MTTTA1 Tilastomenetelmien perusteet 5 op
Luento 10.1.2019

1 Kokonaisuudet johon opintojakso kuuluu

https://www10.uta.fi/opas/opintojakso.htm?rid=14600&idx=1&uiLang=fi&lang=fi&lvv=2018

2 Osaamistavoitteet

https://www10.uta.fi/opas/opintojakso.htm?rid=14600&idx=1&uiLang=fi&lang=fi&lvv=2018

- Opiskelija osaa käyttää opintojaksolla esiteltyjä tilastollisia menetelmiä sekä ymmärtää niihin liittyvän teorian.
- Hän osaa annetussa tutkimustilanteessa suorittaa tilastollisen päättelyn joko valmiiksi annettujen tai itse laskemiensa tulosten perusteella.
- Hän osaa valita asetettuun tutkimusongelmaan liittyen sopivan menetelmän, suorittaa tilanteeseen sopivalla ohjelmistolla kyseisen analyysin sekä tulkita saadut tulokset.

Esim. Tampereella keväällä 2006 myynnissä olleita kerrostalohuoneistoja, aineisto http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Tutkimuskohteita

1) Vaikuttaako sijainti neliöhintaan?

y = neliöhinta
x = sijainti

SPSS-harj. 1 teht. 3a

2) Vaikuttaako huoneiden lukumäärä neliöhintaan? Miten sijainti vaikuttaa tähän riippuvuuteen?

y = neliöhinta
x = huoneiden lukumäärä (luokiteltuna)
z = sijainti

SPSS-harj. 1 teht. 3b

3) Vaikuttaako sijainti huoneiden lukumäärään?

y = huoneiden lukumäärä (luokiteltuna)
x = sijainti

SPSS-harj. 2 teht. 3

4) Miten huoneiston koko vaikuttaa hintaan?
Miten sijainti vaikuttaa tähän riippuvuuteen?

y = hinta
x = neliöt
z = sijainti

SPSS-harj. 3 teht. 2

3 Kurssin kotisivu

https://coursepages.uta.fi/mttta1/kevat-2019/

- Opetus
- Kurssi-info (sisältö, tentit, harjoitushyvitys)
- Luennot, luentorunko, kaavat, taulukot
- Harjoitukset, tehtävät, ohjeet (Moodle), ratkaisut
- Esimerkkiaineistoja
- Oheiskirjallisuutta
- Usein kysyttyä
- Linkkejä
- Palaute

4 Kertausta

Seuraaviin kohtiin 1) – 8) on koottu lyhyesti olennaisimmat asiat, jotka oletetaan opintojaksolla tunnetuiksi aiempien opintojen perusteella.

1) Empiiriset jakaumat
- yksiulotteiset
  taulukot, graafiset esitykset, tunnusluvut
- kaksiulotteiset
  ristiintaulukko, pisteparvi, korrelaatiokerroin
- ehdolliset jakaumat
  riippuvuus, ehdolliset tunnusluvut, laatikko-jana-kuvio
- toteutus SPSS:llä (tai muulla ohjelmistolla)

2) Satunnaismuuttuja X

- todennäköisyysjakauma, tiheysfunktio f(x)
- kertymäfunktio $F(x) = P(X \leq x)$
- $E(X) = \mu$, $Var(X) = \sigma^2$

3) Todennäköisyysjakaumia

- $X \sim N(\mu, \sigma^2)$, $Z = (X - \mu)/\sigma \sim N(0, 1)$.
  Jos $Z \sim N(0, 1)$, niin merkitään
  $F(z) = P(Z \leq z) = \Phi(z)$.

Olkoon $Z \sim N(0, 1)$. Määritellään z_α siten, että $P(Z \geq z_\alpha) = \alpha$. Vastaavalla tavalla $z_{\alpha/2}$ siten, että $P(Z \geq z_{\alpha/2}) = \alpha/2$.

Esim.

$z_{0,05} = 1,64$, koska $\Phi(1,64) = 0,9495$

$z_{0,025} = 1,96$, koska $\Phi(1,96) = 0,9750$

$z_{0,005} = 2,58$, koska $\Phi(2,58) = 0,9951$

- Studentin t-jakauma

  $P(t_{df} \geq t_{\alpha,df}) = \alpha$, $P(t_{df} \geq t_{\alpha/2,df}) = \alpha/2$

  Esim. $t_{0,05,\ 10} = 1,812$, $t_{0,05,\ 30} = 1,697$

  $t_{0,01,\ 10} = 2,764$, $t_{0,01,\ 30} = 2,457$

4) $X_1, X_2, \ldots, X_n$ on satunnaisotos, jos X_i:t ovat riippumattomia ja noudattavat samaa jakaumaa.

Sanonta
”$X_1, X_2, \ldots, X_n$ on satunnaisotos $N(\mu, \sigma^2)$:sta” tarkoittaa, että jokainen $X_i \sim N(\mu, \sigma^2)$ ja X_i:t ovat riippumattomia.

5) Otossuure, otosjakauma

Esim. Otossuure $\overline{X} \sim N(\mu, \sigma^2/n)$, jos satunnaisotos normaalijakaumasta.

6) Estimointi

- Estimointi on populaation tuntemattoman parametrin arviointia otossuureen avulla. Voidaan myös muodostaa väli (luottamusväli), jolla arvioidaan tuntemattoman parametrin olevan.
- Estimaattori otossuure, jolla estimoidaan tuntematonta parametria.
- Estimaatti on estimaattorin arvo.
- Harhaton estimaattori
- Estimaattorin keskivirhe (= estimaattorin keskihajonta)

7) Testaus

- Tilastollinen hypoteesi väite populaatiosta, usein populaation jakauman parametrista.

- Hypoteesin testaus on väitteen tutkimista otoksen perusteella.

- Asetetaan nollahypoteesi (H_0) ja vaihtoehtoinen hypoteesi (H_1).

- Testisuure on otossuure, jota käytetään hypoteesin tutkimisessa.

- Testisuureen jakauma tunnetaan nollahypoteesin ollessa tosi.

- Otoksesta lasketun testisuureen arvon perusteella nollahypoteesi hyväksytään tai hylätään kiinnitetyllä riskitasolla.

- p-arvo on pienin riskitaso, jolla H_0 voidaan hylätä.

8) Joitain testaustilanteita

- $H_0 : \pi = \pi_0$

  Prosenttiosuuden tutkiminen Z-testillä,
  kaava (5.3) MTTTP5

  Esim. Ystäväsi väittää, että suomalaisista on 10% vasenkätisiä. Tutkit asiaa ja valitset satunnaisesti 400 suomalaista, joista vasenkätisiä on 47. Uskotko ystäväsi väitteen?

  $$H_0 : \pi = 10$$

  $$H_1 : \pi > 10$$

Jos Ho tosi,

$$Z = \frac{p - 10}{\sqrt{10(100 - 10)/n}} \sim N(0,1), likimain.$$

Otoksesta laskettu z:n arvo on

$$z = \frac{11{,}75 - 10}{\sqrt{10(100 - 10)/400}} = 1{,}17$$

Pienin riskitaso, jolla H_0 voidaan hylätä yksisuuntaisessa testissä, on $P(Z > 1{,}17) = 1 - \Phi(1{,}17) = 1 - 0{,}8790 = 0{,}121$. Uskotaan siis ystävän väite.

Jos valitaan 5 %:n riskitaso, niin yksisuuntaisessa testissä kriittinen arvo on $z_{0,05} = 1,64$ (koska $\Phi(1,64) = 0,9495$) ja kaksisuuntaisessa testissä $z_{0,05/2} = 1,96$ (koska $\Phi(1,96) = 0,975$).

- $H_0 : \mu_1 = \mu_2$

Riippumattomien otosten t-testi odotusarvojen yhtäsuuruuden testaamiseksi,
kaava (5.5) MTTTP5

$$t = \frac{\overline{X} - \overline{Y}}{s\sqrt{1/n + 1/m}} \sim t(n+m-2), \qquad \text{kun } H_0 \text{ tosi,}$$

$$s^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{n+m-2},$$

Esim 1.4.3. Testi lasten kehityshäiriön tunnistamiseen

Suoritusajat testissä ryhmittäin

Normaali

204, 218, 197, 183, 227, 233, 191

Kehityshäiriö

243, 228, 261, 202, 343, 242, 220, 239

$H_0 : \mu_N = \mu_K$

$H_1 : \mu_N < \mu_K$

$\sum x_i = 204 + \ldots + 191 = 1453$

$\sum x_i^2 = 204^2 + \ldots + 191^2 = 303737$

$SS_N = 303737 - 7 \cdot (1453/7)^2 = 2135{,}714$

$\sum y_i = 243 + \ldots + 239 = 1978$

$\sum y_i^2 = 243^2 + \ldots + 239^2 = 501692$

$SS_K = 501692 - 8 \cdot (1978/8)^2 = 12631{,}5$

$$s^2 = \frac{2135{,}714 + 12631{,}5}{7 + 8 - 2} = 1135{,}94, s = 33{,}7$$

$$t_{hav.} = \frac{\frac{1453}{7} - \frac{1978}{8}}{33{,}7\sqrt{\frac{1}{7} + \frac{1}{8}}} = -2{,}28$$

$-t_{0,01;13} = -2{,}65 < t_{hav.} < -2{,}16 = -t_{0,025;13}$

H_0 voidaan hylätä 2,5 %:n riskitasolla, mutta ei 1%:n riskitasolla.

SPSS-tulos

Group Statistics

	ryhmä	N	Mean	Std. Deviation	Std. Error Mean
testi	normaali	$n_N = 7$	207,5714 $\bar{x}_N$	18,86670 s_N	7,13094
	kehityshäiriö	$n_K = 8$	247,2500 $\bar{x}_K$	42,47941 s_K	15,01874

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means	
		$H_0: \sigma_N^2 = \sigma_K^2$		$H_0: \mu_N = \mu_K$	
		F	Sig.	t	df
testi	Equal variances assumed	,926	,353	-2,275	13
	Equal variances not assumed	H_0 hyv.		-2,387	9,923

Independent Samples Test

		t-test for Equality of Means		
		Sig. (2-tailed)	Mean Difference	Std. Error Difference
testi	Equal variances assumed	,041	-39,67857	17,44332
	Equal variances not assumed	,038	-39,67857	16,62567

$H_1: \mu_N < \mu_K$

p-arvo $0{,}041/2 = 0{,}0205$

Luentorungon luvussa 1 http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=3 lyhyt kertaus olennaisimmista asioista, jotka oletetaan opintojaksolla tunnetuiksi aiempien opintojen perusteella.

Tarvittaessa kertaukseen ja tietojensa täydentämiseen voi käyttää kurssien

- MTTTP1 (https://coursepages.uta.fi/mtttp1/syksy-2018/)
- MTTTP5 (https://coursepages.uta.fi/mtttp5/syksy-2018/)

materiaaleja.

MTTTA1 Tilastomenetelmien perusteet
Luento 15.1.2019

Luku 2

Varianssianalyysi

2.1 Yksisuuntainen varianssianalyysi

Esim. 2.1.1 Tutkitaan golfpallojen keskimääräisiä lentomatkoja, saadaan tulokset:

Merkki	Keskiarvo	Keskihajonta	Lukumäärä
A	251,28	5,977	10
B	261,06	3,866	10
C	269,95	4,501	10

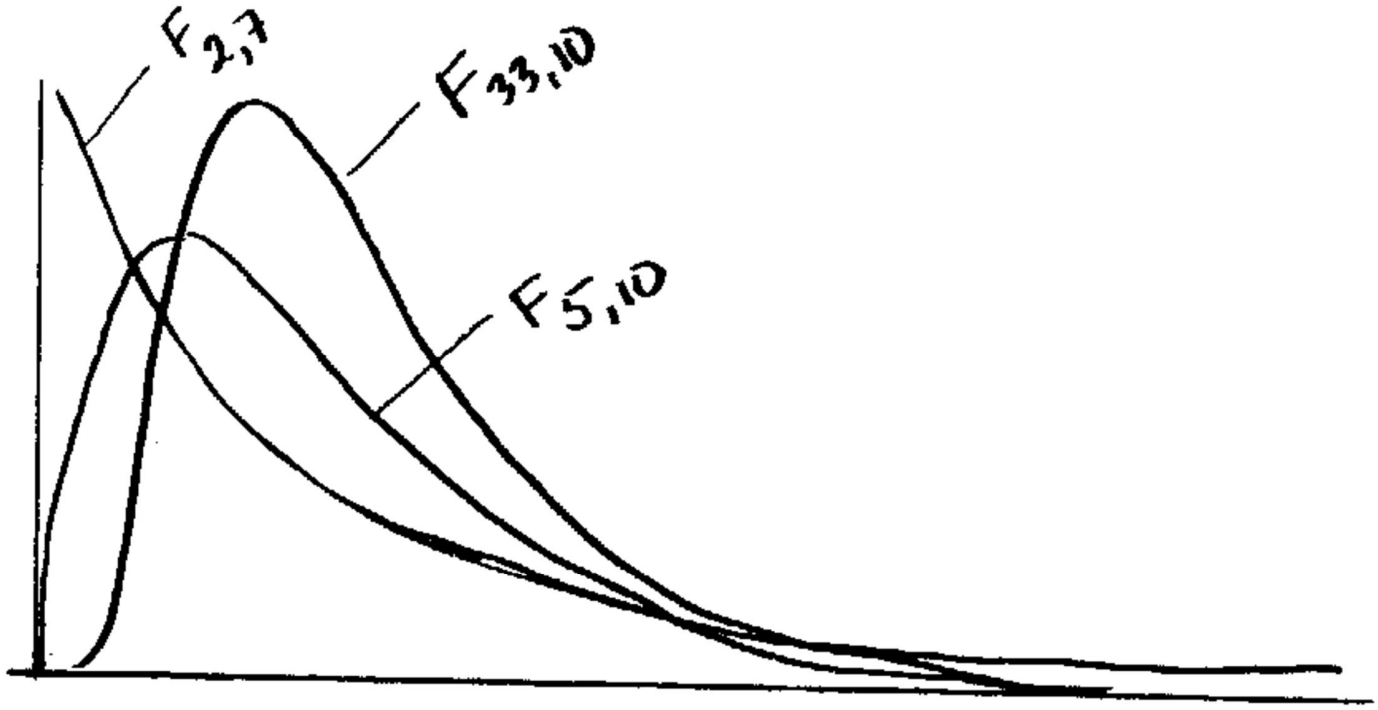
H_0: $\mu_A = \mu_B = \mu_C$
H_1: kaikki μ:t eivät samoja

F-testisuure H_0:n testaamiseksi

Annettujen lukujen perusteella voidaan laskea testisuureelle arvo, saadaan $F_{hav.} = 36{,}87$ ja p-arvo $< 0{,}0001$.

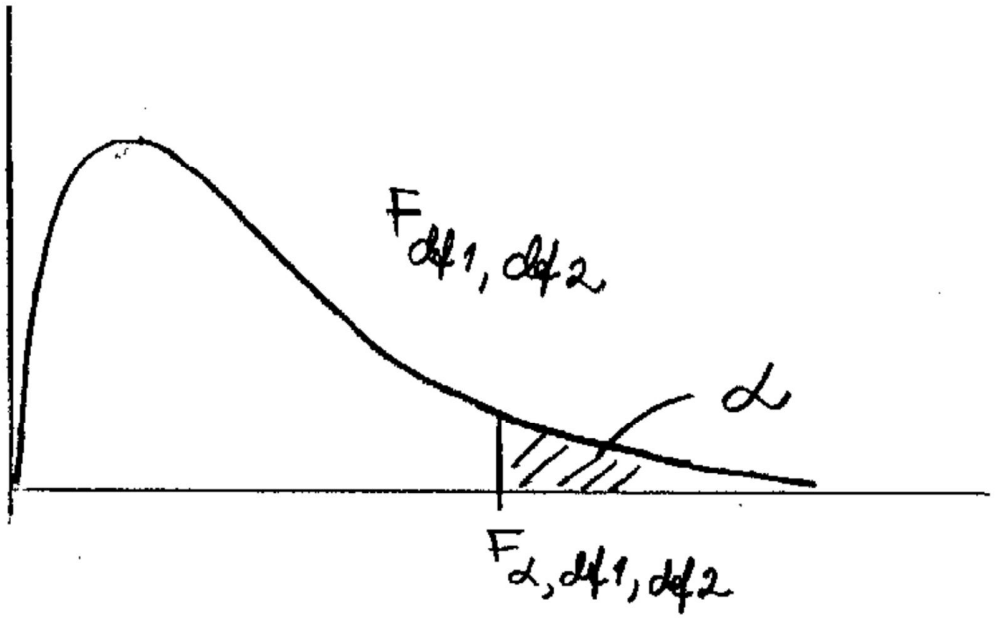
Hylätään H_0 ja päätellään odotusarvoissa olevan eroja.

Fisherin F-jakauman tiheysfunktion kuvaajia



F-jakauma määritellään kaksin vapausastein, $F_{df1,df2}$

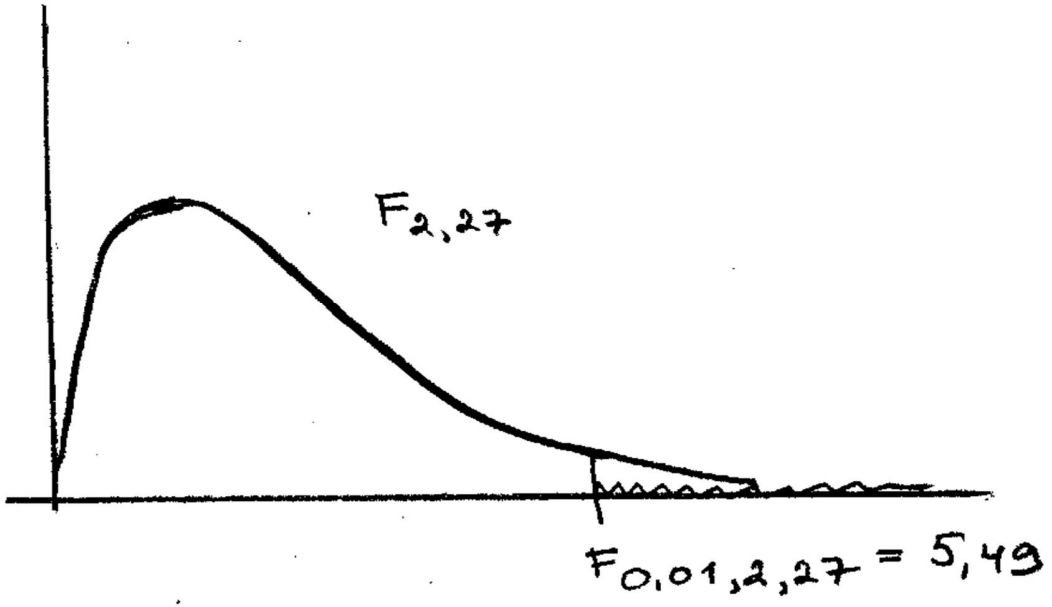
Määritellään $F_{\alpha;df1,\,df2}$ siten, että $P(F_{df1,df2} > F_{\alpha;df1,df2}) = \alpha$.



Näitä arvoja taulukosta
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/F_jakauma.pdf,
kun α = 0,01 tai α = 0,05.

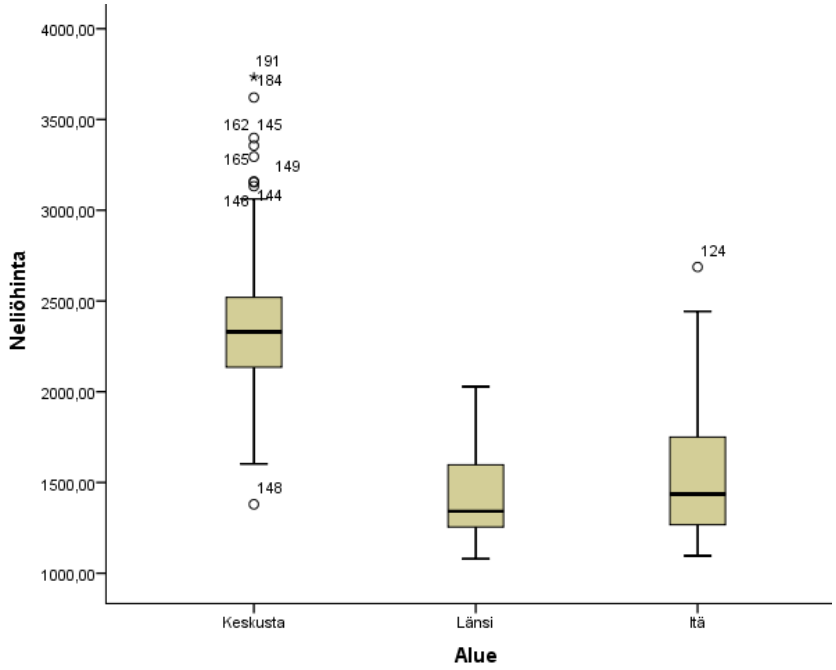
Esim. 2.1.1 Testisuure noudattaa H_0:n ollessa tosi F-jakaumaa vapausastein 2 ja 27.

$F_{0,01;2,27} = 5,49 < F_{hav.} = 36,87$, joten H_0 hylätään 1 %:n riskitasolla.



Esim. 2.1.6 Tutkitaan keskimääräisiä neliöhintoja Tampereen keskustassa, Länsi- ja Itä-Tampereella

Aineisto
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav
sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



H_0: $\mu_K = \mu_L = \mu_I$

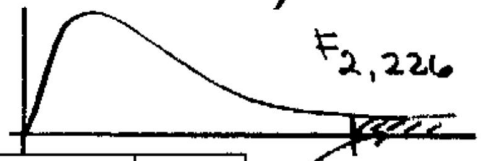
H_1: kaikki μ:t eivät samoja

Neliöhinta

Alue	Mean	N	Std. Deviation
Keskusta	2397,6072	103	408,02462
Länsi	1414,2870	34	260,39544
Itä	1536,1328	92	341,69439
Total	1905,5176	229	575,61088

$H_0: \mu_K = \mu_I = \mu_L$

H_1: kaikki μ:t eivät samoja



ANOVA

Neliöhinta

	Sum of Squares	*df*	Mean Square	*F*	Sig.
Between Groups	45699081	2	22849540.291	173.035	.000
Within Groups	29843678	226	132051.673		
Total	75542759	228			

Koska p-arvo < 0,001, H_0 hylätään ja päätellään eroja olevan. Päättely taulukkoarvon (http://www.sis.uta.fi/tilasto/mttta1/kevat2019/F_jakauma.pdf) perusteella: $F_{0,01;\ 2,\ 226} \approx 4,61 < F_{hav.} = 173,035$, joten H_0 hylätään 1 %:n riskitasolla.

Onko kaikkien alueiden välillä eroja?

Multiple Comparisons

Dependent Variable: Neliöhinta
Bonferroni

(I) Alue	(J) Alue	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Keskusta	Länsi	983,32022*	71,87439	,000	809,9641	1156,6764
	Itä	861,47438*	52,12868	,000	735,7435	987,2052
Länsi	Keskusta	-983,32022*	71,87439	,000	-1156,6764	-809,9641
	Itä	-121,84584	72,93296	,289	-297,7552	54,0635
Itä	Keskusta	-861,47438*	52,12868	,000	-987,2052	-735,7435
	Länsi	121,84584	72,93296	,289	-54,0635	297,7552

*. The mean difference is significant at the .05 level.

Länsi- ja Itä-Tampereen välillä ei eroja, muissa on. Tutkitaan odotusarvojen yhtäsuuruutta pareittain, päättely p-arvon tai luottamusvälin perusteella.

Varianssianalyysin liittyvät oletukset ja laskukaavat

$Y_{11}, Y_{12}, \ldots, Y_{1n1}$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta
$Y_{21}, Y_{22}, \ldots, Y_{2n2}$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta
.
.
.
$Y_{I1}, Y_{I2}, \ldots, Y_{InI}$ satunnaisotos $N(\mu_I, \sigma_I^2)$:sta

Oletetaan, että $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 = \sigma^2$ ja otokset riippumattomia.

H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$
H_1: kaikki μ:t eivät samoja

$$SST = \sum\sum (Y_{ij} - \bar{Y})^2 \,,\ \bar{Y} = \frac{1}{n}\sum\sum Y_{ij} \,,\ n = n_1 + \cdots + n_I$$

$$\text{SSB} = \sum n_i (\bar{Y}_i - \bar{Y})^2 \quad ,\ \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$$

$$SSW = \sum\sum (Y_{ij} - \bar{Y}_i)^2 = (n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2$$

$SST = SSB + SSW$
$MSB = SSB/(I\text{-}1)$
$MSW = SSW/(n\text{-}I)$

E(MSW) = σ^2 aina
E(MSB) = σ^2, jos H_0 tosi

$F = MSB/MSW \sim F_{I\text{-}1,\ n\text{-}I}$, kun H_0 tosi
H_0 hylätään riskitasolla α, jos $F_{hav} > F_{\alpha;\ I\text{-}1,\ n\text{-}I}$.

Esim. 2.1.3 Valmennusmenetelmien vaikutus urheilusuoritukseen

H_0: $\mu_1 = \mu_2 = \mu_3$
H_1: kaikki odotusarvot eivät samoja

Urheilusuoritukset menetelmittäin

Menetelmä 1: 6, 4, 6, 4
Menetelmä 2: 14, 9, 10, 11
Menetelmä 3: 5, 11, 8, 8

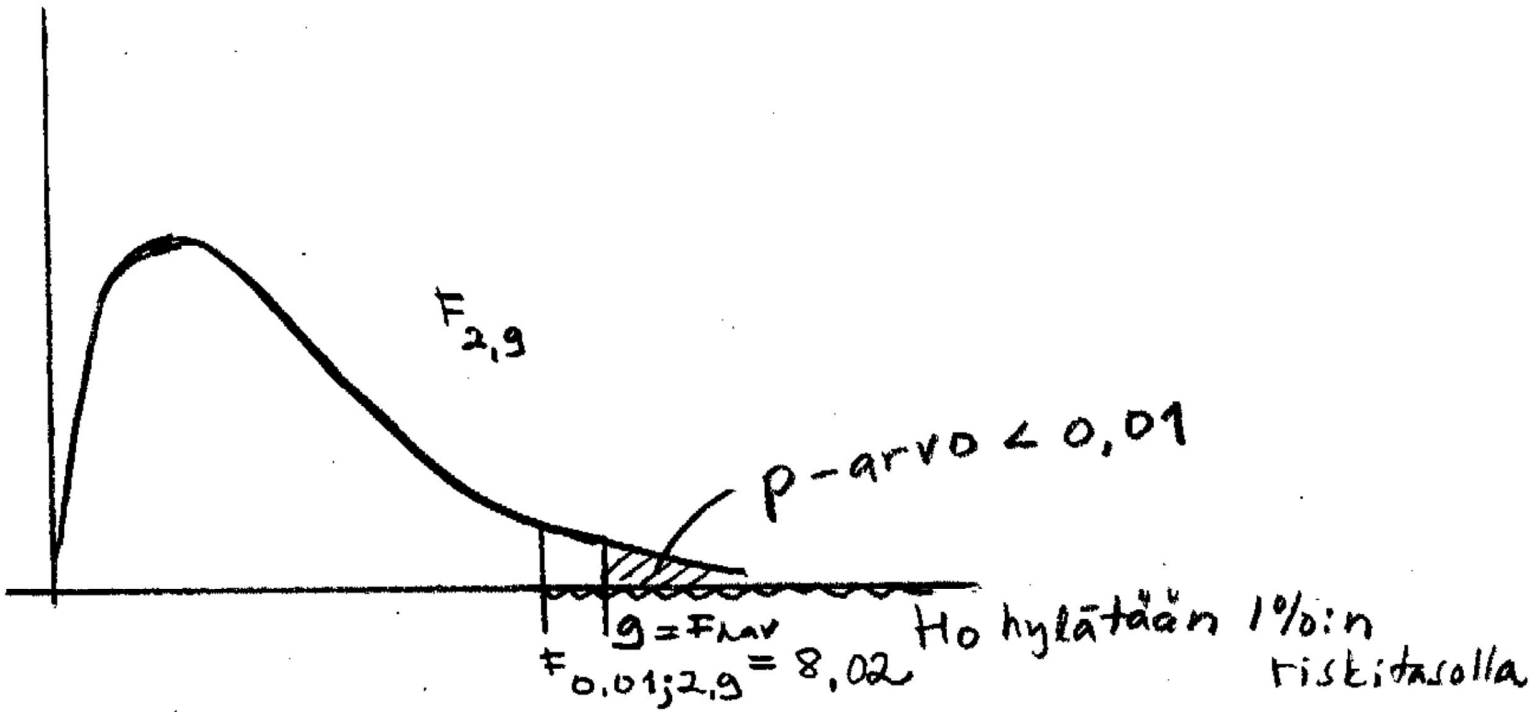
$$\overline{y}_1 = 5, \quad \overline{y}_2 = 11, \quad \overline{y}_3 = 8, \quad \overline{y} = 8,$$
$$n_1 = n_2 = n_3 = 4, \quad n = 12,$$

$$
\begin{aligned}
SST &= \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2 \\
&= (6-8)^2+\cdots+(8-8)^2 = 108, \\
SSB &= \sum_{i=1}^{3} n_i(\overline{y}_i-\overline{y})^2 \\
&= 4(5-8)^2+4(11-8)^2+4(8-8)^2 = 72, \\
SSW &= \sum_{i=1}^{3}\sum_{j=1}^{n_i}(y_{ij}-\overline{y}_i)^2 \\
&= (6-5)^2+(4-5)^2+(6-5)^2+(4-5)^2 \\
&\qquad +(14-11)^2+(9-11)^2+(10-11)^2+(11-11)^2 \\
&\qquad +(5-8)^2+(11-8)^2+(8-8)^2+(8-8)^2 = 36,
\end{aligned}
$$

$$MSB = SSB/(I-1) = 72/(3-1) = 36,$$
$$MSW = SSW/(n-I) = 36/(12-3) = 4,$$
$$F = MSB/MSW = 36/4 = 9,$$

$F_{0,01;\ 2,\ 9} = 8{,}02 < F_{hav.} = 9$, joten H_0 hylätään 1 %:n riskitasolla.

Voidaan sanoa, että p-arvo = $P(F_{2,9} > 9) < 0{,}01$.


$F_{2,9}$
p-arvo < 0,01
$g = F_{havainto}$
$F_{0,01;2,9} = 8,02$
Ho hylätään 1%:n riskitasolla

SPSS-tulos

ANOVA

Suoritus

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	72,000	2	36,000	9,000	,007
Within Groups	36,000	9	4,000		
Total	108,000	11			

Jos H_0: $\mu_1 = \mu_2 = \ldots = \mu_I$ hylätään, niin voidaan tutkia mitkä odotusarvot poikkeavat toisistaan. Tutkitaan odotusarvoja pareittain testin tai luottamusvälin avulla.

Esim. 2.1.3 Vain menetelmien 1 ja 2 välillä eroja.

Multiple Comparisons

Dependent Variable: Suoritus
Bonferroni

(I) Menetelmä	(J) Menetelmä	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-6,000*	1,414	,006	-10,15	-1,85
	3	-3,000	1,414	,189	-7,15	1,15
2	1	6,000*	1,414	,006	1,85	10,15
	3	3,000	1,414	,189	-1,15	7,15
3	1	3,000	1,414	,189	-1,15	7,15
	2	-3,000	1,414	,189	-7,15	1,15

*. The mean difference is significant at the 0.05 level.

Oletusta varianssien yhtäsuuruudesta voidaan myös testata (Levenen testi). Tällöin H_0: $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2$.

Jos variansseja ei voida olettaa samoiksi (Levenen testin p-arvo < 0,05), niin käytetään Welchin tai Brown-Forsythen testejä odotusarvojen yhtäsuuruuden testaamisessa.

Esim. 2.1.3 Varianssien yhtäsuuruuden testaaminen

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

Test of Homogeneity of Variances

Suoritus

Levene Statistic	df1	df2	Sig.
,214	2	9	,811

Hyväksytään H_0, koska p-arvo = 0,811 > 0,05.
Voidaan siis olettaa varianssit yhtä suuriksi.

Nimitys varianssianalyysi tulee siitä, että testisuure on kahden varianssiestimaattorin osamäärä.

Jos $I = 2$, niin $H_0: \mu_1 = \mu_2$. Tällöin $t^2 = F$.

MTTTA1 Tilastomenetelmien perusteet
Luento 17.1.2019

Kertausta ja täydennystä 1-VA

$H_0: \mu_1 = \mu_2 = \ldots = \mu_I$

H_1: kaikki μ:t eivät samoja

Oletetaan riippumattomat otokset:

$Y_{11}, Y_{12}, \ldots, Y_{1n1}$ satunnaisotos $N(\mu_1, \sigma_1^2)$:sta
$Y_{21}, Y_{22}, \ldots, Y_{2n2}$ satunnaisotos $N(\mu_2, \sigma_2^2)$:sta
.
.
.
$Y_{I1}, Y_{I2}, \ldots, Y_{InI}$ satunnaisotos $N(\mu_I, \sigma_I^2)$:sta

Oletetaan lisäksi, että $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_I^2 = \sigma^2$.

Neliösummat:

$$\underbrace{\sum\sum(Y_{ij} - \bar{Y})^2}_{SST} = \underbrace{\sum n_i(\bar{Y}_i - \bar{Y})^2}_{SSB} + \underbrace{\sum\sum(Y_{ij} - \bar{Y}_i)^2}_{SSW}$$

vaihtelu	neliö-summat (SS)	vapaus-asteet (df)	keskineliö-summat (MS)	F-arvo	p-arvo
välinen	SSB	$I-1$	$MSB = \frac{SSB}{I-1}$	$F = \frac{MSB}{MSW}$	$P(F \geq F_{\text{hav.}})$
sisäinen (jäännös)	SSW	$n-I$	$MSW = \frac{SSW}{n-I}$	$\sim F(I-1, n-I)$ kun H_0 tosi	
kokonais	SST	$n-1$			

ks. kaavakokoelma
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/kaavat.pdf

Esim. 2.1.4 Tutkitaan autotyyppien A, B, ja C kulutusta (miles per gallon), http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=15

A-autot	B-autot	C-autot
22.2	24.6	22.7
19.9	23.1	21.9
20.3	22.0	23.3
21.4	23.5	24.1
21.2	23.6	22.1
21.0	22.1	23.4
20.3	23.5	

	n	Mean	Std. Deviation
A	7	20.9000	.79162
B	7	23.2000	.90921
C	6	22.9167	.84004
Total	20	22.3100	1.33610

Test of Homogeneity of Variances

Kulutus (miles/gallon)

Levene Statistic	df1	df2	Sig.
,036	2	17	,965

$H_0: \sigma_A^2 = \sigma_B^2 = \sigma_C^2$

H_0 hyväksytään, koska p-arvo > 0,05

ANOVA

Kulutus (miles/gallon)

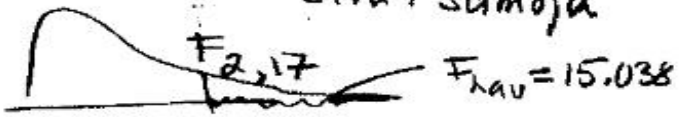
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	SSB 21,670	I-1 2	MSB 10,835	15,038	,000
Within Groups	SSW 12,248	n-I 17	MSW ,720	MSB/MSW	
Total	33,918	19			

$H_0: \mu_A = \mu_B = \mu_C$

H_1: kaikki odotusarvot eivät samoja

Post Hoc Tests

I = 3, n = 20



$6{,}11 = F_{0.01,2,17}$

H_0 hylätään

Multiple Comparisons

Kulutus (miles/gallon)
Bonferroni

(I) Autot yyppi	(J) Autot yyppi	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
A	B	-2,3000*	,4537	,000	-3,505	-1,095
	C	-2,0167*	,4722	,002	-3,270	-,763
B	A	2,3000*	,4537	,000	1,095	3,505
	C	,2833	,4722	1,000	-,970	1,537

*. The mean difference is significant at the 0.05 level.

B ja C eivät eroa toisistaan

Esim. 2.1.1 Tutkitaan golfpallojen keskimääräisiä lentomatkoja, saadaan tulokset:

Merkki	Keskiarvo	Keskihajonta	Lukumäärä
A	251,28	5,977	10
B	261,06	3,866	10
C	269,95	4,501	10

H_0: $\mu_A = \mu_B = \mu_C$

H_1: kaikki μ:t eivät samoja

$\bar{y}$= 260,76, n = 30, I = 3, $n_1 = n_2 = n_3 = 10$

$$SSW = \sum\sum (Y_{ij} - \bar{Y}_i)^2 = (n_1 - 1)s_1^2 + \cdots + (n_I - 1)s_I^2$$

$SSW = (10 - 1)5{,}977^2 + (10 - 1)3{,}866^2 + (10 - 1)4{,}501^2$
$= 638{,}36$

$$SSB = \sum n_i(\bar{Y}_i - \bar{Y})^2$$

$SSB = 10(251{,}28\text{-}260{,}76)^2 + 10(261{,}06\text{-}260{,}76)^2 +$
$10(269{,}95\text{-}260{,}76)^2 = 1744{,}17$

$MSB = SSB/(I-1)$

$MSB = 1744,17/2 = 872,08$

$MSW = SSW/(n-I)$

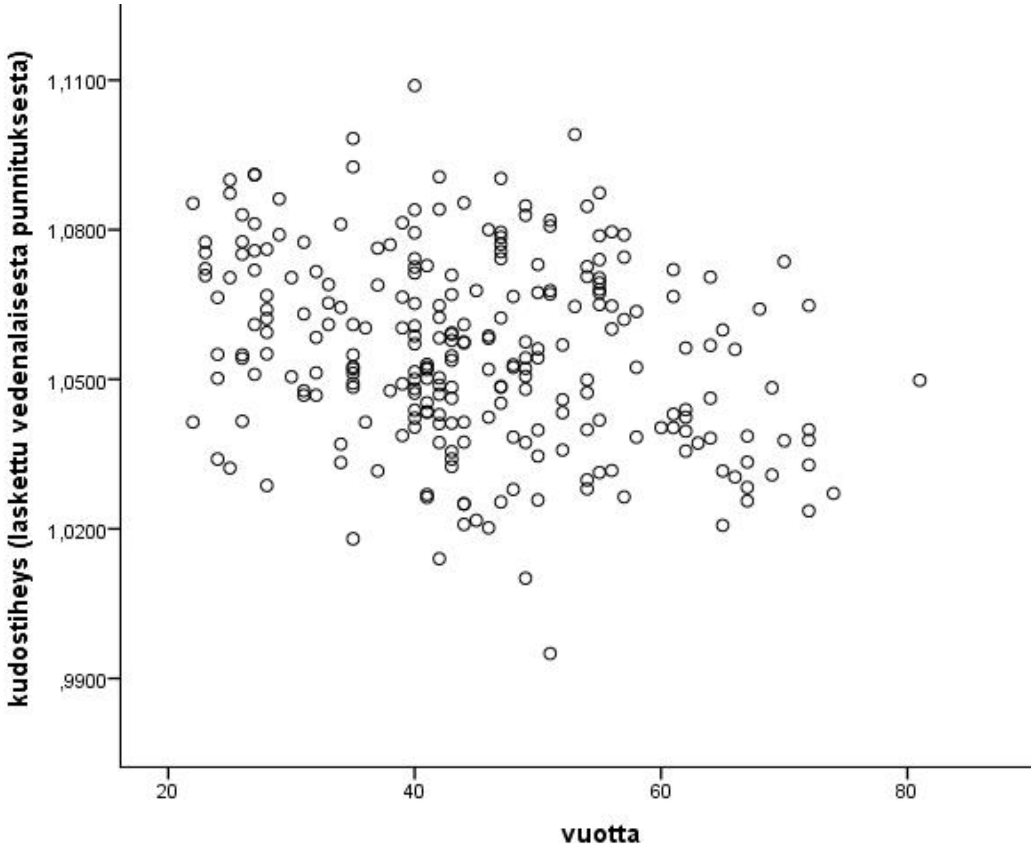
$MSW = 638,36/27 = 23,64$

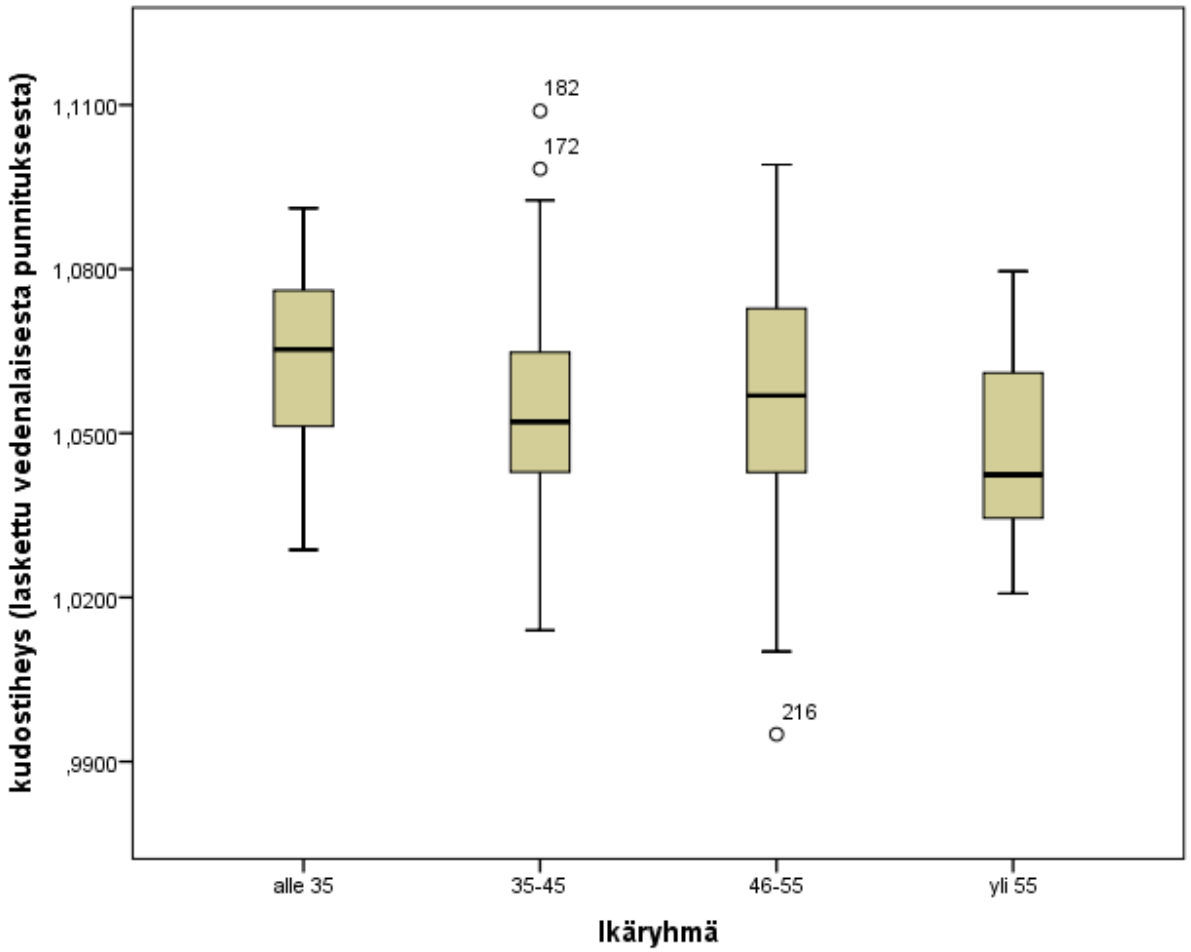
$F = MSB/MSW \sim F_{I-1,\ n-I}$, kun H_0 tosi

$F_{hav.} = 872,08/23,64 = 36,87 > F_{0,01;2,27} = 5,49$, joten H_0 hylätään 1 %:n riskitasolla. Päätellään odotusarvoissa olevan eroja. Voidaan sanoa, että p-arvo = $P(F_{2,27} > 36,87) < 0,01$.

Esim. Miehillä iän vaikutus kudostiheyteen

Aineisto rasvaprosentti.sav sivulta
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/




kudostiheys (laskettu vedenalaisesta punnituksesta)
1,1100
1,0800
1,0500
1,0200
,9900
182
172
216
alle 35
35-45
46-55
yli 55
Ikäryhmä

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1: kaikki odotusarvot eivät samoja

ANOVA

kudostiheys (laskettu vedenalaisesta punnituksesta)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	,007	3	,002	7,247	,000
Within Groups	,084	248	,000		
Total	,091	251			

Koska p-arvo <0,001, niin H_0 hylätään, päätellään eroja olevan. Monivertailusta huomataan, että kaikkien ikäryhmien välillä ei kuitenkaan ole eroja.

Multiple Comparisons

Dependent Variable: kudostiheys (laskettu vedenalaisesta punnituksesta)
Bonferroni

(I) Ikäryhmä	(J) Ikäryhmä	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
alle 35	35-45	,0098402*	,0032131	,015	,001295	,018386
	46-55	,0072854	,0033748	,191	-,001690	,016261
	yli 55	,0168150*	,0036783	,000	,007032	,026598
35-45	alle 35	-,0098402*	,0032131	,015	-,018386	-,001295
	46-55	-,0025547	,0029992	1,000	-,010531	,005422
	yli 55	,0069748	,0033371	,226	-,001900	,015850
46-55	alle 35	-,0072854	,0033748	,191	-,016261	,001690
	35-45	,0025547	,0029992	1,000	-,005422	,010531
	yli 55	,0095295*	,0034930	,041	,000240	,018820
yli 55	alle 35	-,0168150*	,0036783	,000	-,026598	-,007032
	35-45	-,0069748	,0033371	,226	-,015850	,001900
	46-55	-,0095295*	,0034930	,041	-,018820	-,000240

*. The mean difference is significant at the 0.05 level.

Test of Homogeneity of Variances

kudostiheys (laskettu vedenalaisesta punnituksesta)

Levene Statistic	df1	df2	Sig.
1,254	3	248	,291

Populaatioiden varianssit voitiin olettaa samoiksi (H_0: $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$ hyväksytään, koska p-arvo 0,291>0,05), joten varianssianalyysin käyttö sallittua.

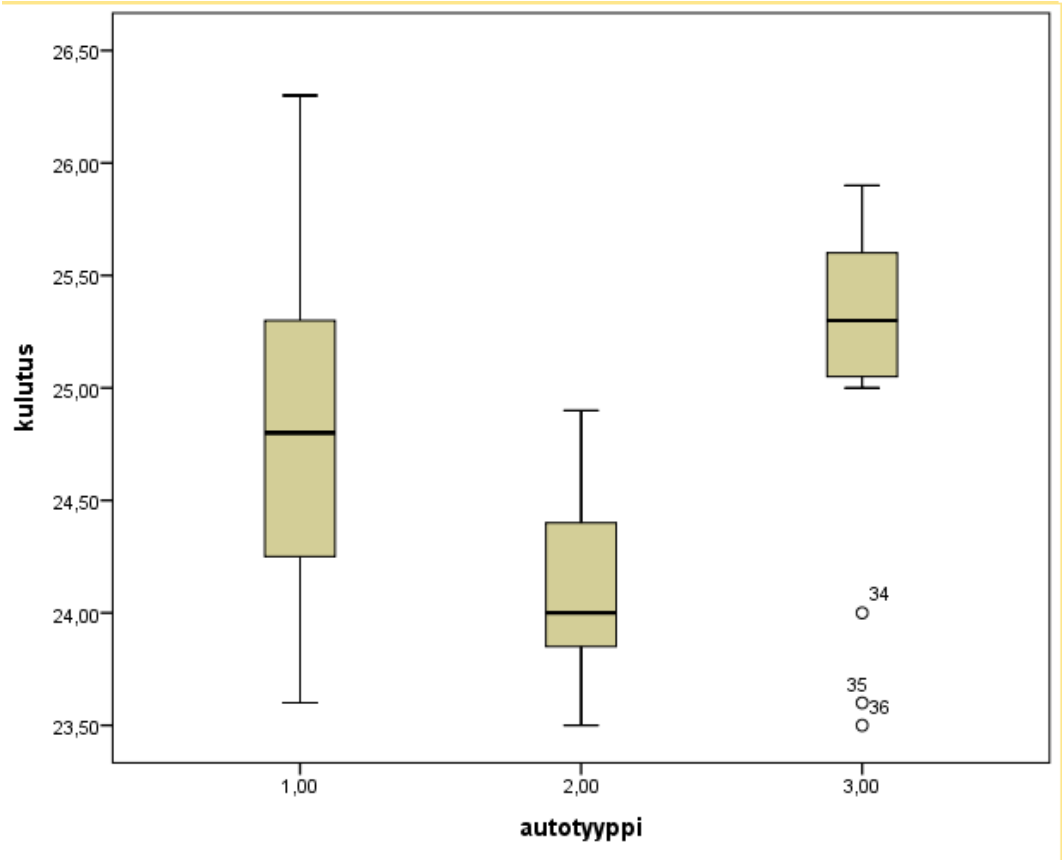
Varianssianalyysi nettilaskurilla:
http://vassarstats.net/ - > ANOVA ->
http://vassarstats.net/anova1u.html

2.2 Kaksisuuntainen varianssianalyysi

Esim. Tutkitaan kolmen autotyypin polttoaineen kulutusta (kulutus = mailit/gallona) huomioiden kuljettajan ikä (5 ikäryhmää), aineisto http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/Aineistoja/autotNB2va.sav

Tehdään aluksi yksisuuntaiset varianssianalyysit.

y = kulutus
x = autotyyppi



Test of Homogeneity of Variances

KULUTUS

Levene Statistic	df1	df2	Sig.
2,302	2	42	,113

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2$

H_0 hyväksytään, koska $p = 0,113 > 0,05$

ANOVA

KULUTUS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	7,156	2	3,578	7,186	,002
Within Groups	20,912	42	,498		
Total	28,068	44			

$H_0: \mu_1 = \mu_2 = \mu_3$

H_0 hylätään, koska $p < 0,01$, kaikki odotusarvot eivät samoja

Post Hoc Tests

Multiple Comparisons

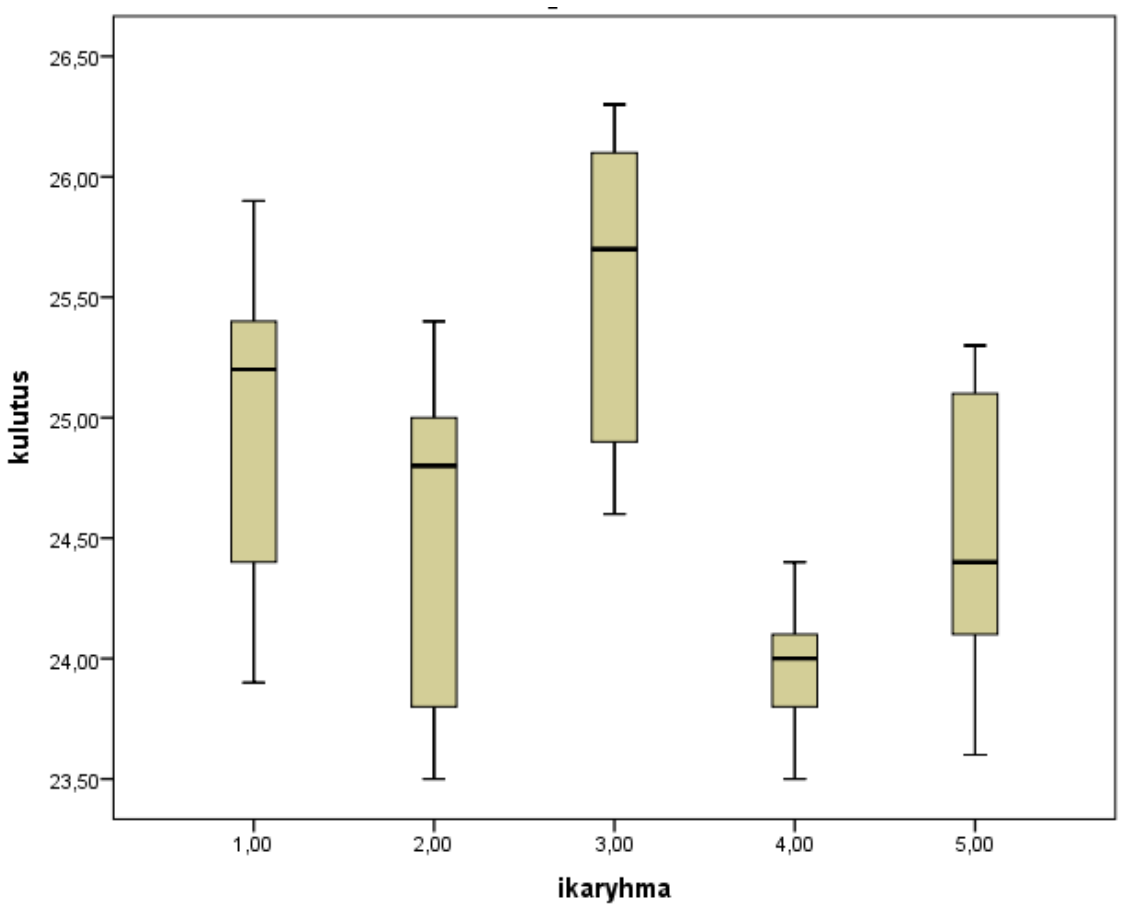
Dependent Variable: KULUTUS
Bonferroni

Autotyyppien 1 ja 2 sekä 2 ja 3 odotusarvot erisuuret

(I) AUTO	(J) AUTO	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1,00	2,00	,7000*	,25766	,029	,0575	1,3425
	3,00	-,2400	,25766	1,000	-,8825	,4025
2,00	1,00	-,7000*	,25766	,029	-1,3425	-,0575
	3,00	-,9400*	,25766	,002	-1,5825	-,2975
3,00	1,00	,2400	,25766	1,000	-,4025	,8825
	2,00	,9400*	,25766	,002	,2975	1,5825

*. The mean difference is significant at the .05 level.

y= kulutus
x = ikäryhmä



Oneway

Test of Homogeneity of Variances

KULUTUS

Levene Statistic	df1	df2	Sig.
2,016	4	40	,111

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \sigma_5^2$

H_0 hyväksytään, koska $p = 0,111 > 0,05$

ANOVA

KULUTUS

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	13,148	4	3,287	8,812	,000
Within Groups	14,920	40	,373		
Total	28,068	44			

$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_0 hylätään, päätellään, että kaikki odotusarvot ei samoja

Post Hoc Tests

Multiple Comparisons

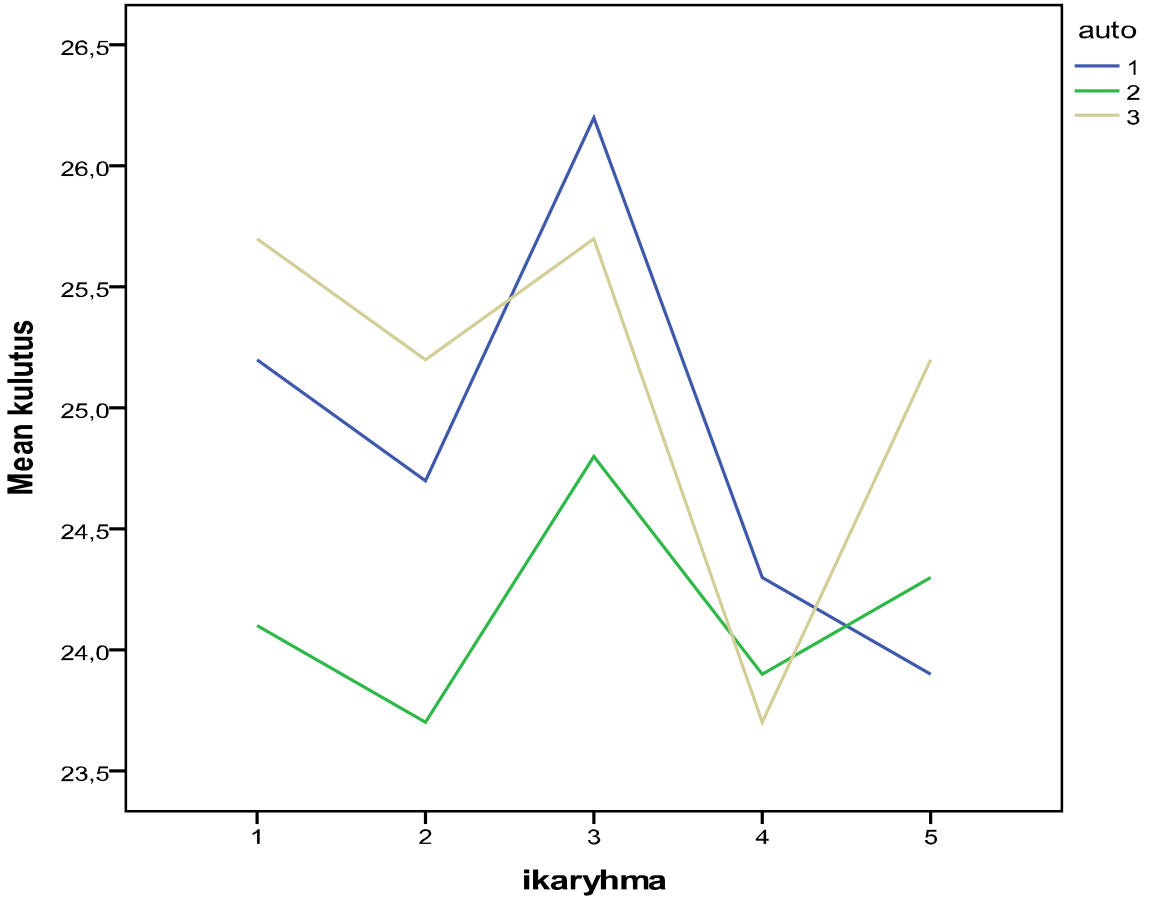
Dependent Variable: KULUTUS
Bonferroni

(I) IKARYHMA	(J) IKARYHMA	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1,00	2,00	,4667	,28790	1,000	-,3887	1,3221
	3,00	-,5667	,28790	,560	-1,4221	,2887
	4,00	1,0333*	,28790	,009	,1779	1,8887
	5,00	,5333	,28790	,713	-,3221	1,3887
2,00	1,00	-,4667	,28790	1,000	-1,3221	,3887
	3,00	-1,0333*	,28790	,009	-1,8887	-,1779
	4,00	,5667	,28790	,560	-,2887	1,4221
	5,00	,0667	,28790	1,000	-,7887	,9221
3,00	1,00	,5667	,28790	,560	-,2887	1,4221
	2,00	1,0333*	,28790	,009	,1779	1,8887
	4,00	1,6000*	,28790	,000	,7446	2,4554
	5,00	1,1000*	,28790	,005	,2446	1,9554
4,00	1,00	-1,0333*	,28790	,009	-1,8887	-,1779
	2,00	-,5667	,28790	,560	-1,4221	,2887
	3,00	-1,6000*	,28790	,000	-2,4554	-,7446
	5,00	-,5000	,28790	,901	-1,3554	,3554
5,00	1,00	-,5333	,28790	,713	-1,3887	,3221
	2,00	-,0667	,28790	1,000	-,9221	,7887
	3,00	-1,1000*	,28790	,005	-1,9554	-,2446
	4,00	,5000	,28790	,901	-,3554	1,3554

*. The mean difference is significant at the .05 level.

Eroja ikäryhmien 1 & 4, 2 & 3, 3 & 4, 3 & 5 välillä

Kulutuksen ehdolliset keskiarvot ryhmitellen sekä ikäryhmän että autotyypin mukaan



Nyt

y= kulutus

x_1 = autotyyppi

x_2 = ikäryhmä

Suoritetaan kaksisuuntainen varianssianalyysi. Halutaan selvittää miten autotyyppi ja ikäryhmä yhdessä vaikuttavat kulutukseen. Tutkitaan autotyypin ikäryhmästä riippumatonta vaikutusta (omavaikutusta), ikäryhmän autotyypistä riippumatonta vaikutusta (omavaikutusta) sekä autotyypin ja ikäryhmän yhdysvaikutusta.

Ks.
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=21

Tests of Between-Subjects Effects

Dependent Variable: kulutus

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	26,908[a]	14	1,922	49,707	,000
Intercept	27468,872	1	27468,872	710401,862	,000
auto	7,156	2	3,578	92,534 F_A	,000
ikaryhma	13,148	4	3,287	85,009 F_B	,000
auto * ikaryhma	6,604	8	,826	21,349 F_{AB}	,000
Error	1,160	30	,039		
Total	27496,940	45			
Corrected Total	28,068	44			

a. R Squared = ,959 (Adjusted R Squared = ,939)

yhtysvaikutus

H_0: ei yhdysvaikutusta

H_1: on -"-

$F_{hav} = 21,349$, $p < 0,001$

H_0 hylätään

Päätellään: ikäryhmittäin kuljettajien väliset erot erilaiset eri autotyypeillä. Myös molemmilla selittäjillä on omavaikutusta (p-arvot <0,001).

SPSS-ohjeet

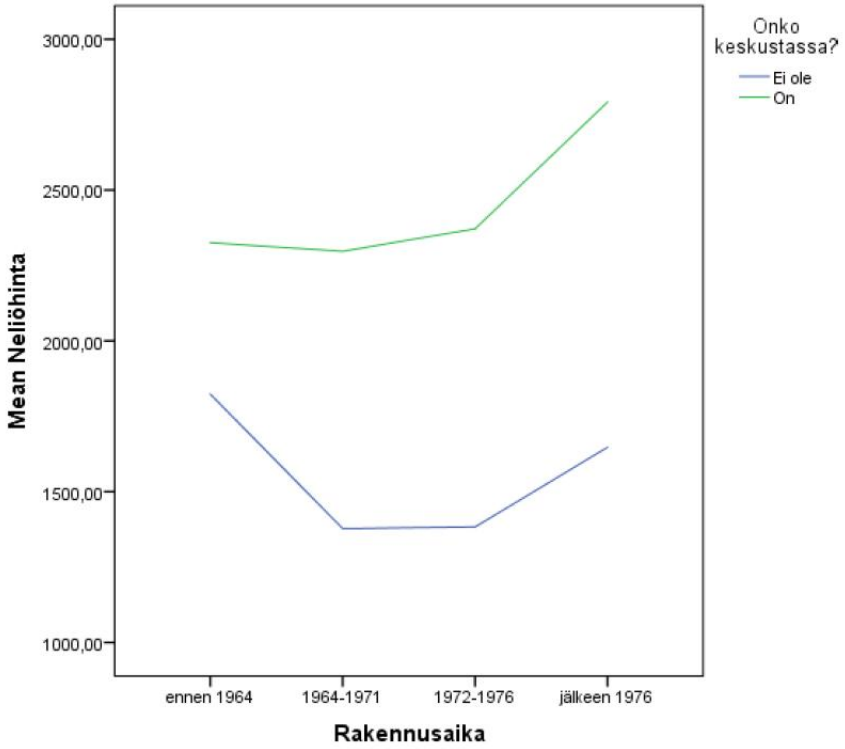
Ehdolliset keskiarvot graafisesti

Graphs-> Line-> Multiple-> Variable -> kulutus-> Category Axis-> ikaryhma -> Define line by->auto

2-VA

General Linear Model -> Univariate -> Dependent -> kulutus-> Fixed Factors ->auto, ikaryhma, Model -> auto, ikaryhma, interaction...

Esim. Rakennusajan ja sijainnin vaikutus keskineliöhintaan, SPSS-monisteen
http://www.uta.fi/sis/reports/index/R55_2017.pdf
esimerkki 19



Tests of Between-Subjects Effects

Dependent Variable: Neliöhinta

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	51473967,7[a]	7	7353423,957	67,519	,000
Intercept	699593758,9	1	699593758,9	6423,680	,000
Onko keskustassa?	34440038,88	1	34440038,88	316,229	,000
Rakennusaika	4400240,537	3	1466746,846	13,468	,000
Onko keskustassa * Rakennusaika	2298538,654	3	766179,551	7,035	,000
Error	24068790,94	221	108908,556		
Total	907041110,3	229			
Corrected Total	75542758,64	228			

a. R Squared = ,681 (Adjusted R Squared = ,671)

Aineisto
http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Asunnot_2006.sav
sivulta https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

MTTTA1 Tilastomenetelmien perusteet
Luento 22.1.2019

Luku 3

χ^2-yhteensopivuus- ja riippumattomuustestit

3.1 χ^2-yhteensopivuustesti

H_0: otos peräisin tietystä jakaumasta

H_1: otos ei peräisin tästä jakaumasta

Esim. H_0: otos peräisin normaalijakaumasta
H_0: otos peräisin tasajakaumasta

Esim. Eräällä kurssilla opiskelijat generoivat satunnaislukuja vastaamalla kysymyksiin:

1. Ravistele päätäsi ja arvo yksi kokonaisluku

	1	2	3	4	5	6	7	8	9	10	
heittotulos:	*2*	*3*	*6*	*3*	*4*	*7*	*6*	*5*	*3*	*1*	*n=40*

2. Ravistele päätäsi uudelleen ja arvo yksi kokonaisluku

	1	2	3	4	5	6	7	8	9	10	
heittotulos:	*1*	*2*	*9*	*7*	*5*	*4*	*2*	*5*	*4*	*1*	*n=40*

3. Ravistele päätäsi ja heitä rahaa

	klaava	kruuna	
heittotulos:	*21 (52,5 %)*	*19*	*n=40*

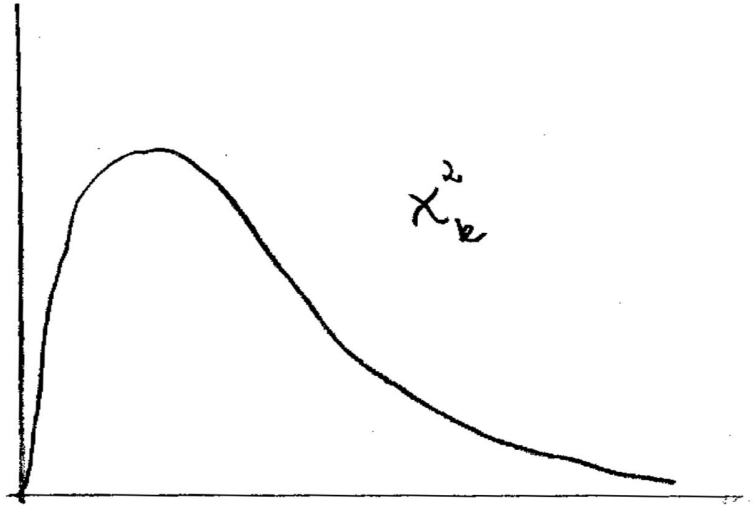
4. Ravistele päätäsi uudelleen ja heitä rahaa

	klaava	kruuna	
heittotulos:	*13 (32,5 %)*	*27*	*n=40*

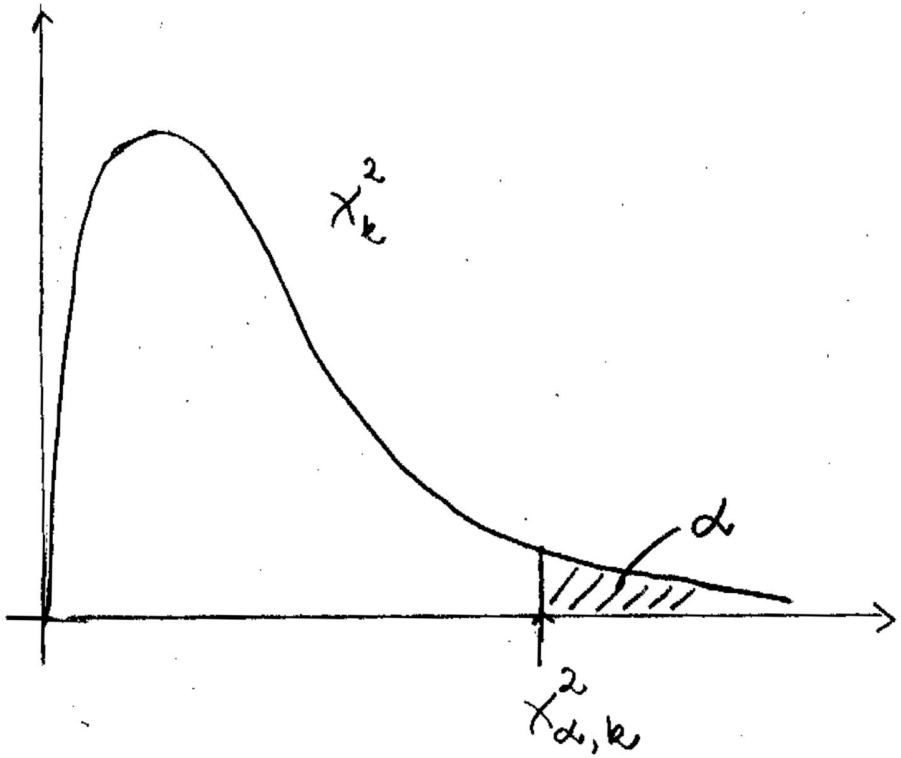
Voidaanko ajatella, että ensimmäinen kokonaisluvun valinta on otos diskreetistä tasajakaumasta? Jos olisi, niin jokainen numero olisi esiintynyt 4 kertaa. Voidaanko ajatella, että rahanheiton tulos on otos jakaumasta, jossa klaavoja 50 %? Jos olisi, niin klaavoja pitäisi olla 20 ja kruunia 20.

Olkoot riippumattomat Z_i ~N(0, 1), i = 1, …, k. Tällöin $Z_1^2 + Z_2^2 + \ldots + Z_k^2$ noudattaa nk. χ^2 – jakaumaa vapausastein k, merkitään χ_k^2. Tällöin $E(\chi_k^2) = k$, $Var(\chi_k^2) = 2k$.

χ^2 – jakauman tiheysfunktion kuvaaja, muoto riippuu vapausasteista



Määritellään $\chi^2_{\alpha,k}$ siten, että $P\left(\chi^2_k \geq \chi^2_{\alpha,k}\right) = \alpha$.



Näitä arvoja on taulukoitu,
ks. http://www.sis.uta.fi/tilasto/mttta1/kevat2019/chi.pdf

Tarkastellaan muuttujan frekvenssijakaumaa. Oletetaan, että jakaumassa on k kappaletta luokkia ja näiden luokkien frekvenssit f_1, f_2, ..., f_k.

Testataan sitä, ovatko havaitut frekvenssit sopusoinnussa H_0:n mukaisten nk. teoreettisten eli odotettujen frekvenssien e_1, e_2, ..., e_k kanssa.

Jos

H_0: otos peräisin tietystä jakaumasta

on tosi, niin

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2_{k-1}.$$

H_0 hylätään riskitasolla α, jos $\chi^2_{havaittu} > \chi^2_{\alpha,k-1}$.

Testiä voidaan käyttää, jos kaikki teoreettiset frekvenssit ovat > 1 ja enintään 20 % < 5.

Esim. Rahanheitto

H_0: Otos peräisin jakaumasta, jossa klaavoja ja kruunia yhtä paljon

1. rahanheitto

	f_i	e_i
klaavoja	21	20
kruunia	19	20

$$\chi^2_{havaittu} = \frac{(21-20)^2}{20} + \frac{(19-20)^2}{20} = 0{,}1$$

$\chi^2_{0.05,2-1}$ = 3,84 > $\chi^2_{havaittu}$ = 0,1 , H_0 hyväksytään 5%:n riskitasolla. Voidaan siis ajatella, että rahanheitto tehty satunnaisesti.

2. rahanheitto

	f_i	e_i
klaavoja	13	20
kruunia	27	20

$\chi^2 = \frac{(13-20)^2}{20} + \frac{(27-20)^2}{20} = 4,9$

Koska

$\chi^2_{0.05,2-1}$ = 3,84 < $\chi^2_{havaittu}$ = 4,9 < $\chi^2_{0.025,2-1}$ = 5,02 ,

niin 0,025 < p-arvo < 0,05.

Esim. Ystäväsi väittää, että suomalaisista 10 % on vasenkätisiä. Tutkit asiaa ja valitset satunnaisesti 400 suomalaista, joista 56 on vasenkätisiä. Uskotko ystäväsi väitteen?

H_0: 10 % suomalaisista on vasenkätisiä

	f_i	e_i
vasenkätisiä	56	0,1·400 = 40
ei-vasenkätisiä	344	0,9·400 = 360

$$\chi^2_{havaittu} = \frac{(56-40)^2}{40} + \frac{(344-360)^2}{360} = 7{,}11$$

$\chi^2_{0,01,1} = 6{,}63$
$\chi^2_{0,005,1} = 7{,}88$

H_0 hylätään 1 %:n riskitasolla, mutta ei 0,5 %:n riskitasolla, siis 0,005 < p-arvo < 0,01.

Laskuri http://vassarstats.net/csfit.html ja p-arvon arviointi http://vassarstats.net/csqsamp.html, p ≈ 0,008151

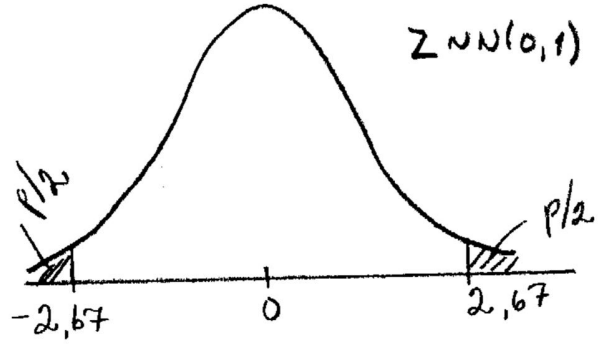
Toisin

$H_0: \pi = 10$

$H_1: \pi \neq 10$

$$z = \frac{14 - 10}{\sqrt{10 \cdot 90/400}} = 2{,}67$$

p-arvo = $2(1-\Phi(2{,}67)) = 2(1-0{,}9962) = 0{,}0076$



Jos χ^2-yhteensopivuustestissä luokkien lukumäärä on kaksi, niin $\chi^2 = Z^2$. Edellisessä esimerkissä $7{,}11 \approx 2{,}67^2$.

Esim. 3.1.4 Nopanheitto,
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=26

H_0: Otos peräisin Tasd(1, 6):sta

silmäluku	f_i	e_i
1	8	122/6 = 20,3
2	5	122/6
3	17	122/6
4	27	122/6
5	26	122/6
6	39	122/6

$$\chi^2_{havaittu} = \frac{(8 - 20{,}3)^2}{20{,}3} + \cdots + \frac{(39 - 20{,}3)^2}{20{,}3} = 40{,}6$$

$$> \chi^2_{0.005,6-1} = 16{,}75$$

H_0 hylätään, nopanheitto ei ole tapahtunut satunnaisesti.

Esim. 3.1.2 Asiakkaiden laskujen maksutavat, http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=25

H_0: ei tapahtunut muutosta

H_1: on tapahtunut muutos

	f_i	e_i	
ajoissa	287	0,8x400	= 320
1 kk myöhässä	49	0,1x400	= 40
2 kk myöhässä	30	0,06x400	= 24
yli 2 kk myöhässä	34	0,04x400	= 16

$$\chi^2_{hav.} = \frac{(287 - 320)^2}{320} + \cdots + \frac{(34 - 16)^2}{16} = 27{,}58 > \chi^2_{0.005,4-1} = 12{,}84$$

Päätellään muutosta tapahtuneen.

Laskuri http://vassarstats.net/csfit.html

Pelkän p-arvon määrittäminen
http://onlinestatbook.com/2/calculators/chi_square_prob.html

Esim. 3.1.5 Onko painoindeksi normaalisti jakautunut?
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=26

H_0: Otos peräisin $N(25.58, 4.66^2)$:sta

Painoindeksi	frekv.	odotettu frekv.
alle 20,1	9	11,5 = e_1
20,1-21,4	15	6,3
21,4-25,5	26	30,3
25,5-28,5	23	23,6
28,5-32,2	15	18,1
yli 32,2	9	7,5
	97	97

$e_1 = 97 \cdot P(X \leq 20{,}1) = 97 \cdot \Phi((20{,}1-25{,}58)/4{,}66)$
$= 97 \cdot \Phi(-1{,}18) = 97 \cdot (1- \Phi(1{,}18)) = 97 \cdot 0{,}119 =$
$11{,}5$

Vastaavalla tavalla lasketaan muidenkin luokkien odotetut frekvenssit.

Saadaan

$$\chi^2_{havaittu} = \frac{(9 - 11{,}5)^2}{11{,}5} + \cdots + \frac{(9 - 7{,}5)^2}{7{,}5} = 13{,}94$$

$> \chi^2_{0.005,6-2-1} = 12{,}84$

Päätellään, että otos ei peräisin normaalijakaumasta.

Huom! Vapausasteet pienenevät estimoitujen parametrien verran.

Laskurin http://vassarstats.net/csfit.html antama tulos, vapausasteissa ei huomioitu estimointia.

Chi-Square "Goodness of Fit" Test

The logic and computational details of chi-square tests are described in Chapter 8 of Concepts and Applications.

This unit will calculate the value of chi-square for a one-dimensional "goodness of fit" test, for up to 8 mutually exclusive categories labeled A through H. To enter an observed cell frequency, click the cursor into the appropriate cell, then type in the value. Expected values can be entered as either frequencies or proportions. If you enter the expected values as proportions, the entries can take the form of either decimal fractions such as .25, or common fractions such as 1/4. Whenever possible, it is better to enter common fractions rather than rounded decimal fractions: 1/3 rather than .3333; 1/6 rather than .1667; and so forth.

When all observed and expected values have been entered, click the «Calculate» button. To perform a new analysis with a new set of data, click the «Reset» button.

Cate-gory	Observed Frequency	Expected Frequency	Expected Proportion	Percentage Deviation	Standardized Residuals
A	9	11.5	0.1185567	-21.74%	-0.74
B	15	6.3	0.0649484	+138.1%	+3.47
C	26	30	0.3092783	-13.33%	-0.73
D	23	23.6	0.2432989	-2.54%	-0.12
E	15	18.1	0.1865979	-17.13%	-0.73
F	9	7.5	0.0773195	+20%	+0.55
G				----	----
H				----	----

Sums:

Observed Frequencies: 97

Expected Frequencies: 97

Expected Proportions: 1.0

Reset Calculate

[Note that for df=1, the calculated value of chi-square is corrected for continuity.]

[For df=1, this is the uncorrected value of chi-square.]

chi-square = 13.94

df = 5

P = 0.016

[P is non-directional]

MTTTA1 Tilastomenetelmien perusteet
Luento 24.1.2019

Kertausta ja täydennystä χ^2-yhteensopivuustestistä

H_0: otos peräisin tietystä jakaumasta

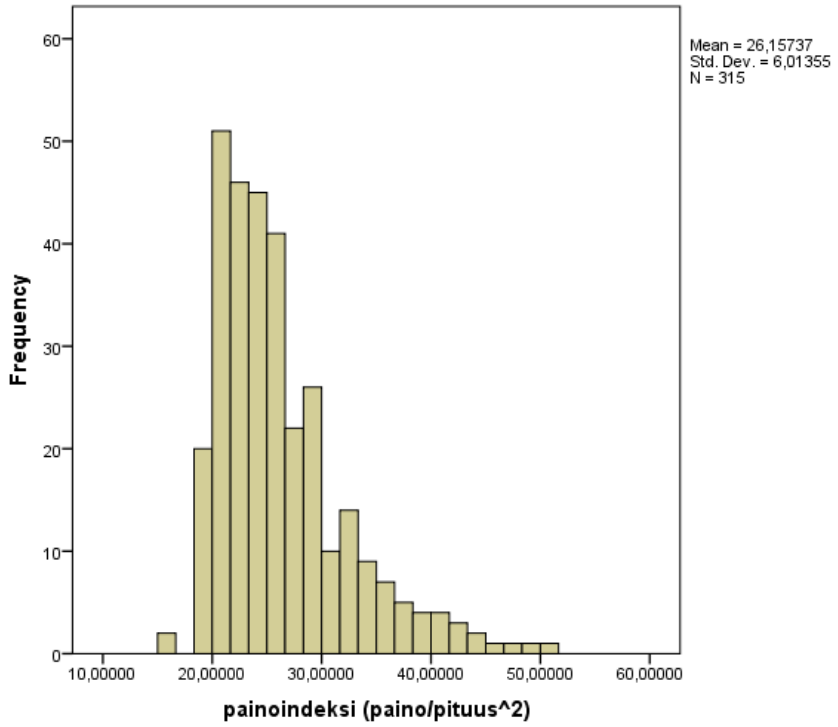
H_1: otos ei peräisin tästä jakaumasta

Jos H_0: otos peräisin tietystä jakaumasta on tosi, niin

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - e_i)^2}{e_i} \sim \chi^2_{k-1}.$$

Esim. Plasma-aineisto, sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = painoindeksi (paino/pituus^2)



H_0: Otos peräisin $N(26.16, 6.01^2)$:sta

Vaihtoehtoisia testejä normaalisuuden testaamiseksi:

SPSS -> Analyze -> Descriptive Statistics -> Explore ...Plots -> Normality plots with tests

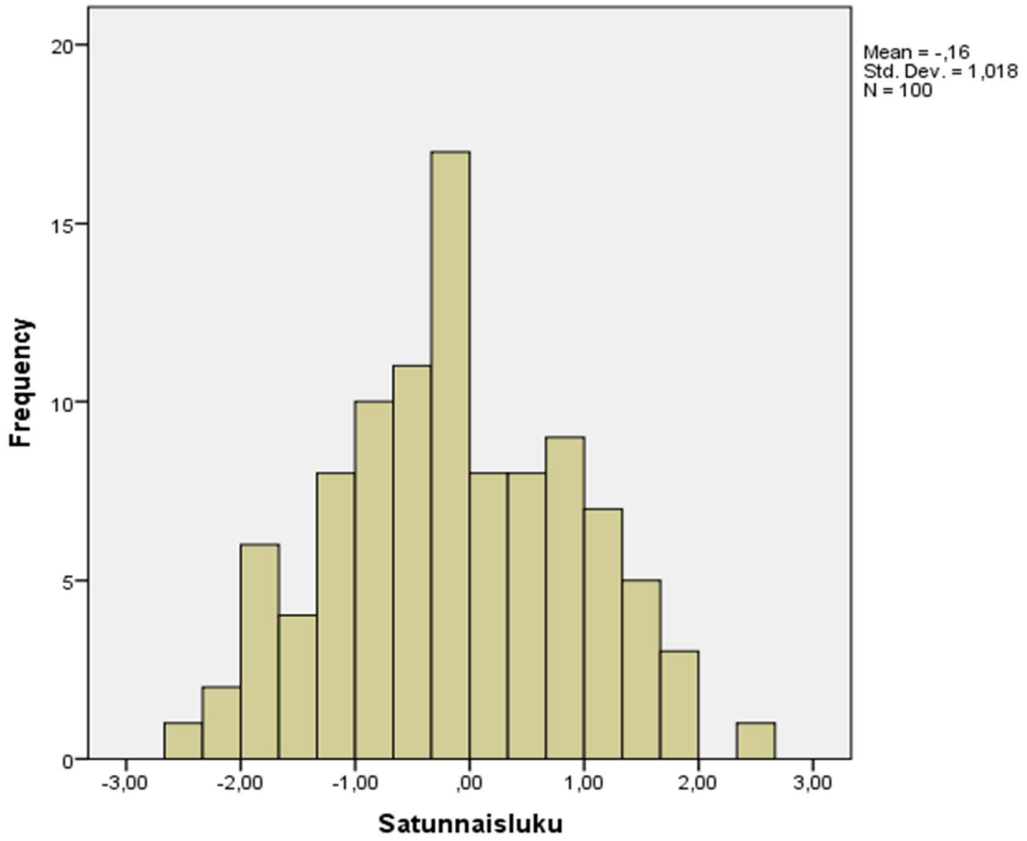
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
painoindeksi (paino/pituus^2)	,127	315	,000	,889	315	,000

a. Lilliefors Significance Correction

H_0 hylätään molemmilla testeillä, koska p-arvot < 0,001. Otos ei peräisin normaalijakaumasta.

Esim. Generoitu 100 lukua N(0, 1):stä, SPSS-funtio RV.NORMAL(0,1)



H_0: Otos peräisin $N(-0{,}16,\ 1{,}018^2)$:sta

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Satunnaisluku	,049	100	,200*	,990	100	,694

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

H_0 hyväksytään molemmilla testeillä, koska p-arvot > 0,05. Satunnaislukugeneraattori OK.

3.2 χ^2-riippumattomuustesti

Ristiintaulukon perusteella riippumattomuuden testaaminen

H_0: X ja Y ovat riippumattomia

H_1: X ja Y ovat riippuvia

Esim. Tampereella myydyt pienet asunnot, aineisto http://www.sis.uta.fi/tilasto/tiltp_aineistoja/Tre_myydyt_asunnot_2009.sav sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Kunto * Sijainti Crosstabulation

% within Sijainti

		Sijainti			
		Keskustassa	Alle 5 km keskustasta	Yli 5 km keskustasta	Total
Kunto	Hyvä	42,9%	29,8%	30,0%	32,7%
	Tyydyttävä	38,1%	42,6%	53,3%	44,9%
	Huono	19,0%	27,7%	16,7%	22,4%
Total		100,0%	100,0%	100,0%	100,0%

H_0: Kunto ja sijainti ovat riippumattomia

H_1: Kunto ja sijainti ovat riippuvia

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,725[a]	4	,605
Likelihood Ratio	2,667	4	,615
Linear-by-Linear Association	,134	1	,714
N of Valid Cases	98		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,71.

H_0 hyväksytään, koska p-arvo on 0,605 > 0,05.

Tarkastellaan yleisesti ristiintaulukkoa

		x				
		1	2	$\dots$	J	
y	1	f_{11}	f_{12}	$\dots$	f_{1J}	$f_{1\cdot}$
	2	f_{21}	f_{22}	$\dots$	f_{2J}	$f_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	f_{I1}	f_{I2}	$\dots$	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	$\dots$	$f_{\cdot J}$	n

Määritetään ristiintaulukkoon teoreettiset frekvenssit e_{ij} siten, että oletetaan H_0: X ja Y riippumattomia on tosi. Tällöin oltava

$$\frac{e_{ij}}{f_{\cdot j}} = \frac{f_{i\cdot}}{n} \qquad \text{eli} \qquad e_{ij} = \frac{f_{i\cdot} f_{\cdot j}}{n}$$

Jos H_0 on tosi, niin

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} \sim \chi^2 (I-1)(J-1)$$

Nyt H_0 hylätään riskitasolla α, jos

$$\chi^2_{havaittu} > \chi^2_{\alpha,(I-1)(J-1)}$$

Jos I =2 ja J = 2 (nelikenttä), niin testisuure voidaan laskea myös kaavalla

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{\cdot 1}f_{\cdot 2}f_{1\cdot}f_{2\cdot}}$$

Testiä voidaan käyttää:

a) (I-1)(J-1) = 1
   - n >40
   - 20 ≤ n ≤ 40

   kaikkien teoreettisten frekvenssien oltava ≥ 5.

b) (I-1)(J-1) > 1
   - kaikkien teoreettisten frekvenssien oltava > 1 ja enintään 20 % saa olla alle 5.

Esim. Edellisestä ristiintaulukosta testisuureen laskeminen.

Kunto * Sijainti Crosstabulation

			Sijainti			
			Keskustassa	Alle 5 km keskustasta	Yli 5 km keskustasta	Total
Kunto	Hyvä	Count	9	14	9	32
		Expected Count	6,9 $= e_{11}$	15,3 $= e_{12}$	9,8	32,0
		% within Sijainti	42,9%	29,8%	30,0%	32,7%
	Tyydyttävä	Count	8	20	16	44
		Expected Count	9,4	21,1	13,5	44,0
		% within Sijainti	38,1%	42,6%	53,3%	44,9%
	Huono	Count	4	13	5	22
		Expected Count	4,7	10,6	6,7 $= e_{33}$	22,0
		% within Sijainti	19,0%	27,7%	16,7%	22,4%
Total		Count	21	47	30	98
		Expected Count	21,0	47,0	30,0	98,0
		% within Sijainti	100,0%	100,0%	100,0%	100,0%

$$e_{11} = \frac{32 \cdot 21}{98}, \; e_{12} = \frac{32 \cdot 47}{98}, \ldots, \; e_{33} = \frac{22 \cdot 30}{98}$$

Chi-Square Tests

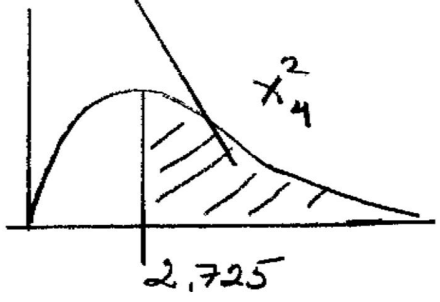
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	2,725[a]	4	,605
Likelihood Ratio	2,667	4	,615
Linear-by-Linear Association	,134	1	,714
N of Valid Cases	98		

a. 1 cells (11,1%) have expected count less than 5. The minimum expected count is 4,71.

H_0: ei riippuvuutta

$$X^2 = \frac{(9-6,9)^2}{6,9} + \dots + \frac{(5-6,7)^2}{6,7} = 2,725$$

(3-1)(3-1)



H_0 hyväksytään, koska p-arvo $> 0,05$.
Kunnon ja sijainnin välillä ei riippuvuutta.

Päätely taulukon avulla

$$\chi^2_{0,05,4} = 9,49 > 2,725 = \chi^2_{havaittu}$$, jotem H_0 hyväksytään 5%:n riskitasolla

Esim. Monisteesta Leppälä, R., Ohjeita tilastollisen tutkimuksen toteuttamiseksi IBM SPSS Statistics -ohjelmiston avulla, http://urn.fi/URN:ISBN:978-952-03-0501-7, esimerkki 13

Kyselylomake
http://www.sis.uta.fi/tilasto/tiltp3/kevat2003/Aineistoja/arviointi_lomake.pdf

Y = Opintojakson työläys
X = Opintosuunta

H_0: X ja Y ovat riippumattomia
H_1: X ja Y ovat riippuvia

Opintojakson työläys * opsuunta Crosstabulation

			opsuunta		
			hallinto	taloust	Total
Opintojakson työläys	työläs	Count	13	16	29
		Expected Count	8,5	20,5	29,0
		% within opsuunta	68,4%	34,8%	44,6%
	sopiva	Count	5	15	20
		Expected Count	5,8	14,2	20,0
		% within opsuunta	26,3%	32,6%	30,8%
	vähätöinen	Count	1	15	16
		Expected Count	4,7	11,3	16,0
		% within opsuunta	5,3%	32,6%	24,6%
Total		Count	19	46	65
		Expected Count	19,0	46,0	65,0
		% within opsuunta	100,0%	100,0%	100,0%

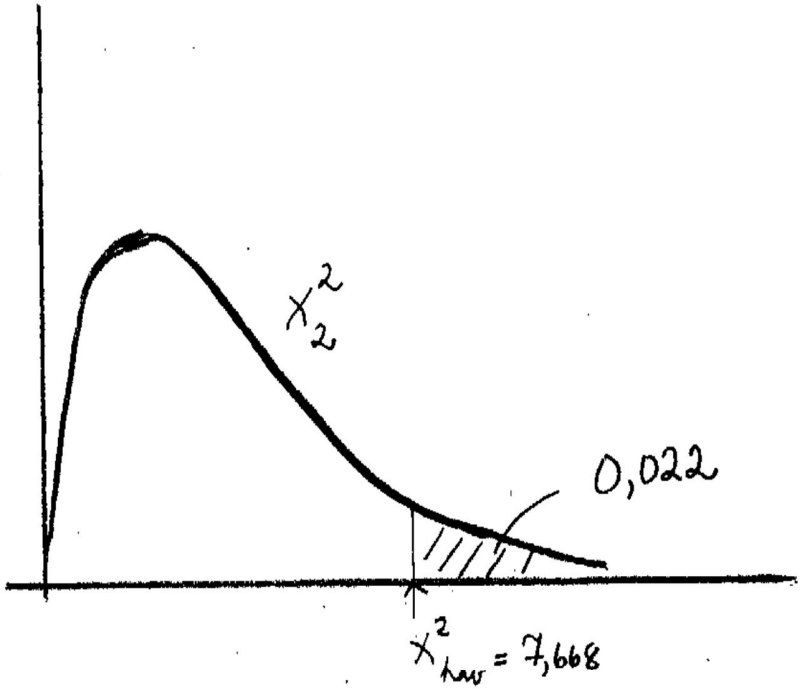
Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	7,668[a]	2	,022
Likelihood Ratio	8,680	2	,013
Linear-by-Linear Association	7,548	1	,006
N of Valid Cases	65		

a. 1 cells (16,7%) have expected count less than 5. The minimum expected count is 4,68.

Testin käyttöön liittyvät oletukset tällä luokituksella kunnossa, vain 16,7 % (1/6) odotetusta frekvensseistä alle 5 ja kaikki > 1.

Pienin riskitaso, jolla H_0 voidaan hylätä, on 0,022. Tätä suuremmilla riskeillä H_0 hylätään, pienemmillä hyväksytään.



$\chi^2_{0,01,2} = 9,21 > 7,668$, H_0 hyväksytään

$\chi^2_{0,025,2} = 7,38 < 7,668$, H_0 hylätään

MTTTA1 Tilastomenetelmien perusteet
Luento 29.1.2019

Kertausta ja täydennystä
χ^2-riippumattomuustesti

Ristiintaulukon perusteella riippumattomuuden testaaminen

H_0: X ja Y ovat riippumattomia

H_1: X ja Y ovat riippuvia

Ristiintaulukkoa

		x				
		1	2	...	J	
y	1	f_{11}	f_{12}	...	f_{1J}	$f_{1\cdot}$
	2	f_{21}	f_{22}	...	f_{2J}	$f_{2\cdot}$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	I	f_{I1}	f_{I2}	...	f_{IJ}	$f_{I\cdot}$
		$f_{\cdot 1}$	$f_{\cdot 2}$	...	$f_{\cdot J}$	n

Jos H_0 on tosi, niin

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(f_{ij}-e_{ij})^2}{e_{ij}} \sim \chi^2(I-1)(J-1) \qquad e_{ij} = \frac{f_{i\cdot}f_{\cdot j}}{n}.$$

Nyt H_0 hylätään riskitasolla α, jos

$$\chi^2_{hav.} > \chi^2_{\alpha,(I-1)(J-1)}$$

Jos I =2 ja J = 2 (nelikenttä), niin testisuure voidaan laskea myös kaavalla

$$\chi^2 = \frac{n(f_{11}f_{22} - f_{12}f_{21})^2}{f_{\cdot 1}f_{\cdot 2}f_{1\cdot}f_{2\cdot}}$$

Esim. 3.2.3 Naisten ja miesten tenttimenestyminen
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=31

	Miehet	Naiset	Yht.
Hylätty	34	15	49
Hyväksytty	59	23	82
Yht.	93	38	131

H_0: ei riippuvuutta

$$\chi^2_{hav.} = \frac{(34 \cdot 23 - 59 \cdot 15)^2 \cdot 131}{93 \cdot 38 \cdot 49 \cdot 82} = 0{,}09787 < 3{,}84 = \chi^2_{0{,}05;1}$$

H_0 hyväksytään, ei riippuvuutta.

Esim. Tutkimuksessa vertailtiin erään kasvaimen yleisyyttä kahdella rottalajilla A ja B. Valittiin satunnaisesti molemmista ryhmistä 100 samanikäistä rottaa. Rotat pidettiin samankaltaisissa olosuhteissa vuoden ajan. Vuoden seurannan jälkeen kasvain löytyi 25:ltä lajin A rotalta ja 15:ltä lajin B rotalta. Onko kasvaimen yleisyys samanlaista molemmilla lajeilla?

H_0: ei riippuvuutta

	Laji A	Laji B	
On kasvain	25	15	40
Ei kasvainta	75	85	160
	100	100	200

$$\chi^2_{hav.} = \frac{(25 \cdot 85 - 75 \cdot 15)^2 \cdot 200}{100 \cdot 100 \cdot 40 \cdot 160} = 3{,}125 < 3{,}84 = \chi^2_{0{,}05;1}$$

H_0 hyväksytään, yleisyys samanlaista.

$P(\chi^2_1 > 3{,}125) = 0{,}0771,$

ks.
http://onlinestatbook.com/2/calculators/chi_square_prob.html

Laskureita

http://www.physics.csbsju.edu/stats/contingency_NROW_NCOLUMN_form.html

http://vassarstats.net/newcs.html

Missä mennään?

Menetelmien valinnasta
http://www.fsd.uta.fi/menetelmaopetus/menetelma/menetelmatyypit.html

Luku 4

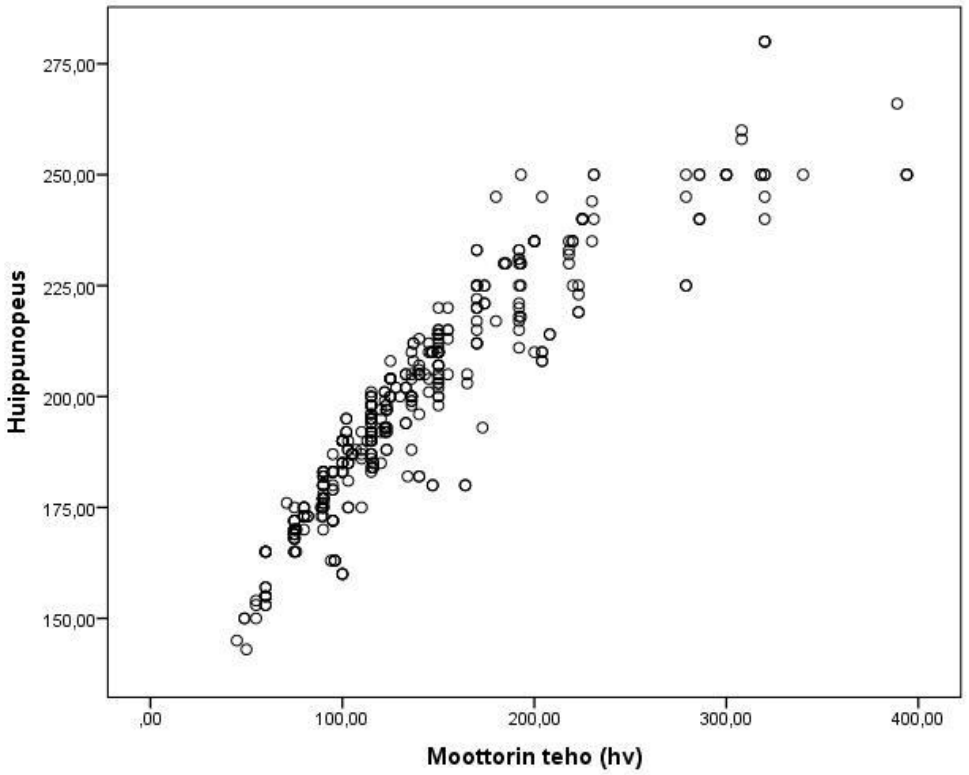
Regressioanalyysi

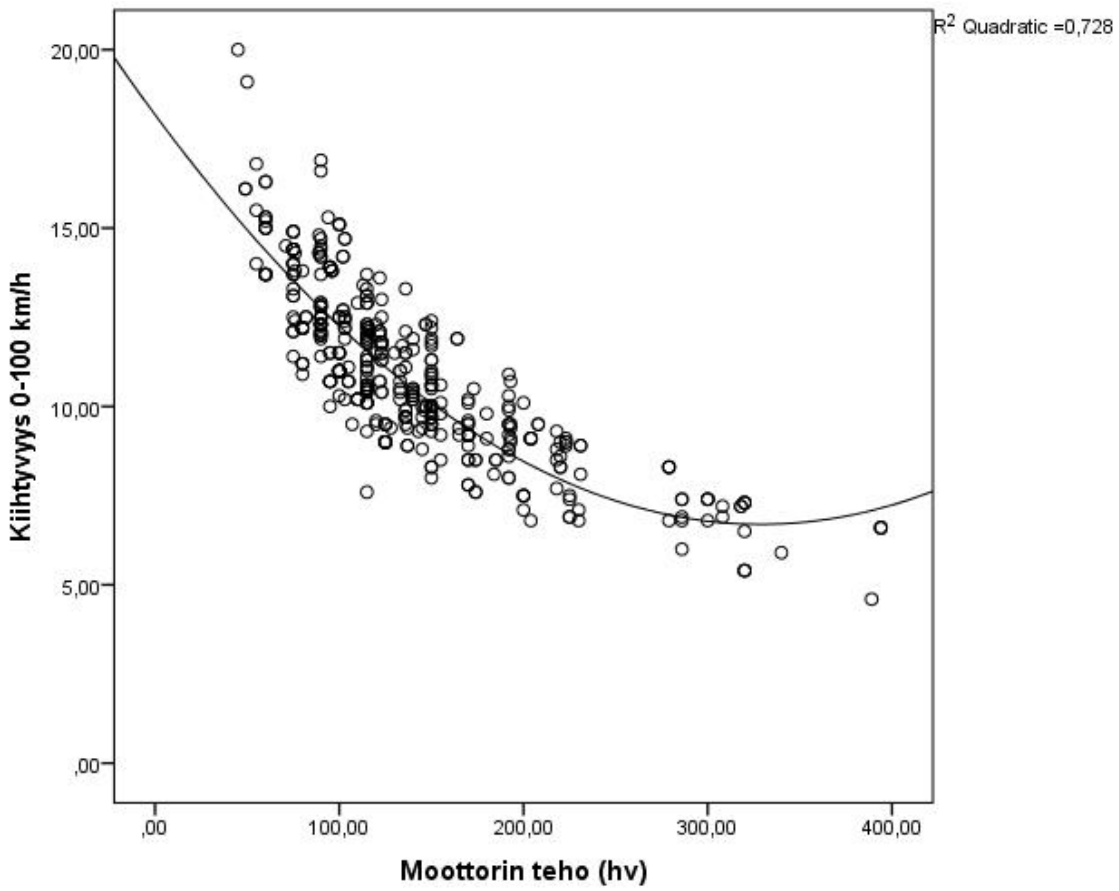
Voidaanko y:n vaihtelua selittää samanaikaisesti useammalla muuttujalla?

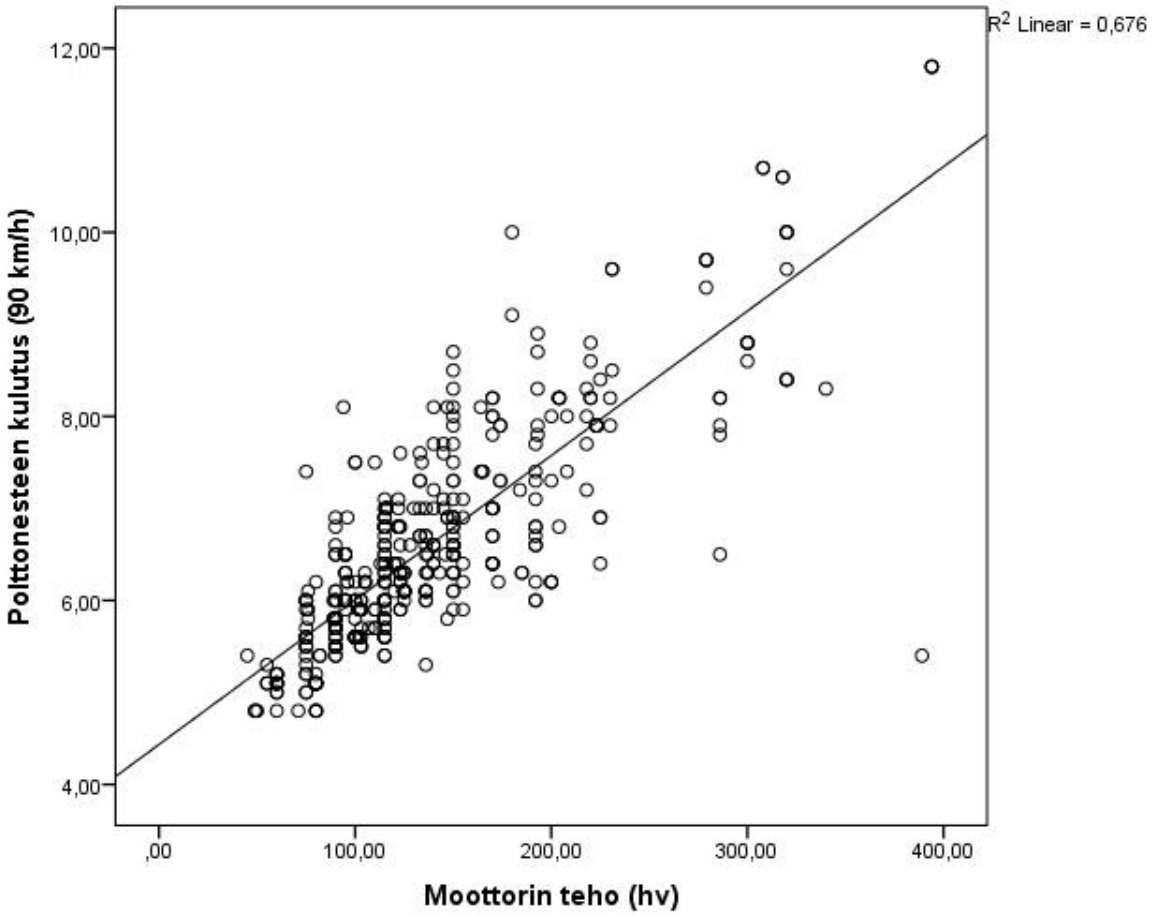
Voidaanko tätä riippuvuutta mallintaa?

Tarkastellaan tilanteita, joissa sekä selitettävä että selittäjät ovat kvantitatiivisia.

Esim. Erilaisia pisteparvia, tilastoyksikkönä auto




R2 Quadratic =0,728
20,00
15,00
10,00
5,00
,00
Kiihtyvyys 0-100 km/h
,00
100,00
200,00
300,00
400,00
Moottorin teho (hv)


R2 Linear = 0,676
12,00
10,00
8,00
6,00
4,00
Polttonesteen kulutus (90 km/h)
,00
100,00
200,00
300,00
400,00
Moottorin teho (hv)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,822[a]	,676	,675	,69270

a. Predictors: (Constant), Moottorin teho (hv)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	4,435	,081		54,873	,000
	Moottorin teho (hv)	,016	,001	,822	29,744	,000

a. Dependent Variable: Polttonesteen kulutus (90 km/h)

Malli

$$\text{Kulutus} = \beta_0 + \beta_1 \text{Teho} + \varepsilon$$

Estimoidaan mallin parametrit β_0 ja β_1.

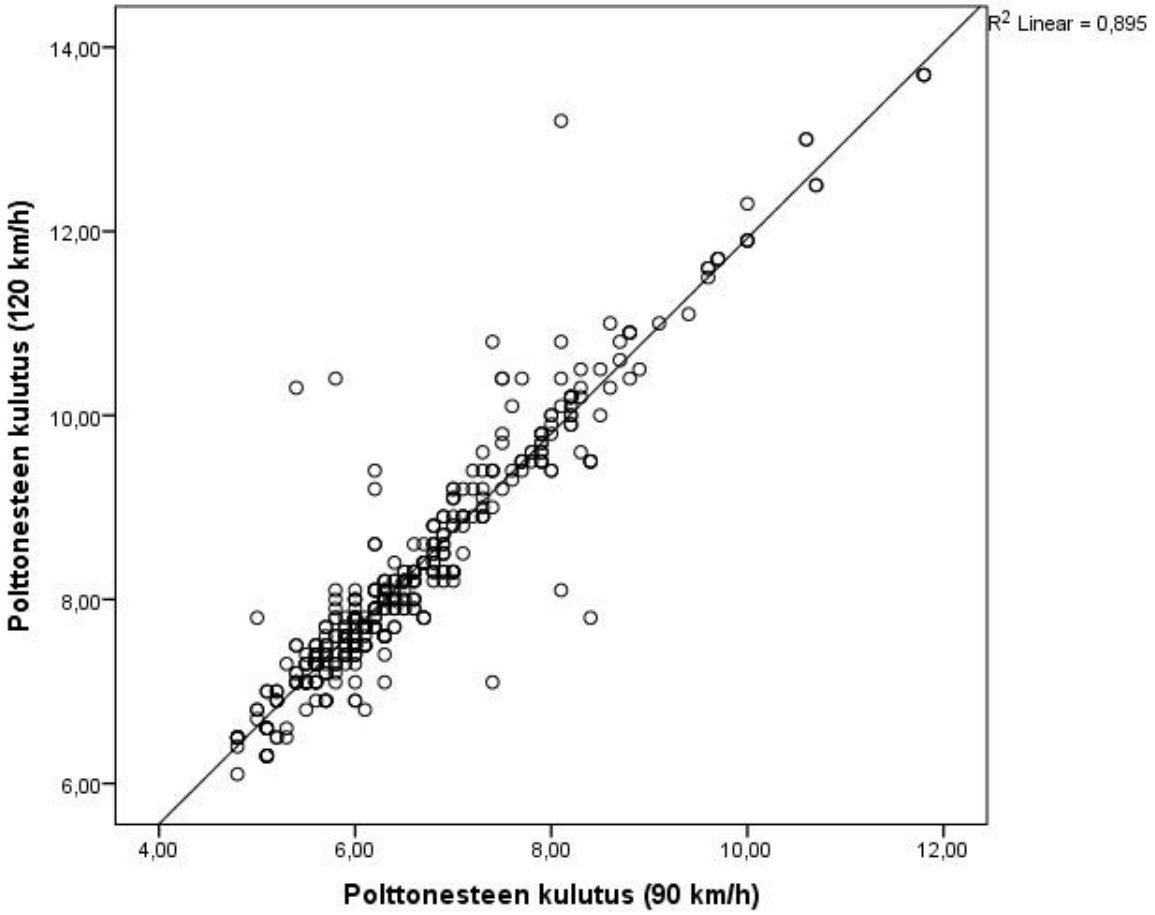
Saadaan

$$\hat{\beta}_0 = 4{,}435, \hat{\beta}_1 = 0{,}016$$

Pisteparveen sovitetun suoran yhtälö

$$y = 4{,}435 + 0{,}016x,$$

y = Kulutus, x = Teho


R2 Linear = 0,895
14,00
12,00
10,00
8,00
6,00
Polttonesteen kulutus (120 km/h)
4,00
6,00
8,00
10,00
12,00
Polttonesteen kulutus (90 km/h)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,946[a]	,895	,894	,44309

a. Predictors: (Constant), Polttonesteen kulutus (90 km/h)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,316	,119		11,059	,000
	Polttonesteen kulutus (90 km/h)	1,061	,018	,946	59,997	,000

a. Dependent Variable: Polttonesteen kulutus (120 km/h)

Merkitään

Y = Polttonesteen kulutus (120 km/h)

x = Polttonesteen kulutus (90 km/h)

Malli

$Y = \beta_0 + \beta_1 x + \varepsilon$

Estimoidaan mallin parametrit β_0 ja β_1.

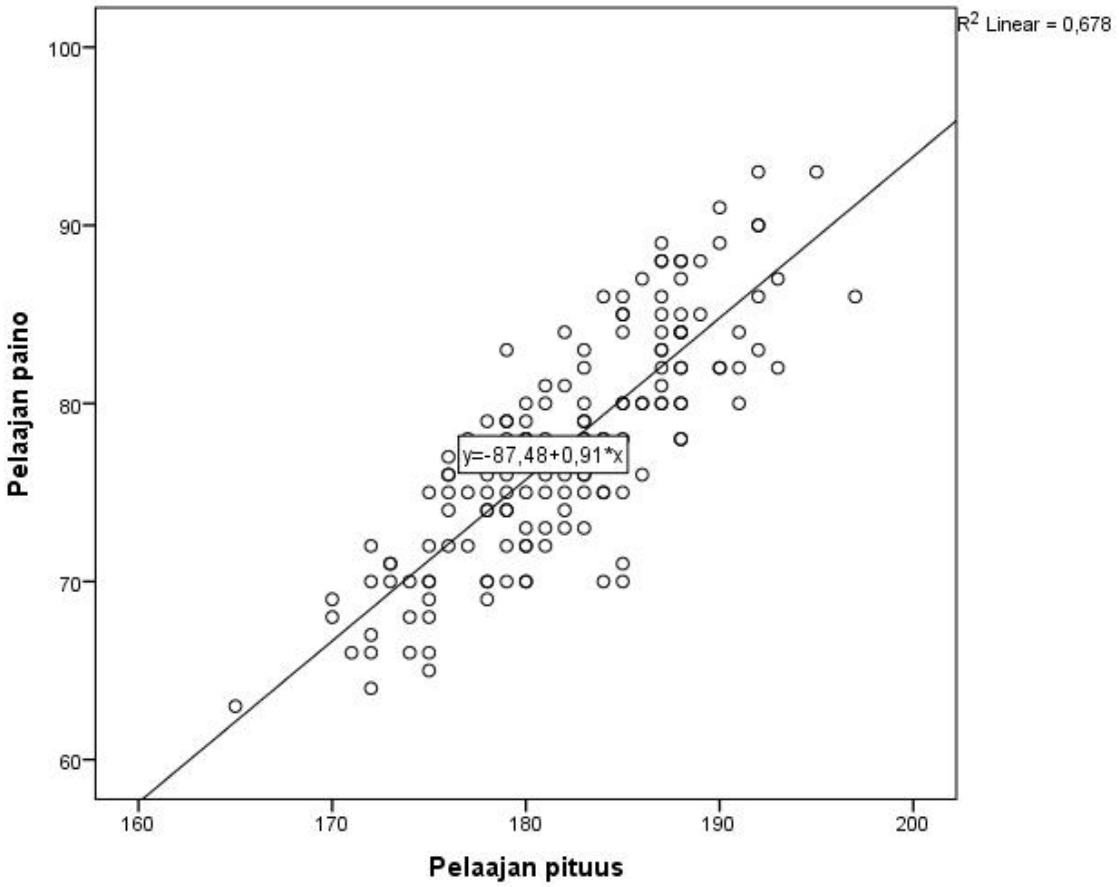
Saadaan

$$\hat{\beta}_0 = 1{,}316, \hat{\beta}_1 = 1{,}061$$

Pisteparveen sovitetun suoran yhtälö

$$y = 1{,}316 + 1{,}061x$$

Esim. Aineisto Jalkapalloilijat_2006 sivulta
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,824[a]	,678	,676	3,682

a. Predictors: (Constant), Pelaajan pituus

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-87,476	9,227		-9,480	,000
	Pelaajan pituus	,907	,051	,824	17,908	,000

a. Dependent Variable: Pelaajan paino

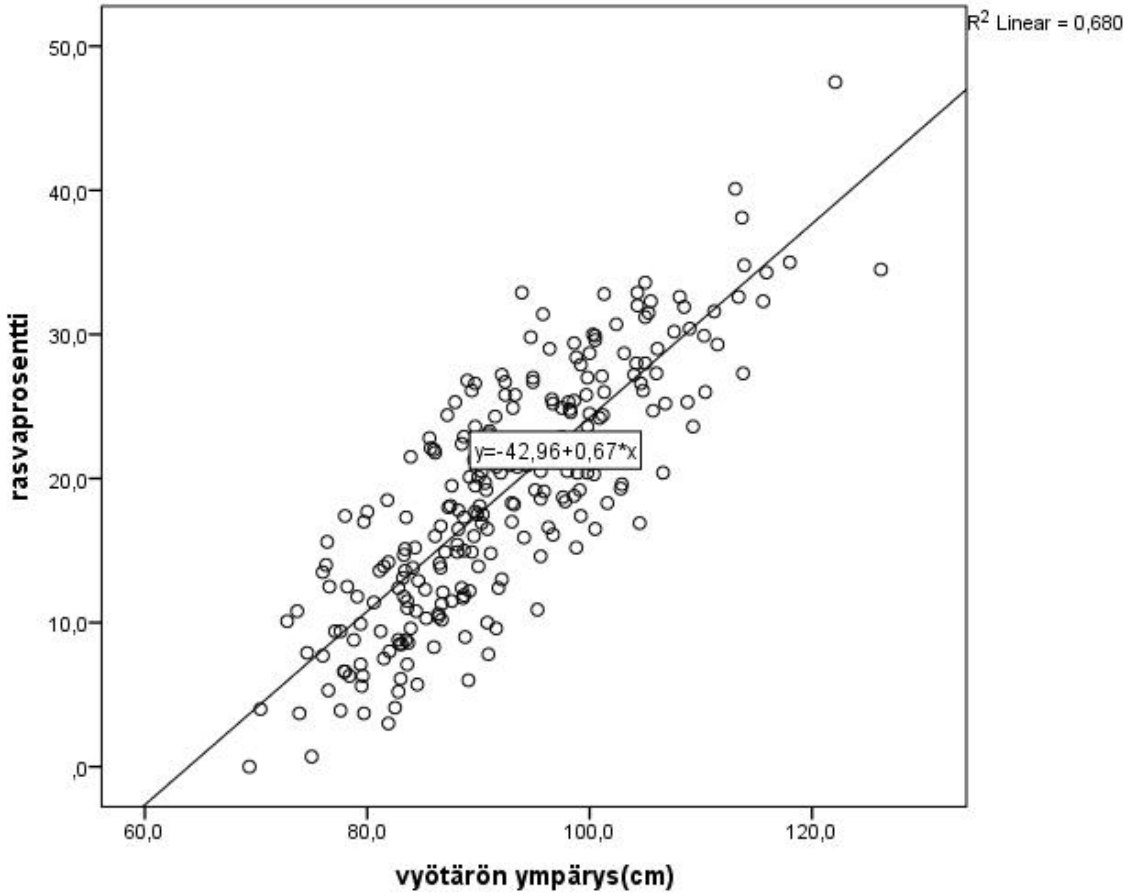
Malli ja estimoinnin tulos:

$$Paino = \beta_0 + \beta_1 \cdot Pituus + \varepsilon$$

$$\widehat{Paino} = -87{,}476 + 0{,}907 \cdot Pituus$$

Esim. Aineisto Rasvaprosentti sivulta

https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,825[a]	,680	,679	4,7174

a. Predictors: (Constant), vyötärön ympärys(cm)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-42,958	2,713		-15,833	,000
	vyötärön ympärys(cm)	,672	,029	,825	23,006	,000

a. Dependent Variable: rasvaprosentti

Malli ja estimoinnin tulos:

$$Rasva\% = \beta_0 + \beta_1 \cdot Vyötärön\ ymp. + \varepsilon$$

$$\widehat{Rasva\%} = -42,958 + 0,672 \cdot Vyötärönymp.$$

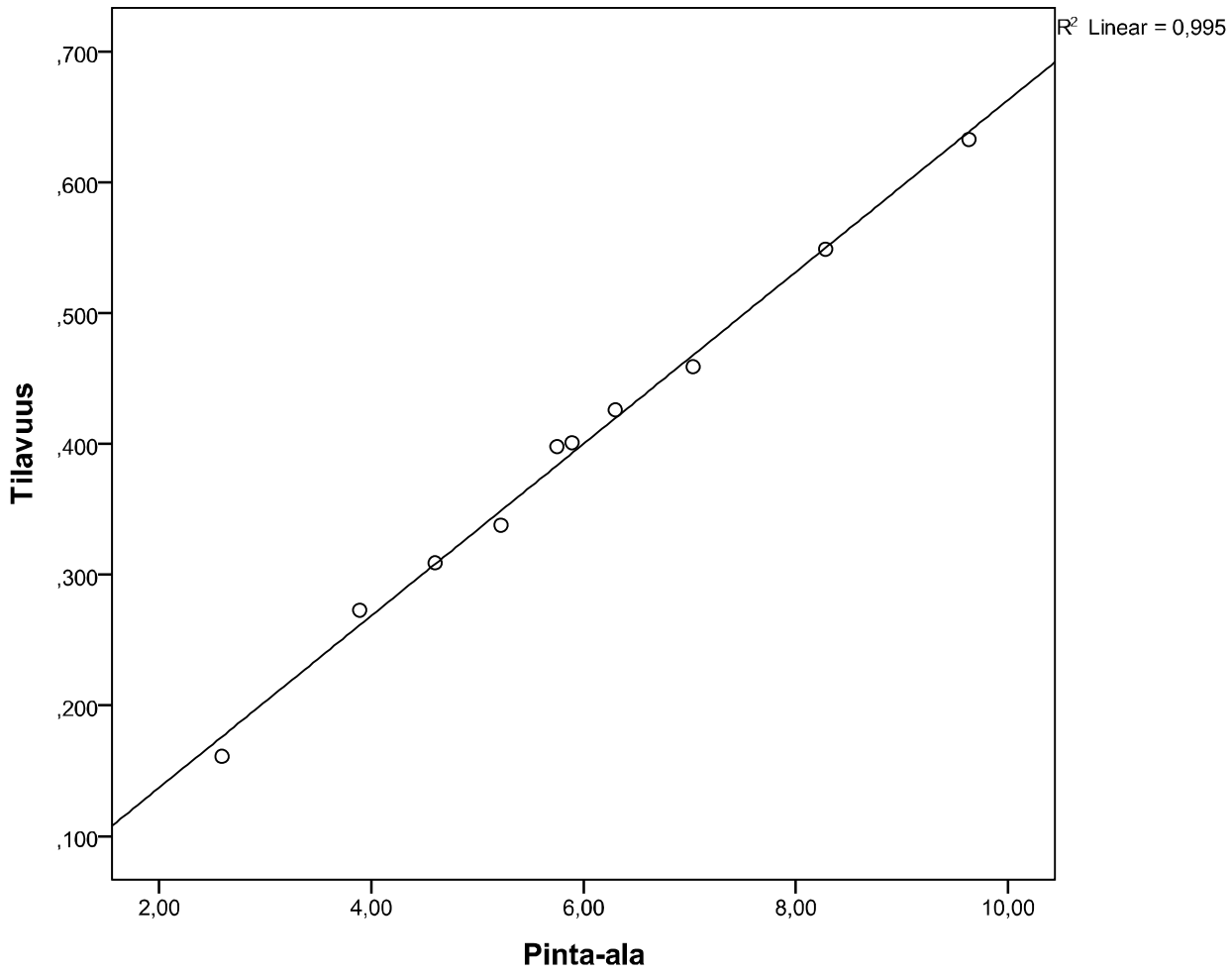
MTTTA1 Tilastomenetelmien perusteet
Luento 31.1.2019

Regressioanalyysi

4.1 Yksi selittävä muuttuja

Esim. 4.1.1 Poimittu samanikäisiä puita, mitattu poikkileikkauspinta-ala sekä puun kuutiomäärä

Pinta-ala	Tilavuus
2,59	0,161
3,89	0,273
...	
9,63	0,633


R² Linear = 0,995
,700
,600
,500
,400
,300
,200
,100
Tilavuus
2,00
4,00
6,00
8,00
10,00
Pinta-ala

Malli

$$\text{Tilavuus} = \beta_0 + \beta_1 \text{Pinta-ala} + \varepsilon$$

Estimointi

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	,006	,011		,547	,599
	Pinta-ala	,066	,002	,997	38,345	,000

a. Dependent Variable: Tilavuus

$$\widehat{\text{Tilavuus}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ Pinta-ala}$$

$\hat{\beta}_0$: 0,006

$\hat{\beta}_1$: 0,066

Jos pinta-ala on 4,60, niin arvioitu tilavuus on
$0{,}006 + 0{,}066 \cdot 4{,}60 = 0{,}310$.

Jos pinta-ala on 4, niin arvioitu tilavuus on
$0{,}006 + 0{,}066 \cdot 4 = 0{,}270$.

Yhden selittäjän regressiomalli

$$Y = \beta_0 + \beta_1 x + \varepsilon,$$

missä

- Y on satunnaismuuttuja, havaittavissa oleva, selitettävä
- x on selittäjä, ei-satunnainen, havaittavissa oleva
- ε on satunnaismuuttuja, ei havaittavissa
- β_0 ja β_1 mallin parametrit, estimoidaan aineiston avulla

Malli voidaan esittää myös muodossa

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (1)$$

Malliin liittyvät oletukset ovat

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Näistä oletuksista seuraa

- $E(Y_i) = E(\beta_0 + \beta_1 x_i + \varepsilon_i)$
  $= E(\beta_0) + E(\beta_1 x_i) + E(\varepsilon_i)$
  $= \beta_0 + \beta_1 x_i$

- $Var(Y_i) = Var(\beta_0 + \beta_1 x_i + \varepsilon_i)$
  $= Var(\varepsilon_i) = \sigma^2$

- Lisäksi $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

  Jokaista x:n arvoa kohden on olemassa Y:n todennäköisyysjakauma, joka on normaalijakauma. Havainnot näistä normaalijakaumista, graafisesti http://www.sis.uta.fi/tilasto/tiltp3/kevat2004/esim_4_1_2.pdf.

Mallin (1) parametrien estimointi

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

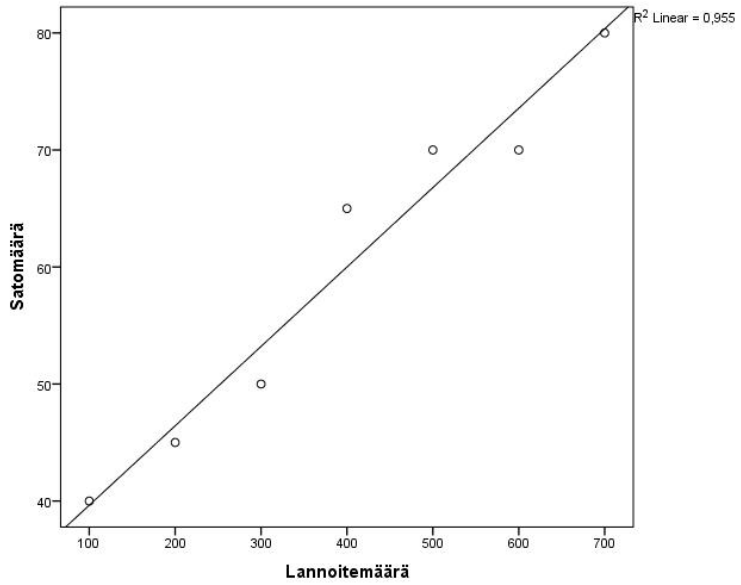
$$= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$= \frac{SP_{xy}}{SS_x}$$

$$= r_{xy} \cdot \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Esim. 4.1.4 Lannoitemäärän vaikutus satoon



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	32,857	2,945		11,157	,000
	Lannoitemäärä	,068	,007	,977	10,304	,000

a. Dependent Variable: Satomäärä

x_i	y_i	$x_i y_i$	x_i^2
100	40	4000	10000
200	45	9000	40000
300	50	15000	90000
400	65	26000	160000
500	70	35000	250000
600	70	42000	360000
700	80	56000	490000
2800	420	187000	1400000

$$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2} = \frac{187000 - \frac{1}{7} \cdot 2800 \cdot 420}{1400000 - \frac{2800^2}{7}}$$

$$= 0{,}06786$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{420}{7} - 0{,}06786 \cdot \frac{2800}{7} = 32{,}857$$

Voidaan osoittaa, että

- $E(\hat{\beta}_1) = \beta_1$
- $E(\hat{\beta}_0) = \beta_0$

Estimoidut y:n arvot saadaan

- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, \ldots, n$

Tämä suoran on Y:n odotusarvon estimaatti

Määritellään residuaalit

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

Esim. 4.1.4 (jatkoa)

x_i	y_i	$\hat{y}_i = 32{,}857+0{,}06786x_i$	$e_i = y_i - \hat{y}_i$	
100	40	$32{,}857+0{,}06786\cdot 100=39{,}64$	40-39,64	= 0,36
200	45	$32{,}857+0{,}06786\cdot 200=46{,}43$	45-46,43	=-1,43
300	50	.		=-3,21
400	65	.		= 5,00
500	70	.		= 3,21
600	70	.		=-3,57
700	80	$32{,}857+0{,}06786\cdot 700=80{,}36$	80-80,36	=-0,36

Neliösummat

$$\underbrace{SST}_{Kokonaisneliösumma} = \underbrace{SSR}_{Regressioneliösumma} + \underbrace{SSE}_{Jäännösneliösumma}$$

$$SST = \sum(y_i - \bar{y})^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum(y_i - \hat{y}_i)^2$$

Selityskerroin

$R^2 = SSR/SST$

Selitysaste, selitysprosentti

$100 \cdot R^2$

Korrelaatiokerroin

$$r_{xy} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Mallin (1) tilanteessa $(r_{xy})^2 = R^2$.

Esim. 4.1.4 (jatkoa)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,977[a]	,955	,946	3,485

a. Predictors: (Constant), Lannoitemäärä

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1289,286	1	1289,286	106,176	,000[b]
	Residual	60,714	5	12,143		
	Total	1350,000	6			

a. Dependent Variable: Satomäärä

b. Predictors: (Constant), Lannoitemäärä

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

$$= (40^2 + \cdots + 80^2) - \frac{420^2}{7} = 26550 - \frac{420^2}{7} = 1350$$

$$\text{SSE} = \sum(y_i - \hat{y}_i)^2$$

$$= 0{,}36^2 + \ldots + (-0{,}36)^2 = 60{,}7$$

$$\text{SSR} = \text{SST} - \text{SSE} = 1350 - 60{,}7 = 1289{,}3$$

$$R^2 = \text{SSR}/\text{SST} = 0{,}955$$

$$r_{xy} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}} = \frac{187000 - \frac{1}{7} \cdot 2800 \cdot 420}{\sqrt{\left(1400000 - \frac{2800^2}{7}\right)\left(26550 - \frac{420^2}{7}\right)}}$$

$$= \frac{19000}{\sqrt{280000 \cdot 1350}} = 0{,}977$$

$$(r_{xy})^2 = R^2$$

$$0{,}977^2 = 0{,}955$$

MTTTA1 Tilastomenetelmien perusteet
Luento 5.2.2019

Regressioanalyysi

4.1 Yksi selittävä muuttuja (kertausta ja jatkoa)

Regressiomalli

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (1)$$

Malliin liittyvät oletukset

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Mallin estimointi

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}$$

$$= \frac{SP_{xy}}{SS_x} = r_{xy} \cdot \frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i,$$

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

Neliösummat

$$\underbrace{SST}_{Kokonaisneliösumma} = \underbrace{SSR}_{Regressioneliösumma} + \underbrace{SSE}_{Jäännösneliösumma}$$

$$SST = \sum(y_i - \bar{y})^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

$$SSE = \sum(y_i - \hat{y}_i)^2$$

$$MSE = SSE/(n-2) = \hat{\sigma}^2$$

Selityskerroin

$$R^2 = SSR/SST$$

Selitysaste, selitysprosentti

$$100 \cdot R^2$$

Korrelaatiokerroin

$$r_{xy} = \frac{SP_{xy}}{\sqrt{SS_x SS_y}}$$

Mallin (1) tilanteessa $(r_{xy})^2 = R^2$.

Testaukset

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-2}, kun\ H_0\ \ tosi,$$

$$s(\hat{\beta}_1) = \sqrt{MSE / SS_x}$$

H_0: $\beta_0 = 0$
H_1: $\beta_0 \neq 0$

$$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-2}, kun\ H_0\ \ tosi,$$

$$s(\hat{\beta}_0) = \sqrt{MSE(\frac{1}{n} + \frac{\bar{x}^2}{SS_x})}$$

Esim. 4.1.4 (jatkoa)

Malli: Satomäärä = $\beta_0 + \beta_1 \cdot$Lannoitemäärä + ε

Kertoimien testaus

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	$\hat{\beta}_0$ = 32,857	$s(\hat{\beta}_0)$ = 2,945		11,157	,000
	Lannoitemäärä	$\hat{\beta}_1$ = ,068	$s(\hat{\beta}_1)$ = ,007	,977	10,304	,000

a. Dependent Variable: Satomäärä

$H_0: \beta_0 = 0$

$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)}$

H_0 hylätään, koska $p < 0{,}001$

$H_0: \beta_1 = 0$

$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)}$

H_0 hylätään, koska $p < 0{,}001$

$$SSE = \sum(y_i - \hat{y}_i)^2 = 0{,}36^2 + \ldots + (-0{,}36)^2 = 60{,}7$$

$$SS_x = 1400000 - 2800^2/7 = 280000$$

$$\bar{x} = 2800/7 = 400$$

$$MSE = 60{,}7/(7-2) = 12{,}143 = \hat{\sigma}^2$$

$$s(\hat{\beta}_1) = \sqrt{12{,}143/280000} = 0{,}007$$

$$s(\hat{\beta}_0) = \sqrt{12{,}143\left(\frac{1}{7} + \frac{400^2}{280000}\right)} = 2{,}945$$

Päättelyt taulukkoarvon perusteella:

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

$t_{0,01/2,7-2} = 4,032 < 10,304$, H_0 hylätään eli lannoitemäärä pidetään mallissa

H_0: $\beta_0 = 0$
H_1: $\beta_0 \neq 0$

$t_{0,01/2,7-2} = 4,032 < 11,157$, H_0 hylätään eli vakio syytä olla mallissa

Regressiomalli ilman vakiokerrointa

$$Y_i = \beta x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (2)$$

Estimointi

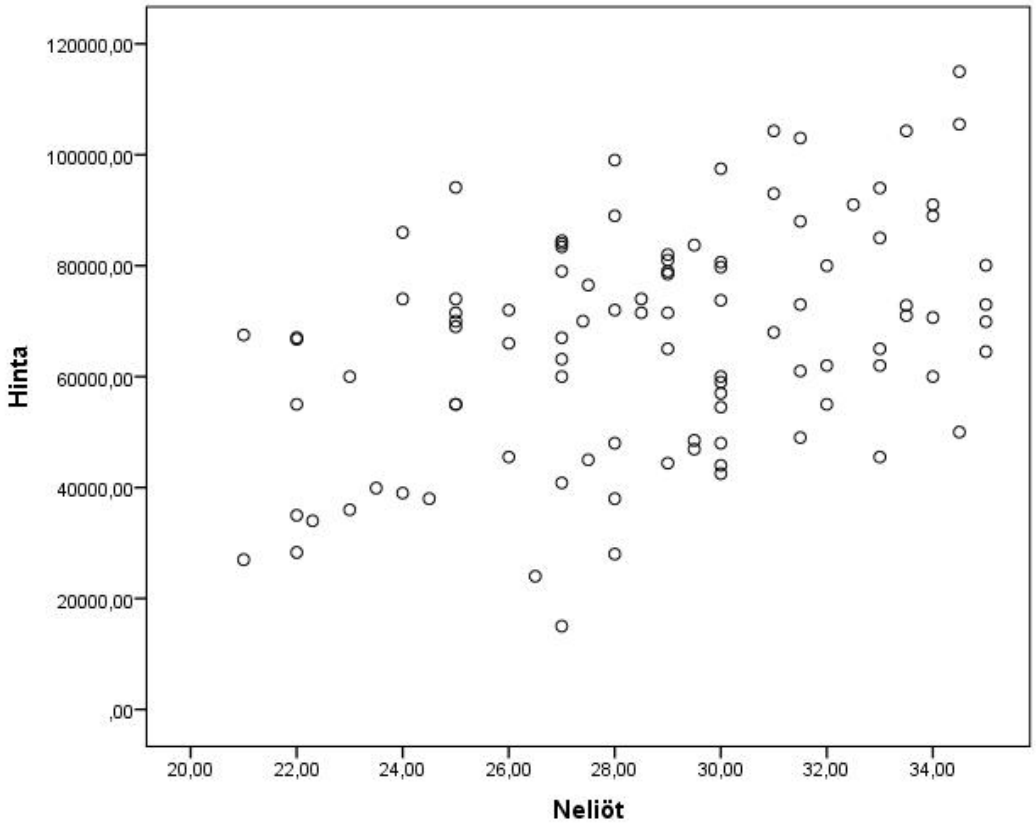
$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

$$\hat{y}_i = \hat{\beta} x_i,$$

Huom! Tällöin R^2 ei ole käytettävissä.

Esim. Aineisto Tre_myydyt_asunnot_2009, sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Malli: Hinta = $\beta_0 + \beta_1 \cdot$Neliöt + ε



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,404[a]	,163	,155	18874,53878

a. Predictors: (Constant), Neliöt

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	3076,025	14759,533		,208	,835
	Neliöt	2205,035	509,399	,404	4,329	,000

a. Dependent Variable: Hinta

Hypoteesi H_0: $\beta_0 = 0$ hyväksytään, vakiokerroin voidaan jättää pois mallista.

Estimoidaan uusi malli

$$\text{Hinta} = \beta \cdot \text{Neliöt} + \varepsilon$$

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	,963[a]	,928	,927	18781,24256

a. Predictors: Neliöt

b. For regression through the origin (the no-intercept model), R Square measures the proportion of the variability in the dependent variable about the origin explained by regression. This CANNOT be compared to R Square for models which include an intercept.

Nyt ei voida laskea selitysprosenttia!

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	Neliöt	2310,309	65,478	,963	35,284	,000

a. Dependent Variable: Hinta

b. Linear Regression through the Origin

Estimoinnin tulos origon kautta kulkeva suora

$$\widehat{Hinta} = 2310{,}309 \cdot \text{Neliöt}$$

Korrelaatiokertoimen testaus

Populaatiossa muuttujien X ja Y välinen korrelaatiokerroin

$$\rho = \mathrm{Cov}(X, Y)/\sigma_X\sigma_Y.$$

Tätä estimoidaan otoskorrelaatiokertoimella

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

$$= \frac{\sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)}{\sqrt{(\sum x_i^2 - \frac{1}{n}(\sum x_i)^2)(\sum y_i^2 - \frac{1}{n}(\sum y_i)^2)}}$$

Testaus

H_0: $\rho = 0$

H_1: $\rho \neq 0$

$$t = \frac{r}{\sqrt{\frac{1 - r^2}{n - 2}}} \sim t_{n-2}, kun\ H_0\ \ tosi$$

Esim. 4.1.9 Esimerkin 4.1.4 muuttujat

y = satomäärä
x = lannoitemäärä

r = 0,977, n = 7

H_0: $\rho = 0$
H_1: $\rho \neq 0$

$$t = \frac{0{,}977}{\sqrt{\frac{1 - 0{,}977^2}{7 - 2}}} = 10{,}304 > t_{0{,}01/2;5} = 4{,}032$$

H_0 hylätään 1 %:n riskitasolla. Päätellään lineaarista riippuvuutta olevan.

Esim. Aineisto Jalkapalloilijat_2006 sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = paino
x = pituus

r_{xy} = 0,823679, n = 154

Correlations

		Pelaajan paino	Pelaajan pituus
Pelaajan paino	Pearson Correlation	1	,824**
	Sig. (2-tailed)		,000
	N	154	154
Pelaajan pituus	Pearson Correlation	,824**	1
	Sig. (2-tailed)	,000	
	N	154	154

**. Correlation is significant at the 0.01 level (2-tailed).

$H_0: \rho = 0$

$H_1: \rho \neq 0$

$$t = \frac{0{,}823679}{\sqrt{\frac{1 - 0{,}823679^2}{154 - 2}}} = 17{,}908 > t_{0{,}005;152} = 2{,}617$$

H_0 hylätään 1 %:n riskitasolla. Päätellään lineaarista riippuvuutta olevan.

Regressiomalli: Paino = β_0 + β_1 Pituus + ε

H_0: $\beta_1 = 0$
H_1: $\beta_1 \neq 0$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-87,476	9,227		-9,480	,000
	Pelaajan pituus	,907	,051	,824	17,908	,000

a. Dependent Variable: Pelaajan paino

$t_{hav.}$ = 17,908

Siis korrelaatiokertoimen testaus on sama kuin regressiomallissa (1) β_1:n testaus!

Esim. Aineisto Jalkapalloilijat_2006 sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

Regressioanalyysin tuloksia
http://www.sis.uta.fi/tilasto/mttta1/kevat2015/RA_jalkapalloilijat.pdf

MTTTA1 Tilastomenetelmien perusteet
Luento 7.2.2019

4.1 Yksi selittävä muuttuja (täydennystä)

Regressiomalli

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (1)$$

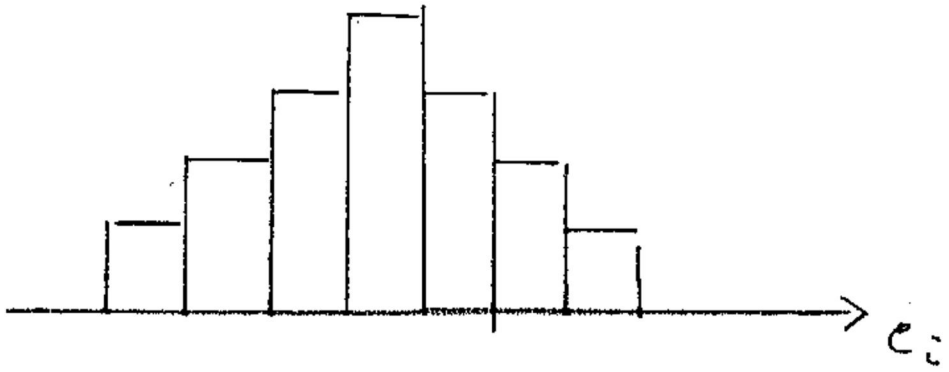
Regressiomallissa (1) oletetaan, että

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

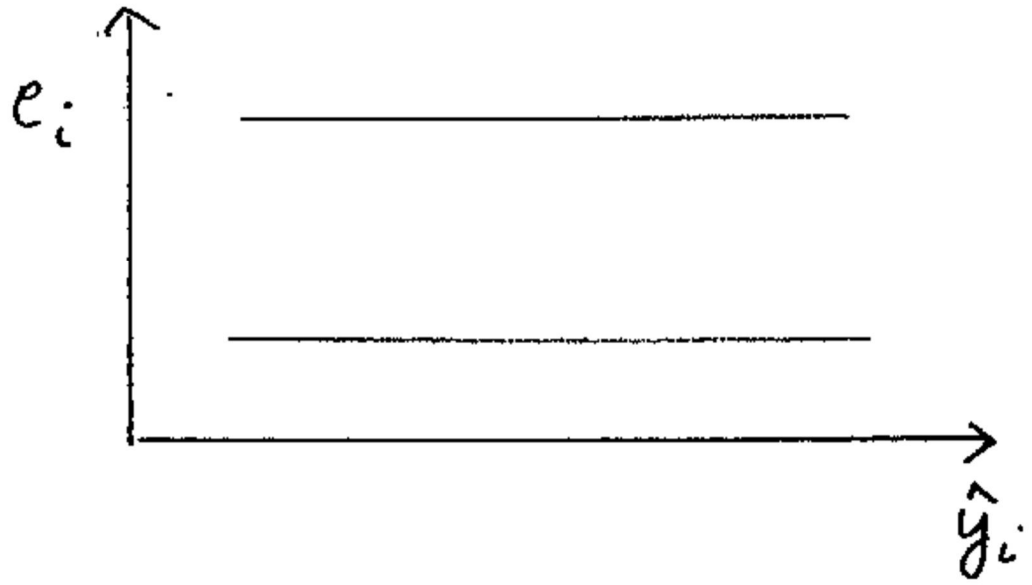
Näiden oletusten voimassaoloa tutkitaan residuaalien avulla. Koska satunnaisvirheistä ε_i ei ole havaintoja, niin estimoidaan niitä estimoidun mallin avulla lasketuilla residuaaleilla $e_i = y_i - \hat{y}_i = \hat{\varepsilon}_i$

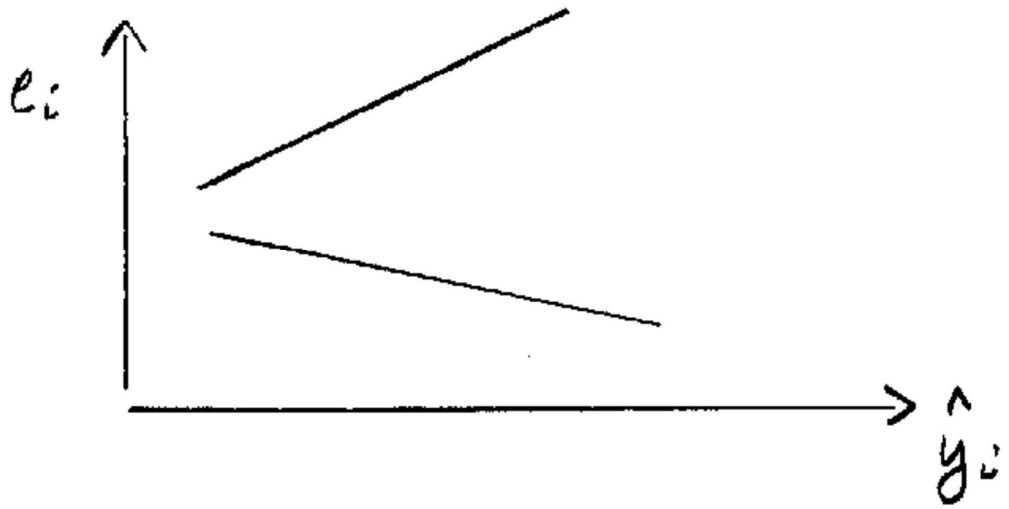
Tutkitaan normaalisuus-, vakiovarianssisuus- ja riippumattomuusoletuksia näiden residuaalien avulla. Voidaan käyttää graafisia esityksiä, esimerkiksi seuraavia:

- Normaalisuusoletuksen tutkiminen esim. histogrammin avulla



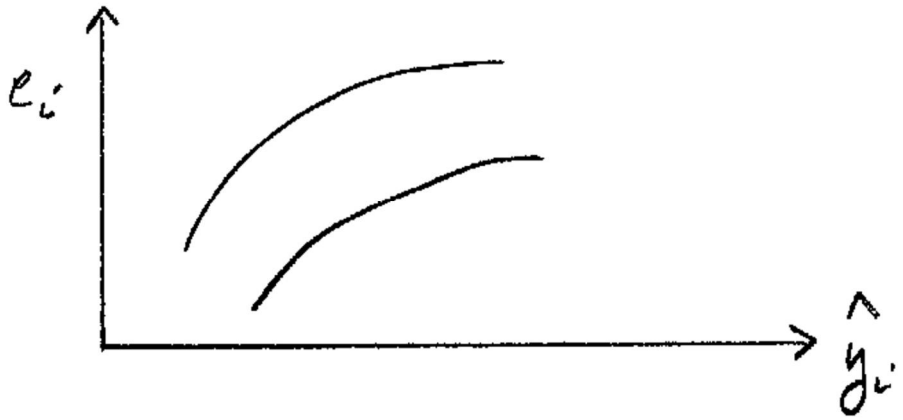
- Vakiovarianssisuuden ja riippumattomuuden tutkiminen pisteparvien avulla





Ei voida olettaa, että $\mathrm{Var}(\varepsilon_i) = \sigma^2$, i = 1, ..., n (heteroskedastisuus).

- Mallin riittävyyden tutkiminen

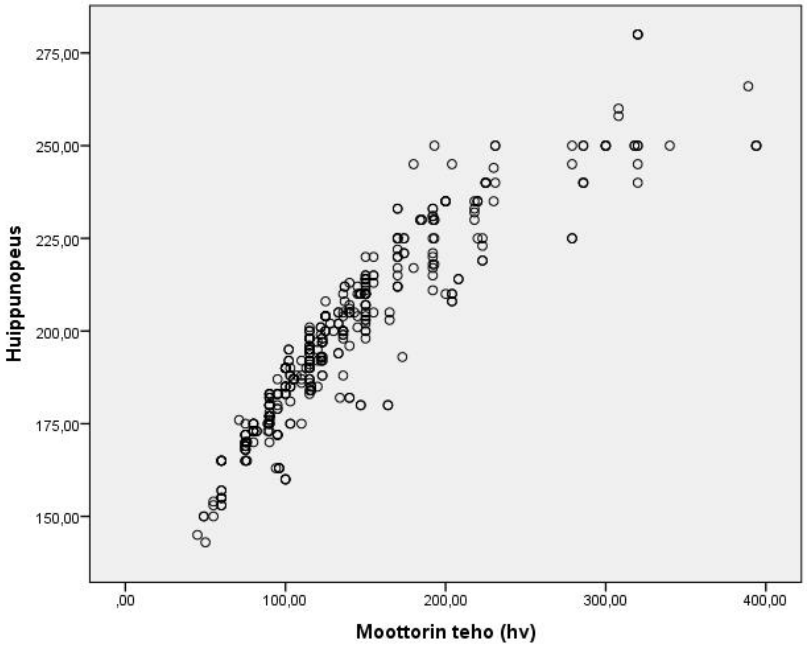


Esimerkki riittämättömästä mallista

Pisteparvissa voidaan käyttää x-akselilla myös selittäjää.

Esim. Autojen ominaisuuksia

Y = Huippunopeus, x = Teho



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	148,577	1,193		124,563	,000
	Moottorin teho (hv)	,361	,008	,914	46,370	,000

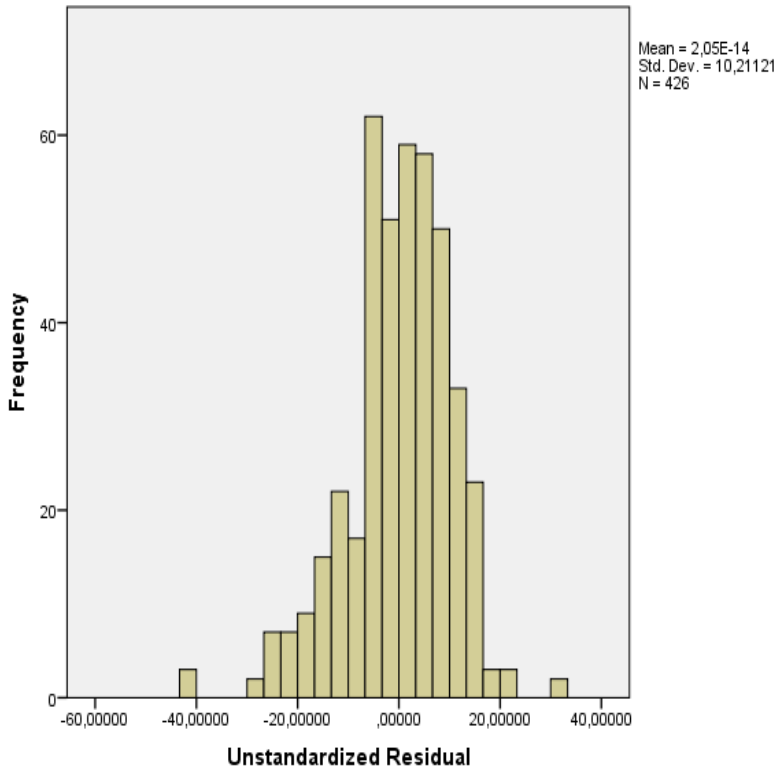
a. Dependent Variable: Huippunopeus

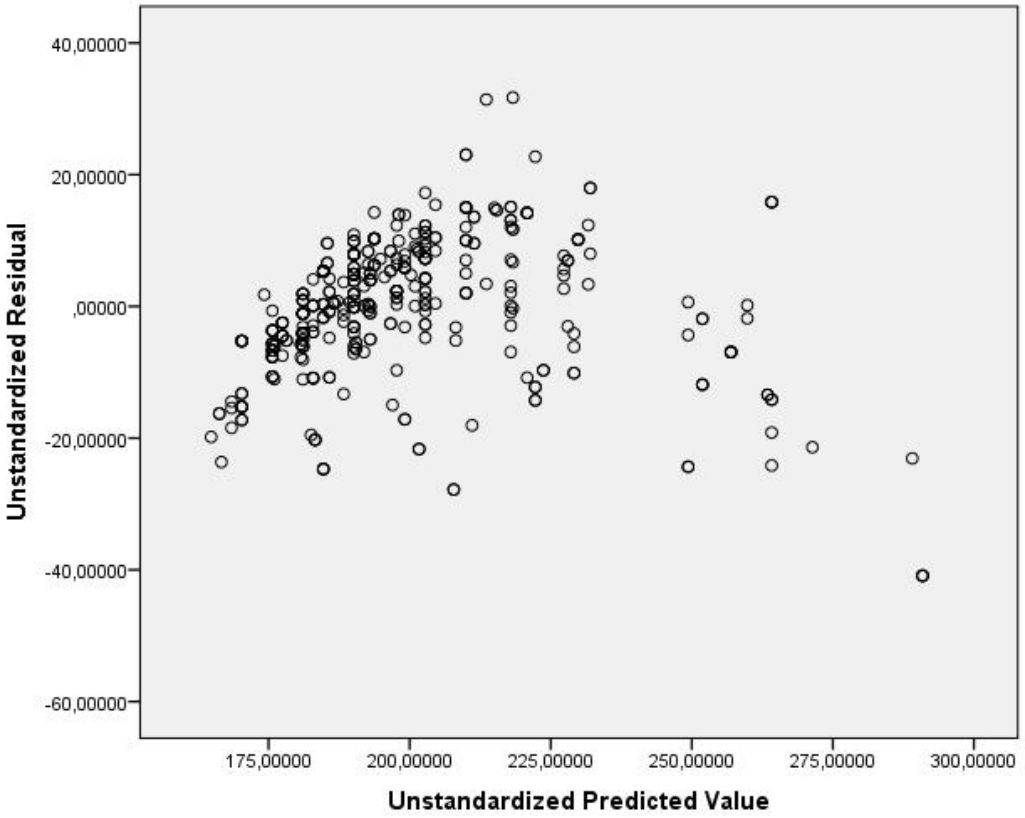
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,914[a]	,835	,835	10,22324

a. Predictors: (Constant), Moottorin teho (hv)

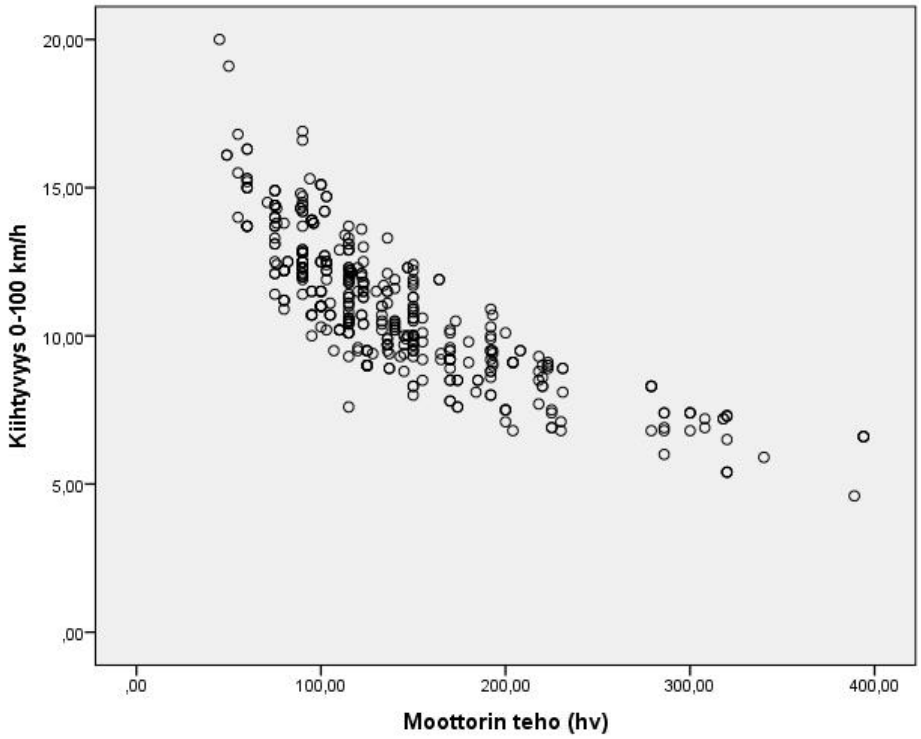
Jäännöstarkastelut





Väärä mallin valinta

Y = Kiihtyvyys, x = Teho



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	15,096	,162		93,342	,000
	Moottorin teho (hv)	-,030	,001	-,808	-28,251	,000

a. Dependent Variable: Kiihtyvyys 0-100 km/h

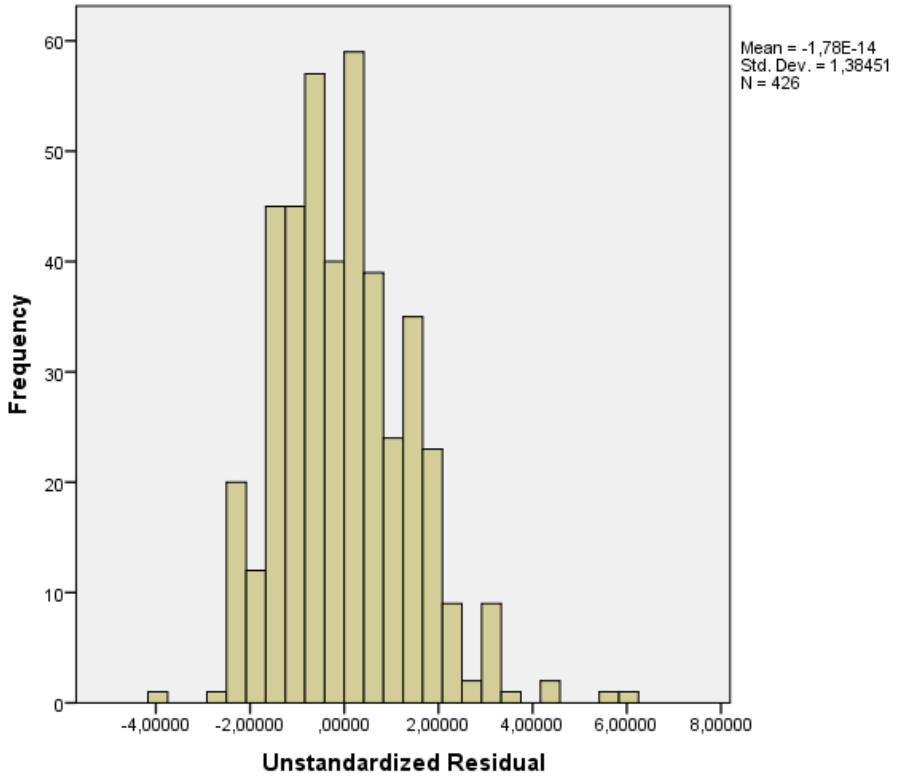
Model Summary[b]

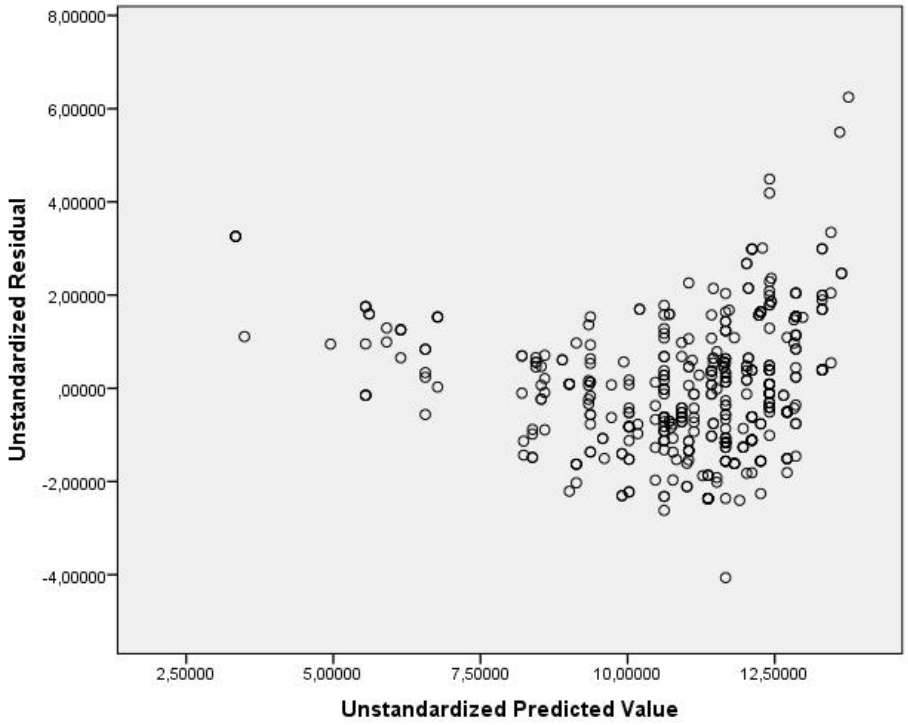
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,808[a]	,653	,652	1,38614

a. Predictors: (Constant), Moottorin teho (hv)

b. Dependent Variable: Kiihtyvyys 0-100 km/h

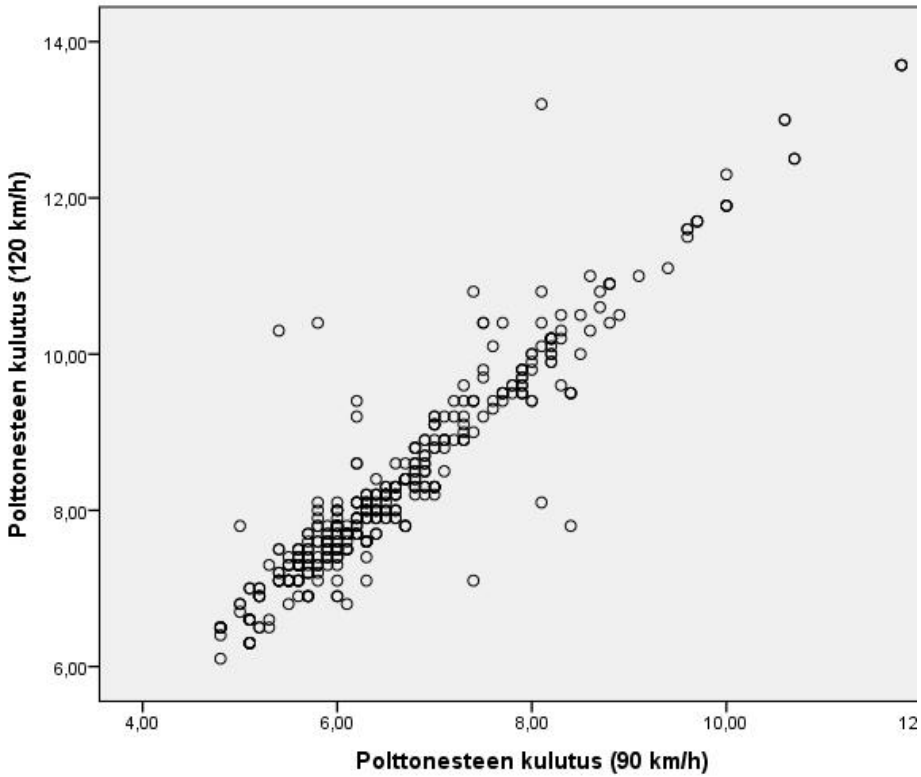
Jäännöstarkastelut





Väärä mallin valinta

Y = Kulutus 120km/h, x = Kulutus 90 km/h



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,316	,119		11,059	,000
	Polttonesteen kulutus (90 km/h)	1,061	,018	,946	59,997	,000

a. Dependent Variable: Polttonesteen kulutus (120 km/h)

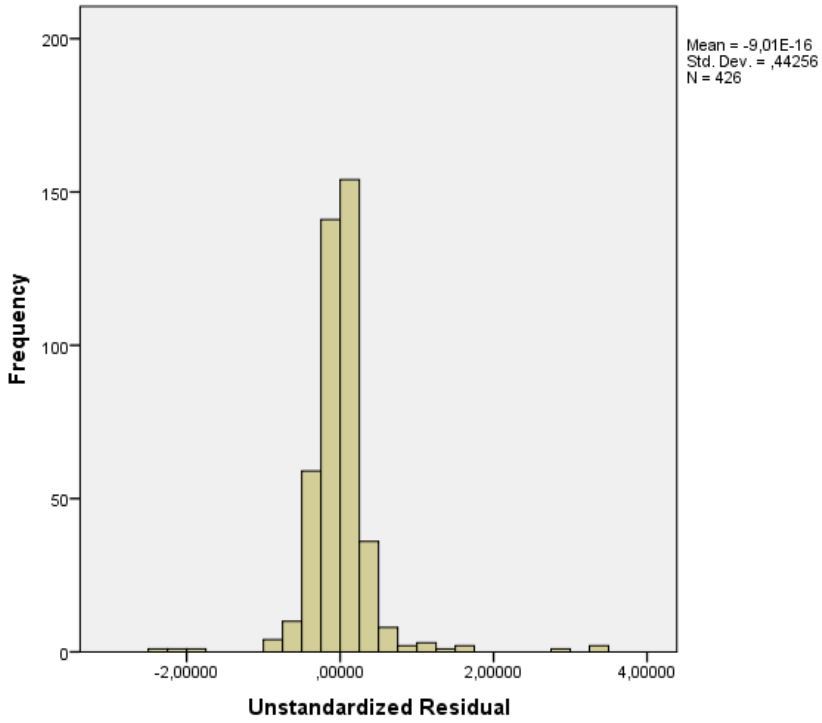
Model Summary[b]

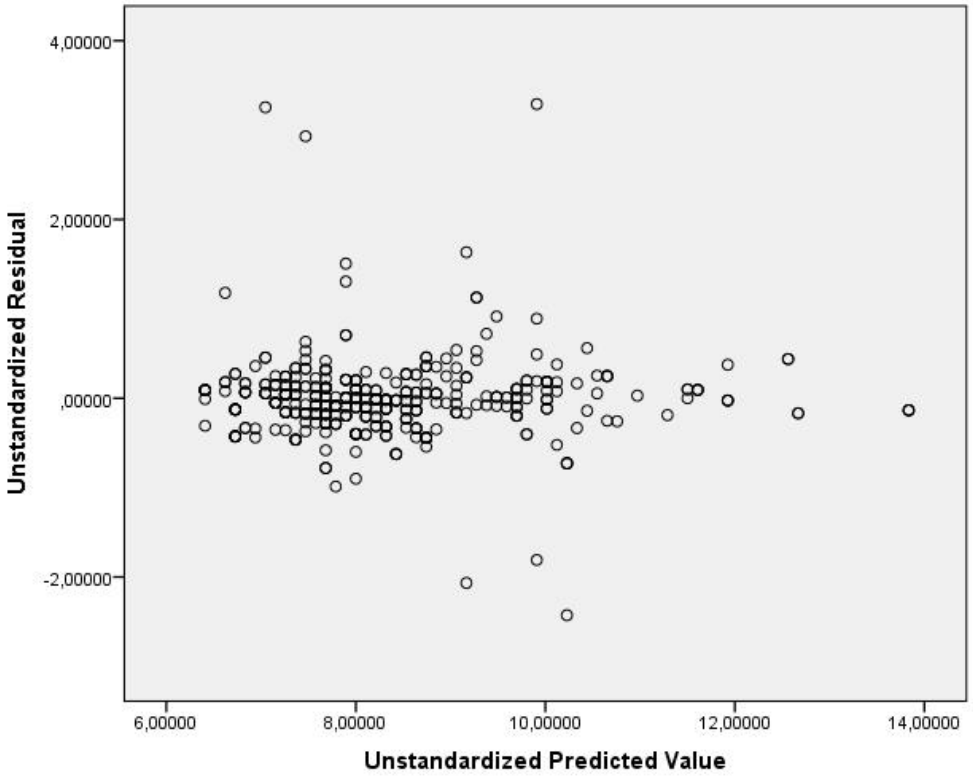
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,946[a]	,895	,894	,44309

a. Predictors: (Constant), Polttonesteen kulutus (90 km/h)

b. Dependent Variable: Polttonesteen kulutus (120 km/h)

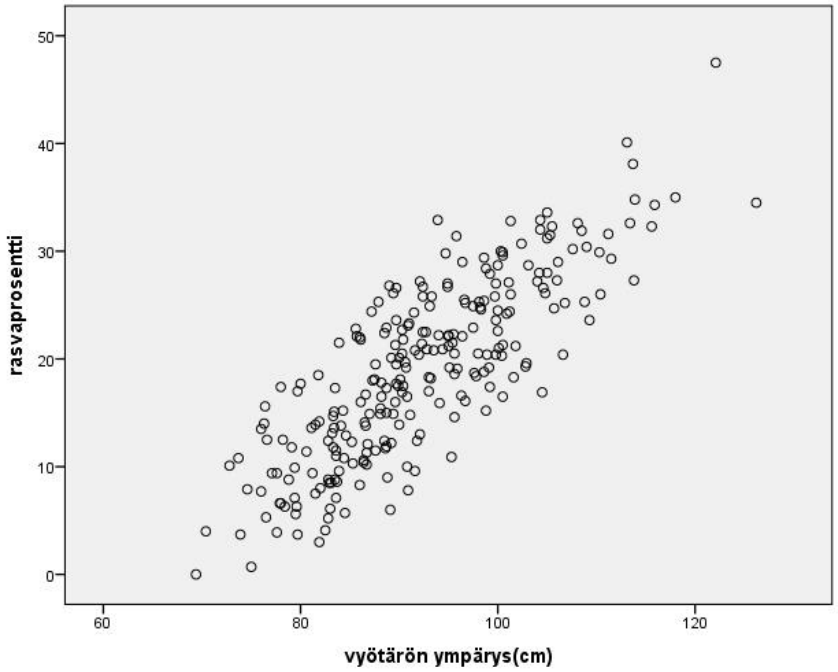
Jäännöstarkastelut





Esim. Aineisto Rasvaprosentti sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = rasvaprosentti
x = vyötärön ympärys



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-42,958	2,713		-15,833	,000
	vyötärön ympärys(cm)	,672	,029	,825	23,006	,000

a. Dependent Variable: rasvaprosentti

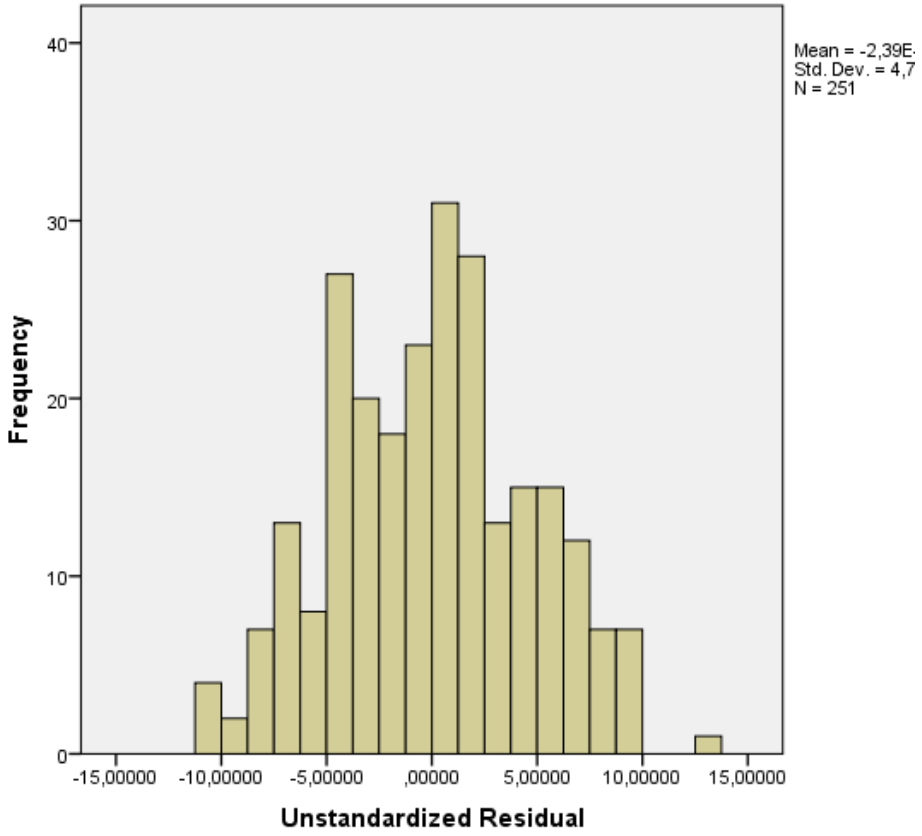
Model Summary

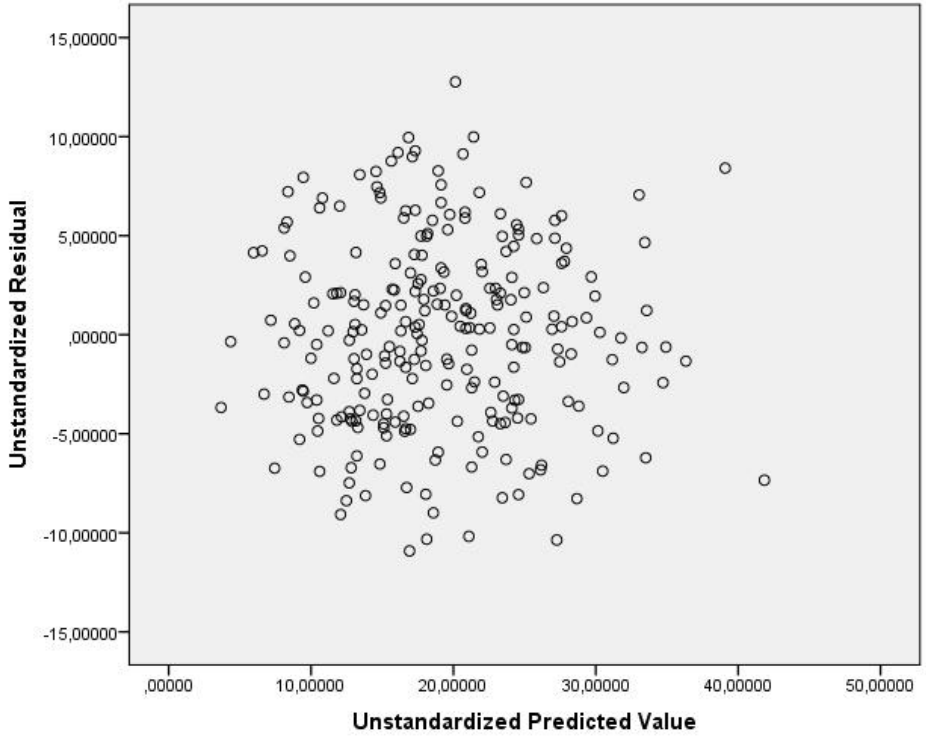
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,825[a]	,680	,679	4,717

a. Predictors: (Constant), vyötärön ympärys(cm)

Ks.
http://www.sis.uta.fi/tilasto/mttta1/kevat2015/RA_rasvaprosentti.pdf

Jäännöstarkastelut





4.2 Useampi selittävä muuttuja

Kaksi selittäjää (2-RA)

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i, \; i = 1, 2, \ldots, n \qquad (2)$$

Malliin liittyvät oletukset

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Estimointi

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{\beta}_2 z_i,$$

Testaukset

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-3}, kun\ H_0\ \ tosi$$

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$$t = \frac{\hat{\beta}_2}{s(\hat{\beta}_2)} \sim t_{n-3}, kun\ H_0\ \ tosi$$

H_0: $\beta_0 = 0$
H_1: $\beta_0 \neq 0$

$$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-3}, kun\ H_0\ \ tosi$$

H_0: $\beta_1 = \beta_2 = 0$
H_1: molemmat eivät nollia

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{2}}{\frac{SSE}{n-3}} \sim F(2, n-3), kun\ H_0\ tosi$$

Neliösummat

SST = SSR + SSE

MSR = SSR/2, MSE = SSE/(n-3) = $\hat{\sigma}^2$

Selityskerroin

R^2 = SSR/SST

Esim. Aineisto Rasvaprosentti sivulla
https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = rasva%
x_1 = vyötärön ympärys
x_2 = ikä

Regressioanalyysin tulokset
http://www.sis.uta.fi/tilasto/mttta1/kevat2015/RA_rasvaprosentti.pdf

Regressiomallissa

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n$$

hypoteesin

$$H_0: \beta_1 = 0$$

testaaminen voidaan tehdä joko t-testillä tai F-testillä, testisuureiden välinen yhteys

$$t^2 = \left(\frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \right)^2 = F = \frac{MSR}{MSE}$$

Esim. Jalkapalloilijat

y = paino

x = pituus

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4347,084	1	4347,084	320,705	,000[b]
	Residual	2060,325	152	13,555		
	Total	6407,409	153			

a. Dependent Variable: Pelaajan paino

b. Predictors: (Constant), Pelaajan pituus

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-87,476	9,227		-9,480	,000
	Pelaajan pituus	,907	,051	,824	17,908	,000

a. Dependent Variable: Pelaajan paino

$t^2 = F$

MTTTA1 Tilastomenetelmien perusteet
Luento 12.2.2019

4.2 Useampi selittävä muuttuja (jatkoa)

Selittäjien lukumäärä k (k-RA)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Malliin liittyvät oletukset

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Estimointi

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Neliösummat

SST = SSR + SSE

MSR = SSR/k, MSE = SSE/(n-k-1) = $\hat{\sigma}^2$

Selityskerroin

R^2 = SSR/SST

Testaukset

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

$$t = \frac{\hat{\beta}_i}{s(\hat{\beta}_i)} \sim t_{n-k-1}, kun\ H_0\ \ tosi$$

H_0: $\beta_1 = \ldots = \beta_k = 0$

H_1: ainakin jokin $\beta_i \neq 0$

$$F = \frac{MSR}{MSE} = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}} \sim F(k, n - k - 1), kun\ H_0\ tosi$$

Esim. CTESTI-aineisto

muuttujien kuvaukset

http://www.sis.uta.fi/tilasto/tiltp1/syksy2004/CTESTI_muuttujienkuvaus.pdf

y = cooper

x_1 = ikä

x_2 = paino

x_3 = hengitystilavuus

Regressioanalyysin tuloksia

http://www.sis.uta.fi/tilasto/mttta1/kevat2015/cooper_3_RA.pdf

Regressioanalyysin taulukko

R^2 = SSR/SST

SSR	k	MSR	F=MSR/MSE
SSE	n-k-1	MSE	~F(k, n-k-1), kun H_0 tosi
SST	n-1		H_0: $\beta_1 = \ldots = \beta_k = 0$

$\hat{\beta}_0$	$s(\hat{\beta}_0)$	$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-k-1}, kun\ H_0: \beta_0 = 0\ tosi$
$\hat{\beta}_1$	$s(\hat{\beta}_1)$	$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-k-1}, kun\ H_0: \beta_1 = 0\ tosi$
...		
$\hat{\beta}_k$	$s(\hat{\beta}_k)$	$t = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \sim t_{n-k-1}, kun\ H_0: \beta_k = 0\ tosi$

Koska

SST = SSR + SSE
1 = SSR/SST + SSE/SST
SSE/SST = 1 – SSR/SST = 1 – R^2, niin

F-testisuure voidaan esittää myös R^2:n avulla

$$F = \frac{SSR/k}{SSE/(n - k - 1)} = \frac{\frac{SSR}{SST}/k}{\frac{SSE}{SST}/(n - k - 1)} = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}}$$

Esim. y = kiinteistön myyntihinta (dollars)
x_1 = asunnon koko (square feet)
x_2 = tontin koko (square feet)
x_3 = makuuhuoneiden lukumäärä
x_4 = kylpyhuoneiden lukumäärä

(Newbold, 1991)

Regressiomalli $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

Estimoinnin tulos (kertoimet ja hajonnat)

$$\hat{y} = 1998{,}5 + 22{,}352\, x_1 + 1{,}4686\, x_2 + 6767{,}3\, x_3 + 2701{,}1\, x_4$$
$$\qquad\quad (2{,}5543) \quad (1{,}4492) \quad (1820{,}8) \quad (1996{,}2)$$

$R^2 = 0{,}9843$, $n = 20$, $k = 4$

$H_0: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

$t = 22{,}352/2{,}5543 = 8{,}75 > t_{0{,}05/2{,}15} = 2{,}131$

H_0 hylätään

$H_0: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

$t = 1{,}4686/1{,}4492 = 1{,}01 < t_{0{,}05/2{,}15} = 2{,}131$

H_0 hyväksytään

$H_0: \beta_3 = 0$
$H_1: \beta_3 \neq 0$

$t = 6767,3/1820,8 = 3,72 > t_{0,05/2,15} = 2,131$

H_0 hylätään

$H_0: \beta_4 = 0$
$H_1: \beta_4 \neq 0$

$t = 2701,1/1996,2 = 1,35 < t_{0,05/2,15} = 2,131$

H_0 hyväksytään

H_0: $\beta_1 = \ldots = \beta_4 = 0$

H_1: ainakin jokin $\beta_i \neq 0$

$$F = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}}$$

$$F_{Hav.} = \frac{\frac{0{,}9843}{4}}{\frac{1 - 0{,}9843}{20 - 4 - 1}} = 235{,}1 > F_{0{,}01;4,15} = 4{,}89$$

H_0 hylätään

Jos selittävät muuttujat ovat keskenään voimakkaasti korreloituneita (multikollineaarisia), saattaa käydä niin, että

$H_0: \beta_1 = ... = \beta_k = 0$

hylätään (tehdään päättely, että ainakin jokin $\beta_i \neq 0$), mutta kaikki hypoteesit

$H_0: \beta_i = 0$ hyväksytään.

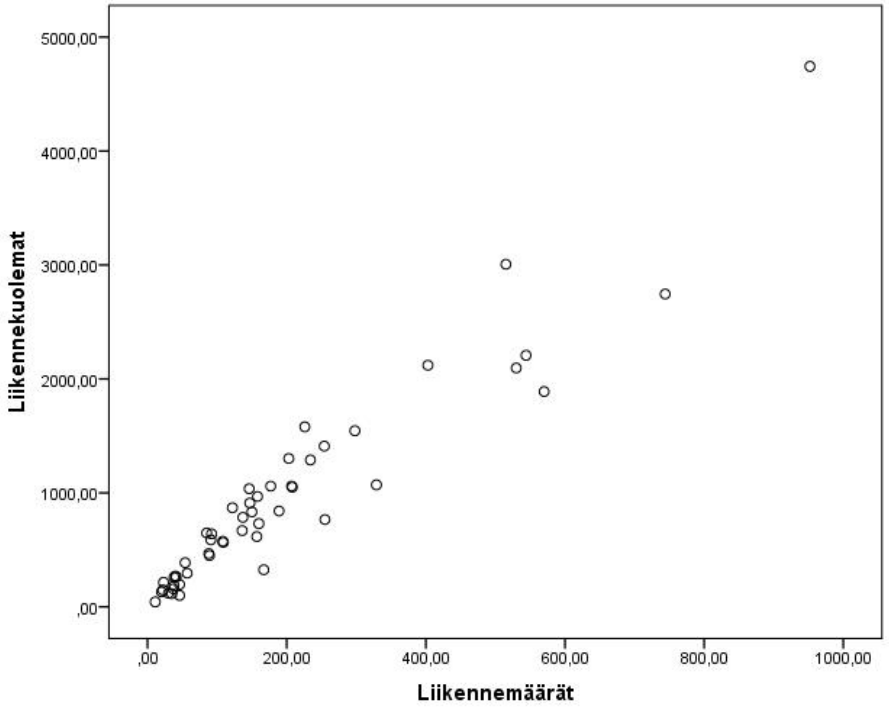
4.3 Selittävien muuttujien valinnasta ja mallin oletuksista

Mallin valinnasta

- Tarpeeksi selittäjiä, mutta käyttötarkoitukseen sopiva, tulkittavissa oleva malli.
- Tarvittaessa muunnokset, jotta mallin oletuksen voimaan.
- Automaattiset mallinvalintamenetelmät
  - etenevä valinta (Forward)
  - taaksepäin eliminointi (Backward)
  - askeltava valinta (Stepwise)

Esim. 4.3.1 Aineisto Liikennekuolemat sivulla
https://coursepages.uta.fi/mttta1/esimerkkiaineistoja/

y = liikennekuolemat
x = liikennemäärät



Malli I

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \ldots, n$$

Mallin oletukset

$\varepsilon_i \sim N(0, \sigma^2)$ ja

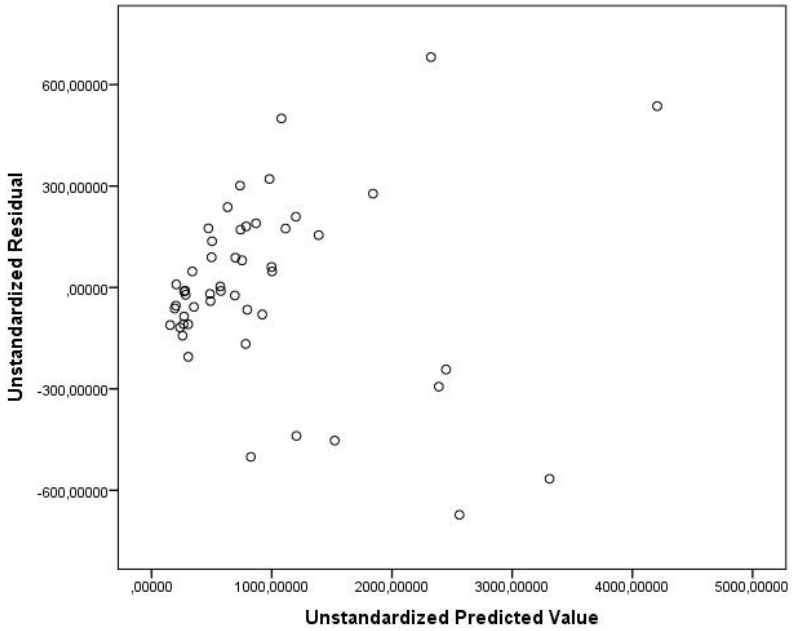
ε_i:t ovat riippumattomia

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	107,029	52,099		2,054	,045
	Liikennemäärät	4,306	,191	,956	22,549	,000

a. Dependent Variable: Liikennekuolemat

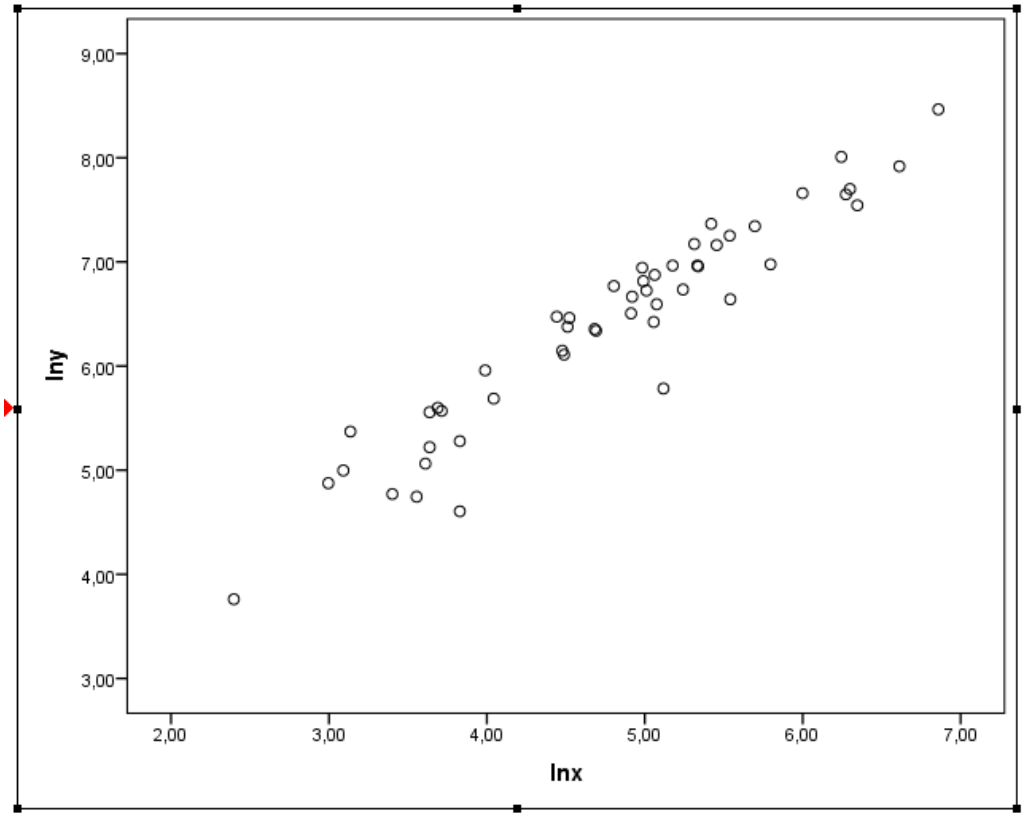
$R^2 = 0{,}914$



Residuaalitarkastelut: ei voi olettaa, että $Var(\varepsilon_i) = \sigma^2$, kun i = 1, 2, …, n.

Malli II

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

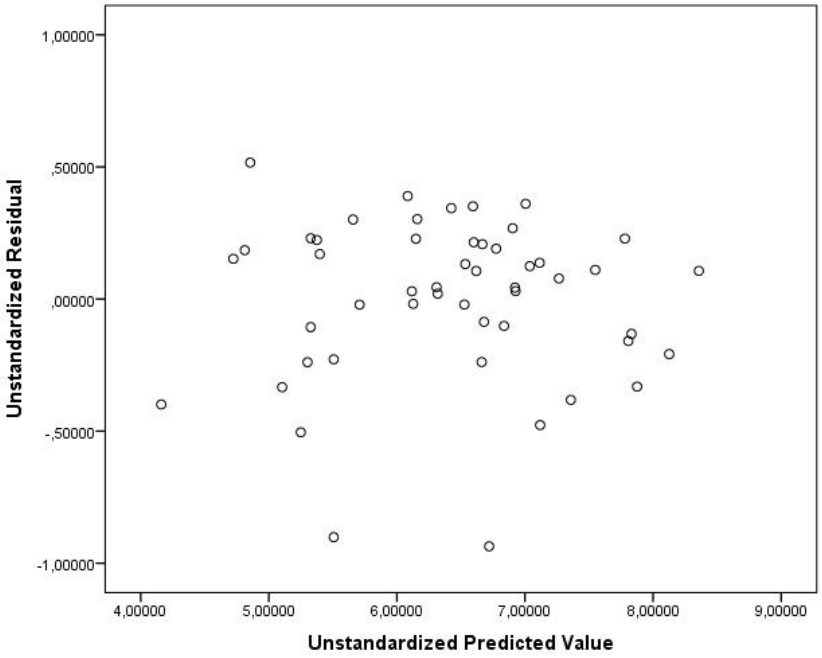


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,904	,209		9,101	,000
	lnx	,941	,043	,954	21,965	,000

a. Dependent Variable: lny

R^2 = 0,910, residuaalitarkastelut OK



Esim. Aineisto Audi_A6 sivulla https://coursepages.uta.fi/mtttp1/esimerkkiaineistoja/

y = auton hinta
x = vuosimalli
z = ajetut kilometrit
v = moottorin tilavuus

Malleja:

- $Y = \beta_0 + \beta_1 x + \varepsilon$
- $\ln(Y) = \beta_0 + \beta_1 \ln(x) + \varepsilon$
- $Y = \beta_0 + \beta_1 z + \varepsilon$
- $Y = \beta_0 + \beta_1 z + \beta_2 z^2 + \varepsilon$
- $Y = \beta_0 + \beta_1 x + \beta_2 v + \varepsilon$
- $Y = \beta_0 + \beta_1 x + \beta_2 v + \beta_3 z + \varepsilon$

MTTTA1 Tilastomenetelmien perusteet
Luento 14.2.2019

4.2 Useampi selittävä muuttuja (kertausta)

Selittäjien lukumäärä k (k-RA)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Malliin liittyvät oletukset

- $\varepsilon_i \sim N(0, \sigma^2)$ ja
- ε_i:t ovat riippumattomia

Regressioanalyysin taulukko

R^2 = SSR/SST

SSR	k	MSR	F=MSR/MSE
SSE	n-k-1	MSE	~F(k, n-k-1), kun H_0 tosi
SST	n-1		H_0: $\beta_1 = \ldots = \beta_k = 0$

$\hat{\beta}_0$	$s(\hat{\beta}_0)$	$t = \frac{\hat{\beta}_0}{s(\hat{\beta}_0)} \sim t_{n-k-1}, kun\ H_0: \beta_0 = 0\ tosi$
$\hat{\beta}_1$	$s(\hat{\beta}_1)$	$t = \frac{\hat{\beta}_1}{s(\hat{\beta}_1)} \sim t_{n-k-1}, kun\ H_0: \beta_1 = 0\ tosi$
...		
$\hat{\beta}_k$	$s(\hat{\beta}_k)$	$t = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \sim t_{n-k-1}, kun\ H_0: \beta_k = 0\ tosi$

Esim. Ilmansaasteille altistumisen vaikutus kuolleisuuteen suurkaupungeissa (Devore&Peck)

y = total mortality rate (deaths per 10000)
x_1 = mean suspended particle reading ($\mu g/m^3$)
x_2 = smallest sulfate reading (($\mu g/m^3$)x10)
x_3 = population density (people/mi^2)
x_4 = (percent nonwhite)x10
x_5 = (percent over 65)x10

Regressiomalli

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

Estimoinnin tuloksia (kertoimet, kertoimien estimoituja hajontoja)

$$\hat{y} = 19{,}607 + \underset{(0{,}016)}{0{,}041 x_1} + 0{,}071 x_2 + 0{,}001 x_3 + \underset{(0{,}007)}{0{,}014\, x_4} + 0{,}687 x_5$$

$R^2 = 0{,}827$, $n = 117$, $k = 5$

$H_0: \beta_1 = 0$

$H_1: \beta_1 > 0$

$t = 0{,}041/0{,}016 = 2{,}5625$

$t_{0{,}01,111} = 2{,}358$, $t_{0{,}005,111} = 2{,}617$

Siis $0{,}005 <$ p-arvo $< 0{,}01$

$H_0: \beta_4 = 0$
$H_1: \beta_4 > 0$

$t = 0{,}041/0{,}007 = 5{,}86 > t_{0{,}005,111} = 2{,}617$

H_0 hylätään, p-arvo $< 0{,}005$

H_0: $\beta_1 = \ldots = \beta_5 = 0$

H_1: ainakin jokin $\beta_i \neq 0$

$$F = \frac{\frac{R^2}{k}}{\frac{1 - R^2}{n - k - 1}} \sim F(k, n - k - 1), \text{ kun } H_0 \text{ tosi}$$

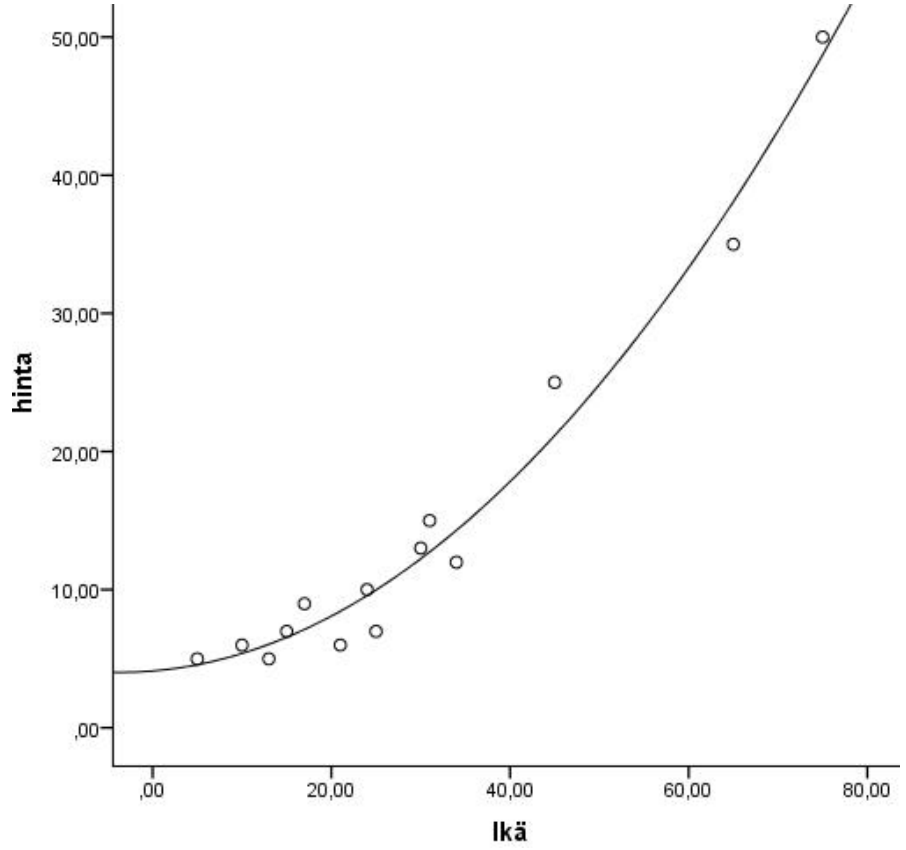
$$F = \frac{\frac{0{,}827}{5}}{\frac{1 - 0{,}827}{117 - 5 - 1}} = 106{,}124 > F_{0{,}01;5,111} = 3{,}02$$

H_0 hylätään

4.3 Selittävien muuttujien valinnasta ja mallin oletuksista (jatkoa)

Esim. Polynomiregressio, y = viinin hinta, x = viinin ikä

hinta	ika	ika2
50,00	75,00	5625,00
35,00	65,00	4225,00
25,00	45,00	2025,00
11,98	34,00	1156,00
15,00	31,00	961,00
13,00	30,00	900,00
6,98	25,00	625,00
10,00	24,00	576,00
5,99	21,00	441,00
8,98	17,00	289,00
6,98	15,00	225,00
4,99	13,00	169,00
5,98	10,00	100,00
4,98	5,00	25,00



Malli $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,114	2,015		2,042	,066
	Ikä	,055	,126	,083	,432	,674
	IkäxIkä	,007	,002	,906	4,699	,001

a. Dependent Variable: hinta

Malli $Y = \beta_0 + \beta_1 x^2 + \varepsilon$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4,918	,745		6,605	,000
	IkäxIkä	,008	,000	,987	21,391	,000

a. Dependent Variable: hinta

$R^2 = 0,974$

Esim. Autoregressio

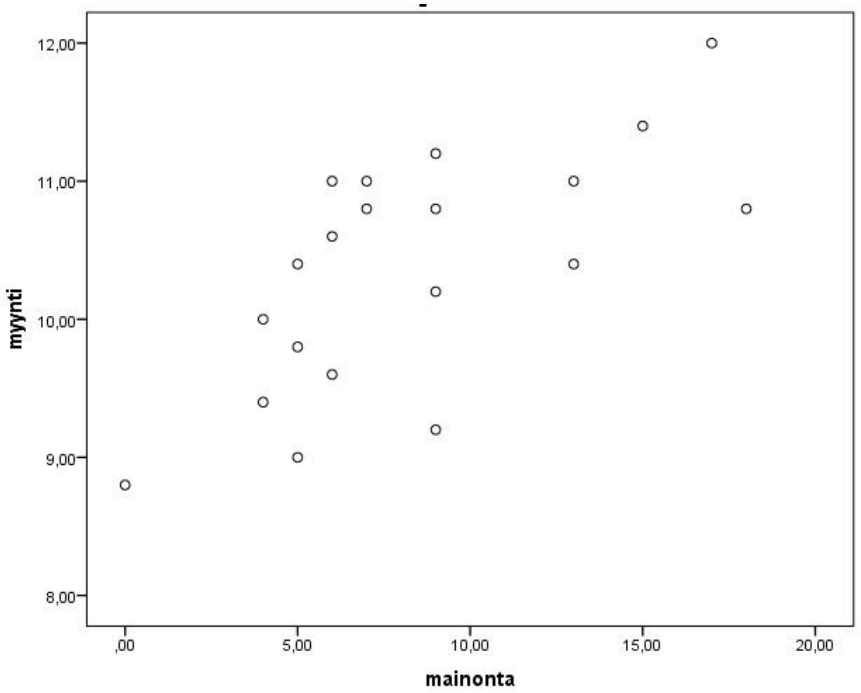
Tutkitaan vaikuttaako TV-mainonta tavaratalon myyntiin. Tarkastellaan viikoittaista myyntiä 20 viikon ajan, aineisto myynti_mainonta.sav sivulla https://coursepages.uta.fi/mttta1/esimerkkiaineistoja/

y = myynti
x = mainonta

viikko	mainonta	myynti
1	,00	8,80
2	5,00	9,00
3	9,00	9,20
4	4,00	10,00
5	6,00	9,60
6	13,00	10,40

...

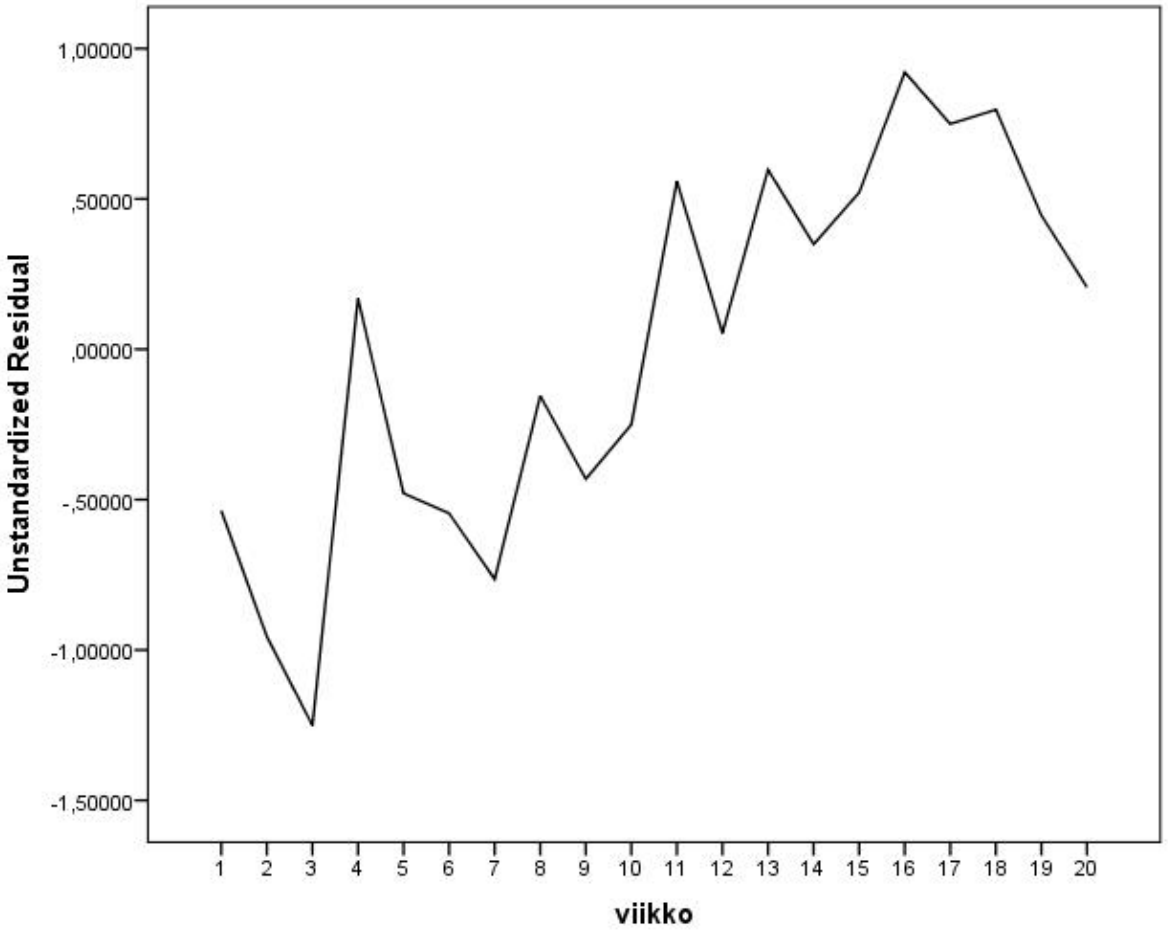
Malli I: $Y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	9,336	,301		31,020	,000
	mainonta	,124	,032	,678	3,916	,001

a. Dependent Variable: myynti



Autokorreloituneet residuaalit

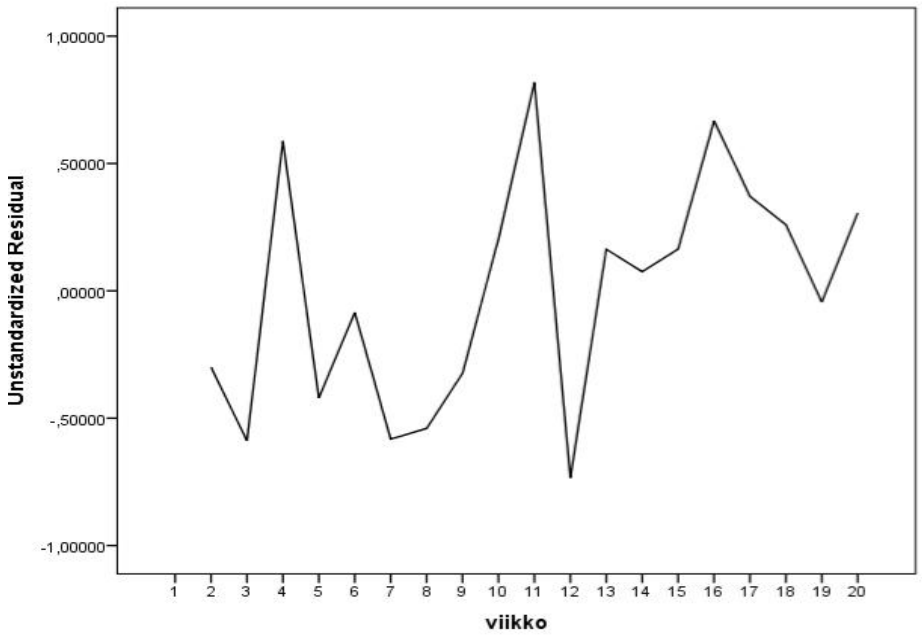
Malli II $Y_t = \beta_0 + \beta_1 x_t + \beta_2 y_{t-1} + \varepsilon_t$

Autoregressio

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t
		B	Std. Error	Beta	
1	(Constant)	4,240	1,401		3,027
	mainonta	,096	,026	,530	3,651
	myynti_1	,520	,137	,553	3,808

a. Dependent Variable: myynti

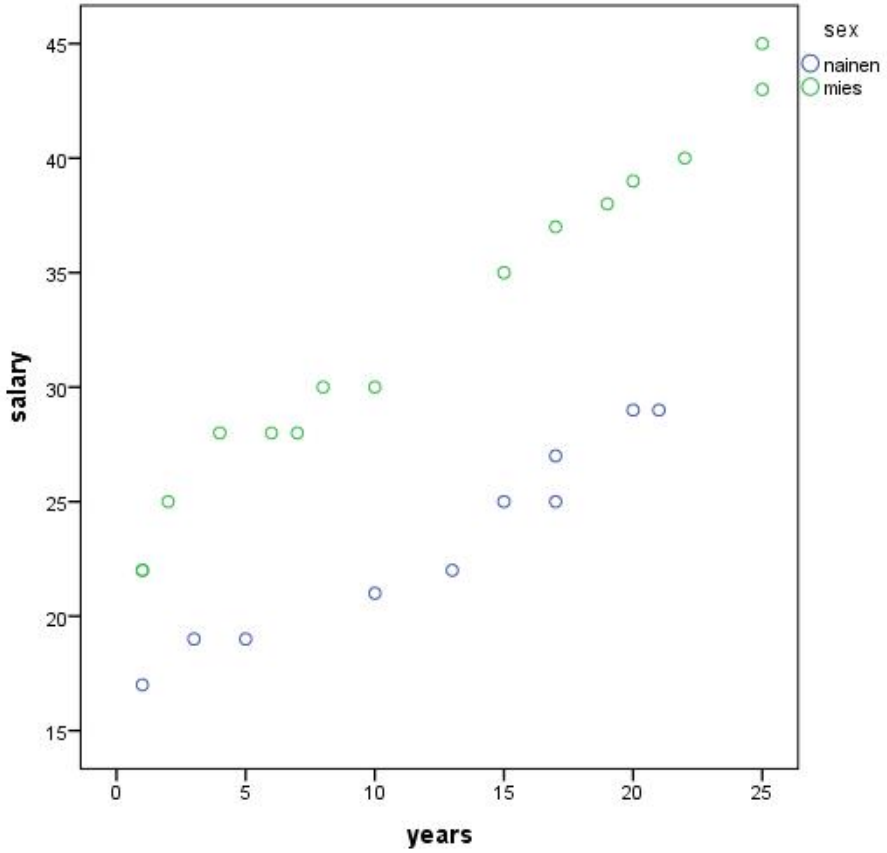


viikko	myynti	mainonta	myynti_1
1	8,80	,00	.
2	9,00	5,00	8,80
3	9,20	9,00	9,00
4	10,00	4,00	9,20
5	9,60	6,00	10,00
6	10,40	13,00	9,60
7	10,80	18,00	10,40

...

Esim. 4.3.3 Dummy-muuttuja selittäjänä mallissa

y = Salary
x = Years
z = Sex (0 = nainen, 1 = mies)



Salary	Years	Sex*
35	15	1
27	17	0
45	25	1
22	13	0
25	2	1
30	10	1
37	17	1
25	17	0
17	1	0
28	4	1
43	25	1
25	15	0
22	1	1

Salary	Years	Sex*
28	6	1
29	20	0
19	3	0
29	21	0
38	19	1
19	5	0
22	1	1
39	20	1
40	22	1
21	10	0
28	7	1
30	8	1

*1 = mies

Malli $Y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$

$E(Y) = \beta_0 + \beta_1 x$, kun Sex=0 (naiset)
$E(Y) = \beta_0 + \beta_1 x + \beta_2$, kun Sex = 1 (miehet)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	13,970	,627		22,277	,000
	years	,765	,036	,782	21,191	,000
	sex	9,418	,577	,603	16,335	,000

a. Dependent Variable: salary

Naisilla $\widehat{Salary} = 13{,}970 + 0{,}765 \cdot Years$

Miehillä $\widehat{Salary} = 13{,}970 + 0{,}765 \cdot Years + 9{,}418$
$= 23{,}388 + 0{,}765 \cdot Years$

4.4 Varianssianalyysimalli

Oletukset yksisuuntaisessa varianssianalyysissä:

$$
\begin{aligned}
&Y_{11}, Y_{12}, \ldots, Y_{1n_1} && \text{satunnaisotos } N(\mu_1, \sigma^2)\text{:sta,}\\
&Y_{21}, Y_{22}, \ldots, Y_{2n_2} && \text{satunnaisotos } N(\mu_2, \sigma^2)\text{:sta,}\\
&\quad\vdots\\
&Y_{I1}, Y_{I2}, \ldots, Y_{In_I} && \text{satunnaisotos } N(\mu_I, \sigma^2)\text{:sta.}
\end{aligned}
$$

Halutaan tutkia ovatko jakaumien odotusarvot yhtä suuret, jolloin

$$
\begin{aligned}
&H_0\colon \mu_1 = \mu_2 = \cdots = \mu_I,\\
&H_1\colon \text{kaikki odotusarvot eivät ole samoja.}
\end{aligned}
$$

Oletuksista seuraa, että varianssianalyysi voidaan ajatella mallina

$$
Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad \text{missä } \varepsilon_{ij} \sim N(0, \sigma^2).
$$

$\mu_1, \mu_2, \ldots, \mu_I$ ovat mallin parametrit. Vaihtoehtoisesti myös $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$.

Luku 5

Epäparametrisista menetelmistä (ei tenttiin)

Ei oletuksia populaatiosta, esim. normaalijakaumaoletusta.

- Mann-Witneyn testi
  Kahden riippumattoman otoksen t-testin epäparametrinen vastine (normaalijakaumaoletus ei voimassa)
- Kruskal-Wallisin testi
  Epäparametrinen vastine yksisuuntaiselle varianssianalyysille (normaalijakaumaoletusta ei

tehdä, selitettävä muuttuja voi olla järjestysasteikollinen)

- Welchin tai Brown-Forsythen testi
  Yksisuuntainen varianssianalyysi, kun oletus varianssien yhtäsuuruudesta ei voimassa

Ks. luentorunko s. 51,
http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luentorunko.pdf#page=52

Tentit

- ti 26.2.2019 klo 12.15-15.00 ls. A1, voi osallistua, jos on tehnyt vähintään 30 % harjoituksista, ilmoittaudu viimeistään 24.2.
- pe 5.4.2019
- pe 3.5.2019
- to 6.6.2019

Osaamistavoitteet

http://www.sis.uta.fi/tilasto/mttta1/kevat2019/luento_10_1_2019.pdf

Mitä jatkoksi?

Matematiikan ja tilastotieteen tutkinto-ohjelman opiskelijat

Matematiikan ja tilastotieteen perusopinnot (tilastotieteen opintopolku)
https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14974&lang=fi&uiLang=fi&lvv=2018

- MTTTP4 Todennäköisyyslaskenta (S2019)

Tilastotieteen aineopinnot (pakolliset)

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14600&lang=fi&uiLang=fi&lvv=2018

- MTTTA2 Matemaattisen tilastotieteen perusteet, (S2019)
- MTTTA4 Tilastollinen päättely 1, (K2020)
- MTTTA14 Tilastotieteen matriisilaskenta ja laskennalliset menetelmät, (S2019)

Tilastotieteen aineopinnot (muut)
https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14611&lang=fi&uiLang=fi&lvv=2018

- MTTTA13 Empiirinen projekti

Tilastotieteen perusopintokokonaisuus valinnaisina opintoina

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14644&lang=fi&uiLang=fi&lvv=2018

- MTTTA13Empiirinen projekti

Tilastotieteen aineopintokokonaisuus valinnaisina opintoina

https://www10.uta.fi/opas/opintoKokonaisuus.htm?rid=14909&lang=fi&uiLang=fi&lvv=2018

Pakolliset

- MTTTP4 Todennäköisyyslaskenta, (S2019)
- MTTTA2 Matemaattisen tilastotieteen perusteet (S2019)
- MTTTA4 Tilastollinen päättely 1, (K2020)
- MTTTA14 Tilastotieteen matriisilaskenta ja laskennalliset menetelmät, (S2019)

Muut (valitaan 3)

- MTTTA5 Monimuuttujamenetelmät
- MTTTA6 Regressioanalyysi
- MTTTA7 Yleistetyt lineaariset mallit 1
- MTTTA9 Tilastollinen ennustaminen
- MTTTA10 Sekamallit
- MTTTA11 Tilastolliset ohjelmistot (esitietona MTTTA1)
- MTTA2 Muu erikseen sovittava opintojakso