REGRESSIOON-ANALÜÜS II

LOENGU TEEMAD

1. Mitmene regressioon
2. Parameetrite tõlgendamine
3. Korrigeeritud determinatsioonikordaja
4. Mudeli ANOVA tabel ja *F*-test
5. Parameetrite statistilise olulisuse kontrollimine
6. Tunnuste valik
   - Edaspidine
   - Tagurpidine
7. Multikollineaarsus
8. Kvalitatiivsed seletavad tunnused
9. Lineariseerimine

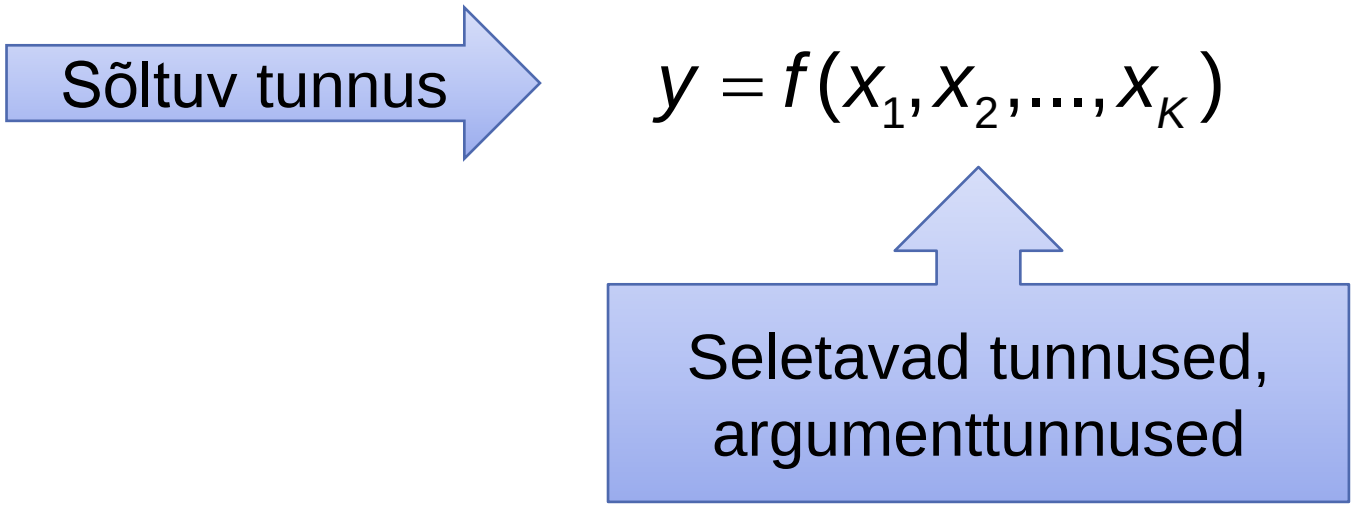
MITU SELETAVAT TUNNUST

Sõltuvat tunnust mõjutab enamasti rohkem kui üks seletav tunnus.

Käivet võib mõjutada
- toote hind;
- kulud reklaamile;
- keskmine palk;
- töötuse määr.

Sündivust võivad mõjutada
- sünnitamisealiste naiste arv;
- elatustase (SKP elaniku kohta);
- töötuse määr;
- naiste osalemise määr tööjõuturul;
- riiklikud ja omavalitsuste poolt eraldatavad toetused (lastetoetus, sünnitoetus, emapalk).

MITMENE REGRESSIOON



Lineaarne mudel, K argumenttunnust

$$y = b + a_1 x_1 + a_2 x_2 + ... + a_K x_K + \varepsilon$$

Vaja hinnata K+1 parameetrit: b, a_1, …, a_K

LINEAARSE MUDELI PARAMEETRITE TÕLGENDUS

$$y = b + a_1 x_1 + a_2 x_2 + \ldots + a_K x_K + \varepsilon$$

b näitab, millega võrdub y, kui kõik argumenttunnused on nullid;

a_1 näitab, kui palju muutub y, kui x_1 suureneb ühiku võrra ja teised argumenttunnused jäävad samaks;

a_2 näitab, kui palju muutub y, kui x_2 suureneb ühiku võrra ja teised argumenttunnused jäävad samaks;

jne

ceteris paribus kõik muu jääb samaks

NÄIDE: RAVIMIPOE KÄIBE MUDEL

Millest sõltub ravimipoe käive? USA ravimimüügi keti Walgreens 27 poe andmetele tuginedes saadi lineaarne mudel:

$$\hat{y} = -18,9 + 16,2x_1 + 0,175x_2 + 11,5x_3 + 13,6x_4 - 5,31x_5$$

$$R^2 = 0,993$$

y	netokäive aastas, tuh $;
x_1	poe pindala, tuh ruutjalga;
x_2	varude maksumus, tuh $;
x_3	reklaamikulud aastas, tuh $;
x_4	piirkonnas elavate perede arv, tuh;
x_5	piirkonnas tegutsevate konkurentide arv.

Parameetrite tõlgendus?

NÄIDE: RAVIMIPOE KÄIBE MUDEL

Millest sõltub ravimipoe käive? USA ravimimüügi keti Walgreens 27 poe andmetele tuginedes saadi lineaarne mudel:

$$\hat{y} = -18,9 + 16,2x_1 + 0,175x_2 + 11,5x_3 + 13,6x_4 - 5,31x_5$$

$$R^2 = 0,993$$

y	netokäive aastas, tuh $;
x_1	poe pindala, tuh ruutjalga;
x_2	varude maksumus, tuh $;
x_3	reklaamikulud aastas, tuh $;
x_4	piirkonnas elavate perede arv, tuh;
x_5	piirkonnas tegutsevate konkurentide arv.

Poel, mille pindala on tuhande ruutjala võrra suurem, on 16,2 tuh $ võrra suurem netokäive aastas.

NÄIDE: RAVIMIPOE KÄIBE MUDEL

Millest sõltub ravimipoe käive? USA ravimimüügi keti Walgreens 27 poe andmetele tuginedes saadi lineaarne mudel:

$$\hat{y} = -18,9 + 16,2x_1 + 0,175x_2 + 11,5x_3 + 13,6x_4 - 5,31x_5$$

$$R^2 = 0,993$$

y	netokäive aastas, tuh $;
x_1	poe pindala, tuh ruutjalga;
x_2	varude maksumus, tuh $;
x_3	reklaamikulud aastas, tuh $;
x_4	piirkonnas elavate perede arv, tuh;
x_5	piirkonnas tegutsevate konkurentide arv.

Poel, kus varude maksumus on tuh $ võrra suurem, on 0,175 tuh $ võrra suurem netokäive aastas.

NÄIDE: RAVIMIPOE KÄIBE MUDEL

Millest sõltub ravimipoe käive? USA ravimimüügi keti Walgreens 27 poe andmetele tuginedes saadi lineaarne mudel:

$$\hat{y} = -18,9 + 16,2x_1 + 0,175x_2 + 11,5x_3 + 13,6x_4 - 5,31x_5$$

$$R^2 = 0,993$$

y	netokäive aastas, tuh $;
x_1	poe pindala, tuh ruutjalga;
x_2	varude maksumus, tuh $;
x_3	reklaamikulud aastas, tuh $;
x_4	piirkonnas elavate perede arv, tuh;
x_5	piirkonnas tegutsevate konkurentide arv.

Poel, kus reklaamikulud on tuh $ võrra suuremad, on 11,5 tuh $ võrra suurem netokäive aastas.

NÄIDE: RAVIMIPOE KÄIBE MUDEL

Millest sõltub ravimipoe käive? USA ravimimüügi keti Walgreens 27 poe andmetele tuginedes saadi lineaarne mudel:

$$\hat{y} = -18,9 + 16,2x_1 + 0,175x_2 + 11,5x_3 + 13,6x_4 - 5,31x_5$$

$$R^2 = 0,993$$

y	netokäive aastas, tuh $;
x_1	poe pindala, tuh ruutjalga;
x_2	varude maksumus, tuh $;
x_3	reklaamikulud aastas, tuh $;
x_4	piirkonnas elavate perede arv, tuh;
x_5	piirkonnas tegutsevate konkurentide arv.

Poel, millel konkurentide arv on ühe võrra suurem, on netokäive aastas 5,31 tuh $ võrra väiksem.

NÄIDE: TÖÖJÕU PAKKUMINE

Uuring USA-s 1966.a, valimis 6000 leibkonda.
Millest sõltub töötatud tundide arv aastas?

Ühe seletava tunnusega mudel. TTASU on tunnitasu, $

$$\text{TUNNID} = 1913 + 80{,}9 \cdot \text{TTASU}, \quad R^2 = 0{,}333$$

Determinatsioonikordajaid võrrelda ei tohi!

Kolme seletava tunnusega mudel. Lisatud VARAD (varade suurus) ja VANUS

$$\text{TUNNID} = 2444{,}8 - 47{,}6 \cdot \text{TTASU} + 0{,}02641 \cdot \text{VARAD} - 8{,}66 \cdot \text{VANUS}, \quad R^2 = 0{,}715$$

Kumb mudel on parem?

Demo: R^2 suurenemine

DETERMINATSIOONIKORDAJA R^2 PUUDUS

1 argumenttunnus $y = b + ax_1 + \varepsilon, \qquad R_1^2$

Lisame mudelisse teise tunnuse $y = b + a_1x_1 + a_2x_2 + \varepsilon, \qquad R_2^2$

Tunnuste lisamisel alati $R_2^2 > R_1^2$

Ka siis, kui x_2 tegelikult tunnusega y seotud ei ole.

Determinatsioonikordaja R^2 puudus: lisades mudelisse uusi tunnuseid, determinatsioonikordaja alati suureneb.

Determinatsioonikordajat R^2 ei saa kasutada erineva tunnuste arvuga mudelite võrdlemiseks.

KORRIGEERITUD DETERMINATSIOONIKORDAJA

Et võrrelda mudeleid, kus on erinev arv tunnuseid, kasutatakse korrigeeritud determinatsioonikordajat (modifitseeritud, *adjusted*):

$$R_a^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

kus *n* on valimi maht ja *k* seletavate tunnuste arv mudelis.

Kui uue tunnuse lisamisel mudelisse R_a^2 suureneb, siis mudel paraneb.

Kui R_a^2 väheneb, siis uue tunnuse lisamine pole õigustatud.

NÄIDE: TÖÖJÕU PAKKUMINE

Uuring USA-s 1966.a, valimis 6000 leibkonda.
Millest sõltub töötatud tundide arv aastas?

Ühe seletava tunnusega mudel

$$\text{TUNNID} = 1913 + 80,9 \cdot \text{TTASU}, \quad R^2 = 0,333, \; R_a^2 = 0,315$$

Teine mudel on parem, sest korrigeeritud determinatsioonikordaja on suurem.

Kolme seletava tunnusega mudel

$$\text{TUNNID} = 2444,8 - 47,6 \cdot \text{TTASU} + 0,02641 \cdot \text{VARAD} - 8,66 \cdot \text{VANUS}, \quad R^2 = 0,715, \; R_a^2 = 0,691$$

NÄIDE: TÖÖJÕU PAKKUMINE

Uuring USA-s 1966.a, valimis 6000 leibkonda.
Millest sõltub töötatud tundide arv aastas?

Ühe seletava tunnusega mudel

$$\text{TUNNID} = 1913 + 80,9 \cdot \text{TTASU}, \quad R^2 = 0,333,\ R_a^2 = 0,315$$

Kolme seletava tunnusega mudel

$$\text{TUNNID} = 2444,8 - 47,6 \cdot \text{TTASU} + 0,02641 \cdot \text{VARAD} - 8,66 \cdot \text{VANUS}, \quad R^2 = 0,715,\ R_a^2 = 0,691$$

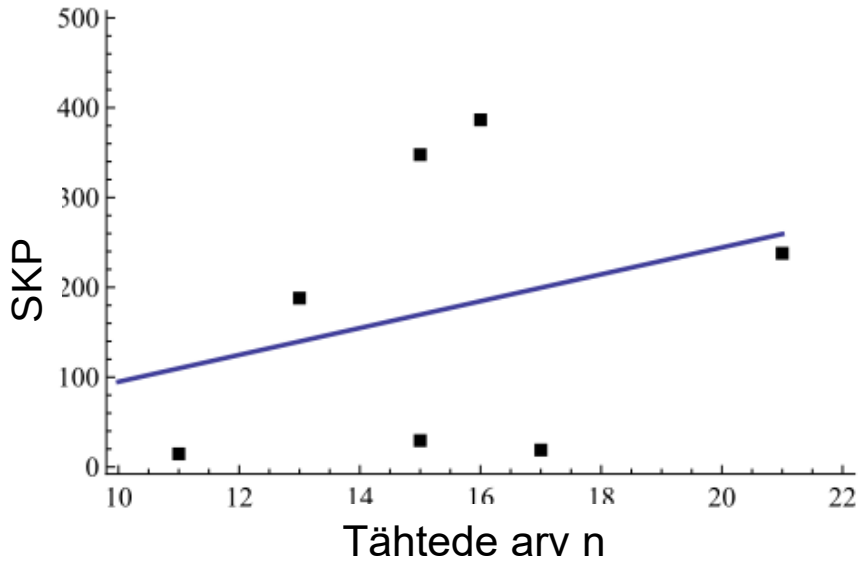
Oluliste seletavate tunnuste lisamine muutis tunnitasu kordaja märki, st muutis tunnitasu mõju suunda.

Kui mudelis puuduvad mõned olulised seletavad tunnused, võib mudel anda VALESID tulemusi.

NÄIDE: PEAMINISTRI NIMI JA SKP

Aastal 2012

Riik	Peaminister		Tähtede arv nimes	SKP, mld €
Eesti	Andrus	Ansip	11	16,0
Läti	Valdis	Dombrovskis	17	20,2
Leedu	Andrius	Kubilius	15	30,7
Soome	Jyrki	Katainen	13	189,4
Rootsi	Fredrik	Reinfeldt	16	387,9
Taani	Helle	Thorning-Schmidt	21	239,2
Norra	Jens	Stoltenberg	15	349,1



Regressioonmudel

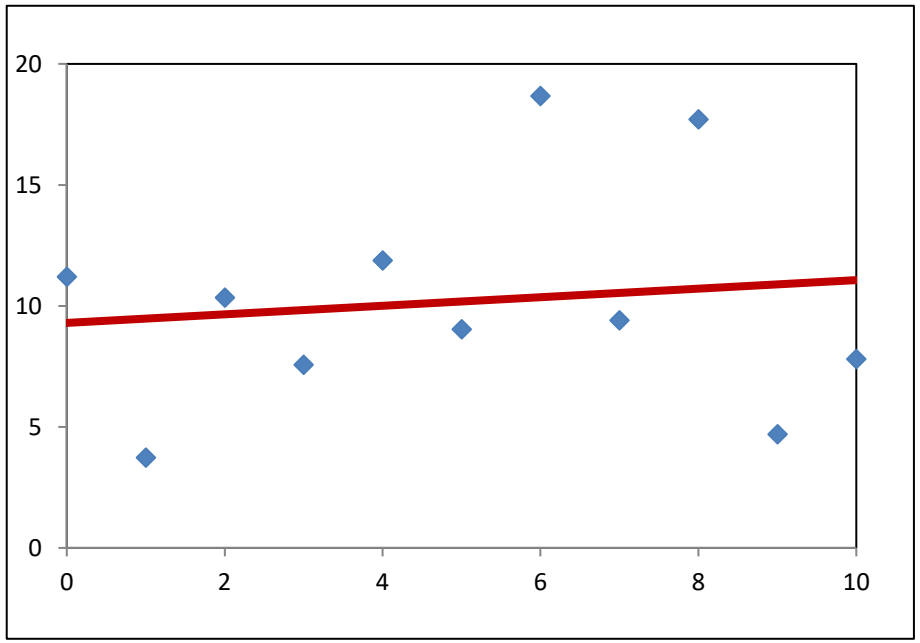
$$SKP = -54,68 + 14,96n, \quad R^2 = 0,089.$$

kus *n* on tähtede arv peaministri nimes.

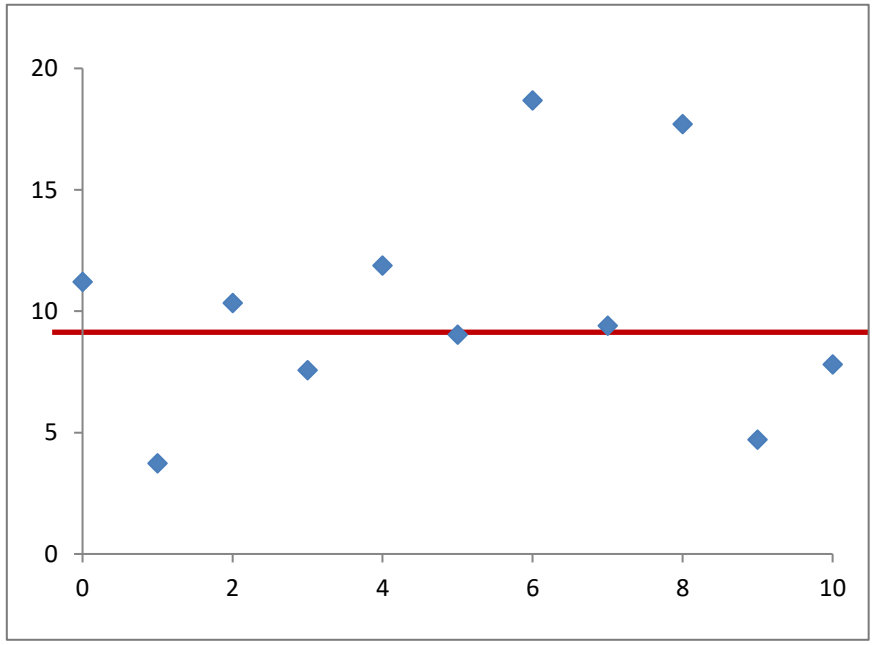
REGRESSIOONI JUHUSLIKKUS

Kuidas otsustada, kas

Y tõesti sõltub tunnusest X

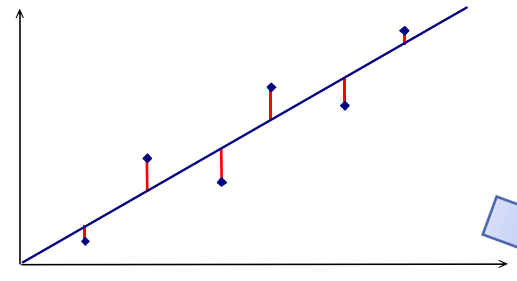


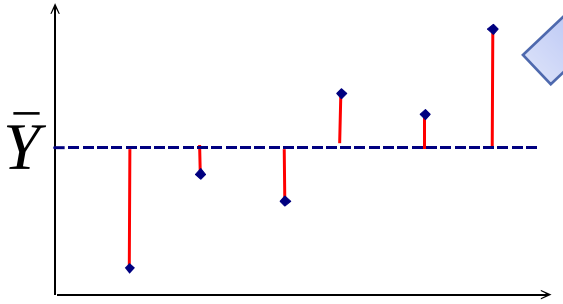
või kõigub juhuslikult ümber keskväärtuse?



Otsustamiseks on ANOVA tabelil põhinev *F*-test.

REGRESSIOONANALÜÜSI ANOVA TABEL





Varieeruvuse allikas	Vabadus-astmete arv df	Hälvete ruutude summa SS	Keskruut MS	*F* statistik	Olulisuse tõenäosus
Regressioon-mudel	K	$SSR=SST-SSE$	$MSR = \frac{SSR}{K}$	$F = \frac{MSR}{MSE}$	p
Jääkliikmed	$n - 1 - K$	$SSE = \sum(y_i - \hat{y})^2$	$MSE = \frac{SSE}{n-1-K}$		
Summaarne	$n - 1$	$SST = \sum(y_i - \bar{y})^2$			

F näitab, mitu korda on regressioonmudeliga ära seletatud hajumine suurem seletamata hajumisest (jääkhajumisest).

REGRESSIOONI ANOVA TABEL EXCELIS

Tööjõu pakkumise mudel

Varieeruvuse allikas	Vabadus-astmete arv	Hälvete ruutude summa	Keskruut	*F* statistik	Olulisuse tõenäosus

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	111393	37130,9	29,3	1,18E-09
Residual	35	44391	1268,3		
Total	38	155783			

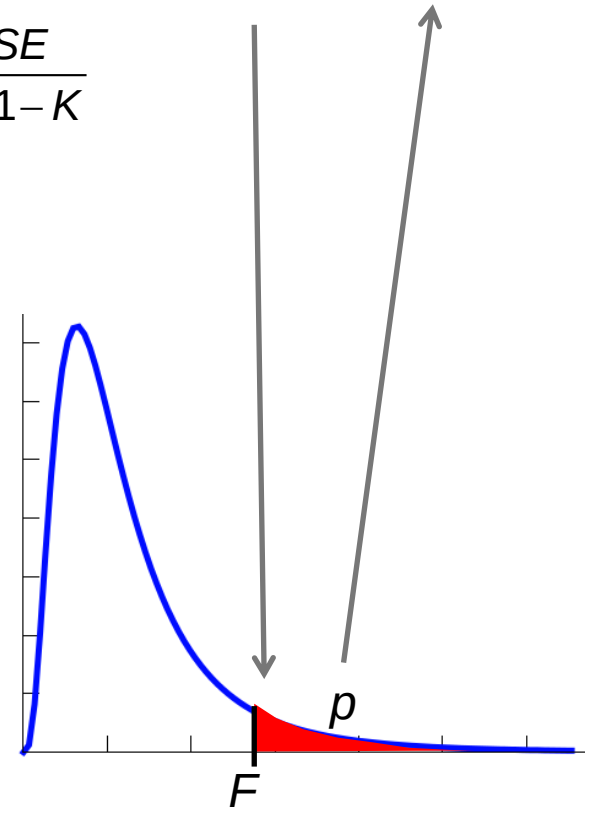
Kui suur peab *F* olema, et mudel oleks statistiliselt oluline?

Nii suur, et sellele vastav olulisuse tõenäosus (*Significance F*) oleks väiksem kriitilisele vastavast olulisuse nivoost 0,05.

REGRESSIOONANALÜÜSI ANOVA TABEL

Varieeruvuse allikas	Vabadus-astmete arv df	Hälvete ruutude summa SS	Keskruut MS	*F* statistik	Olulisuse tõenäosus
Regressioon-mudel	K	$SSR=SST-SSE$	$MSR = \frac{SSR}{K}$	$F = \frac{MSR}{MSE}$	p
Jääkliikmed	$n - 1 - K$	$SSE = \sum(y_i - \hat{y})^2$	$MSE = \frac{SSE}{n-1-K}$		
Summaarne	$n - 1$	$SST = \sum(y_i - \bar{y})^2$			

Olulisuse tõenäosus *p* leitakse *F*-jaotusest.



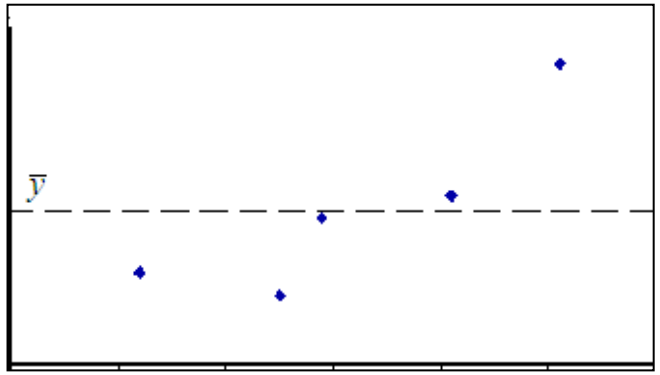
F-TEST JA MUDELI STATISTILINE OLULISUS

Mudeli statistilise olulisuse kontrollimiseks kasutatakse *F*-testi.

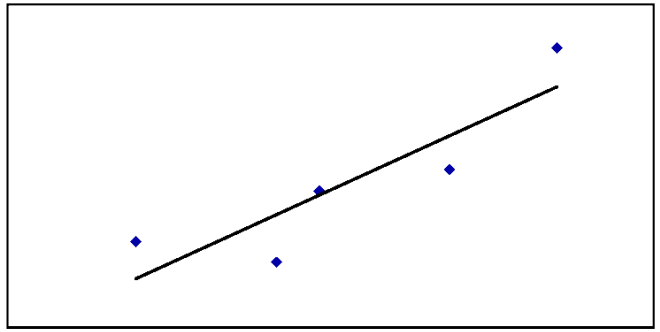
H_0 kõik seletavate tunnuste kordajad on nullid, $a_1=a_2=\ldots=a_k=0$

H_1 vähemalt üks kordaja a_i on nullist erinev

Nullhüpotees: *Y* on määratud oma keskväärtusega:



Sisukas hüpotees



Otsustamiseks võrreldakse empiirilisele väärtusele vastavat olulisuse tõenäosust *p* olulisuse nivooga α.

$p > \alpha$ nullhüpotees, mudel ei ole statistiliselt oluline.
$p < \alpha$ sisukas hüpotees, mudel on statistiliselt oluline.

NÄIDE: *F*-TEST JA MUDELI STATISTILINE OLULISUS

Peaministri nimi ja SKP $SKP = -54,68 + 14,96n$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	13355	13355	0,489	0,516
Residual	5	136670	27334		
Total	6	150026			

0,516> 0,05

Võtta vastu nullhüpotees, vastav mudel ei ole statistiliselt oluline.

Tööjõu pakkumise mudel $\mathrm{TUNNID} = 2444,8 - 47,6 \cdot \mathrm{TTASU} + 0,02641 \cdot \mathrm{VARAD} - 8,66 \cdot \mathrm{VANUS}$

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	111393	37130,9	29,3	1,18E-09
Residual	35	44391	1268,3		
Total	38	155783			

$1,18 \cdot 10^{-9} < 0,05$

Võtta vastu sisukas hüpotees, vastav mudel on statistiliselt oluline.

NÄIDE: TÖÖJÕU PAKKUMINE

Uuring USA-s 1966.a, valimis 6000 leibkonda.
Millest sõltub töötatud tundide arv aastas?

$$\text{TUNNID} = 2444{,}8 - 47{,}6 \cdot \text{TTASU} + 0{,}02641 \cdot \text{VARAD} - 8{,}66 \cdot \text{VANUS}, \qquad R^2 = 0{,}715$$

TUNNID	töötundide arv aastas
TTASU	tunnitasu, $
VARAD	varade suurus, tuh $
VANUS	vanus aastates

Mudeli statistilise olulisuse testimine:
F-testi olulisuse tõenäosus $p = 1{,}18 \cdot 10^{-9} < 0{,}05$.
Mudel on statistiliselt oluline: vähemalt ühe tunnuse kordaja on nullist erinev.

Kas kõik kordajad on statistiliselt olulised?
Kas kõik mudelis olevad seletavad tunnused mõjutavad töötundide arvu?

PARAMEETRITE STATISTILISE OLULISUSE TESTIMINE

$$y = b + a_1 x_1 + a_2 x_2 + \ldots + a_K x_K + \varepsilon$$

Iga parameetri jaoks eraldi:

Nullhüpotees H_0: $a_j = 0$

Sisukas hüpotees H_1: $a_j \neq 0$

t-test, teststatistik $t = \frac{a_j}{se(a_j)} \sim t(n - K - 1)$

$se(a_j)$ on parameetri standardviga

Kui $t > t_{kriitiline}$ ($p < \alpha$), on vastav parameeter oluliselt nullist erinev: tunnuse lülitamine mudelisse on põhjendatud.

Vastupidisel juhul tuleks tunnus mudelist eemaldada, st

viia läbi uue mudeli hindamine ilma selle tunnuseta.

Demo: parameetri t-test

NÄIDE: TÖÖJÕU PAKKUMINE, 1

Aruanne Excelis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,846
R Square	0,715
Adjusted R Square	0,691
Standard Error	35,61
Observations	39

ANOVA

	df	SS	MS	F	Significance F
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2444,8	93,62	26,113	1,54E-24	2254,7	2634,9
TTASU	-47,6	23,01	-2,070	0,0459	-94,32	-0,91
VARAD	0,02641	0,00393	6,720	8,8E-08	0,018	0,034
VANUS	-8,66	1,706	-5,078	1,27E-05	-12,13	-5,20

Mudel on statistiliselt oluline
$1{,}18 \times 10^{-9} < 0{,}05$

NÄIDE: TÖÖJÕU PAKKUMINE, 2

Aruanne Excelis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,846
R Square	0,715
Adjusted R Square	0,691
Standard Error	35,61
Observations	39

Kõik parameetrid on statistiliselt olulised.
Vastavad olulisuse tõenäosused:

vabaliige $1{,}54 \times 10^{-24} < 0{,}05$

TTASU $0{,}0459 < 0{,}05$

VARAD $8{,}8 \times 10^{-8} < 0{,}05$

VANUS $1{,}27 \times 10^{-5} < 0{,}05$

ANOVA

	df	SS	MS	F	Significance F
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2444,8	93,62	26,113	1,54E-24	2254,7	2634,9
TTASU	-47,6	23,01	-2,070	0,0459	-94,32	-0,91
VARAD	0,02641	0,00393	6,720	8,8E-08	0,018	0,034
VANUS	-8,66	1,706	-5,078	1,27E-05	-12,13	-5,20

NÄIDE: TÖÖJÕU PAKKUMINE, 3

Aruanne Excelis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,846
R Square	0,715
Adjusted R Square	0,691
Standard Error	35,61
Observations	39

ANOVA

	df	SS	MS	F	Significance F
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2444,8	93,62	26,113	1,54E-24	2254,7	2634,9
TTASU	-47,6	23,01	-2,070	0,0459	-94,32	-0,91
VARAD	0,02641	0,00393	6,720	8,8E-08	0,018	0,034
VANUS	-8,66	1,706	-5,078	1,27E-05	-12,13	-5,20

$$TUNNID = 2444{,}8 - 47{,}6 \cdot TTASU + \\ +0{,}02641 \cdot VARAD - 8{,}66 \cdot VANUS$$

Kui tunnused on olulised, võime esitada mudeli.

NÄIDE: TÖÖJÕU PAKKUMINE, 4

Aruanne Excelis

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,846
R Square	0,715
Adjusted R Square	0,691
Standard Error	35,61
Observations	39

$$TUNNID = 2444{,}8 - 47{,}6 \cdot TTASU + \\ +0{,}02641 \cdot VARAD - 8{,}66 \cdot VANUS$$

Determinatsioonikordaja R^2=0,715.
Mudel seletab ära 71,5% töötundide arvu varieerumisest.

ANOVA

	df	SS	MS	F	Significance F
Regression	3	111392,6	37130,9	29,28	1,18E-09
Residual	35	44390,6	1268,3		
Total	38	155783,2			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2444,8	93,62	26,113	1,54E-24	2254,7	2634,9
TTASU	-47,6	23,01	-2,070	0,0459	-94,32	-0,91
VARAD	0,02641	0,00393	6,720	8,8E-08	0,018	0,034
VANUS	-8,66	1,706	-5,078	1,27E-05	-12,13	-5,20

NÄIDE: NUTIKELLADE HIND, 1

Millest sõltub nutikella hind?
Analüüsi aluseks 29 nutikella Euronicsi veebipoe kodulehelt aastal 2023.
12 Apple ja 17 Samsungi nutikella.

Tunnused:

HIND	hind eurodes;
VANUS	mudeli vanus aastates;
EKDGN	ekraani diagonaal tollides;
KAAL	kaal grammides;
AKU	aku mahutavus, mAh.

Lineaarne mudel

$$\text{HIND} = b + a_1\text{VANUS} + a_2\,\text{EKDGN} + a_3\,\text{KAAL} + a_4\,\text{AKU} + \varepsilon$$

Allikas: Hansar, H., Suur, S.-M., (2023) Ökonomeetria kodutöö „Nutikella hinna sõltuvus tehnilistest näitajatest“

NÄIDE: NUTIKELLADE HIND, 2

Regression Statistics	
Multiple R	0,842
R Square	0,709
Adjusted R Square	0,661
Standard Error	60,50
Observations	29

Mudel on statistiliselt oluline
$3{,}5 \times 10^{-6} < 0{,}05$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	214133,8	53533	14,625	3,5E-06
Residual	24	87852,39	3661		
Total	28	301986,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	45,7	119	0,39	0,7034	-199,3	290,8
VANUS	-62,9	18,2	-3,46	0,0020	-100,4	-25,4
EKDGN	215,4	51,8	4,16	0,0004	108,4	322,3
KAAL	3,80	1,41	2,69	0,0129	0,88	6,71
AKU	-0,130	0,140	-0,93	0,3632	-0,418	0,159

NÄIDE: NUTIKELLADE HIND, 3

Regression Statistics	
Multiple R	0,842
R Square	0,709
Adjusted R Square	0,661
Standard Error	60,50
Observations	29

Parameetrite statistilise olulisuse kontroll.
Vastavad olulisuse tõenäosused:

VANUS	0,0020 < 0,05	on oluline
EKDGN	0,0004 < 0,05	on oluline
KAAL	0,0129 < 0,05	on oluline
AKU	0,3632 > 0,05	ei ole oluline

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	214133,8	53533	14,625	3,5E-06
Residual	24	87852,39	3661		
Total	28	301986,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	45,7	119	0,39	0,7034	-199,3	290,8
VANUS	-62,9	18,2	-3,46	0,0020	-100,4	-25,4
EKDGN	215,4	51,8	4,16	0,0004	108,4	322,3
KAAL	3,80	1,41	2,69	0,0129	0,88	6,71
AKU	-0,130	0,140	-0,93	0,3632	-0,418	0,159

NÄIDE: NUTIKELLADE HIND, 4

Regression Statistics	
Multiple R	0,842
R Square	0,709
Adjusted R Square	0,661
Standard Error	60,50
Observations	29

ANOVA

Regression	
Residual	
Total	

	Co
Intercept	
VANUS	
EKDGN	
KAAL	
AKU	

Parameetrite statistilise olulisuse kontroll.
Vastavad olulisuse tõenäosused:

VANUS	0,0020 < 0,05	on oluline
EKDGN	0,0004 < 0,05	on oluline
KAAL	0,0129 < 0,05	on oluline
AKU	0,3632 > 0,05	ei ole oluline

Kuna AKU pole statistiliselt oluline, tuleb läbi viia uue mudeli hindamine ilma tunnuseta AKU.

$$\text{HIND} = b + a_1\,\text{VANUS} + a_2\,\text{EKDGN} + a_3\,\text{KAAL} + \varepsilon$$

NÄIDE: NUTIKELLADE HIND, 5

Regression Statistics	
Multiple R	0,836
R Square	0,699
Adjusted R Square	0,663
Standard Error	60,33
Observations	29

Korrigeeritud determinatsiooni-kordaja suurenes, enne oli 0,661.

Parameetrite statistilise olulisuse kontroll.
Vastavad olulisuse tõenäosused:

VANUS	0,0021 < 0,05	on oluline
EKDGN	0,0002 < 0,05	on oluline
KAAL	0,0177 < 0,05	on oluline

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	210988,5	70329,5	19,322	1,07E-06
Residual	25	90997,67	3639,91		
Total	28	301986,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13,4	113	0,12	0,907	-219,7	246,4
VANUS	-61,9	18,1	-3,42	0,0021	-99,2	-24,7
EKDGN	220,5	51,4	4,29	0,0002	114,7	326,3
KAAL	3,24	1,27	2,54	0,0177	0,61	5,86

Parameetrid eelmises mudelis.

45,7
-62,9
215,4
3,80

NÄIDE: NUTIKELLADE HIND, 6

Regression Statistics	
Multiple R	0,836
R Square	0,699
Adjusted R Square	0,663
Standard Error	60,33
Observations	29

ANOVA

	df	*SS*			
Regression	3	210988,5	70329,5	19,322	1,07E-06
Residual	25	90997,67	3639,91		
Total	28	301986,2			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13,4	113	0,12	0,907	-219,7	246,4
VANUS	-61,9	18,1	-3,42	0,0021	-99,2	-24,7
EKDGN	220,5	51,4	4,29	0,0002	114,7	326,3
KAAL	3,24	1,27	2,54	0,0177	0,61	5,86

$$HIND = 13,4 - 61,9\,VANUS + 220,5\,EKDGN + 3,24\,KAAL + \varepsilon$$

Ühe aasta võrra vanemal mudelil hind 61,9 € väiksem, kui muud tunnused samad.
Kui ekraani diagonaal on 1 tolli võrra suurem, on hind 220,5 € suurem, kui muud tunnused samad.
Ühe grammi võrra raskema nutikella hind on 3,24 € võrra suurem, kui muud tunnused samad.

NÄIDE: NUTIKELLADE HIND, 7

Regression Statistics	
Multiple R	0,836
R Square	0,699
Adjusted R Square	0,663
Standard Error	60,33
Observations	29

$$\text{HIND} = 13,4 - 61,9\,\text{VANUS} + 220,5\,\text{EKDGN} + 3,24\,\text{KAAL} + \varepsilon$$

Determinatsioonikordaja R^2=0,699.
Mudel seletab ära 69,9% hinna varieerumisest.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	210988,5	70329,5	19,322	1,07E-06
Residual	25	90997,67	3639,91		
Total	28	301986,2			

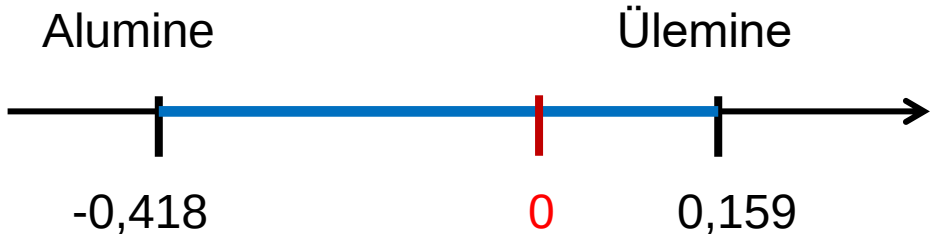
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13,4	113	0,12	0,907	-219,7	246,4
VANUS	-61,9	18,1	-3,42	0,0021	-99,2	-24,7
EKDGN	220,5	51,4	4,29	0,0002	114,7	326,3
KAAL	3,24	1,27	2,54	0,0177	0,61	5,86

NÄIDE: PARAMEETRITE OLULISUS JA USALDUSPIIRID

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	45,7	119	0,39	0,7034	-199,3	290,8
VANUS	-62,9	18,2	-3,46	0,0020	-100,4	-25,4
EKDGN	215,4	51,8	4,16	0,0004	108,4	322,3
KAAL	3,80	1,41	2,69	0,0129	0,88	6,71
AKU	-0,130	0,140	-0,93	0,3632	-0,418	0,159

Aku mahutavus mitteoluline

Usalduspiirid aku mahutavuse parameetri jaoks, usaldatavusega 95%.



Kui parameetri väärtus võib olla 0, siis järelikult vastav tunnus mudelisse ei kuulu.

VABALIIKME STATISTILINE MITTEOLULISUS

$$y = b + a_1x_1 + a_2x_2 + \ldots + a_Kx_K + \varepsilon$$

Vabaliige *b* võib olla statistiliselt mitteoluline (H_0).

Seda mudelist ei eemaldata.

- Va juhul, kui selle puudumine on loogiliselt põhjendatud (vt õpikust ptk 9.16 „Lineaarse mudeli vabaliige ja nullpunkti läbiv regressioonsirge“).

Seepärast ei pea vabaliikme statistilist olulisust üldjuhul testima.

Vabaliikme olemasolu garanteerib selle, et jääkliikmete summa on 0.

Näide: nutikella hind

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	13,4	113	0,12	0,907	-219,7	246,4
VANUS	-61,9	18,1	-3,42	0,0021	-99,2	-24,7
EKDGN	220,5	51,4	4,29	0,0002	114,7	326,3
KAAL	3,24	1,27	2,54	0,0177	0,61	5,86

TUNNUSTE VALIK

TUNNUSTE VALIK

Potentsiaalseid seletavaid tunnuseid võib olla väga palju.

Kuidas valikut teha?

Edaspidine valik:

liigutakse väiksema tunnuste arvuga mudelist suurema poole, st tunnuseid lisatakse ükshaaval.

Tagurpidine valik:

alustatakse suure arvu tunnustega ning ülearused eemaldatakse ükshaaval.

EDASPIDINE VALIK

Millises järjekorras tunnuseid lisada?

Potentsiaalsed tunnused tuleb reastada tähtsuse järjekorras. Aluseks näiteks korrelatsioonikordajad funktsioontunnusega.

1. Leitakse korrelatsioonimaatriks.
2. See X, mille korrelatsioonikordaja absoluutväärtus on kõige suurem, lisatakse mudelisse esimesena.
3. Kui see on statistiliselt oluline, jäetakse mudelisse.
4. Lisatakse tähtsuselt järgmine tunnus, mille korrelatsioonikordaja absoluutväärtus on suuruselt järgmine, ja kontrollitakse selle statistilist olulisust.
5. Jätkatakse seni, kuni järgmine tunnus ei ole enam statistiliselt oluline või kui korrigeeritud determinatsioonikordaja väheneb.

NÄIDE: KÄIBE MUDEL, 1

Sõltuv tunnus käive Y.

Potentsiaalsed seletavad tunnused:
toote hind *P;*
turunduskulud *M;*
majanduskasvu indeks *E;*
tooteühiku kulude indeks *C*.

Korrelatsioonimaatriks

	Y	*P*	*M*	*E*	*C*
Y	1				
P	0,137	1			
M	0,636	0,656	1		
E	0,846	-0,073	0,285	1	
C	0,237	0,855	0,738	0,0155	1

Reastatud tähtsuse järjekorras

1. E
2. M
3. C
4. P

NÄIDE: KÄIBE MUDEL, 2

Multiple R	0,846
R Square	0,716
Adjusted R Square	0,700
Standard Error	123,85
Observations	20

	Coefficients	Standard Error	t Stat	P-value
Intercept	-3989	774	-5,154	6,66E-05
E	50,13	7,45	6,732	2,61E-06

Parameetrite olulisuse hindamiseks kasutame nivood 0,1.

1. E
2. M
3. C
4. P

Multiple R	0,941
R Square	0,885
Adjusted R Square	0,872
Standard Error	80,93
Observations	20

Suurenes

	Coefficients	Standard Error	t Stat	P-value
Intercept	-3527	514	-6,86	2,76E-06
E	42,88	5,08	8,45	1,73E-07
M	332,9	66,4	5,02	0,000106

NÄIDE: KÄIBE MUDEL, 3

Multiple R	0,952
R Square	0,906
Adjusted R Square	0,888
Standard Error	75,72
Observations	20

Suurenes

	Coefficients	Standard Error	t Stat	P-value
Intercept	-2642	679	-3,89	0,00130
E	40,10	4,98	8,05	5,1E-07
M	469,7	96,6	4,86	1,7E-04
C	-6,29	3,40	-1,85	0,083

Multiple R	0,953
R Square	0,908
Adjusted R Square	0,883
Standard Error	77,364
Observations	20

Vähenes

Ei sobi!

	Coefficients	Standard Error	t Stat	P-value
Intercept	-2639	693	-3,805	0,0017
E	39,46	5,21	7,571	1,7E-06
M	478,5	99,9	4,791	2,4E-04
C	-4,35	4,85	-0,897	0,38
P	-68,31	119	-0,574	0,57

C ja P mitte-olulised

Parameetrite olulisuse hindamiseks kasutame nivood 0,1.

1. E
2. M
3. C
4. P

NÄIDE: KÄIBE MUDEL, 4

Lõplikus mudelis on seletavateks tunnusteks
majanduskasvu indeks *E;*
turunduskulud *M;*
tooteühiku kulude indeks *C*.

$$\hat{Y} = -2642 + 40{,}10E + 469{,}7M - 6{,}29C, \quad R^2 = 0{,}906$$

TAGURPIDINE VALIK

Alustatakse mudelist, milles on kõik potentsiaalsed seletavad tunnused.

1. Kui on tunnuseid, mis ei ole statistiliselt olulised ette võetud nivool, siis eemaldatakse tunnus, mille olulisuse tõenäosus on kõige suurem.
2. Eemaldatakse järgmine mitteoluline tunnus, mille olulisuse tõenäosus on järelejäänud tunnuste hulgas kõige suurem.
3. Jätkatakse, kuni kõik mudelisse jäänud tunnused on valitud nivool statistiliselt olulised.

NÄIDE: REISIJATE ARV LINNALIINI BUSSIDES, 1

40 USA linna andmed aastast 1988.

Sõltuv tunnus
REISIJAD reisijate arv (tuh reisijat tunnis);

Potentsiaalsed seletavad tunnused
HIND bussipileti hind ($);
BENSIIN galloni bensiini hind ($);
TULU keskmine sissetulek elaniku kohta ($);
RHV linna rahvaarv (tuh);
TIHEDUS rahvastiku tihedus (tuh elanikku ruutmiili kohta);
PINDALA linna pindala (ruutmiili).

NÄIDE: REISIJATE ARV LINNALIINI BUSSIDES, 2

Adjusted R Square	0,907

	Coefficients	Standard Error	t Stat	P-value
Intercept	2745	2642	1,039	0,306
HIND	-239	452	-0,528	0,601
BENSIIN	522	2658	0,196	0,845
TULU	-0,195	0,0649	-3,001	0,0051
RHV	1,711	0,231	7,397	1,69E-08
TIHEDUS	0,116	0,0596	1,954	0,059
PINDALA	-1,155	1,80	-0,641	0,526

Kõige suurem

Eemaldame tunnuse BENSIIN.

Adjusted R Square	0,909

Suurenes

	Coefficients	Standard Error	t Stat	P-value
Intercept	3216	1090	2,949	0,00573
HIND	-226	440	-0,512	0,612
TULU	-0,1957	0,0638	-3,069	0,00420
RHV	1,717	0,226	7,581	8,33E-09
TIHEDUS	0,1182	0,0580	2,037	0,049
PINDALA	-1,20	1,77	-0,677	0,503

Kõige suurem

Eemaldame tunnuse HIND.

NÄIDE: REISIJATE ARV LINNALIINI BUSSIDES, 3

Adjusted R Square	0,911	Suurenes

	Coefficients	Standard Error	t Stat	P-value
Intercept	3021	1011	2,99	0,00512
TULU	-0,1939	0,0630	-3,08	0,00404
RHV	1,731	0,222	7,79	3,81E-09
TIHEDUS	0,1159	0,0572	2,03	0,0505
PINDALA	-1,41	1,70	-0,83	0,413

Kõige suurem

Eemaldame tunnuse PINDALA.

Adjusted R Square	0,912	Suurenes

	Coefficients	Standard Error	t Stat	P-value
Intercept	2816	976	2,88	0,00659
TULU	-0,2013	0,0621	-3,24	0,00257
RHV	1,577	0,121	13,07	3,10E-15
TIHEDUS	0,1534	0,0349	4,40	9,34E-05

Kõik tunnused on statistiliselt olulised nivool 0,01.

Lõplik mudel

$$REISIJAD = 2816 - 0,2013 \cdot TULU + 1,577 \cdot RHV + 0,1534 \cdot TIHEDUS$$

$$R^2 = 0,919$$

MULTIKOLLINEAARSUS

NÄIDE: HÕIVATUTE OSAKAAL ERINEVATES SEKTORITES, 1

26 OECD riigi kohta on järgmised andmed (2012. a):

GDPPC	SKP elaniku kohta (tuh $);
AGR	põllumajanduses hõivatute osakaal (%);
IND	tööstuses hõivatute osakaal (%);
SER	teeninduses hõivatute osakaal (%).

Hindame lineaarset mudelit

$$\text{GDPPC} = b + a_1 \cdot \text{AGR} + a_2 \cdot \text{IND} + a_3 \cdot \text{SER} + \varepsilon$$

NÄIDE: HÕIVATUTE OSAKAAL ERINEVATES SEKTORITES, 2

Mudel tervikuna on statistiliselt oluline.

ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	2861,5	953,8	6,553	0,00247	
Residual	22	3202,2	145,6			
Total	25	6063,7				
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	213228,6	379626,2462	0,5617	0,5800	-574068	1000525
AGR	-213315	379620,1428	-0,5619	0,5799	-1000599	573969
SER	-213158	379626,2797	-0,5615	0,5801	-1000455	574138
IND	-213256	379624,6269	-0,5618	0,5800	-1000549	574037

Ükski parameeter pole statistiliselt oluline.

NÄIDE: HÕIVATUTE OSAKAAL ERINEVATES SEKTORITES, 4

Sõltumatud tunnused SER ja AGR ning SER ja IND pole sõltumatud!

Korrelatsioonimaatriks

	GDPPC	*AGR*	*SER*	*IND*
GDPPC	1			
AGR	-0,568	1,000		
SER	0,665	-0,693	1,000	
IND	-0,416	0,071	-0,768	1

Sõltumatute tunnuste omavahelised seosed on tugevamad kui nende seos sõltuva tunnusega.

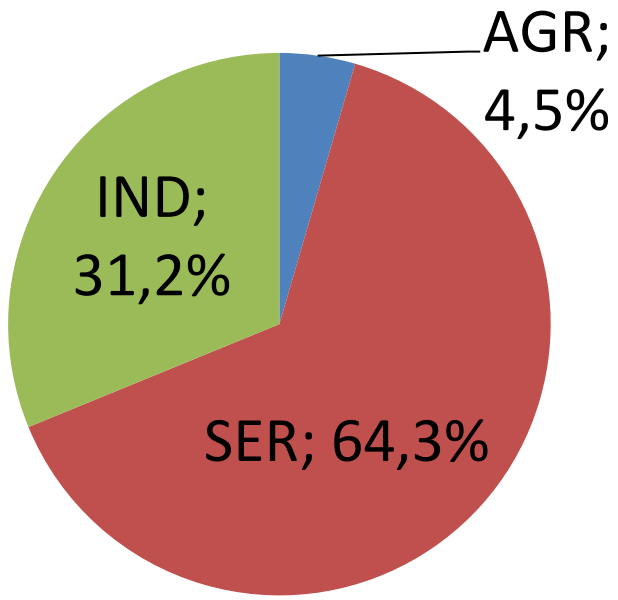
MULTIKOLLINEAARSUS

NÄIDE: HÕIVATUTE OSAKAAL ERINEVATES SEKTORITES, 5

Sõltumatud tunnused AGR, IND ja SER pole sõltumatud!

$$AGR + IND + SER = 100\%$$

Näiteks Eestis



NÄIDE: HÕIVATUTE OSAKAAL ERINEVATES SEKTORITES, 6

Multiple R	0,665
R Square	0,442
Adjusted R Square	0,419
Standard Error	11,872
Observations	26

Mudelisse jätame ainult teeninduses hõivatute osakaalu SER, sest selle korrelatsioon sõltuva tunnusega on kõige suurem.

ANOVA

	df	SS	MS	F	Significance F
Regression	1	2681,0	2681,0	19,02	0,00021
Residual	24	3382,7	140,9		
Total	25	6063,7			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-50,9	20,2	-2,52	0,019	-92,5	-9,3
SER	125,7	28,8	4,36	0,00021	66,2	185,2

$$\text{GDPPC} = -50,9 + 125,7 \cdot \text{SER} \qquad R^2 = 0,442$$

MULTIKOLLINEAARSUSE TUNNUSED

Mõned multikollineaarsuse tunnused:

1. Mõne sõltumatute tunnuste paari omavaheline korrelatsioon on tugevam kui korrelatsioon sõltuva muutujaga.
2. Mudeli parameetritel on väga suured standardvead.
3. Regressioonmudeli ühe või mitme parameetri märk on ebaloogiline.

KVALITATIIVSED SELETAVAD TUNNUSED

KVALITATIIVSED SELETAVAD TUNNUSED

Isikuküsitlustel

sugu	nimiskaalas
ametikoht	nimiskaalas
haridustase	järjestusskaalas

Ettevõtted

tegevusala	nimiskaalas
omandivorm	nimiskaalas

Kuidas mudelisse panna?

NÄIDE: IPOD-IDE HINNAD EBAY OKSJONITEL, 1

Apple iPod mini oksjonite andmed 27. juuni kuni 18. juuli 2008, kokku 1225 oksjonit.

HIND lõpphind dollarites;
ALGH alghind dollarites;
ARV pakkumiste arv;

Intervallskaalas

SKORD müügil oleva iPod-i seisukord:

- halb (mõrane klaas, katkine kõrvaklappide ühendus, aku mahutavus väike);
- keskmine (kriimustused);
- hea.

Järjestusskaalas, kvalitatiivne

Kuidas lõpphind sõltub alghinnast, pakkumiste arvust ja seisukorrast?

NÄIDE: IPOD-IDE HINNAD EBAY OKSJONITEL, 2

SKORD tasemete jaoks tuleb luua kaheväärtuselised fiktiivsed tunnused:

$$D_1 = \begin{cases} 1, & \text{kui seisukord on keskmine} \\ 0, & \text{kui seisukord on muu} \end{cases}$$

$$D_2 = \begin{cases} 1, & \text{kui seisukord on halb} \\ 0, & \text{kui seisukord on muu} \end{cases}$$

Seisukord "hea" on baasväärtus.

Kahe fiktiivse tunnusega on kodeeritud 3 taset.
Näiteks 3 iPod-i andmed

HIND	ALGH	ARV	D1	D2	SKORD
135	0,99	8	0	0	hea
46	0,99	7	1	0	keskmine
56	20	4	0	1	halb

Mudelit otsime kujul $\text{HIND} = b + a_1 \cdot \text{ALGH} + a_2 \cdot \text{ARV} + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$

NÄIDE: IPOD-IDE HINNAD EBAY OKSJONITEL, 3

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	59,76	3,35	17,8	1,69E-63	53,19	66,33
ALGH	0,554	0,0250	22,2	5,87E-92	0,51	0,60
ARV	6,569	0,341	19,2	2,59E-72	5,90	7,24
D1	-16,11	2,07	-7,8	1,44E-14	-20,17	-12,05
D2	-40,47	3,24	-12,5	8,67E-34	-46,83	-34,12

keskmine

halb

$$HIND = 59,76 + 0,554 \cdot ALGH + 6,569 \cdot ARV - 16,11 D_1 - 40,47 D_2$$

Heas seisukorras iPod: $D_1 = 0, \quad D_2 = 0$

$$HIND = 59,76 + 0,554 \cdot ALGH + 6,569 \cdot ARV$$

Keskmises seisukorras iPod: $D_1 = 1, \quad D_2 = 0$

$$HIND = 59,76 + 0,554 \cdot ALGH + 6,569 \cdot ARV - 16,11$$

Hind on 16,11 $ väiksem kui heas seisukorras iPod-il.

Halvas seisukorras iPod: $D_1 = 0, \quad D_2 = 1$

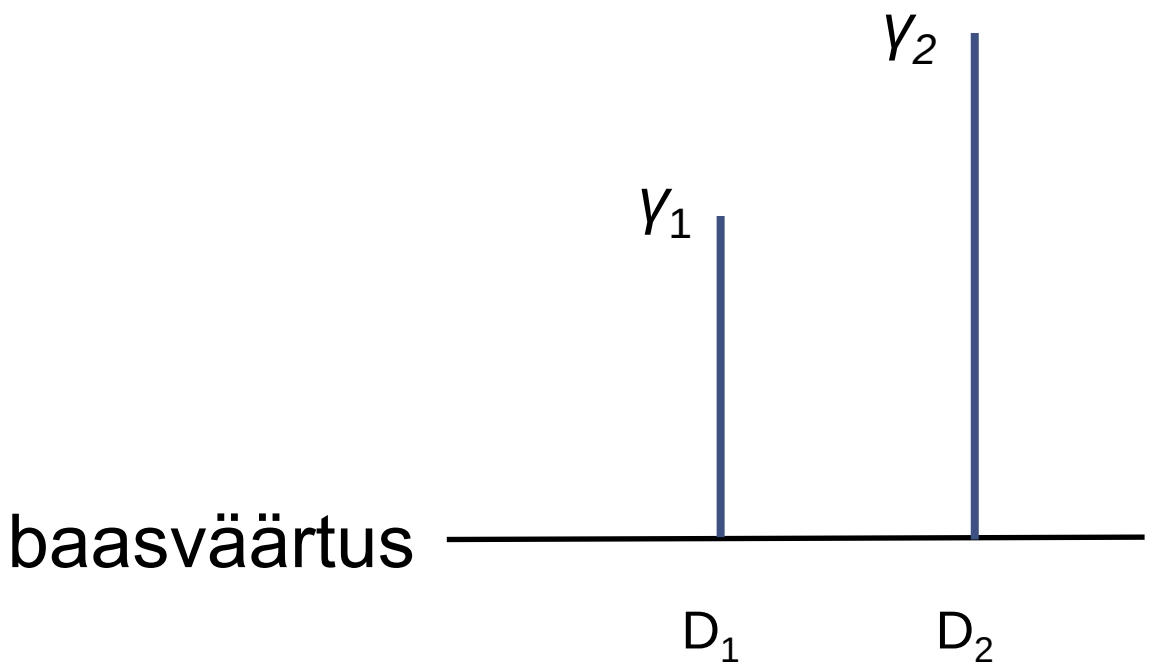
$$HIND = 59,76 + 0,554 \cdot ALGH + 6,569 \cdot ARV - 40,47$$

Hind on 40,47 $ väiksem kui heas seisukorras iPod-il.

FIKTIIVSETE TUNNUSTE KORDAJAD

Fiktiivsete tunnuste kordajad näitavad sõltuva tunnuse muutust baasväärtusega võrreldes.

$$\text{HIND} = b + a_1 \cdot \text{ALGH} + a_2 \cdot \text{ARV} + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$$



NÄIDE: IPOD-IDE HINNAD EBAY OKSJONITEL, 4

$$\text{HIND} = 59,76 + 0,554 \cdot \text{ALGH} + 6,569 \cdot \text{ARV} - 16,11 D_1 - 40,47 D_2$$

Kui palju erineb halvas seisukorras oleva iPod-i hind keskmises seisukorras oleva iPod-i hinnast, kui muud tunnused on samad?

Halvas seisukorras: $D_1=0$, $D_2=1$
Keskmises seisukorras: $D_1=1$, $D_2=0$

$$-40,47 - (-16,11) = -24,36$$

Hind on 24,36 $ väiksem kui keskmises seisukorras iPod-il.

Fiktiivsete tunnuste kordajate omavaheline erinevus näitab sõltuva tunnuse muutust teineteisega võrreldes.

FIKTIIVSED TUNNUSED

Fiktiivne tunnus on kaheväärtuseline tunnus, mis võib omada väärtusi 0 või 1 ning mis vastab kvalitatiivse tunnuse kindlale tasemele.

Fiktiivsete tunnuste arv mudelis on **ühe võrra väiksem** kvalitatiivse tunnuse tasemete arvust.
Väärtus, mille fiktiivset tunnust mudelis pole, on **baasväärtus**.

Fiktiivsed on need tunnused seepärast, et vastavad ühe kvalitatiivse tunnuse erinevatele väärtustele.

MIKS VAJA FIKTIIVSEID TUNNUSEID?

Kui iPod-i seisukord X=1 (hea), 2 (keskmine), 3 (halb)

ja mudel

$$\text{HIND} = b + a_1 \cdot \text{ALGH} + a_2 \cdot \text{ARV} + a_3 X + \varepsilon$$

Siis

hea (X=1) ja keskmise (X=2) hinnavahe a_3

keskmise (X=2) ja halva (X=3) hinnavahe a_3

Ühesugune hinnavahe?
See pole õigustatud!

Aga mudelis

$$\text{HIND} = b + a_1 \cdot \text{ALGH} + a_2 \cdot \text{ARV} + \gamma_1 D_1 + \gamma_2 D_2 + \varepsilon$$

hea ja keskmise hinnavahe γ_1

keskmise ja halva hinnavahe $\gamma_1 - \gamma_2$

NÄIDE: EESTI NOORTE SÄÄSTMISHARJUMUSED, 1

Sihtgrupp 18-35 aastased noored, 1006 vastajat, aasta 2015. Küsimustikus oli küsimus „Kui suur on Teie hädareserv ootamatute kulude katteks (eurodes)?“

Seletavad tunnused:
ST viimase 6 kuu keskmine sissetulek (eurot);
SP mitu protsenti on keskmiselt säästetud viimase 6 kuu sissetulekust;
laste arv L;

haridus:
alg- või põhiharidus, baasväärtus
keskharidus, D1
keskeriharidus, D2
kõrgharidus. D3

	Coefficients	Standard Error	t Stat	P-value
Intercept	-383	137	-2,80	0,0052
Sissetulek ST	0,535	0,056	9,58	7,6E-21
Säästuprotsent SP	59,66	3,46	17,22	2,1E-58
Laste arv L	125,6	55,9	2,25	0,025
D1	245	144	1,70	0,089
D2	224	161	1,40	0,162
D3	632	145	4,37	1,3E-05

Kasutame olulisuse nivood 0,1.

Ei ole oluline

Allikas: Tiitso, R. „Eesti noorte säästmisharjumused“, TTÜ, Magistritöö 2015.

NÄIDE: EESTI NOORTE SÄÄSTMISHARJUMUSED, 2

	Coefficients	Standard Error	t Stat	P-value
Intercept	-383	137	-2,80	0,0052
Sissetulek ST	0,535	0,056	9,58	7,6E-21
Säästuprotsent SP	59,66	3,46	17,22	2,1E-58
Laste arv L	125,6	55,9	2,25	0,025
D1	245	144	1,70	0,089
D2	224	161	1,40	0,162
D3	632	145	4,37	1,3E-05

Kasutame olulisuse nivood 0,1.

Baasväärtus
alg- või põhiharidus
keskharidus
keskeriharidus
kõrgharidus

Ei ole oluline

Kui mõni fiktiivne tunnus on statistiliselt mitteoluline, siis seda eraldi mudelist eemaldada EI TOHI.

Mis juhtub, kui eemaldame D2?
Mudelisse jäävad D1 ja D3.

	D1	D3
Alg- ja põhiharidus	0	0
Keskharidus	1	0
Keskeriharidus	0	0
Kõrgharidus	0	1

Alg- ja põhiharidus ning keskeriharidus lähevad kokku.

Fiktiivne tunnus D2 ei ole tunnus, vaid kvalitatiivse tunnuse „haridus“ üks väärtus.

NÄIDE: MITU KVALITATIIVSET TUNNUST

Nutikellade hind, varasem mudel

$$\text{HIND} = 13,4 - 61,9\,\text{VANUS} + 220,5\,\text{EKDGN} + 3,24\,\text{KAAL} + \varepsilon$$

$$R^2 = 0,699, \quad R_a^2 = 0,663$$

Lisame 2 kvalitatiivset tunnust.

Andmeside olemasolu

$$\text{AS} = \begin{cases} 1, & \text{kui on andmeside} \\ 0, & \text{kui andmeside puudub} \end{cases}$$

Tootja

$$\text{APPLE} = \begin{cases} 1, & \text{kui tootja on Apple} \\ 0, & \text{kui tootja on Samsung} \end{cases}$$

$$\text{HIND} = 247 - 77,4\,\text{VANUS} + 4,43\,\text{KAAL} + 68,9\,\text{AS} + 112\,\text{APPLE} + \varepsilon$$

$$R^2 = 0,824, \quad R_a^2 = 0,795$$

Ekraani diagonaal EKDGN ei ole enam statistiliselt oluline, eemaldatud.

Andmeside olemasolu suurendab hinda 68,9 € võrra, kui muud tunnused on samad. Apple kellad on keskmiselt 112 € võrra Samsungi omadest kallimad, kui muud tunnused on samad.

MITU KVALITATIIVSET TUNNUST

Mudelisse võib panna mitu kvalitatiivset tunnust.
Sellisel juhul on igal tunnusel oma fiktiivsete tunnuste komplekt.

Näiteks isiku-uuringutel

Kvalitatiivne tunnus	Tasemeid	Fiktiivsete tunnuste arv	Fiktiivsete tunnuste komplekt mudelis
Sugu	2	1	S
Haridustase	3	2	H_1, H_2
Maakond	15	14	M_1, M_2, …,M_{14}
Tegevusala (EMTAK järgi)	22	21	T_1, T_2, …, T_{21}

Näide: palgaregressioon

LINEARISEERIMINE

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 1

Ettevõte	Varad K (mld $)	Töötajaid L	Käive q (mld $)
General Motors	228888	608000	178174
Ford	279097	363900	153627
Toyota	103894	159000	95137
Daimler Benz	76191	300000	71561
Daewoo	44861	265000	71526
Volkswagen AG	45417	280000	65328
Chrysler Corp.	60418	121000	61147
Nissan Motor Co	59121	137000	53478
Fiat S.p.a.	69028	239500	52269
Honda Motor Co	36110	109000	48876
Renault	37750	141315	35264
BMW	29629	118000	34692
Peugeot S.A	31472	140000	32004
Mitsubishi Motors	25276	27300	30429
…….	……	……	……

20 suurimat autode ja nende tarvikute tootjat, aprill 1999.

Cobbi-Douglase tootmisfunktsioon

$$q = AK^{\alpha}L^{\beta}$$

Vaja hinnata parameetrite A, α, β arvväärtused.

LINEARISEERIMINE

$$q = AK^{\alpha}L^{\beta}$$

$$\ln q = \ln(AK^{\alpha}L^{\beta})$$

$$\ln q = \ln A + \ln K^{\alpha} + \ln L^{\beta}$$

$$\ln q = \ln A + \alpha \ln K + \beta \ln L$$

y	b	x_1	x_2	uued tähistused

$$y = b + \alpha x_1 + \beta x_2$$ lineaarne mudel

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 2

Nende põhjal viime läbi regressioonmudeli hindamise

Ettevõte	Varad K (mld $)	Töötajaid L	Käive q (mld $)	ln K	ln L	ln q
General Motors	228888	608000	178174	12,341	13,318	12,091
Ford	279097	363900	153627	12,539	12,805	11,942
Toyota	103894	159000	95137	11,551	11,977	11,463
Daimler Benz	76191	300000	71561	11,241	12,612	11,178
Daewoo	44861	265000	71526	10,711	12,487	11,178
Volkswagen AG	45417	280000	65328	10,724	12,543	11,087
Chrysler Corp.	60418	121000	61147	11,009	11,704	11,021
Nissan Motor Co	59121	137000	53478	10,987	11,828	10,887
Fiat S.p.a.	69028	239500	52269	11,142	12,386	10,864
Honda Motor Co	36110	109000	48876	10,494	11,599	10,797
Renault	37750	141315	35264	10,539	11,859	10,471
BMW	29629	118000	34692	10,297	11,678	10,454
Peugeot S.A	31472	140000	32004	10,357	11,849	10,374
Mitsubishi Motors	25276	27300	30429	10,138	10,215	10,323
.....	.....	.....	.....	.....	.....	.....

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 3

$$q = AK^{\alpha}L^{\beta}$$

Lineariseeritud $\ln q = \ln A + \alpha \ln K + \beta \ln L$

Uutes tähistustes $y = b + \alpha x_1 + \beta x_2$

Lineaarse mudeli hinnangud

	Coefficients
Intercept	2,002
x1	0,674
x2	0,130

b = 2,002
α = 0,674
β = 0,130

A = **?**

$b = \ln A$ → $A = e^{b} = e^{2,002} \approx 7,4$

Autotööstuse tootmisfunktsioon $q = 7,4K^{0,67}L^{0,13}$

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 4

Ettevõte	Varad K (mld $)	Töötajaid L	Käive q (mld $)	Käibe mudelväärtus	Erinevus, %	Standardi-seeritud jääk
General Motors	228888	608000	178174	171192	4%	0,243
Ford	279097	363900	153627	183018	-16%	-1,062
Toyota	103894	159000	95137	84437	13%	0,724
Daimler Benz	76191	300000	71561	74419	-4%	-0,238
Daewoo	44861	265000	71526	51249	40%	2,023
Volkswagen AG	45417	280000	65328	52047	26%	1,379
Chrysler Corp.	60418	121000	61147	56555	8%	0,474
Nissan Motor Co	59121	137000	53478	56643	-6%	-0,349
Fiat S.p.a.	69028	239500	52269	67617	-23%	-1,562
Honda Motor Co	36110	109000	48876	39442	24%	1,301
Renault	37750	141315	35264	42037	-16%	-1,066
BMW	29629	118000	34692	34879	-1%	-0,033
Peugeot S.A	31472	140000	32004	37144	-14%	-0,904
Mitsubishi Motors	25276	27300	30429	25901	17%	0,978
.....	.....	.....	.....	.....	.....	-0,157

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 5

Võrdleme Daimler Benziga

Ettevõte	Varad *K* (mld $)	Töötajaid *L*	Käive *q* (mld $)	*Käibe mudel-väärtus*	*Erinevus ehk jääk*	*Suhteline erinevus*
Daimler Benz	76191	300000	71561	74419	-2858	-4%
Daewoo	44861	265000	71526	51249	20277	40%

Varad Daewool oluliselt väiksemad.

Töötajate arv ja käive ligikaudu samad.

Kas see, et Daewoo tegelik käive oluliselt suurem kui mudelväärtus, oli hea või halb?

NÄIDE: AUTOTÖÖSTUSE TOOTMISFUNKTSIOON, 6

Mudel oli hinnatud 1999. a. aprilli andmete põhjal.

ÄRIPÄEV 17.aug. 1999

Daewoo lammutatakse

Finantsraskustes Lõuna-Korea tööstusgrupi Daewoo kreeditorid otsustasid eile lammutada riigi suuruselt teise konglomeraadi ja jätta alles vaid kuus autotootmisüksust.

Praegu kuulub Daewoo gruppi 22 firmat. Tänavune restruktureerimisplaan näeb ette grupi võlgade ja omakapitali suhte kahandamise 1998. a 527 protsendilt 196 protsendile.

LINEARISEERIMINE

Paljusid mittelineaarseid mudeleid saab lineariseerida.

Paraboolne mudel

$$y = b + a_1 x + a_2 x^2 \quad \rightarrow \quad y = b + a_1 x + a_2 z, \quad z = x^2$$

Arvutatakse tunnus *z* ja mudelisse pannakse *x* ning *z.*

Eksponentsiaalne mudel

$$y = a e^{rx}$$

$$\ln y = \ln a + r \ln x$$

$$w = \ln y, \quad z = \ln x$$

Arvutatakse tunnused w ja z ja hinnatakse mudelit $w = \tilde{a} + r\, z$

Pärast arvutatakse esialgne kordaja $a = e^{\tilde{a}}$

REGRESSIOONMUDELITE KASUTAMINE

KÕIGE OLULISEM EESMÄRK

Regressioonmudeli argumenttunnuste valik annab informatsiooni, millised tunnused mõjutavad funktsioontunnust Y.

1. Teoreetiliste seisukohtade kontrollimine praktikas.
2. Uue info saamine võimalike seoste kohta ja sealt uued teoreetilised seisukohad.

Seejärel

- Suuruste vaheliste seoste uurimine:
  - mis suunas üks suurus teist mõjutab;
  - kui palju mõjutab;
  - kas mõju on lineaarne või mittelineaarne.
- Prognoosimine.
- Erindite kindlakstegemine.