Kursus: Mitmemõõtmeline statistika

Seminar IV: Seoste analüüs

Lineaarne regressioon

Õppejõud: Katrin Niglas
PhD, dotsent
informaatika instituut

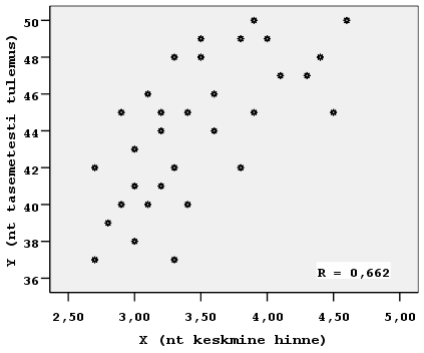
TALLINNA ÜLIKOOL

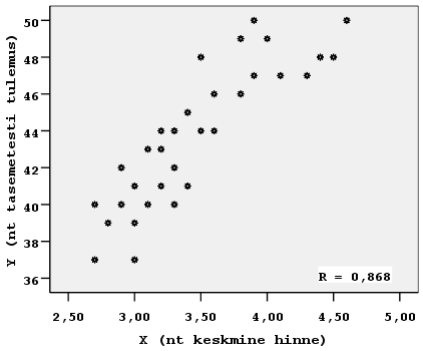
Seoste analüüs

I. Põhjuslikkus
II. Kuju
III. Suund
IV. Tugevus / kirjeldusvõime
***
Statistiline olulisus
***
DV variatiivsuse modelleerimine
– optimaalne kirjeldamine ja prognoosimine
Prognoosi täpsus
Multikollineaarsus / "puhas" ja "lisanduv" kirjeldusvõime

TALLINNA ÜLIKOOL

Korrelatsiooniväli seose kirjeldamiseks





TALLINNA ÜLIKOOL

Millest sõltub analüüsimeetodi valik?

I. Küsimuse tüübist

e mis tüüpi on küsimus, millele tahame analüüsiga vastust saada – nt Kas kolm gruppi **erinevad**? Kas kaks nähtust on **seotud**?

II. Andmete tüübist

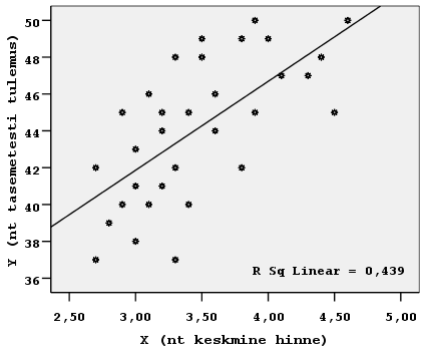
Kas nimi-, järjestus, arv- või binaarsed tunnused

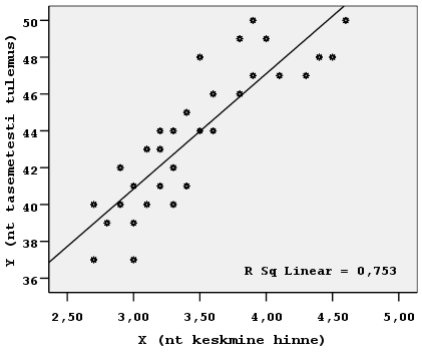
III. Sihtrühmast

Kui suurt teadlikkust statistiliste meetodite osas võib eeldada? Milline esitlusviis on selle rühma puhul köitev ja sobilik?

TALLINNA ÜLIKOOL

Nii sõltuv kui sõltumatu(d) on arvtunnused
Seose kuju – lineaarne seos





TALLINNA ÜLIKOOL

Seosekordajad

Lineaarse korrelatsioonikordaja r väärtuste tõlgendamine:

- r = +1 tähendab maksimaalse tugevusega positiivset seost
- r = 0 tähendab seose puudumist
- r = -1 tähendab maksimaalse tugevusega negatiivset seost

***

- $|r| < 0.30$ olematu, väga nõrk
- $0.30 < |r| < 0.70$ keskmise tugevusega
- $0.70 < |r|$ tugev

Determinatsioonikordaja r^2 – kirjeldusvõime osakaaluna koguvariatiivsusest

PS! Samade põhimõtete järgi saab tõlgendada ka teisi seosekordajaid!

TALLINNA ÜLIKOOL

Seoste analüüs

I. Põhjuslikkus
II. Kuju
III. Suund
IV. Tugevus / kirjeldusvõime
***
Statistiline olulisus
***
DV variatiivsuse modelleerimine
– optimaalne kirjeldamine ja prognoosimine
Prognoosi täpsus
Multikollineaarsus / "puhas" ja "lisanduv" kirjeldusvõime

TALLINNA ÜLIKOOL

Statistilise olulisustesti põhisammud:

E I: Analüüsisin olemasolevaid andmeid kirjeldava statistika meetodite abil ning leidsin midagi „huvitavat" (nt. erinevuse või seose jne)

E II : Tekkis küsimus: „Kas võib üldistada?"
Ü → V → Ü (v eksperimentaalne disain)

I. **Õige olulisustesti valik** (lähtuvalt probleemist ja andmetüübist)

II. **Valitud olulisustesti eelduste kontroll:** **ei** / **jah**

III. **Hüpoteesid:** sisukas hüpotees H_1:
nullhüpotees H_0:
Olulisuse nivoo α
(„Kui väike peab olema H_0 kehtimise tõenäosus, et me võiks ilma suurema riskita ta mittekehtivaks tunnistada?")

IV. **Arvutused**
eesmärgiks hinnata H_0 kehtimise tõenäosust *p*
(„Kui suur on tõenäosus, et olukorras, kus H_0 kehtib, tekkis valmis olnud erinevus v seos juhuse tõttu?")
(*p = Sig = olulisuse tõenäosus*)

V. **Otsus tulemuse kohta**:
$p > \alpha$ H_0 jääb kehtima - statistiliselt mitte oluline (ei üldista)
$p \leq \alpha$ H_1 tõestatud - statistiliselt oluline (võib üldistada)

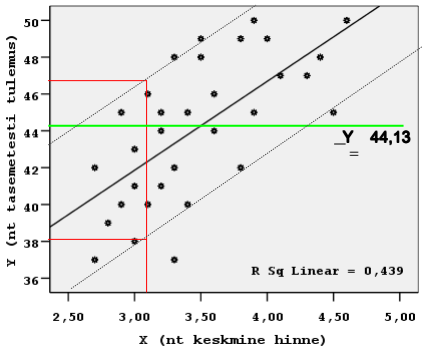
VI. **Järelduse sõnastamine**

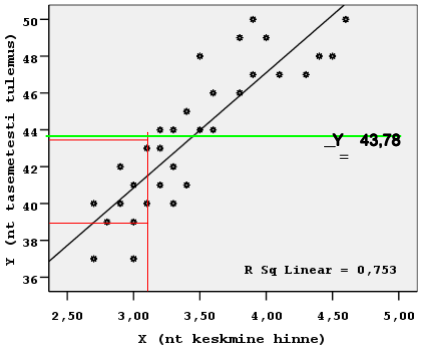
Seoste analüüs

I. Põhjuslikkus
II. Kuju
III. Suund
IV. Tugevus / kirjeldusvõime
***
Statistiline olulisus
***
DV variatiivsuse modelleerimine
– optimaalne kirjeldamine ja prognoosimine
Prognoosi täpsus
Multikollineaarsus / "puhas" ja "lisanduv" kirjeldusvõime

TALLINNA ÜLIKOOL

Mida tugevam seos, seda täpsem prognoos





TALLINNA ÜLIKOOL

Regressioonimudelid

Sirge võrrand: $y = ax + b$

Lineaarne regressioonimudel e -võrrand:

$$Y = b_0 + b_1 X$$

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_m X_m$$

Y	sõltuv tunnus e funktsioontunnus
X_i	sõltumatu tunnus e argumenttunnus e prediktor
b_0	vabaliige (prognoositav väärtus, kui kõik IV'd =0)
b_i	regressioonikordja (DV muut, kui antud IV muutub 1 ühiku võrra ja teised IV'd jäävad samaks)

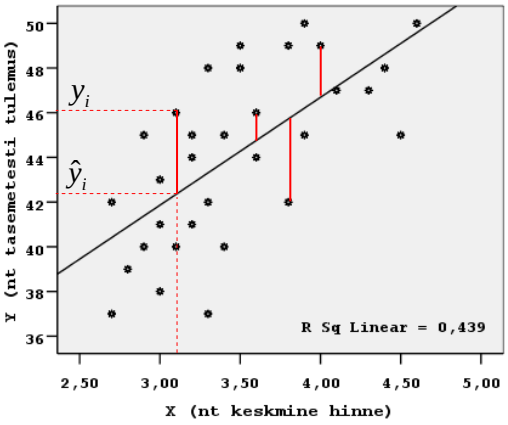
TALLINNA ÜLIKOOL

Üksikobjekti tulemuse prognoosi viga e jääkliige

Eeldus: jääkliikmete vastavus normaaljaotusele ning ühtlane hajuvus (dispersioon)

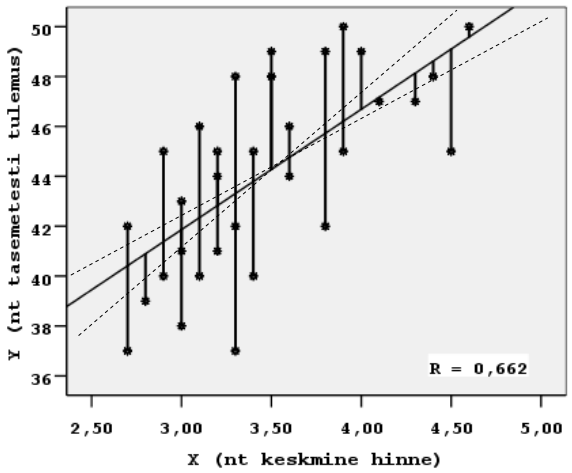
$$y_i - \hat{y}_i$$

NB! Ekstreemsed väärtused mõjutavad seost (standardiseeritud jääkliige > ±3)



TALLINNA ÜLIKOOL

Vähimruutude meetod



TALLINNA ÜLIKOOL

Regressioonikordaja on seotud korrelatsioonikordajaga

$$Y = b_0 + b_1 X \qquad \hat{y}_* = b_0 + b_1 x_*$$

$$r = \frac{1}{N} * \sum_{i=1}^{N} \frac{x_i - \bar{x}}{st.h._x} * \frac{y_i - \bar{y}}{st.h._y} \qquad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 * \sum (y_i - \bar{y})^2}}$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad b_1 = r * \frac{\sqrt{\sum (y_i - \bar{y})^2}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

TALLINNA ÜLIKOOL

Standardiseerimata ja standardiseeritud mudel

$$Y = b_0 + b_1 X \qquad Z(Y) = \beta_1 * Z(X)$$

$$\hat{y}_* = b_0 + b_1 x_* \qquad Z(Y) = \beta_1 * Z(X_1) + \beta_2 * Z(X_2) + \ldots$$

Coefficients[a]

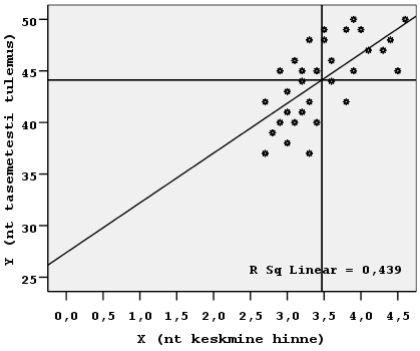
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	27,385	3,497		7,832	,000
	X	4,826	,997	,662	4,841	,000

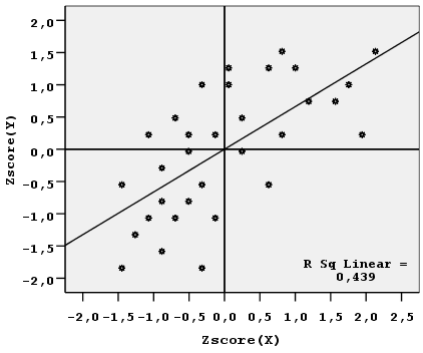
a. Dependent Variable: Y

$$Y = 27{,}385 + 4{,}826 * X \qquad Z(Y) = 0{,}662 * Z(X)$$

TALLINNA ÜLIKOOL

Standardiseerimata ja standardiseeritud mudel





TALLINNA ÜLIKOOL

Regressioonimudeli täpsus

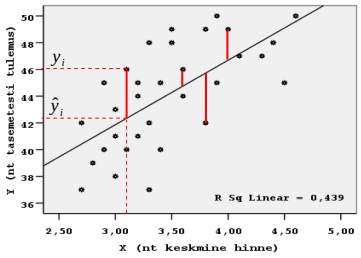
Mudeli kirjeldusvõime: determinatsioonikordaja R^2

Mudeli/prognoosi st.viga: $s = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,662[a]	,439	,420	2,945

a. Predictors: (Constant), X



TALLINNA ÜLIKOOL

Kirjeldusvõime mitmeses regressioonimudelis

DV variatiivsuse modelleerimine
– optimaalne kirjeldamine ja prognoosimine

Prognoosi täpsus

$$Y = b_0 + b_1X_1 + b_2X_2 + ... + b_mX_m$$

NB! Multikollineaarsus / "lisanduv" ja "puhas" kirjeldusvõime

Osakorrelatsioon ja pseudo-osakorrelatsioon =>
sisuliselt on tegu ühise kirjeldusvõime osa jagamisega

TALLINNA ÜLIKOOL

Regressioonimudeli üldistamine (stat. olulisus)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_m X_m$$

NB! Kordajad on täpsed antud konkreetse valimi jaoks!

Olulisustestid kordajate ning kogu mudeli jaoks:

H_0: $B_i=0$

H_0: Mudel e sõltumatute muutujate lineaarkombinatsioon ei aita kirjeldada sõltuva muutuja variatiivsust üldkogumis

Üldkogumi kordajaid saab esitada hinnanguliselt kasutades usaldusintervalle!

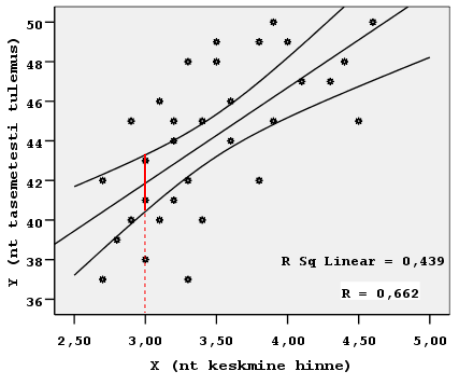
$$95\%\_B_i \in b_i \pm 2{,}5 * st.viga$$

TALLINNA ÜLIKOOL

Regressioonimudeli üldistamine (stat. olulisus)

Standardviga mudeli/prognoosi usaldusintervalli arvutamiseks:

$$st.v._* = s\sqrt{\frac{1}{N} + \frac{(x_* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$



TALLINNA ÜLIKOOL

Seoste analüüs

I. Põhjuslikkus
II. Kuju
III. Suund
IV. Tugevus / kirjeldusvõime
***
Statistiline olulisus

***

DV variatiivsuse modelleerimine
– optimaalne kirjeldamine ja prognoosimine

Prognoosi täpsus

Multikollineaarsus / "puhas" või "lisanduv" kirjeldusvõime

TALLINNA ÜLIKOOL