Esim. Pistemäärä-aineisto oli seuraava:

```
The decimal point is 1 digit(s) to the right of the |

0 | 9
1 | 0112233334444
1 | 566777799
2 | 001
```

Leikattu keskiarvo: Lasketaan esimerkiksi pistemäärän 10 % leikattu keskiarvo $\bar{t}_{10}$.
Lasketaan ensin se montako havaintoa 26:sta havainnosta on 10 %:

$$100 \cdot \tfrac{x}{26} = 10 \;\Rightarrow x = 2.6$$

$\Rightarrow$ Poistetaan aineistosta kolme pienintä ja kolme suurinta havaintoa, joten $n = 26-3-3 = 20$.

$$\Rightarrow \; \bar{t}_{10} = \tfrac{1}{20} \cdot (11 + 12 + 12 + \cdots + 17 + 19 + 19) \approx 14.8$$

Winsoroitu keskiarvo: Lasketaan esimerkiksi pistemäärän 10 % winsoroitu keskiarvo $\bar{w}_{10}$.
Korvataan leikattua keskiarvoa laskettaessa poisjätetyt havainnot seuraavasti:

- 3 pienintä havaintoa korvataan pienimmällä aineistoon jääneellä arvolla
- 3 suurinta havaintoa korvataan suurimmalla aineistoon jääneellä arvolla

Lasketaan keskiarvo "manipuloidulle" aineistolle, jonka $n = 26 - 3 - 3 + 3 + 3 = 26$.

$$\Rightarrow \; \bar{w}_{10} = \tfrac{1}{26} \cdot (11+11+11+11+12+\cdots+17+19+19+19+19+19) \approx 14.8$$

Esimerkki harmonisen keskiarvon käytöstä:

Herra B tekee autollaan edestakaisen matkan 120 kilometrin päässä sijaitsevalle paikkakunnalle. Menomatkalla keskinopeus oli 60 km/h ja paluumatkalla 120 km/h. Mikä oli hänen keskinopeutensa edestakaisella matkalla?

Jos keskinopeus lasketaan aritmeettisen keskiarvon avulla eli

$$\frac{60+120}{2} = 90 \quad km/h$$

saadaan virheellinen tulos!

Herra B:n ajoaika menomatkalla oli 120/60 = 2 h ja paluumatkalla 120/120 = 1 h. Oikea keskinopeus on siis

$$\frac{120+120}{3} = 80 \quad km/h$$

Tulos on harmoninen keskiarvo nopeuksista 60 km/h ja 120 km/h, sillä

$$H = \frac{2}{1/60 + 1/120} = 80 \quad km/h$$

Esimerkki geometrisen keskiarvon käytöstä:

Yhtiön A tulos 3-kertaistui vuonna 2002, 4-kertaistui vuonna 2003 ja 1.5-kertaistui vuonna 2004. Koko tänä aikana tulos siis $3 \cdot 4 \cdot 1.5 = 18$ -kertaistui. Koska tutkittava ilmiö on kerrannainen, keskimääräinen vuotuinen kasvu saadaan geometrisena keskiarvona:

$$\sqrt[3]{3 \cdot 4 \cdot 1.5} = 2.62$$

Kutakin vuotta kohti tulos siis keskimäärin 2.62-kertaistui. (Aritmeettinen keskiarvo antaisi tässä virheellisen tuloksen 2.83).

YHDEN MUUTTUJAN TARKASTELU:

MITTA-ASTEIKKO			
LUOKITTELUASTEIKKO	**JÄRJESTYSASTEIKKO**	**VÄLIMATKA-ASTEIKKO**	**SUHDEASTEIKKO**
TAULUKOITA:			
- frekvenssijakauma *(frequency distribution)*	- frekvenssijakauma - summajakauma *(cumulative frequency distribution)*	- frekvenssijakauma - summajakauma	- frekvenssijakauma - summajakauma
GRAAFISIA ESITYKSIÄ:			
- pylväsdiagrammi *(bar chart)* - piirakkakuvio *(pie chart)*	- pylväsdiagrammi - piirakkakuvio	**jatkuva / diskreetti** - histogrammi *(histogram)* / janadiagrammi - frekvenssimonikulmio - summakäyrä / porraskuvio - runko-lehti -kuvio *(stem and leaf plot)* - laatikko-jana -kuvio *(boxplot)* - pistekuvio *(dot plot)*	
TUNNUSLUKUJA:			
- moodi *(mode)*	- moodi - mediaani *(median)*	- moodi - mediaani *(median)* - aritmeettinen keskiarvo *(mean)* - leikattu keskiarvo *(trimmed mean)* - winsoroitu keskiarvo *(winsorized mean)* - geometrinen keskiarvo - harmoninen keskiarvo	
(- entropiasuhde)	- vaihteluväli - kvartiiliväli	- vaihteluväli - vaihteluvälin pituus *(range)* - kvartiiliväli - kvartiilivälin pituus *(IQR=interquartile range)* - keskipoikkeama *(mean deviation)* - keskihajonta *(standard deviation)* - varianssi *(variance)* - variaatiokerroin *(coefficient of variation, vain suhdeasteikolle)* - vinousmitta *(coefficient of skewness, skewness)* - huipukkuusmitta *(coefficient of kurtosis)*	

ESIMERKKI vakioinnista kun muuttujien välistä riippuvuutta on tutkittu ristiintaulukoinnin avulla:

Tutkimuksessa selvitettiin erään ihmisjoukon kiinnostusta urheiluun (= x) ja politiikkaan (= y). Muuttujien välille saatiin seuraavanlainen ristiintaulukko:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	55	36	91
politiikkaan	vähäinen	51	54	105
Yhteensä		106	90	196

Tutkitaan muuttujien välistä riippuvuutta ehdollisten prosenttijakaumien avulla:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	52 %	40 %	46 %
politiikkaan	vähäinen	48 %	60 %	54 %
Yhteensä		100 %	100 %	100 %

Koska ehdolliset prosenttijakaumat poikkeavat toisistaan, näyttää tarkasteltavien muuttujien muuttujien x ja y välillä olevan riippuvuutta: urheilusta paljon kiinnostuneiden joukossa on suhteellisesti enemmän paljon politiikasta kiinnostuneita.

Sukupuolen (= z) epäillään kuitenkin vaikuttavan x:n ja y:n väliseen riippuvuuteen. Tämän takia kaksiulotteiset jakaumat (eli ristiintaulukot) lasketaan sukupuoli-muuttujan eri luokissa erikseen ts. *vakioidaan sukupuoli-muuttuja.*

miehet:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	45	20	65
politiikkaan	vähäinen	25	11	36
Yhteensä		70	31	101

naiset:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	10	16	26
politiikkaan	vähäinen	26	43	69
Yhteensä		36	59	95

Tutkitaan muuttujien välistä riippuvuutta ehdollisten prosenttijakaumien avulla eri sukupuoli-luokissa:

miehet:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	64 %	65 %	64 %
politiikkaan	vähäinen	36 %	35 %	36 %
Yhteensä		100 %	100 %	100 %

- ehdolliset prosenttijakaumat ovat likimain samat, muuttujien välillä ei ole riippuvuutta.

naiset:

		kiinnostus urheiluun		
		suuri	vähäinen	Yhteensä
kiinnostus	suuri	28 %	27 %	27 %
politiikkaan	vähäinen	72 %	73 %	73 %
Yhteensä		100 %	100 %	100 %

- ehdolliset prosenttijakaumat ovat likimain samat, muuttujien välillä ei ole riippuvuutta.

Nyt nähdään, että alussa havaittu riippuvuussuhde muuttujien x ja y välillä on syntynyt niin, että miehet olivat usein kiinnostuneita sekä urheilusta että politiikasta, kun taas naisia ei kiinnostanut useimmiten kovin paljoa kumpikaan.

Pearsonin tulomomenttikorrelaatiokertoimen r arvojen tulkinnassa voi käyttää (tässä vaiheessa kurssia) seuraavia adjektiiveja:

$-1 \leq r < -0.8$: vahva (voimakas) negatiivinen lineaarinen riippuvuus _ .

$-0.8 \leq r < -0.6$: melko vahva (melko voimakas) negatiivinen lineaarinen riippuvuus

$-0.6 \leq r < -0.4$: kohtalainen (melkoinen) negatiivinen lineaarinen riippuvuus

$-0.4 \leq r < -0.2$: melko heikko negatiivinen lineaarinen riippuvuus

$-0.2 \leq r < -0.1$: heikko negatiivinen lineaarinen riippuvuus

$-0.1 \leq r \leq 0.1$: ei *lineaarista* riippuvuutta

$0.1 < r \leq 0.2$: heikko positiivinen lineaarinen riippuvuus

$0.2 < r \leq 0.4$: melko heikko positiivinen lineaarinen riippuvuus

$0.4 < r \leq 0.6$: kohtalainen (melkoinen) positiivinen lineaarinen riippuvuus

$0.6 < r \leq 0.8$: melko vahva (melko voimakas) positiivinen lineaarinen riippuvuus

$0.8 < r \leq 1$: vahva (voimakas) positiivinen lineaarinen riippuvuus

Huom. Samat adjektiivit sopivat myös Spearmannin ρ:ta ja suhteellista kontingenssikerrointa C/C_{max} tulkittaessa:

Esim.,

$\rho = 0.45$: muuttujien välillä vallitsee kohtalainen positiivinen riippuvuus

$C/C_{max} = 0.45$: muuttujien välillä vallitsee kohtalainen riippuvuus

Esim. Erään tilastotieteen kurssin opiskelijoiden sukupuolen ja koulutusohjelman ristiintaulukko oli seuraava:

	sukupuoli		
koulutusohjelma	mies	nainen	Yhteensä
kansantaloustiede	9	9	18
laskentatoimi	8	15	23
markkinointi	2	16	18
jokin muu	4	3	7
yhteensä	23	43	66

Ovatko sukupuoli ja koulutusohjelma toisistaan riippumattomia?

Ehdolliset prosenttijakaumat ovat:

	sukupuoli		
koulutusohjelma	mies	nainen	Yhteensä
kansantaloustiede	39	21	27
laskentatoimi	35	35	35
markkinointi	9	37	27
jokin muu	17	7	11
yhteensä	100	100	100

Ehdolliset prosenttijakaumat näyttävät poikkeavan jonkin verran toisistaan, joten koulutusohjelman ja sukupuolen välillä vallitsee kohtalainen riippuvuus. Miehistä kansantaloustieteen koulutusohjelmaan kuuluu noin 39 % ja naisista noin 21 %. Vastaavasti esim. markkinoinnin koulutusohjelmaan kuuluu noin 9 % miehistä ja 37 % naisista.

Mikäli tarkasteltavien muuttujien välillä ei olisi riippuvuutta ollenkaan, olisivat ehdolliset prosenttijakaumat seuraavat:

	sukupuoli		
koulutusohjelma	mies	nainen	Yhteensä
kansantaloustiede	27	27	27
laskentatoimi	35	35	35
markkinointi	27	27	27
jokin muu	11	11	11
yhteensä	100	100	100

Edellä olevaan tilanteeseen päädyttäisiin, jos ristiintaulukon havaitut solufrekvenssit (f_{ij}) olisivat olleet alunperin seuraavat:

	sukupuoli		
koulutusohjelma	mies	nainen	Yhteensä
kansantaloustiede	6.27	11.73	18
laskentatoimi	8.02	14.98	23
markkinointi	6.27	11.73	18
jokin muu	2.44	4.56	7
yhteensä	23	43	66

Edellä olevassa taulukossa olevat teoreettiset frekvenssit ovat ns. *odotettuja frekvenssejä* (e_{ij}) tilanteessa, jossa muuttujat sukupuoli ja koulutusohjelma ovat toisistaan riippumattomia ja ristiintaulukon reunafrekvenssit ovat ne, mitkä on havaittu.

Lasketaan seuraavaksi muuttujien välistä riippuvuutta kuvaavan tunnusluvun (*suhteellisen kontingenssikertoimen*) arvo.

Edellisessä taulukossa esiintyvät riippumattomuus-oletuksen mukaiset odotetut frekvenssit e_{ij} saadaan laskettua kaavalla $e_{ij} = \frac{f_{i.} f_{.j}}{n}$, missä indeksi i viittaa ristiintaulukon i. riville ja indeksi j viittaa ristiintaulukon j. sarakkeelle. Esimerkkiaineistossa

$$e_{11} = \frac{f_{1.} f_{.1}}{n} = \frac{18*23}{66} \approx 6.27, \quad e_{12} = \frac{f_{1.} f_{.2}}{n} = \frac{18*43}{66} \approx 11.73,$$
$$e_{21} = \frac{f_{2.} f_{.1}}{n} = \frac{23*23}{66} \approx 8.02, \quad e_{22} = \frac{f_{2.} f_{.2}}{n} = \frac{23*43}{66} \approx 14.98,$$
$$e_{31} = \frac{f_{3.} f_{.1}}{n} = \frac{18*23}{66} \approx 6.27, \quad e_{32} = \frac{f_{3.} f_{.2}}{n} = \frac{18*43}{66} \approx 11.73,$$
$$e_{41} = \frac{f_{4.} f_{.1}}{n} = \frac{7*23}{66} \approx 2.44, \quad e_{42} = \frac{f_{4.} f_{.2}}{n} = \frac{7*43}{66} \approx 4.56$$

χ^2-tunnusluvun havaittu arvo on

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{r} \frac{(f_{ij} - e_{ij})^2}{e_{ij}} = \frac{(9-6.27)^2}{6.27} + \frac{(9-11.73)^2}{11.73} + \ldots + \frac{(3-4.56)^2}{4.56} \approx 7.82$$

$$\Rightarrow C = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{7.82}{66+7.82}} \approx 0.33 \qquad \Rightarrow C_{MAX} = \sqrt{\frac{q-1}{q}} = \sqrt{\frac{2-1}{2}} = \sqrt{\frac{1}{2}}$$

Edellä q = min(m,r) = min(4,2) = 2.

$$\Rightarrow \frac{C}{C_{MAX}} \approx \frac{0.33}{\sqrt{0.5}} \approx 0.47$$

Muuttujien välillä vallitsee kohtalainen riippuvuus.

Vastaavasti voitaisiin laskea ns. cramerin V:

$$V = \sqrt{\frac{\chi^2}{nu}} \approx \sqrt{\frac{7.82}{66 \cdot 1}} \approx 0.34$$

, missä u = min(m-1,r-1) = min (4-1,2-1) = 1

Muuttujien välillä vallitsee melko heikko riippuvuus.

Esimerkki ristitulosuhteen käytöstä:

Tutkittaessa odottavan äidin runsaan tupakoinnin (yli 10 savuketta/päivä) vaikutusta lapsen syntymäpainoon prof. Paula Rantakallion havaintoaineiston avulla:

	Äidin tupakointi		
Lapsen syntymäpaino	**yli 10 savuketta/päivä**	**ei ollenkaan**	**Yhteensä**
alle 2500 g	23	345	368
vähintään 2500 g	253	9360	9613
Yhteensä	276	9705	9981

Tutkitaan muuttujien välistä riippuvuutta

- *ehdollisten prosenttijakaumien* avulla:

  Ehdolliset prosenttijakaumat kannattaa laskea nyt sarakkeittain, sillä äidin tupakoinnin voidaan ajatella olevan riski (syy) lapsen alipainoisuudelle (seuraus). Ehdolliset prosenttijakaumat ovat:

	Äidin tupakointi		
Lapsen syntymäpaino	**yli 10 savuketta/päivä**	**ei ollenkaan**	**Yhteensä**
alle 2500 g	8.3 %	3.6 %	3.7 %
vähintään 2500 g	91.7 %	96.4 %	96.3 %
Yhteensä	100 %	100 %	100 %

  Ehdolliset prosenttijakaumat poikkeavat toisistaan, joten äidin runsaan tupakoinnin ja lapsen syntymäpainon välillä näyttää olevan riippuvuutta. Paljon tupakoivien äitien lapsista noin 8.3 % on alipainoisia, kun taas tupakoimattomien äitien lapsista alipainoisia on noin 3.6 %. Paljon tupakoivat äidit saavat siis suhteellisesti enemmän alipainoisia lapsia kuin tupakoimattomat äidit.

- *ristitulosuhteen (OR) ja riskisuhteen (RR)* avulla:

$$OR = \frac{f_{11} f_{22}}{f_{12} f_{21}} = \frac{23 \cdot 9360}{345 \cdot 253} \approx 2.47$$

  OR:n tulkinta: muuttujien välillä on (selvä) positiivinen riippuvuus.

$$RR = \frac{\pi_1}{\pi_0} = \frac{f_{11}/f_{.1}}{f_{12}/f_{.2}} = \frac{23/276}{345/9705} \approx 2.34$$

  RR:n tulkinta: Paljon tupakoivilla äideillä on noin 2.34 kertainen riski saada alipainoinen lapsi tupakoimattomiin äiteihin verrattuna.

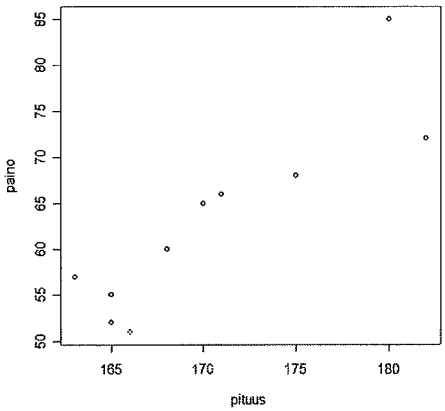
  riskiero (RD):

$$RD = \pi_1 - \pi_0 = \frac{23}{276} - \frac{345}{9705} \approx 0.048.$$

Tarkastellaan seuraavaksi pituuden (=x, cm) ja painon (=y, kg) välistä riippuvuutta 10 henkilön aineistossa:

henkilö (i)	1	2	3	4	5	6	7	8	9	10
pituus (x_i)	175	180	165	170	168	166	165	171	163	182
paino (y_i)	68	85	52	65	60	51	55	66	57	72

Pituuden ja painon välinen korrelaatiodiagrammi on seuraava:



Muuttujien välillä näyttää vallitsevan lineaarista riippuvuutta, joten korrelaatiokerroin r on mielekästä laskea.

Nyt n=10,

$$\sum_{i=1}^{10} x_i = 1705, \bar{x} = 170.5, \sum_{i=1}^{10} x_i^2 = 291089$$

$$\sum_{i=1}^{10} y_i = 631, \bar{y} = 63.1, \sum_{i=1}^{10} y_i^2 = 40793,$$

$$\sum_{i=1}^{10} x_i y_i = (175 \cdot 68) + (180 \cdot 85) + \ldots + (182 \cdot 72) = 108132$$

$$s_x = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{10} x_i^2 - n\bar{x}^2\right)} = \sqrt{\frac{1}{10-1}\left(291089 - 10 \cdot 170.5^2\right)} \approx 6.553,$$

$$s_y = \sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{10} y_i^2 - n\bar{y}^2\right)} = \sqrt{\frac{1}{10-1}\left(40793 - 10 \cdot 63.1^2\right)} \approx 10.418,$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{10} x_i y_i - n\bar{x}\bar{y}\right) = \frac{1}{10-1}\left(108132 - 10 \cdot 170.5 \cdot 63.1\right) \approx 60.722$$

$$\Rightarrow r_{xy} = \frac{s_{xy}}{s_x s_y} \approx \frac{60.722}{6.553 \cdot 10.418} \approx 0.889$$

r_{xy}:n tulkinta: Muuttujien välillä vallitsee voimakas positiivinen lineaarinen riippuvuus.

ESIMERKKI **Sovitetaan** jo aiemmin esillä olleeseen (pituus,paino)-aineistoon **regressiosuora** $y = a + bx$, missä y = paino ja x = pituus. Aineisto oli seuraava:

henkilö (i)	1	2	3	4	5	6	7	8	9	10
pituus (x_i)	175	180	165	170	168	166	165	171	163	182
paino (y_i)	68	85	52	65	60	51	55	66	57	72

Edellisessä ko. aineistoa käsitelleessä esimerkissä havaittiin korrelaatiodiagrammin avulla, että muuttujien välillä vallitsee voimakas lineaarinen riippuvuus. Lisäksi aineistosta on laskettu mm. seuraavat tunnusluvut:

$$n = 10, \quad \bar{x} = 170.5, \quad s_x \approx 6.553,$$
$$\bar{y} = 63.1, \quad s_y \approx 10.418 \quad s_{xy} \approx 60.722 \quad ja \quad r_{xy} \approx 0.889.$$

$$\Rightarrow b = \frac{s_{xy}}{s_x^2} \approx \frac{60.722}{6.553^2} \approx 1.414$$

$$a = \bar{y} - b\bar{x} \approx 63.1 - 1.414 \cdot 170.5 \approx -177.987$$

=> Regressiosuora on

$$\hat{y} = -177.987 + 1.414\, x.$$

Regressioyhtälön antama painon ennuste 178 cm pitkälle henkilölle on siis

$$\hat{y} = -177.987 + 1.414 \cdot 178 \approx 73.7 kg$$

Aineiston ensimmäisellä tilastoyksiköllä muuttujien x ja y havaitut arvot olivat 175 cm ja 68 kg. Regressioyhtälön antama ennustearvo y_1 ko. tilastoyksikön arvoksi muuttujalle y on

$$\hat{y}_1 = -177.987 + 1.414 \cdot 175 \approx 69.5 kg$$

ja jäännöstermi

$$e_1 = y_1 - \hat{y}_1 \approx 68 - 69.5 = -1.5 kg$$

Mallissa on nyt vain yksi selittävä muuttuja, joten regressiomallin selitysaste

$$\hat{R}^2 = r_{xy}^2 \approx 0.889^2 \approx 0.79.$$

Regressiokertoimien tulkinta:

b: Kun henkilön pituus kasvaa sentillä, kasvaa henkilön paino keskimäärin noin 1.414 kg.

a: Kun henkilön pituus on nolla senttiä, on henkilön paino keskimäärin noin -178 kg!

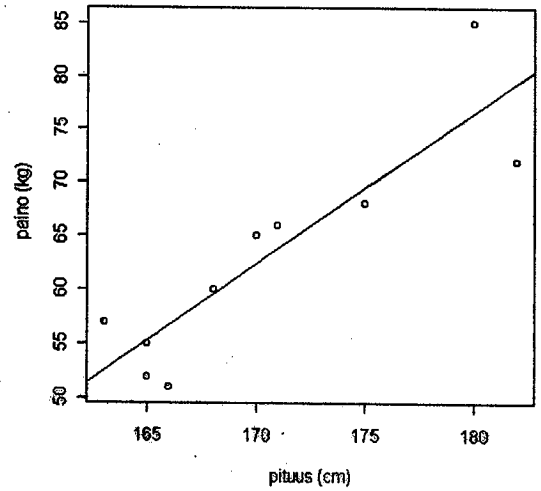
Determinaatiokertoimen eli selitysasteen tulkinta:

R^2: Regressiomallilla (henkilön pituudella) voidaan selittää noin 79 % painon kokonaisvaihtelusta.

Regressiosuoran piirtämiseksi korrelaatiodiagrammiin tarvitaan suoralta kaksi pistettä:

$$x = 165 : \hat{y} = -177.987 + 1.414 \cdot 165 \approx 55.3$$
$$x = 180 : \hat{y} = -177.987 + 1.414 \cdot 180 \approx 76.5$$

Pisteet (165,55.3) ja (180,76.5) merkitään korrelaatiodiagrammiin ja yhdistetään ne suoralla.



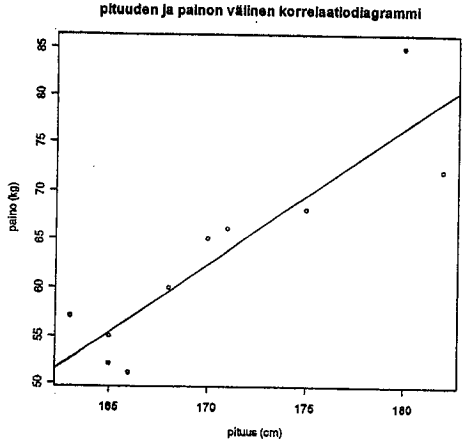
Neliösummahajotelmassa $SST = SS_{reg} + SS_{res}$: SST = vasteen kokonaisvaihtelu

SS_{reg} = regressiomallin selittämä osuus vasteen kokonaisvaihtelusta

SS_{res} = regressiomallilla selittämättä jäänyt osuus vasteen kokonaisvaihtelusta

$\hat{R}^2 = 0.6213$:n tulkinta: Regressiomallilla (eli muuttujilla AGE15_64 ja LIT_MALE) voidaan selittää noin 62 % vasteen eli BKT/asukas-muuttujan kokonaisvaihtelusta

Yhden selittävän muuttujan regressioanalyysi R-ohjelmalla:




```
> pituus <- c(175,180,165,170,168,166,165,171,163,182)
> paino <- c(68,85,52,65,60,51,55,66,57,72)

> lm(paino~pituus)

Call:
lm(formula = paino ~ pituus)

Coefficients:
(Intercept)      pituus
   -177.982        1.414

> abline(lm(paino~pituus))

> summary(lm(paino~pituus))

Call:
lm(formula = paino ~ pituus)

Residuals:
     Min       1Q   Median       3Q      Max
-7.36067 -2.85809  0.05589  2.50349  8.46727

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -177.982     43.841  -4.060 0.003635 **
pituus         1.414      0.257   5.503 0.000572 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 5.052 on 8 degrees of freedom
Multiple R-Squared: 0.791,      Adjusted R-squared: 0.7649
F-statistic: 30.28 on 1 and 8 DF,  p-value: 0.000572
```

lm(vaste ~ selittäjä) ← komento muotoa

a, b

$\Rightarrow \hat{y} = -177.982 + 1.414x$

saadaan lisää tietoa mallinnuksesta

regressiokertoimet a ja b

determinaatiokerroin $\hat{R}^2$

ESIMERKKI kahden selittäjän regressioanalyysistä, aineistona WORLD2005-aineiston Afrikan maat.

Muuttujat:

y = GDP_CAPI: BKT/asukas, US dollaria
x_1 = AGE15_64: 15-64 -vuotiaiden prosentuaalinen osuus maan väestöstä
x_2 = LIT_MALE: miesten lukutaitoprosentti

Komennot:

```
malli_1 <- lm(GDP_CAPI~AGE15_64+LIT_MALE, data=afrikka)
summary(malli_1)
```

$\Rightarrow y = b_0 + b_1x_1 + b_2x_2$

Tulostus:

```
Call:
lm(formula = GDP_CAPI ~ AGE15_64 + LIT_MALE, data = afrikka)

Residuals:
    Min      1Q  Median      3Q     Max
-3741.6 -1056.8  -320.3   731.3  5478.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -22154.89    2753.06  -8.047 1.41e-10 ***
AGE15_64       432.31      52.58   8.222 7.55e-11 ***
LIT_MALE        13.22      16.56   0.799    0.428
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 1799 on 50 degrees of freedom
Multiple R-Squared: 0.6213,     Adjusted R-squared: 0.6062
F-statistic: 41.02 on 2 and 50 DF,  p-value: 2.863e-11
```

b_0, b_1 ja b_2 $\Rightarrow \hat{y} = -22154.89 + 432.31 \cdot AGE15\_64 + 13.22 \cdot LIT\_MALE$

$\hat{R}^2$

Tulkinnat: b_0: Kun työikäisten suhteellinen osuus on nolla ja miesten lukutaitoprosentti on nolla, on maan BKT/asukas keskimäärin noin -22154 dollaria (ei käytännössä mielekäs tulos!)

b_1: Kun työikäisten suhteellinen osuus kasvaa yhdellä prosenttiyksiköllä ja samaan aikaan miesten lukutaitoprosentti ei muutu, kasvaa maan BKT/asukas keskimäärin noin 432 dollarilla.

b_2: Kun miesten lukutaitoprosentti kasvaa yhdellä prosenttiyksiköllä ja samaan aikaan työikäisten suhteellinen osuus ei muutu, kasvaa maan BKT/asukas keskimäärin noin 13 dollaria