

TALLINNA ÜLIKOOL

Regressioonanalüüs

Korrelatsioonanalüüs

Lineaarne regressioon

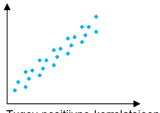
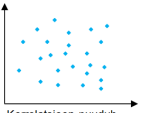
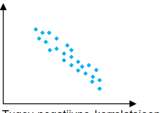
Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Korrelatsioonanalüüs

- Seose tugevus
- Seose suund
- Seose kuju
- Põhjuslikkus
- Statistiline olulisus

Seose visuaalne hindamine

Tugev positiivne korrelatsioon Korrelatsioon puudub Tugev negatiivne korrelatsioon

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Korrelatsioonanalüüs

Seose analüütiline hindamine

-1	-0.9	-0.7	-0.5	-0.3	-0.1	0	0.1	0.3	0.5	0.7	0.9	+1
Väga tugev seos	Tugev seos	Keskmine seos	Nõrk seos	Olematu, väga nõrk seos	Olematu, nõrk seos		Olematu, nõrk seos	Keskmine seos	Tugev seos	Väga tugev seos		
Negatiivsed seosed							Positiivsed seosed					

- Korrelatsioonikordaja valik sõltub
 - andmete tüübist

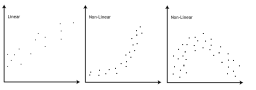
Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Korrelatsioonikordajad

- **Pearson's r**
- Standardiseeritud kahe tunnuse vahelise seose kordaja
- Pearsoni kordaja puudused
 - lineaarne seos: tunneb punktipilve, mis on venitatud piki sirget.
 - tundlik erandite suhtes: paar üksikut erandit väikeses valimis võivad kahekordistada kordaja väärtust.

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




Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Pearsoni r

- Kuidas on seotud vanus ja majapidamistöödeks kuluv aeg?

	vanus	Tunde majapidamistöödeks (tööpäeviti)
vanus	1	,198**
		,000
		873
Tunde majapidamistöödeks (tööpäeviti)	,198**	1
	,000	
	873	873

***. Correlation is significant at the 0.01 level (2-tailed).

vanus	idpt
66	0
69	4
51	2
70	1
66	1
37	1
57	1
46	1
56	2
50	0
66	0
48	4
55	1
49	2
66	1
63	1
63	0

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Korrelatsioonikordaja ruut

	vanus	Tunde majapidamistöödeks (tööpäeviti)
vanus	1	,198**
		,000
		873
Tunde majapidamistöödeks (tööpäeviti)	,198**	1
	,000	
	873	873

***. Correlation is significant at the 0.01 level (2-tailed).

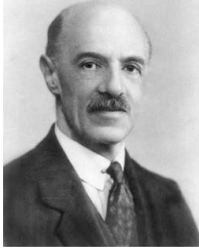
- $R=0,198$
- $R^2=0,039$ e 3,9%
- 3,9% majapidamistöödeks kuluva aja variatiivsusest kirjeldab vanus

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOO

Korrelatsioonikordajad

- Spearman's ρ**
- MPAR korrelatsioonikordaja
- Astakkorrelatsioonikordaja
 - intervalltunnused ei vasta normaaljaotusele (ka erandlikud väärtused)
 - Järjestustunnus(ed)
 - asendab väärtused järjekorranumbritega ning kasutab Pearsoni kordaja valemit => Spearmani kordaja < Pearsoni kordaja (üldjuhul)



Mitmemõtteline statistika Kairi Osula 2013/kevad

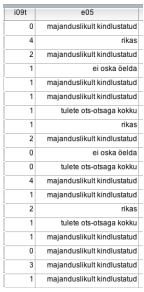
TALLINNA ÜLIKOO

Spearmani roo

- Kuidas on seotud majanduslik kindlustatus ja majapidamistöödeks kuluv aeg?

	sissetulek viimast kuul	Tunde majapidamistöödeks (tööpäeviti)
Spearman's rho	Correlation Coefficient	-.367**
	Sig. (2-tailed)	.000
	N	821
Tunde majapidamistöödeks (tööpäeviti)	Correlation Coefficient	-.367**
	Sig. (2-tailed)	.000
	N	821

** Correlation is significant at the .01 level (2-tailed).

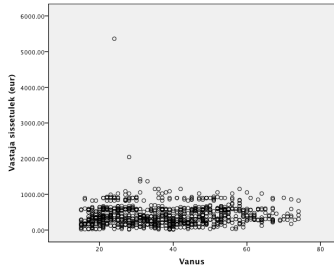


Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOO

Näide (m2008.sav)

- Kuidas on seotud vanus ja sissetulek?



	Vastaja sissetulek (eur)	Vanus
Pearson Correlation	1	.071
Sig. (2-tailed)		.028
N	957	957

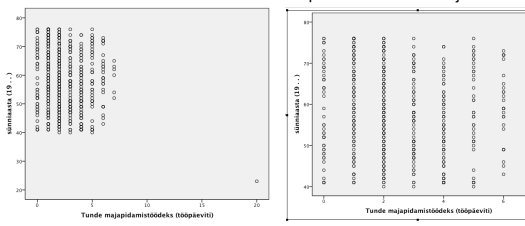
	Vastaja sissetulek (eur)	Vanus
Correlation Coefficient	1.000	.128**
Sig. (2-tailed)		.000
N	957	957

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOO

Pseudoseos

Erandluskord – ekstreemne väärtus hoopis suurendab kordaja väärtust



Pearsoni kordaja = -0,156
Spearmani kordaja = -0,143

Pearsoni kordaja = -0,109
Spearmani kordaja = -0,138

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOO

Korrelatsioonikordajad

- Kendall's τ**
- MPAR korrelatsioonikordaja
- Astakkorrelatsioonikordaja
 - kui on vähe andmeid ja palju sarnaseid väärtuseid



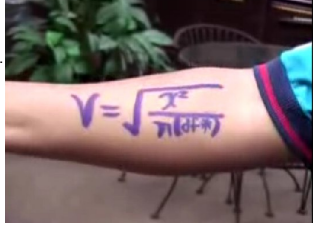
Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOO

Korrelatsioonikordajad

- Cramer's V**
 - Nimitunnuste seose tugevuse uurimiseks.
 - Kordaja ei näita seose suunda, ainult tugevust.

N: Eriala ja tööl käimise seos $V=.45$



- Phi**
 - 2x2 binaarsete tunnuste korral
 - Ruutjuur hii-ruudu väärtuse jagatisest valimi suurusega

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Crameri V

- Kuidas on seotud sugu ja majapidamistöödeks kuluv aeg?

Symmetric Measures			
		Value	Approx. Sig.
Nominal by Nominal	Phi	.380	.000
	Cramer's V	.380	.000
N of Valid Cases		873	

IDR	sugu
0	mees
4	mees
2	naine
1	mees
1	mees
1	mees
1	mees
1	mees
2	mees
0	mees
0	mees
4	mees
1	mees
2	mees
1	mees
1	mees
0	mees
3	mees
1	mees

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Milline kordaja valida?

KORDAJA	ANDMED	LISATINGIMUS
PEARSON	I + I	Seose kuju – lineaarne Erandlikud väärtused puuduvad (ei domineeri)
SPEARMAN	I + I, I + J, J + J	Seose kuju ei ole lineaarne ja jaotusel on erandlikud väärtused (I + I)
KENDALL	J + J	Väike valim ja palju sarnaseid väärtuseid
CRAMER	I + N, I + B, N + B, N + N, N + J, B + J	Tõlgendatakse vaid seose tugevust, mitte suunda
PHI	B + B	Tõlgendatakse vaid seose tugevust, mitte suunda

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Ülesanne

- Sõnasta tabelis olevate tunnuste ning nende korral arvutatud korrelatsioonikordajate põhjal järeldused.

Tunnused	Arvutatud korrelatsioonikordaja	Järeldus
* Vanus	0.72	
* Milliseks Te hindate oma arvutikasutusoskusi? (1-väga head ... 9-väga kehvad, ma ei oska üldse arvutit kasutada).		
* Kui tihti võtad sõna toorimises sinu jaoks olulistel teemadel? (0-mitte kunagi ... 5-alati)	-0.25	
* Koolis käidud aastate arv		
* Mitu korda oled sellel õppeaastal kooli hilinenud?	0.00	
* Mtu päeva oled sellel õppeaastal koolist puudunud?		
* Kui suurt tähelepanu pöörad oma välismusele? (0-üldse mitte ... 5-väga palju)	-0.48	
* Vanus		

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

korrelatsioon >>> regressioon

- Korrelatsioonanalüüsis kirjeldasime **kahe** tunnuse vahelist seost
- Miks me kasutame/vajame lisaks regressioonanalüüsi?
 - Soovime teha ennustusi e **prognoosida** tunnuse väärtuseid
 - Kirjeldada** võimalikult hästi ja usaldusväärselt sõltuvat tunnust sõltumatute tunnuste kaudu

Üksikisiku jaoks me reeglina väga täpseid prognoose ei saa, kuid saame kirjeldada mitme tunnuse vahelist seost (milliselt tunnusest sõltub üks tunnus enam ja millisest vähem).

- KUI meil on **1 tunnus**, siis väärtuste prognoosimisel annab kõige täpsema tulemuse kui prognoosida jaotuse keskvaartust
- Kuidas saaksime mudeli **paremaks/täpsemaks**? Lisame sõltumatuid tunnuseid:
 - Sõltumatud tunnused võiksid olla omavahel mitteseotud, kuid seotud sõltuva tunnusega
 - Mida tugevamalt on sõltumatud tunnused seotud sõltuva tunnusega, seda täpsem saab prognoos olema

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Kahemõtteline regressioonanalüüs

Andmed **rekord2.sav**

- Kuidas on seotud **müügitulemus** (intervall) & **reklaamile kulunud summa** (intervall)
- Hajuvusdiagramm & Pearsoni $r = 0,578$

Müügid plaatide arv (320 000)

Reklaamile kulunud summa (2271,86)

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLICOOL

Kuidas on seotud müük ja reklaam?

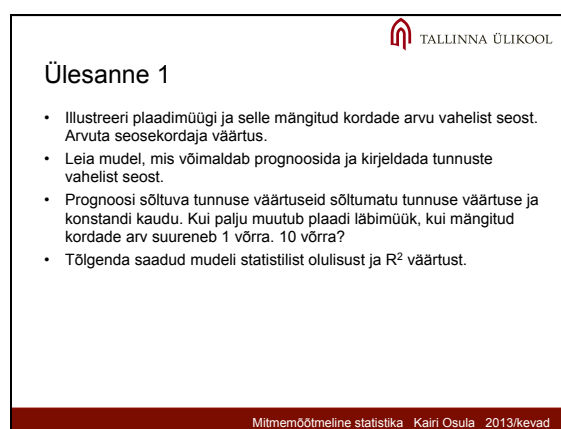
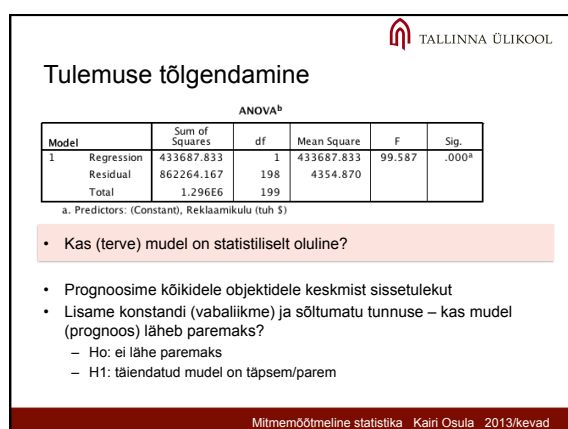
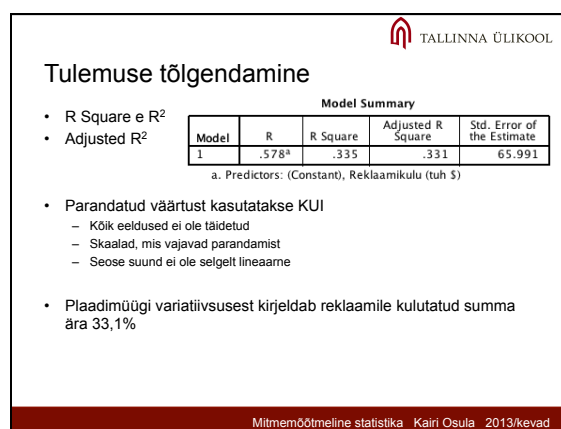
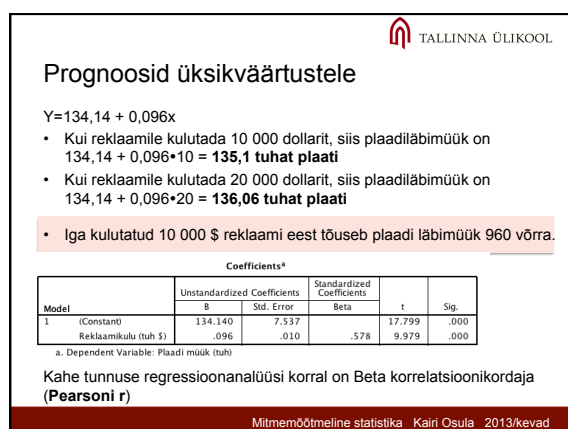
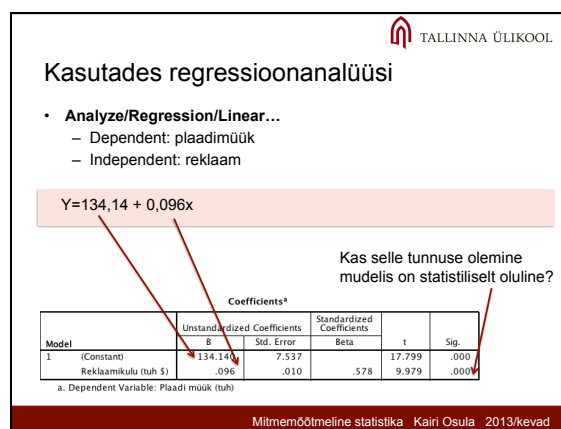
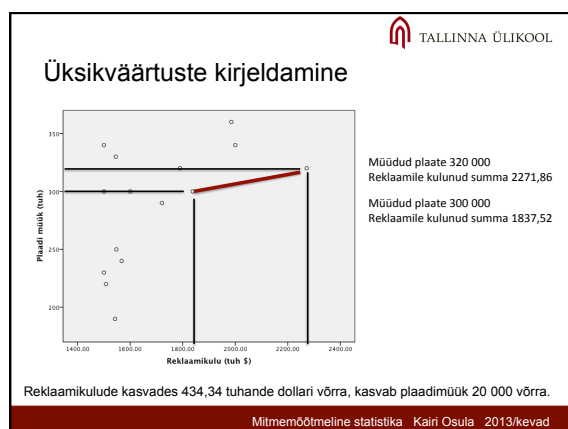
- Pearsoni $r = 0,578$

Seost kirjeldav sirge - asukoht niisugune, et erinevused **joone** ja **vaatlustulemuste** vahel oleksid minimaalsed e prognoosivead tuleksid minimaalsed e kõik punktid oleksid joonele võimalikult lähedal.

$Y = a + bx$
 y sõltuv tunnus
 x sõltumatu tunnus
 a vabaliige

Kui vabaliiget ei oleks, peaks sirge läbima 0-punkti. Sellega ei oleks enam sirge asukoha optimaalsuse nõue täidetud.

Mitmemõtteline statistika Kairi Osula 2013/kevad



Mitmemõõtmeline regressioonanalüüs

- Lisame mudelisse tunnuseid, eesmärgiga suurendada kirjeldusvõimet ja prognoosi täpsust.
- Andmed **Rekord2.sav**

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

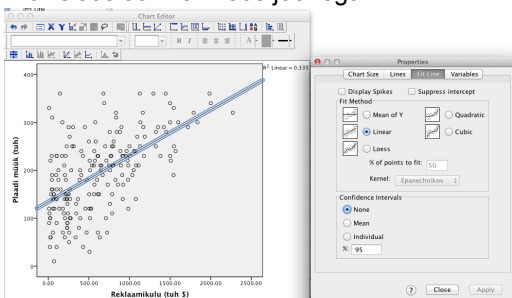
Korrelatsioonikordajad

Correlations						
	Reklaamikulu (tuh \$)	Plaadi müük (tuh)	Mitu korda mängitud kohalikus raadiojaamas viimasel nädalal	Koosseisu atraktiivsus	rik	
Reklaamikulu (tuh \$)	1	.578 ^{**}	.102	.081	-.087	
	Pearson Correlation					
	Sig. (2-tailed)	.000	.151	.256	.219	
	N	200	200	200	200	
Plaadi müük (tuh)	.578 ^{**}	1	.599 ^{**}	.326 ^{**}	-.120	
	Pearson Correlation					
	Sig. (2-tailed)	.000	.000	.000	.090	
	N	200	200	200	200	
Mitu korda mängitud kohalikus raadiojaamas viimasel nädalal	.102	.599 ^{**}	1	.182 ^{**}	-.117	
	Pearson Correlation					
	Sig. (2-tailed)	.151	.000	.010	.098	
	N	200	200	200	200	
Koosseisu atraktiivsus	.081	.326 ^{**}	.182 ^{**}	1	-.028	
	Pearson Correlation					
	Sig. (2-tailed)	.256	.000	.010	.697	
	N	200	200	200	200	
rik	-.087	-.120	-.117	-.028	1	
	Pearson Correlation					
	Sig. (2-tailed)	.219	.090	.098	.697	
	N	200	200	200	200	

**. Correlation is significant at the 0.01 level (2-tailed).

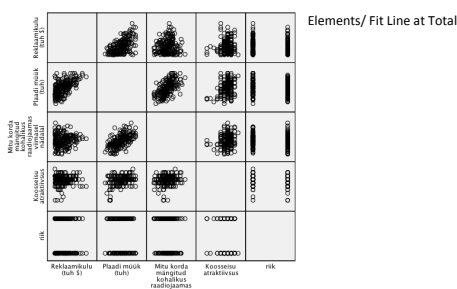
Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Korrelatsiooniväli koos joonega



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

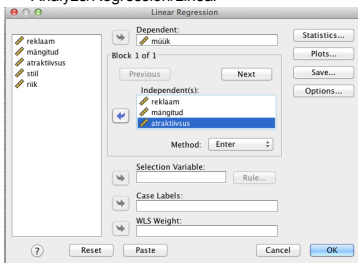
Matrix Scatter



Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Mitmemõõtmeline regressioonanalüüs

- Analyze/Regression/Linear



Enter – kõik tunnused lisatakse mudelisse koos/korrige

Stepwise (ka forward ja backward) – tunnused lisatakse mudelisse ükshaaval.

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Väärtuste prognoosimine

Regression/Linear/Save
Predicted Values☒ Unstandardized ☒ Standardized

Residuals

☒ Unstandardized

	müük	mängitud	atraktiivsus	stiil	rik	PRE_1	RES_1	ZPR_1
1	320	31	5	Pop	USA	352.52131	-32.52131	3.41281
2	340	31	7	Rock	UK	326.38891	13.61109	2.85303
3	360	35	5	Rock	USA	324.95848	35.04152	2.82238
4	300	40	5	Rock	USA	310.77022	-10.77022	2.51846
5	320	30	9	Rock	UK	306.16999	13.83001	2.41992
6	290	32	7	Pop	UK	299.55153	-9.55153	2.27815
7	300	24	9	Pop	USA	287.93912	12.06088	2.02940
8	240	29	6	Rock	USA	284.81968	-44.81968	1.96258
9	250	28	8	Klassika	USA	282.85980	-32.85980	1.92059
10	330	42	7	Pop	UK	282.64256	47.35744	1.91594
11	190	33	8	Pop	USA	282.39552	-92.39552	1.91065
12	220	37	7	Rock	UK	279.09297	-59.09297	1.83991

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Tulemuste tõlgendamine

$R^2=0,66$
 Kogu mudel kirjeldab
 plaadimüügist 66%
 Keskmise prognoosiviga on
 47,1 tuhat plaati

Kas see mudel tervikuna
 aitab statistiliselt määrata
 kirjeldada plaadimüügi
 kogust e kas mudel on
 statistiliselt oluline?

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.813 ^a	.665	.660	47.087

a. Predictors: (Constant), Koostisosa atraktiivsus, Reklaamikulu (tuh \$), Müü korda mängitud kohalikes raadiojaamas viimasei nädal

Model	Sum of Squares	df	Mean Square	F	Sig.
1	Regression Residual Total	881377.418 434574.582 129666	3 196 199	287125.806 2217.217	129.498 .000 ^b

a. Predictors: (Constant), Koostisosa atraktiivsus, Reklaamikulu (tuh \$), Müü korda mängitud kohalikes raadiojaamas viimasei nädal
 b. Dependent Variable: Plaadi müük (tuh)

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	-26.613	17.350		-1.534	.127
	Reklaamikulu (tuh \$)	.085	.007	.511	12.261	.000
	Müü korda mängitud kohalikes raadiojaamas viimasei nädal	3.367	.278	.512	12.123	.000
	Koostisosa atraktiivsus	11.086	2.438	.192	4.548	.000

a. Dependent Variable: Plaadi müük (tuh)

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Tulemuste tõlgendamine

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	-26.613	17.350		-1.534	.127
	Reklaamikulu (tuh \$)	.085	.007	.511	12.261	.000
	Müü korda mängitud kohalikes raadiojaamas viimasei nädal	3.367	.278	.512	12.123	.000
	Koostisosa atraktiivsus	11.086	2.438	.192	4.548	.000

a. Dependent Variable: Plaadi müük (tuh)

Regressioonimudel (standardiseerimata)
 $Y = b_0 + b_1 \text{reklaam} + b_2 \text{mängitud} + b_3 \text{atraktiivsus}$
 Plaadimüük = -26,613 + 0,085 reklaam + 3,367 mängitud + 11,086 atraktiivsus

Reklaam ($b=0,085$) => kui suurendada reklaamikulu ühe ühiku võrra, suureneb plaadimüük 0,085 ühiku võrra.
 Iga reklaamile kulutatud 1000 \$ müüakse lisaks 85 plaati (juhu kui mängitud kordade arv ja atraktiivsus on muutumatud)
 ...

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Tulemuste tõlgendamine

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	-26.613	17.350		-1.534	.127
	Reklaamikulu (tuh \$)	.085	.007	.511	12.261	.000
	Müü korda mängitud kohalikes raadiojaamas viimasei nädal	3.367	.278	.512	12.123	.000
	Koostisosa atraktiivsus	11.086	2.438	.192	4.548	.000

a. Dependent Variable: Plaadi müük (tuh)

Regressioonimudel (standardiseeritud)
 $Z(Y) = 0,511 z(\text{reklaam}) + 0,512 z(\text{mängitud}) + 0,192 z(\text{atraktiivsus})$

Reklaam ($z(b)=0,511$) => kui suurendada reklaamikulu ühe standardhälbe võrra ($SD=485,655$) võrra, suureneb plaadimüük 0,511 standardhälbe võrra.
 Plaadimüügi SD = 80,699
 Iga reklaamile kulutatud 485,655 tuhande \$ korral müüakse lisaks (0,511 x 80,699) 41,24 plaati (juhu kui mängitud kordade arv ja atraktiivsus on muutumatud)

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Väärtuste standardiseerimine

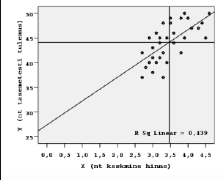
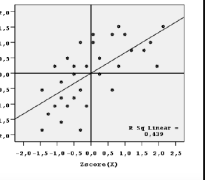
- Vajalik, sest tunnused on mudelisse lisatud erinevate ühikutega
- Jääkliikmed on mudelis oma ühikuga.
- Eesmärk on hinnata, millal jääkliige on liiga suur
- Standardiseeritud jaotuse puhul on see teostatav (norm.jaotus, ± 3)

- Standardiseerimata mudel**
 - Kasutame väärtuste prognoosimisel
- Standardiseeritud mudel**
 - Kasutame seose kirjeldamisel (mitte 1:1 seosekordajad, seoste ühiosa on maha arvestatud)

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Standardiseerimata ja standardiseeritud mudel

K.Niglas
loenguslaidid

Mitmemõtteline statistika Kairi Osula 2013/kevad

TALLINNA ÜLIKOOL

Ülesanne 3

- A study was carried out to explore the relationship between Aggression and several potential predicting factors in 666 children that had an older sibling.
- Variables measured were
 - Parenting_style (high score = bad parenting practices)
 - Computer_Games (high score = more time spent playing computer games)
 - Television (high score = more time spent watching television)
 - Diet (high score = the child has a good diet low in E-numbers)
 - Sibling_Aggression (high score = more aggression seen in their older sibling).
- Past research indicated that parenting style and sibling aggression were good predictors of the level of aggression in the younger child.
- The data are in the file **Child Aggression.sav**
- Analyse them with multiple regression.

Mitmemõtteline statistika Kairi Osula 2013/kevad

Ülesanne 4

- A fashion student was interested in factors that predicted the salaries of catwalk models. She collected data from 231 models.
- For each model she asked them their
 - Salary per day on days when they were working (salary),
 - their age (age),
 - how many years they had worked as a model (years)
 - and then got a panel of experts from modelling agencies to rate the attractiveness of each model as a percentage with 100% being perfectly attractive (beauty).
- The data are in the file **Supermodel.sav**.
- Unfortunately, this fashion student bought some substandard statistics text and so doesn't know how to analyse her data. Can you help her out by conducting a multiple regression to see which variables predict a model's salary? How valid is the regression model?

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Dummy variables (grupp binaarseid tunnuseid)

/libatunnus, fiktiivne tunnus, indikaator-tunnus/

- Problem:** sõltumatu tunnus on mõõdetud nimi- või järjestusskaalal.
- Regressioonianalüüsis **saab kasutada** vaid binaarseid või intervallskaalal mõõdetud tunnuseid.
- Lahendus:** moodustame ühest sõltumatu tunnusest grupi binaarseid tunnuseid (*dummy variables*).
- Mitu binaarset tunnust tekib? Sõltumatu tunnuse väärtuste arv – 1

Perekonnaseis

- abielus
- lahutatud
- vallaline

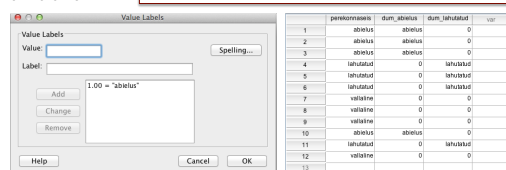
dum_abielus väärtused 1=abielus; 0=lahutatud või vallaline*dum_lahutatud* väärtused 1=lahutatud; 0=abielus või vallaline

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Dummy variables (grupp binaarseid tunnuseid)

Perekonnaseis

- abielus
- lahutatud
- vallaline

dum_abielus väärtused 1=abielus; 0=lahutatud või vallaline*dum_lahutatud* väärtused 1=lahutatud; 0=abielus või vallaline

Transform/Recode into Different Variables

Perekonnaseis -> dum_abielus & 1->1; 2->0; 3->0
 Variable View – Values (1 = abielus)

Mitmemõõtmeline statistika Kairi Osula 2013/kevad

Dummy variables (grupp binaarseid tunnuseid)

Kõrgeim omandatud haridustase

- Lõpetamata algharidus
- Algharidus
- Põhiharidus
- Keskharidus, kutseharidus pärast põhihariduse omandamist
- Keskoolijärgne haridus, kuid mitte kõrgharidus
- Lõpetamata kõrgharidus, bakalaureus
- Viis aastat kõrgharidust, magistriõpe, doktoriõpe

Mitmemõõtmeline statistika Kairi Osula 2013/kevad