

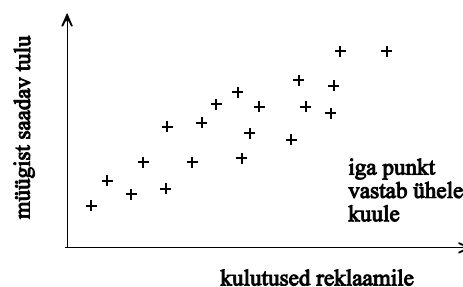
6. NÄHTUSTEVAHELISTE SEOSTE UURIMINE

Korrelatsiooni mõiste

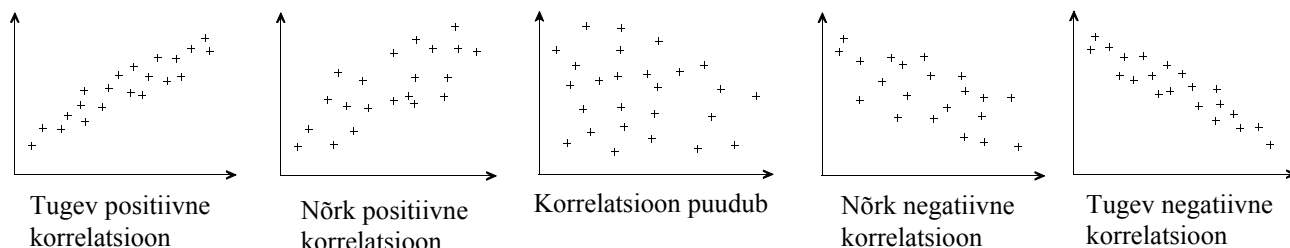
Olgu meil andmed reklaamikulude ja müügist saadud tulude kohta erinevatel kuudel. Nende andmete põhjal konstrueerime diagrammi (joonis 1).

Saadud diagrammi nimetatakse **hajumisdiagrammiks** (*scatter diagram*).

Diagrammilt on näha, et üldiselt reklaamikulude suurenemisel müügist saadav tulu suureneb. Öeldakse, et nende kahe suuruse vahel on olemas **korrelatsioon**.



Joonis 1



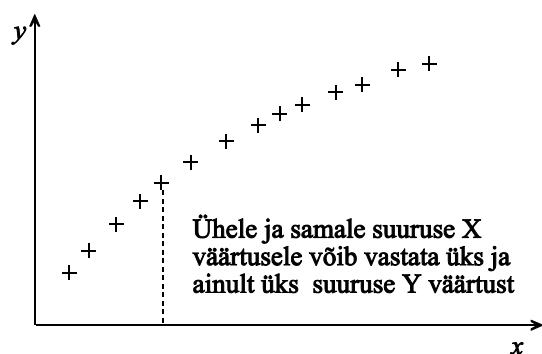
Positiivne korrelatsioon — ühe suuruse kasvades teine suurus samuti kasvab.

Negatiivne korrelatsioon — ühe suuruse kasvades teine suurus kahaneb.

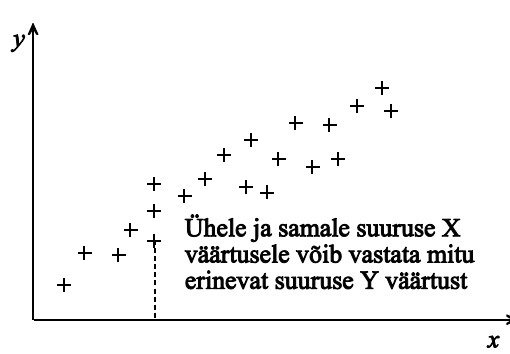
Seos kahe juhusliku suuruse vahel võib olla kaht tüüpi:

1. **Funktsionaalse seose** korral vastab argumendi x mingile väärtusele üks ja ainult üks funktsiooni y väärtus.

2. **Statistilise (korrelatiivse) seose** puhul võib ühe suuruse X mingile väärtusele vastata mitu teise suuruse Y väärtust, mida täpselt ei saa kindlaks määrata. Statistiline seos väljendub ühe juhusliku suuruse Y keskvaertuse sõltuvuses teise juhusliku suuruse X väärtustest.



Joonis 7 Funktsionaalne seos

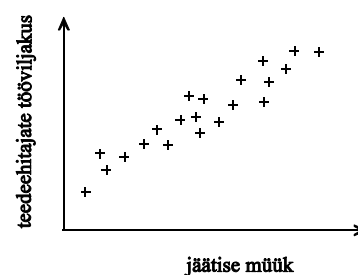


Joonis 8 Korrelatiivne seos

Seose olemasolu ei tähenda, et suurused on omavahel põhjuslikult seotud.

Põhjuslik seos ehk **deterministlik seos** – seos, mille korral üks nähtus on põhjus ja teine tagajärg. Põhjus avaldab mõju tagajärjele, põhjuslik seos on alati kindla suunaga.

NÄIDE 6.1. Kogudes andmeid jäätise müügi ja tee-ehitajate tööviljakuse kohta erinevatel suvepäevadel, võime avastada, et nende suuruste vahel on korrelatiivne seos (joon. 9). Kuid põhjuslik seos puudub. Ühe suuruse muutmine ei põhjusta teise suuruse muutumist.



Joonis 9. Korrelatiivne seos on, põhjuslik seos puudub.

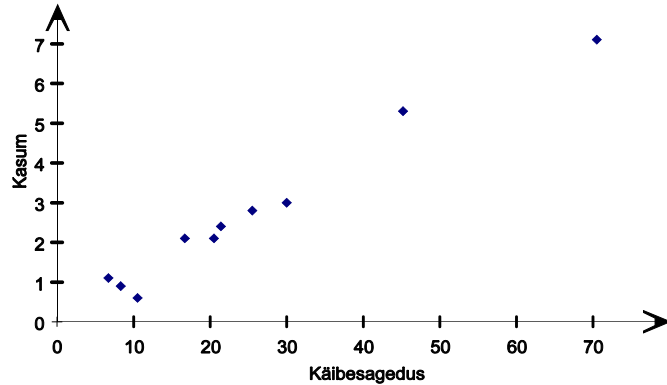
Üks ja sama põhjus võib esile kutsuda rohkem kui ühe tagajärje, mis võivad omavahel seotud olla, kuigi põhjuslik seos nende vahel puudub.

Kui korrelatiivne seos on tugev, vihjab see küll põhjusliku seose võimalusele, ent ei tõesta veel selle olemasolu. Seda saab tõestada vaid nähtuste **kvalitatiivse analüüsi** teel.

NÄIDE 6.2.

Tabelis on toodud kümne ettevõtte käibesagedus (näitab, mitu ringi teevad vahendid antud ettevõttes perioodi jooksul) ja kasum. Samad andmed on esitatud juuresoleval diagrammil (joonis 10).

Käibesagedus	Kasum
30,0	3,0
25,5	2,8
6,7	1,1
45,2	5,3
10,5	0,6
16,7	2,1
20,5	2,1
21,4	2,4
8,3	0,9
70,5	7,1



Joonis 10 Tegemist on tugeva positiivse korrelatsiooniga. Võib väita, et käibesageduse suurenedes on oodata ka suuremat kasumit.

Regressioonanalüüs

Kui kahe suuruse vahel on olemas seos, siis järgmiseks eesmärgiks on selle seose modelleerimine. Näiteks ökonomeetrias võib huvi pakkuda majapidamise sissetuleku suuruse ja erinevatele kaubagruppidele tehtavate kulutuste vaheline seos. Ettevõtjale pakub huvi seos müügist saadava tulu ja reklaamikulude vahel. Investeerijale pakub huvi seos väärtpaberituru indeksi ja SKT (sisemajanduse kogutoodangu) vahel.

Seost kirjeldava mudeli leidmiseks kasutatakse **regressioonanalüüsi**. Mudel võib sisaldada kas ühte või mitut argumentsuurust. Näiteks müügitulu võib sõltuda peale reklaamikulude veel paljudest muudest suurustest – hinnad, üldine majandusolukord, ajategur (aasta algus, keskpäev või -lõpp).

- ▶ üks argumentsuurus, $y = f(x)$ – lihtne regressioon;
- ▶ mitu argumentsuurust, $y = f(x_1; x_2; x_3; \dots)$ – mitmene regressioon.

Saadav mudel ei võimalda arvutada funktsioonsuuruse Y täpset väärtust. Kuigi mudelisse on võimalik lisada mitmeid suursi, pole täpse müügitulu väärtuse leidmine mudeli põhjal ikkagi võimalik.

Regressioonanalüüsi eesmärgiks on leida arvutusvalem, mis võimaldab argumentsuuruse X väärtuse põhjal välja arvutada funktsioonsuuruse Y vastavat väärtust \hat{y} . Suuruse Y täpne väärtus y_i on juhuslik suurus, mille võimalikud väärtused erinevad mudeli põhjal leitud väärtustest \hat{y}_i juhusliku

komponendi ε võrra :

$$y_i = \hat{y}_i + \varepsilon_i$$

Juhusliku komponendi ε tekitab mudelist väljajäetud suurustest põhjustatud variatsioon (seletamata variatsioon).

Lihtsaim regressioonmudel on lineaarne mudel, mille graafikuks on sirge parameetritega a ja b .

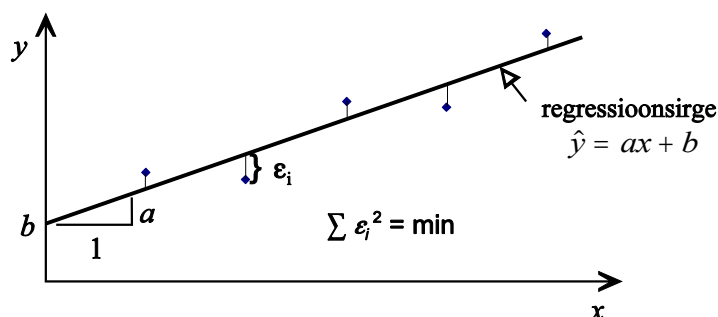
Lineaarse mudeli üldkuju

$$y = ax + b$$

Konkreetses mudeli leidmiseks tuleb empiiriliste andmete põhjal leida mudeli parameetrid a ja b . Parameetrite arvvaartuste leidmiseks kasutatakse **vähimruutude meetodit**. Tuleb leida sellise sirge võrrand, mille korral empiiriliste punktide ja sirge vastavate punktide vaheliste kauguste ruutude summa on minimaalne

$$\sum_i \varepsilon_i^2 = \min$$

s.t. otsitakse sellist mudelit, mille korral seletamata variatsioon on minimaalne (joonis 11).



Joonis 11 Vähimruutude meetod

Sellest tingimusest saadakse võrrandsüsteem, mille lahendamine annab mudelite parameetrite leidmiseks järgmised valemid:

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{n \sum_i x_i y_i - \sum_i x_i \sum_i y_i}{n \sum_i x_i^2 - \left(\sum_i x_i \right)^2}$$

$$b = \bar{y} - a\bar{x}$$

Siin x_i ja y_i on suuruste X ja Y arvvaartused, n väärtuste paaride arv ning \bar{x} ja \bar{y} aritmeetilised keskmised.

Koefitsiendi a leidmiseks kasutatakse *MS Excelis* funktsiooni **SLOPE** (ingl.k. tõus). Vabaliikme b leidmiseks kasutatakse funktsiooni **INTERCEPT** (ingl. k. vabaliige).

Kui mudeli parameetrid on leitud, saab mudelit kasutada funktsioon suuruse Y väärtuste prognoosimiseks suvalise argumentsuuruse X väärtuse korral:

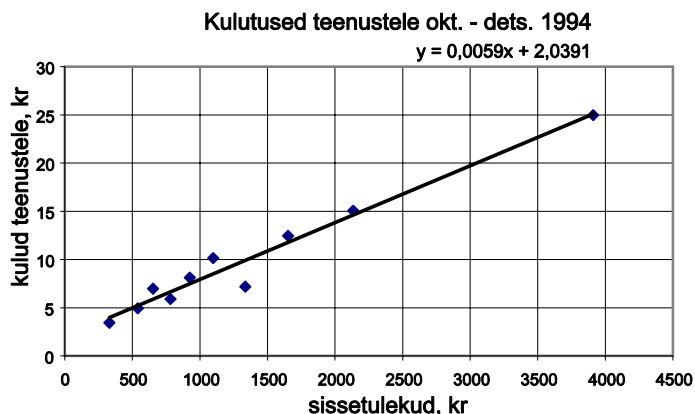
$$\hat{y}_i = ax_i + b$$

MS Excelis leiab lineaarse regressioonmudeli abil leitud väärtused \hat{y}_i funktsioon **TREND**.

NÄIDE 6.3. Tarbimismudelid Alates 1992. aastast tehakse Eestis statistikaameti tellimusel pere-eelarve uuringuid. Perevaatluste läbiviijaks on Eesti Turu- ja Arvamusteuringute Keskuse AS EMOR. Andmed saadakse peredest päevikumeetodil ja intervjuudest. Alates aprillist 1993 on püsivastajaskonna suurus 2028 peret. Uuringute andmed avaldatakse Eesti Statistikaameti poolt väljaantavates statistika aastaraamatutes. Neid andmeid on võimalik kasutada Eesti majanduse modelleerimisel, sealhulgas ka tarbimismudelite konstrueerimisel. Tabelis on toodud tulugruppide lõikes andmed Eesti perede tulude kohta ja kulutused teenindusele oktoobris-detsembris 1994 [Eesti Statistika aastaraamat 1995, lk.148-151]. Nii tulud kui kulud on antud ühe pereliikme kohta kuus.

Eesti perede tulud ja kulutused teenindusele ühe pereliikme kohta kuus okt.-dets. 1994

Tulugrupp	Tulud, kr	Kulud teenindusele, kr
1	330,65	3,44
2	540,78	4,95
3	653,87	6,98
4	781,01	5,9
5	925,91	8,12
6	1097,46	10,15
7	1335,47	7,19
8	1652,48	12,47
9	2133,96	15,07
10	3909,12	24,96



Joonis 12 Tarbimismudel

Püstitame ülesande uurida, kuidas perede kulutused teenindusele sõltuvad nende rahalistest sissetulekutest. Selleks konstrueerime andmete põhjal lineaarse tarbimismudeli [Paas, T. Sissejuhatus ökonomeetriasse, lk.173]. Mudelit otsime kujul $y = ax + b$,

kus on kasutatud järgmisi tähistusi: y - kulud teenindusele kroonides; x - tulud kroonides; a , b - mudeli parameetrid.

Tegemist on lineaarse regressioonülesandega ning mudeli parameetrite a ja b leidmiseks kasutame programmi *MS Excel* vastavaid funktsioone. Lahendus annab parameetrite väärtusteks järgmised arvud:

$a = 0,0059$ kr/kr; $b = 2,0391$ kr.

Seega perede kulutused teenindusele on modelleeritavad järgmise võrrandiga: $y = 0,0059x + 2,0391$.

Rahaliste sissetulekute suurenemisel 1 krooni võrra suurenevad kulutused teenindusele 0,0059kr ehk 0,59 senti. Sissetulekute suurenemisel 1000 kr võrra suurenevad kulutused teenindusele 5 kr 90 senti inimese kohta.

Analüüsisid samal moel kulutusi toidukaupadele, saame $z = 0,109x + 215$

Sissetulekute suurenemisel 1000 kr võrra suurenevad kulutused toidukaupadele 109 kr inimese kohta.

Seega 1994.a. lõpus kasvasid suuremate sissetulekutega peredes kulutused toidukaupadele oluliselt rohkem kui kulutused teenindusele. Millise järelduse võib sellest teha?

Regressioonanalüüs on üks **ökonomeetrias** kasutatavatest põhimeetoditest. Regressioonanalüüsi kasutamine võimaldab konstrueerida ökonomeetrilisi mudeleid, mis on aluseks majandusteoreetiliste hüpoteeside kontrollimisele ja majandusprotsesside võimaliku arengu prognoosimisele.

NÄIDE 6.4. Nõudlusfunktsioon.

Tabelis on toodud ühe kauba erinevad hinnad ja neile hindadele vastavad nõutavad kogused (toodud kõrvalolevas tabelis). Andmed on esitatud ka joonisel 13 toodud diagrammil. On vaja saada nõudlusfunktsiooni kujul hinna p sõltuvus kogusest q , kusjuures kasutame lineaarset mudelit:

$$p(q) = aq + b$$

I variant. Kasutame lineaarset regressiooni, kus võtame koguse argumentiks ja hinna funktsiooniks. Mudeliks saame (joonis 13)

$$p(q) = -0,2082q + 1192,7$$

II variant. Arvestame, et andmete kogumisel on hind ette antud, m ääratud, ja nõutav kogus on sellest sõltuv suurus, mis sisaldab juhuslikku komponenti. Seega regressioonmudelit otsime kujul, kus hind on argument ja kogus funktsioon

Hind p	Kogus q
1000	1010
975	1100
950	1100
925	1170
900	1470
875	1500
850	1650

$$q(p) = cp + d$$

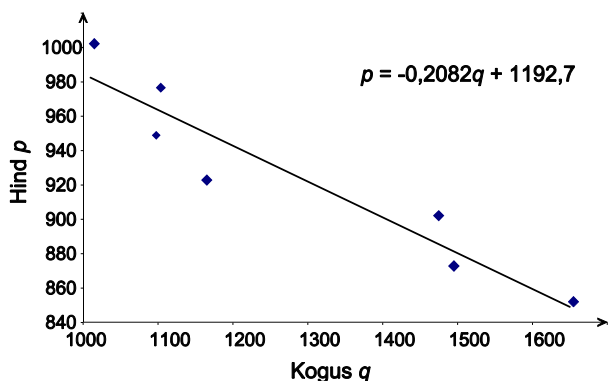
Regressioanalüüs annab mudeliks (joonis 14):

$$q(p) = -4,4143p + 5368,9$$

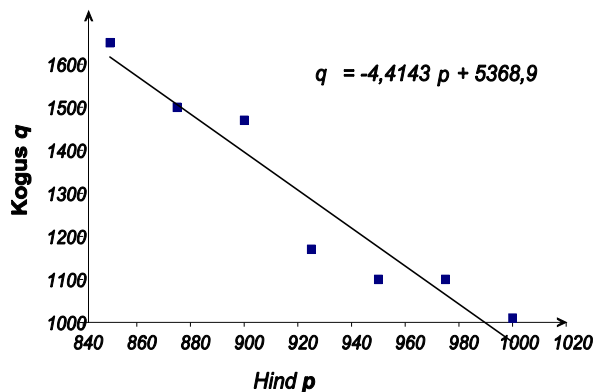
Saadud mudelist avaldame hinna:

$$p(q) = -0,2265q + 1216,3$$

Nagu näha, ei lange I ja II variandi korral leitud hinna avaldised kokku. Õige on II variant.



Joonis 13 Hinna sõltuvus kogusest.



Joonis 14 Koguse sõltuvus hinnast

Regressioanalüüs **ei ole pööratav**. Saadud avaldise kuju sõltub sellest, kas suurus y on suuruse x funktsioon või vastupidi, (joonis 15):

x - argumentsuurus, y - funktsioonsuurus: $y(x) = ax + b$;

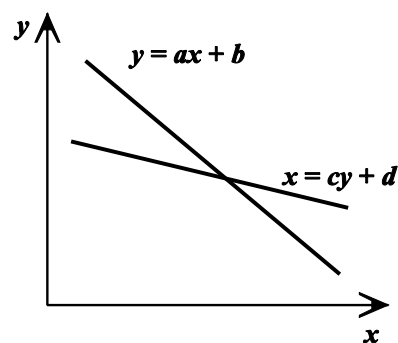
x - funktsioonsuurus, y - argumentsuurus: $x(y) = cy + d$.

Avaldame viimasest $y(x) = \frac{1}{c}x - \frac{d}{c}$.

Mida nõrgem on suuruste x ja y vaheline seos, seda suurem on suuruste a ja $\frac{1}{c}$ erinevus. Mida tugevam seos suuruste vahel

on, seda lähedasemad on funktsioonid teineteisele ja vastavad sirged lähenevad. Funktsionaalse seose korral sirged langevad

kokku, $a = \frac{1}{c}$.



Joonis 15 Regressioanalüüs pole pööratav

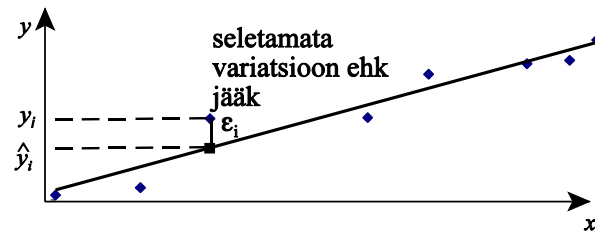
Seose tugevus. Korrelatsioonikordaja.

Olgu meil kahe suuruse X ja Y väärtuste paarid. Leiame suuruse, mis kirjeldab nende vahelise seose rangust.

Vaatleme nende suuruste variatsioonridasid $x_1; x_2; \dots; x_n$ ja $y_1; y_2; \dots; y_n$. Neid variatsioonridasid kirjeldatakse aritmeetiliste keskmiste \bar{x} , \bar{y} ja standardhälvete σ_x , σ_y abil. Mõlema variatsioonrea puhul võib varieerumist (väärtuse muutumist) vaadelda kui kahe varieerumise summana:

summaarne varieerumine = varieerumine, mis on tingitud kahe suuruse koosmuutumisest tänu nende vahelisele seosele + selgitamata varieerumine

X	Y
x_1	y_1
x_2	y_2
...	...
x_n	y_n



Joonis 16 Jäägid

Esimest komponenti, seosest tingitud varieerumist, võib kirjeldada regressioonanalüüsi abil saadud mudeliga ning see annab meile suuruse Y silutud väärtused

$$\hat{y}_i = ax_i + b$$

Vaatlusega saadud väärtused ehk empiirilised väärtused on siis modelleeritavad silutud väärtuse y_i ja juhusliku komponendi, mudeliga seletamata osa ehk **jäägi** ε_i summana:

$$y_i = \hat{y}_i + \varepsilon_i$$

Mida lähemal on vaatluspunktid regressioonjoonele, seda väiksemad on juhusliku komponendi väärtused, seda väiksem on hajuvus ja seda rangem on seos kahe suuruse vahel.

Summaarset varieerumist kirjeldab dispersioon, mille arvutamisel kasutatakse vaatluspunktide hälbeid aritmeetilisest keskmisest:

$$\sigma_T^2 = \frac{\sum (y_i - \bar{y})^2}{n}$$

Selgitamata varieerumist kirjeldab dispersioon, mille arvutusel kasutatakse vaatluspunktide hälbeid regressioonjoonest

$$\sigma_E^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n}$$

Nende kahe dispersiooni vahe $\sigma_T^2 - \sigma_E^2$ on **seletatud dispersioon**, mis on tingitud kahe suuruse vahelisest seosest, teise suuruse varieerumisest. Seose rangust iseloomustab seosega seletatud varieerumisest põhjustatud dispersiooni suhe summaarsesse dispersiooni ja vastavat suurust nimetatakse **determinatsioonikordajaks**:

$$r^2 = \frac{\sigma_T^2 - \sigma_E^2}{\sigma_T^2} = 1 - \frac{\sigma_E^2}{\sigma_T^2}$$

Näites 6.2 (lk. 54) on determinatsioonikordaja $r^2 = 0,9748$ ehk ca 97,5%. See tähendab, et 97,5% kasumi muutustest on seletatav käibesageduse muutumisega, ülejäänud varieerumine aga muudest põhjustest.

Determinatsioonikordaja näitab, kui suure osa summaarsest varieerumisest kirjeldab ära seosega seletatud varieerumine.

Sagedamini kasutatakse lineaarse seose ranguse iseloomustamiseks **korrelatsioonikordajat** r , mille absoluutväärtus on ruutjuur determinatsioonikordajast. Peale matemaatilisi teisendusi saadakse

korrelatsioonikordaja arvutamiseks järgmine valem.

Lineaarne korrelatsioonikordaja:

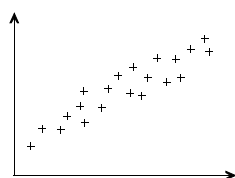
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n \sigma_X \sigma_Y}$$

Siin n on korreleeruvate suuruste X ja Y väärtuste x_i ja y_i paaride arv, \bar{x} ja \bar{y} aritmeetilised keskmised ning σ_X ja σ_Y vastavad standardhälbed.

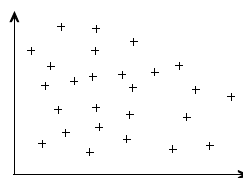
MS Excelis kasutatakse korrelatsioonikoeffitsiendi leidmiseks funktsiooni **CORREL** või vahendit *Correlation* komplektist *Data Analysis*.

Korrelatsioonikordaja on ühikuta suurus ja selle väärtus on -1 ja 1 vahel, $-1 < r < 1$:

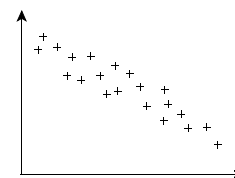
- $r = 0$ korrelatsioon puudub;
- $|r| = 1$ tegemist on täielikult korreleeruvate suurustega;
- $0 < r < 1$ positiivne korrelatsioon, ühe suuruse kasvades kasvab ka teine suurus;
- $-1 < r < 0$ negatiivne korrelatsioon, ühe suuruse kasvades teine suurus kahaneb.



Positiivne korrelatsioon
 $r > 0$



Korrelatsioon puudub
 $r = 0$



Negatiivne korrelatsioon
 $r < 0$

Kui determinatsioonikordaja iseloomustab vaid seose tugevust, siis korrelatsioonikordaja näitab ära ka seose suuna.

NÄIDE 6.5. Korrelatsioonanalüüs väärtpaberiturul
Korrelatsioonanalüüsi kasutamiseiga väärtpaberite turul võib tutvuda raamatus "Kunsing, S., Tuusis, D. Väärtpaberite portfelli-analüüs, TÜ Kirjastus". Leheküljel 62 on räägitud rahvusvahelisest diversifitseerimisest:

"USAs on aktsiate ja turuindeksite tulumäärade vahelised korrelatsioonikoeffitsiendid 50% ringis. Harudevahelised korrelatsioonikoeffitsiendid on umbes 60% ringis, kuid USA ja USA-väliste turgude korrelatsioonid on reeglina väga madalad. Ibbotsoni, Carri ja Robinsoni 1982. aastal korraldatud uurimuse tulemused näitasid, et aastatel 1960 - 1980 olid väga tugevad korrelatiivsed seosed vaid kolme Aasia riigi -- Jaapani, Hongkongi ja Singapuri -- aktsiaindeksite vahel ($r_{ij} = 90\%$), samuti mõnede Euroopa riikide indekse vahel."

Korrelatsioonikoeffitsiendid USA ja mõnede välisturgude aktsiaindeksite vahel

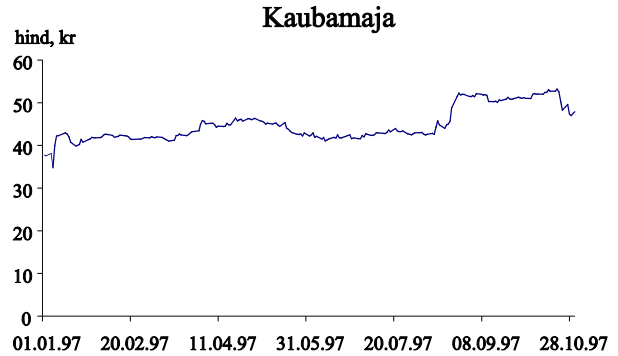
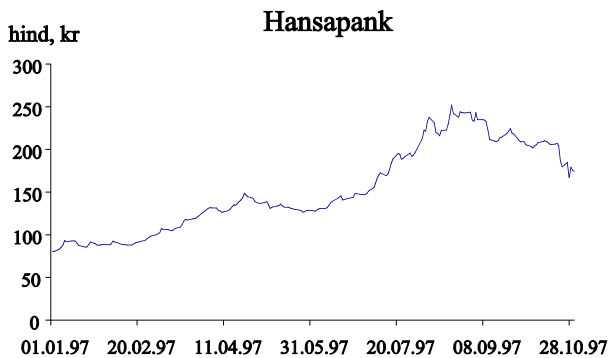
	Aktsiad	Võlakirjad
Kanada	0,71	0,63
Jaapan	0,22	0,10
Suurbritannia	0,62	0,08
Saksamaa LV	0,21	0,10
Euroopa (kokku)	0,63	-
Aasia (kokku)	0,26	-

Autokorrelatsiooni (*serial correlation*) kasutatakse mingi suuruse muutumise juhuslikkuse või mittejuhuslikkuse hindamisel. Kui suuruse X muutuste rida on $\Delta x_1; \Delta x_2; \dots; \Delta x_p; \dots; \Delta x_n$, siis autokorrelatsiooni koeffitsiendi leidmiseks moodustatakse $n - 1$ järjestikuste muutuste paari $(\Delta x_i; \Delta x_{i+1})$ ja leitakse vastav korrelatsioonikoeffitsient. Kui see on nullilähedane, on muutused juhuslikud. Kui aga korrelatsioonikoeffitsient on suur, siis muutus Δx_{i+1} on suures osas põhjustatud

eelmisest muutusest Δx_i , tegemist on autokorrelatsiooniga.

Näiteks, kui aktsia hinna muutused ei ole juhuslikud, siis hinnatõusule ühel päeval peaks järgnema hinnatõus ka järgmisel päeval. Efektive turu hüpotees väidab, et aktsiate hindade muutus on juhuslik.

NÄIDE 6.6. Hansapanga ja Kaubamaja aktsiate hindade muutmist ajavahemikul 1.1.97 - 31.10.97.



Leiame mõlema aktsia hinnamuutuse autokorrelatsiooni koefitsiendi. Selleks leiame hindade muutused (päeva keskmine hind - eelmise päeva keskmine hind) ja moodustame paarid (hinna muutus, eelmine hinnamuutus). Paaride arv on 212.

Hanspanga aktsia korral saadakse korrelatsioonikoefitsiendiks 0,18.

Kaubamaja aktsia korrelatsioonikoefitsiendiks tuleb 0,0049.

Seega Kaubamaja aktsia hinna muutmises on juhuslikkust rohkem.

Kuupäev	HP keskmine hind, kr	Hinna muutus, kr	Eelmine hinna muutus, kr
02.01.97	80,704115		
03.01.97	80,738212	0,0340965	
06.01.97	83,050807	2,312595	0,0340965
07.01.97	85,024523	1,973716	2,312595
08.01.97	87,7928	2,768277	1,973716
09.01.97	93,395178	5,6023785	2,768277
...

Kovariatsioon. Dispersioonide liitmine.

Lisaks determinatsiooni- ja korrelatsioonikordajale kasutatakse samadimensionaalsete suuruste (suurused, mida mõõdetakse samades ühikutes) korral seose ranguse ja suuna kirjeldamiseks **kovariatsiooni**:

$$cov_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Kovariatsiooni ühikuks on vastava suuruse ühik ruudus. Näiteks kui aktsiate hindasid mõõdetakse kroonides, siis kahe aktsia vahelise kovariatsiooni ühik on kr^2 . Kovariatsioon võib olla vahemikus

$$-\sigma_X \sigma_Y < cov_{XY} < \sigma_X \sigma_Y$$

Kovariatsiooni arvutusvalem on sarnane dispersiooni arvutusvalemile ja tihtilugu kasutatakse ka vastavat tähistust $\sigma_{XY}^2 = cov_{XY}$. Mingi suuruse dispersioon on tema kovariatsioon iseendaga,

$$\sigma_X^2 = cov_{XX}$$

Korrelatsioonikoefitsiendi ja kovariatsiooni vaheline seos:

$$r = \frac{cov_{XY}}{\sigma_X \sigma_Y}$$

MS Excelis kasutatakse kovariatsiooni leidmiseks funktsiooni **COVAR** või vahendit *Covariance* komplektist *Data Analysis*.

Kovariatsiooni kasutatakse kahest kogumist moodustatud summaarse kogumi dispersiooni leidmisel.

Kahe suuruse **summaarne dispersioon** on võrdne nende suuruste dispersioonide summaga, millele on liidetud kahekordne nende suurustevaheline kovariatsioon:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2 \text{cov}_{XY}$$

Näiteks tuleb summaarset dispersiooni kasutada väärtpaberiportfelli tulumäära standardhälbe (portfelliriski) arvutamisel. Kahest väärtpaberist koosneva portfelli tulumäära dispersioon:

$$\sigma_P^2 = w_A^2 \sigma_A^2 + (1 - w_A)^2 \sigma_B^2 + 2 w_A (1 - w_A) \text{cov}_{AB}$$

kus σ_A^2 ja σ_B^2 on vastavate väärtpaberite tulumäärade dispersioonid, w_A väärtpaberi A osakaal portfellis ja cov_{AB} tulumäärade vaheline kovariatsioon.

ÜLESANDED

6.1 Ajavahemikul 1.03.98-17.04.98 oli Hansapanga lihtaktsia keskmine tulumäär Tallinna väärtpaberibörsil -0,18% standardhלבega 1,81% ja AS Pennu Computers aktsia keskmine tulumäär -1,62% standardhלבega 4,31%. Nendevaheline kovariantsus oli 0,02%. Kui portfellis oli HP aktsiate osakaal 60% ja Pennu aktsiate osakaal 40%, kui suur oli portfelli tulumäär ja risk?

6.2 Näidata, et kui kahe võrdse standardhלבega väärtpaberi tulumäärad muutuvad vastupidistes suundades (korrelatsioonikoefitsient $r = -1$), on nendest võrdsetes osades moodustatud portfelli risk võrdne nulliga (absoluutselt riskivaba portfelli).

6.3 *) Aktsia A tulumäära standardhלבega on 1,80% ja aktsia B oma 2,4%. Aktsiate tulumäärade vaheline korrelatsioonikoefitsient on -0,1. Milline peaks olema kummagi aktsia osakaal portfellis, et portfelli risk oleks minimaalne? (Näpunäide: kasutada diferentsiaalarvutust funktsiooni miinimumkoha leidmiseks).

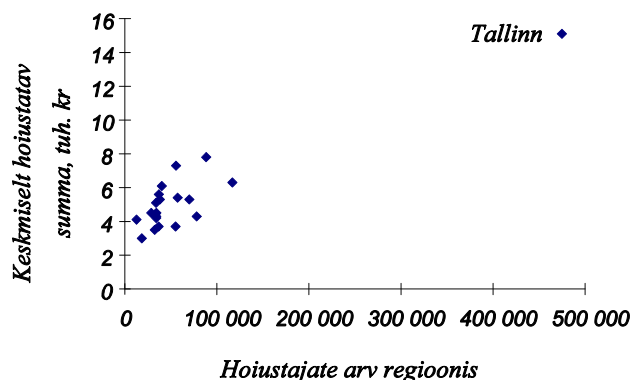
VASTUSED 6.1 Tulumäär -0,76%, risk 2,04% 6.3 A: 62,7%, B: 37,2%.

Lineaarse korrelatsioonikordaja puudused

Lineaarse korrelatsioonikordaja kasutamisel tuleb arvestada mõningaid nüansse. Alati ei pruugi lineaarse korrelatsioonikordaja suurus anda meile objektiivst informatsiooni.

NÄIDE 6.7.

Kasutades Eesti Pangast saadud statistilisi andmeid hoiuste kohta erinevates regioonides (30. nov. 1997), uuritakse seost hoiustajate arvu ja keskmise hoiuse suuruse vahel. Andmed on 21 regiooni (linnad ja maakonnad) kohta. Korrelatsioonikordajaks saadi 0,9208, mis näitab tugevat korrelatsiooni. Uurides aga hajumisdiagrammi, on näha, et korrelatsioonikordaja muudab suureks Tallinn. Jättes Tallinna andmed kõrvale ja leides korrelatsioonikordaja ülejäänud regioonide kohta, saadakse 0,5925. Seos hoiustajate arvu ja hoiuse suuruse vahel on oluliselt nõrgem.



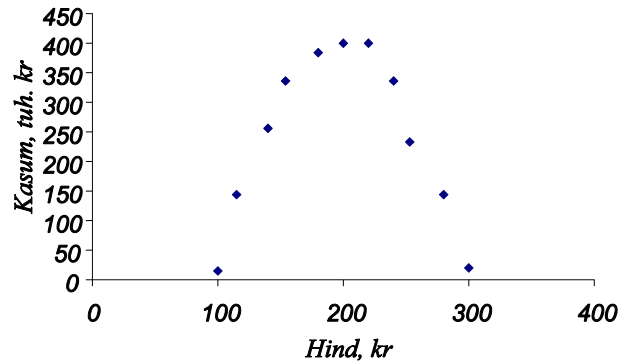
Joonis 22

Antud näites oli seose tugevus genereeritud ühe, teistest tugevasti erineva vaatluse poolt. See on **erind**. Erindi kõrvalejätmine oleneb olukorrast. Analüüsija peab otsustama, kas erind sobib uuritavasse kogumisse või mitte.

Lineaarne korrelatsioonikordaja on kergesti mõjutatav erindite poolt. Seetõttu tuleb lisaks korrelatsioonikordaja arvutamisele analüüsida alati ka hajuvusdiagrammi. Iga erindit tuleb analüüsida ja otsustada, kas see kuulub samasse kogumisse ülejäänud vaatlustega.

Lineaarse korrelatsioonikordaja väärtus võib olla eksitav ka teises suunas: tunnuste vahel võib olla väga tugev seos, kuid lineaarne korrelatsioonikordaja ei viita selle olemasolule.

NÄIDE 6.8. Ettevõttes on analüüsitud toote hinna ja saadud kasumi vahelist seost. Hajumisdiagrammilt on näha, et vaatluspunktid asuvad piki paraboolset kõverat. Ka teoreetilistest arvutustest on teada, et kui kulufunktsioon ja nõudlusfunktsioon on lineaarsed, avaldub kasumi sõltuvus hinnast ruutfunktsioonina. Seose tugevuse leidmiseks arvutatakse korrelatsiooni-koefitsient. Kuna tegemist on peaaegu funktsionaalse sõltuvusega, peaks korrelatsiooni-koefitsiendi absoluutväärtus olema lähedane ühele. Arvutused näitavad aga, et korrelatsioonikordaja on nullilähedane, $r = -0,0058$.



Joonis 23

Lineaarse korrelatsioonikordaja arvutusvalemi tuletamisel lähtuti sellest, et kahe tunnuse vaheline seos on modelleeritav lineaarse mudeliga. Seepärast lineaarne korrelatsioonikordaja "tunneb ära" punktide kogumi, mis on välja venitatud piki sirget. Kui punktide kogum asub piki mingit keerulisemat kõverat, või tekkida olukord, kus lineaarne korrelatsioonikordaja omandab nullilähedase väärtuse. Seda ka siis, kui tegemist on funktsionaalse seosega.

Mõlemat tüüpi eksitust on võimalik vältida hajuvusdiagrammide uurimisel.

Lineaarse korrelatsioonikordaja puuduste tõttu kasutatakse ka teisi seosekordajaid.

Järjestikskaalas mõõdetavate tunnuste vaheline seos.

Küsitluste puhul palutakse tihti reastada mingid suurused näiteks meeldivuse, olulisuse vms. järgi. Näiteks palutakse ostjatel reastada tegurid, mis mõjutavad kaubavalikut. Nendeks teguriteks võivad olla hind, tootjafirma, pakend, kättesaadavus jne. Sellisel juhul kasutatakse järjestikskaalat. Kui soovime analüüsida, kui hästi langevad hinnangud kokku erinevatel vastajate gruppidel, tuleb võrrelda erinevate tegurite järjenumbreid ehk **astakuid** erinevatel gruppidel. Selleks kasutatakse **astakkorrelatsiooni**.

Järjenumbrite korrelatsioonikordaja, mida nimetatakse **Spearmani korrelatsiooni-koefitsiendiks**, leitakse valemiga

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

kus d_i on erinevates gruppides kõrvuti olevate järjekorranumbrite vahe ja n väärtuspaaride arv.

NÄIDE 6.9.

Firmadel paluti tähtsuse järjekorras reastada tegurid, mis mõjutavad firma edukust. Küsitlus viidi läbi nii Ameerikas kui Euroopas. Tulemused on toodud järgnevas tabelis. Küsimus: Kui suur on seos Ameerika ja Euroopa firmade arvamusel firma edukust mõjutavate tegurite tähtsusele?

	Ameerika	Euroopa	d_i
Uuringud ja arendustegevus	4	2	2
Tootmisvahendite vanus	2	4	-2
Kulutused reklaamile	3	1	2
Müügistrateegiad	7	6	1
Toodete pakendamine	8	7	1
Töötasud	5	8	-3
Maa infrastruktuur	6	5	1
Majandushinnangud	1	3	-2

Seose tugevuse hindamiseks kasutame Spearmani korrelatsioonikoefitsienti, mille väärtuseks saame $r_s = 0,6667$.

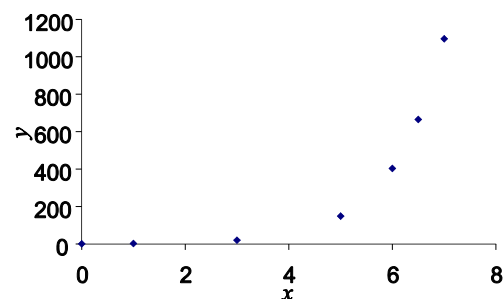
Astakkorrelatsiooni kasutatakse ka intervallskaalas mõõdetud kvantitatiivse tunnuste vahelise seose kirjeldamisel, kui on vaja vähendada erindite mõju. Selleks leitakse mõlema tunnuse korral vastava variandi järjenumbr (astak) ning leitakse korrelatsioonikoefitsient astakute vahel. Kui kahe või enama mõõtmistulemuse väärtused on aga võrdsed, määratakse võrdsetele väärtusetele sama astak, mis arvuliselt on võrdne vastavate astakute keskvärtusega.

Astaku leidmiseks *MS Excelis* kasutatakse funktsiooni **RANK**.

On võimalik näidata, et kui astakud on kõik erinevad, siis Spearmani korrelatsioonikoefitsiendi arvutusvalem langeb kokku lineaarse korrelatsioonikoefitsiendi valemiga. Sel juhul võib *MS Excelis* Spearmani koefitsiendi leidmiseks kasutada samuti funktsiooni **CORREL**.

NÄIDE 6.10. Leiame funktsiooni $y = e^x$ väärtused erinevate x väärtuste korral ja kanname saadud punktid graafikule (joonis 24). Leiame lineaarse korrelatsioonikoefitsiendi $r = 0,8196$. Kuigi on tegemist funktsionaalse sõltuvusega, pole see 1. Seejärel leiame Spearmani korrelatsioonikoefitsiendi $r_s = 1$.

x	y	x astak	y astak
0	1,00	1	1
1	2,72	2	2
3	20,09	3	3
5	148,41	4	4
6	403,43	5	5
6,5	665,14	6	6
7	1096,63	7	7



Joonis 24

- ▶ Lineaarne korrelatsioonikoefitsient mõõdab lineaarse seose tugevust.
- ▶ Spearmani korrelatsioonikoefitsient mõõdab monotoonse seose tugevust.

Sõltuvust nimetatakse **monotoonseks**, kui ühe tunnuse kasvamine toob kaasa teise tunnuse kasvamise ning ühe tunnuse kahanemine toob kaasa teise tunnuse kahanemise.

Regressioonsirge parameetrite usalduspiirid.

Regressioonimudel peegeldab tegelikkuses valitsevat seost. Kuid mudeli abil leitud väärtused, silutud väärtused, ei pruugi vaatlustulemustega täpselt kokku langeda, sest esinevad juhuslikud hälbed, jäägid.

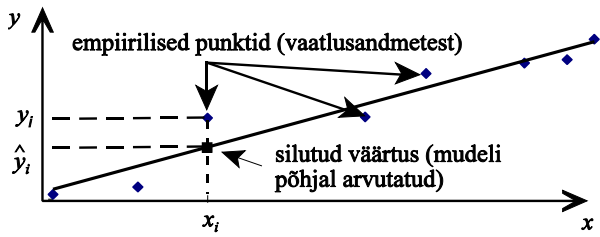
Kui regressioonimudel on

$$y = ax + b$$

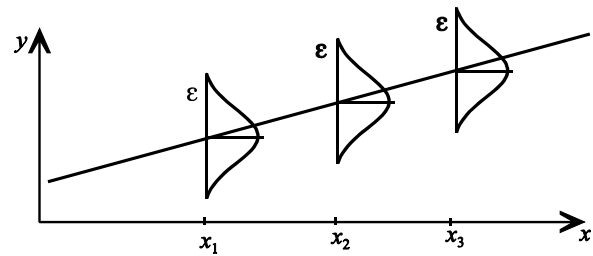
siis tegelik seos võib olla

$$y = \alpha x + \beta$$

kus α ja β on regressioonsirge parameetrite tõelised väärtused. Vähimruutude meetodiga määratud parameetrid a ja b on nende tegelike väärtuste hinnangud. Korrekse analüüsi korral tuleb leida nende hinnangute usalduspiirid.



Joonis 25 Kõrvalekalde mudelist



Joonis 26 Prognoosijääkide normaaljaotus

Mudeli parameetrite usalduspiiride leidmisel arvestatakse seda, et mida suurem on prognoosijääkide $\varepsilon = y - \hat{y}$

hajuvus, seda suurem on ka viga parameetrite a ja b määramisel.

Parameetrite usalduspiiride leidmisel eeldatakse, et prognoosijäägid ε on normaaljaotusega, mille keskvärtus on null ja standardhälve σ konstantne kõigi argumentsuuruse X väärtuste jaoks (joonis 26). Praktikas on jääkide jaotuse standardhälve teadmata ja selle hinnanguks kasutatakse vaatlusandmete kui valimi standardhälvet s . Jääkstandardhälve ehk lineaarse **regressioonmudeli standardviga**:

$$s_e = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}}$$

Standardviga iseloomustab funktsioontunnuse Y väärtuste kõrvalekallet regressioonvõrrandiga määratud väärtustest y_i . *MS Excelis* leiab selle suuruse funktsioon **STEYX**.

Kui jäägid alluvad normaaljaotusele, on regressioonsirge parameetrite hinnangud samuti normaaljaotusega. See võimaldab leida parameetrite hinnangute standardvead ja vastavad usalduspiirid.

Sirge tõusu a standardviga

$$s_a = \frac{s_e}{\sqrt{\sum x_i^2 - n(\bar{x})^2}}$$

Usaldusvahemiku laius parameetri a jaoks

$$\Delta a = t_{\beta, n-2} s_a$$

kus $t_{\beta, n-2}$ on Studenti koefitsient, β usaldatavus ja n empiiriliste punktide arv.

Vabaliikme b standardviga

$$s_b = s_e \sqrt{\frac{\sum x_i^2}{n \sum x_i^2 - (\sum x_i)^2}}$$

Usaldusvahemiku laius parameetri b jaoks

$$\Delta b = t_{\beta, n-2} s_b$$

Regressioonmudeli parameetrite usalduspiiride leidmiseks ja mudeli headuse kontrollimiseks kasutatakse *MS Excelis* vahendit *Regression* analüüsikomplektist *Data Analysis*.

NÄIDE 6.11. Kulude mudel. Vt. järgmine lk.

Prognoosi usalduspiirid

Üks regressioonanalüüsi enamlevinumaid kasutusalasid on prognoosimine. Tuleb arvestada sellega, et regressioonimudel kehtib vaid piirkonnas, mis on kaetud vaatlusandmetega. Ekstrapoleerimist sellest piirkonnast väljapoole võib lubada vaid vähesel määral ning see on seotud suure riskiga.

Lineaarfunktsiooni ekstrapoleerimise korral ekstrapoleeritud väärtus $y_p = ax_p + b$.

Selle usalduspiirid

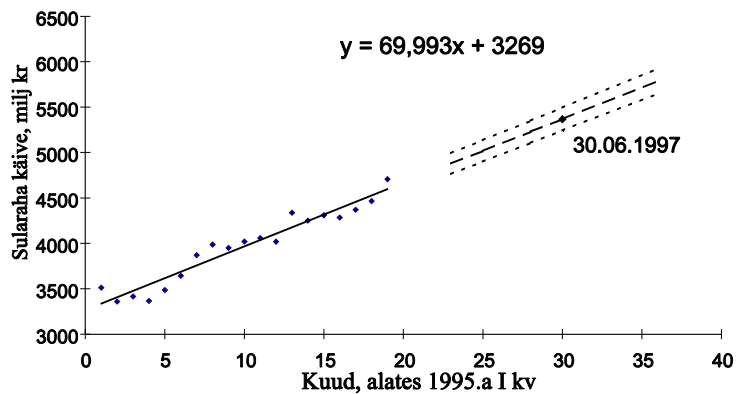
$$y_p \pm t_{\beta, n-2} s_p$$

kus s_p on prognoosi standardviga :

$$s_p = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum x_i^2 - n\bar{x}^2}}$$

Siin s_e on mudeli standardviga, x_i vaatlusandmed ja x_p väärtus, mille jaoks prognoos tehakse. Leides usalduspiirid paljude punktide jaoks, saame **veakoridori**.

NÄIDE 6.12. Kasutades empiirilisi andmeid 1995 kuni 1996. nov, leiti käibel oleva sularaha hulga muutumise mudel. Mudeli põhjal prognoositud väärtus 1997. a. juunis oli 5369 milj. kr. Usalduspiiride leidmine andis prognoosiks 5240 - 5500 milj. kr. Tegelik väärtus oli (EP andmetel) 5366,9 milj. kr



Mittelineaarne regressioon

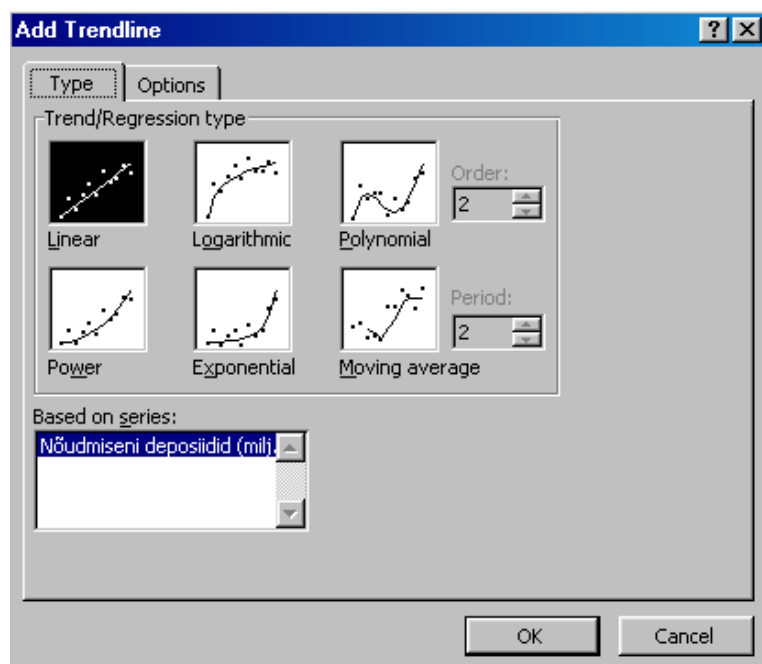
Mittelineaarse regressiooni puhul leitakse mudeli parameetrid sama loogika põhjal, mis lineaarse mudeli korral, kasutatakse vähimruutude meetodit.

Kõige sagedamini kasutatavad **regressioonmudelite tüübid**:

lineaarne	$y = ax + b$
parabool	$y = ax^2 + bx + c$
n -astme polünoom	$y = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x^1 + a_0$
eksponentsiaalne	$y = a e^{bx}$
logaritmiline	$y = a \ln x + b$
astmefunktsioon	$y = ax^b$

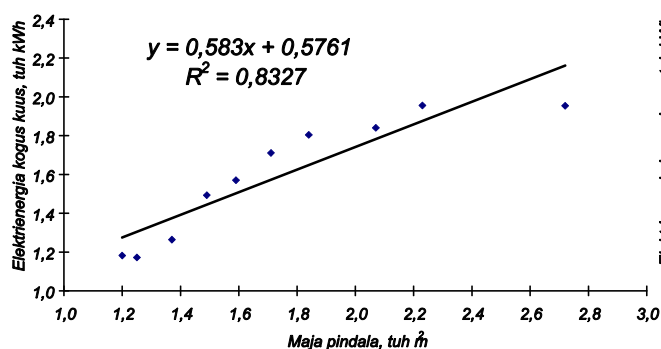
MS Excelis on lihtne otse diagrammile lisada mudel ja vastav lähendusjoon. Selleks tuleb empiiriliste andmete põhjal koostatud diagrammil märkida ära vastav punktikogum (Data Series) ning hiire parempoolse klahviga avatavast objektmenüüst valida *Add Trendline*.

Aknas *Add Trendline* lehel *Type* valida sobiv mudeli tüüp. Lehel *Options* märkida ära valikud *Display equation on Chart* (näita valemit diagrammil) ja *Display R-squared value on chart* (näita diagrammil R^2).

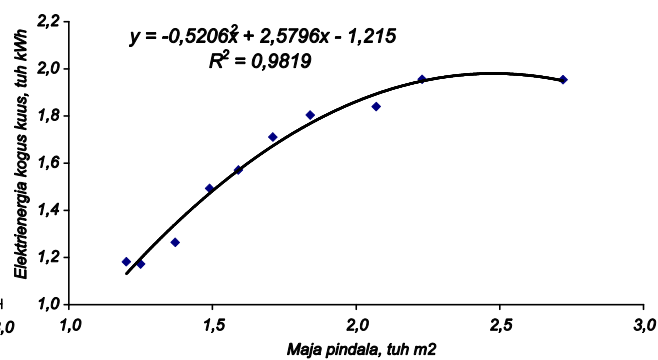


NÄIDE 6.13. Mittelineaarne regressioon

USA-s uuriti, kuidas pere elektrienergia tarbimine sõltub maja pindalast. Joonisel 30 on leitud vastav lineaarne mudel, millest selgub, et maja pindala suurenemisel 1000 m² võrra suureneb elektrienergia tarbimine kuus 583 kWh võrra. Kuid on näha, et lineaarse mudeli kasutamine pole siin õigustatud, parema tulemuse annab lähendamine parabooliga (joonis 31) (determinatsioonikoefitsient R^2 on suurem). Selgub, et pindala suurenedes kasvab elektrienergia tarbimine üha aeglasemalt ning maksimaalne on 2,5 tuh m² suurustes majades ning siis hakkab vähenema.



Joonis 30



Joonis 31