

Biomeetria

Kahe arvtunnuse ühine käitumine, korrelatsioon- ja regressioonanalüüs

Lineaarne e Pearsoni korrelatsioonikordaja



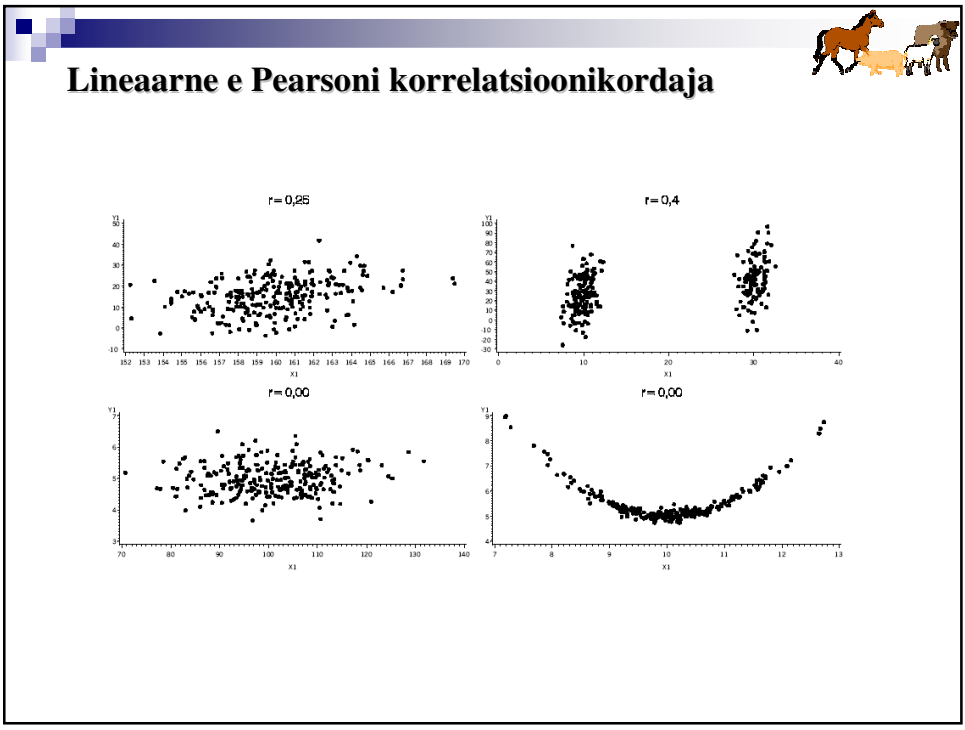
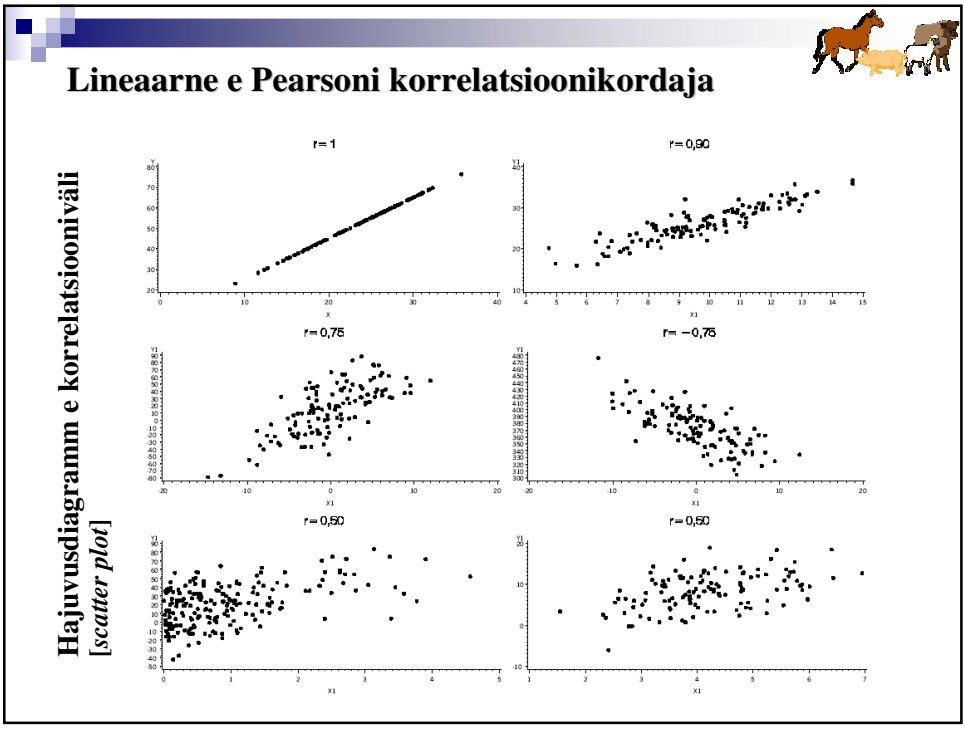
Millal kasutada ja mida näitab?

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



Kokkuleppedelised piirid seose tugevuse iseloomustamiseks:

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

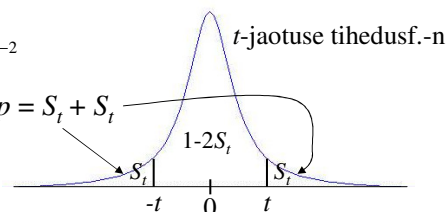
Seose statistiline olulisus

$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

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

$$\text{Olulisuse tõenäosus } p = S_t + S_t$$



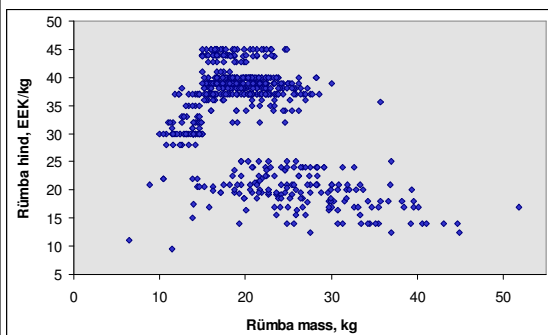
Näiteks *MS Excelis* saab olulisustõenäosuse p leidmiseks kasutada funktsiooni $\text{TDIST}(t;n-2;2)$.

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 tõestatult 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.

Lineaarne e Pearsoni korrelatsioonikordaja



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			<.0001	0.0060
Tiiva_p	0.06082	0.21007	1.00000	0.25273
Tiiva_p				<.0001
Iminokk	-0.03518	0.07390	0.25273	1.00000
Iminokk				

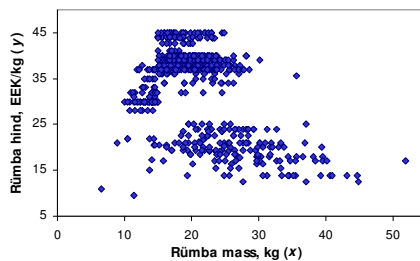
Osa SAS-i protseduuri
CORR väljundist

Lineaarne regressioonanalüüs



Millal kasutada ja mida näitab?

Kasutatakse prognoosimaks ühe arvutunnuse väärtusi teis(t)e järgi.



Regressioonivõrrand: $y_i = a + bx_i$

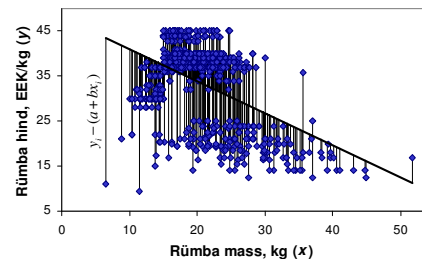
Tunnust y nimetatakse funktsioon- ja tunnust x argumenttunnuseks.

Näide.

Rümbe 1 kg hind = $a + b$ *Rümbe mass

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

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



Lineaarne regressioonanalüüs



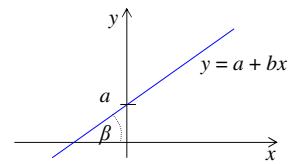
Regressioonivõrrandi parameetrite a ja b vähimruutude hinnangud:

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

Regressioonivõrrandi kordajate geomeetriline tähendus:

vabaliige a märgib kohta, kus regressioonisirge lõikab y -telge, ning regressioonikordaja b iseloomustab regressioonisirge ja x -telje vahelist nurka (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.



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.

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	Tegelik rümba 1 kg hind	Proгноositud rümba 1 kg hind	Proгноosi jäägid
(i)	(y_i)	($a + b \cdot x_i$)	$y_i - (a + b \cdot x_i)$
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 $SS_1 = \sum_{i=1}^n [(a + bx_i) - \bar{y}]^2$ 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!

Hüpooteeside 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

Mitme 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			

Hüpooteeside kontroll 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).

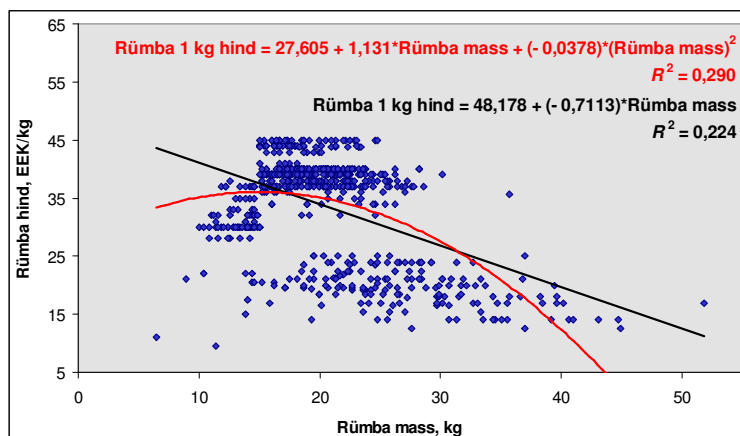
Ükskõik kumma eelduse rikutuse korral ei pruugi mudeli kohta käivate hüpoteeside kontrollimisel arvatavate teststatistikute jaotusseadused kehtida, mistõttu ei pruugi õiged olla ka otsustused mudeli sobivuse ja rakendatavuse üle.

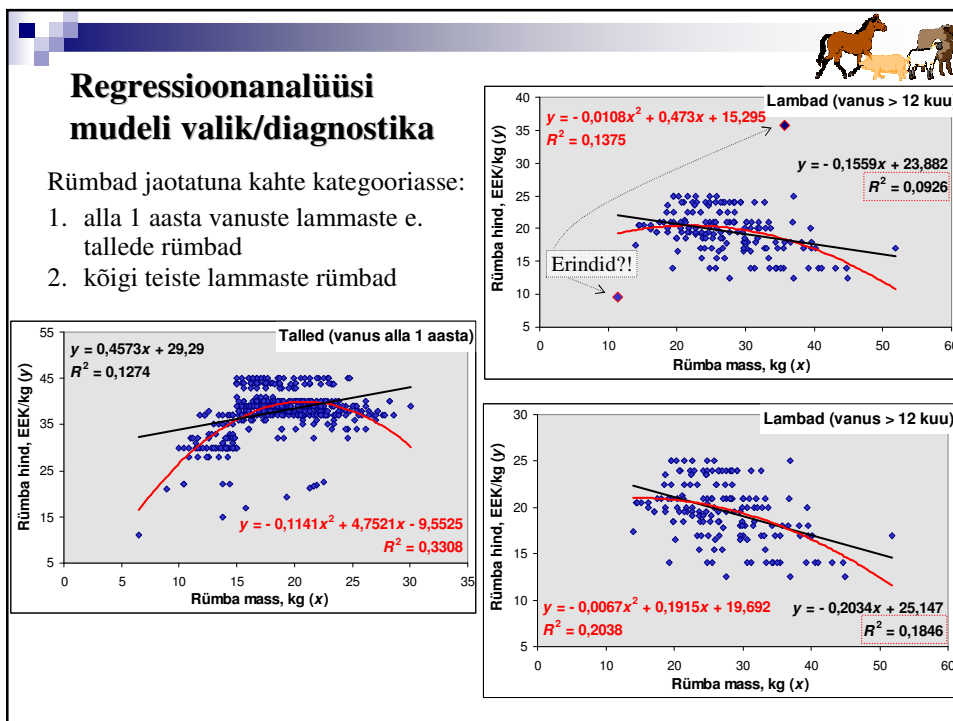
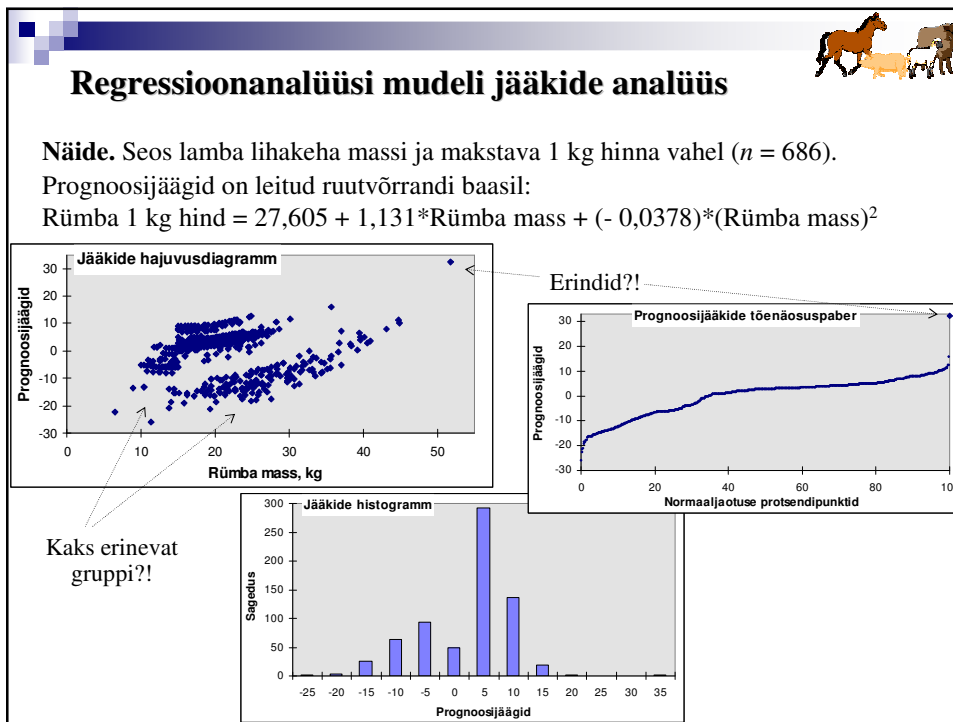
Eelkõige teise eelduse paikapidamatus võib vihjata mittesobivale mudelile (vale matemaatiline funktsioon, mõni arvestamata jäänud argument vmt).


Regressioonanalüüsi mudeli valik/diagnostika



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







Mitmene regressioonanalüüs

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

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

The REG Procedure
Model: MODEL1
Dependent Variable: PIIM_KG PIIM_KG

Analysis of Variance

Otsus: mudel on parem võrreldes konstantse mudeliga (H_1)

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
Dependent Mean 9654.08511
Coeff Var 11.46992

R-Square 0.3928
Adj R-Sq 0.3504

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

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
POEGASTA	POEGASTA	1	-353.33163	291.29395	-1.21	0.2318

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, $kg = a + b_1 * \text{Söödakulud (EEK/1 kg piim)} + b_2 * \text{Lakt_nr}$