

*Aki Taanila*

# LINEAARISET REGRESSIOMALLIT

---

*17.6.2010*

# SISÄLLYSLUETTELO

0 Johdanto .....	1
1 Keskiarvo ennustemallina .....	2
2 Yhden selittävän muuttujan malli .....	3
3 Useamman selittävän muuttujan malli .....	6
4 Excel ja mallin laskeminen.....	7
5 SPSS ja mallin laskeminen.....	9
6 Mallin käyttäminen ennustamiseen ja selittämiseen .....	10
Ennustaminen .....	10
Selittäminen.....	10
7 Poikkeavat havainnot .....	11
8 Mallin tilastollinen merkitsevyys.....	11
Edeltävyys ehdot.....	11
Lineaarinen riippuvuus ja varianssien yhtä suuruus .....	12
Jäännösten normaalijakautuneisuus .....	14
Jäännösten riippumattomuus .....	15
Koko mallin merkitsevyys.....	15
Yksittäisten selittävien muuttujien merkitsevyys .....	16
Regressiokertoimien luottamusväli.....	17
Ennusteen luottamusväli .....	17
9 Selittävien muuttujien valitseminen.....	18
Selityskerroin ja korjattu selityskerroin .....	18
T-testin p-arvo .....	19
Kolinearisuus ja multikolinearisuus .....	19
10 Kategorisen muuttujan käyttö selittävänä muuttujana .....	21
11 SPSS:n valinta-algoritmit .....	23
Esimerkki valintaalgoritmin käytöstä .....	23
12 Harjoitus .....	26

## 0 Johdanto

Tässä monisteessa esitetään perusteet useamman selittävän muuttujan lineaarisista regressiomalleista. Lukijan oletetaan tuntevan entuudestaan ainakin seuraavat tilastolliset tunnusluvut: keskiarvo, keskihajonta, varianssi ja korrelaatio. Edelleen lukijan oletetaan hallitsevan tilastollisen päättelyn periaatteet, erityisesti p-arvon tulkinnan.

Laskentaan liittyviä matemaattisia yksityiskohtia esitetään vain kohdissa, joissa siitä arvellaan olevan ymmärtämisen kannalta hyötyä. Muilta osin laskennassa luotetaan tilasto-ohjelmien tai taulukkolaskennan antamiin valmiisiin tuloksiin. Usean selittävän muuttujan regressiomallien laatiminen on vaativa ja laaja aihepiiri. Tässä monisteessa aihepiiriä käsitellään pintapuolisesti, karttaen regressiomallien laskennassa olennaista matriisilaskentaa ja merkitsevyytestauksen taustalla vaikuttavia todennäköisyysjakaumia.

Joissain tapauksissa regressiomalli voidaan laskea taulukkolaskentaohjelmalla. Vakavammin otettavaan mallintamiseen tarvitaan tilasto-ohjelmaa (esim. SPSS). Tässä monisteessa annetaan tarpeelliset Excel ja SPSS ohjeet. Monisteessa viitatu aineistot löytyvät osoitteesta <http://myy.haaga-helia.fi/~taaak/m/>

Lineaarisen regressiomallin avulla voidaan mallintaa yhden tai useamman selittävän muuttujan ja yhden selitettävän muuttujan välistä riippuvuutta. Mallissa mukana olevien muuttujien täytyy olla määrällisiä. Selittävien muuttujien joukossa voi olla myös kategorisia muuttujia.

### Ennustava malli

Regressiomallia voidaan käyttää ennustavana mallina. Esimerkiksi kiinteistövälittäjä voi laatia mallin, jonka avulla voidaan ennustaa myyntiin tulevien kesämökkien hintoja. Selittävinä muuttujina voivat olla rantaviivan pituus, rakennusten pinta-ala ja tieto siitä onko mökki sähköistetty vai ei.

### Selittävä malli

Regressiomallia voidaan käyttää selittävänä mallina. Esimerkiksi asiakastyytyvyyden eri osa-alueiden yhteyttä kokonaistyytyvyyteen voidaan tutkia regressiomallin avulla. Hyvin laadittu malli paljastaa kokonaistyytyvyyteen merkittävästi vaikuttavat osa-alueet. Lisäksi voidaan arvioida eri osa-alueiden suhteellista tärkeyttä kokonaistyytyvyyden määräytymisessä.

### Otoskoko

Regressiomalleja laadittaessa otoskoon on oltava riittävä. Yhden selittävän muuttujan malleja voidaan menestyksellisesti laatia jo otoskoosta 20 alkaen. Useamman selittävän muuttujan mallien taakse suositellaan vähintään otoskokoa 50, mutta mielellään 100. Tämän monisteen esimerkeissä käytetään havainnollisuuden takia suosituksia pienempiä otoksia.

## 1 Keskiarvo ennustemallina

Tarkastellaan esimerkkinä aineistoa, joka sisältää eräällä alueella myytyjen kesämökkien myyntihintoja:

Myyntihintoja tuhansina euroina: 95, 95, 80, 100, 135, 100, 210, 160, 150, 150.

Myyntihintojen keskiarvoksi saadaan 127,5 (tuhatta euroa). Jos käytettävissä ei ole muuta tietoa kuin myyntihinnat, niin myyntihintoja voidaan mallintaa keskiarvon 127,5 avulla. Jos pitäisi ennustaa myytävänä olevan mökin myyntihinta, niin keskiarvoa voitaisiin käyttää ennusteena. Tarkastellaan, miten keskiarvo olisi toiminut ennusteena aineistossa olevien myyntihintojen kohdalla. Taulukkoon 1 on laskettu myyntihinnan ja ennusteen (keskiarvo) erotukset eli jäännökset ja jäännösten neliöt.

TAULUKKO 1. Kesämökkien myyntihinnat, keskiarvot ja jäännökset (n=10)

nro	myyntihinta (tuhatta euroa)	myyntihintojen keskiarvo	jäännös	jäännöksen neliö
1	95	127,50	-32,50	1056,25
2	95	127,50	-32,50	1056,25
3	80	127,50	-47,50	2256,25
4	100	127,50	-27,50	756,25
5	135	127,50	7,50	56,25
6	100	127,50	-27,50	756,25
7	210	127,50	82,50	6806,25
8	160	127,50	32,50	1056,25
9	150	127,50	22,50	506,25
10	150	127,50	22,50	506,25
Jäännösneliösumma				<b>14812,50</b>

Yleensä mallin antamien ennusteiden hyvyttä arvioidaan jäännösten neliöiden summan eli jäännösneliösumman avulla. Esimerkkinä olevan aineiston tapauksessa jäännösneliösummaksi saadaan 14812,50. Voidaan osoittaa, että mikään muu luku keskiarvon tilalla ei johda pienempään jäännösneliösummaan. Tässä mielessä keskiarvo on siis paras mahdollinen yksittäinen tunnusluku ennusteeksi.

Tässä yhteydessä on hyvä tarkastella jäännösneliösumman yhteyttä tuttuihin vaihtelua mittaaviin tunnuslukuihin varianssiin ja keskihajontaan. Myyntihinnat vaihtelevat keskiarvon molemmin puolin ja jäännösneliösumma kuvaa tätä vaihtelua. Vaihtelua keskiarvon molemmin puolin mitataan yleisemmin varianssin tai keskihajonnan avulla. Varianssi saadaan jakamalla jäännösneliösumma vapausasteiden lukumäärällä. Tässä vapausasteita on 9 (havaintoja on 10), koska yksi vapausaste on menetetty keskiarvon laskennassa.

Varianssi =  $14812,50/9 \approx 1645,83$  (keskihajonta saadaan ottamalla neliöjuuri varianssista).

Seuraavassa kehitämme mallin, johon otamme mukaan myyntihinnoissa esiintyvää vaihtelua selittäviä muuttujia. Tavoitteena on malli, joka selittäisi merkittävän osan myyntihintojen varianssista (vaihtelusta) ja tuottaisi pienemmän jäännösneliösumman.

## 2 Yhden selittävän muuttujan malli

Otetaan äskeiseen aineistoon mukaan rantaviivan pituus metreinä ja rakennusten pinta-ala neliömetreinä. Kyseiset muuttujat arvatenkin selittävät merkittävän osan myyntihintojen vaihtelusta.

TAULUKKO 2. Kesämökkien rantaviivat, pinta-alat ja myyntihinnat (n=10)

rantaviiva (m)	rakennusten pinta-ala m <sup>2</sup>	myyntihinta (tuhatta euroa)
30	50	95
35	42	95
40	25	80
50	30	100
55	45	135
60	24	100
60	60	210
70	34	160
80	32	150
85	28	150

Myyntihinnan riippuvuutta rantaviivan pituudesta ja rakennusten pinta-alasta voidaan laskeamalla korrelaatiokertoimet.

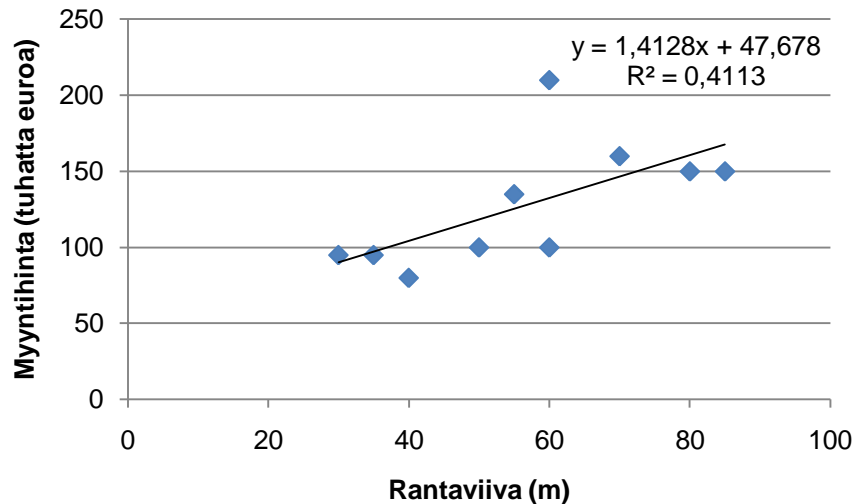
TAULUKKO 3. Rantaviivan ja pinta-alan korrelaatiot myyntihinnan kanssa

	Korrelaatio myyntihinnan kanssa
Rantaviiva	0,64
Pinta-ala	0,48

Myyntihinta korreloi voimakkaammin rantaviivan pituuden kanssa. Valitaan tällä perusteella rantaviiva selittäväksi muuttujaksi ja laaditaan lineaarinen regressiomalli rantaviivan pituuden ja myyntihinnan välille.

Riippuvuutta mallintavan regressiosuoran yhtälö lasketaan yleensä siten että jäännösneliösumma saadaan mahdollisimman pieneksi (pienimmän neliösumman menetelmä). Kaikkien suorien joukosta etsitään siis se, jonka kohdalla jäännösneliösumma saa pienimmän mahdollisen arvon. Tässä ei selitetä regressiomallin laskentaan liittyviä yksityiskohtia. Voimme luottaa siihen, että tietokoneohjelmat osaavat laskea pienimmän neliösumman regressiosuoran.

Seuraavaan hajontakuviioon on piirretty pienimmän neliösumman suora. Lisäksi kuviossa on näkyvillä regressiosuoran yhtälö ja mallin selityskerroin ( $R^2$ ). Kuvion perusteella lineaarisen regressiomallin käyttö on perusteltua, koska rantaviivan yhteys myyntihintaan vaikuttaa lineaariselta.



KUVIO 1. Myyntihinnan riippuvuus rantaviivan pituudesta

Kuviossa ylimpänä oleva havaintopiste on selvästi muista poikkeava. Muista poikkeavien havaintojen kohdalla kannattaa miettiä, onko havainto syytä pudottaa pois mallin laskennasta. Tässä asiaan ei kiinnitetä sen enempää huomiota ja poikkeava havainto pidetään mukana.

Regressiosuoraksi saadaan (kertoimia pyöristetty)  $y = 1,4128 \cdot x + 47,678$ , missä  $y$  on myyntihinta (tuhatta euroa) ja  $x$  on rantaviivan pituus (m).

Rantaviivan pituuden edessä oleva kerroin 1,4128 (suoran kulmakerroin) on tulkittavissa seuraavasti: 1 metri lisää rantaviivaa merkitsee 1,4128 yksikköä eli 1412,8 euroa lisää hintaa (regressiomallin mukaan).

Mallin avulla voidaan laskea, minkälaisia ennusteita malli olisi antanut aineistossamme oleville mökeille. Saamme ennusteet sijoittamalla malliin arvon  $x$  paikalle rantaviivan pituuksia.

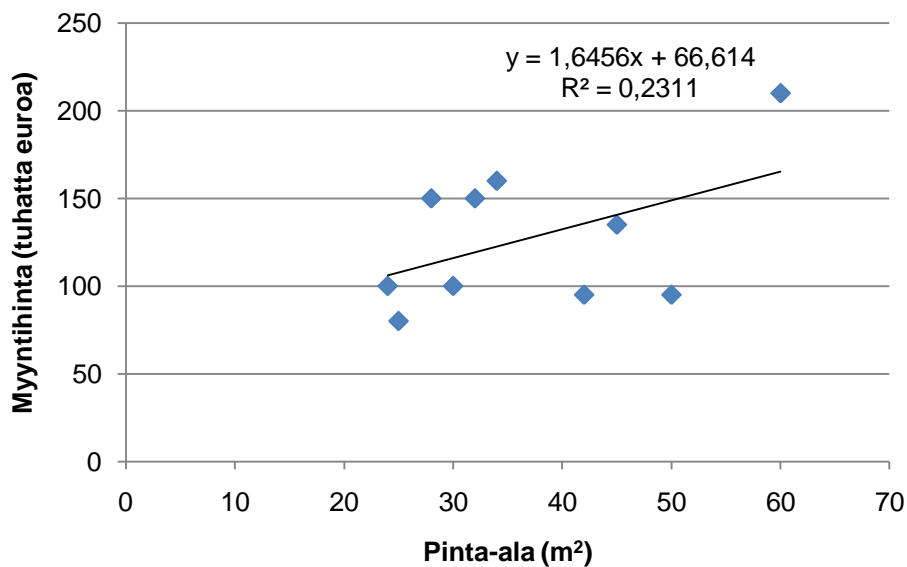
TAULUKKO 4. Rantaviiva selittävänä muuttujana

rantaviiva (m)	myyntihinta (tuhatta euroa)	mallin ennuste	jäännös	jäännös <sup>2</sup>
30	95	90,06	4,94	24,39
35	95	97,13	-2,13	4,52
40	80	104,19	-24,19	585,12
50	100	118,32	-18,32	335,51
55	135	125,38	9,62	92,53
60	100	132,44	-32,44	1052,66
60	210	132,44	77,56	6014,82
70	160	146,57	13,43	180,30
80	150	160,70	-10,70	114,50
85	150	167,76	-17,76	315,56
Jäännösneliösumma				<b>8719,90</b>

Esimerkiksi ensimmäiseen myyntihintaan liittyvä ennuste saadaan laskemalla:  
 $1,4128 \cdot 30 + 47,678 \approx 90,06$  (tuhatta euroa).

Taulukon 4 jäännösneliösumma on 8719,90. Aiemmassa mallissa, jossa ennusteena käytettiin keskiarvoa, jäännösneliösumma oli 14812,50. Näin ollen jäännösneliösumma on  $14812,50 - 8719,90 = 6092,60$  pienempi kuin keskiarvomallissa. Tämä merkitsee sitä, että osa hintojen vaihtelusta on pystytty selittämään rantaviivan pituuden avulla. Mallin selitystasetta kuvaava nk. selityskerroin saadaan jakamalla selitetty jäännösneliösumma kokonaisneliösummalla:  $6092,60 / 14812,50 \approx 0,411 \approx 41,1\%$ . Voimme todeta, että noin 41,1 % myyntihinnan vaihtelusta voidaan selittää rantaviivan pituuden vaihtelulla.

Jos selittäväksi muuttujaksi otetaan rantaviivan sijasta rakennusten pinta-ala, niin tilanne näyttää seuraavalta:



KUVIO 2. Myyntihinnan riippuvuus rakennusten pinta-alasta

Lineaarinen yhteys ei tässä tapauksessa ole yhtä selkeä kuin rantaviivan tapauksessa, mutta jonkinlaista lineaarista yhteyttä on havaittavista (tästä kertoo myös korrelaatiokerroin 0,48). Ylimmäinen havainto on jossain määrin muista poikkeava, mutta pidetään se kuitenkin mukana mallin laskennassa.

Malliksi saadaan (kertoimia pyöristetty)  $y = 1,6456 \cdot x + 66,614$ , missä  $y$  on myyntihinta (tuhatta euroa) ja  $x$  rakennusten pinta-ala (m<sup>2</sup>).

Pinta-alan edessä oleva kerroin 1,6456 (suoran kulmakerroin) on tulkittavissa seuraavasti: 1 m<sup>2</sup> lisää rakennusten pinta-alaa merkitsee 1,6456 yksikköä eli 1645,6 euroa lisää hintaa (regressiomallin mukaan).

Mallin avulla voidaan laskea, minkälaisia ennusteita malli olisi antanut aineistossamme oleville mökeille. Saamme ennusteet sijoittamalla malliin arvon  $x$  paikalle rakennusten pinta-aloja. Esimerkiksi ensimmäiseen myyntihintaan liittyvä ennuste:  $1,6456 \cdot 50 + 66,614 \approx 148,89$ .

TAULUKKO 5. Rakennusten pinta-ala selittävänä muuttujana

rakennusten pinta- ala m <sup>2</sup>	myyntihinta (tuhatta euroa)	ennuste	erotus	erotuksen neliö
50	95	148,89	-53,89	2904,39
42	95	135,73	-40,73	1658,76
25	80	107,75	-27,75	770,24
30	100	115,98	-15,98	255,39
45	135	140,66	-5,66	32,09
24	100	106,11	-6,11	37,30
60	210	165,35	44,65	1993,79
34	160	122,56	37,44	1401,51
32	150	119,27	30,73	944,20
28	150	112,69	37,31	1392,05
Jäännösneliösumma				<b>11389,72</b>

Jäännösneliösumma on 11389,72. Mallissa, jossa ennusteena käytettiin keskiarvoa, jäännösneliösumma oli 14812,50. Näin ollen jäännösneliösumma on  $14812,50 - 11389,72 = 3422,78$  pienempi kuin keskiarvomallissa. Mallin selityksastetta kuvaava selityskerroin saadaan jakamalla selitetty jäännösneliösumma kokonaisneliösummalla:  $3422,78/14812,50 \approx 0,231 \approx 23,1\%$ . Voimme todeta, että noin 23,1 % myyntihinnan vaihtelusta voidaan selittää rakennusten pinta-alan vaihtelulla. Vertailun vuoksi palautettakoon mieliin, että rantaviivan kohdalla selityskerroin oli 44,1 %.

Nyt olemme valmiit rakentamaan mallin, johon otetaan yhtä aikaa mukaan sekä rantaviivan pituus ja rakennusten pinta-ala. Arvatenkin näin saadaan selityskertoimen valossa parempi malli kuin mikään edellä käsitellyistä.

### 3 Useamman selittävän muuttujan malli

Regressiomalliin voidaan ottaa useita selittäviä muuttujia. Itse asiassa hyvässä mallissa pitäisi olla mukana kaikki olennaisesti tarkasteltavaan muuttujaan vaikuttavat selittävät muuttujat. Useamman selittävän muuttujan malli lasketaan siten että jäännösneliösumma saadaan mahdollisimman pieneksi. Tämä tarkoittaa sitä, että malli selittää mahdollisimman suuren osan selitettävän muuttujan vaihtelusta keskiarvonsa molemmin puolin. Jos myyntihintaa selittäviksi muuttujiksi otetaan sekä rantaviivan pituus että rakennusten pinta-ala, niin malliksi saadaan (kertoimia pyöristetty):

$$y = 1,9149 \cdot x_1 + 2,5545 \cdot x_2 - 75,210$$

Mallissa  $y$  on myyntihinta tuhansina euroina,  $x_1$  rantaviivan pituus metreinä ja  $x_2$  rakennusten pinta-ala neliömetreinä. Mallin kertoimet ovat tulkittavissa seuraavasti:

- 1 metri lisää rantaviivaa merkitsee 1,9149 yksikköä eli 1914,9 euroa lisää hintaa, jos oletetaan rakennusten pinta-alan pysyvän samana.
- 1 neliometri lisää rakennuksen pinta-alaa merkitsee 2,5545 yksikköä eli 2554,5 euroa lisää hintaa, jos oletetaan rantaviivan pituuden pysyvän samana.

Kannattaa panna merkille, että rantaviivaan liittyvä regressiokerroin ei ole samansuuruinen kuin aiemmassa mallissa, jossa selittävänä muuttujana oli ainoastaan rantaviiva. Mallin kertoimet siis vaihtelevat sen mukaan mitä muita muuttujia malliin otetaan mukaan. Tämä on seurausta siitä, että selittävät muuttujat korreloivat keskenään ja tätä kautta voivat selittää samaa vaihtelua. Selittävien muuttujien korrelaation takia regressiokertoimien tulkinta vaikeutuu, koska emme voi tietää missä suhteessa molempien muuttujan selittämä vaihtelu on jaettu muuttujien kertoimiin.

Taulukkoon 6 on laskettu mallin jäännösneliösumma, joka on huomattavasti pienempi kuin yhden selittäjän regressiomallissa. Kahden selittäjän malli selittää tässä tapauksessa myyntihintojen vaihtelua selvästi paremmin kuin kumpikaan tarkastelluista yhden muuttujan malleista.


TAULUKKO 6. Rantaviiva ja rakennusten pinta-ala selittävinä muuttujina

rantaviiva (m)	rakennusten pinta-ala (m <sup>2</sup> )	myyntihinta (tuhatta euroa)	Ennuste	jäännös	jäännös <sup>2</sup>
30	50	95	109,96	-14,96	223,93
35	42	95	99,10	-4,10	16,83
40	25	80	65,25	14,75	217,57
50	30	100	97,17	2,83	8,00
55	45	135	145,06	-10,06	101,28
60	24	100	100,99	-0,99	0,99
60	60	210	192,96	17,04	290,48
70	34	160	145,69	14,31	204,85
80	32	150	159,73	-9,73	94,62
85	28	150	159,08	-9,08	82,51
jäännösneliösumma					<b>1241,06</b>

Jäännösneliösumma on 1241,06. Jäännösneliösumma on  $14812,50 - 1241,06 = 13571,44$  pienempi kuin keskiarvomallissa. Mallin selitysastetta kuvaava selityskerroin on  $13571,44/14812,50 \approx 0,916 \approx 91,6\%$ . Voimme todeta, että noin 91,6 % myyntihinnan vaihtelusta voidaan selittää rantaviivan pituuden ja rakennusten pinta-alan vaihtelulla.

## 4 Excel ja mallin laskeminen

Excelissä voit laskea regressiomallin analyysityökalujen Regression-toiminnolla. Jos käytät analyysityökaluja ensimmäistä kertaa, niin:

1. Napsauta  **Office** -painiketta ja valitse alhaalta **Excel Options/Excelin asetukset** (Excel2010: **File-Options/Tiedosto/Asetukset**).
2. Valitse vasemmalta **Add Ins/Apuohjelmat** ja valitse sitten alhaalta **Manage/Hallinta** -ruudusta **Excel Add Ins/Excel-apuohjelmat**.
3. Valitse **Go/Siirry**.
4. Valitse luettelosta **Analysis ToolPak/Analyysityökalut** ja valitse **OK**. Jos Excel huomauttaa, että analyysityökaluja ei ole asennettu, niin valitse **Yes/Kyllä** asentaaksesi ne.

Kun olet ottanut käyttöön analyysityökalut, voit käyttää **Data/Tiedot**-välilehden **Analysis/Analyysi**-ryhmässä olevaa **Data Analysis/Tietojen analysointi** -komentoa.

Kun valitset analyysityökaluista työkalun **Regression/egressio**, niin pääset Regression-valintaikkunaan:

Regression-valintaikkunassa:

1. **Input Y Range:/Y-syöttöalue** -ruutuun viittaus selitettävän muuttujan arvoihin.
2. **Input X Range:/X-syöttöalue** -ruutuun viittaus selittäviin muuttujiin (muuttujien on hyvä olla vierekkäisissä sarakkeissa).
3. Valitse **Labels/Otsikot**, jos sisällytit otsikot syöttöalueiden viittauksiin.
4. Jos et halua tulosteita uuteen laskentataulukkoon, niin määritä tulostusalue.

Regression Statistics/Regressiotunnusluvut-taulukosta löytyy korrelaatiokerroin (R) ja selityskerroin (R Square). Excelin suomenkielisessä versiossa on virhe: selityskerrointa kutsutaan virheellisesti korrelaatiokertoimeksi. Seuraavassa tulosteessa korrelaatiokertoimen arvo on 0,957 ja selityskertoimen arvo on 0,916.

TAULUKKO 7. Excelin tulostamat regressiotunnusluvut

<i>Regressiotunnusluvut</i>	
Kerroin R	0,957
Korrelaatiokerroin	0,916
Tarkistettu korrelaatiokerroin	0,892
Keskivirhe	13,315
Havainnot	10

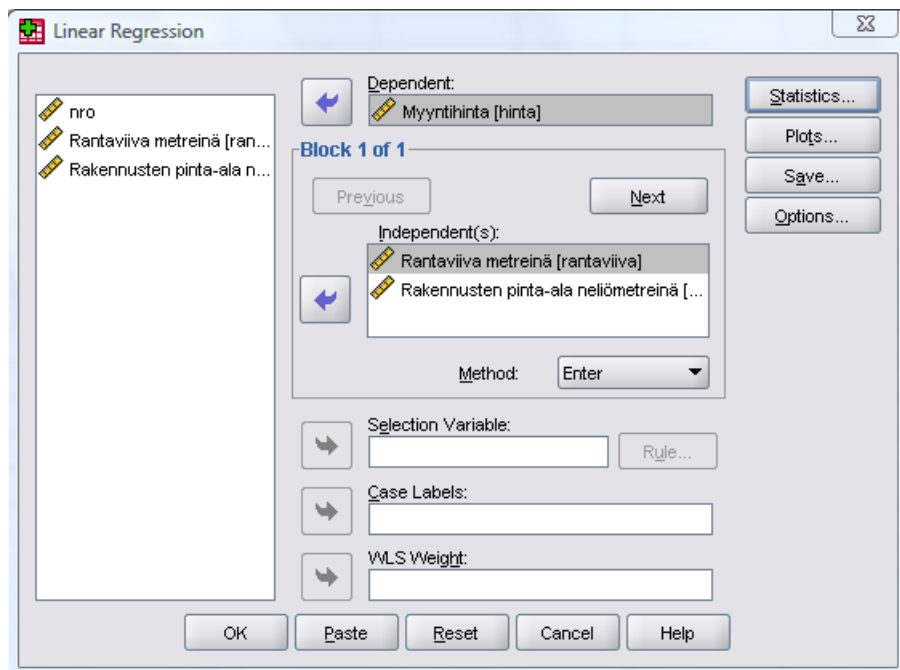
TAULUKKO 8. Excelin tulostamat kertoimet

	Kertoimet	Keskivirhe	t Tunnusluvut	P-arvo	Alin 95%	Ylin 95%
Leikkauspiste	-75,210	23,688	-3,175	0,016	-131,224	-19,196
rantaviiva (m)	1,915	0,253	7,566	0,000	1,316	2,513
rakennusten pinta-ala m2	2,555	0,393	6,495	0,000	1,624	3,485

Viimeisestä tulostaulukosta saadaan regressiomallin kertoimet. Esimerkkitulosteemme perusteella regressiomallin lauseke on  $y=1,915x_1+2,555x_2-75,210$ , missä  $y$  on myyntihinta tuhansina euroina,  $x_1$  on rantaviivan pituus metreinä ja  $x_2$  on rakennusten pinta-ala neliömetreinä.

## 5 SPSS ja mallin laskeminen

1. Valitse **Analyze – Regression – Linear...**
2. Siirrä selitettävä muuttuja **Dependent**-ruutuun. Siirrä selittävät muuttujat **Independent(s)**-ruutuun.
3. OK.



Model Summary –taulukosta löydät korrelaatiokertoimen (R) ja selityskertoimen (R Square).

TAULUKKO 9. SPSS:n Model Summary -taulukko

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,957 <sup>a</sup>	,916	,892	13,315

a. Predictors: (Constant), Rakennusten pinta-ala neliömetreinä, Rantaviiva metreinä

Regressiomallin kertoimet löytyvät Coefficients-taulukon B-sarakkeesta.

TAULUKKO 10. SPSS:n Coefficients-taulukko

		Coefficients <sup>a</sup>				
		Unstandardized Coefficients		Standardized Coefficients		
Model		B	Std. Error	Beta	t	Sig.
1	(Constant)	-75,210	23,688		-3,175	,016
	Rantaviiva metreinä	1,915	,253	,869	7,566	,000
	Rakennusten pinta-ala neliömetreinä	2,555	,393	,746	6,495	,000

a. Dependent Variable: Myyntihinta

## 6 Mallin käyttäminen ennustamiseen ja selittämiseen

### Ennustaminen

Mallin avulla voidaan laskea ennusteita. Ennustamisessa on huomioitava mallin käyttöalue. Mallin käyttöalue on kutakuinkin se alue, jolta oli havaintoja käytettävissä mallia laskettaessa. Jos mallia laskettaessa pisin rantaviiva oli 85 metriä, niin mallia ei ole syytä käyttää esim. 150 metrin rantaviivalle. Emme nimittäin tiedä jatkuuko rantaviivan ja myyntihinnan riippuvuus samanlaisena havaintoalueen ulkopuolella.

Esim. Jos mökin rantaviivan pituus on 70 metriä ja rakennusten pinta-ala 30 m<sup>2</sup>, niin mallin mukainen hintaennuste on  $y = 1,9149 \cdot 70 + 2,5545 \cdot 30 - 75,210 \approx 135,5 \approx 135\,500$  euroa.

Ennusteiden hyvydestä saa jonkinlaisen arvion selityskertoimesta. Mitä isompi selityskerroin, sitä paremmin malli kykenee ennustamaan.

### Selittäminen

Se, että selittävät muuttujat istuvat hyvin malliin vahvistaa tietenkin käsitystä siitä, että kyseiset muuttujat selittävät selitettävää muuttujaa. Selittävien muuttujien painoarvoa on kuitenkin vaikea nähdä regressiokertoimista, jos muuttujien arvot ovat eri suuruusluokkaa. Tällöin kannattaa hyödyntää normitettuja (standardoituja) regressiokertoimia eli nk. beta-kertoimia. SPSS laskee beta-kertoimet Coefficientitaulukon Beta-sarakkeeseen.

Esimerkissämme rantaviivan beta-kerroin on 0,869 ja rakennusten pinta-alan beta-kerroin 0,746. Tämän perusteella voimme sanoa, että rantaviivan painoarvo myyntihintaa selitettäessä on suurempi kuin pinta-alan painoarvo. Tarkempi tulkinta beta-arvoille on seuraava:

- Jos rantaviiva kasvaa yhden keskihajonnan verran (kyse on rantaviivan keskihajonnasta), niin myyntihinta kasvaa 0,869 tuhatta euroa olettaen, että rakennusten pinta-ala ei samalla muutu.
- Jos rakennusten pinta-ala kasvaa yhden keskihajonnan verran (kyse on rakennusten pinta-alan keskihajonnasta), niin myyntihinta kasvaa 0,746 tuhatta euroa olettaen, että rantaviiva ei samalla muutu.

Beta-kertoimia kuten regressiokertoimiakin tulkittaessa on myös syytä pitää mielessä, että kertoimen arvo riippuu siitä mitä muita muuttujia malliin on otettu mukaan. Erityisesti, jos selittävien muuttujien välillä on voimakkaita korrelaatioita, niin sekä regressiokertoimien ja beta-kertoimien tulkintaan on suhtauduttava varoen.

## 7 Poikkeavat havainnot

Selvästi muista poikkeavat havainnot vaikuttavat mallin kertoimiin. Tämän takia poikkeaviin havaintoihin täytyy suhtautua kriittisesti:

- ovatko poikkeavat havainnot virheellisiä tietoja
- ovatko poikkeavat havainnot väärin syötettyjä tietoja
- jos kyse ei ole virheestä, niin löytyykö poikkeaville arvoille luonnollinen selitys?

Kun poikkeavien havaintojen alkuperä selviää, niin seuraavaksi mietitään onko hyvä pitää havainnot mukana vai olisiko perusteltua jättää ne pois tarkasteluista:

- jos virheelliset tai väärin syötetyt tiedot voidaan korjata, niin ne voidaan pitää tarkasteluissa mukana
- jos virheellisille tai väärin syötetyille tiedoille ei syystä tai toisesta saada korjattuja arvoja, niin ne on syytä pudottaa pois tarkasteluista
- jos poikkeavuudelle löytyy luonnollinen selitys, niin asiaa on ajateltava tarkasteltavan ilmiön kannalta.

Poikkeavat havainnot voidaan tunnistaa laatimalla kullekin selittävälle muuttujalle hajontakuviota, jossa selittävä muuttuja on x-akselilla ja selitettävä muuttuja y-akselilla. Mallin laskemisen jälkeen kannattaa vielä silmäillä jäännöskuviota (selitetään seuraavassa luvussa) mahdollisten poikkeavien havaintojen tunnistamiseksi.

## 8 Mallin tilastollinen merkitsevyys

Malli selittää selitettävän muuttujan vaihtelua sitä paremmin mitä korkeampi selityskerroin. Pelkkä selityskerroin ei kuitenkaan takaa mallin käyttökelpoisuutta. Käytännön sovelluksissa on tärkeintä, että malli toimii käytännössä ja sen takana ovat riittävät käytännön tilanteen tuntemuksesta ja/tai teoriasta johdettavat perustelut. Erityisesti selittävien muuttujien valinta täytyy olla hyvin perusteltu. Varsinkin jos malliin mukaan otettavien selittävien muuttujien valintaan liittyy epävarmuutta tai jos halutaan varmistua mallin tilastollisesta merkitsevyydestä, niin seuraavassa esitettävät tarkastelut ovat tarpeellisia.

### Edeltävyys ehdot

Mallin edeltävyys ehtojen voimassaolo takaa merkitsevyydestäuksen pätevyyden. Lineaaristen regressiomallien edeltävyys ehdot ovat:

- Selittävien muuttujien ja selitettävän muuttujan välillä on lineaarinen riippuvuus.
- Jäännösten varianssi on yhtä suuri kaikilla selittävien muuttujien arvoilla.
- Jäännökset noudattavat normaalijakaumaa.
- Jäännökset ovat toisistaan riippumattomia.

Seuraavassa tarkastellaan joitain helppokäyttöisiä tapoja edeltävyys ehtojen voimassaolon tutkimiseen.

## Lineaarinen riippuvuus ja varianssien yhtä suuruus

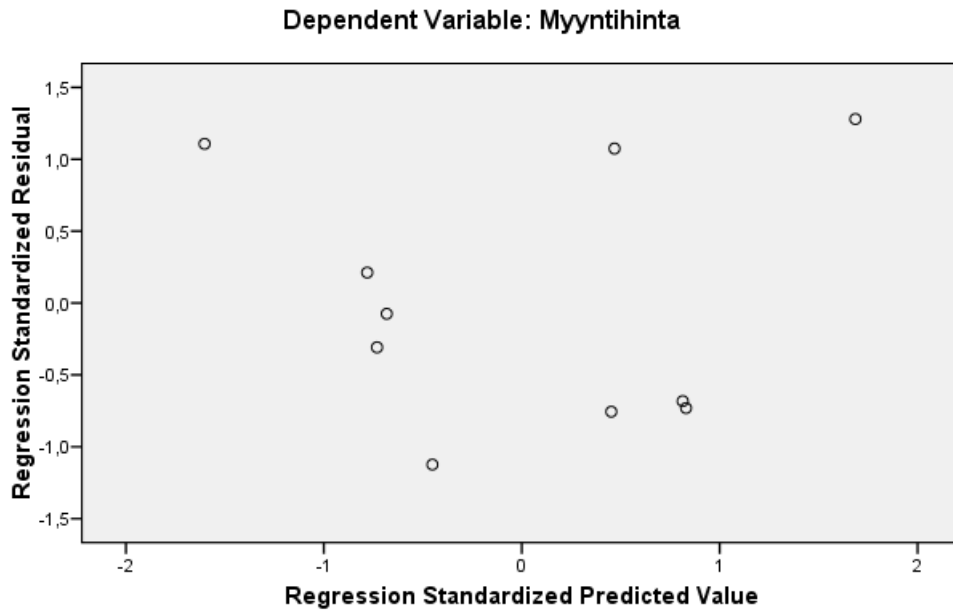
Lineaarisuutta ja jäännösten varianssien yhtä suuruutta kannattaa tarkastella jäännöskuvion avulla. Parhaiten tähän sopii hajontakuviot, jossa x-akselilla on mallin ennustamat arvot ja y-akselilla vastaavat jäännökset. Jäännösten jakauma näyttää kuviossa kutakuinkin samalta käytettiinpä jäännöksiä sellaisenaan tai normitettuna. On kuitenkin suositeltavampaa käyttää normitettuja (standardoituja) jäännöksiä. Normitettujen jäännösten keskiarvo on 0 ja keskihajonta 1. SPSS tarjoaa mahdollisuuden esittää myös ennusteet normitettuina.

Excelissä Regression-valintaikkunasta voidaan valita laskettavaksi **Residuals/Jäännökset** ja **Standardized Residuals/Normalisoidut jäännökset**.

Excel tulostaa jäännökset yhdessä myyntihintojen kanssa siistiin taulukkoon, jonka pohjalta on helppo laatia hajontakuviot.

SPSS:ssä Linear Regression –valintaikkunassa napsautetaan **Plots**-painiketta.

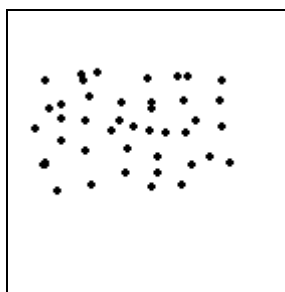
Plots-valintaikkunassa siirretään **ZPRED** (normitettu ennuste) x-akselin muuttujaksi ja **ZRESID** (normitettu jäännös) y-akselin muuttujaksi. Seuraava kuvio on SPSS:n tulostama.



KUVIO 3. Jäännöskuvio

Esimerkkinä käyttämämme havaintoaineisto on pieni. Tämän vuoksi jäännöskuvion perusteella ei voi sanoa paljonkaan. Jäännöskuviosta ei kuitenkaan ole nähtävissä mitään selkeitä lineaarisuuden tai varianssien yhtä suuruuden rikkomuksia.. Pisteet ovat kutakuinkin satunnaisesti jakaantuneet jäännöskuvioon.

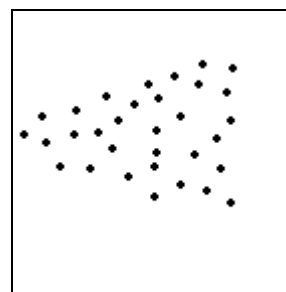
Seuraavassa on joitain esimerkkejä jäännöskuvioiden tulkinnasta.



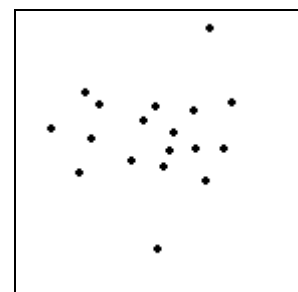
Siisti jäännöskuvio



Yhteys ei ole lineaarinen



Jäännösten varianssien yhtä suuruus ei täyty



Kaksi poikkeavaa havaintoa

KUVIO 4. Esimerkkejä jäännöskuvioista

Siistissä jäännöskuviossa pisteet ovat satunnaisesti jakaantuneet ilman mitään säännönmukaisuutta.

Jos riippuvuus ei ole lineaarinen, niin se yleensä näkyy pisteiden jonkin asteisena säännönmukaisuutena. Kuvion 4 toisessa kuviossa on nähtävissä selvä säännönmukaisuus, joten yhteys ei ole lineaarinen. Jäännöstermien tarkastelua ei pidä sekoittaa hajontakuviioon, jossa x-akselilla on selittävän muuttujan arvot ja y-akselilla selitettävän muuttujan arvot (kyseisessä kuviossa kuvion 4 toisen kuvion kaltainen säännönmukaisuus ilmentäisi nimenomaan lineaarista riippuvuutta).

Jos pistejoukko on kiilamaisesti ryhmittynyt, niin se viittaa varianssien yhtä suuruus ehdon rikkomuksiin. Kuvion 4 kolmannessa kuviossa jäännöstermien varianssi kasvaa selvästi x-akselilla olevien ennusteiden kasvaessa.

Jäännöskuviota kannattaa silmäillä myös mahdollisten poikkeavien havaintojen varalta. Yksittäiset poikkeavat pisteet viittaavat poikkeaviin havaintoihin, joiden kohdalla on mietittävä mahdollista poisjättämistä koko mallista. Kuvion 4 viimeisessä kuviossa ylimmäinen ja alimmainen havainto ovat selvästi muista poikkeavia.

Kannattaa tarkastella myös erikseen kunkin selittävän muuttujan kohdalla hajontakuviota, jossa x-akselilla on selittävän muuttujan arvot ja y-akselilla selitettävän muuttujan arvot tai jäännökset. Näistä kuvioista paljastuu mihin selittävään muuttujaan mahdolliset edeltävyysehtorikkomukset tai poikkeavat havainnot liittyvät. Esimerkkiimme liittyvät kyseiset hajontakuviot on jo esitetty luvussa 2.

### Jos ehdot eivät täyty

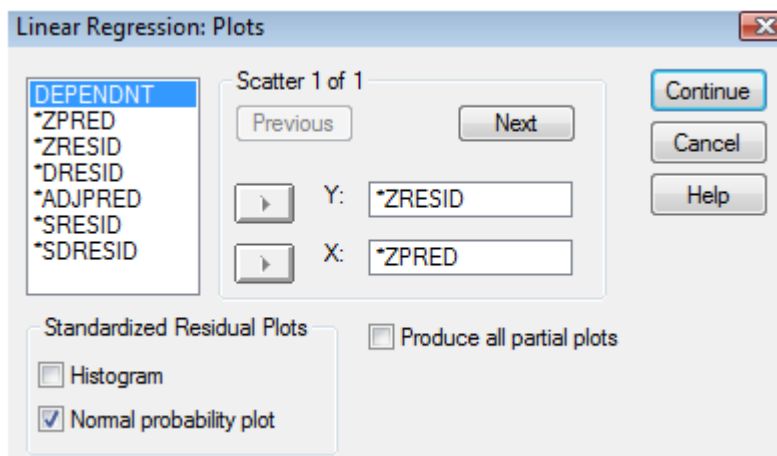
Jos lineaarisuutta rikotaan, niin on mahdollista lisätä malliin ylimääräisiä muuttujia, esimerkiksi muuttujan neliö (ei käsitellä tässä monisteessa).

Jos jäännösten varianssien yhtä suuruutta rikotaan, niin sopivilla muuttujiin kohdistettavilla muunnoksilla tilannetta voidaan korjata. Tässä monisteessa ei käsitellä muuttuja muunnoksia.

Malli voi käytännössä toimia vaikka jäännösten varianssien yhtä suuruutta jossain määrin rikotaankin. Merkitysevyydestaukset ja ennusteille laskettavat luottamusvälit eivät kuitenkaan tällöin ole päteviä.

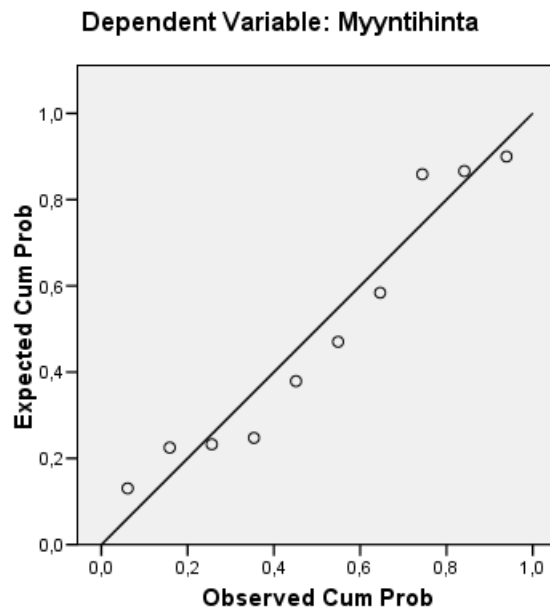
## Jäännösten normaalijakautuneisuus

Jäännöstermien normaalijakautuneisuuden tarkasteluun sopii parhaiten normaalijakaumakuviot. Excel ei tarjoa valmistoimintoa kunnollisen normaalijakaumakuviot piirtämiseen. SPSS:ssä Linear Regression –valintaikkunassa napsautetaan **Plots**-painiketta. Plots-valintaikkunasta valitaan **Normal probability plot**.



Normaalijakaumakuviot akselit on sopivien muunnosten avulla laadittu sellaisiksi, että normaalijakaumaa noudattavat havainnot sijoittuvat kuviossa suoralle viivalle. Jos kuviossa ei ole havaittavissa isoja poikkeamia suoralta viivalta, niin jakauma voidaan olettaa normaaliksi.

### Normal P-P Plot of Regression Standardized Residual



KUVIO 5. Jäännösten normaali jakaumakuvi

Jonkin verran poikkeamia on nähtävissä, mutta ei mitään vakavia rikkomuksia.

#### Jos ehtoa rikotaan

Pienet poikkeamat normaalisuudesta eivät yleensä ole vakavia. Jos poikkeamat ovat isoja, niin merkitsevyystestit ja ennusteille laskettavat luottamusvälit eivät ole päteviä. Sopivilla muuttujiin kohdistettavilla muunnoksilla tilannetta voidaan korjata. Tässä monisteessa ei käsitellä muuttujamuunnoksia.

#### Jäännösten riippumattomuus

Jäännösten mahdollinen riippuvuus (nk. autokorrelaatio) voi muodostua ongelmaksi aikasarjojen kohdalla. Aikasarjojen kohdalla tilanne voidaan tarkistaa hajontakuviolla, jossa x-akselilla on havainnon järjestysnumero (ajallisen järjestyksen mukaan) ja y-akselilla jäännökset (mielellään normitetut). Jos pistejoukossa ei näy satunnaisuudesta poikkeavaa säännönmukaisuutta, niin jäännökset voidaan olettaa toisistaan riippumattomiksi. Tässä monisteessa ei käsitellä aikasarja-analyyseja eikä näin ollen myöskään menetelmiä autokorrelaation huomioimiseen.

#### Koko mallin merkitsevyys

Edellä tehtyjen tarkastelujen perusteella voimme todeta, että esimerkissämme edeltävyysehdot ovat siinä määrin voimassa, että voimme käyttää merkitsevyystestejä mallin merkitsevyyden testaamiseen.

F-testillä testataan onko malli kokonaisuutena tilastollisesti merkitsevä. F-testin tulokset ovat luettavissa ANOVA-tilastustaulukosta.

Nollahypoteesi: Mallin kaikkien selittävien muuttujien regressiokertoimet ovat nollia.  
Vaihtoehtoinen hypoteesi: Ainakin yksi regressiokertoimista on nolasta poikkeava.

Jos F-testin p-arvo on alle 0,050, niin nollahypoteesi voidaan hylätä ja mallia voidaan tältä osin pitää merkitseväenä.

TAULUKKO 11. SPSS:n ANOVA-tilaus

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	13571,437	2	6785,719	38,274	,000 <sup>a</sup>
	Residual	1241,063	7	177,295		
	Total	14812,500	9			

a. Predictors: (Constant), Rakennusten pinta-ala neliömetreinä, Rantaviiva metreinä

b. Dependent Variable: Myyntihinta

Esimerkissämme p-arvo (**Sig.**) 0,000 on pienempi kuin 0,050, joten malli on tilastollisesti merkitsevä. Excelin suomenkielisessä versiossa p-arvo löytyy Excelin ANOVA-tilauksen kohdasta F:n tarkkuus.

## Yksittäisten selittävien muuttujien merkitsevyys

Yksittäisen selittävän muuttujan osalta testataan t-testillä nollahypoteesia:

Nollahypoteesi: Selittävään muuttujaan liittyvä regressiokerroin on nolla.

Vaihtoehtoinen hypoteesi: Kerroin on nolasta poikkeava.

Jos t-testin p-arvo on alle 0,050, niin nollahypoteesi voidaan hylätä. Jos nollahypoteesi jää voimaan, niin muuttuja yleensä jätetään mallista pois, ellei ole painavia teoreettisia tai käytännöllisiä perusteluja muuttujan pitämiseksi mallissa. Täytyy kuitenkin huomioida, että t-testin p-arvo vaihtelee sen mukaan, mitä muita muuttujia malliin on otettu.

TAULUKKO 12. SPSS:n Coefficients-tilaus

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-75,210	23,688		-3,175	,016
	Rantaviiva metreinä	1,915	,253	,869	7,566	,000
	Rakennusten pinta-ala neliömetreinä	2,555	,393	,746	6,495	,000

a. Dependent Variable: Myyntihinta

Esimerkissämme kumpaankin selittävään muuttujaan liittyvät p-arvot (**Sig.**) ovat alle 0,050 ja siltä osin muuttujat sopivat malliin.

## Regressiokertoimien luottamusväli

Jos merkitsevyydestäukseen liittyvät edeltävyys ehdot ovat täytetty, niin voimme määrittää regressiomallin kertoimille luottamusvälit. Excel tulostaa luottamusvälit suoraan. SPSS:ssä täytyy valita Regression-valintaikkunassa **Statistics**-painikkeen alta **Confidence intervals**. 95 % luottamusväli muodostetaan tulosteessa annetun alarajan ja ylärajan avulla.

TAULUKKO 13. Regressiokertoimien luottamusvälit SPSS:n Coefficients-taulukossa

Model		Coefficients <sup>a</sup>						
		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta				
1	(Constant)	-75,210	23,688		-3,175	,016	-131,224	-19,196
	Rantaviiva metreinä	1,915	,253	,869	7,566	,000	1,316	2,513
	Rakennusten pinta-ala neliömetreinä	2,555	,393	,746	6,495	,000	1,624	3,485

a. Dependent Variable: Myyntihinta

Esimerkiksi rantaviivaan liittyvä kerroin on mallissa 1,915 ja sen 95 % luottamusväli on 1,316 – 2,513.

## Ennusteen luottamusväli

On tärkeää erottaa toisistaan kaksi ennustetta, keskiarvoennuste ja yksittäistapauksen ennuste:

1. Tiettyihin selittävän muuttujan arvoihin voi käytännössä liittyä erilaisia selitettävän muuttujan arvoja. Esimerkimmme tapauksessa kaksi mökkiä voivat olla erihintaisia vaikka niillä on samanpituinen rantaviiva ja yhtä suuri rakennusten pinta-ala. Mallin avulla voidaan ennustaa tiettyihin selittävän muuttujan arvoihin liittyvää keskiarvoa (esimerkiksi **tietyn rantaviivan ja rakennusten pinta-alan omaavien mökkien keskiarvohintaa**).
2. Voimme myös ennustaa tiettyihin selittävän muuttujan arvoihin liittyvää yksittäistapausta (esim. **tietyn mökin hintaa**).

Ennustimme pa sitten keskiarvoa tai yksittäistapausta, niin ennusteen arvo on sama. Sen sijaan luottamusväli on keskiarvon kohdalla pienempi kuin yksittäistapauksen kohdalla. Keskiarvo saadaan siis ennustettua tarkemmin kuin yksittäistapaus.

Ennusteen luottamusväli vaihtelee myös sen mukaan kuinka kaukana havaintojen keskiarvosta ollaan. Mitä kauempana havaintojen keskiarvosta ollaan, sitä epätarkemmiksi ennusteet käyvät. SPSS osaa laskea havaintojen perusteella laskettujen ennusteiden luottamusvälit. Näiden avulla voit arvioida mallin antamien ennusteiden tarkkuutta.

- Napsauta SPSS:n Regression-valintaikkunassa **Save**-painiketta.
- Valitse Save-valintaikkunasta **Predicted Values** –otsikon alta **Unstandardized** sekä **Prediction Intervals** -otsikon alta **Mean** ja/tai **Individual**.

Tuloksena saat SPSS-aineistoon uusia muuttujia:

TAULUKKO 14. SPSS-aineistoon lasketut ennusteiden luottamusvälit

	nro	rantaviiva	ala	hinta	PRE 1	LMCI 1	UMCI 1	LICI 1	UICI 1
1	1	30	50	95	109,9642	90,47825	129,4502	72,93672	146,9918
2	2	35	42	95	99,10241	83,29822	114,9066	63,87307	134,3317
3	3	40	25	80	65,24972	45,53841	84,96103	28,10312	102,3963
4	4	50	30	100	97,17139	84,05182	110,2910	63,06190	131,2809
5	5	55	45	135	145,0640	132,7669	157,3611	111,2623	178,8656
6	6	60	24	100	100,9931	85,68847	116,2978	65,98503	136,0012
7	7	60	60	210	192,9566	168,6981	217,2151	153,2098	232,7034
8	8	70	34	160	145,6875	133,1008	158,2742	111,7794	179,5956
9	9	80	32	150	159,7274	143,0358	176,4189	124,0911	195,3636
10	10	85	28	150	159,0837	139,7729	178,3946	122,1481	196,0194

- PRE\_1 = Havaintoja vastaavat mallin antamat ennusteet.
- LMCI = Lower Mean Confidence Interval eli ennustetun keskiarvon luottamusvälin alaraja.
- UMCI = Upper Mean Confidence Interval eli ennustetun keskiarvon luottamusvälin yläraja.
- LICI = Lower Individual Confidence Interval eli ennustetun arvon luottamusvälin alaraja.
- UICI = Upper Individual Confidence Interval eli ennustetun arvon luottamusvälin alaraja.

Esimerkiksi mökin nro 1 tiedoilla mallin antaman keskiarvohinnan luottamusväli on 90,478 - 129,450 (tuhansina euroina). Ensimmäisen mökin tiedoilla mallin antaman yksittäisen mökin hinnan luottamusväli on 72,937 - 146,992 (tuhansina euroina). Luottamusvälit ovat esimerkissämme isoja. Tämä on ymmärrettävää, koska esimerkkimme otos (n=10) on liian pieni käyttökelpoisen mallin laatimiseen.

## 9 Selittävien muuttujien valitseminen

Mieluiten regressiomalliin mukaan otettavat selittävät muuttujat ovat teorian ja/tai käytännön perusteella etukäteen tiedossa. Jos taas malliin mukaan otettavat selittävät muuttujat täytyy valita isommasta joukosta mahdollisia selittäviä muuttujia, niin tarvitaan kriteerit muuttujien valinnalle. Tarjolla on erilaisia kriteerejä joiden käyttö voi johtaa erilaisiin malleihin. Kyse onkin viime kädessä kompromissista eri kriteerien, käytännöllisen toimivuuden ja edeltävyyssehtojen välillä.

Seuraavassa tarkastellaan selityskertoimen ja t-testin käyttöä muuttujien valinnan tukena sekä varoitetaan kolineaarisuuden ja multikolineaarisuuden vaaroista.

### Selityskerroin ja korjattu selityskerroin

Kuten aiemmin on todettu, niin selityskerroin ilmaisee kuinka monta prosenttia jäännösneliösummasta saadaan mallin avulla selitettyä. Toisin sanottuna selityskerroin on mitta selitetyn vaihtelun prosenttiosuudelle. Näin ollen selityskerroin on mitta mallin hyvyydelle. Mitä suurempi selityskerroin on, sitä parempi malli on. Selityskerrointa voidaan hyödyntää otettaessa malliin uusi selittävä muuttuja. Jos uuden selittävän muuttujan mukaan tuominen kasvattaa selityskerrointa merkittävästi, niin muuttujan mukaan ottaminen on perusteltua.

Yleensä uuden muuttujan tuominen malliin kasvattaa aina selityskerrointa (ei ainakaan pienennä sitä). Selityskertoimen perusteella malliin voidaan ottaa sellaisiakin muuttujia, jotka eivät oikeasti malliin kuulu. Tämän välttämiseksi selityskertoimen lisäksi kannattaa tarkastella myös tarkistettua selityskerrointa (Adjusted R Square). Tarkistettu selityskerroin huomioi myös malliin mukaan otettujen selittävien muuttujien lukumäärän. Tarkistettu selityskerroin voi jopa pienentyä uuden muuttujan myötä. Monet tutkijat suosittelivat korjatun selityskertoimen käyttöä päätettäessä kannattaako uusi muuttuja ottaa malliin mukaan. Jos uusi muuttuja lisää korjatun selityskertoimen arvoa merkittävästi, niin muuttuja kannattaa ottaa malliin mukaan.

## T-testin p-arvo

Muuttujaa, jonka regressiokertoimeen liittyvän t-testin p-arvo on yli 0,050 ei voida pitää tilastollisesti merkitsevinä selittäjinä. Jos on vahvat teoreettiset tai käytännöstä lähtevät perusteet, niin malliin voidaan ottaa mukaan hieman suuremman p-arvon omaavia muuttujia. T-testin p-arvoa seurattaessa kannattaa pitää mielessä, että muuttujaan liittyvä p-arvo riippuu siitä, mitä muita muuttujia mallissa on mukana. Tämän vuoksi ennen lopullisia päätöksiä kannattaa kokeilla erilaisia selittävien muuttujien yhdistelmiä sisältäviä malleja.

## Kolinearisuus ja multikolinearisuus

Kolinearisuus tarkoittaa kahden muuttujan välistä korrelaatiota ja multikolinearisuus useamman muuttujan välistä korrelaatiota. Selittävien muuttujien väliset korrelaatiot vaikeuttavat regressiomallin tulkintaa.

Havainnollistetaan kolinearisuuteen liittyviä ongelmia kärjistetyn esimerkin avulla. Otetaan esimerkkiaineistoomme mukaan tontin pinta-ala.

TAULUKKO 15. Aineisto, jossa voimakas kolinearisuus

rantaviiva (m)	rakennusten pinta-ala m <sup>2</sup>	tontin pinta-ala m <sup>2</sup>	myyntihinta (tuhatta euroa)
30	50	1900	95
35	42	2200	95
40	25	2600	80
50	30	3050	100
55	45	3300	135
60	24	3500	100
60	60	3250	210
70	34	3700	160
80	32	4500	150
85	28	4800	150

Tontin pinta-ala korreloi voimakkaasti rantaviivan pituuden kanssa. Korrelaatiokerroin on 0,99. Edellä olemme todenneet, että rantaviivan pituus selittää melko hyvin myyntihinnan vaihtelua. Koska tontin pinta-ala on lähes suorassa yhteydessä rantaviivan pituuteen, niin myös tontin pinta-ala selittää myyntihinnan vaihtelua. Itse asiassa rantaviivan pituus ja tontin pinta-ala selittävät pitkälti samaa myyntihinnassa esiintyvää vaihtelua.

Jos laskemme regressiomallin, jossa myös tontin pinta-ala on selittävänä muuttujana, niin saamme omituisia tuloksia.

TAULUKKO 16. Eikö tontin pinta-alan kasvu kasvatakaan mökin hintaa

		Coefficients <sup>a</sup>				
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-23,304	38,197		-,610	,564
	Rantaviiva metreinä	4,880	1,827	2,215	2,672	,037
	Rakennusten pinta-ala neliömetreinä	2,104	,448	,615	4,695	,003
	Tontin pinta-ala	-,062	,038	-1,401	-1,636	,153

a. Dependent Variable: Myyntihinta

Tontin pinta-alaan liittyvä regressiokerroin on negatiivinen (-0,062). Tämä on ristiriidassa sen kanssa, että myyntihinta yleensä kasvaa tontin pinta-alan kasvaessa. Selitys omituisuuteen on se, että tontin pinta-alan selittämä vaihtelu on sisäänrakennettu rantaviivan pituuden regressiokertoimeen. Tästä on syytä oppia, että regressiokertoimien tulkinnat voivat mennä täysin pieleen, jos selittävien muuttujien välillä on voimakkaita korrelaatioita.

Toinen merkille pantava asia on tontin pinta-alan regressiokertoimeen liittyvä t-testin p-arvo 0,153. Tämä on suurempi kuin 0,050, joten tontin pinta-ala ei selitä tilastollisesti merkitsevästi myyntihinnan vaihtelua. Selitys tähänkin omituisuuteen on sama. Jos tontin pinta-alan selittämä vaihtelu on sisäänrakennettu rantaviivan pituuden regressiokertoimeen, niin tontin pinta-ala ei enää näytä selittävän myyntihinnan vaihtelua. Jos kuitenkin jättäisimme rantaviivan pituuden pois mallista, niin huomaisimme tontin pinta-alan selittävän tilastollisesti merkitsevästi myyntihinnan vaihtelua.

TAULUKKO 17. Tontin pinta-ala selittää myyntihinnan vaihtelua

		Coefficients <sup>a</sup>				
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-101,806	33,441		-3,044	,019
	Rakennusten pinta-ala neliömetreinä	2,798	,500	,817	5,598	,001
	Tontin pinta-ala	,038	,006	,869	5,952	,001

a. Dependent Variable: Myyntihinta

Tästä on syytä oppia, että t-testin käyttö on kyseenalaista, jos selittävien muuttujien välillä on voimakkaita korrelaatioita.

Mahdollinen kolineaarisuus selviää tarkastelemalla selittävien muuttujien välisiä korrelaatiokertoimia. Yksikäsitteistä rajaa mallin kannalta haitallisille korrelaatiolle ei ole. Asiaan kannattane kiinnittää huomiota, jos selittävien muuttujien välillä on itseisarvoltaan yli 0,70 suuruisia korrelaatioita.

SPSS:llä voidaan laskea kullekin selittävälle muuttujalle tunnusluvut, jotka mittaavat kolineaarisuutta/multikolineaarisuutta. Jos valitset SPSS:n Regression-valintaikkunassa **Statistics** ja valitset laskettavaksi **Collinearity diagnostics**, niin saat tulostaulukkoon tunnusluvut Tolerance ja VIF. Tunnusluvut ovat yhteydessä toisiinsa seuraavasti: Tolerance = 1 / VIF.

TAULUKKO 18. Toleranssit

		Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-23,304	38,197		-,610	,564		
	Rantaviiva metreinä	4,880	1,827	2,215	2,672	,037	,014	71,203
	Rakennusten pinta-ala neliömetreinä	2,104	,448	,615	4,695	,003	,564	1,774
	Tontin pinta-ala	-,062	,038	-1,401	-1,636	,153	,013	75,962

a. Dependent Variable: Myyntihinta

Toleranssi on se osuus selittävän muuttujan vaihtelusta, jota muut mallissa olevat selittävät muuttujat eivät selitä. On tietenkin sitä parempi, mitä lähempänä ykköstä toleranssin arvo on. Yksikäsitteistä rajaa mallin kannalta haitalliselle toleranssille ei ole. Asiaan kannattaa kiinnittää huomiota ainakin jos toleranssi on pienempi kuin 0,20 (VIF suurempi kuin 5). Edellä rantaviivan ja tontin pinta-alan toleranssit ovat selvästi pienempiä kuin 0,20.

Jos mallia käytetään ainoastaan ennusteiden laskemiseen, niin kolinearisuus/multikolinearisuus ei ole haitallista.

## 10 Kategorisen muuttujan käyttö selittävänä muuttujana

Pääsääntöisesti lineaarisen regressiomallin selittävät muuttujat ovat määrällisiä muuttujia. Jos kategorista muuttujaa käytetään selittävänä muuttujana, niin se koodataan dikotomisiksi (kaksiarvoiseksi) nk. dummy-muuttujaksi.

Esimerkkinä käyttämässämme tapauksessa selittävänä muuttujana voisi tulla kyseeseen mökin sähköliittymä. Sähköliittymän mahdollinen olemassaolo koodataan dummy-muuttujaksi siten, että 0=ei sähköliittymää ja 1=sähköliittymä.

TAULUKKO 19. Mukana dummy-muuttuja

rantaviiva (m)	rakennusten pinta-ala m <sup>2</sup>	sähköliittymä	myyntihinta (tuhatta euroa)
30	50	0	95
35	42	0	95
40	25	1	80
50	30	1	100
55	45	0	135
60	24	1	100
60	60	1	210
70	34	1	160
80	32	0	150
85	28	0	150

Ottamalla sähköliittymä rantaviivan ja rakennusten pinta-alan ohella selittäväksi muuttujaksi saadaan malli:  $y = 1,9750 \cdot x_1 + 2,7758 \cdot x_2 + 20,2988 \cdot x_3 - 96,941$ , missä  $x_1$  = rantaviiva,  $x_2$  = rakennusten ala ja  $x_3$  = sähköliittymä.

TAULUKKO 20. Malli, jossa on mukana kategorinen selittävä muuttuja

## Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,991 <sup>a</sup>	,982	,973	6,682

a. Predictors: (Constant), Sähkö, Rantaviiva metreinä, Rakennusten pinta-ala neliömetreinä

ANOVA<sup>b</sup>

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14544,595	3	4848,198	108,580	,000 <sup>a</sup>
	Residual	267,905	6	44,651		
	Total	14812,500	9			

a. Predictors: (Constant), Sähkö, Rantaviiva metreinä, Rakennusten pinta-ala neliömetreinä

b. Dependent Variable: Myyntihinta

Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-96,941	12,767		-7,593	,000
	Rantaviiva metreinä	1,975	,128	,897	15,470	,000
	Rakennusten pinta-ala neliömetreinä	2,776	,203	,811	13,674	,000
	Sähkö	20,299	4,348	,264	4,668	,003

a. Dependent Variable: Myyntihinta

Mallin mukaan sähköliittymä lisää mökin hintaa 20,299 tuhatta eli 20299 euroa. Mallin jäännösneliösumma on 267,905 (vrt. 1241,06 mallissa, jossa selittäjinä olivat rantaviiva ja rakennusten pinta-ala). Selityskertoimeksi saadaan 98,2 % (vrt. 91,6 % mallissa, jossa selittäjinä olivat rantaviiva ja rakennusten pinta-ala). Sähkön regressiokertoimelle lasketun t-testin p-arvo on 0,003. Näin ollen sähkö sopii hyvin mallimme selittäväksi muuttujaksi.

Jos selittäväksi muuttujaksi halutaan useampia kategorioita sisältävä kategorinen muuttuja, niin se koodataan useammaksi dummy-muuttujaksi.

Esim. Jos ottaisimme esimerkissämme selittäväksi muuttujaksi mökin rannan ilmansuunnan (etelä, pohjoinen, itä, länsi), niin se voidaan koodata kolmen dummy-muuttujan avulla:

- pohjoinen, joka saa arvon 1 ilmansuunnan ollessa pohjoinen, muutoin arvon 0.
- itä, joka saa arvon 1 ilmansuunnan ollessa itä, muutoin arvon 0.
- länsi, joka saa arvon 1 ilmansuunnan ollessa länsi, muutoin arvon 0.

Huomaa, että dummy-muuttujia tarvitaan yksi vähemmän kuin kategorioita. Äskeisessä esimerkissä etelä tulee koodatuksi siten, että muuttujat pohjoinen, itä ja länsi saavat kaikki arvon 0. Regressiokertoimia tulkitessa esim. pohjoisen regressiokerroin ilmoittaisi hintaeron verrattaessa etelään.

Jos useita kategorioita sisältävä muuttuja koodataan dummy-muuttujiksi, niin yksittäisten dummy-muuttujien regressiokertoimien t-testien p-arvoihin ei kannata kiinnittää huomiota (koska eri kategoriat sisältäviä dummy-muuttujia on tarkasteltava kokonaisuutena). Sen sijaan voidaan tarkastella selityskertoimen ja korjatun selityskertoimen kasvua verrattuna malliin, jossa dummy-muuttujia ei ole mukana. Jos selityskerroin kasvaa merkittävästi, niin dummy-muuttujien mukaan ottaminen on perusteltua.

## 11 SPSS:n valinta-algoritmit

SPSS sisältää valmiita algoritmeja selittävien muuttujien valintaan. Näitä kannattaa hyödyntää, jos malliin otettavat selittävät muuttujat eivät määräydy teorian tai käytännön kautta. Algoritmi valitaan SPSS:n Regression valintaikkunan **Method**-kohdasta. Aiemmissa esimerkeissä menetelmänä on ollut **Enter**, jolloin kaikki Independent(s)-ruutuun siirretyt muuttujat otetaan malliin mukaan. Tarkastellaan seuraavassa lyhyesti **Stepwise**-, **Forward**- ja **Backward**-algoritmeja.

### Stepwise

Independent(s)-ruutuun siirretyistä muuttujista valitaan ensimmäiseen malliin se, joka selittää selitettävän muuttujan vaihtelua parhaiten. Seuraaviin malleihin lisätään aina se jäljellä olevista muuttujista, joka parhaiten selittää selitettävän muuttujan vaihtelua.

Multikolinearisuuden/kolinearisuuden takia jo aiemmin mukaan otettujen muuttujien selitysvoima voi laskea uusien selittävien muuttujien mukaan otton myötä. Tämän vuoksi jo mukaan otettu muuttuja voidaan myöhemmässä vaiheessa pudottaa mallista. Kun mikään jäljelle jääneistä muuttujista ei enää selitä vaihtelua riittävästi, niin malli on valmis.

### Forward

Independent(s)-ruutuun siirretyistä muuttujista valitaan ensimmäiseen malliin se, joka selittää selitettävän muuttujan vaihtelua parhaiten. Seuraaviin malleihin lisätään aina se jäljellä olevista muuttujista, joka parhaiten selittää selitettävän muuttujan vaihtelua. Kun mikään jäljelle jääneistä muuttujista ei enää selitä vaihtelua riittävästi, niin malli on valmis.

Olenainen ero Stepwise-algoritmiin on se, että Forward-algoritmissa kerran malliin otettua muuttujaa ei enää myöhemmissä vaiheissa voida poistaa mallista.

### Backward

Independent(s)-ruutuun siirretyistä muuttujista valitaan ensimmäiseen malliin kaikki. Seuraavasta mallista poistetaan muuttuja, joka selittää selitettävän muuttujan vaihtelua huonoiten. Näin jatketaan kunnes malliin jää vain muuttujia, jotka merkittävästi selittävät selitettävän muuttujan vaihtelua.

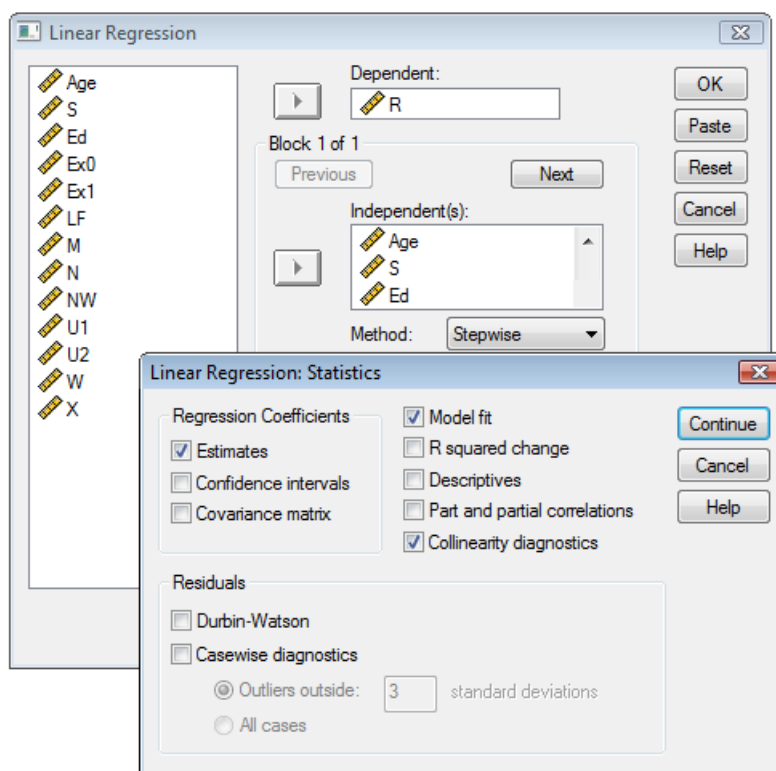
## Esimerkki valintaalgoritmin käytöstä

Tarkastellaan esimerkkinä Data and Story Library -sivuilta löytyvää aineistoa <http://lib.stat.cmu.edu/DASL/Stories/USCrime.html>. Löydät aineiston SPSS-muodossa osoitteesta <http://myy.haaga-helia.fi/~taaak/m/>.

Aineisto sisältää USAn rikostiheyden ja rikollisuutta mahdollisesti selittäviä muuttujia vuodelta 1960. Aineiston muuttujat ovat:

1. R: Rikostiheys: poliisille ilmoitettujen rikosten määrää miljoonaa asukasta kohden
2. Age: 14-24 vuotiaiden miesten määrä tuhatta asukasta kohden
3. S: Etelävaltio (0 = Ei, 1 = Kyllä)
4. Ed: Keskiarvo koulutuksen pituus x 10 alle 25-vuotiailla
5. Ex0: 1960 rahankäyttö henkeä kohden poliisiin ja paikallishallintoon
6. Ex1: 1959 rahankäyttö henkeä kohden poliisiin ja paikallishallintoon
7. LF: 14-24 vuotiaiden työllisten miesten määrä 1000 asukasta kohden
8. M: Miesten lukumäärä 1000 naista kohden
9. N: Osavaltion asukasluku satoina tuhansina
10. NW: Ei-valkoisten määrä 1000 asukasta kohden
11. U1: Työttömiä 14-24 vuotiaita kaupunkilaismiehiä 1000 asukasta kohden
12. U2: Työttömiä 35-39 vuotiaita kaupunkilaismiehiä 1000 asukasta kohden.
13. W: Perheiden mediaanitulo kymmeninä dollareina.
14. X: Mediaanitulon puolikkaan alapuolella olevien perheiden määrä 1000 perhettä kohden.

Otetaan rikostiheys (R) selitettäväksi muuttujaksi ja kaikki muut muuttujat ehdolle selittäviksi muuttujiksi. Valitaan valinta-algoritmiksi Stepwise. Valitaan Statistics-painikkeen alta laskettavaksi Collinearity diagnostics.



SPSS:n Coefficients taulukosta voimme seurata, missä järjestyksessä SPSS on lisännyt muuttujia malliin. Viimeiseen 5. malliin on kelpuutettu 5 selittävää muuttujaa.

TAULUKKO 21. Mallin muodostus Stepwise-algoritmilla

		Coefficients <sup>a</sup>						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	14,446	12,669		1,140	,260		
	Ex0	,895	,141	,688	6,353	,000	1,000	1,000
2	(Constant)	-94,466	34,395		-2,747	,009		
	Ex0	1,241	,164	,954	7,582	,000	,602	1,660
	X	,410	,122	,422	3,357	,002	,602	1,660
3	(Constant)	-327,541	76,914		-4,259	,000		
	Ex0	1,243	,148	,955	8,408	,000	,602	1,660
	X	,751	,151	,774	4,978	,000	,321	3,110
	Ed	1,579	,477	,457	3,312	,002	,409	2,444
4	(Constant)	-424,922	85,851		-4,950	,000		
	Ex0	1,298	,144	,997	9,029	,000	,584	1,711
	X	,641	,153	,661	4,197	,000	,287	3,478
	Ed	1,661	,458	,480	3,626	,001	,406	2,460
	Age	,760	,344	,247	2,209	,033	,570	1,754
5	(Constant)	-524,374	95,116		-5,513	,000		
	Ex0	1,233	,142	,948	8,706	,000	,557	1,796
	X	,635	,147	,655	4,324	,000	,287	3,480
	Ed	2,031	,474	,587	4,283	,000	,351	2,853
	Age	1,020	,353	,331	2,887	,006	,501	1,998
	U2	,914	,434	,199	2,105	,041	,734	1,363

a. Dependent Variable: R

Toleranssiarvot ovat suurempia kuin 0,2, joten kolineaarisuuden/multikolineaarisuuden osalta ei ole syytä erityiseen huoleen.

Lukijaa kehoitetaan täydentämään esimerkkiä seuraavasti:

- Edeltävyysehtojen tarkastelu jäännöskuvion ja jäännöstermien normaalijakaumakuvion avulla.
- Mallin merkitsevyyden tarkastelu F-testin avulla.
- Yksittäisten malliin mukaan otettujen selittäjien merkitsevyys t-testin avulla.
- Forward- ja Backward-algoritmien kokeilu. Tämän esimerkin tapauksessa Backward-algoritmi jättää malliin yhden muuttujan enemmän kuin Stepwise- ja Forward-algoritmit.
- Muuttujien välisten korrelaatioiden laskeminen.
- Pohdiskelua selittäivistä muuttujista (vrt. <http://lib.stat.cmu.edu/DASL/Stories/USCrime.html>).

## 12 Harjoitus

Löydät harjoitukseen liittyvän SPSS-aineiston HBAT.sav osoitteesta

<http://myy.haaga-helia.fi/~taaak/m/>

Aineistossa on muuttuja X19, joka mittaa asiakastytyväisyyttä asteikolla 1-10 ja muuttujat X6-X18, jotka kuvaavat kokemuksia yrityksestä asteikolla 1-10.

---

X6 Product Quality	X13 Competitive Pricing
X7 E-commerce	X14 Warranty & Claims
X8 Technical Support	X15 New Products
X9 Complaint Resolution	X16 Ordering & Billing
X10 Advertising	X17 Price Flexibility
X11 Product Line	X18 Delivery Speed
X12 Salesforce Image	

---

Laadi regressiomalli, jossa selitettävä muuttujia on X19 ja selittävät muuttujat valitaan muuttujista X6-X18.

Aineisto HBAT.SAV on peräisin kirjasta:

Hair, Black, Babin, Anderson, Tatham. 2006. *Multivariate Data Analysis*. Sixth edition. Prentice Hall.

Kirja on sopivaa jatkolukemista tälle monisteelle. Kirjasta löytyy tietoa mm. muuttujamuunnoksista, joilla voidaan korjata edeltävysehtojen rikkomuksia. Kirjasta löytyy myös ratkaisu yllä annettuun harjoitustehtävään.