Meenutuseks: dispersioonanalüüs

```
> mudel=lm(y~factor(grupp))
> summary(mudel)

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.12598    0.08045   1.566   0.1177
factor(grupp)B -0.15165    0.11786  -1.287   0.1985
factor(grupp)C -0.15047    0.11624  -1.294   0.1958
factor(grupp)D -0.13851    0.11602  -1.194   0.2328
factor(grupp)E -0.15785    0.11517  -1.371   0.1708
factor(grupp)F -0.23271    0.11397  -2.042   0.0414 *
factor(grupp)G -0.13462    0.11249  -1.197   0.2317
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.982 on 993 degrees of freedom
Multiple R-squared: 0.004437,   Adjusted R-squared: -0.001579
F-statistic: 0.7376 on 6 and 993 DF,  p-value: 0.6194
```


Mitmene testimine

Meil on 100 „kahtlusalust“ geeni, mis võivad olla seotud meid huvitava fenotüübiga, Ja siis, milles on probleem? Kogu lihtsalt andmed ja hakka testima geenide seotust mõõdetud tunnuste/näitajatega, kasutades olulisuse nivood 0,05?

No nii ei lähe mitte...
Kujuta hetkeks ette, et ükski neist sajast geenist ei seostu meid huvitava näitajaga. Mitu statistilist testi siis „leiavad“ statistiliselt olulise tulemuse (mitu valepositiivset, eksitavat vastust me saame)?

Milline on oodatav valepositiivsete tulemuste arv?
Vastus: 5!

Mis on tõenäosus saada vähemalt ühte (või enamat) valepositiivset tulemust (100 testi korral)?

$1-(1-0{,}05)^{100} = 0{,}994$

Seega testides 100 geeni leiame peaaegu kindlalt mõne, mis justnagu näiks olevat seotud meid huvitava näitajaga!

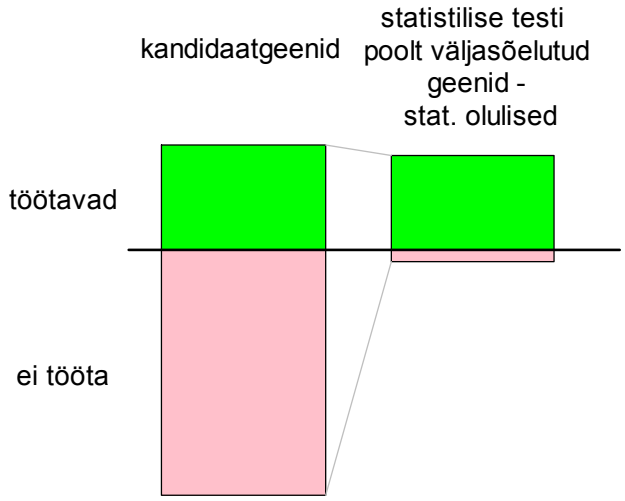
Kus kerkib esile mitmese testimise probleem?

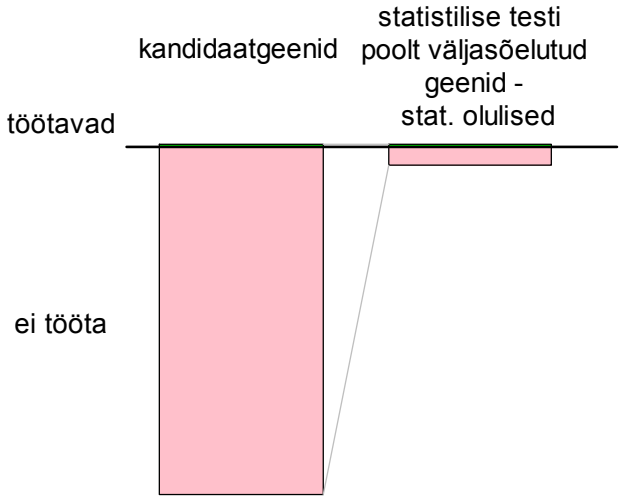
„Pühendunud teadlane“ kordab oma eksperimenti 50 korda saamaks soovitud andmeid;

„Teaduslik“ ajakiri eelistab avaldada statistiliselt olulisi tulemusi (pisikesi p-väärtuseid) sisaldavaid artikleid (nn *publication bias*);

Nõutu ja ideedetu teadlane teeb lihtsalt 100 ekserimenti – “miski võib ju välja tulla”;

Hoolikas doktorant kogub veidi andmeid ja katsetab siis, kas õnnestub tõestada teda huvitavat hüpoteesi. Kui alternatiivse hüpoteesi tõestamine ebaõnnestub, siis teeb ta täiendavalt mõned mõõtmised. Ikka veel pole võimalik saada statistiliselt olulist tulemust? Pole probleemi! Lisame veel mõned vaatlused ja testime uuesti!





Mitmene testimine – lahendus I – Bonferroni meetod

Tehes 10 testi, kui kindel võid olla, et nähtud statistiliselt olulised tulemused ikka tegelikult kehtivad? Kui teed kümme testi (olulisuse nivool 0,05) siis ka juhul, kui kõigil kümnel juhul oli õige nullhüpotees (seost tunnuste vahel polnud), „tõestavad“ üks või enam testi alternatiivse hüpoteesi tõenäosusega 0,4. Seda on palju enam, kui 0,05, mida enamasti peetakse vastuvõetavaks. Mida teha?

Bonferroni' meetod

Väga lihtsa lahenduse pakkus välja Bonferroni (1936). Nimelt soovitas ta kasutada üksiktesti tegemisel rangemat olulisusnivood – üksiktesti tegemisel peaksime kasutama olulisuse nivood, mille saame, kui jagame soovitava olulisuse nivoo (näiteks α=0,05) sooritatud testide arvuga (k):

$$\alpha_\beta = \alpha / k$$

Tehes 10 testi, peaksime iga üksiktesti tegemisel kasutama olulisuse nivood $\alpha_\beta = 0.05 / 10 = 0.005$. Seega juhul, kui kasutame iga üksiktesti tegemisel olulisuse nivood 0.005 võime olla kindlad, et vale testitulemuse – ekslikult tõestatud alternatiivse hüpoteesi – saamise tõenäosus 10 testi peale kokku on 5% või väiksem.

Bonferroni meetodi modifikatsioonid

Järjestikusel testimisel võime „kulutada“ oma olulisuse nivood ebaühtlaselt. Näiteks võime teha esimese statistilise testi peale 10 vaatluse tegemist, kuid kasutame siis olulisuse tõenäosust 0,001 – kui vastus on ilmne (olulisuse nivool 0,001), siis lõpetame andmete edasise kogumise ja loeme H1 tõestatuks. Kui peale 10 vaatluse tegemist pidime jääma nullhüpoteesi juurde, siis võime teha veel 100 vaatlust ja testida siis (kasutades olulisuse nivood 0,004), kas võime juba andmete kogumise lõpetada. Kui ei õnnestunud ikka nullhüpoteesi kummutada, võime koguda veel täiendavalt andmeid 400 objekti kohta ja teha siis viimane statistiline test olulisuse nivool 0,045. Sellise testprotseduuri korral (kus võime teha kuni kolm testi) ei kerki summarne I-liiki vea tegemise tõenäosus suuremaks kui 0,05, sest

$$0{,}001+0{,}004+0{,}045=0{,}05$$

Bonferroni-Holmi meetod

Sorteeri kõik p-value'd mis saad tehes k testi: $p_{(1)} \le p_{(2)} \le \ldots \le p_{(k)}$. Kui $p_{(1)} > 0{,}05/k$ siis peatu ja jää kõigi kontrollitud hüpoteeside puhul H_0 juurde. Kui $p_{(1)} \le 0{,}05/k$, siis loe $p_{(1)}$ seonduva testi puhul H_1 tõestatuks ja jätka. Kas $p_{(2)} > 0{,}05/(k\text{-}1)$? Kui jah, siis jää ülejäänud k-1 testi puhul nullhüpoteesi juurde. Kui $p_{(2)} \le 0{,}05/(k\text{-}1)$ siis loe H1 tõestatuks testi jaoks, mille olulisustõenäosus oli $p_{(2)}$ ja jätka…

Mitmene testimine R-is

```
> p.adjust(c(0.01,0.04,0.2,0.001, 0.5), "bonferroni")
[1] 0.050 0.200 1.000 0.005 1.000

> p.adjust(c(0.01,0.04,0.2,0.001, 0.5), "holm")
[1] 0.040 0.120 0.400 0.005 0.500

> p.adjust(c(0.01,0.04,0.2,0.001, 0.5), "fdr")
[1] 0.02500000 0.06666667 0.25000000 0.00500000 0.50000000
```

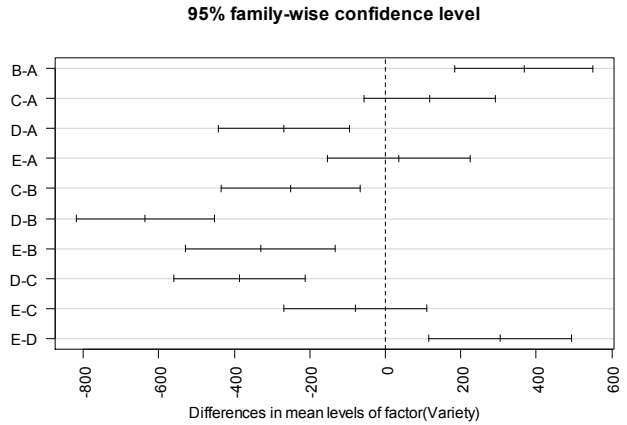
Viimane meetod – False Discovery Rate – võimaldab määrata, kui suur võib olla ekslikult „avastatud“ või „tõestatud“ H1-tede arv kõigi statistiliselt oluliseks loetud tulemuste seas.

Mitmene testimine ja ANOVA. Kui igal faktori tasemel on tehtud samapalju (või peaaegu samapalju) vaatluseid...

```
> TukeyHSD(aov(lm(SAAK~factor(SORT))))
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = lm(SAAK ~ factor(SORT)))

$`factor(Variety)`
          diff        lwr        upr     p adj
B-A  367.43434  184.41977  550.44892 0.0000083
C-A  116.81818  -56.80469  290.44105 0.3260638
D-A -268.72727 -442.35014  -95.10440 0.0006073
E-A   36.42045 -152.78068  225.62159 0.9817649
C-B -250.61616 -433.63074  -67.60159 0.0028804
D-B -636.16162 -819.17619 -453.14704 0.0000000
E-B -331.01389 -528.86864 -133.15914 0.0001945
D-C -385.54545 -559.16832 -211.92259 0.0000011
E-C  -80.39773 -269.59886  108.80341 0.7471023
E-D  305.14773  115.94659  494.34886 0.0003375
```



```
> plot(TukeyHSD(aov(lm(Yield~factor(Variety)))), las=2)
```



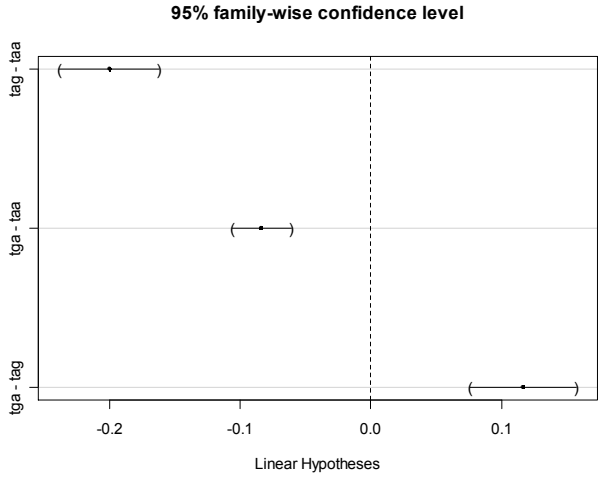
Lineaarsed mudelid (ja üldistatud lineaarsed mudelid, glm)

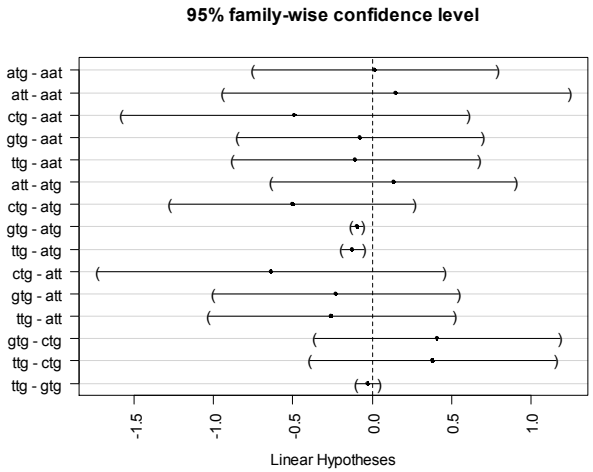
Sobib kasutada lisamoodulit multcomp

```
> library(multcomp)
> m1=lm(log(cai)~ factor(last_codon))
> a=glht(m1, linfct=mcp("factor(last_codon)"="Tukey"))
> summary(a)

         Simultaneous Tests for General Linear Hypotheses
Multiple Comparisons of Means: Tukey Contrasts
Fit: aov(formula = lm(log(cai) ~ last_codon))

Linear Hypotheses:
               Estimate Std. Error t value p value
tag - taa == 0 -0.200617   0.016548 -12.124  <1e-10 ***
tga - taa == 0 -0.083736   0.009632  -8.694  <1e-10 ***
tga - tag == 0  0.116881   0.017542   6.663  <1e-10 ***
 (Adjusted p values reported)
```



```
> confint(a)
> plot(confint(a))
```






```
library(multcomp); sortF=factor(sort);
summary(simint(saak~sortF, whichf="sortF", type="Tukey"))

Coeficients:
              Estimate    2.5 %   97.5 % t value Std.Err. p raw p Bonf p adj
sortFB-sortFA  367.434  181.576  553.293   5.615    65.436 0.000  0.000 0.000
sortFC-sortFA  170.045  -22.096  362.187   2.514    67.647 0.016  0.156 0.105
sortFD-sortFA -216.455 -392.776  -40.133  -3.487    62.078 0.001  0.011 0.009
sortFE-sortFA   33.000 -143.321  209.321   0.532    62.078 0.598  1.000 0.984
sortFC-sortFB -197.389 -398.319    3.541  -2.790    70.742 0.008  0.077 0.056
sortFD-sortFB -583.889 -769.748 -398.030  -8.923    65.436 0.000  0.000 0.000
sortFE-sortFB -334.434 -520.293 -148.576  -5.111    65.436 0.000  0.000 0.000
sortFD-sortFC -386.500 -578.642 -194.358  -5.713    67.647 0.000  0.000 0.000
sortFE-sortFC -137.045 -329.187   55.096  -2.026    67.647 0.049  0.487 0.270
sortFE-sortFD  249.455   73.133  425.776   4.018    62.078 0.000  0.002 0.002
```


Dispersioonanalüüsi eeldused

Dispersioonanalüüsi mudeli hindamisel ja hüpoteeside testimisel tehakse samu eelduseid, mis regressioonmudeli hindamisel ja testimisel. Seega:

Mudeli jäägid peavad olema normaaljaotusega (muidu arvutab arvuti olulisustõenäosused ja usaldusintervallid valesti välja);

Uuritava tunnuse hajuvus peab iga faktortunnuse taseme korral olema samasuur (testid, usaldusintervallid muidu valed)

Valim esindav, sisestusvigu pole.

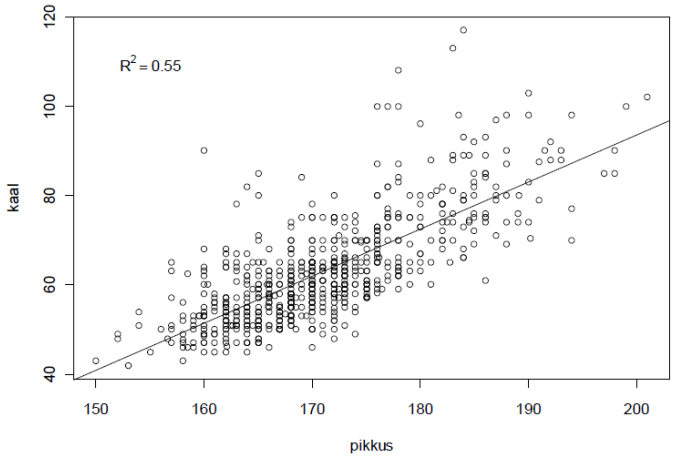
Kui eeldused pole täidetud, võib proovida samu lahendusteid, mida regressioonanalüüsigi korral – näiteks uuritava tunnuse transformeerimist.

Determinatsioonikordajaga manipuleerimine

Determinatsioonikordaja on eelkõige interpreteeritav siis, kui uuritav valim on tõepoolest juhuslik valim mingist populatsioonist. Kui uuritavad andmed on kogutud eksperimenteerides (olukorras, kus me ise otsustame, millised saavad olema *X*-tunnuse väärtused) on determinatsioonikordaja R^2 väärtus teatavates piirides eksperimentaatori enda valida/otsustada.

Vaatame paari näidet.

Tudengite kaalu prognoosimine pikkuse järgi. Valim - R^2=0,55.



Kuidas suurendada või vähendada determinatsioonikordajat?

$$KAAL = c0 + c1\ PIKKUS + e.$$

Determinatsioonikordaja

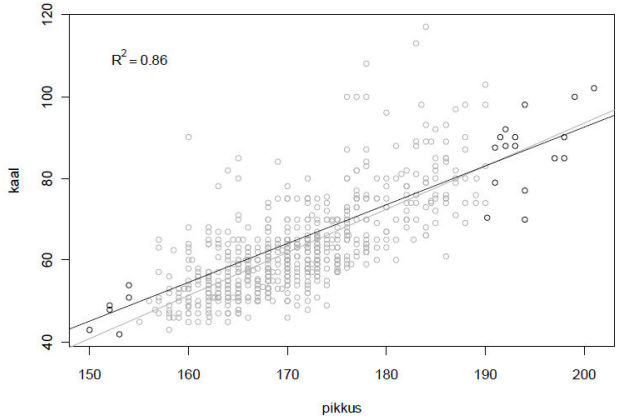
$$R^2 = 1 - D(e)/D(KAAL).$$

Vaja oleks kas muuta D(e) või D($KAAL$) väärtust. Antud näites on lihtsam muuta D($KAAL$) väärtust:

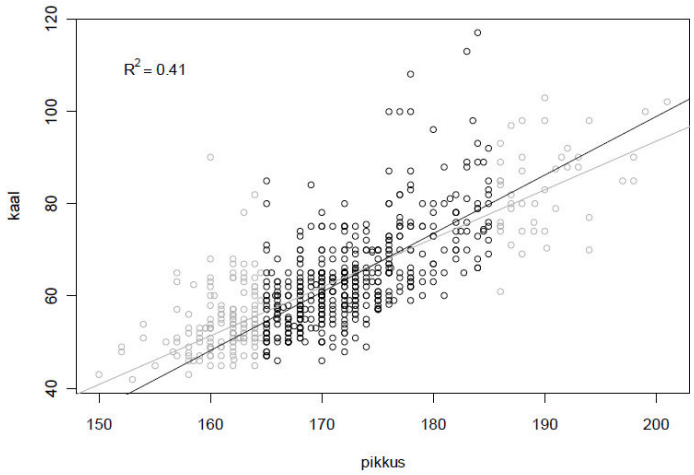
$$D(KAAL) = c1^2\ D(PIKKUS) + D(e)$$

Valides uuringusse väga erinevate pikkustega tudengeid saame suure R^2 väärtuse, valides sarnase pikkusega tudengeid saame väikese R^2 väärtuse.

Tudengite kaal ja pikkus –tudengeid, kelle pikkus<155cm või pikkus >190cm



Tudengite kaal ja pikkus – tudengeid, kelle 170<pikkus<180cm



Determinatsioonikordaja kasutamisel on mõtet, kui tegemist on juhusliku valimiga uuritavast populatsioonist. Determinatsioonikordaja kasutamine (interpretatsioon) on tugevalt kaheldav, kui tegemist on planeeritud katsega või kui uuritavad on valitud uurija isikliku suva järgi.

Determinatsioonikordajate võrdlemisel on mõtet samuti eelkõige siis, kui tegemist on sama populatsiooni käsitlevate uuringutega (NB! eestlased ja 20-65a. vanused eestlased on erinevad populatsioonid!!!)

Mitmene regressioonanalüüs, kovariatsioonanalüüs

Koosmõjud, multikollineaarsus, mudeli valik

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 arvessevõtmine 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 – täpsemad prognoosid!

```
> summary(lm(y~x1))
[...]
Residual standard error: 22.08 on 998 degrees of freedom
Multiple R-Squared: 0.2109,     Adjusted R-squared: 0.2101

> summary(lm(y~x2))
 [...]
Residual standard error: 12.41 on 998 degrees of freedom
Multiple R-Squared: 0.751,      Adjusted R-squared: 0.7507

> summary(lm(y~x1+x2))
 [...]
Residual standard error: 5.078 on 997 degrees of freedom
Multiple R-Squared: 0.9583,     Adjusted R-squared: 0.9582
```

$\max(R_{X1}^2, R_{X2}^2) \leq R_{X1,X2}^2 \leq R_{X1}^2 + R_{X2}^2$

Mitmene regressioon – interpretatsioon – pidevad tunnused

```
> summary(lm(SVR~pikkus+kaal))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 50.74997   13.58055   3.737 0.000211 ***
pikkus       0.27696    0.09636   2.874 0.004251 **
kaal         0.30591    0.06921   4.420 1.25e-05 ***

Residual standard error: 11.62 on 432 degrees of freedom
Multiple R-Squared: 0.1925,     Adjusted R-squared: 0.1887
F-statistic: 51.48 on 2 and 432 DF,  p-value: < 2.2e-16
```

SVR = 50,75 + 0,277 pikkus + 0,306 kaal + jääk (prognoosiviga)

pikkus = 170cm, kaal = 60 kg:
SVR = 50,75 + 0,277*170 + 0,306*60 = 116,2 mmHg
pikkus = 171cm, kaal=60 kg:
SVR = 50,75 + 0,277*171 + 0,306*60 = 116,477 mmHg

Kahe sama kaaluga tudengi korral on 1cm võrra pikemal tudengil keskmiselt 0,277 mmHg kõrgem vererõhk.

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
```

10 cm pikem mees kaalub lühemast mehest keskmiselt 8,9 kg rohkem.
10 cm pikem naine kaalub lühemast naisest keskmiselt 8,9 kg rohkem.

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

Naise kaal = -90,47 + 0,89125*Pikkus + prognoosiviga
Mehe kaal = -90,47 + 4,818 + 0,89125*Pikkus + prognoosiviga

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 arvatavasti 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:

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

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

Multikollineaarsus

Multikollineaarsus – argumenttunnuste kõrge korreleeritus (sageli valitakse kriitiliseks piiriks 0,7 – kui argumenttunnuste korrelatsioon on suurem kui 0,7 – siis on tegemist multikollineaarsusega).

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

Multikollineaarsus muutub probleemiks, kui üritame saadud mudeli parametreid (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:

Taimede või loomade kasvuperioodil pikkus ≈ kaal ≈ vanus
Rääkides pikkuse mõjust ühele või teisele näitajale peame silmas ehk hoopis vanust?

Haridus ja sissetulek on sageli tugevalt seotud, seega võib hariduse lisamine mudelile sissetuleku mõju hoopistükkis tagurpidiseks pöörata;

Vanus ja sünniaasta käivad sageli käsikäes. Seega võib vanemate loomade tervis olla kehvem sellepärast, et vanasti neid halvasti hooldati, aga ka sellepärast, et nad on lihtsalt vanad.

Ilmastikunähtused võivad olla sageli omavahel tugevas seoses – kas saak tuli kehv sellepärast, et sadas palju, või sellepärast, et pilvede vahelt paistis vähe päikest, või sellepärast, et oli jahe suvi?

...

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 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

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
```

μ=180,0365

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

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

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

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

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

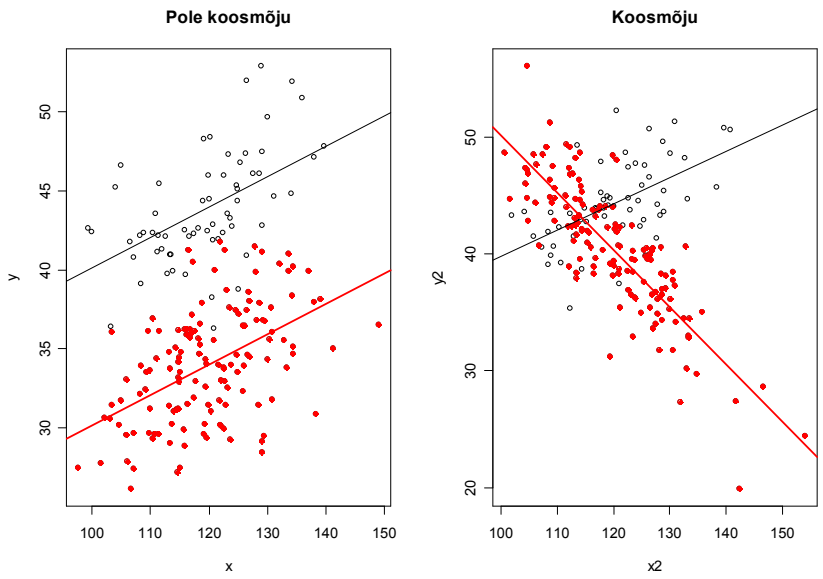
Faktortunnuse koosmõjud pideva tunnusega

Näide:

vanuse ja soo koosmõju

tõlgendus 1:
vanus mõjub naistele ja meestele erinevalt;

tõlgendus 2:
naiste ja meeste erinevus muutub vanuse kasvades;



Faktortunnuse koosmõjud teise faktortunnusega