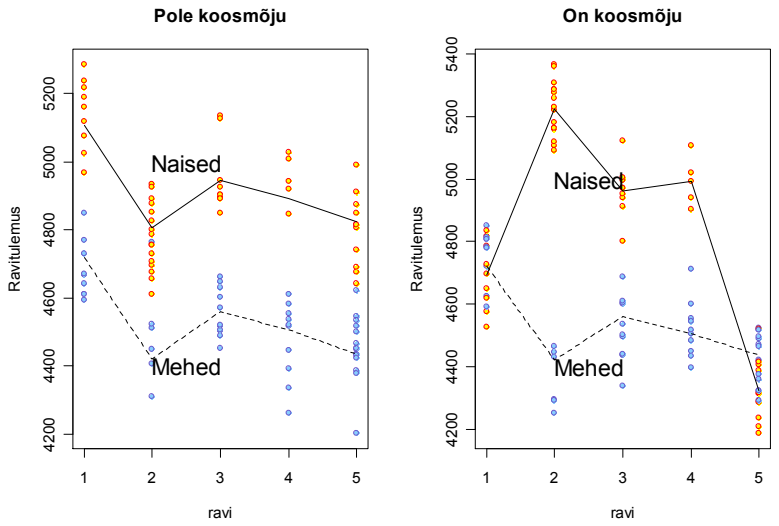
tõlgendus 2:
naiste ja meeste erinevus muutub vanuse kasvades;




Kaalu prognoosime pikkuse ja soo abil. Koosmõjudega mudel



## Faktortunnuse koosmõjud teise faktortunnusega




## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 1 (aditiivne)

vabaliige 2

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 2 |
| A+ | 2 | 2 |

$EY = 2$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 1 (aditiivne)

| | |
|---|---|
| vabaliige | 2 |
| B+ | 2 |

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 4 |
| A+ | 2 | 4 |

$EY = 2 + 2I_{B+}$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 1 (aditiivne)

| | |
|---|---|
| vabaliige | 2 |
| B+ | 2 |
| A+ | 4 |

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 4 |
| A+ | | |

$EY = 2 + 2I_{B+} + 4I_{A+}$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 1 (aditiivne)

| | |
|---|---|
| vabaliige | 2 |
| B+ | 2 |
| A+ | 4 |

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 4 |
| A+ | 6 | **8** |

$$EY = 2 + 2I_{B+} + 4I_{A+}$$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 1 (aditiivne)

| | |
|---|---|
| vabaliige: | 2 |
| B+ | 2 |
| A+ | 4 |
| A+ : B+ | 4 |

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 4 |
| A+ | 6 | 8+4 |

$$EY = 2 + 2I_{B+} + 4I_{A+} + 4I_{A+B+}$$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 1. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 12 |

Mudel 2 (multiplikatiivne)

| | |
|---|---|
| konstant | 2 |
| B+ | 2 |
| A+ | 3 |

Maailm mudeli järgi

| | B- | B+ |
|---|---|---|
| A- | 2 | 2·2=4 |
| A+ | 2·3=6 | 2·2·3=12 |

$$EY = 2 \cdot 2^{I(B+)} \cdot 3^{I(A+)}$$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

- Lineaarse mudeli korral on reaalsuse kirjeldamiseks tarvis koosmõju;
- Multiplikatiivse mudeli korral pole reaalsuse korrektseks kirjeldamiseks koosmõju vajalik…

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

Näide 2. Tinglikud keskväärtused

| | ei saa ravimit B (B-) | saab ravimit B (B+) |
|---|---|---|
| ei saa ravimit A (A-) | 2 | 4 |
| saab ravimit A (A+) | 6 | 8 |

Mudel 1 (lineaarne – ilma koosmõjuta)

$$EY = 2 + 2I_{B+} + 4I_{A+}$$

Mudel 2 (multiplikatiivne - koosmõjuga)

$$EY = 2 \cdot 2^{I(B+)} \cdot 3^{I(A+)} \cdot (8/12)^{I(A+,\ B+)}$$

## Koosmõju olemasolu või puudumine sõltub valitud mudeli matemaatilisest kujust!

- Nüüd saab lineaarne mudel hakkama ilma koosmõjudeta…
- Kuid multiplikatiivne mudel vajab koosmõjude abi…

**Koosmõju olemasolu või puudumine sõltub valitud matemaatilise mudeli kujust!**

Biomeetria bioloogidele
5. loeng

osa II

# Mudeli valikust

## Kas eesmärgiks on

a)uuritava tunnuse väärtuseid võimalikult täpselt prognoosiv mudel?

*või*

b)uuritavat tunnust mõjutavate tegurite võimalikult täpne kirjeldamine (põhjusliku mudeli loomine)?

# Kui eesmärgiks on täpsed prognoosid, siis…

- Ära kasuta tunnuseid, mille väärtuseid sa prognoosimise hetkel ei tea!

Tarvis on prognoosida kurjategija pikkust, kurjategija kohta teame vaid tema jalajälje mõõtmeid.

Kasutuks osutuvad kõik mudelid, mis prognoosivad inimese pikkust tema kehakaalu, silmavärvi või naeratuse laiuse järgi – me lihtsalt ei tea nende tunnuste väärtuseid sellel inimesel, kelle jaoks me soovime leida prognoosi.

Kasulikud on vaid mudelid, mis prognoosimiseks kasutavad jalajäljest saadavat informatsiooni – näiteks jalanumbrit.

# Kui eesmärgiks on täpsed prognoosid, siis…

Järelejäänuid tunnuseid kasutades koosta mudel või mudelid. Paremaks pea mudelit, millele vastav AIC-väärtus on väiksem (vahel võid ka kõhutunnet – arvutile mitteteadaolevat lisainformatsiooni – kasutada).

NB! Veendu, et mõlemad võrreldavad mudelid on hinnatud samasid vaatluseid (uuritavaid) kasutades!

```
> mudel1=lm(pikkus~factor(sugu))
> mudel2=lm(pikkus~factor(sugu)+kaal)

> AIC(mudel1)
[1] 4249.781
> AIC(mudel2)
[1] 3973.655
> AIC(mudel1, mudel2)
       df      AIC
mudel1  3 4249.781
mudel2  4 3973.655
Warning message:
In AIC.default(mudel1, mudel2) :
  models are not all fitted to the same number of observations

> mudel1=lm(pikkus~factor(sugu), data=tudengid[!is.na(kaal),])
> AIC(mudel1, mudel2)
       df      AIC
mudel1  3 4222.667
mudel2  4 3973.655
```

# Võid kasutada ka sammregressiooni abi

Näide sammregressioonist

```
m1=lm(SVR~vanus+pikkus+kaal+factor(olu)+factor(sugu)+factor(sport)+factor(viirus))
m2=step(m1)
summary(m2)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   99.52868    3.44233  28.913  < 2e-16 ***
kaal           0.25166    0.05663   4.444 1.12e-05 ***
factor(sugu)2  9.36896    1.59482   5.875 8.41e-09 ***
```

Kui kontrollitavaid argumenttunnuseid on vähe, võib sammregressioonil „lihtsustada“ lastavasse mudelisse kohe ka koosmõjud sisse panna. Rohkemate tunnuste puhul võib aga lasta sammregressioonil välja otsida kõige tähtsamad peamõjud ja alles siis hakata kontrollima, kas nende tähtsate peamõjude omavahelised koosmõjud tuleks ka mudelisse lisada või mitte.

# Mudelid põhjuslikele mõjudele