Matemaatiline statistika ja modelleerimine

Kahe arvtunnuse ühine käitumine, korrelatsioon- ja regressioonanalüüs

EMÜ doktorikool
DK.0007

Tanel Kaart

Lineaarne e Pearsoni korrelatsioonikordaja

Millal kasutada ja mida näitab?
Mõõdab kahe pideva (normaaljaotusega) tunnuse vahelise lineaarse seose tugevust ja suunda.

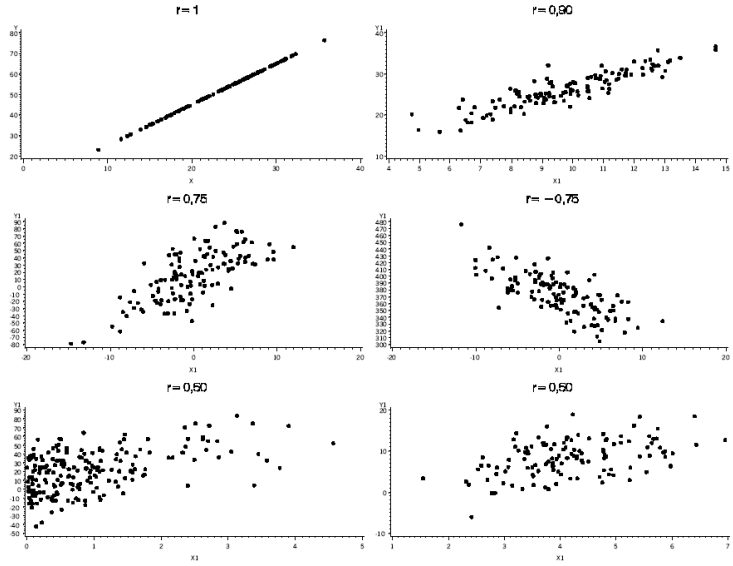
Arvutusvalem: $$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_i - \bar{y})^2}}$$

Omadused:

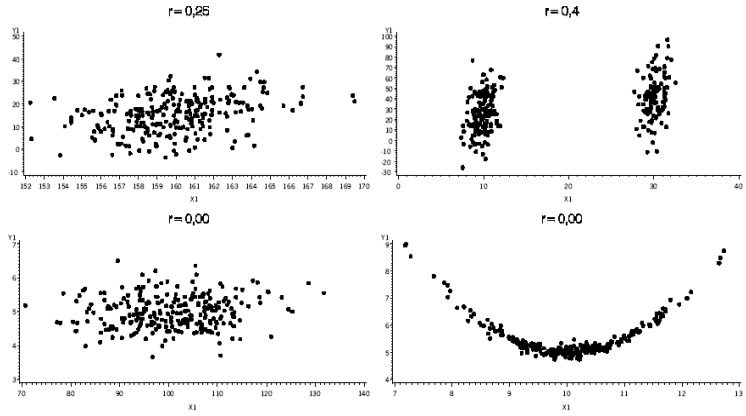
- $-1 \le r \le 1$;
- kui $r > 0$, siis tunnuse X suurenedes suureneb keskmiselt ka tunnus Y; kui $r < 0$, siis X-i suurenedes Y keskmiselt kahaneb ja X-i kahanedes Y keskmiselt suureneb;
- kui tunnused X ja Y on sõltumatud, siis $r = 0$;
- kui tunnuste X ja Y vahel on täpne lineaarne seos, siis $|r| = 1$;
- 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



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.

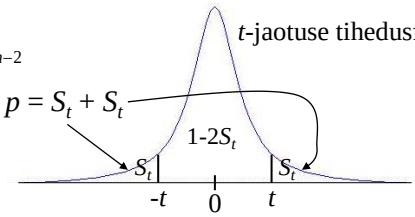
Seose statistiline olulisus

$H_0 : r = 0$

$H_1 : r \neq 0$

Teststatistik $t = r\sqrt{n-2}\Big/\sqrt{1-r^2} \underset{H_0}{\sim} t_{n-2}$

Olulisustõenäosus $p = S_t + S_t$

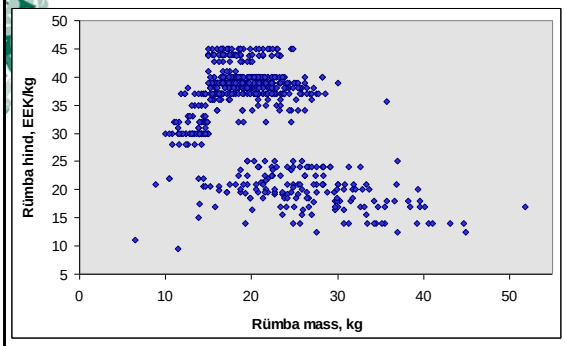


Näiteks *MS Excel*is saab olulisustõenäosuse *p* leidmiseks kasutada funktsiooni *TDIST*(*t*;*n*-2;2).

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

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

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

$= -14{,}039,$

millest $|t| = 14{,}039$.

Viimase alusel leitav olulisuse tõenäosus $p = 1{,}577 * 10^{-37} < 0{,}05$,

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 kehamõõtude 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***

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

Osa SAS-i protseduuri CORR väljundist

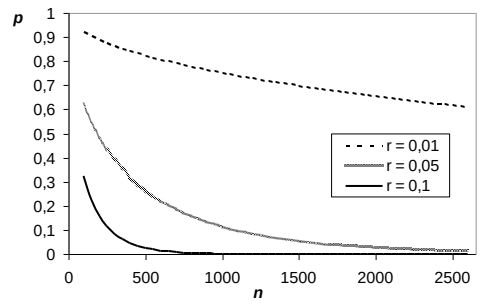
Pearson Correlation Coefficients, N = 1380
Prob > |r| under H0: Rho=0

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

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 $p < 0,05$.



Astakkorrelatsioonikordaja 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:

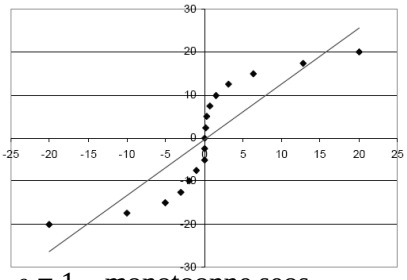
$$\rho = 1 - \frac{6\sum_{i=1}^{n}(x_{(i)} - y_{(i)})^2}{n(n^2-1)},$$

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

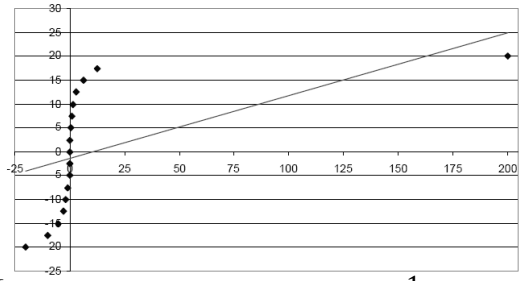
Omadused:

- $-1 \le \rho \le 1$;
- 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 $|\rho| = 1$.

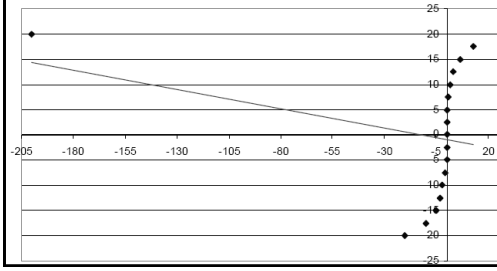
Lineaarne *versus* astakkorrelatsioonikordaja



$\rho = 1$ – monotoonne seos
$r = 0{,}873$ – lineaarne seos

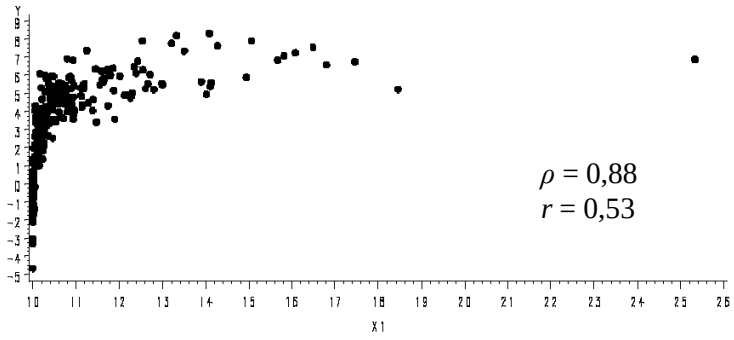


$\rho = 1$
$r = 0{,}513$



$\rho = 0{,}667$
$r = -0{,}295$

Lineaarne *versus* astakkorrelatsioonikordaja



$\rho = 0{,}88$
$r = 0{,}53$

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

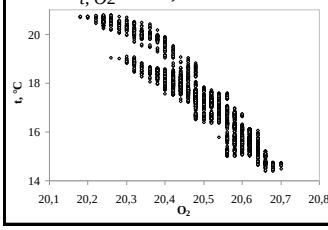
Millal kasutada ja mida näitab?

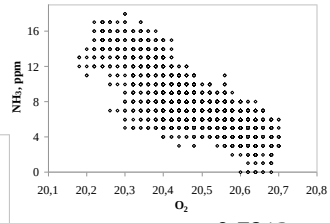
Kasutatakse kirjeldamaks kahe tunnuse vahelist seost elimineerides kolmanda tunnuse mõju.

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

Näide.

$r_{t,\,O2} = -0{,}9538$

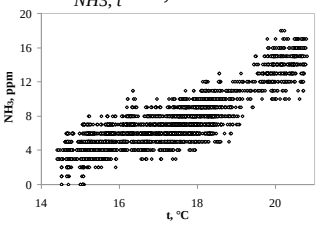




$r_{NH3,\,O2} = -0{,}7912$

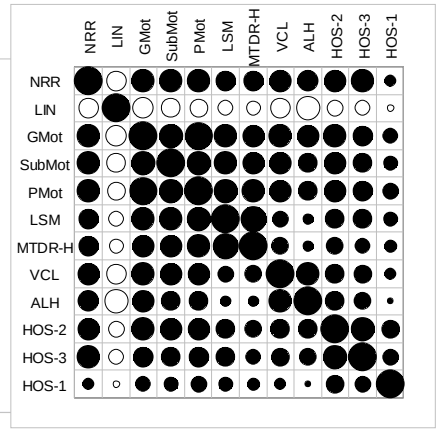
$r_{NH3,\,O2\,|\,t} = 0{,}1998$

$r_{NH3,\,t} = 0{,}8614$



Korrelatsioonimaatriks

	T	HOT1	HOT2	HOT3	Submot	mot
T	1.00000	0.15422	0.62903	0.65005	0.68489	0.68314
T		0.3118	<.0001	<.0001	<.0001	<.0001
HOT1	0.15422	1.00000	0.39188	0.32162	0.26502	0.30416
HOT1	0.3118		0.0078	0.0312	0.0785	0.0422
HOT2	0.62903	0.39188	1.00000	0.71678	0.58923	0.64757
HOT2	<.0001	0.0078		<.0001	<.0001	<.0001
HOT3	0.65005	0.32162	0.71678	1.00000	0.46806	0.50912
HOT3	<.0001	0.0312	<.0001		0.0012	0.0004
Submot	0.68489	0.26502	0.58923	0.46806	1.00000	0.73547
Submot	<.0001	0.0785	<.0001	0.0012		<.0001
mot	0.68314	0.30416	0.64757	0.50912	0.73547	1.00000
mot	<.0001	0.0422	<.0001	0.0004	<.0001	
P_Mot	0.65639	0.34532	0.58423	0.49611	0.71525	0.95432
P_Mot	<.0001	0.0202	<.0001	0.0005	<.0001	<.0001
VCL	0.61413	0.15815	0.47451	0.42156	0.61050	0.66258
VCL	<.0001	0.2995	0.0010	0.0039	<.0001	<.0001
LIN	-0.52542	-0.05709	-0.34396	-0.30667	-0.43722	-0.53114
LIN	0.0002	0.7095	0.0207	0.0405	0.0027	0.0002
ALH	0.57381	0.03706	0.47624	0.37158	0.45994	0.62856
ALH	<.0001	0.8030	0.0009	0.0120	0.0015	<.0001
L_st_m	0.52203	0.24141	0.44450	0.43701	0.67219	0.66186
L_st_m	0.0002	0.1101	0.0022	0.0027	<.0001	<.0001
MTDR_H	0.50658	0.20733	0.37936	0.30456	0.62168	0.59183
MTDR_H	0.0004	0.1718	0.0102	0.0419	<.0001	<.0001



Tiinestumise (NRR) ja sperma kvaliteedinäitajate vahelised seosed.

Ringide suurus (pindala) näitab seose tugevust

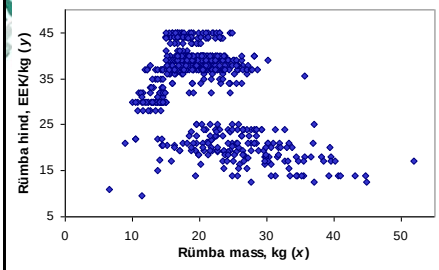
(diagonaalil on tunnuste seosed iseendaga, st seos on maksimaalse tugevusega ehk lineaarne korrelatsioonikordaja $r = 1$) ja

värvus seose suunda

(mustad ringid vastavad positiivsele ja valged negatiivsele seosele).

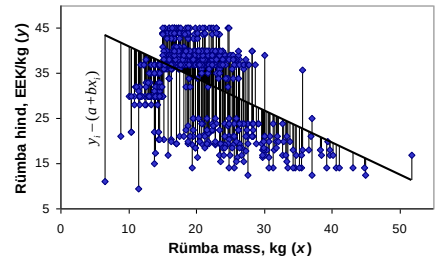
Lineaarne regressioonanalüüs

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



Regressioonivõrrand: $y_i = a + bx_i$
Tunnust *y* nimetatakse **funktsioon-** ja tunnust *x* **argumenttunnuseks**.

Näide.
Rümba 1 kg hind = *a* + *b**Rümba mass



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

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

Lineaarne regressioonanalüüs

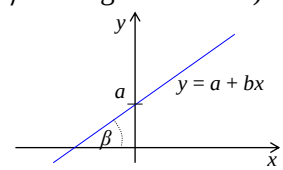
Regressioonivõrrandi parameetrite *a* ja *b* vähimruutude hinnangud:

$$b = \sum_{i=1}^{n} (x_i y_i - \overline{xy}) \Big/ \sum_{i=1}^{n} (x_i - \overline{x})^2, \quad a = \overline{y} - b\overline{x}$$

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(β) = *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!

$y = a + bx \quad \not\Leftrightarrow \quad x = -a/b + y/b$

Lineaarne regressioonanalüüs

Näide.

Rümba 1 kg hind = a + b*Rümba mass = 48,178 + (–0,7113)*Rümba mass

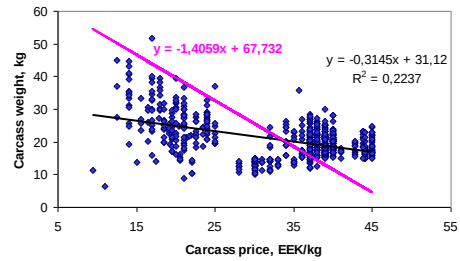
Seega kaasneb lamba lihakeha massi suurenemisega 1 kg võrra 0,71-kroonine hinnalangus 1 kg liha eest.

Aga prognoosides teistpidi:

Rümba mass = 31,12 – 0,3145*Rümba 1 kg hind

$\neq$ (–48,178 + Rümba 1 kg hind) / –0,7113

= 67,732 – 1,406*Rümba 1 kg hind



Lineaarne regressioonanalüüs – jäägid

Näide.

Rümba 1 kg hind = a + b*Rümba mass
= 48,178 + (–0,7113)*Rümba mass

Vaatluse jrk nr (i)	Tegelik rümba 1 kg hind (y_i)	Prognoositud rümba 1 kg hind ($a+b*x_i$)	Prognoosi jäägid $y_i-(a+b*x_i)$
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 varieeruvu-sest mudel ära kirjeldab, $0 \le R^2 \le 1$. 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*Rümba mass = 48,178 + (–0,7113)*Rümba mass

MS Exceli protseduuri *Regression* väljund:

SUMMARY OUTPUT

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

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

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			

	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

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

Regressioonanalüüsi eeldused

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

Mudeli täpsuse ja **statistilise olulisuse** hindamiseks peavad:

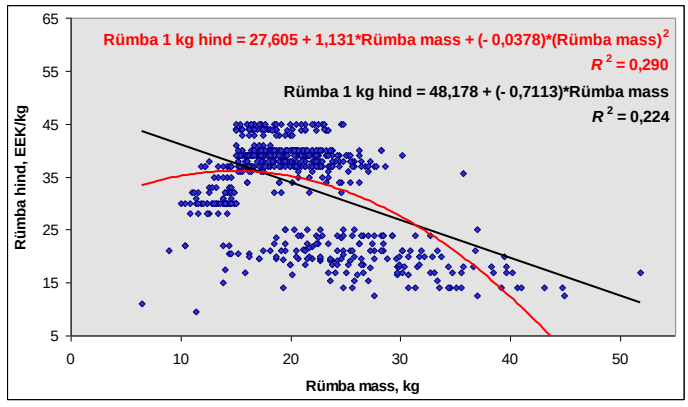
- mudeli (prognoosi) jäägid olema ligikaudu normaaljaotusega (kontrollimiseks histogramm, tõenäosuspaber);
- mudeli jäägid olema ühtlase varieeruvusega (hajuvusdiagramm).

Ükskõik kumma eelduse rikutuse korral ei pruugi mudeli kohta käivate hüpoteeside kontrollimisel arvutatavate 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).

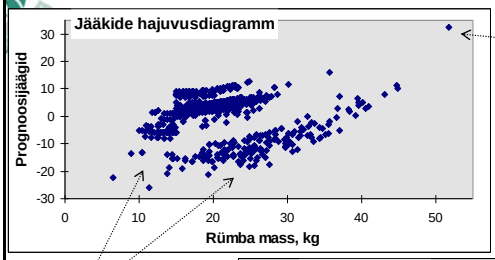
Regressioonanalüüsi mudeli valik/diagnostika

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

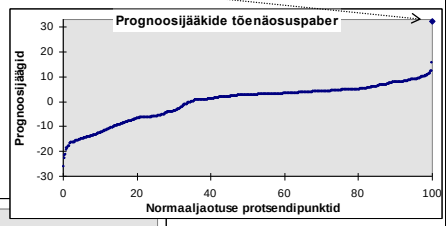


Regressioonanalüüsi mudeli jääkide analüüs

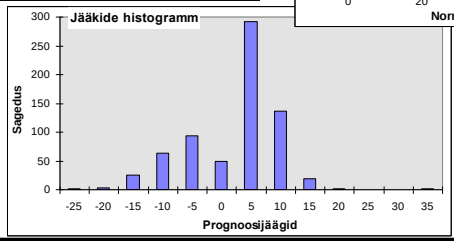
Näide. Seos lamba lihakeha massi ja makstava 1 kg hinna vahel (n = 686).
Prognoosijäägid on leitud ruutvõrrandi baasil:
Rümba 1 kg hind = 27,605 + 1,131*Rümba mass + (- 0,0378)*(Rümba mass)^2



Erindid?!



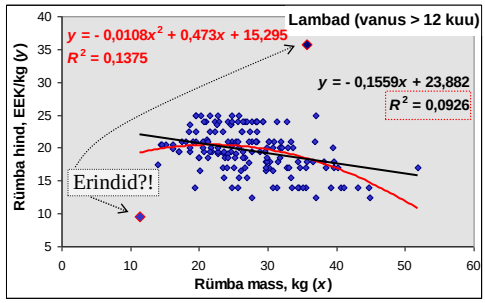
Kaks erinevat gruppi?!

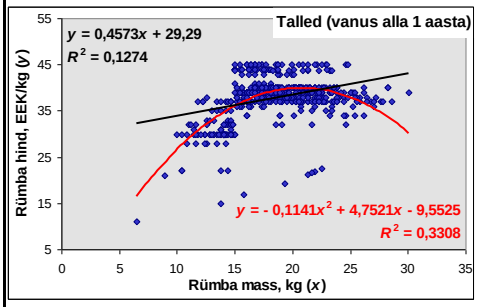


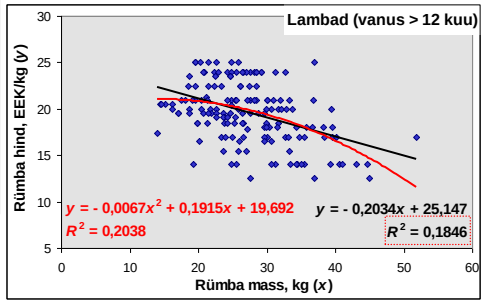
Regressioonanalüüsi mudeli valik/diagnostika

Rümbad jaotatuna kahte kategooriasse:

1. alla 1 aasta vanuste lammaste e. tallede rümbad
2. kõigi teiste lammaste rümbad







Mitmene regressioonanalüüs

Piimatoodang, kg = $a + b_1$ * Söödakulud (EEK/1 kg piim) + b_2*Lakt_nr + b_3*Poeg_aasta

The SAS System 10:47 Saturday, October 23, 2004

The REG Procedure
Model: MODEL1
Dependent Variable: PIIM_KG PIIM_KG

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	34109062	11369687	9.27	<.0001
Error	43	52724352	1226148		
Corrected Total	46	86833414			

Root MSE	1107.31554	R-Square	0.3928
Dependent Mean	9654.08511	Adj R-Sq	0.3504
Coeff Var	11.46992		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	711091	583034	1.22	0.2292
SOODAK_1	soodak_1	1	88.10517	31.79147	2.77	0.0082
LAKT_NR	lakt	1	1558.83257	451.21172	3.45	0.0013
PGAASTA	PGAASTA	1	-353.33163	291.29395	-1.21	0.2318

Otsus: mudel on parem võrreldes konstantse mudeliga (H_1)

Determinatsioonikordaja R^2 (kohandatud) väärtus ütleb, et konstrueeritud mudeli abil on võimalik ära kirjeldada 35% piimatoodangu varieeruvusest

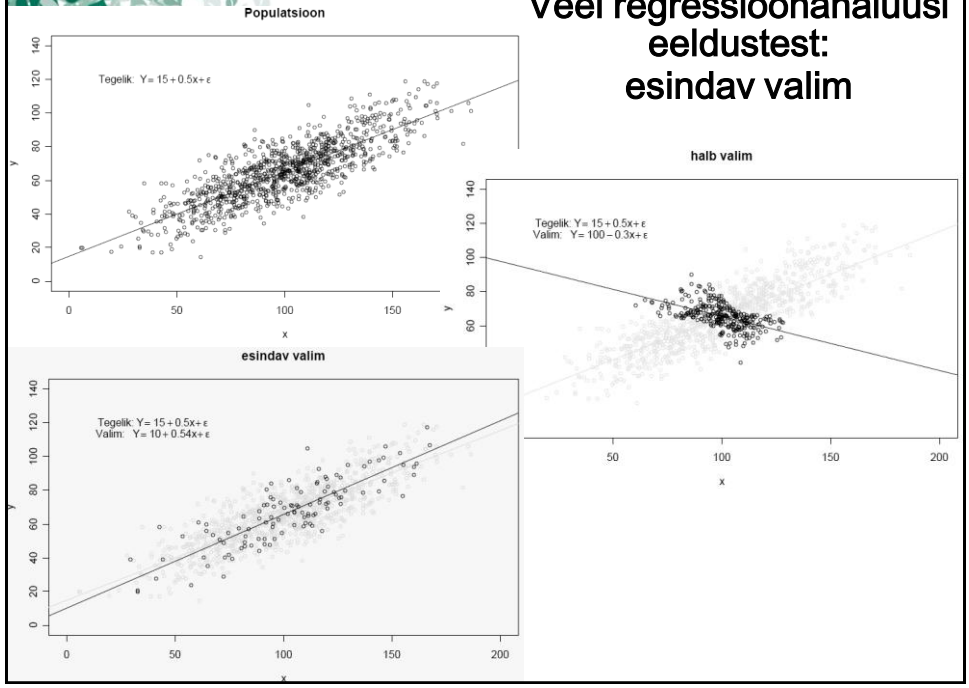
Hüpoteeside kontroll mudeli iga parameetri kohta:
$b_1 \neq 0$ (H_1);
$b_2 \neq 0$ (H_1);
$b_3 = 0$ (H_0)
(vabaliiget a tavaliselt mudelist välja ei jäeta)

Piimatoodang, kg
= 711091 + 88,11 * Söödakulud (EEK/1 kg piim)
+ 1558,83*Lakt_nr – 353,33*Poeg_aasta

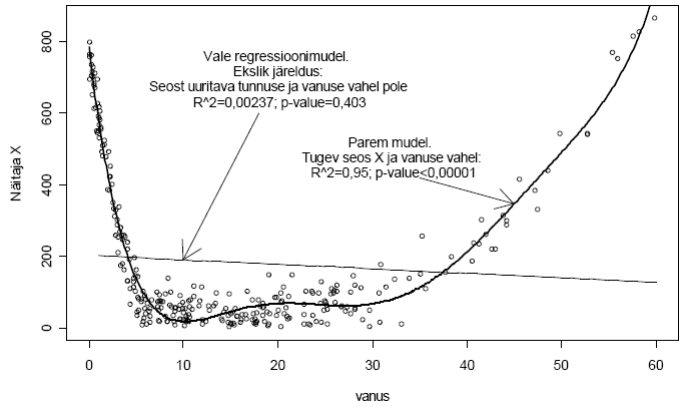
Otsus: uus mudel kujul

Piimatoodang, kg = $a + b_1$ * Söödakulud (EEK/1 kg piim) + b_2*Lakt_nr

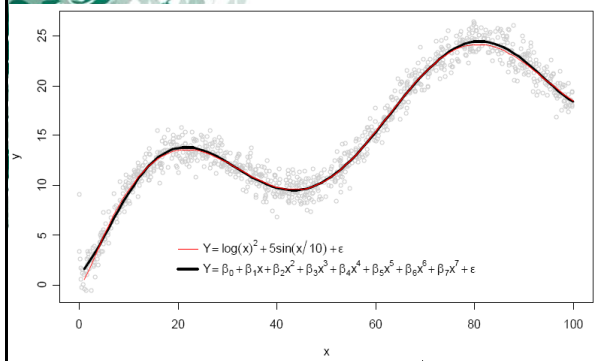
Veel regressioonanalüüsi eeldustest: esindav valim

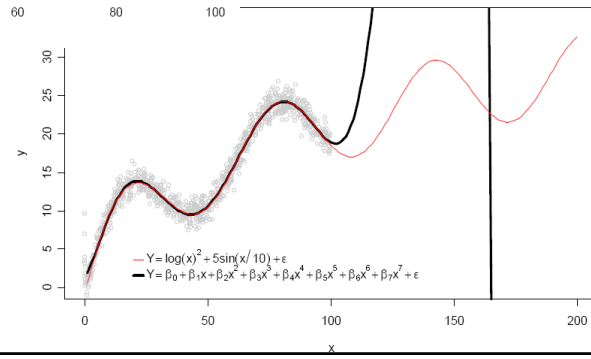


Veel regressioonanalüüsi eeldustest: mittelineaarne seos

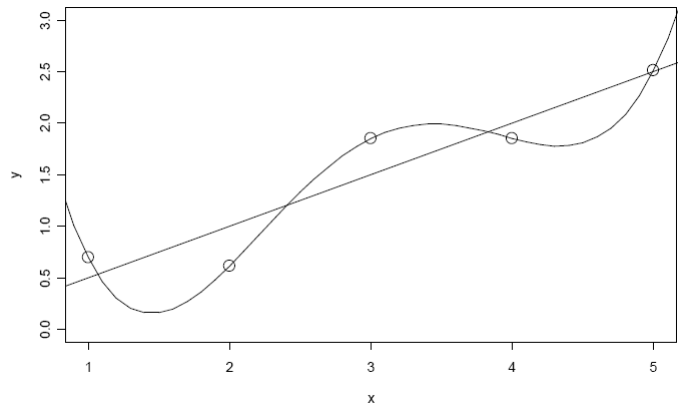


Mittelineaarne seos: lähendamine polünoomiga

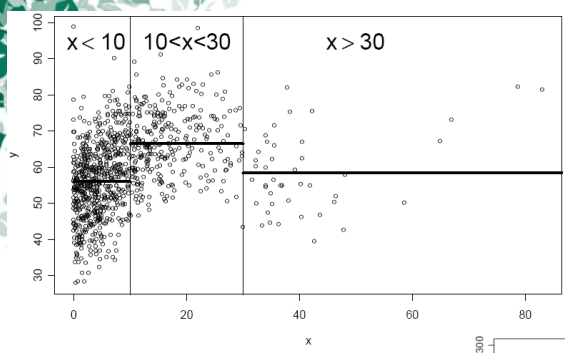




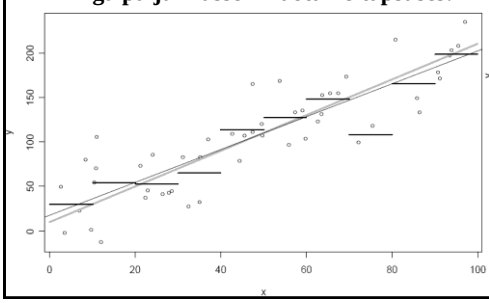
Küllastunud mudel (*overfitting*)

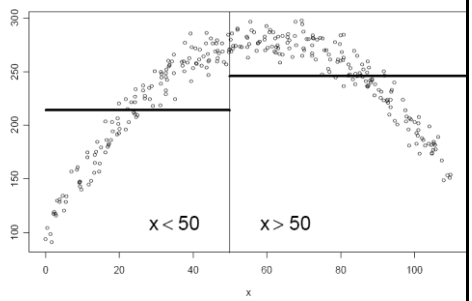


Mittelineaarne seos: jagame *X*-i väärtused klassidesse



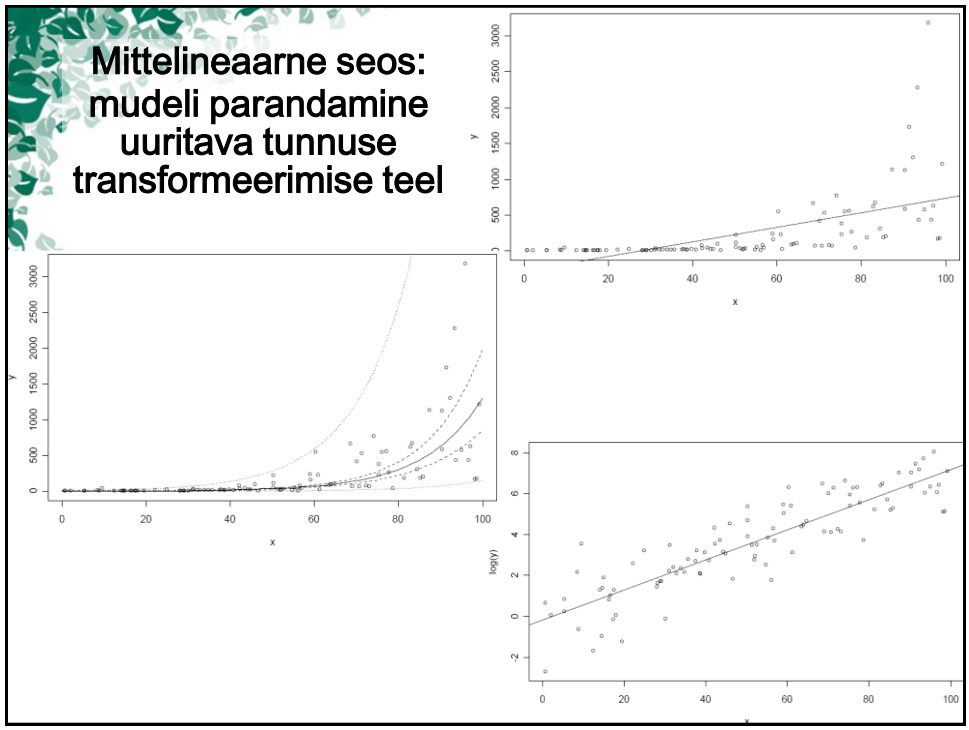
Liiga palju klasse – kaotame täpsuses!



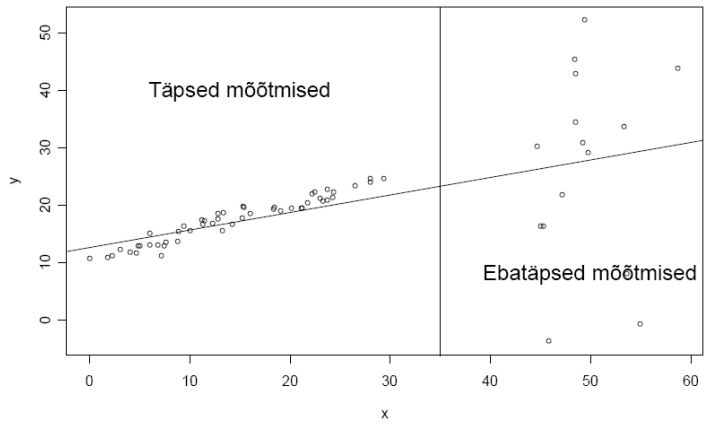


Liiga vähe klasse – kaotame informatsiooni!

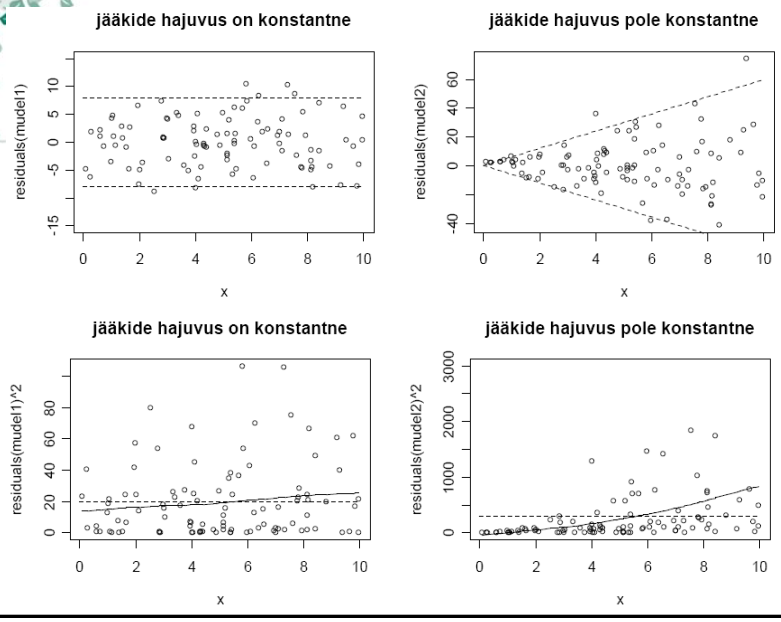
Mittelineaarne seos: mudeli parandamine uuritava tunnuse transformeerimise teel



Veel regressioonanalüüsi eeldustest: vaatluste hajuvus



Veel regressioonanalüüsi eeldustest: jääkide hajuvus



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

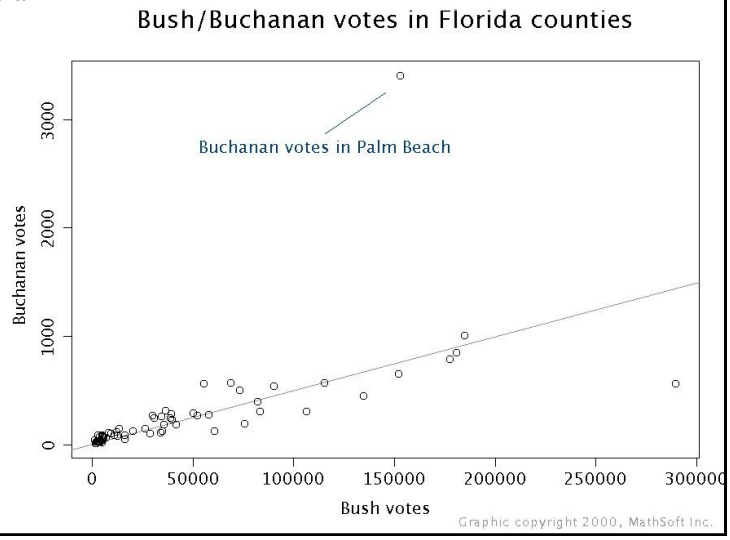
Mõõduks:

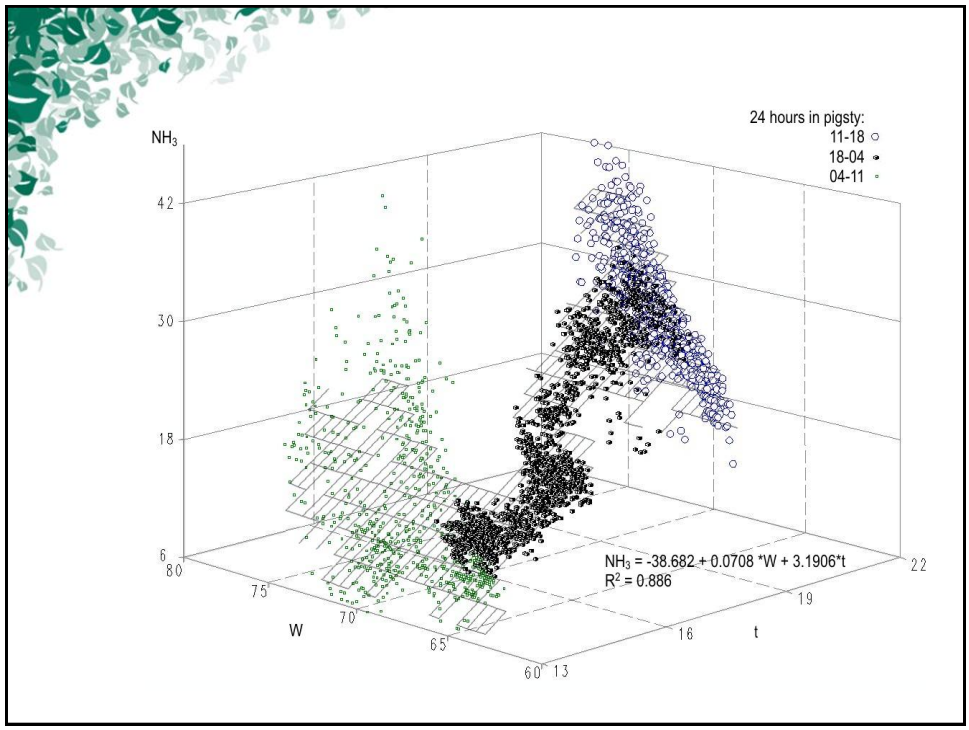
- **tolerants** (*tolerance*) TOL on multikollineaarsuse mõõt, mis näitab kui suur osa argumendi varieeruvusest jääb ülejäänud argumentide poolt kirjeldamata;
- **varieeruvusindeks** ehk dispersiooni mõju faktor (*variance inflation factor*) VIF näitab argumendi mõju regressiooniparameetri hajuvusele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

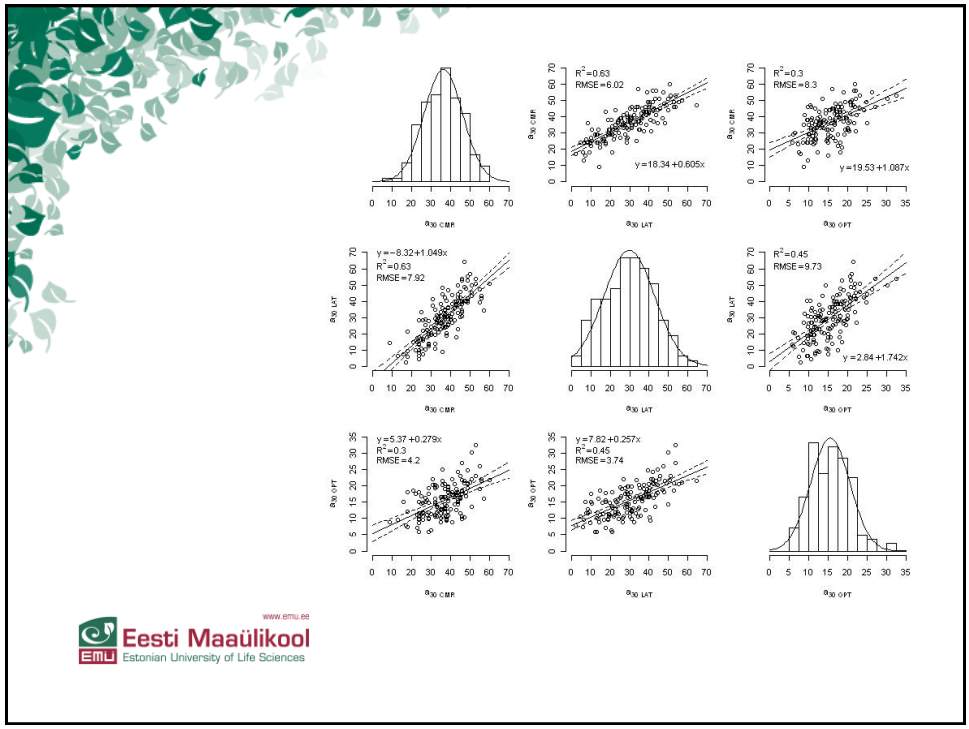
Empiiriline kriteerium: kui TOL < 0,15 või VIF > 10 on tegemist multikollineaarsusega.

Näide. USA 2000. aasta presidendivalimised (Bush *versus* Gore) otsustasid valijate eelistused Florida osariigis. Palm Beach'i maakonna valijad protestisid tulemused väites, et valimissedeli 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 häältega aastaiks 2000-2004 USA presidendiks …




24 hours in pigsty:
11-18
18-04
04-11
NH3
42
30
18
6
80
75
70
65
60
W
13
16
19
22
t
NH3 = -38.682 + 0.0708 *W + 3.1906*t
R2 = 0.886


R2=0.63
RMSE=6.02
y=18.34+0.605x
R2=0.3
RMSE=8.3
y=19.53+1.087x
y=-8.32+1.049x
R2=0.63
RMSE=7.92
R2=0.45
RMSE=9.73
y=2.84+1.742x
y=5.37+0.279x
R2=0.3
RMSE=4.2
y=7.82+0.257x
R2=0.45
RMSE=3.74
a30 CMF
a30 LAT
a30 OPT
www.emu.ee
Eesti Maaülikool
Estonian University of Life Sciences