



**Statistika ja modelleerimise
baaskoolitus õppejõududele ja
juhendajatele**

Loeng 2

Eesti Maaülikool
27.-30. august 2012
Tanel Kaart



Korrelatsioonanalüüs

 **Eesti Maaülikool**
Estonian University of Life Sciences

www.emu.ee

Lineaarne e Pearsoni korrelatsioonikordaja

Millal kasutada ja mida näitab?

Mõõdab kahe pideva (normaaljaotusega) tunnuse vahelise lineaarse seose tugevust ja suunda.

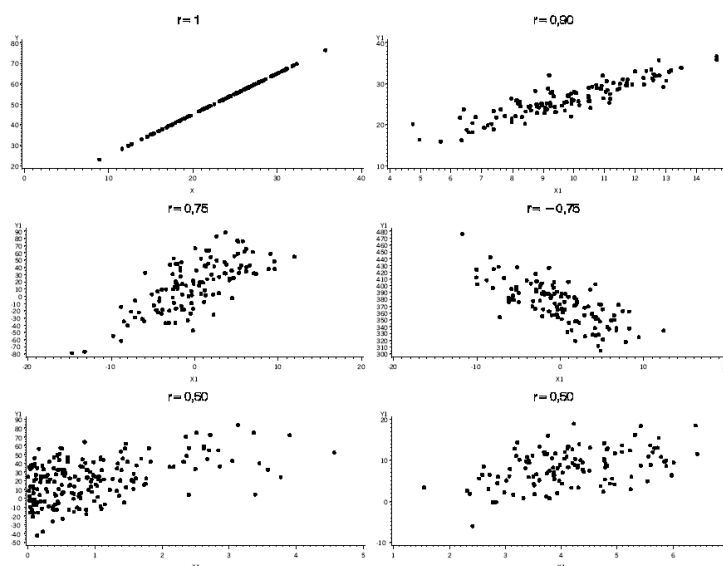
Arvutusvalem:
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{j=1}^n (y_j - \bar{y})^2}}$$

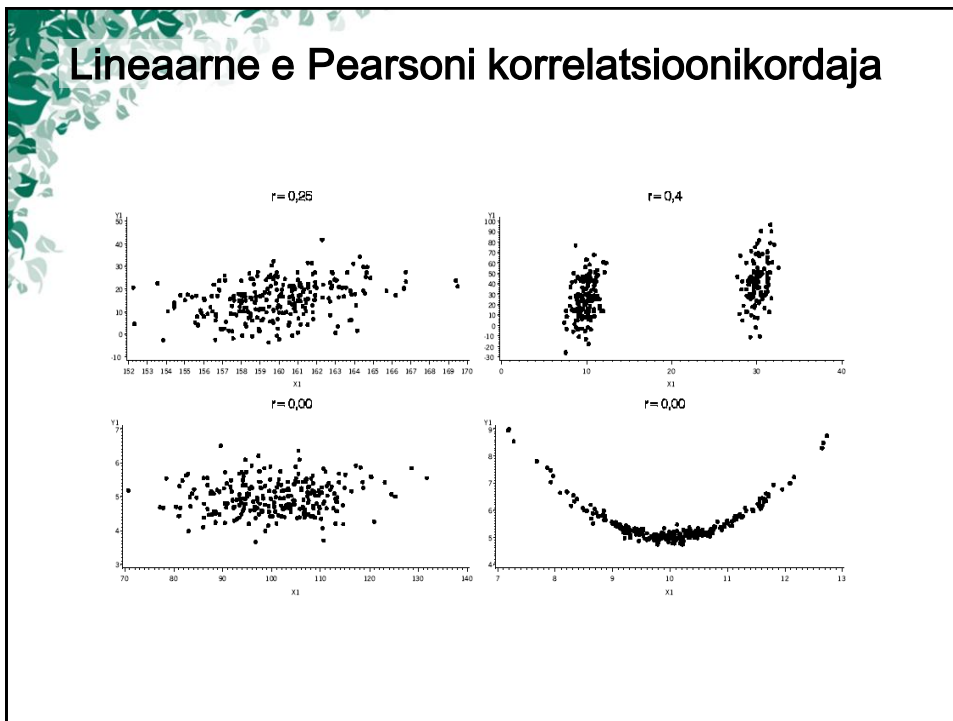
Omadused:

- $-1 \leq r \leq 1$;
- kui $r > 0$, siis tunnuse X suurenedes suureneb keskmiselt ka tunnus Y ;
kui $r < 0$, siis X -i suurenedes Y keskmiselt kahaneb ja X -i kahanedes Y keskmiselt suureneb;
- kui tunnused X ja Y on sõltumatud, siis $r = 0$;
- kui tunnuste X ja Y vahel on täpne lineaarne seos, siis $|r| = 1$;
- mida suurem on korrelatsioonikordaja absoluutväärtus, seda tugevam on korrelatiivne seos tunnuste vahel.

Lineaarne e Pearsoni korrelatsioonikordaja

Hajuvusdiagramm e korrelatsiooniväli
[scatter plot]





Lineaarne e Pearsoni korrelatsioonikordaja

Kokkuleppelised piirid seose tugevuse iseloomustamiseks:

- $|r| \leq 0,3$ – nõrk seos;
- $0,3 < |r| < 0,7$ – keskmine seos;
- $|r| \geq 0,7$ – tugev seos.

Seose statistiline olulisus

$H_0 : r = 0$
 $H_1 : r \neq 0$

Teststatistik $t = r\sqrt{n-2} / \sqrt{1-r^2} \sim t_{n-2}$

Olulisustõenäosus $p = S_t + S_t$

t-jaotuse tihedusf.-n

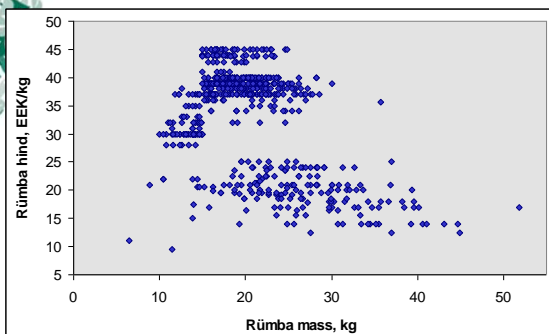
1-2 S_t

-t 0 t

Näiteks *MS Excelis* saab olulisustõenäosuse p leidmiseks kasutada funktsiooni $TDIST(t;n-2;2)$.

Lineaarne e Pearsoni korrelatsioonikordaja

Näide. Lineaarne seos lamba lihakeha massi ja makstava 1 kg hinna vahel ($n=686$).
Andmed aastast 2002.



$$r = -0,473$$

Teststatistik

$$t = r\sqrt{n-2} / \sqrt{1-r^2}$$

$$= -0,473\sqrt{686-2} / \sqrt{1-(-0,473)^2}$$

$$= -14,039,$$

millest $|t|=14,039$.

Viimase alusel leitav olulisuse tõenäosus $p = 1,577 \cdot 10^{-37} < 0,05$,
mistõttu võime lugeda tõestatuks negatiivse seose olemasolu lamba lihakeha massi
ja rümba 1 kg hinna vahel ($H_1: r \neq 0$) – mida suurem on tapamajja viidav lammas,
seda vähem ühe kg liha eest makstakse.

Korrelatsioonimaatriks

Näide. Lineaarsed seosed mesilaste peamiste kehameetrite vahel ($n=1380$).

	Tergiit	Tiiva laius	Tiiva pikkus
Tiiva laius	0,052		
Tiiva pikkus	0,061*	0,210***	
Iminokk	-0,035	0,074**	0,253***

* – $p < 0,05$; ** – $p < 0,01$; *** – $p < 0,001$



Pearson Correlation Coefficients, N = 1380
Prob > |r| under H0: Rho=0

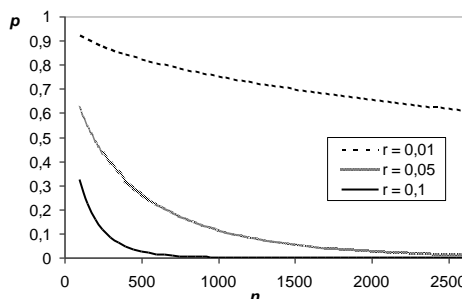
	Tergiit	Tiiva_l	Tiiva_p	Iminokk
Tergiit	1.00000	0.05226	0.06082	-0.03518
Tergiit		0.0523	0.0239	0.1915
Tiiva_l	0.05226	1.00000	0.21007	0.07390
Tiiva_l		0.0523	<.0001	0.0060
Tiiva_p	0.06082	0.21007	1.00000	0.25273
Tiiva_p		0.0239	<.0001	<.0001
Iminokk	-0.03518	0.07390	0.25273	1.00000
Iminokk		0.1915	<.0001	

Osa SAS-i protseduuri
CORR väljundist

Lineaarne e Pearsoni korrelatsioonikordaja

NB!

- ✓ Lineaarne korrelatsioonikordaja mõõdab üksnes lineaarset seost.
- ✓ Korrelatsioonikordaja ei ütle midagi seose **põhjuslikkuse** kohta.
- ✓ Korrelatsioonikordajale vastav p -väärtus ei puugi anda mingit infot seose tugevuse kohta – näiteks $n = 39000$ korral on ka korrelatsioonikordajale $r = 0,01$ vastav $p < 0,05$.



Astakorrelatsioonikordaja e Spearmani korrelatsioonikordaja

Millal kasutada ja mida näitab?

Mõõdab kahe arvtunnuse vahelise monotoonse seose tugevust ja suunda.

Ei ole tundlik erindite suhtes ega eelda tunnuste normaalset jaotumist (on põhimõtteliselt kasutatav ka järjestustunnuste puhul).

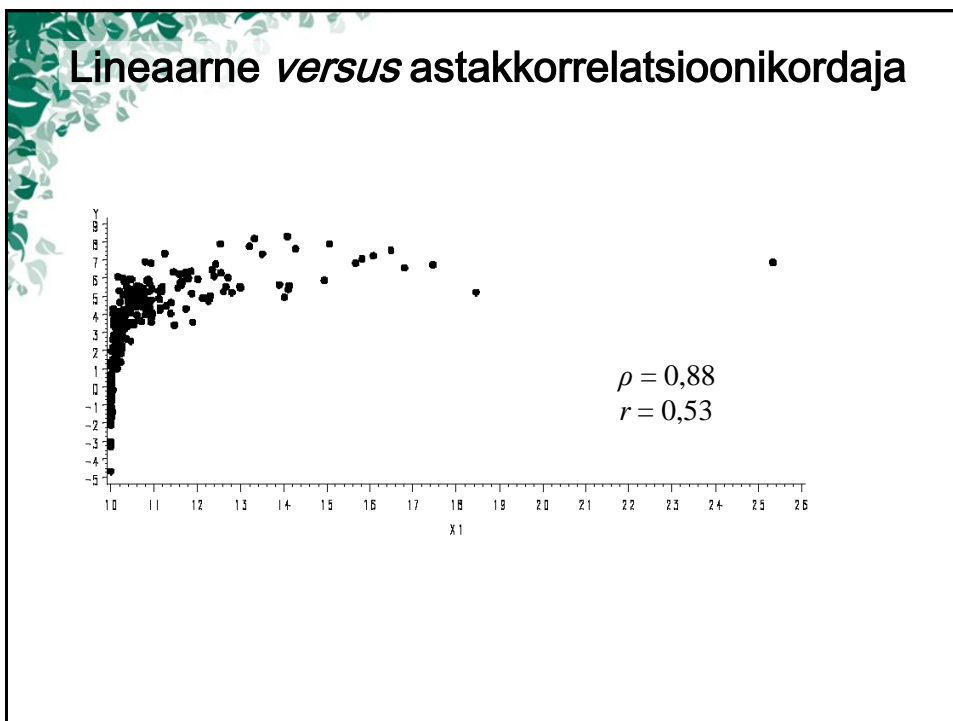
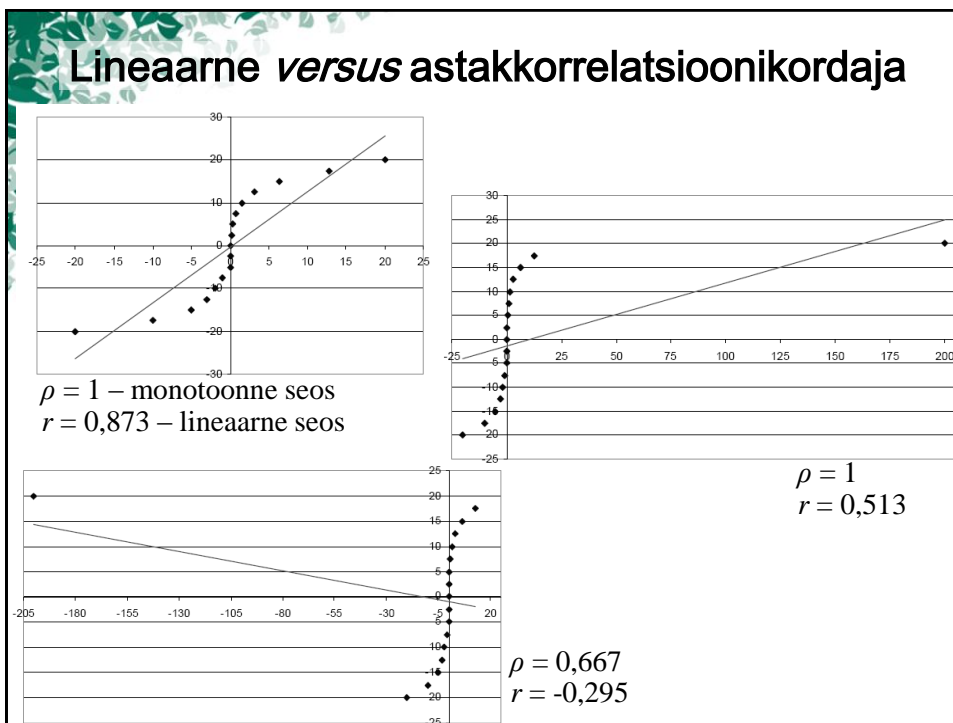
Arvutamine: leitakse kui seos vaatluste järjekorranumbrite e astakute vahel:

$$\rho = 1 - \frac{6 \sum_{i=1}^n (x_{(i)} - y_{(i)})^2}{n(n^2 - 1)},$$

kus $x_{(i)}$ on tunnuse X väärtuse x_i astak
ja $y_{(i)}$ on tunnuse Y väärtuse y_i astak.

Omadused:

- $-1 \leq \rho \leq 1$;
- kui tunnuste vahel on kasvav monotoonne seos, siis $\rho > 0$;
- kui tunnuste vahel on kahanev monotoonne seos, siis $\rho < 0$;
- kui tunnused X ja Y on sõltumatud, siis $\rho = 0$;
- kui tunnuste X ja Y vahel on funktsionaalne monotoonne seos, siis $|\rho| = 1$.



Kendalli korrelatsioonikordaja

Millal kasutada ja mida näitab?

Mõõdab kahe arv- või järjestustunnuse vahelise monotoonse seose tugevust ja suunda.

Ei ole tundlik erindite suhtes ega eelda tunnuste pidevust ja normaalset jaotumist (minimaalne nõue on vaatluste järjestatavus).

Arvutamine: vaadeldakse kõikvõimalikke väärtuste paare x_i-x_j ja y_i-y_j – kokku on selliseid $N=n(n-1)/2$ – ning loetakse kokku samasuunaliste ja vastassuunaliste erinevustega paarid, vastavalt n_s ja n_v .

Kendalli korrelatsioonikordaja avaldatakse seosest

$$\tau = \frac{n_s}{N} - \frac{n_v}{N} = 1 - \frac{2n_v}{N}.$$

Omadused: analoogsed astakkorrelatsioonikordaja omadustele.

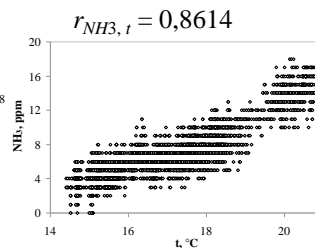
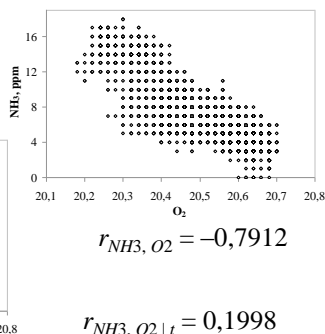
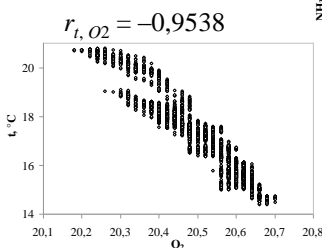
Osakorrelatsioonikordaja

Millal kasutada ja mida näitab?

Kasutatakse kirjeldamiseks kahe tunnuse vahelist seost elimineerides kolmanda tunnuse mõju.

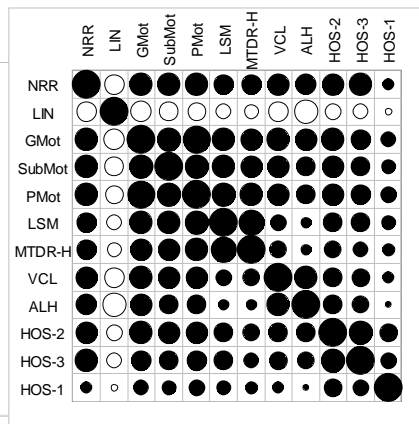
$$r_{XY.Z} = \frac{r_{XY} - r_{XZ}r_{YZ}}{\sqrt{(1-r_{XZ}^2)(1-r_{YZ}^2)}}$$

Näide.



Korrelatsioonimaatriks

	T	HOT1	HOT2	HOT3	Submot	mot
T	1.00000	0.15422 0.3118	0.62903 0.0001	0.65005 0.0001	0.68489 0.0001	0.68314 0.0001
HOT1	0.15422 0.3118	1.00000	0.39188 0.0078	0.32162 0.0312	0.26502 0.0785	0.30416 0.0422
HOT2	0.62903 0.0001	0.39188 0.0078	1.00000	0.71678 0.0001	0.58923 0.0001	0.64757 0.0001
HOT3	0.65005 0.0001	0.32162 0.0312	0.71678 0.0001	1.00000	0.46806 0.0012	0.50912 0.0004
Submot	0.68489 0.0001	0.26502 0.0785	0.58923 0.0001	0.46806 0.0012	1.00000	0.73547 0.0001
mot	0.68314 0.0001	0.30416 0.0422	0.64757 0.0001	0.50912 0.0004	0.73547 0.0001	1.00000
P_Mot	0.65639 0.0001	0.34532 0.0202	0.58423 0.0001	0.49611 0.0005	0.71525 0.0001	0.95432 0.0001
VCL	0.61413 0.0001	0.15815 0.2995	0.47451 0.0010	0.42156 0.0039	0.61050 0.0001	0.66258 0.0001
ALH	-0.52542 0.0002	-0.05709 0.7095	-0.34396 0.0207	-0.30667 0.0405	-0.43722 0.0027	-0.53114 0.0002
HOS-2	0.57381 0.0001	0.03706 0.8090	0.47624 0.0003	0.37158 0.0120	0.45994 0.0015	0.62856 0.0001
HOS-3	0.52203 0.0002	0.24141 0.1101	0.44450 0.0022	0.43701 0.0027	0.67219 0.0001	0.66186 0.0001
MTDR_H	0.50658 0.0004	0.20733 0.1718	0.37936 0.0102	0.30456 0.0419	0.62168 0.0001	0.59103 0.0001

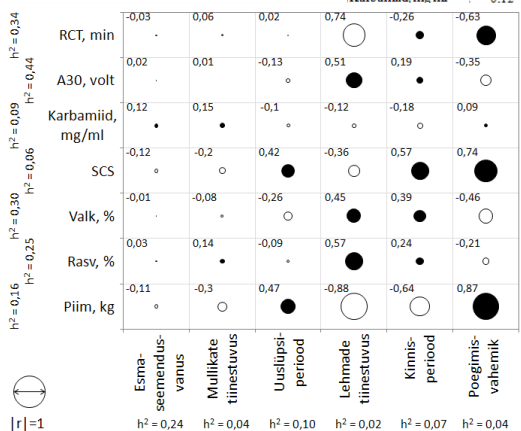


Tiinestumise (NRR) ja sperma kvaliteedinäitajate vahelised seosed. Ringide suurus (pindala) näitab seose tugevust (diagonaalil on tunnuste seosed iseendaga, st seos on maksimaalse tugevusega ehk lineaarne korrelatsioonikordaja $r = 1$) ja värvus seose suunda (mustad ringid vastavad positiivsele ja valged negatiivsele seosele).

Tabel 2. Sigimismäitajate ning piimatoodangu, -kvaliteedi ja -laapumismäitajate päritavused (h^2) ning geneetilised korrelatsioonid (sulgudes standardvead).

Tunnus	Ln(AF) $h^2 = 0,24$	hCR $h^2 = 0,04$	Ln(CF) $h^2 = 0,10$	cCR $h^2 = 0,02$	Ln(DP) $h^2 = 0,07$	Ln(CI) $h^2 = 0,04$
Piim, kg $h^2 = 0,16$	-0,11 (0,05)	-0,30 (0,12)	0,47 (0,07)	-0,88 (0,09)	-0,64 (0,07)	0,87 (0,06)
Rasv, % $h^2 = 0,25$	0,03 (0,03)	0,14 (0,03)	-0,09 (0,05)	0,57 (0,25)	0,24 (0,07)	-0,21 (0,12)
Valk, % $h^2 = 0,30$	-0,01 (0,03)	-0,08 (0,09)	-0,26 (0,06)	0,45 (0,16)	0,39 (0,06)	-0,46 (0,11)
SCS $h^2 = 0,06$	-0,12 (0,07)	-0,20 (0,14)	0,42 (0,10)	-0,36 (0,22)	0,57 (0,08)	0,74 (0,17)
Karbamiid, mg ml ⁻¹	0,12	0,15	-0,10	-0,12	-0,18	0,09
		(0,12)	(0,07)	(0,17)	(0,09)	(0,12)
		-0,01	-0,10	-0,33	0,11	0,22
		(0,08)	(0,06)	(0,17)	(0,07)	(0,12)
		0,01	-0,13	0,51	0,19	-0,35
		(0,07)	(0,04)	(0,18)	(0,07)	(0,09)
		-0,06	-0,02	-0,74	0,26	0,63
		(0,08)	(0,05)	(0,17)	(0,05)	(0,14)

endamisel; hCR – mulkikate tiinestuvus esmasseemendusel; ma de tiinestuvus esmasseemendusel; DP – kinnis-periood; g (SCC/100 000) = 3



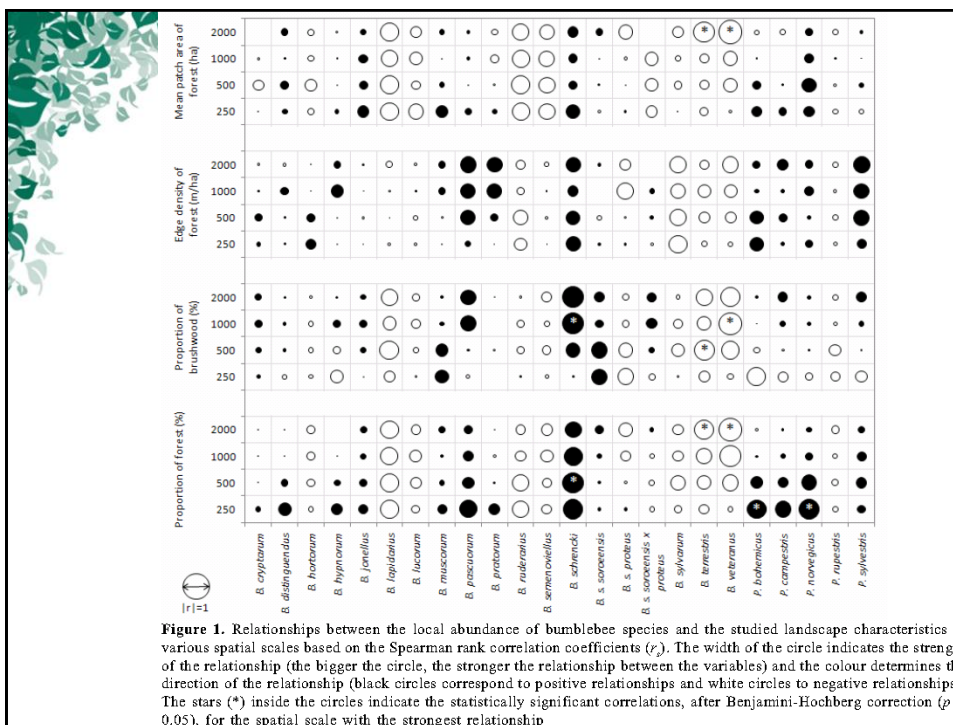


Figure 1. Relationships between the local abundance of bumblebee species and the studied landscape characteristics at various spatial scales based on the Spearman rank correlation coefficients (r_s). The width of the circle indicates the strength of the relationship (the bigger the circle, the stronger the relationship between the variables) and the colour determines the direction of the relationship (black circles correspond to positive relationships and white circles to negative relationships). The stars (*) inside the circles indicate the statistically significant correlations, after Benjamini-Hochberg correction ($p < 0.05$), for the spatial scale with the strongest relationship.

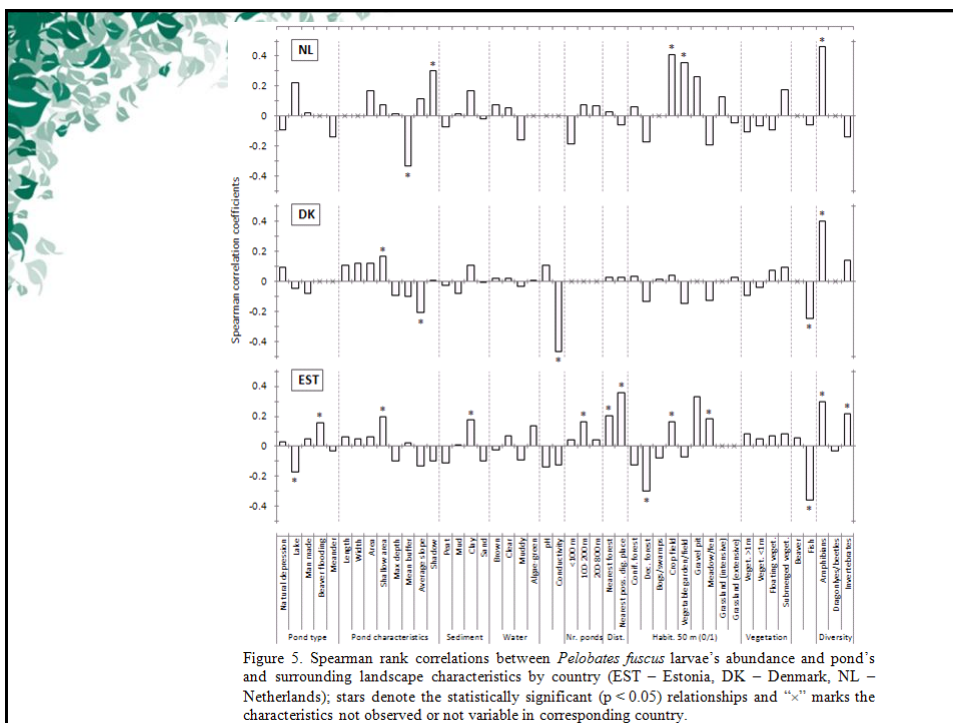
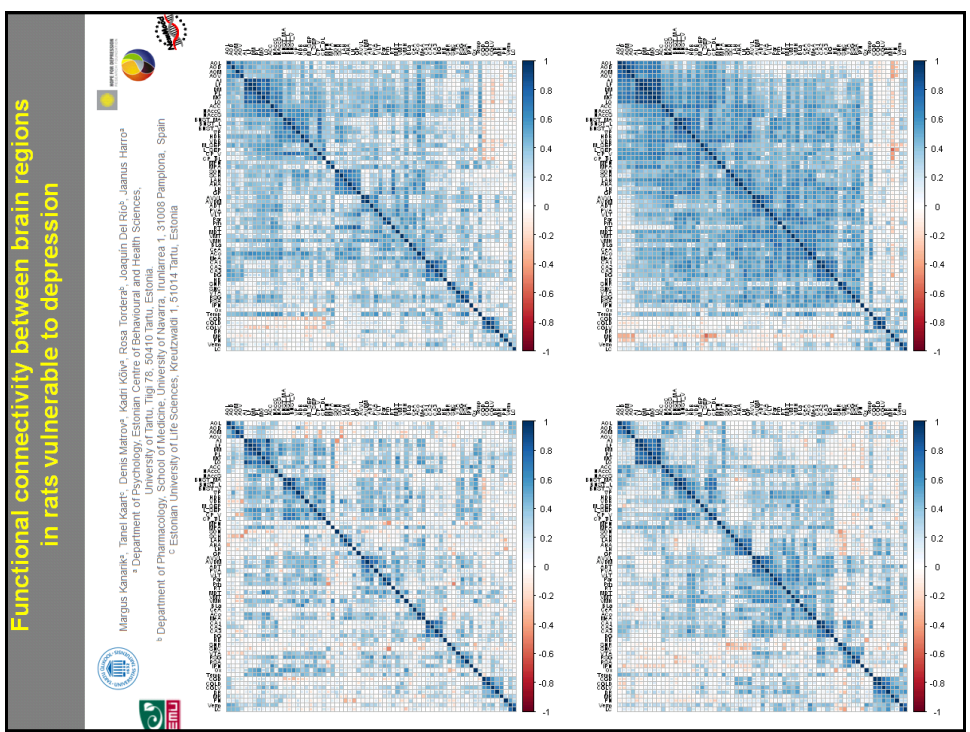
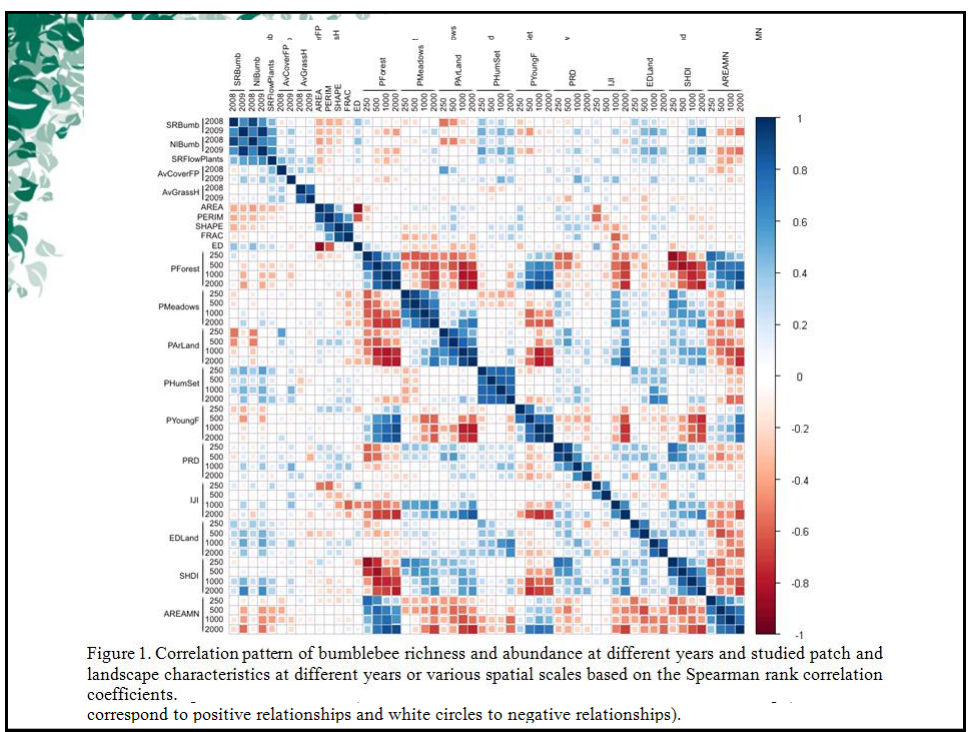


Figure 5. Spearman rank correlations between *Pelobates fuscus* larvae's abundance and pond's and surrounding landscape characteristics by country (EST – Estonia, DK – Denmark, NL – Netherlands); stars denote the statistically significant ($p < 0.05$) relationships and "x" marks the characteristics not observed or not variable in corresponding country.






Regressioonanalüüs

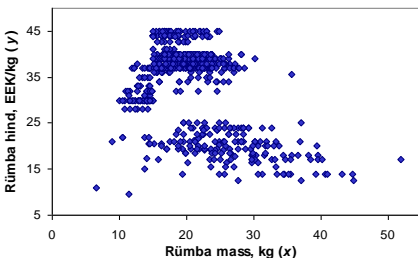
 **Eesti Maaülikool**
Estonian University of Life Sciences

www.emu.ee



Lineaarne regressioonanalüüs

Millal kasutada ja mida näitab?
Kasutatakse prognoosimaks ühe arvtunnuse väärtusi teis(t)e järgi.



Rümbe hind, EEK/kg (y)

Rümbe mass, kg (x)

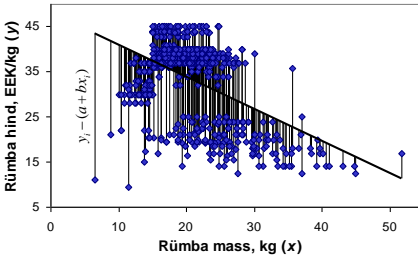
Regressioonivõrrand: $y_i = a + bx_i$

Tunnust y nimetatakse **funktsioon-** ja tunnust x **argumenttunnuseks**.

Näide.
Rümbe 1 kg hind = $a + b \cdot$ Rümbe mass

Regressioonivõrrandi parameetrid a ja b hinnatakse **vähimruutude meetodil**, st et minimeeritakse prognoosi jäägid:

$$\sum_{i=1}^n y_i - (a + bx_i)^2 \Rightarrow \min$$



Rümbe hind, EEK/kg (y)

Rümbe mass, kg (x)

$y_i - (a + bx_i)$

Lineaarne regressioonanalüüs

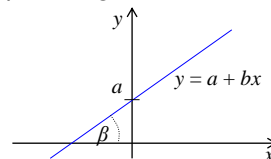
Regressioonivõrrandi parameetrite a ja b vähimruutude hinnangud:

$$b = \frac{\sum_{i=1}^n (x_i y_i - \bar{x}\bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad a = \bar{y} - b\bar{x}$$

Regressioonivõrrandi kordajate geomeetriline tähendus:

vabaliige a märgib kohta, kus regressioonisirge lõikab y -telge, ning regressioonikordaja b iseloomustab nurka, mille alla regressioonisirge kulgeb x -telje suhtes (matemaatilisemalt väljendudes $\tan(\beta) = b$, kus β on sirge tõusunurk).

Sisulise tähenduse kohaselt näitab regressioonikordaja b , kui mitme ühiku võrra muutub funktsioontunnuse väärtus, kui argumenttunnus muutub 1 ühiku võrra.



NB!

$$y = a + bx \quad \neq \quad x = -a/b + y/b$$

Lineaarne regressioonanalüüs

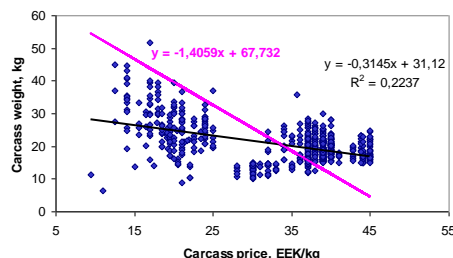
Näide.

Rümba 1 kg hind = $a + b \cdot \text{Rümba mass} = 48,178 + (-0,7113) \cdot \text{Rümba mass}$
Seega kaasneb lamba lihakeha massi suurenemisega 1 kg võrra
0,71-kroonine hinnalangus 1 kg liha eest.

Aga prognoosides teistpidi:

Rümba mass = $31,12 - 0,3145 \cdot \text{Rümba 1 kg hind}$

$$\neq \frac{-48,178 + \text{Rümba 1 kg hind}}{-0,7113}$$
$$= 67,732 - 1,406 \cdot \text{Rümba 1 kg hind}$$



Lineaarne regressioonanalüüs – jäägid

Näide.

Rümba 1 kg hind = $a + b \cdot \text{Rümba mass}$
= $48,178 + (-0,7113) \cdot \text{Rümba mass}$

Vaatluse jrk nr (i)	Tegelik rümba 1 kg hind (y_i)	Prognoositud rümba 1 kg hind ($a + b \cdot x_i$)	Prognoosi jäägid $y_i - (a + b \cdot x_i)$
1	39	31,250	7,750
2	39	37,367	1,633
3	40	37,083	2,917
4	39	30,824	8,176
5	39	33,171	5,829
6	37	34,593	2,407
7	39	32,175	6,825
8	39	32,744	6,256
9	37	31,250	5,750
10	34	31,677	2,323
11	37	35,304	1,696
12	33	37,936	-4,936
13	33	39,216	-6,216
14	37	33,597	3,403
15	39	32,175	6,825
...

Regressioonimudeli sobivus

Determinatsioonikordaja R^2 ütleb, kui suure osa uuritava tunnuse varieeruvusest mudel ära kirjeldab, $0 \leq R^2 \leq 1$. Mida suurem, seda parem!

Leitakse kui mudelile vastava hajuvuskomponendi $SS_1 = \sum_{i=1}^n (a + bx_i) - \bar{y}$ ja uuritava tunnuse koguhajuvust kirjeldava hälvete ruutude summa $SS = \sum_{i=1}^n (y_i - \bar{y})^2$ jagatis: $R^2 = SS_1/SS$.

Mudeli standardviga SE on mudeli prognoosijäägi standardhälve. Mida väiksem, seda parem!

Hüpoteeside kontroll

1) Hüpotees mudeli, kui terviku kohta (võrreldakse konstrueeritud mudeli ja nn konstantse mudeli $y=a$ jääkide varieeruvust):

H_0 : mudel ei ole parem võrreldes konstantse mudeliga,

H_1 : mudel on parem võrreldes konstantse mudeliga.

2) Hüpoteesid mudeli parameetrite kohta – kontrollitakse väidet iga parameetri nullist erinemise kohta: $H_0: a = 0$ $H_0: b = 0$

$H_1: a \neq 0$ $H_1: b \neq 0$

Regressioonimudeli sobivus

Näide.

Rümba 1 kg hind = $a + b \cdot \text{Rümba mass} = 48,178 + (-0,7113) \cdot \text{Rümba mass}$

MS Exceli protseduuri Regression väljund:

SUMMARY OUTPUT

Mitmene korrelatsioonikordaja – mõõdab uuritava tunnuse ja tema prognoositud väärtuste vahelist korrelatsiooni. Mida suurem, seda parem!

Regression Statistics	
Multiple R	0,4730
R Square	0,2237
Adjusted R Square	0,2226
Standard Error	7,8450
Observations	686

Determinatsioonikordaja R^2 ja selle väikeste valimite tarvis kohandatud [adjusted] väärtus

Mudeli standardviga

H_0 : mudel ei ole parem võrreldes konstantse mudeliga
 H_1 : mudel on parem võrreldes konstantse mudeliga

ANOVA

	df	SS	MS	F	Significance F
Regression	1	12130,21573	12130,2157	197,0999	1,57702E-39
Residual	684	42095,74764	61,5435		
Total	685	54225,96337			

Hüpoteeside kontrollimiseks mudeli iga parameetri kohta:
 H_0 : $a = 0$
 H_1 : $a \neq 0$
 H_0 : $b = 0$
 H_1 : $b \neq 0$

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	48,1782	1,0843	44,4341	8,2819E-204	46,0493	50,3070
R_mass	-0,7113	0,0507	-14,0392	1,5770E-39	-0,8107	-0,6118

Mudeli parameetrite hinnangud

Regressioonanalüüsi eeldused

Regressioonivõrrandi parameetrite hindamine **ei eelda** tunnuste jaotumist vastavalt normaaljaotuse seaduspäradele!

Mudeli täpsuse ja statistilise olulisuse hindamiseks peavad:

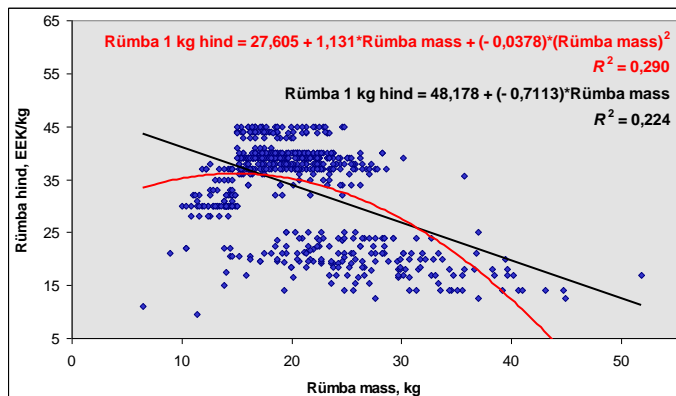
- mudeli (prognoosi) jäägid olema ligikaudu normaaljaotusega (kontrollimiseks histogramm, tõenäosuspaber);
- mudeli jäägid olema ühtlase varieeruvusega (hajuvusdiagramm).

Ükskõik kumma eelduse rikutuse korral ei pruugi mudeli kohta käivate hüpoteeside kontrollimisel arvatavate teststatistikute jaotusseadused kehtida, mistõttu ei pruugi õiged olla ka otsustused mudeli sobivuse ja rakendatavuse üle.

Eelkõige teise eelduse paikapidamatus võib vihjata mittesobivale mudelile (vale matemaatiline funktsioon, mõni arvestamata jäänud argument vmt).

Regressioonanalüüsi mudeli valik/diagnostika

Näide. Seos lamba lihakeha massi ja makstava 1 kg hinna vahel ($n = 686$).

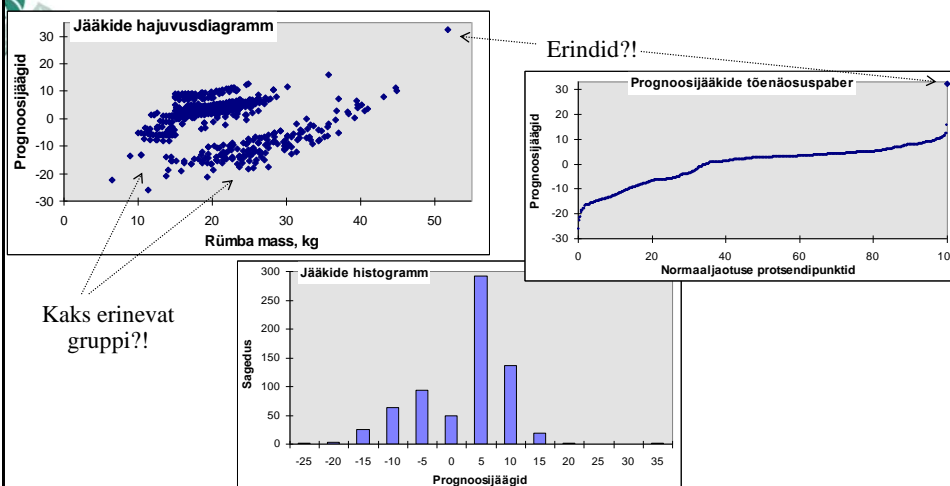


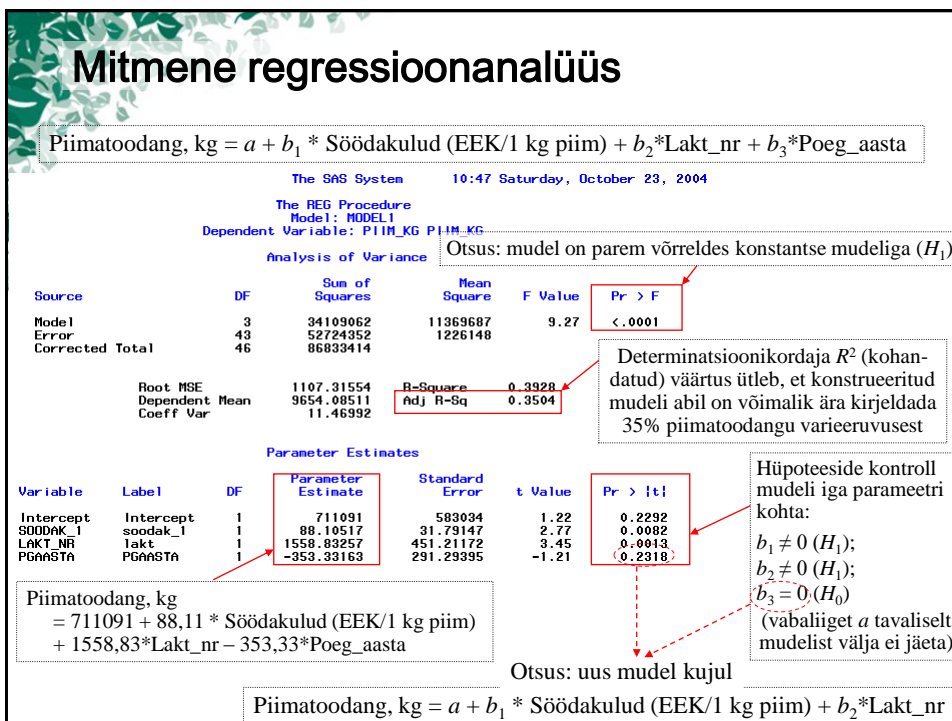
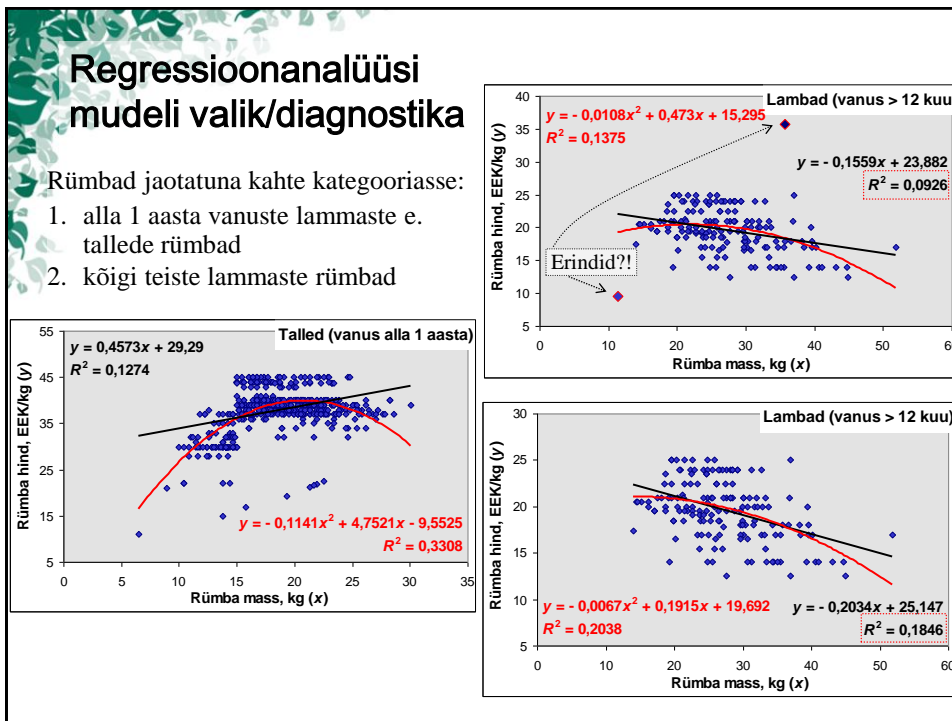
Regressioonanalüüsi mudeli jääkide analüüs

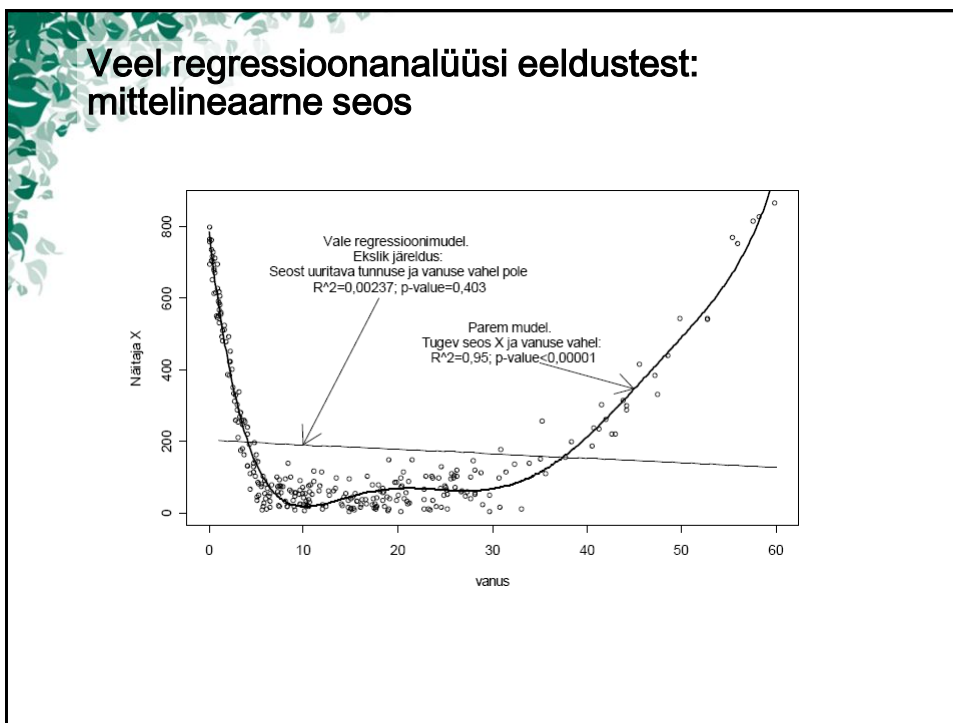
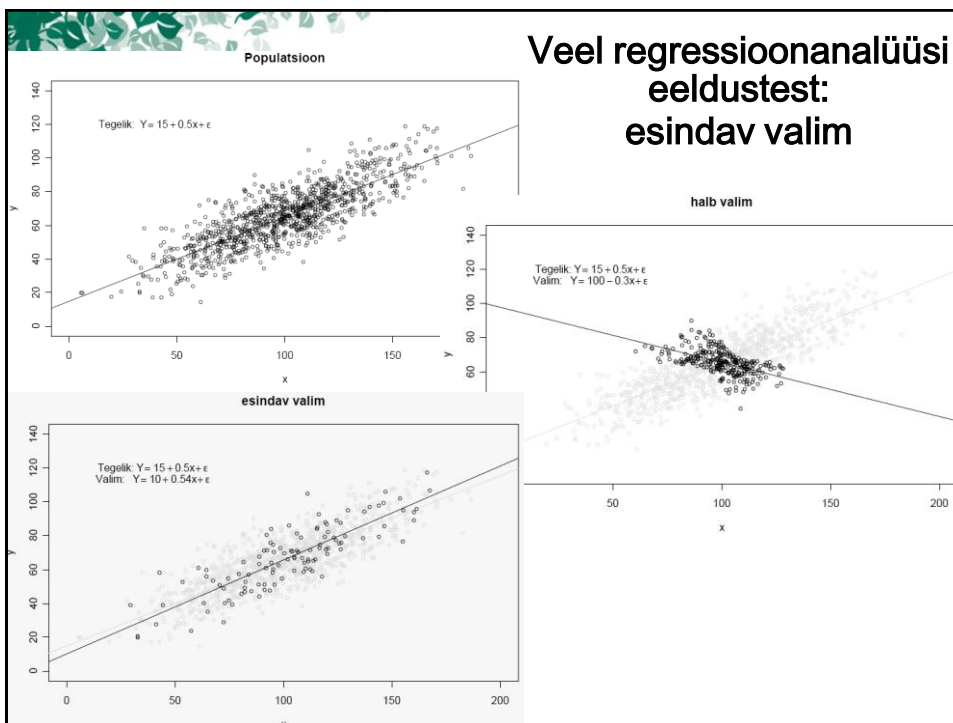
Näide. Seos lamba lihakeha massi ja makstava 1 kg hinna vahel ($n = 686$).

Prognosijäägid on leitud ruutvõrrandi baasil:

$$\text{Rümba 1 kg hind} = 27,605 + 1,131 \cdot \text{Rümba mass} + (-0,0378) \cdot (\text{Rümba mass})^2$$

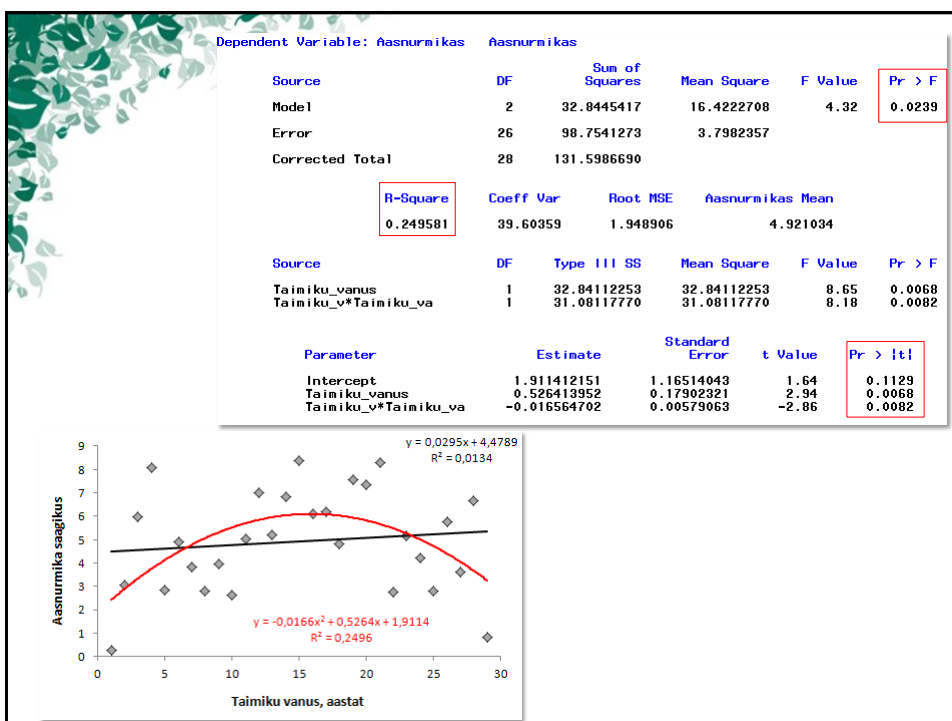
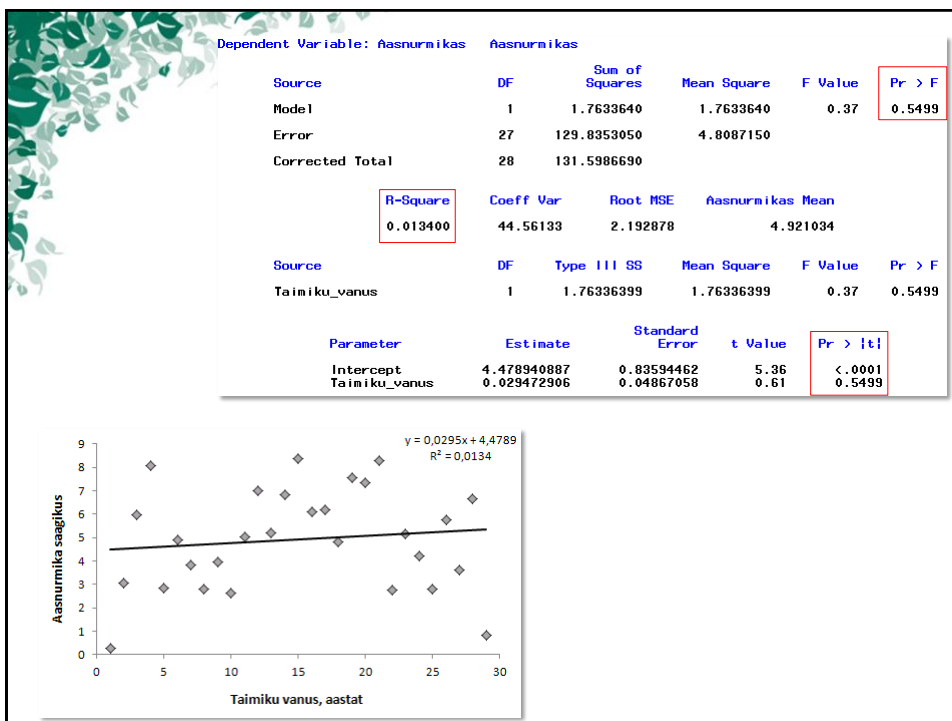


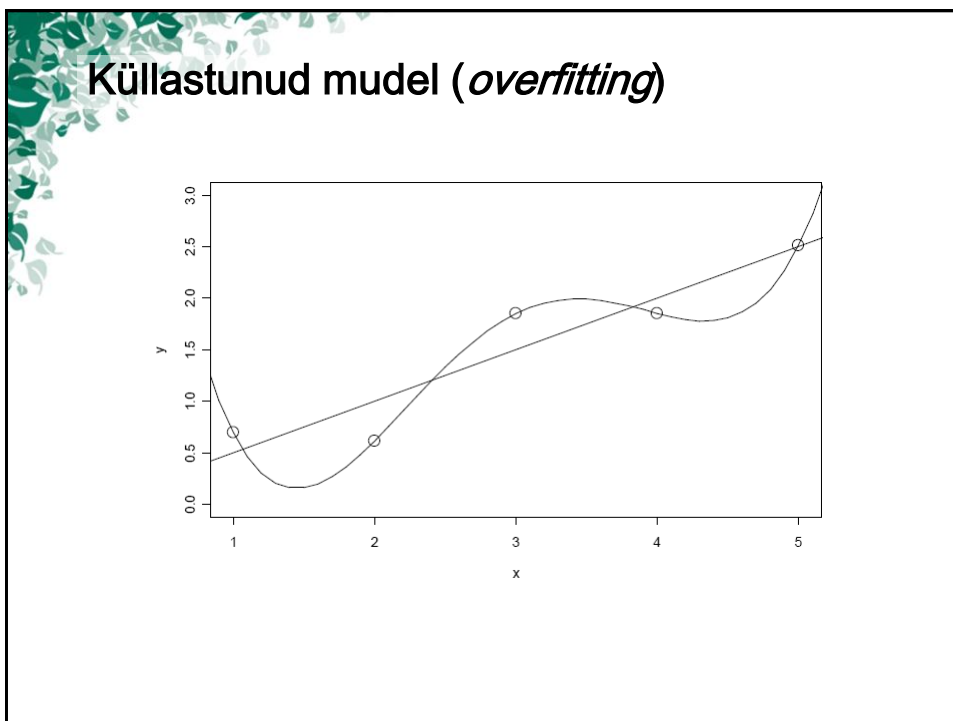
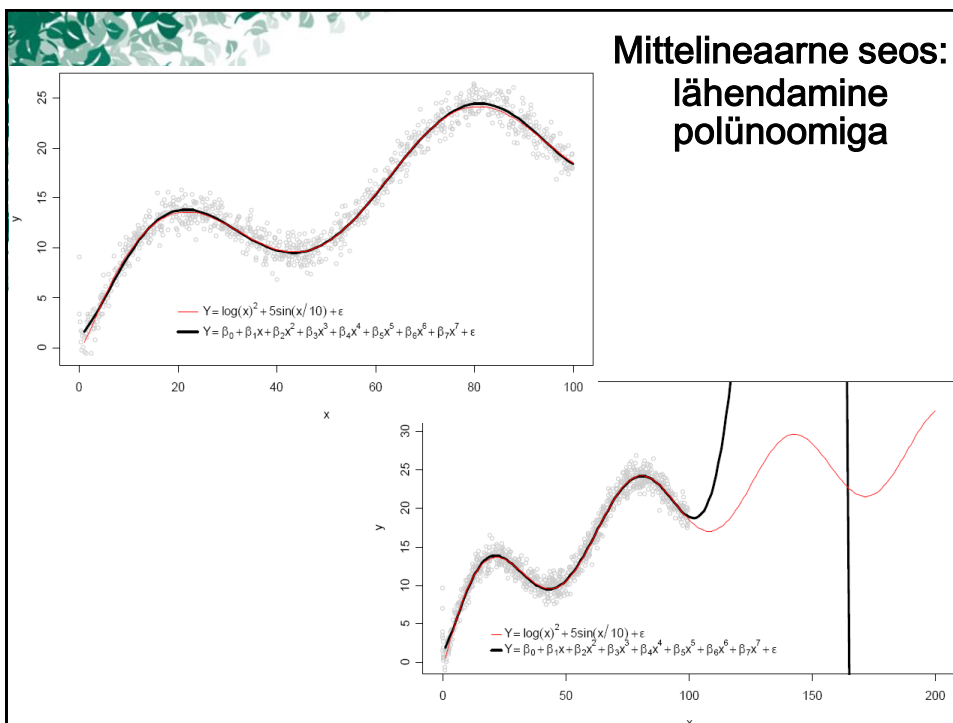


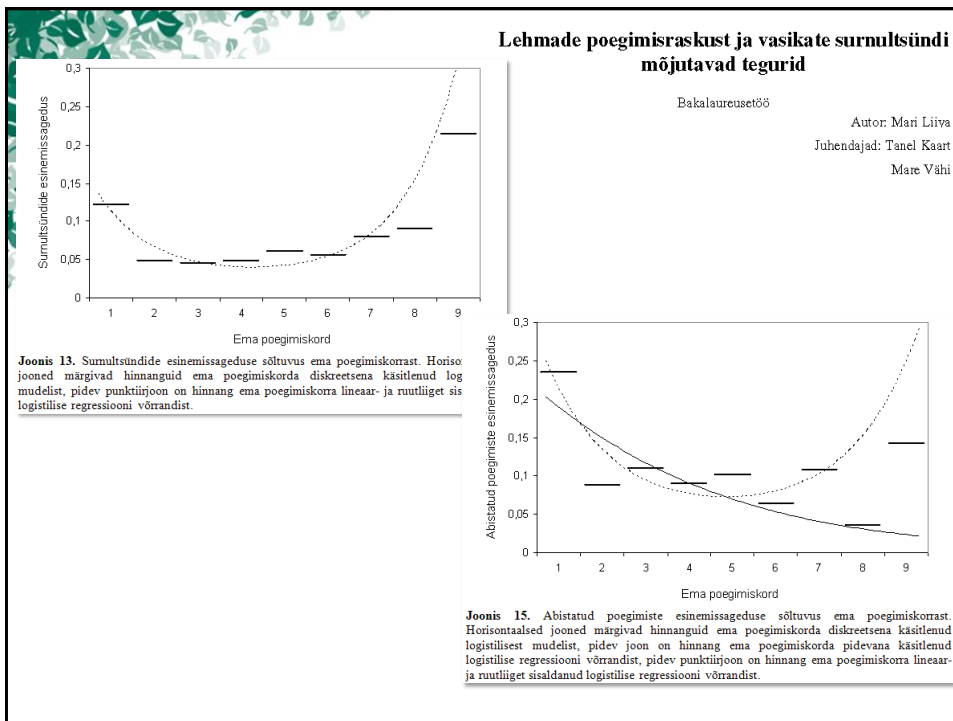
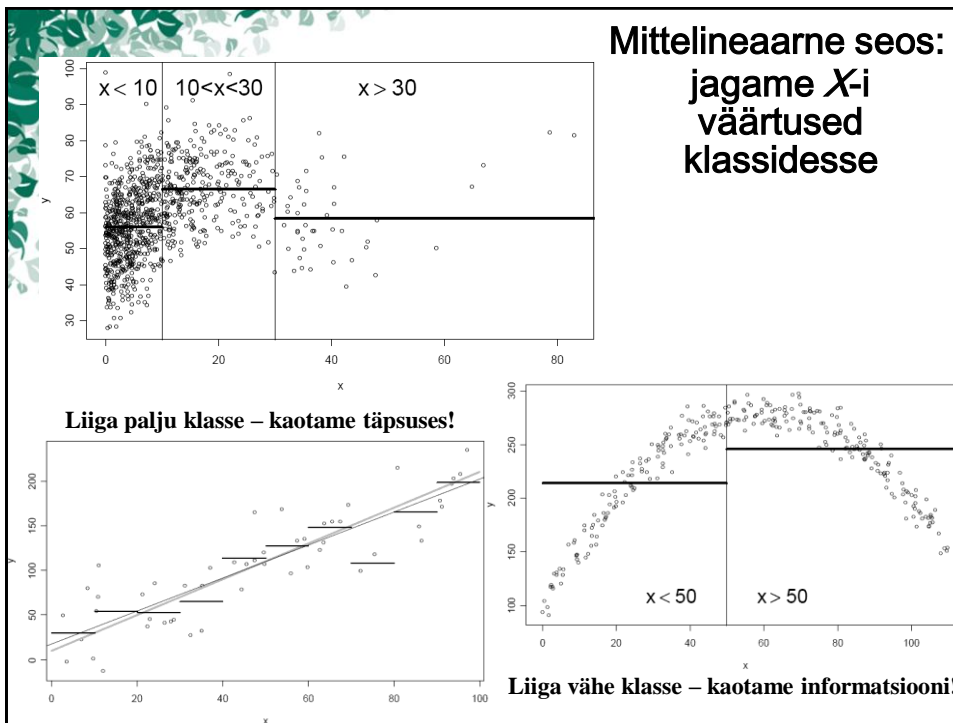


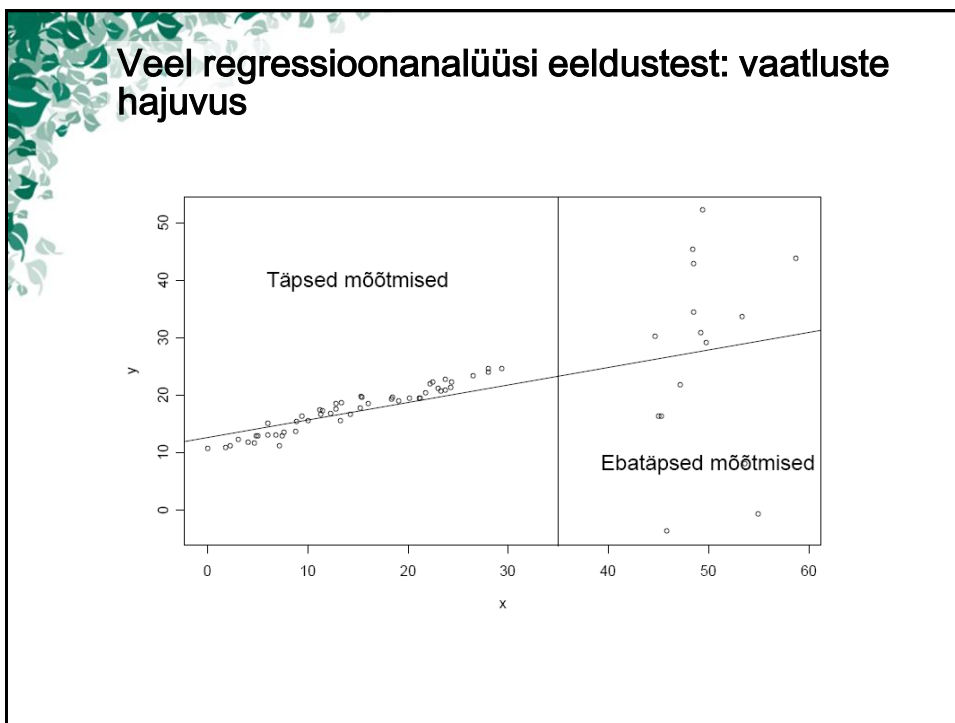
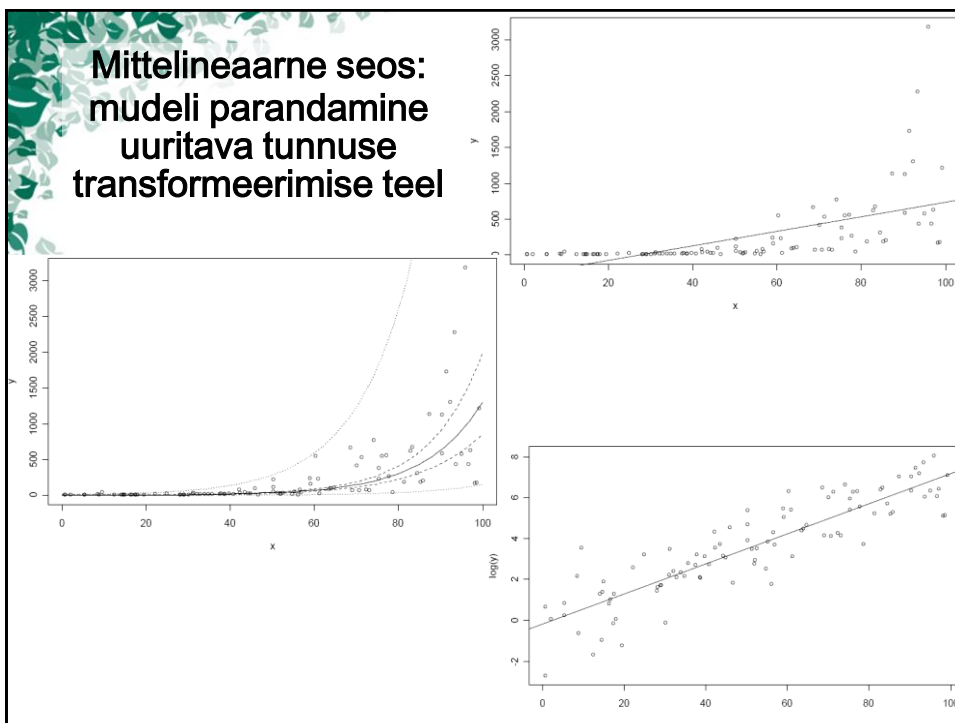
Statistika ja modelleerimise baaskoolitus

õppejõududele ja juhendajatele



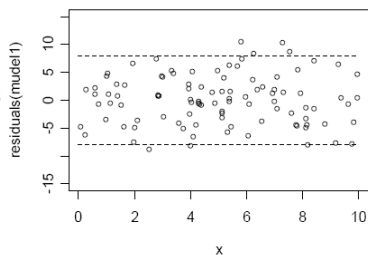




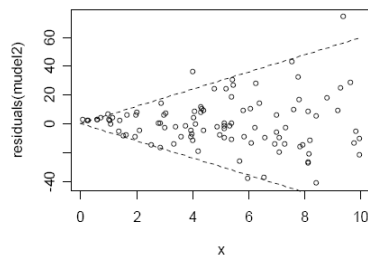


Veel regressioonanalüüsi eeldustest: jääkide hajuvus

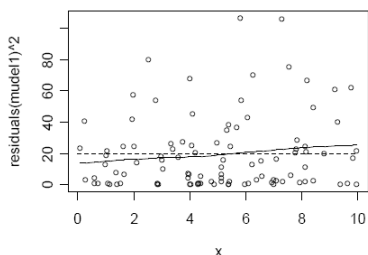
jääkide hajuvus on konstantne



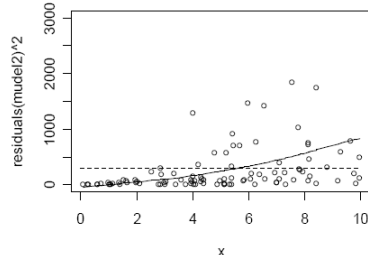
jääkide hajuvus pole konstantne



jääkide hajuvus on konstantne



jääkide hajuvus pole konstantne



Multikollineaarsus

Olukorda, kus argumenttunnused (sõltumatud muutujad) on omavahel küllalt tugevalt seotud, nimetatakse **multikollineaarsuseks**.

Sellisel on tulemuseks ebatäpsed hinnangud (võivad olla isegi vale märgiga) ja seega ka ebatäpsed prognoosid.

Lisaks tekib probleeme regressioonikordajate tõlgendamisel – kui argumenttunnuste vahel pole sõltuvust, määrab parameeter keskmise sõltuva tunnuse muutuse kui vastav argumenttunnus muutub ühiku võrra ja teised argumenttunnused ei muutu. Argumenttunnuste sõltuvuse korral aga muutuvad argumendid üheaegselt.

Multikollineaarsust loetakse suureks, kui argumentide vaheline korrelatsioon on suurem kui samade argumentide ja uuritava tunnuse vaheline korrelatsioon.

Multikollineaarsus

Mõõduks:

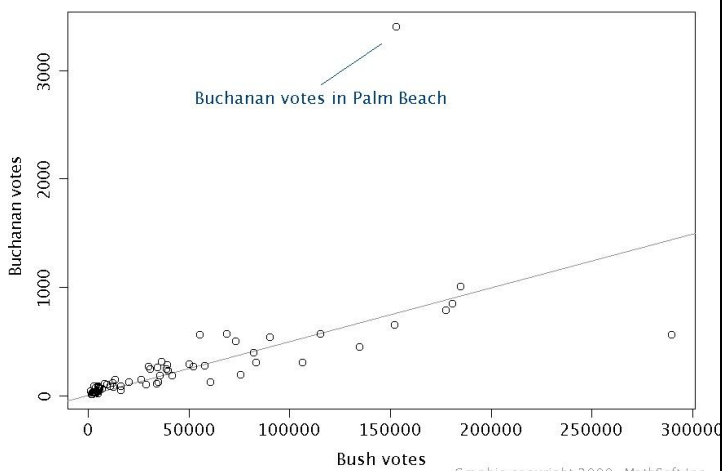
- o **tolerants** (*tolerance*) TOL on multikollineaarsuse mõõt, mis näitab kui suur osa argumenti varieeruvusest jääb ülejäänud argumentide poolt kirjeldamata;
- o **varieeruvusindeks** ehk dispersiooni mõju faktor (*variance inflation factor*) VIF näitab argumenti mõju regressiooniparameetri hajuvusele ja on tolerantsi pöördväärtus.

Empiiriline kriteerium: kui $TOL < 0,15$ või $VIF > 10$ on tegemist multikollineaarsusega.

Näide. USA 2000. aasta presidendivalimised (Bush *versus* Gore) otsustasid valijate eelistused Florida osariigis. Palm Beach'i maakonna valijad protestisid tulemused väites, et valimisedeli mitmetimõistetavuse tõttu hääletati Gore asemel 3. kandidaadi, Buchanan'i poolt.

USA kohus proteste ei arvestanud ja kordusvalimisi ei tulnud, Georg W. Bush võitis Al Gore'i Floridas <1000 häälega ning valiti Florida valijameeste häälega aastaiks 2000-2004 USA presidendiks ...

Bush/Buchanan votes in Florida counties



Graphic copyright 2000, MathSoft Inc.

