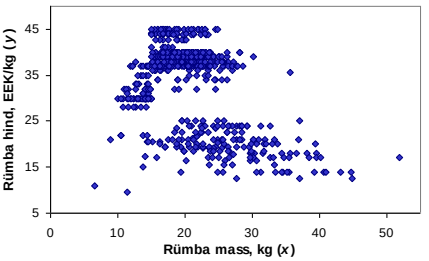
Biomeetria

Kahe arvtunnuse ühine käitumine, regressioonanalüüs

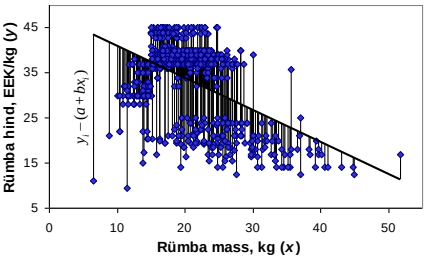
Lineaarne regressioonanalüüs

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



Regressioonivõrrand: $\hat{Y} = a + bX$
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:

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

Lineaarne regressioonanalüüs

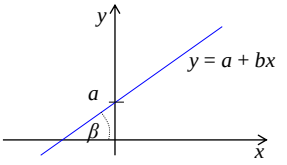
Regressioonivõrrandi (ehk regressioonimudeli) parameetrite a ja b vähimruutude hinnangud:

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

Regressioonivõrrandi kordajate geomeetriline tähendus:

vabaliige a märgib kohta, kus regressioonisirge lõikab y-telge, ning regressioonikordaja b iseloomustab 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*Rümba mass = 48,18 + (–0,711)*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*Rümba mass
= 48,18 + (–0,711)*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 varieeruvusest mudel ära kirjeldab, $0 \le R^2 \le 1$. Mida suurem, seda parem!

Leitakse kui mudelile vastava hajuvuskomponendi $SS_1 = \sum_{i=1}^{n}[(a+bx_i)-\bar{y}]^2$ 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 standardhälbe hinnang, mis näitab keskmist prognoosiviga. 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 = 48,18 + (–0,711)*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			

	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

Regressioonimudeli parameetrite hindamine ei eelda tunnuste jaotumist vastavalt normaaljaotuse seaduspäradele!

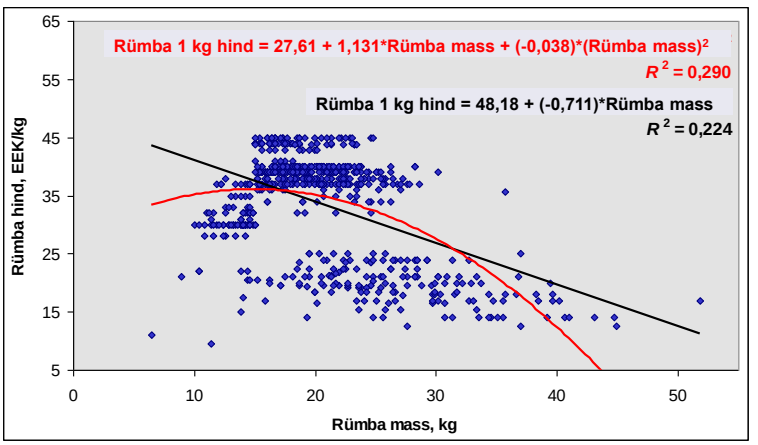
Mudeli täpsuse ja statistilise olulisuse hindamiseks peavad:

- 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).

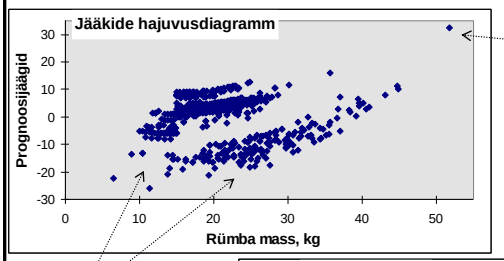
Regressioonanalüüsi mudeli valik/diagnostika

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

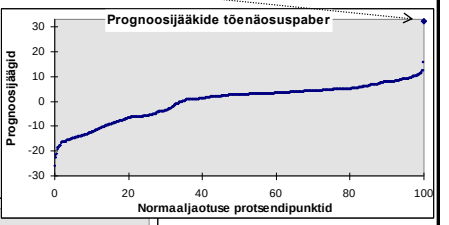


Regressioonanalüüsi mudeli jääkide analüüs

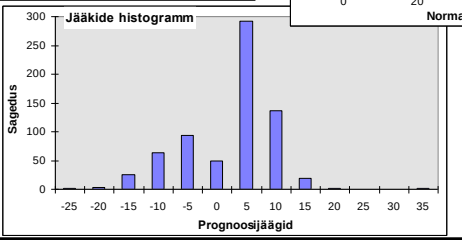
Näide. Seos lamba lihakeha massi ja makstava 1 kg hinna vahel (n = 686).
Prognoosijäägid on leitud ruutvõrrandi baasil:
Rümba 1 kg hind = 27,61 + 1,131*Rümba mass + (-0,038)*(Rümba mass)^2



Erindid?!



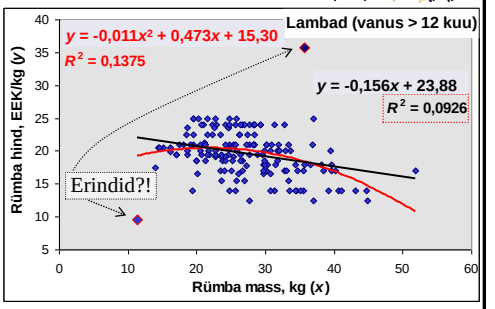
Kaks erinevat gruppi?!

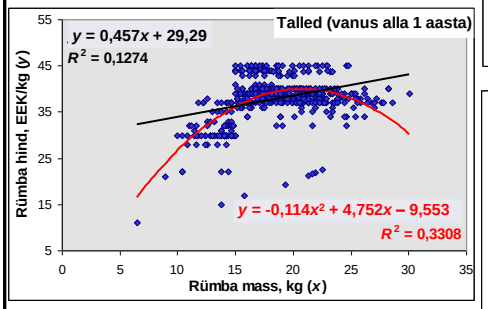


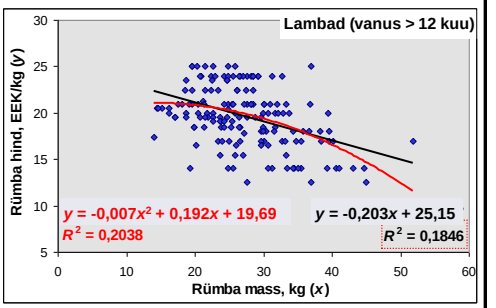
Regressioonanalüüsi mudeli valik/diagnostika

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