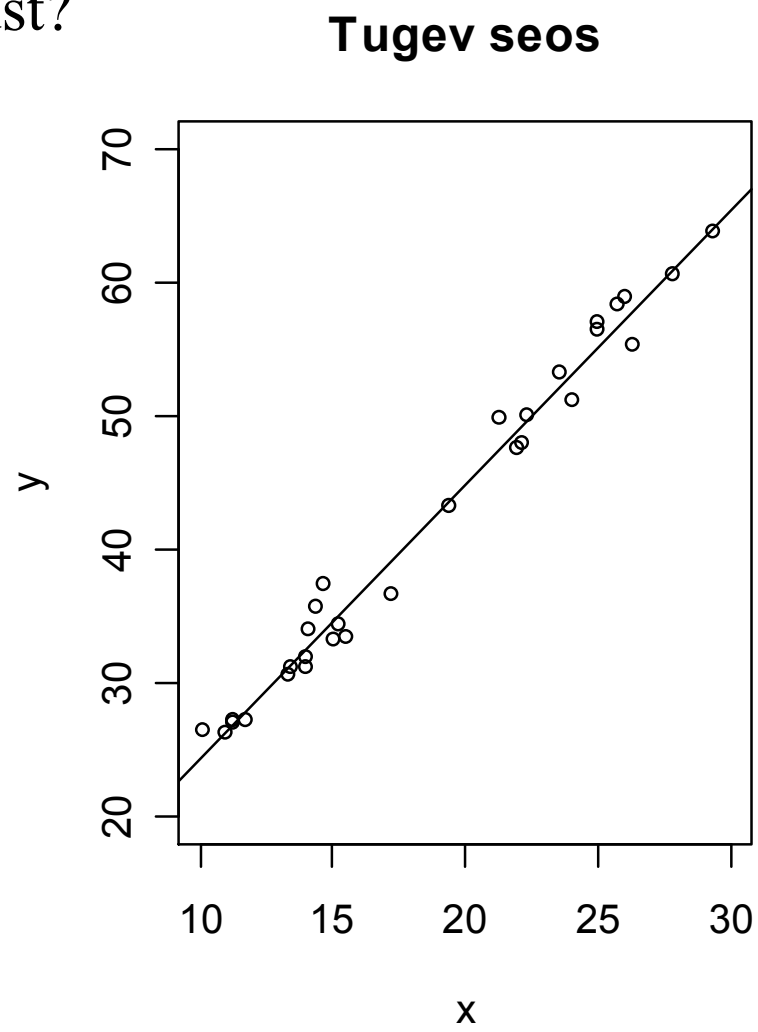
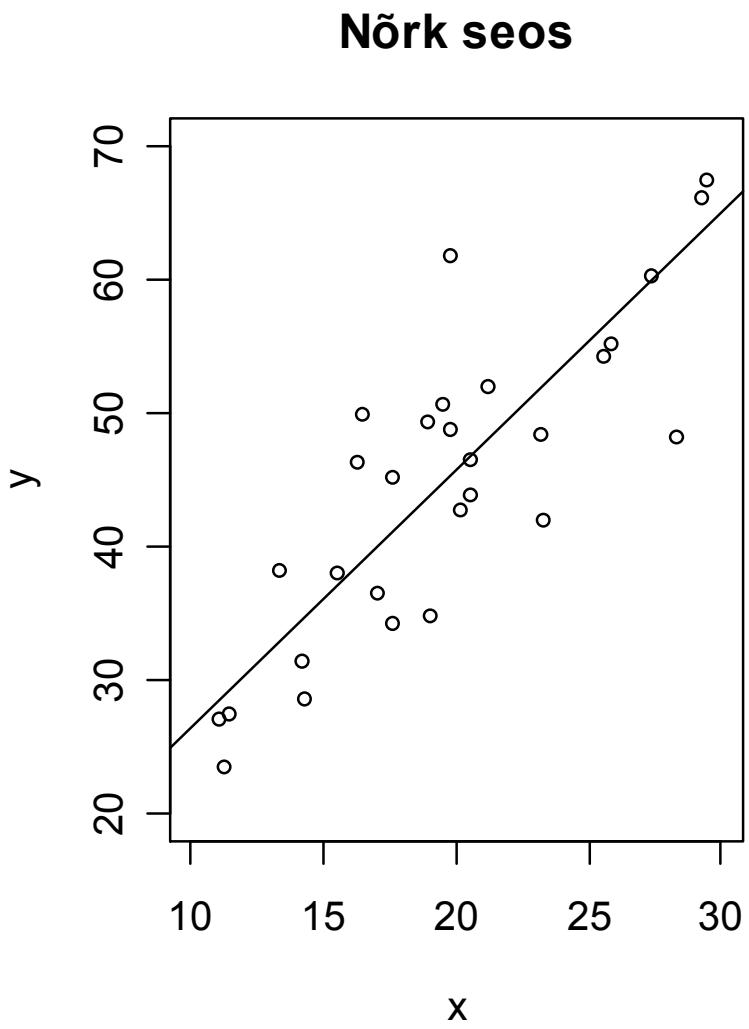
Seose tugevuse iseloomustamine

Mõnikord on üht tunnust teades võimalik väga suure täpsusega ära arvata teise tunnuse värtust, teinekord mitte. Kuidas mõõta seose tugevust?





Kuidas mõõta meie ennustuste headust - seose tugevust? Juhul, kui me poleks mõõtnud sõltumatu tunnuse (X) väärtuseid, oleks parim prognoos, mida me sõltuvale tunnusele (Y) anda suudaksime, sõltuva tunnuse väärtuste keskmine. Sellise "prognoosi" täpsust saab iseloomustada "ennustusvigade" dispersiooni abil ehk tunnuse Y dispersiooni kasutades (tähistame tunnuse Y dispersiooni tähega $D(Y)$). Kasutades sõltuva tunnuse (Y) prognoosimiseks sõltumatu tunnuse (X) väärtuseid, saame teha täpsemaid prognoose. Võime taas mõõta oma prognoosi täpsust, kasutades selleks mõõtmisvigade dispersiooni (tähistame saadava suuruse sümboliga $D(Y|X)$). Järelikult teades tunnuse X väärtust, on meil võimalik prognoosida tunnuse Y väärtust $D(Y)$-$D(Y|X)$ võrra täpsemalt (väiksema dispersiooniga). Prognoosi täpsuse suurenemise jagatist prognoositava tunnuse dispersiooniga kutsutakse determinatsioonikordajaks (R^2) ja seda kasutatakse sageli tunnustevahelise seose tugevuse mõõtmiseks:

$$R^2 = [\ D(Y) - D(Y|X)\ ] / D(Y)$$

Determinatsioonikordajat esitatakse vahel ka protsentides – näidates, mitu protsenti õnnestus tänu kasutatavale regressioonimudelile tõsta prognoosi täpsust.

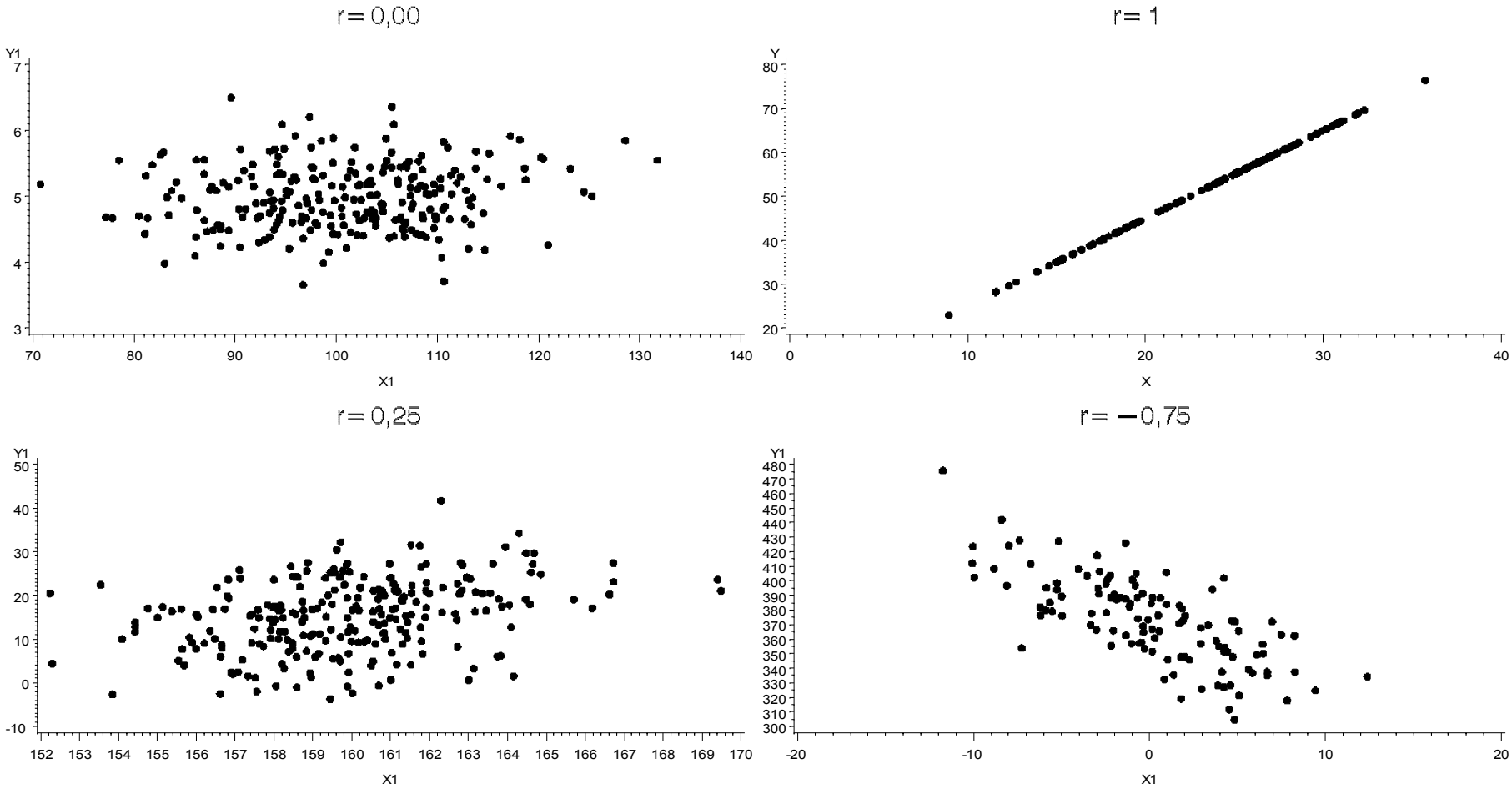
Lineaarne korrelatsioonikordaja

Arvatavasti kõige sagedamini kasutatakse kahe tunnuse vahelise seose tugevuse iseloomustamiseks lineaarset korrelatsioonikordajat (Pearsoni korrelatsioonikordajat). Korrelatsioonikordaja ruut on determinatsioonikordaja, kusjuures lineaarne korrelatsioonikordaja on positiivne, kui ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad samuti kasvama. Kui aga ühe tunnuse väärtuste kasvades teise tunnuse väärtused kipuvad kahanema siis on korrelatsioonikordaja negatiivne. Lineaarset korrelatsioonikordajat tähistatakse tähega r.

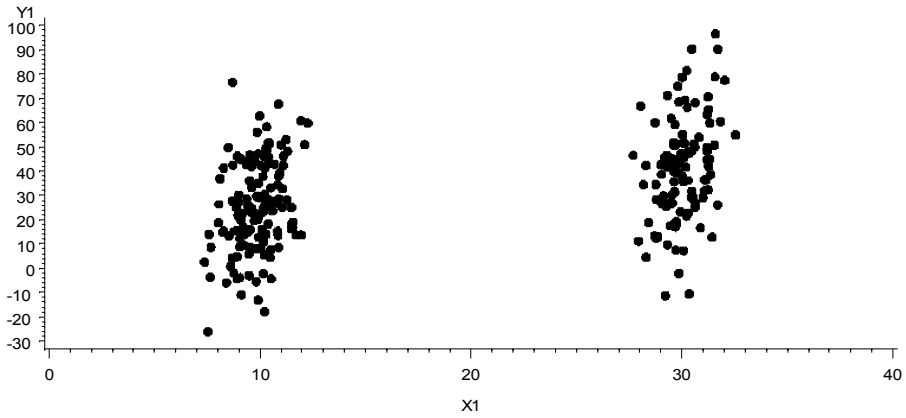
Pearsoni korrelatsioonikordaja omadused:

- Kui tunnuste X ja Y vahel on lineaarne funktsionaalne seos $Y=a+bX$ (ehk täpne lineaarne seos), siis on korrelatsioonikordaja väärtus kas 1 või -1 vastavalt kordaja b märgile.
- Kui $r>0$, siis ühe tunnuse suurenedes keskmiselt teine tunnus kasvab ja vastupidi - ühe vähenedes väheneb ka teine.

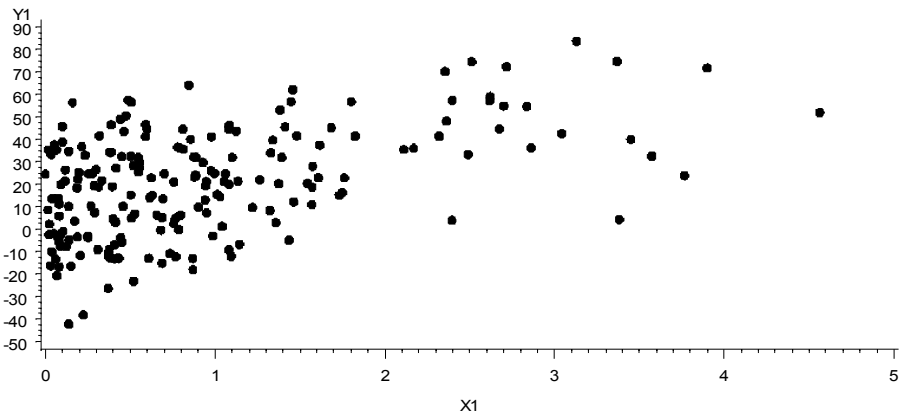
- Kui $r<0$, siis ühe tunnuse väärtuste suurenedes keskmiselt teise tunnuse väärtused kahanevad ja vastupidi - ühe kahanedes teine kasvab.
- Kui tunnused on lineaarselt sõltumatud (tunnuste vahel võib aga olla mittelineaarne sõltuvus), siis on korrelatsioonikordaja null $r = 0$.
- Korrelatsioonikordaja ruut r^2 ehk determinatsioonikordaja näitab, kui suur osa ühe tunnuse hajuvusest (dispersioonist) on kirjeldatud teise poolt.
- Mida suurem on korrelatsioonikordaja absoluutväärtus, seda tugevam on korrelatiivne seos tunnuste vahel.
- Mõõtühiku (lineaarne) vahetus ei muuda korrelatsioonikordaja suurust (Korrelatsioonikordaja ei muutu, kui mõõdame temperatuuri Celsiuse kraadide Co asemel Farenheitides Fo, samuti võime pikkust mõõta sentimeetrites või meetrites- korrelatsioonikordaja jääb samaks).


r= 0,00
Y1
X1
r= 1
Y
X
r= 0,25
Y1
X1
r= −0,75
Y1
X1

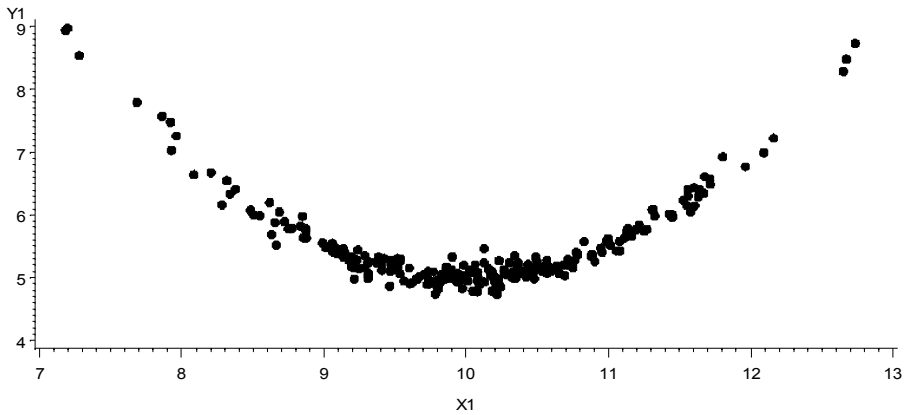
r= 0,4



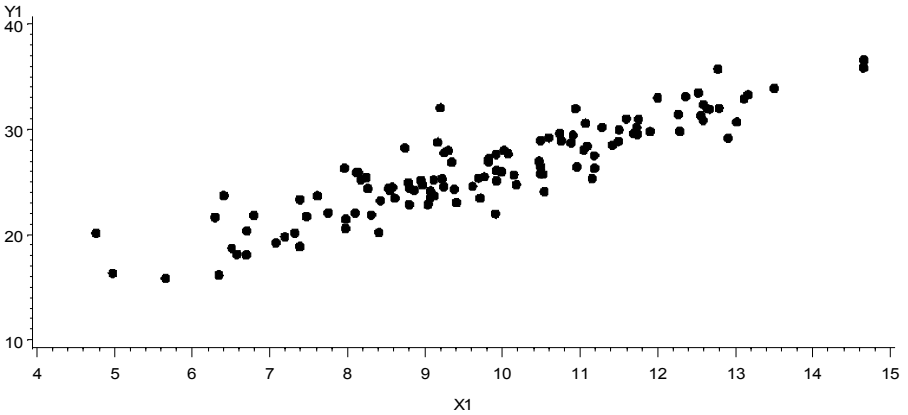
r= 0,50



r= 0,00



r= 0,90

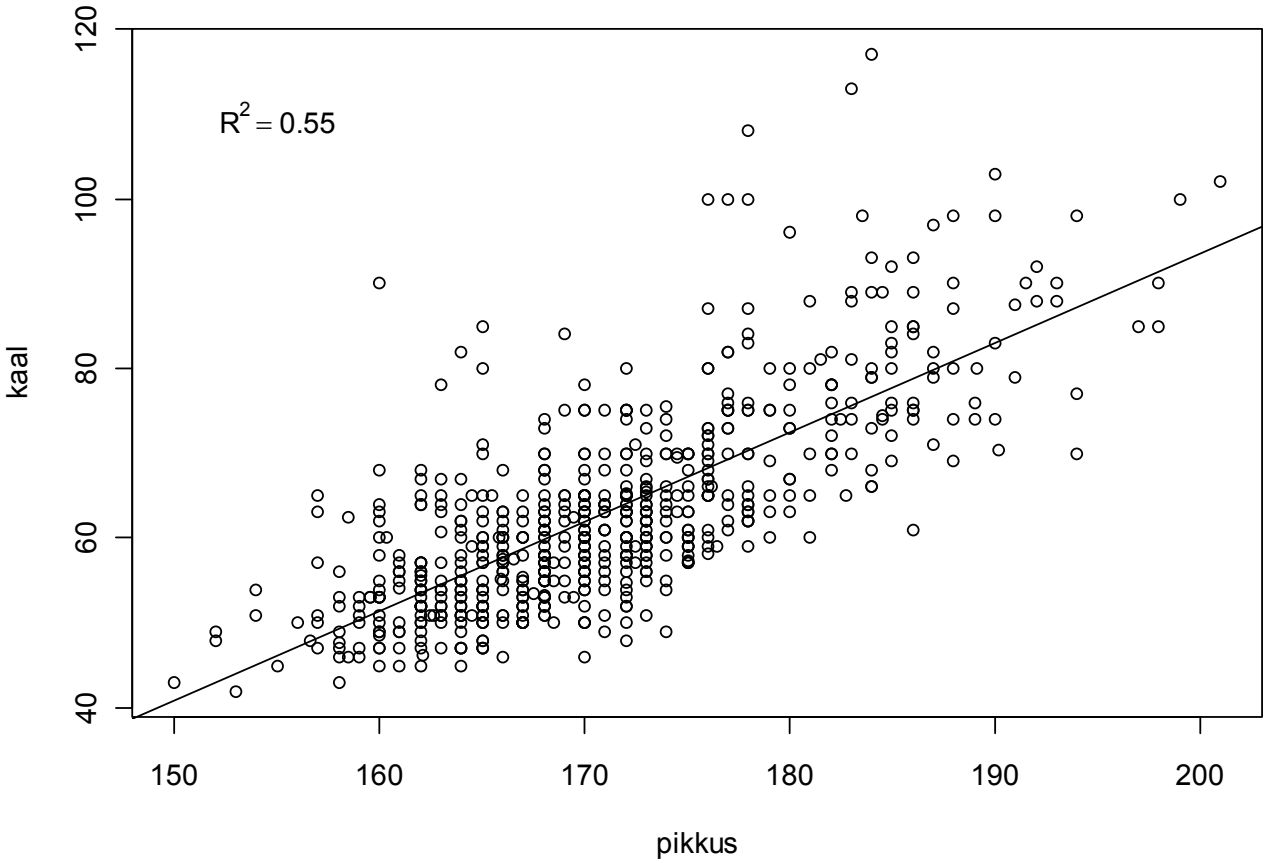


Determinatsioonikordajaga manipuleerimine

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 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

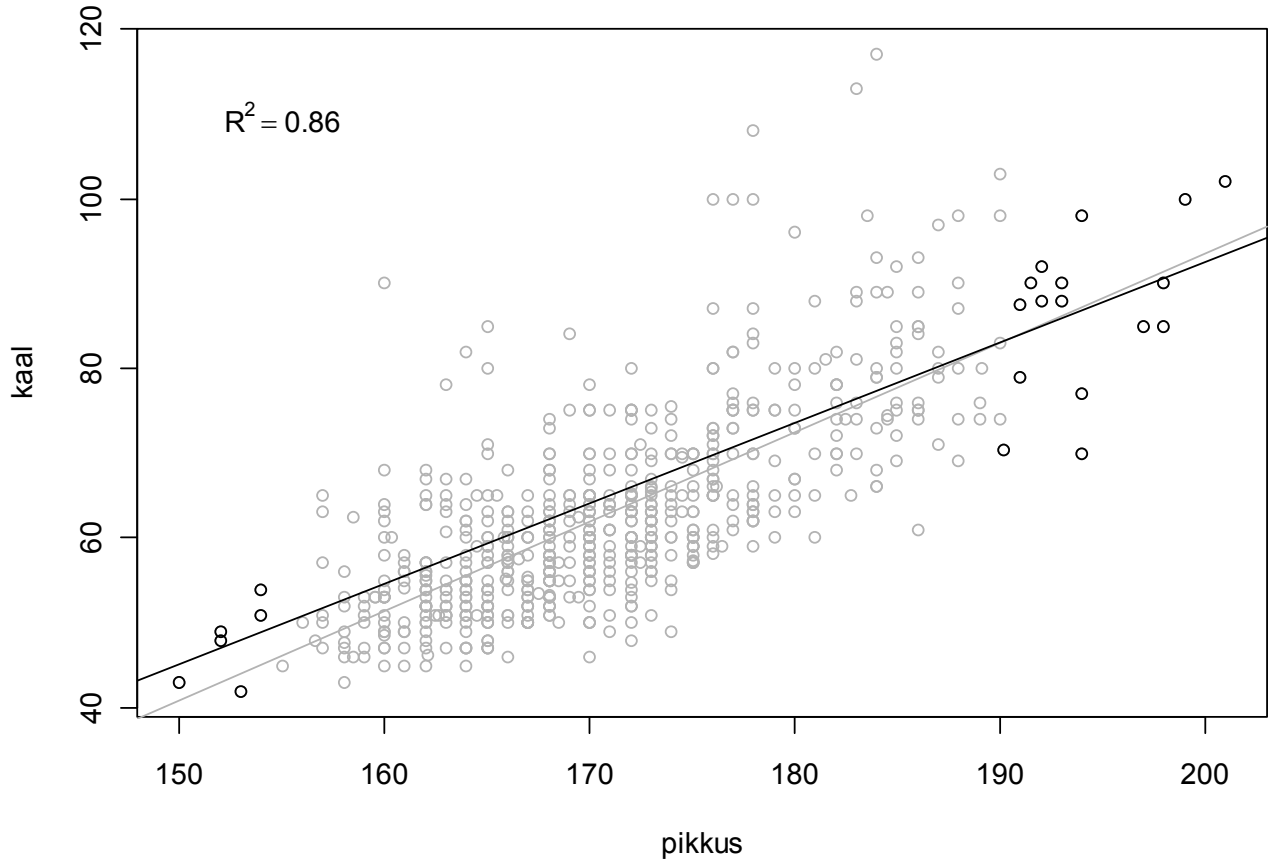
$$R^2 = 1 - D(e)/D(KAAL).$$

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

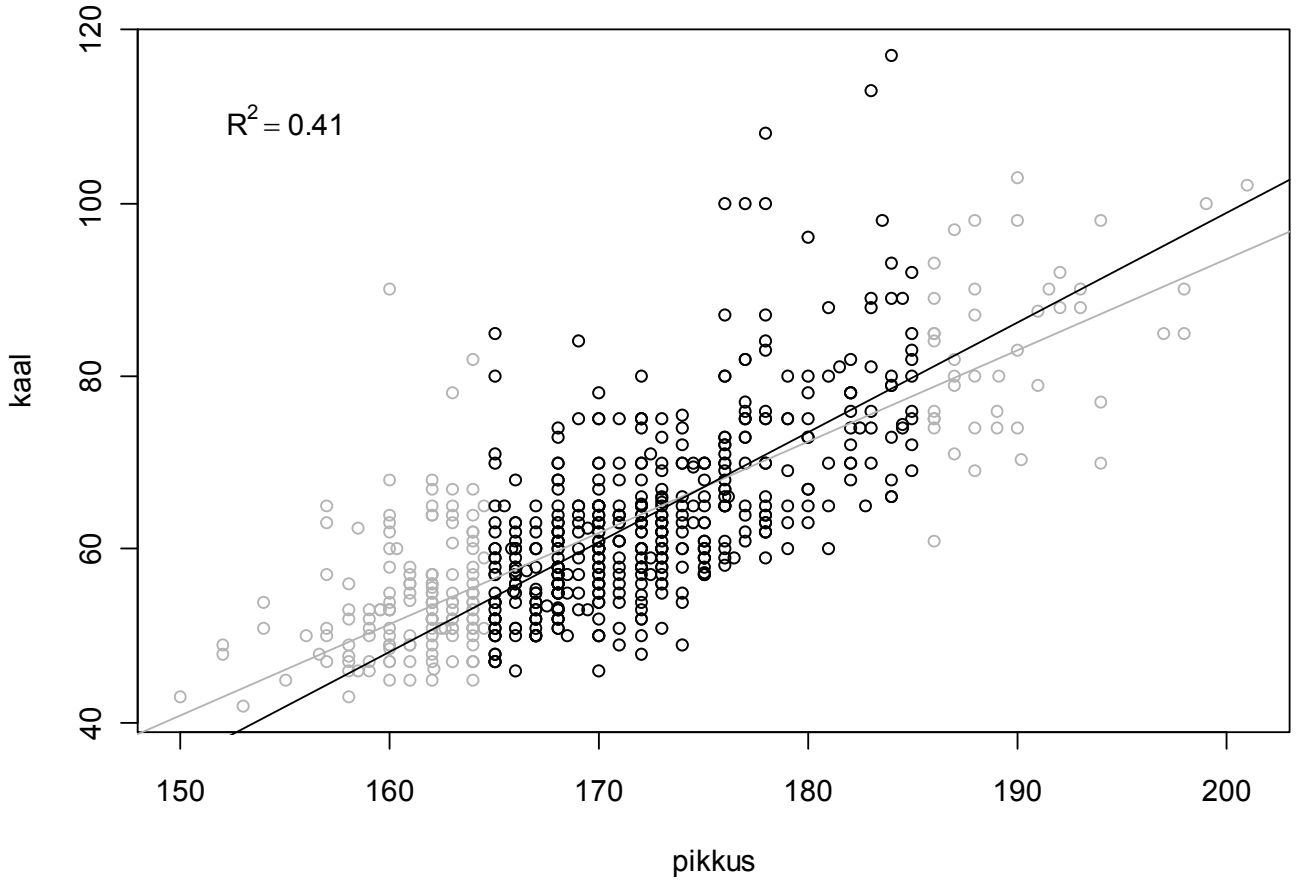
$$D(KAAL) = c1^2\ D(PIKKUS) + D(e)$$

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



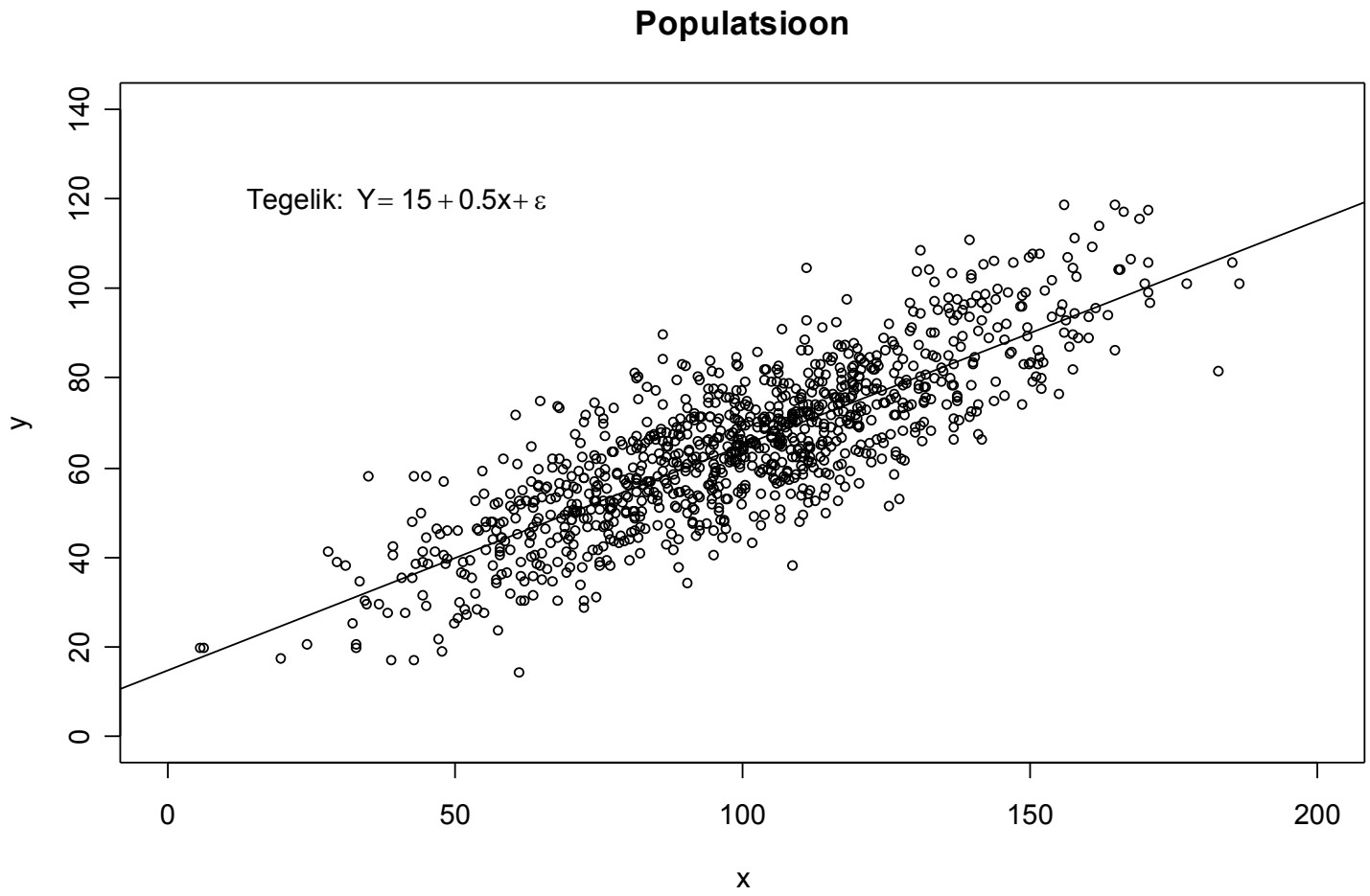
Tudengite kaal ja pikkus – tudengeid, kelle 170<pikkus<180cm



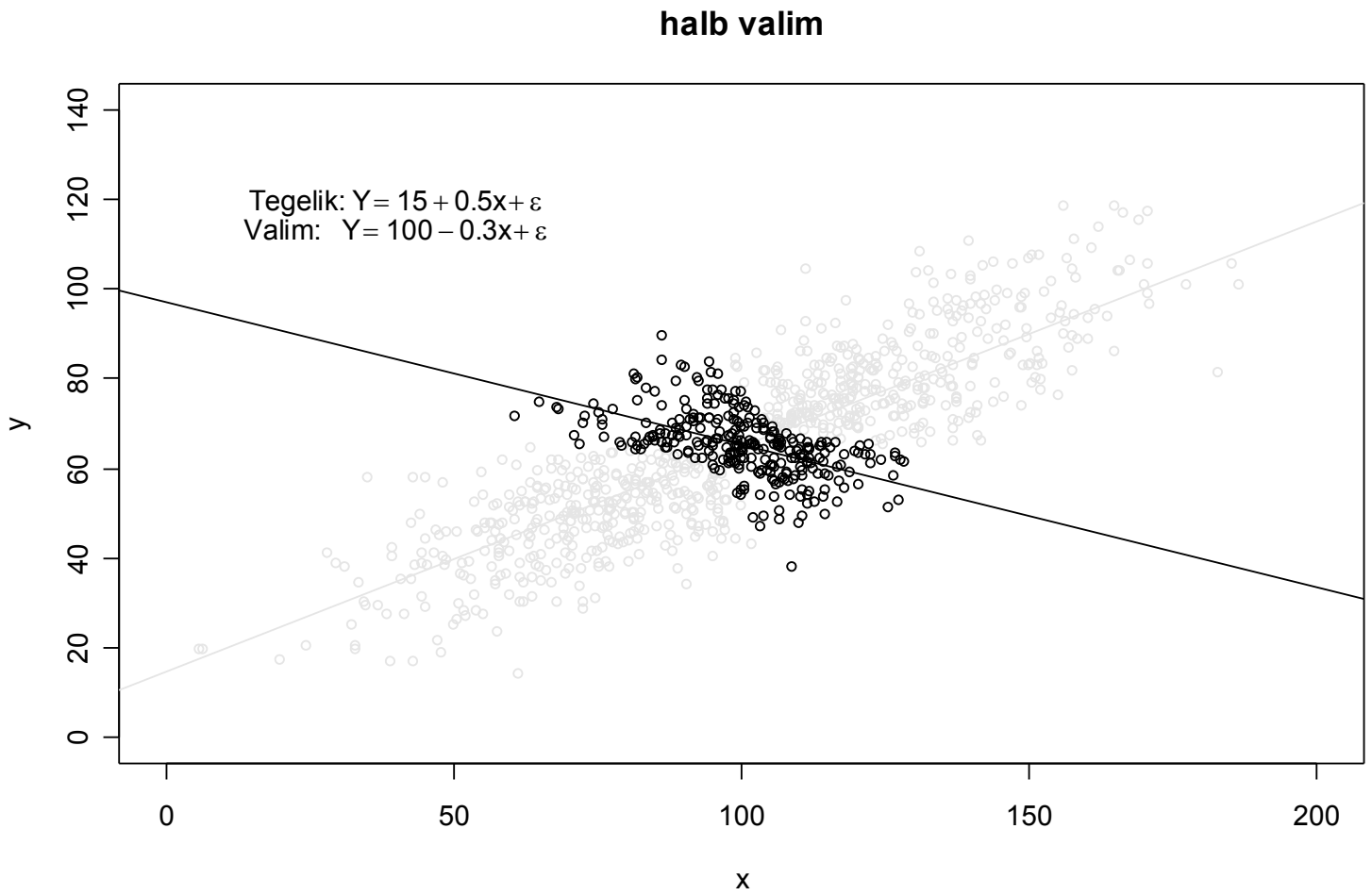
III osa

Lineaarse regressiooni eeldused.

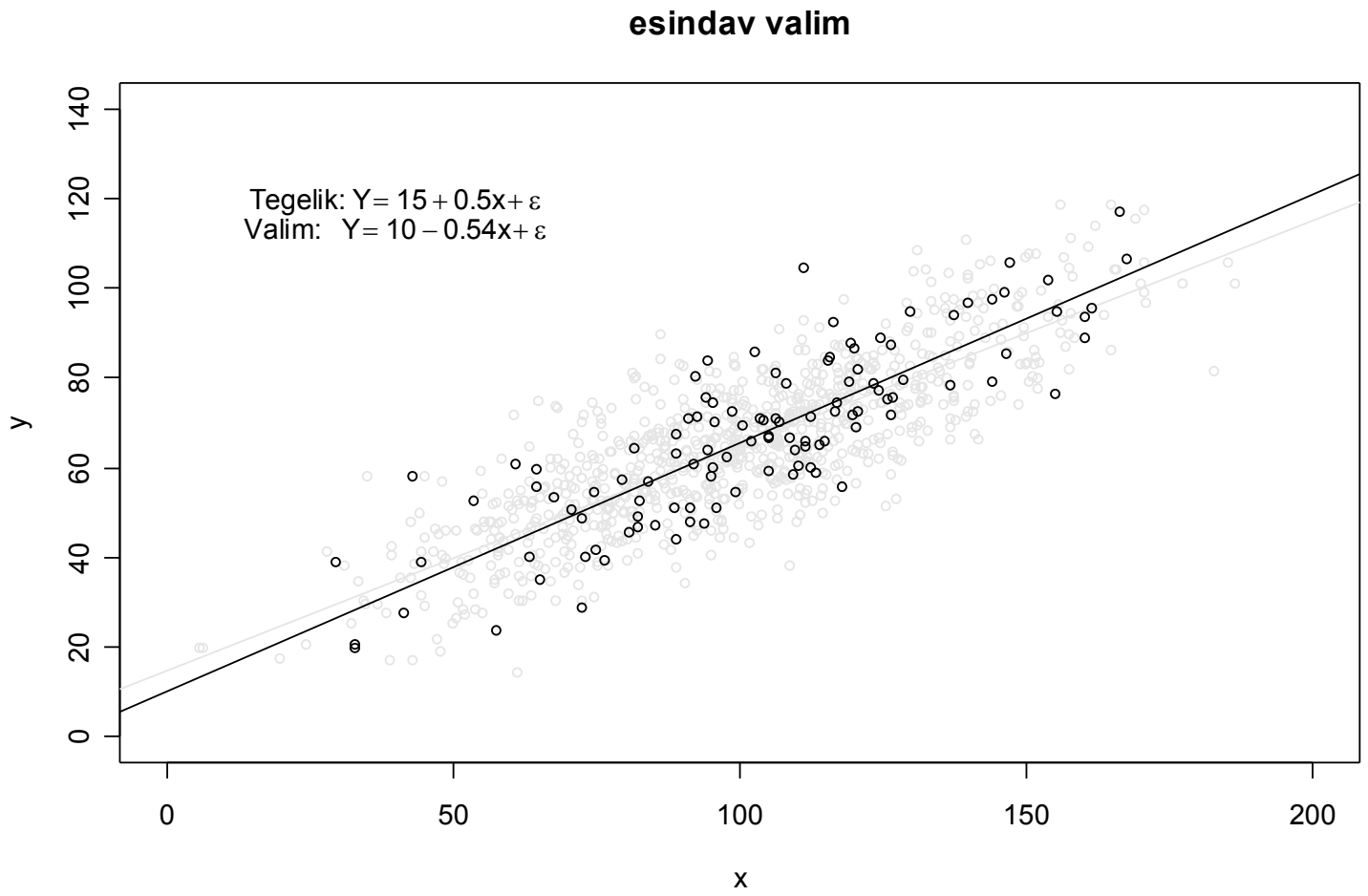
Eeldus I. Esindav valim



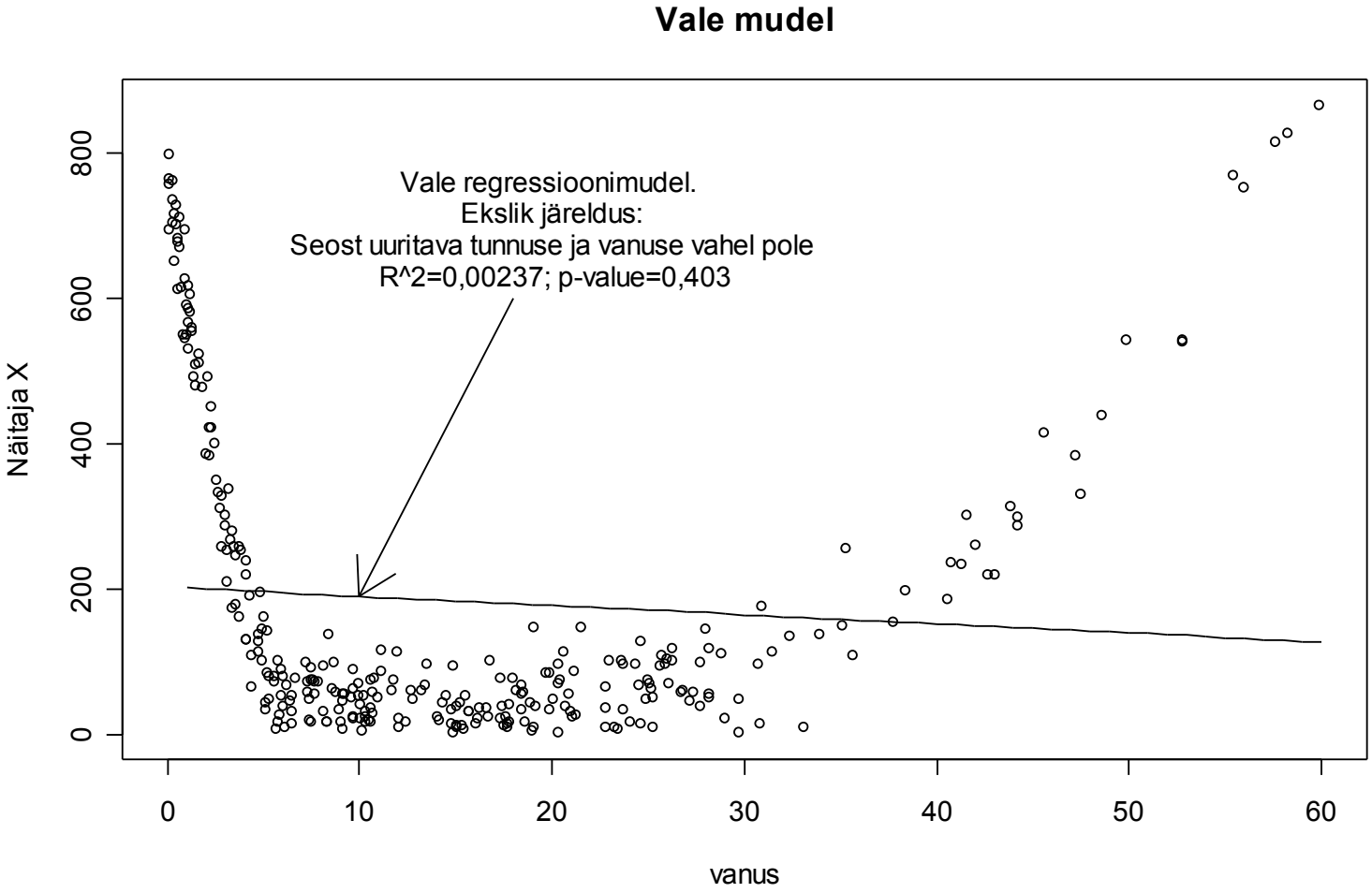
Eeldus I. Esindav valim



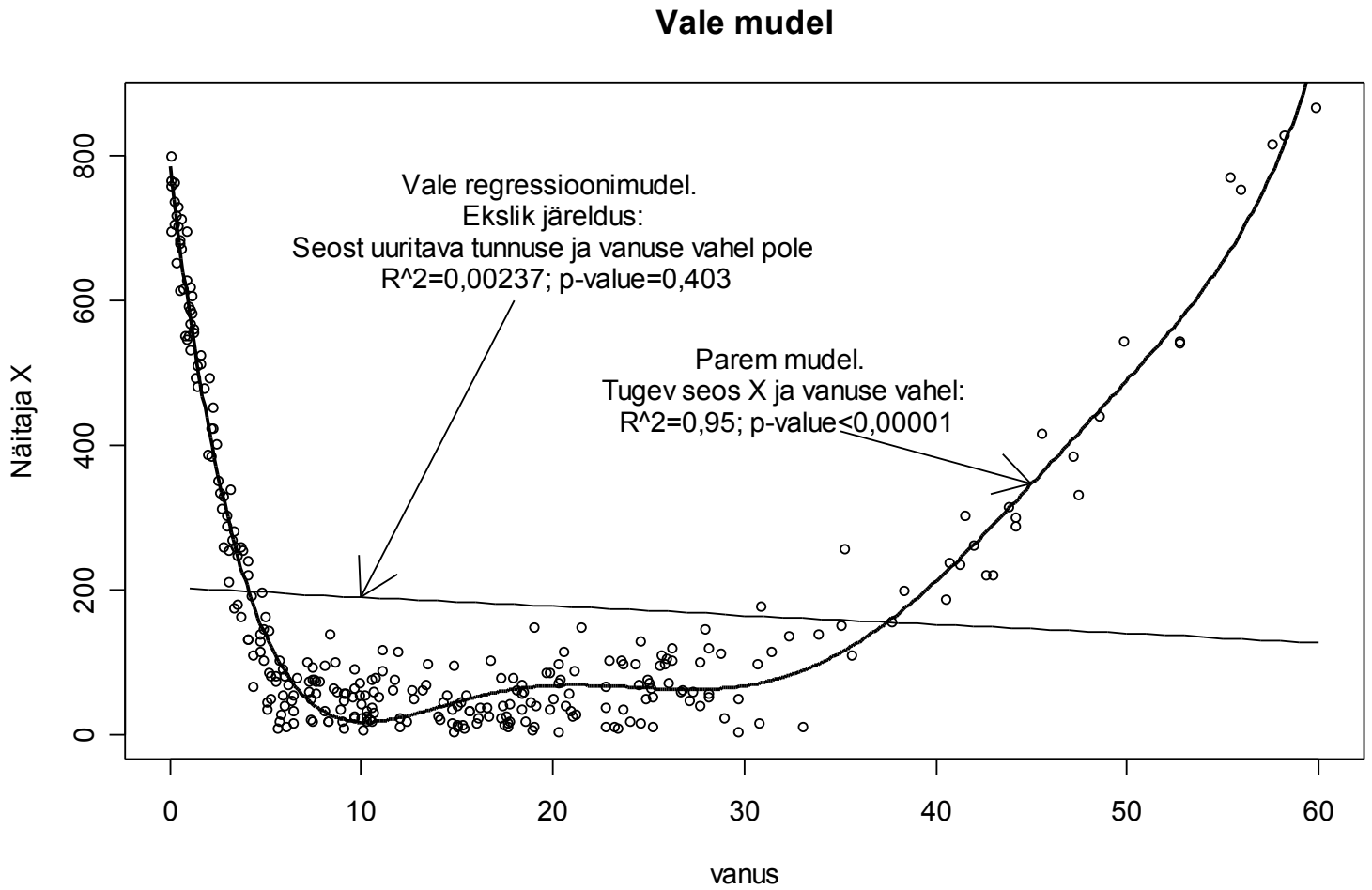
Eeldus I. Esindav valim



Eeldus II. Seos tunnuste vahel peab olema lineaarne.



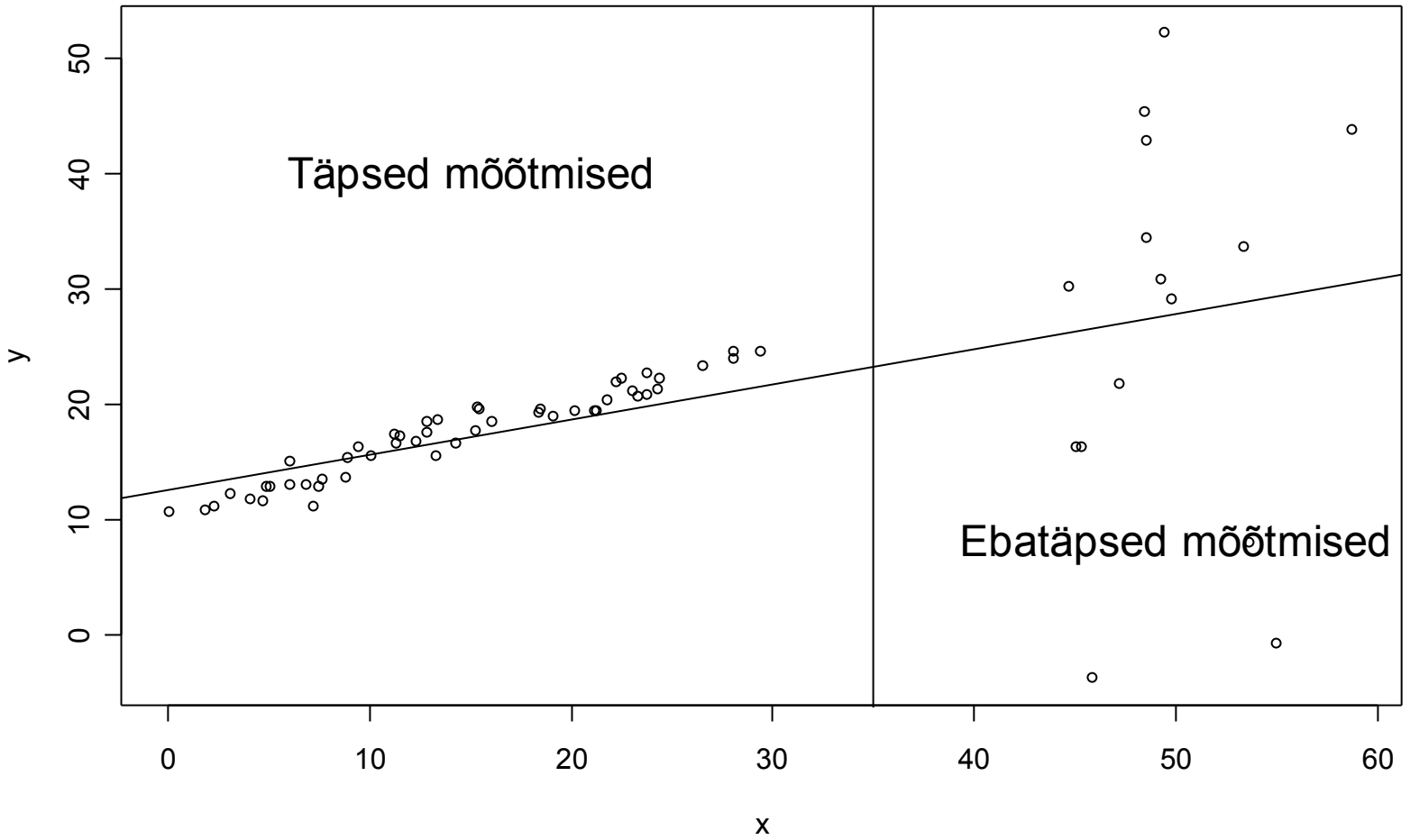
Eeldus II. Seos tunnuste vahel peab olema lineaarne.



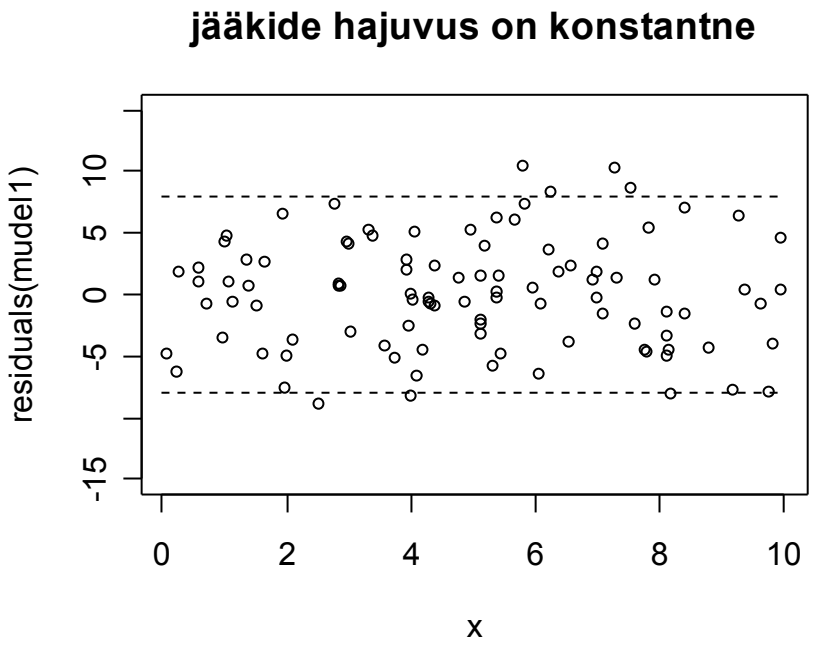
Veel eelduseid: uuritava tunnuse hajuvus I

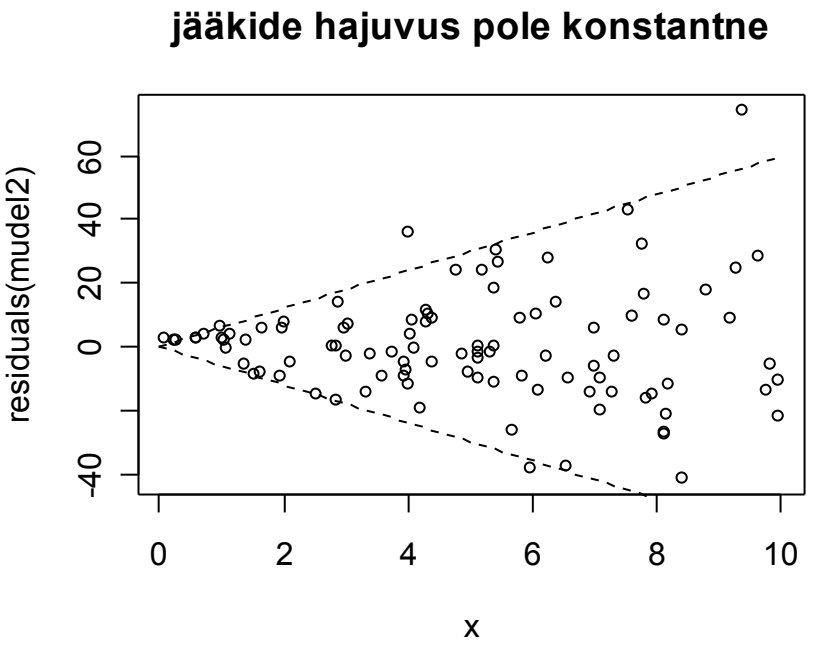
Harilik regressioonanalüüs eeldab, et uuritava tunnuse hajuvus on konstantne, st. mudeli jääkide hajuvus peaks olema ligikaudu samasuur (Näiteks peaksime suurte X-väärtuste korral saame ligikaudu samatäpsed prognoosid kui väikeste X väärtuste korral).

Kui uuritava tunnuse dispersioon pole konstantne, siis võivad arvuti poolt leitud usaldus- ja prognoosi (tolerantsi)intervallid osutuda eksitavaks. Samuti võivad suure hajuvusega vaatlus või vaatlused (näiteks teistest ebatäpsemalt mõõdetud vaatlused) oluliselt mõjutada regressioonanalüüsi tulemusi – regressioonisirge võib hakata liiga palju tähelepanu pöörama vigaselt mõõdetud väärtustele!

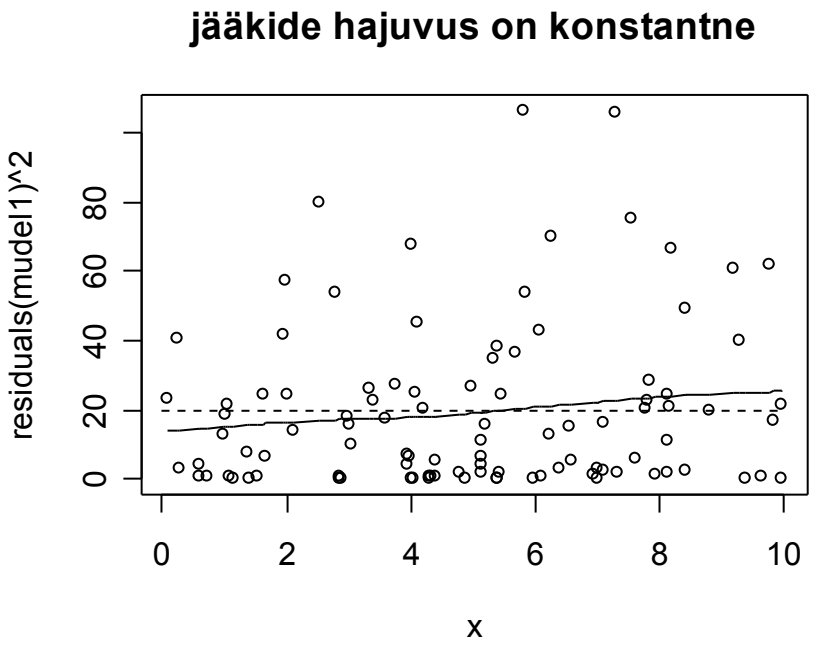

Täpsed mõõtmised
Ebatäpsed mõõtmised
y
x
0
10
20
30
40
50
60

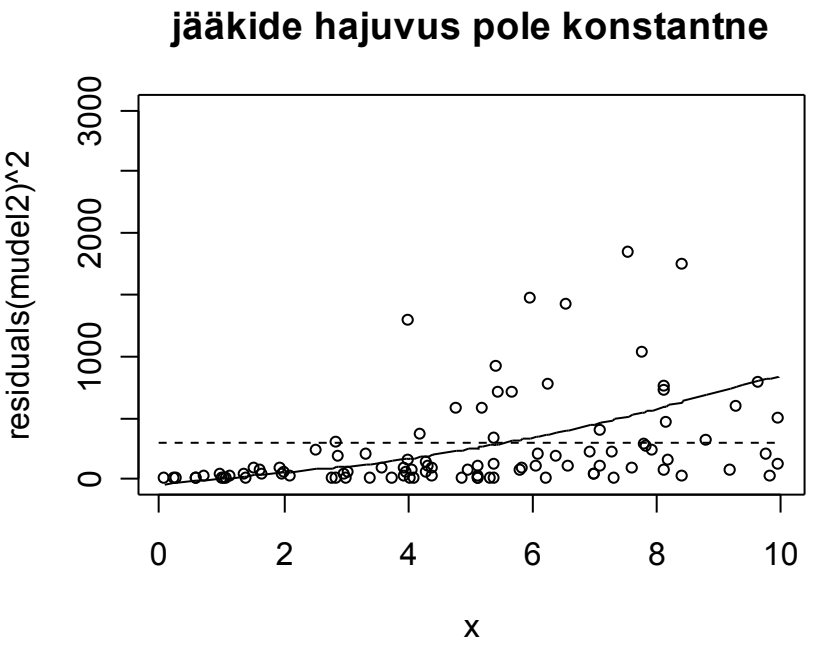
Tunnuse hajuvuse konstantsuse nõuet saab regressioonanalüüsi korral kontrollida vaadates mudeli jääkide ja argumenttunnuste või funktsioontunnuse hajuvusgraafikut





Juhul, kui argumenttunnuse mõningaid väärtuseid esineb märgatavalt sagedamini kui teisi, võivad ülaltoodud pildid kergesti silma eksitada. Parem on uurida jääkide ruutude ja argumenttunnuse vahelist seost:





Esmalt hindame mudeli:

```
> mudel1 = lm(y1~x+x2+factor(sugu))
```

Joonistame jääkide ruutude ja argumenttunnuse hajuvusgraafiku

```
> plot(x,residuals(mudel1)^2, main="Jääkide hajuvuse
konstantsuse kontrollimine")
```

Abistamaks silma lisame punkte siluva joone. Kui joon tuleb liiga hüplev, saab *spar*-parameetri väärtust suurendades teha joont siluvamaks

```
> lines(smooth.spline(x,residuals(mudel1)**2,spar=1.1))
```

Kasutatav mudel teeb eelduse jääkide hajuvuse kohta. Lisame mudeli eeldust kirjeldava katkendjoone (lty=2) hajuvusgraafikule:

```
> lines(c(0,10),rep(summary(mudel1)$sigma^2,2),lty=2)
```

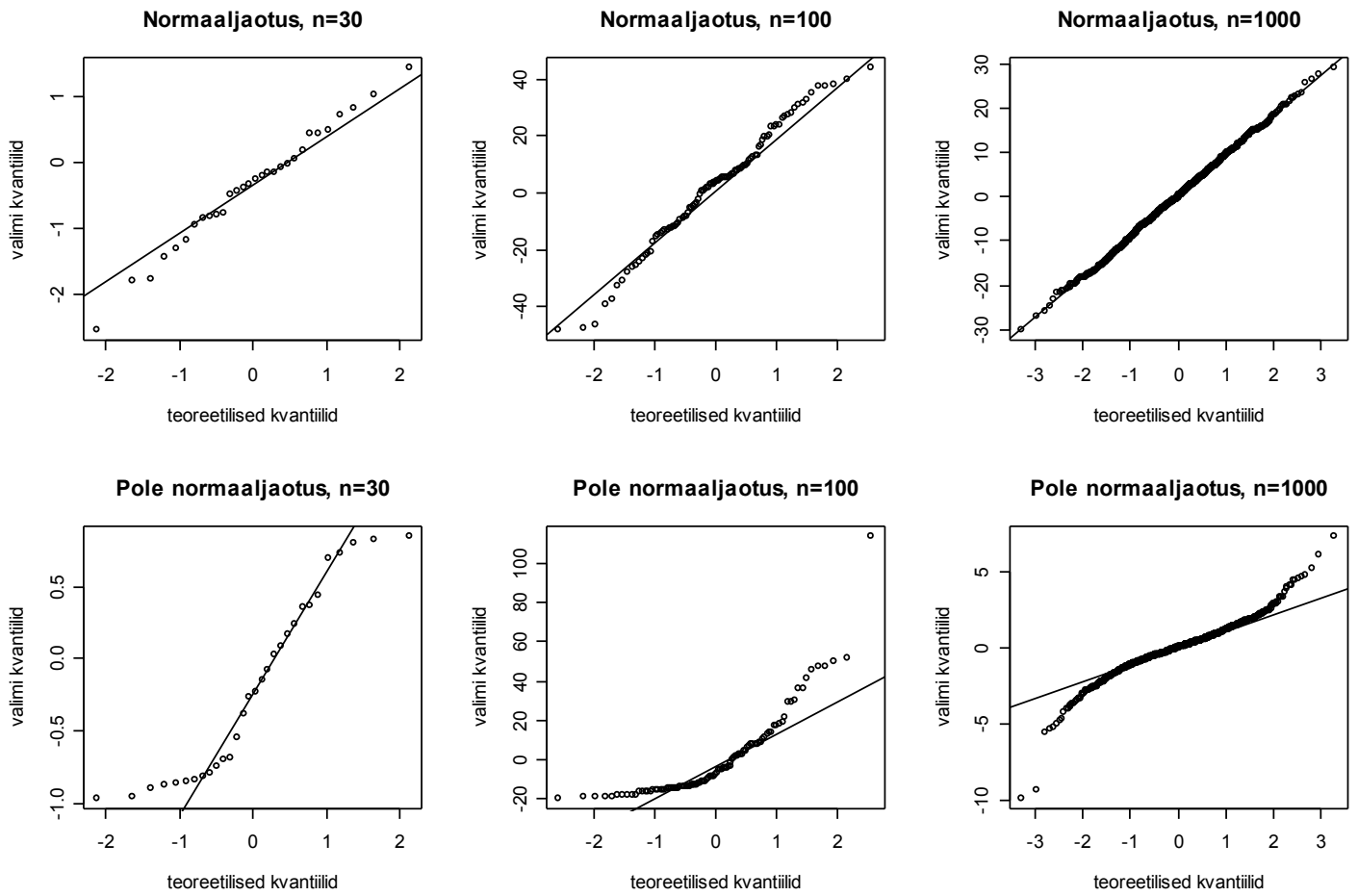
Mida teha, kui ilmnevad probleemid?

- Vahel aitab uuritava tunnuse teisendamine, näiteks logaritmi võtmine;

- lihtsa regressioon- või dispersioonanalüüsi asemel võib kasutada üldistatud lineaarseid mudeleid (Generalized Linear Models) kui mudeli jäägid pole normaaljaotusega või segamudeleid (Mixed Models); segamudelid sobivad näiteks siis, kui jääkide hajuvus pole konstantne (mõõtmistäpsus on andmete kogumise jooksul muutunud vms);

- kui kõrvalekalle eeldustest on väike, siis võib seda ka ignoreerida (probleemid eelkõige prognoosiintervalliga ja veidi langenud hinnangute täpsusega – viimast võib aga kompenseerida lihtsalt rohkem mõõtmisi tehes…).

Normaaljaotuse eeldus



Uuritava tunnuse jaotus – kasulikud käsud

Tõenäosuspaberi joonistamine

```
> qqnorm(residuals(mudel1))
```

Abistava joone lisamine pildile

```
> qqline(residuals(mudel1))
```

Shapiro-Wilki test kontrollimaks, kas mudeli jäägid on normaaljaotusega

```
> shapiro.test(jaak)

        Shapiro-Wilk normality test
data:  jaak
W = 0.9864, p-value = 0.3963
```

Uuritava tunnuse jaotus – mida teha probleemide ilmnemisel?

- Mudeli jäägid pole normaaljaotusega ka siis, kui on eksitud mudeli valikul – ehk annab mudelit parandades saada ka normaaljaotusega jääke?

- Vahel aitab uuritava tunnuse teisendamine, näiteks logaritmimine;

- Kui erinevus normaaljaotusest on väike, võib probleemi enamasti ka ignoreerida;

- Tavalise regressioon- või dispersioonanalüüsi asemel võib kasutada üldistatud lineaarseid mudeleid (näiteks logistilist või poissoni regressiooni).

- Dispersioonanalüüsi asemel võib kaaluda mitteparameetriliste testide, näiteks Kruskall-Wallise testi kasutamist:

```
kruskall.test(sissetulek~factor(rahvus))
```

Muud võimalikud probleemid

- erindid

  Cook’i kaugused (Cook’s distance):

  ```
  > plot(cooks.distance(mudel))
  > identify(cooks.distance(mudel))
  ```

- sõltuvad vaatlused
- mõõtmisvead
- mõõtmata jäänud olulised tunnused
- ...

Ühefaktoriline dispersioonanalüüs

(*One-way ANOVA*)

Vaatame, mida teha, kui soovime prognoosimisel kasutada sellise tunnuse abi, mis polegi pidev.

Faktortunnused

Vahel soovime prognoosida tunnuse *Y* väärtuseid, aga tunnus, mille abil me prognoosime, *X*, pole pidev. Näiteks soovime prognoosida põllult saadavat saaki, teades, kumba sorti – A-d või B-d sellel põllul kasvatati.

Mida saame teha?

```
> by(saak, sort, mean)
INDICES: A
[1] 3205.455
------------------------------------------------------------
INDICES: B
[1] 3572.889
```

Mudel 1:

Saak = 3205,5*I(sort==“A“) + 3572,9*I(sort==“B“) + prognoosiviga

Mudel 1 prognoos:

sort A saagikusele: 3205,5

sort B saagikusele: 3572,9

Mudel 2 (samaväärne):

Saak = 3205,5 + 367,4*I(sort==“B“) + prognoosiviga

Mudel 2 prognoos:

sort A saagikusele: 3205,5

sort B saagikusele: 3205,5+367,4 = 3572,9

Kuidas leida 2. mudeli parameetreid?

Teeme esmalt abitunnuse

$$I_{sortB} = \begin{cases} 0, \text{ kui sort} \neq \text{B} \\ 1, \text{ kui sort} = \text{B} \end{cases}.$$

Kasutame nüüd saadud abitunnust regressioonimudelis argumenttunnusena.

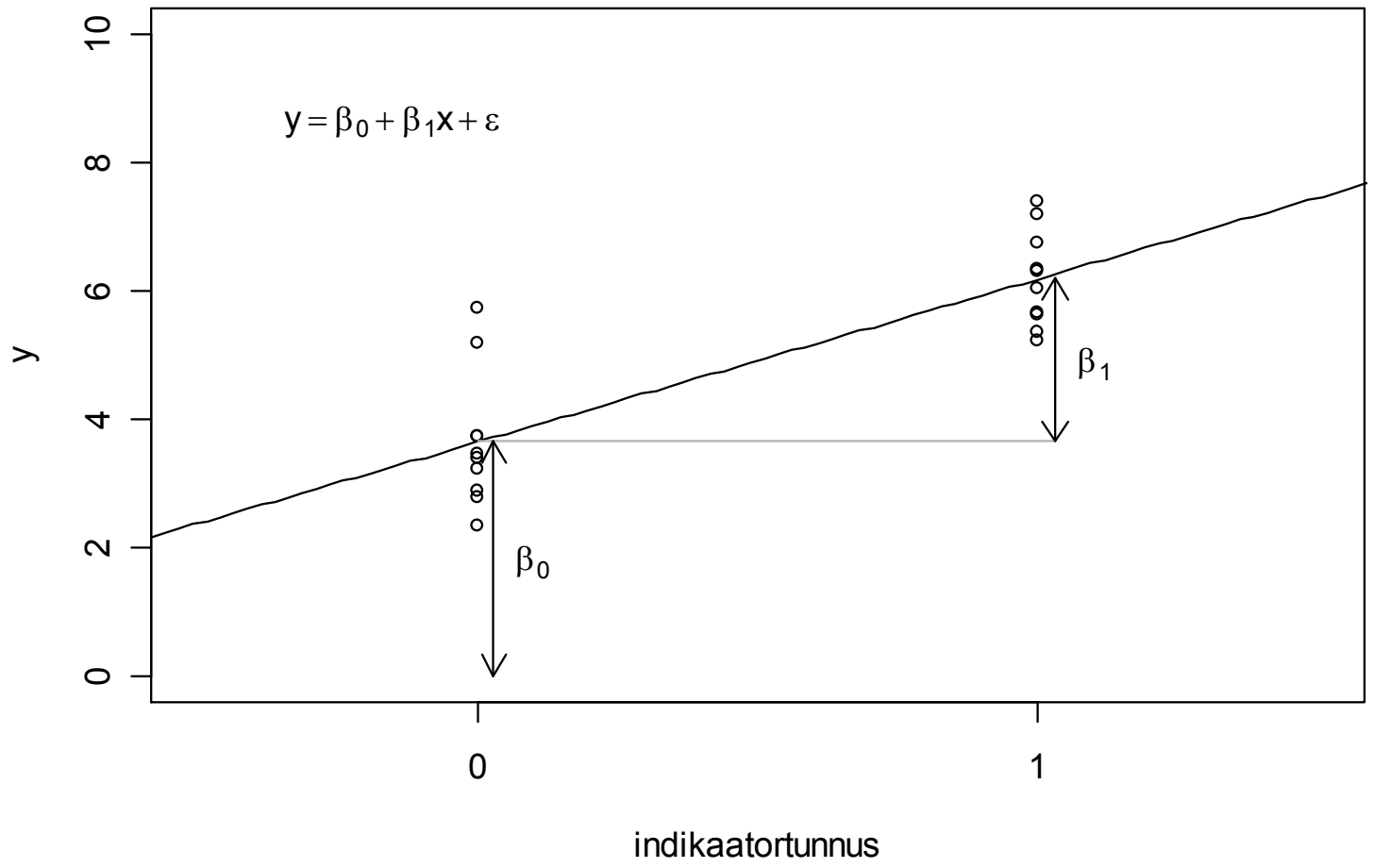
```
> I_sortB=1*(sort=="B")
> lm(saak~I_sortB)

Coefficients:
(Intercept)      I_sortB
     3205.5        367.4
```

Võime lasta ka statistikaprogrammil endal indikaatortunnuse teha:

```
> lm(saak~factor(sort))

Coefficients:
  (Intercept)  factor(sort)B
       3205.5          367.4
```


$y = \beta_0 + \beta_1 x + \varepsilon$
β_0
β_1
y
0
2
4
6
8
10
0
1
indikaatortunnus

Millist rühma valida võrdluse aluseks?

Paneme tähele, et vaatlustulemuste kirjeldamise seisukohast on järgmised kolm mudelit täpselt sama head:

Mudel 1:

Saak = 3205*I(sort==“A“) + 3573*I(sort==“B“) + prognoosiviga

Mudel 2 (võrdluse aluseks sort A):

Saak = 3205 + 367*I(sort==“B“) + prognoosiviga

Mudel 3 (võrdluse aluseks sort B):

Saak = 3573 - 367*I(sort==“A“) + prognoosiviga

Mudel 4 (võrdluse aluseks valimi keskmine):

Saak = 3371 - 166*I(sort==“A“) + 202*I(sort==“B“) + prognoosiviga

Mudel 5 (võrdluse aluseks sortide keskmiste saakide keskmine):

Saak = 3389 - 184*I(sort==“A“) + 184*I(sort==“B“) + prognoosiviga

Tähelepanu!

Erinevad programmid valivad võrdluse aluse erinevalt! Kui analüüsime sama andmestikku erinevate statistikaprogrammide abil, võime saada vastuseks vägagi erinevaid numbreid. Tulemuste interpretatsioon jääb aga alati samaks.

SAS

```
proc glm data=saak;
  class sort;
  model saak=sort /solution; run;
```

```
Parameter              Estimate              Error    t Value    Pr > |t|
 Intercept          3572.888889 B     35.72819340     100.00      <.0001
 sort       A       -367.434343 B     48.17588615      -7.63      <.0001
 sort       B          0.000000 B              .          .           .
```

R

```
summary(lm(saak~factor(sort)))
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
factor(sort)B   367.43      48.18   7.627 4.81e-07 ***
```

Testimine – faktortunnusel 2 taset

Kui faktortunnusel on kõigest 2 taset, jõuame samade tulemusteni nii t-testi, ANOVA kui ka regressioonanalüüsi abil:

```
> t.test(SAAK~factor(SORT), var.equal=TRUE)
```

Studenti t-test

```
        Two Sample t-test
data:  SAAK by factor(SORT)
t = -7.6269, df = 18, p-value = 4.805e-07
mean in group A mean in group B
       3205.455        3572.889
```

```
> summary(lm(SAAK~factor(SORT)))
```

ANOVA ehk dispersioonanalüüs

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
factor(SORT)B   367.43      48.18   7.627 4.81e-07 ***

Residual standard error: 107.2 on 18 degrees of freedom
Multiple R-Squared: 0.7637,     Adjusted R-squared: 0.7506
F-statistic: 58.17 on 1 and 18 DF,  p-value: 4.805e-07
```

```
> summary(lm(saak~I_sortB))
```

regressioonanalüüs

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      32.32  99.187  < 2e-16 ***
I_sortB         367.43      48.18   7.627 4.81e-07 ***
```

Rohkem kui kaks faktortunnuse taset

Kui faktortunnuses on rohkem kui kaks taset (lisaks sortidele A ja B on vaadeldud ka sorte C, D ja E), siis tuleb teha ka rohkem indikaatortunnuseid.

Andmestiku näide (võrreldakse sordiga A):

Saak	sort	I_{sortB}	I_{sortC}	I_{sortD}	I_{sortE}
3450	A	0	0	0	0
3567	A	0	0	0	0
3256	B	1	0	0	0
3345	B	1	0	0	0
3890	C	0	1	0	0
3925	C	0	1	0	0
3300	D	0	0	1	0
3123	D	0	0	1	0
3800	E	0	0	0	1
3850	E	0	0	0	1

Üks vähegi viisakas statistikaprogramm teeb indikaatortunnused muidugi ise ka valmis.

Näide

```
> summary(lm(saak~factor(sort)))

Residuals:
    Min      1Q  Median      3Q     Max
-309.00 -108.12   13.52   90.05  290.50

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    3205.45      43.90  73.025  < 2e-16 ***
factor(sort)B   367.43      65.44   5.615 1.16e-06 ***
factor(sort)C   170.05      67.65   2.514  0.01559 *
factor(sort)D  -216.45      62.08  -3.487  0.00110 **
factor(sort)E    33.00      62.08   0.532  0.59762
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 145.6 on 45 degrees of freedom
Multiple R-Squared: 0.6578,     Adjusted R-squared: 0.6274
F-statistic: 21.63 on 4 and 45 DF,  p-value: 5.249e-10
```

Kordajad on tõestatavalt nullist erinevad, sortide B, C ja D keskmine saagikus erineb sordi A keskmisest saagikusest

Sordi E saagikuse keskväärtus võib olla sama, mis sordi A keskmine saagikus

Testimine – kuid me soovisime testida midagi muud?

Esimene küsimus, millele otsime vastust: kas üldse eksisteerib erinevust sortide vahel?

Kui sort A saagikus on $Saak_1$ js sort B saagikust tähistame $Saak_2$ jne, siis meid huvitava hüpoteesi võiks sõnastada järgmiselt:

H_0: E $Saak_1$ = E $Saak_2$ = E $Saak_3$ = … = E $Saak_k$
H_1: H_0 ei kehti (eksisteerivad vähemalt kaks sorti, mille keskmised saagikused pole võrdsed).

Alljärgnevalt vaatame erinevaid võimalusi kontrollida hüpoteese keskväärtuste võrdsuse kohta R-is.

Kolm sarnast testi

```
> summary(lm(SAAK~factor(SORT)))
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       3205.45      43.21  74.189  < 2e-16 ***
factor(SORT)B      367.43      64.41   5.705 8.55e-07 ***
factor(SORT)C      116.82      61.10   1.912   0.0623 .
factor(SORT)D     -268.73      61.10  -4.398 6.62e-05 ***
factor(SORT)E       36.42      66.59   0.547   0.5871
Residual standard error: 143.3 on 45 degrees of freedom
Multiple R-Squared: 0.694,      Adjusted R-squared: 0.6668
F-statistic: 25.52 on 4 and 45 DF,  p-value: 4.456e-11

> drop1(lm(SAAK~factor(SORT)), test="F")
               Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                      924079     501
factor(SORT)    4   2095865 3019944     552  25.516 4.456e-11 ***

> summary(aov(lm(SAAK~factor(SORT))))
               Df  Sum Sq Mean Sq F value    Pr(>F)
factor(Variety)  4 2095865  523966  25.516 4.456e-11 ***
Residuals       45  924079   20535
```

Kas mudel on hea?

Kas me vajame tunnust „SORT“?

„Traditsiooniline“ ANOVA tabel, võib osutuda eksitavaks mittetasakaaluliste andmestike korral (aga ühefaktorilise dispersioonanalüüsi korral on OK.)

Testimine – üks tähelepanek

```
> m1=lm(Saak~factor(Sort)); summary(m1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     4370.6      152.8  28.607   <2e-16 ***
factor(Sort)B   -202.3      161.0  -1.256    0.212
factor(Sort)C    162.0      162.4   0.997    0.321
factor(Sort)D    103.8      162.9   0.637    0.526
factor(Sort)E   -162.5      160.8  -1.011    0.315

Residual standard error: 264.6 on 98 degrees of freedom
Multiple R-Squared: 0.2693,     Adjusted R-squared: 0.2394
F-statistic: 9.028 on 4 and 98 DF,  p-value: 2.993e-06


> drop1(m1, test="F")
Single term deletions

Model:
Saak ~ factor(Sort)
             Df Sum of Sq     RSS     AIC F value     Pr(F)
<none>                    6862583    1154
factor(Sort)  4   2528847 9391430    1178  9.0282 2.993e-06 ***
```

Võrdleme teise faktortunnuse tasemega (sort „B“-ga)

```
> m1=lm(Saak~C(factor(Sort), base=2)); summary(m1)

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                4168.32      50.93  81.849  < 2e-16 ***
C(factor(Sort), base = 2)1   202.26     161.05   1.256 0.212142
C(factor(Sort), base = 2)3   364.29      75.09   4.852 4.62e-06 ***
C(factor(Sort), base = 2)4   306.01      76.00   4.026 0.000112 ***
C(factor(Sort), base = 2)5    39.79      71.38   0.558 0.578455
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 264.6 on 98 degrees of freedom
Multiple R-Squared: 0.2693,     Adjusted R-squared: 0.2394
F-statistic: 9.028 on 4 and 98 DF,  p-value: 2.993e-06
```

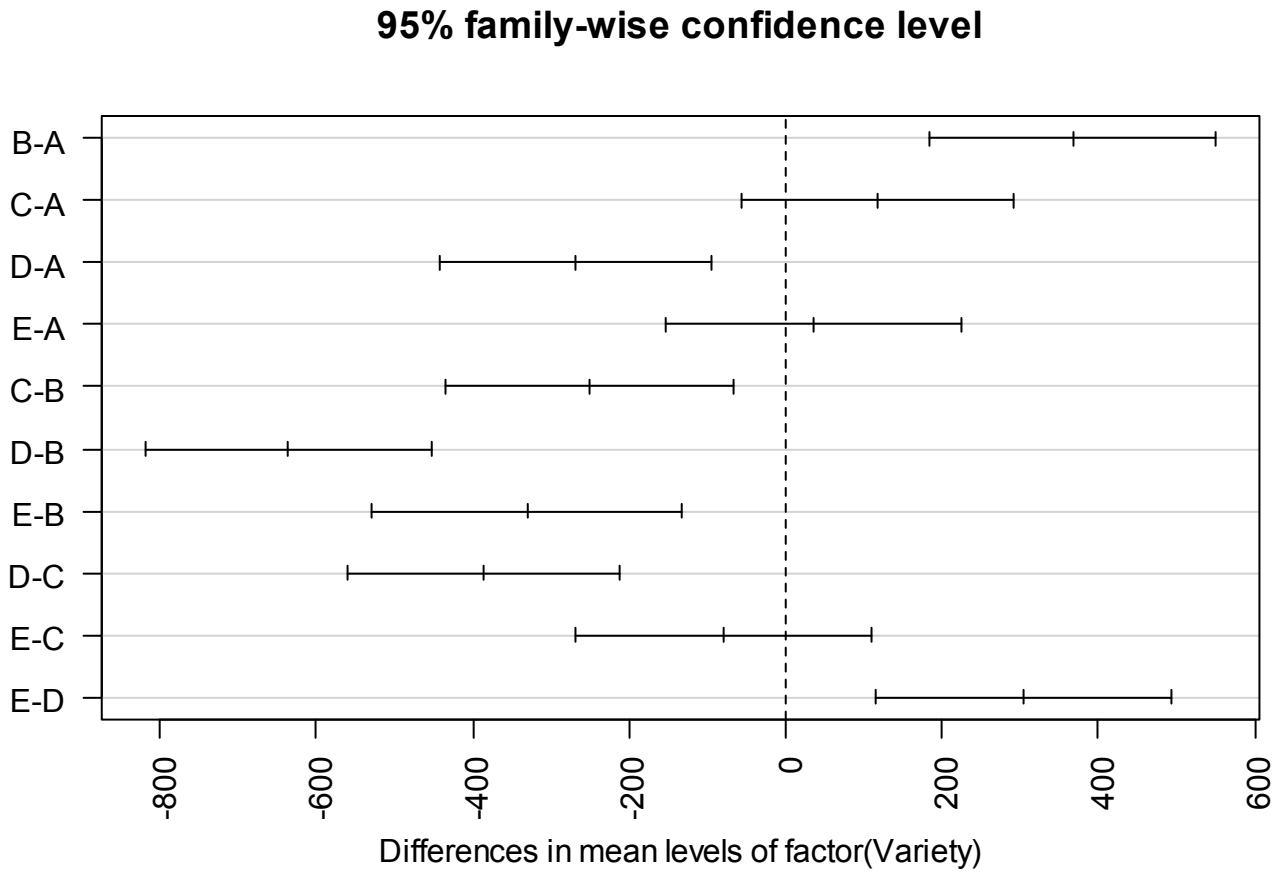
Võib kasutada ka relevel-käsku!

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)
```



Mittetasakaaluline andmestik

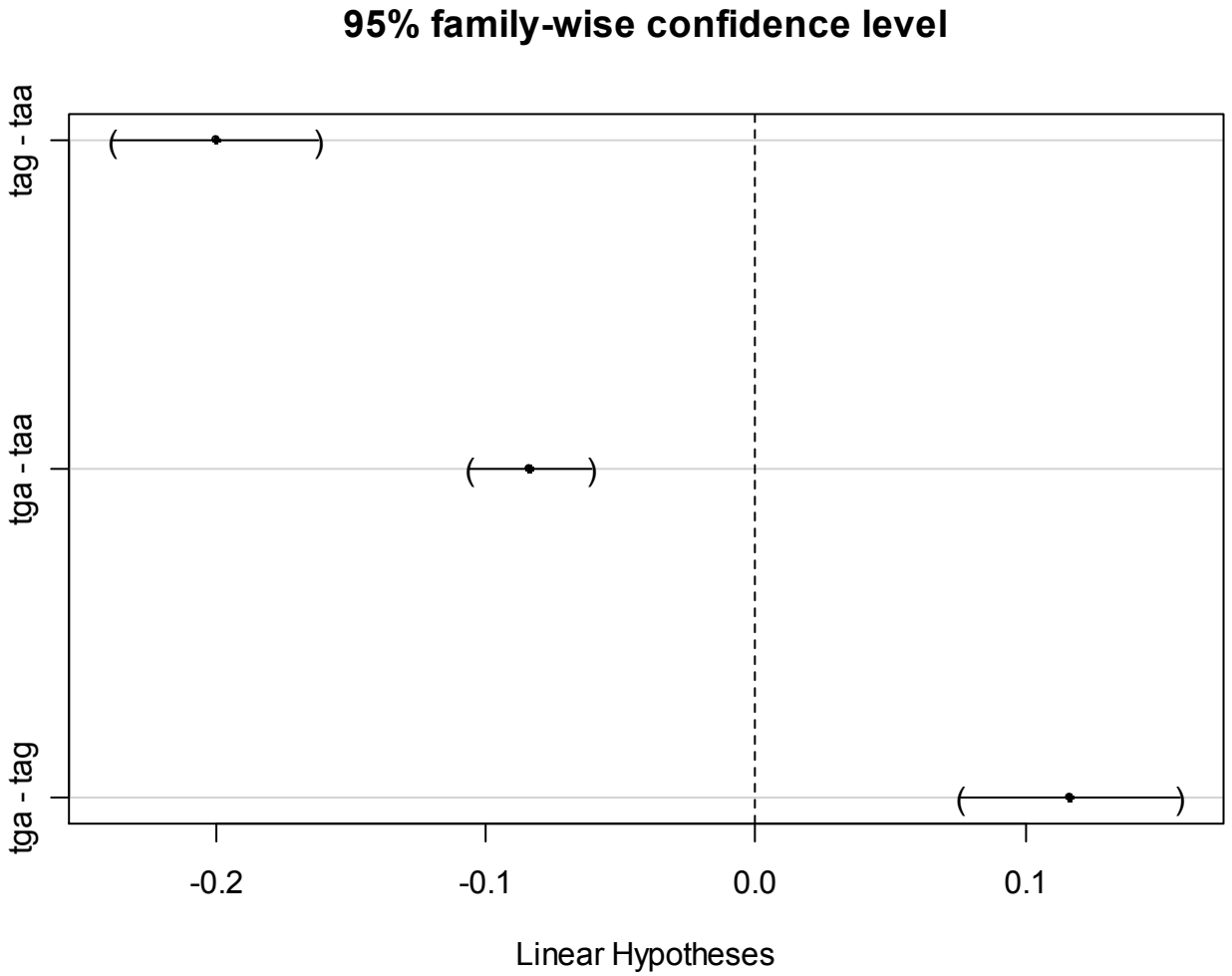
Tuleb kasutada lisamoodulit multcomp

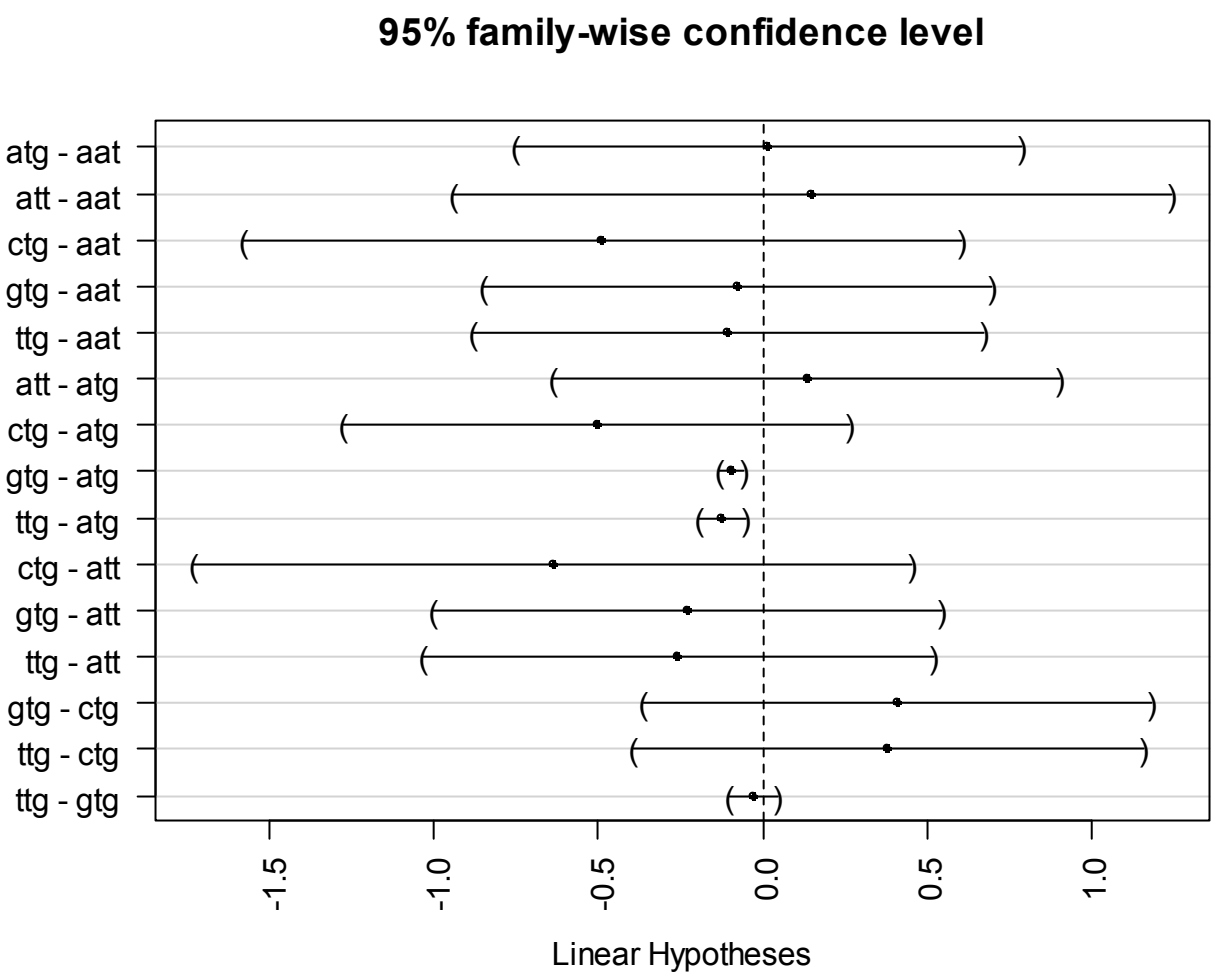
```
> library(multcomp)
> flast=factor(last_codon)
> m1=aov(lm(log(cai)~flast))
> a=glht(m1, linfct=mcp(flast="Tukey"))
> summary(a)

         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))
```




95% family-wise confidence level
atg - aat
att - aat
ctg - aat
gtg - aat
ttg - aat
att - atg
ctg - atg
gtg - atg
ttg - atg
ctg - att
gtg - att
ttg - att
gtg - ctg
ttg - ctg
ttg - gtg
-1.5
-1.0
-0.5
0.0
0.5
1.0
Linear Hypotheses

```
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
```

Veel determinatsioonikordajast

Vaatame hüpoteetilist olukorda – taime X saagikus sõltub a) talumehe hoolikusest; b) sordis; c) aastast; d) põllust; e) sordist; f) 4m^2 –sel katselapil antud aastal kasvavate taimede eripärast. Eeldame, et koosmõjusid mainitud tunnuste vahel pole. Saagikust talupõllul kirjeldab siis järgmine mudel:

$$Y = c_0 + f_1(hoolikus) + f_2(sort) + f_3(aasta) + f_4(põld) + e.$$

Oletame hüpoteetiliselt, et kõik mainitud tunnused on teineteisest sõltumatud. Sellisel juhul talupõldudelt saadud saagi dispersioon DY koosneb järgmistest komponentidest:

$$\mathrm{D}Y = \mathrm{D}(f_1) + \mathrm{D}(f_2) + \mathrm{D}(f_3) + \mathrm{D}(f_4) + \mathrm{D}(e). \qquad (1)$$

Oletame, et üks põlluharija teeb sortide võrdluskatse ühel aastal ühel põllul (mitmel katselapil). Tema katses on saagikuse dispersioon

$$DY = D(f_2) + D(e). \qquad (2)$$

Eeldame näitlikustamise huvides, et $D(f_1) = D(f_2) = D(f_3) = D(f_4) = D(e) = 1$.

Siis

Talupõldudel tehtud katse puhul tuleb

$$R^2 = D(f_2)/ (D(f_1)+D(f_2)+D(f_3)+D(f_4)+D(e)) = 1/5 = 0{,}2.$$

Põldkatse puhul tuleb

$$R^2 = D(f_2)/ (D(f_2) + D(e)) = ½ = 0{,}5.$$

Oleks katset tehtud 16 m^2-tel katselappidel, oleksime saanud tulemuseks

$$R^2 = D(f_2)/ (D(f_2) +D(e)/4) = 0{,}8.$$

Seega: kuigi sortidevaheline erinevus on sama (sordi mõju alati samasuur) võib determinatsioonikordaja tulla vägagi erinev! Determinatsioonikordaja väärtus sõltub väga tugevalt sellest, kuidas me defineerime prognoositava „üksuse“ – ehk kuivõrd me kontrollime katsetingimusi.

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.