Regressioonanalüüs

Regressioonanalüüs võimaldab luua matemaatilise mudeli kirjeldamaks tunnuste vahelisi seoseid. Regressioonanalüüsi puhul vaatleme üht tunnust kui sõltuvat[1] ning püüame leida tunnuseid, mille põhjal oleks võimalik kirjeldada ning ühtlasi ka prognoosida selle sõltuva tunnuse väärtusi. Kirjeldav mudel ning prognoos on seda täpsem, mida tugevamini sõltumatu(d) tunnus(ed) sõltuva tunnusega seotud on.

Regressioonanalüüsil on mitmeid erinevaid metoodilisi variatsioone sõltuvalt kasutada olevate tunnuste tüübist ning jaotuse parameetritest.[2] Vaatleme edaspidises vaid kõige klassikalisemat regressioonanalüüsi meetodit – **lineaarset regressiooni**.

Eeldused:

- arvtunnused (arvskaalaga samaväärseks võib pidada võrdsete vahemikega järjestusskaalat) ja/või binaarsed e kahe väärtusega tunnused;[3]
- lineaarne seose kuju sõltuva ja sõltumatute tunnuste vahel;
- sõltumatud tunnused võimalikult tugevalt seotud sõltuva tunnusega kuid samas võimalikult vähe seotud omavahel.

Eesmärk:

Koostada lineaarne regressioonivõrrand üldkujuga:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{K} + b_n x_n$$

kus y on sõltuv tunnus ning $x_1 \mathrm{K}\ x_n$ on sõltumatud tunnused.

Regressioonanalüüsi idee:

Sobivate kordajate leidmiseks kasutatakse vähimruutude meetodit. **Vähimruutude meetodi idee** seisneb selles, et seost iseloomustavat punktiparve valitakse esindama selline sirge, millest kõikide üksikpunktide kauguste ruutude summa on minimaalne.

Kuna regressioonimudel põhineb tunnuste vahelistel seostel, siis on ühe sõltumatu muutujaga mudeli puhul kordaja b_1 seotud sõltuva ja sõltumatu tunnuse vahelise korrelatsioonikordajaga r ; mitme sõltumatu muutujaga mudeli puhul on kordajad b_1, b_2, ... seotud vastavate osakorrelatsioonikordajatega.

Standardiseerimata mudelis on kordajatel aga täita ka skaalasid ühtlustav funktsioon ning seetõttu pole nad omavahel võrreldavad. Standardiseeritud mudelis, kus kasutatakse algtunnuste standardiseeritud väärtusi on aga sõltumatute tunnuste kordajad võrreldavad ning suurem kordaja väljendab tugevamat seost sõltumatu ja sõltuva tunnuse vahel.

Regressioonimudeli põhjal saadud prognoos on üksikobjekti jaoks pea alati (pisut) ebatäpne sest mudel prognoosib keskmist taset. Mudeli "headust" hinnataksegi selle põhjal kui tugev on seos sõltumatute tunnuste komplekti ja sõltuva tunnuse vahel[4], sest see määrab prognoosi täpsuse. Mudeli "headust" väljendab ka **mudeli standardviga**, mis kirjeldab sõltuva tunnuse väärtuste keskmist kõrvalekallet ehk erinevust prognoosist.

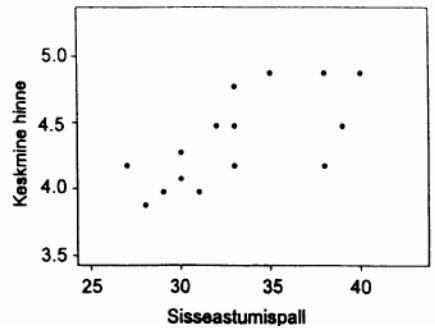
---

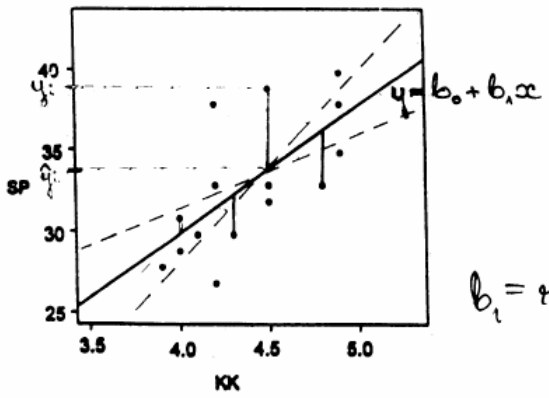
[1] Keerulisema mudeli korral võib sõltuvaid tunnuseid olla ka mitu.
[2] Tuntuim mitteparameetriline regressioonanalüüsi meetod on logistiline regressioon.
[3] Kui soovitakse kasutada võrrandis nominaalseid tunnuseid, siis tuleb need ümber kodeerida grupiks binaarseteks tunnusteks (ingl *dummy variables*).
[4] seda väljendab **mitmene korrelatsioonikordaja**

SP	KK	SP	KK
35	4.9	31	4.0
32	4.5	32	4.5
38	4.2	33	4.5
30	4.3	38	4.9
33	4.8	33	4.2
27	4.2	29	4.0
30	4.1	40	4.9
28	3.9	39	4.5

$n = 16$

$\bar{y} = 0,67$





$\hat{y} = b_0 + b_1 x$

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

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}}$$

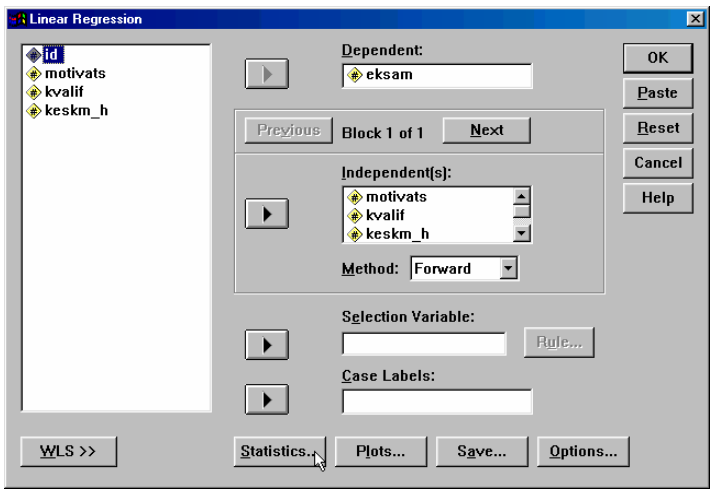
$$b_0 = \bar{y} - b_1 \bar{x}$$

$b_1 \approx 7,87$

$b_0 \approx -1,64$

Regressioonanalüüsi tellimine SPSS'is:

- Vali *Analyze/Regression/Linear...*



- Paiguta sõltuv tunnus väljale *Dependent:*
- Paiguta sõltumatu(d) tunnus(ed) väljale *Independent(s):*
- Vali sõltumatute tunnuste mudelisse võtmise meetod väljal *Method:*

  Enter kõik valitud tunnused pannakse mudelisse

  Forward mudelissse lisatakse sammhaaval need tunnused, mis mõjutavad sõltuvat tunnust statistiliselt olulisel määral

- Kui soovid regressioonimudeli konstrueerimisel lähtuda mitte kõigist objektidest vaid ainult ühest alamosast, siis paiguta selekteerimise aluseks olev tunnus väljale *Selection Variable:* ning vajutades nuppu *Value...* määra, millise väärtusega objekte arvestatakse.
- Kui mingi tunnus sisaldab objektide identifikaatoreid e nimetusi, mida soovid kasutada, siis paiguta see väljale *Label Cases by:*
- Vajutades nuppu *Statistics...* saad tellida mitmeid algtunnuseid ja regressioonimudelit kirjeldavaid arvnäitajaid. Tähtsamad:

  Regr Coef / Estimates regressioonivõrrandi kordajad

  Regr Coef / Confidence intervals kordajate usaldusintervallid

  Model fit mitmene korrelatsioonikordaja R, R^2, parandatud R^2, mudeli standardviga

  Descriptives algtunnuste kirjeldavad arvnäitajad

  Part and partial correlations korrelatsioonimaatriks

  Residuals / Casewise diagnostics jääkliikmete analüüs; tuuakse välja objektid, mille puhul tegelik väärtus erineb prognoosist palju.

- Nupu *Plots...* abil saad tellida mitmeid jääkliikmete jaotust iseloomustavaid diagramme.

- Nupu *Save...* abil saad tellida mitmeid üksikobjekte kirjeldavaid statistikuid, mis salvestatakse andmestikku uute tunnustena. Tähtsamad:

Predicted Values	
Unstandardized	prognoositavad väärtused (standardiseerimata)
Standardized	prognoositavad väärtused (standardiseeritud)
Residuals	
Unstandardized	jääkliikmed (standardiseerimata)
Standardized	jääkliikmed (standardiseeritud)
Prediction Intervals	
Mean	prognoosi usaldusintervall
Individual	üksikväärtuste usaldusintervall
Confidence Interval	usaldusnivoo eelmiste usaldusintervallide jaoks

- Nupu *Options...* abil saad
  - määrata tingimused tunnuste mudelisse liitmiseks[5];
  - otsustada, kas regressioonivõrrandis on konstant e vabaliige;
  - määrata puuduvate väärtuse käsitlemise viisi:

Exclude cases listwise	objektid, kus esineb puuduvaid väärtusi, jäetakse kogu analüüsist välja
Exclude cases pairwise	objektid, kus esineb puuduvaid väärtusi, jäetakse välja puuduvate väärtustega tunnuste vaheliste seoste arvutamisel
Replace with mean	puuduvad väärtused asendatakse tunnuse keskväärtusega.

Regressioonanalüüsi põhitulem ja selle tõlgendamine:

EELANALÜÜS:
kirjedavad arvnäitajad kõigi tunnuste kohta
ja
korrelatsioonimaatriks

Descriptive Statistics

	Mean	Std. Deviation	N
koondeksami tulemus	204.22	135.79	465
motivatsioonitesti tulemus	7.90	10.95	465
teadmiste tase sisseastumisel	4.97	2.39	465
keskmine hinne ülikoolis	6.12	4.19	465

Correlations

		koondeksami tulemus	motivatsiooni testi tulemus	teadmiste tase sisseastumisel	keskmine hinne ülikoolis
Pearson Correlation	koondeksami tulemus	1.000	.287	.306	.370
	motivatsioonitesti tulemus	.287	1.000	.440	.256
	teadmiste tase sisseastumisel	.306	.440	1.000	.505
	keskmine hinne ülikoolis	.370	.256	.505	1.000
Sig. (1-tailed)	koondeksami tulemus	.	.000	.000	.000
	motivatsioonitesti tulemus	.000	.	.000	.000
	teadmiste tase sisseastumisel	.000	.000	.	.000
	keskmine hinne ülikoolis	.000	.000	.000	.
N	koondeksami tulemus	465	465	465	465
	motivatsioonitesti tulemus	465	465	465	465
	teadmiste tase sisseastumisel	465	465	465	465
	keskmine hinne ülikoolis	465	465	465	465

[5] Seda siis, kui pole kasutatud meetodit *Enter*, mille korral pannakse kõik tunnused mudelisse.

Regression [6]

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	keskmine hinne ülikoolis	.	Forward (Criterion: Probability-of-F-to-enter <= .050)
2	motivatsioonitesti tulemus	.	Forward (Criterion: Probability-of-F-to-enter <= .050)

a. Dependent Variable: koondeksami tulemus

Kirjeldatakse sõltumatud tunnused, mis regressioonimudelisse pannakse ning tunnuste mudelisse lisamise meetod.

NB! Meetodi *Enter* puhul antakse ainult üks mudel, ülejäänud meetodite puhul toimub tunnuste lisamine/eemaldamine mudelisse/st ükshaaval ning seetõttu kirjeldatakse nii mitut mudelit kui mitu lisamist/eemaldamist tehakse. Lõplik mudel on viimane.

Antud näites lisati mudelisse kaks sõltumatut tunnust; seega esimeses mudelis on üks sõltumatu tunnus – keskmine hinne koolis ja teises mudelis kaks sõltumatut tunnust – keskmine hinne ja motivatsioonitesti tulemus. Kolmas tunnus – teadmiste tase sisseastumisel – jäi mudelist välja.

Regressioonimudeli kordajad (b_0, b_1, ...) | Kordajate standardvead | Standardiseeritud kordajad | t-väärtused (H_0: B_i=0) | Kordajate olulisustõenäosus p

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	130.904	10.373		12.620	.000
	keskmine hinne ülikoolis	11.974	1.398	.370	8.564	.000
2	(Constant)	121.178	10.352		11.706	.000
	keskmine hinne ülikoolis	10.273	1.414	.317	7.264	.000
	motivatsioonitesti tulemus	2.549	.542	.206	4.705	.000

a. Dependent Variable: koondeksami tulemus

Tabelist saadud kordajate abil saame välja kirjutada regressioonimudeli kujul:

$$y = 121{,}178 + 10{,}273 * x_1 + 2{,}549 * x_2$$

ehk

$$eksam = 121{,}2 + 10{,}3 * kesk\ \min e\_h + 2{,}5 * motivats$$

Saadud mudeli abil saame prognoosida eksami tulemusi kui meil on teada üliõpilase keskmine hinne ja motivatsioonitesti tulemus. Näiteks olgu Juku keskmine_h=5 ja motivats=10, siis eksamitulemuse prognoos oleks:

$$eksam = 121{,}2 + 10{,}3 * 5 + 2{,}5 * 10 = 197{,}7\ palli$$

Regressioonimudeli saab esitada ka standardiseeritud kujul - sel juhul kasutatakse mudelis algtunnuste standardiseeritud väärtusi. Standardiseeritud mudelis on seega kõik tunnused viidud ühisele universaalsele ühikuta skaalale ning seetõttu on sõltumatute tunnuste kordajad omavahel võrreldavad – mida suurem kordaja, seda suurem seos on sõltumatul tunnusel sõltuva tunnusega. Näeme, et keskmine hinne on eksamitulemusega mudelis enam seotud kui motivatsioonitesti tulemus.

$$Z(eksam) = 0{,}317 * Z(kesk\ \min e\_h) + 0{,}206 * Z(motivats)$$

[6] tabelite järjekorda on muudetud nii et nad oleksid tõlgendamise jaoks loogilises järjekorras.

Regressioonivõrrandi kordajad on leitud valimi tulemuste põhjal ning seetõttu on mudeli laiema kasutamise puhul vajalik kontrollida kordajate statistilist olulisust. Nullhüpotees väidab, et üldkogumis vastava sõltumatu tunnuse kordaja võrdub nulliga[7], mis tähendab, et tunnusel mõju sõltuvale tunnusele puudub.

Kui meetodiks on valitud *Enter*, siis pannakse mudelisse kõik sõltumatud tunnused vaatamata sellele, kas nende kordajad on statistiliselt olulised või mitte. Kõigi teiste meetodite puhul jäävad mudelisse ainult need tunnused, mille kordaja on statistiliselt oluline. Mudelist välja jäänud tunnuseid kirjeldab järgmine tabel.

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance
1	motivatsioonitesti tulemus	.206[a]	4.705	.000	.214	.935
	teadmiste tase sisseastumisel	.159[a]	3.220	.001	.148	.745
2	teadmiste tase sisseastumisel	.086[b]	1.629	.104	.076	.642

a. Predictors in the Model: (Constant), keskmine hinne ülikoolis

b. Predictors in the Model: (Constant), keskmine hinne ülikoolis, motivatsioonitesti tulemus

c. Dependent Variable: koondeksami tulemus

Näeme, et peale kahe sõltumatu tunnuse lisamist mudelisse jääb kolmanda tunnuse (teadmiste tase sisseastumisel) osakorrelatsioon sõltuva tunnusega väga madalaks (Partial Correlation 0,076 ja standardiseeritud kordaja 0,086) ning kordaja ei osutu ka statistiliselt oluliseks (p=0,104). Seetõttu seda tunnust mudelisse ei lisata.

Järgmises tabelis on toodud mudeli parameetrid, mis iseloomustavad mudeli "headust" ning prognoosi täpsust.

Mitmene korrelatsioonikordaja | Mitmese korrelatsioonikordaja ruut e determinatsioonikordaja | Parandatud determinatsioonik. | Mudeli standardviga

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.370[a]	.137	.135	126.30
2	.420[b]	.176	.173	123.52

a. Predictors: (Constant), keskmine hinne ülikoolis

b. Predictors: (Constant), keskmine hinne ülikoolis, motivatsioonitesti tulemus

Lõpliku mudeli puhul on seos sõltuva tunnuse ja sõltumatute tunnuste vahel (nõrgapoolse) keskmise tugevusega (R=0,420). Determinatsioonikordaja põhjal saab väita, et sõltumatud tunnused kirjeldavad koos pisut üle 17% sõltuva tunnuse variatiivsusest, ehk keskmise hinde ja motivatsioonitesti tulemusega on võimalik kirjeldada ligikaudu 17% koondeksami tulemuste variatiivsusest.

Mudeli standardvea põhjal saame väita, et keskmiselt erineb tegelik eksamitulemus mudeli põhjal saadud prognoosist 123,5 punkti võrra.

Dispersioonanalüüs viiakse läbi kontrollimaks kogu mudeli statistilist olulisust – siin on küsimuseks, kas sõltumatutel muutujatel üheskoos on statistiliselt oluline mõju sõltuvale tunnusele (kui me saame rääkida põhjuslikust seosest) või kas sõltumatud muutujad kirjeldavad osa sõltuva tunnuse väärtuste variatiivsusest parandades seega meie prognoosi.

Tavaliselt osutub mudel statistiliselt oluliseks kui ta sisaldab kasvõi ühte sõltumatut tunnust, mille kordaja on statistiliselt oluline. Nii ka meie näites.

[7] H_0: $B_i=0$ ja H_1: $B_i\neq0$

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1169994.438	1	1169994.438	73.342	.000[a]
	Residual	7386008.624	463	15952.502		
	Total	8556003.062	464			
2	Regression	1507751.623	2	753875.812	49.415	.000[b]
	Residual	7048251.439	462	15255.955		
	Total	8556003.062	464			

a. Predictors: (Constant), keskmine hinne ülikoolis

b. Predictors: (Constant), keskmine hinne ülikoolis, motivatsioonitesti tulemus

c. Dependent Variable: koondeksami tulemus

Vahemikhinnangud regressioonanalüüsi puhul

Statistilise olulisuse kõrval võib arvutada ka prognoosi vahemikhinnangu ehk usaldusintervalli. Regressioonanalüüsi korral saame rääkida kahest erinevast usaldusintervallist:

- **prognoosi usaldusintervallist**, mis annab vahemiku, kuhu valitud tõenäosusega jääb üldkogumi tegelik keskmine tase (regressioonivõrrand prognoosib valimi keskmist taset)
- **üksikväärtuste usaldusintervalli**, mis annab vahemiku, kuhu jääb enamus (valitud osa n 95% või 99%) üldkogumi sõltuva tunnuse väärtustest.

Mõlema usaldusintervalli puhul sõltub standardviga lisaks valimi suurusele ka konkreetsetest sõltumatute tunnuste väärtustest ning seetõttu pole usaldusintervallid kogu tunnuste muutumispiirkonnas ühesuguse ulatusega. Viga on seda väiksem, mida keskmisele lähemal asuvad sõltumatute tunnuste väärtused.

Eelmisest tuleneb ka see, et pole võimalik arvutada välja üht standardviga ega usaldusintervalli. SPSS võimaldab arvutada usaldusintervalli otspunktid iga objekti poolt määratud punktis ning salvestada need andmestikku uute tunnustena (vt tellimise kohta eespool).