Tilastolliset menetelmät

Osa 4: Lineaarinen regressioanalyysi

- **Tilastollinen riippuvuus ja korrelaatio**

Tilastollinen riippuvuus ja korrelaatio

>> Tilastollinen riippuvuus, korrelaatio ja regressio

Kahden muuttujan havaintoaineiston kuvaaminen

Pearsonin korrelaatiokertoimen estimointi ja testaus

Muuttujien väliset riippuvuudet tilastollisen tutkimuksen kohteena

- Tieteellisen tutkimuksen *tärkeimmät* ja *mielen-kiintoisimmat kysymykset liittyvät* tavallisesti tutkimuksen kohteena olevaa ilmiötä kuvaavien **muuttujien välisiin riippuvuuksiin**.
- Jos tilastollisen tutkimuksen kohteena olevaan ilmiöön liittyy useampia kuin yksi muuttuja, *yhden muuttujan tilastolliset menetelmät* antavat tavallisesti vain *rajoittuneen kuvan* ilmiöstä.
- Sovellusten kannalta ehkä merkittävin osa tilastotiedettä käsittelee kahden tai useamman muuttujan välisten *riippuvuuksien kuvaamista* ja *mallintamista*.

Eksakti *vs* tilastollinen riippuvuus

- Tarkastelemme tässä esityksessä yksinkertaisuuden vuoksi pääasiassa *kahden* muuttujan välistä riippuvuutta:

  (i) Sanomme, että muuttujien välinen *riippuvuus on* **eksaktia**, jos *toisen arvot voidaan ennustaa tarkasti toisen saamien arvojen perusteella*.

  (ii) Sanomme, että muuttujien välinen *riippuvuus on* **tilastollista**, jos niiden välillä *ei ole eksaktia riippuvuutta*, mutta *toisen muuttujan arvoja voidaan käyttää apuna toisen muuttujan arvojen ennustamisessa*.

Tilastollinen riippuvuus ja korrelaatio

- Kahden muuttujan välistä *lineaarista tilastollista riippuvuutta* kutsutaan tilastotieteessä tavallisesti **korrelaatioksi**.

- *Korrelaation* eli *lineaarisen tilastollisen riippuvuuden voimakkuutta* mittaavia tilastollisia tunnuslukuja kutsutaan **korrelaatiokertoimiksi**.

- Korrelaatiot muodostavat *perustan muuttujien välisten riippuvuuksien ymmärtämiselle*.

Tilastollinen riippuvuus ja regressio

- Vaikka korrelaatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle, riippuvuuksia halutaan tavallisesti *analysoida* myös *tarkemmin*.

- **Regressioanalyysi** on tilastollinen menetelmä, jossa jonkin, ns. *selitettävän muuttujan* tilastollista riippuvuutta joistakin toisista, ns. *selittävistä muuttujista* pyritään mallintamaan **regressiomalliksi** kutsutulla tilastollisella mallilla.

Kahden muuttujan havaintoaineiston kuvaaminen

- Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useamman muuttujan havainto-aineistojen kuvaamiselle muodostaa tutustuminen **havaintoarvojen jakaumaan**.
- Havaintoarvojen jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:
  - Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valituilla **graafisilla esityksillä**.
  - Havaintoarvojen *jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **otostunnusluvuilla**.

Tilastollinen riippuvuus ja korrelaatio

Tilastollinen riippuvuus, korrelaatio ja regressio

>> Kahden muuttujan havaintoaineiston kuvaaminen

Pearsonin korrelaatiokertoimen estimointi ja testaus

Pistediagrammi

- Tarkastellaan tilannetta, jossa tutkimuksen kohteina olevista *havaintoyksiköistä* on mitattu *kahden järjestys-*, *välimatka-* tai *suhdeasteikollisen* muuttujan *x* ja *y* arvot.
- Muuttujien *x* ja *y* arvojen samaan havaintoyksikköön liittyvien *parien* muodostamaa havaintoaineistoa voidaan kuvata graafisesti **pistediagrammilla**.

- Pistediagrammi sopii erityisesti kahden muuttujan välisen *riippuvuuden* havainnollistamiseen.
- Pistediagrammi on keskeinen työväline *korrelaatio-* ja *regressioanalyysissa*.

Pistediagrammi: Määritelmä

- Olkoot x ja y *järjestys-*, *välimatka-* tai *suhdeasteikollisia* muuttujia, joiden havaitut arvot ovat

  $x_1, x_2, \ldots, x_n$

  $y_1, y_2, \ldots, y_n$
- Oletetaan lisäksi, että havaintoarvot x_i ja y_i liittyvät *samaan havaintoyksikköön* kaikille $i = 1, 2, \ldots, n$.
- Havaintoarvojen $x_1, x_2, \ldots, x_n$ ja $y_1, y_2, \ldots, y_n$ parien **pistediagrammi** saadaan esittämällä *lukuparit*

  (x_i, y_i), $i = 1, 2, \ldots, n$

  pisteinä avaruudessa $\mathbb{R}^2$.

Pistediagrammi: Havainnollistus

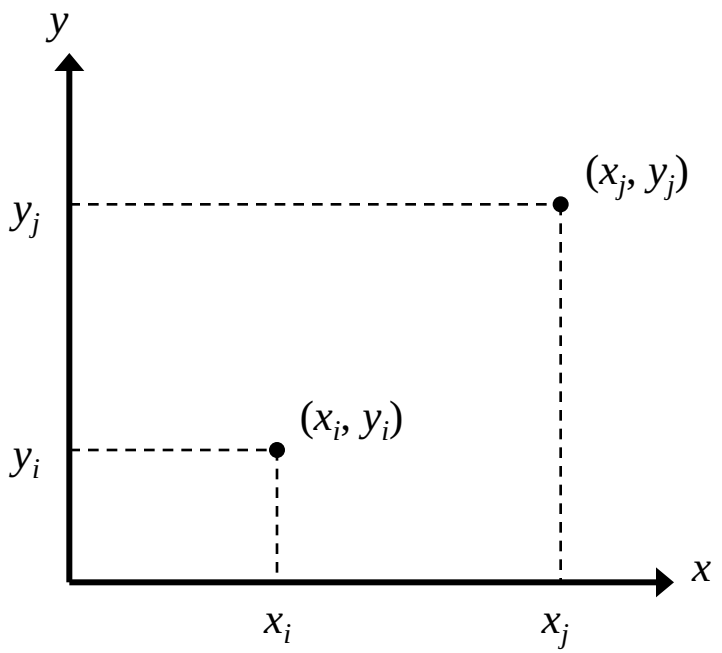
- Kuvio oikealla esittää lukuparien

  (x_i , y_i)

  ja

  (x_j , y_j)

  määrittelemien pisteiden esittämistä tasokoordinaatistossa.



Pistediagrammi: 1. esimerkki – 1/2

- *Hooken lain* mukaan kierrejousen pituus riippuu *lineaarisesti* jouseen ripustetusta painosta.
- Oikealla on tulokset kokeesta, jossa Hooken lain pätevyyttä tutkittiin ripustamalla jouseen 6 erikokoista painoa.
- Merkitään:

  (x_i , y_i) , $i = 1, 2, 3, 4, 5, 6$

  jossa

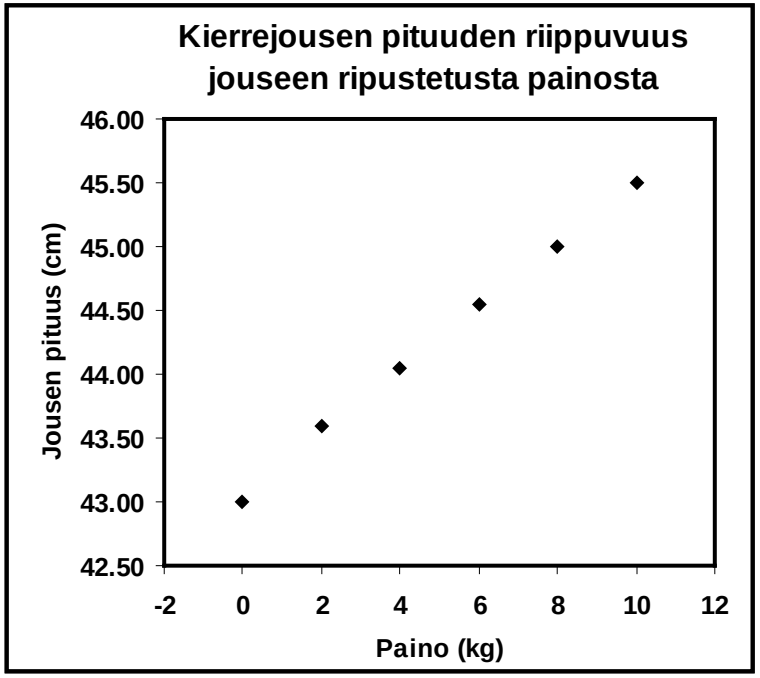
  x_i = paino i

  y_i = jousen pituus, kun painona on x_i

Paino (kg)	Pituus (cm)
0	43.00
2	43.60
4	44.05
6	44.55
8	45.00
10	45.50

Pistediagrammi: 1. esimerkki – 2/2

- Pistediagrammi oikealla havainnollistaa koetuloksia graafisesti.
- Ovatko havainnot *sopusoinnussa* Hooken lain kanssa?
- Vastausta tarkastellaan luvussa **Yhden selittäjän lineaarinen regressiomalli**.



Pistediagrammi: 2. esimerkki – 1/2

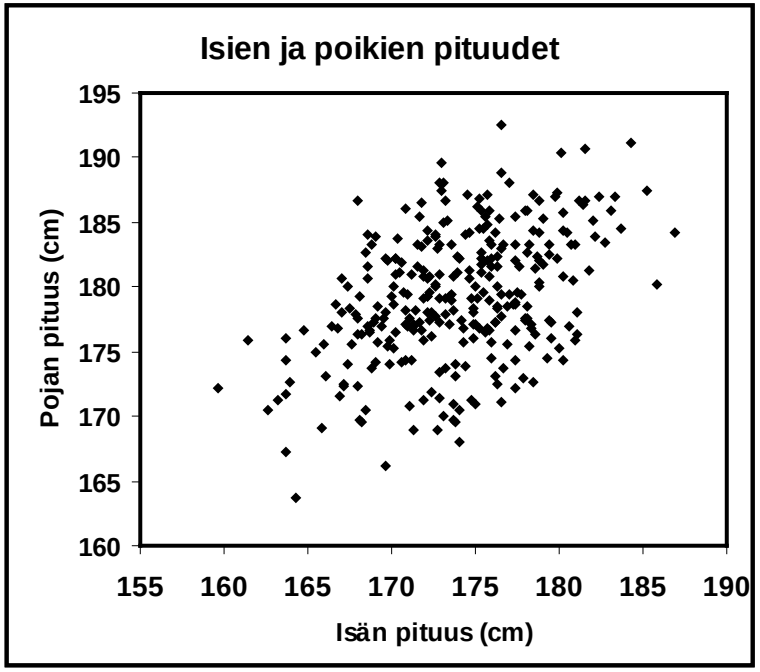
- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:n isän ja heidän poikiensa pituuksien muodostamasta lukuparista

  (x_i , y_i) , $i = 1, 2, \ldots , 300$

  jossa

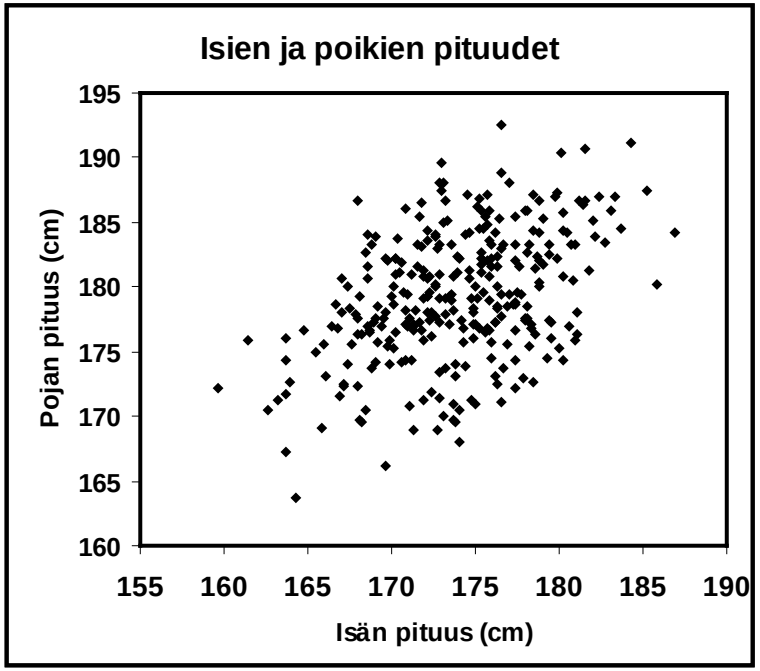
  x_i = isän i pituus

  y_i = isän i pojan pituus
- Ks. pistediagrammia oikealla.



Pistediagrammi: 2. esimerkki – 2/2

- Yhtä pitkillä isillä näyttää olevan monen mittaisia poikia.
- Mutta: Lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkillä isillä ja pitkillä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.
- Tällaisten *tilastollisten riippuvuuksien* analysoimista lineaaristen regressiomallien avulla tarkastellaan luvussa **Yhden selittäjän lineaarinen regressiomalli**.



Tunnusluvut

- Kahden *välimatka-* tai *suhdeasteikollisen* muuttujan havaintoarvojen parien muodostamaa jakaumaa voidaan *karakterisoida* seuraavilla *tunnusluvuilla*:

  - Havaintoarvojen keskimääräistä *sijaintia* kuvataan **aritmeettisilla keskiarvoilla**.
  - Havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* kuvataan (**otos-**) **keskihajonnoilla** tai (**otos-**) **variansseilla**.
  - Havaintoarvojen lineaarista riippuvuutta kuvataan **otoskovarianssilla** ja **otoskorrelaatiokertoimella**.

Havainnot

- Olkoot

  $$x_1, x_2, \ldots, x_n$$

  ja

  $$y_1, y_2, \ldots, y_n$$

  välimatka- tai *suhdeasteikollisten* muuttujien x ja y havaittuja arvoja.
- Oletetaan lisäksi, että havaintoarvot x_i ja y_i liittyvät *samaan havaintoyksikköön i kaikille* $i = 1, 2, \ldots, n$.

Aritmeettiset keskiarvot: Määritelmät

- Havaintoarvojen $x_1, x_2, \ldots, x_n$ **aritmeettinen keskiarvo** on

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- Havaintoarvojen $y_1, y_2, \ldots, y_n$ **aritmeettinen keskiarvo** on

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

Aritmeettiset keskiarvot: Tulkinnat

- Havaintoarvojen pareista

  $$(x_i , y_i) , i = 1, 2, \ldots , n$$

  laskettujen aritmeettisten keskiarvojen $\bar{x}$ ja $\bar{y}$ muodostama lukupari

  $$(\bar{x}, \bar{y})$$

  on havaintoarvojen parien (x_i , y_i) muodostamien pistejoukon *painopiste*.
- Havaintoarvojen aritmeettinen keskiarvo kuvaa havainto-arvojen *keskimääräistä sijaintia*.

Varianssit: Määritelmät

- Havaintoarvojen $x_1, x_2, \ldots, x_n$ (otos-) **varianssi** on

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$$

  jossa $\bar{x}$ on x-havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen $y_1, y_2, \ldots, y_n$ (otos-) **varianssi** on

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2$$

  jossa $\bar{y}$ on y-havaintoarvojen aritmeettinen keskiarvo.

- Havaintoarvojen varianssi mittaa havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havaintoarvojen aritmeettisen keskiarvon suhteen.

Keskihajonnat: Määritelmät

- Havaintoarvojen $x_1, x_2, \ldots, x_n$ (otos-) **keskihajonta** on

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

  jossa $\bar{x}$ on x-havaintoarvojen aritmeettinen keskiarvo.
- Havaintoarvojen $y_1, y_2, \ldots, y_n$ (otos-) **keskihajonta** on

$$s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

  jossa $\bar{y}$ on y-havaintoarvojen aritmeettinen keskiarvo.
- Havaintoarvojen keskihajonta mittaa havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havaintoarvojen aritmeettisen keskiarvon suhteen.

Otoskovarianssi: Määritelmä

- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ laskettu **otoskovarianssi** on

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$$

jossa

$\bar{x}$ = x-havaintoarvojen aritmeettinen keskiarvo

$\bar{y}$ = y-havaintoarvojen aritmeettinen keskiarvo

- x- ja y-havaintoarvojen otoskovarianssit niiden itsensä kanssa ovat niiden *variansseja*:

$$s_{xx} = s_x^2$$

$$s_{yy} = s_y^2$$

Otoskovarianssi: Tulkinta

- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ laskettu otoskovarianssi s_{xy} mittaa x- ja y-havaintoarvojen *yhteisvaihtelua* niiden aritmeettisten keskiarvojen muodostaman pisteen ympärillä.
- Mitä suurempi on otoskovarianssin s_{xy} itseisarvo

  $$|s_{xy}|$$

  sitä voimakkaampaa on x- ja y-havaintoarvojen yhteisvaihtelu.

Otoskovarianssi: Merkin määräytyminen 1/4

- Otoskovarianssin s_{xy} *merkin* määrää summalauseke

  (1) $\sum(x_i - \bar{x})(y_i - \bar{y})$

- Summalausekkeen (1) *i*. termin

  $$(x_i - \bar{x})(y_i - \bar{y})$$

  itseisarvo

  $$|x_i - \bar{x}||y_i - \bar{y}|$$

  on sellaisen *suorakaiteen pinta-ala*, jonka sivujen pituudet ovat

  $$|x_i - \bar{x}|$$

  ja

  $$|y_i - \bar{y}|$$

Otoskovarianssi: Merkin määräytyminen 2/4

- Summalausekkeen (1) *i*. termin

$$(x_i - \bar{x})(y_i - \bar{y})$$

merkki määräytyy seuraavalla tavalla:

$$(x_i - \bar{x})(y_i - \bar{y}) \geq 0 \quad \begin{cases} \text{jos } x_i \geq \bar{x} \text{ ja } y_i \geq \bar{y} \\ \text{jos } x_i \leq \bar{x} \text{ ja } y_i \leq \bar{y} \end{cases}$$

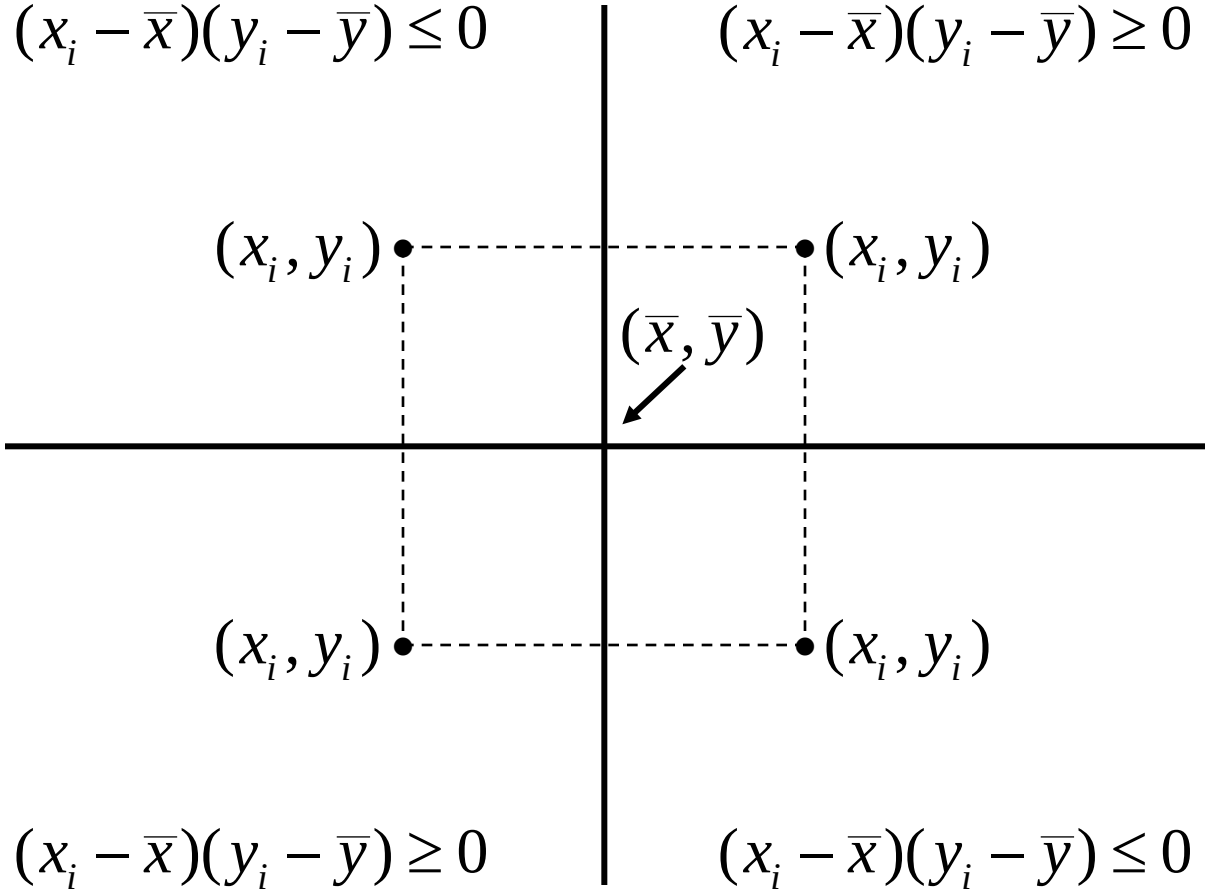
$$(x_i - \bar{x})(y_i - \bar{y}) \leq 0 \quad \begin{cases} \text{jos } x_i \geq \bar{x} \text{ ja } y_i \leq \bar{y} \\ \text{jos } x_i \leq \bar{x} \text{ ja } y_i \geq \bar{y} \end{cases}$$

- Merkin määräytymistä voidaan *havainnollistaa geometrisesti* seuraavalla tavalla (ks. kuviota seuraavalla kalvolla):

  (i) Jaetaan *xy*-taso neljään osaan eli *neljännekseen* pisteen $(\bar{x}, \bar{y})$ kautta piirretyillä koordinaattiakseleiden suuntaisilla suorilla.

  (ii) Termin $(x_i - \bar{x})(y_i - \bar{y})$ *merkin* määrää se, *mihin neljännekseen havaintopiste* (x_i , y_i) *sijoittuu*.

Otoskovarianssi: Merkin määräytyminen 3/4



Otoskovarianssi: Merkin määräytyminen 4/4

- Jos positiiviset termit summalausekkeeseen

  (1) $\sum (x_i - \bar{x})(y_i - \bar{y})$

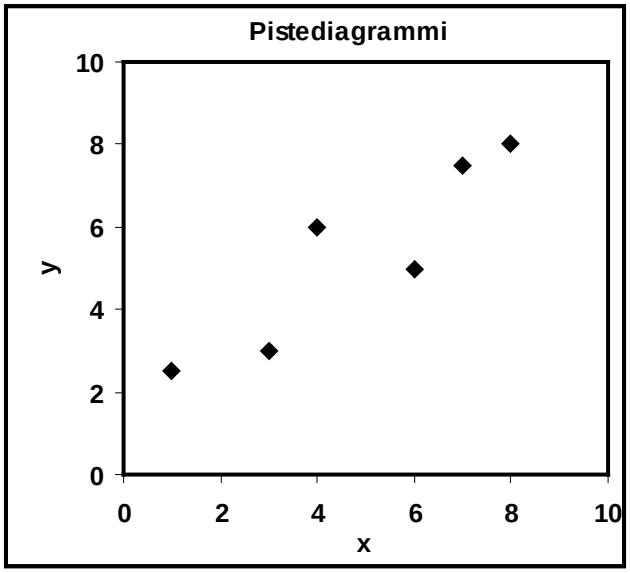
  tuottavien suorakaiteiden yhteenlaskettu pinta-ala on *suurempi* (*pienempi*) kuin negatiiviset termit tuottavien suorakaiteiden yhteenlaskettu pinta-ala, otoskovarianssin s_{xy} merkki on *positiivinen* (*negatiivinen*).

- Siten otoskovarianssilla on taipumus saada *positiivisia* (*negatiivisia*) arvoja, jos havaintopisteiden muodostama pistepilvi tai -parvi *näyttää nousevalta* (*laskevalta*) *oikealle mentäessä*.

Otoskovarianssi: Merkin määräytyminen
Esimerkki 1/2

- Taulukossa oikealla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$).
- Aineistoa kuvaava *pistediagrammi* on oikealla alhaalla.

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8

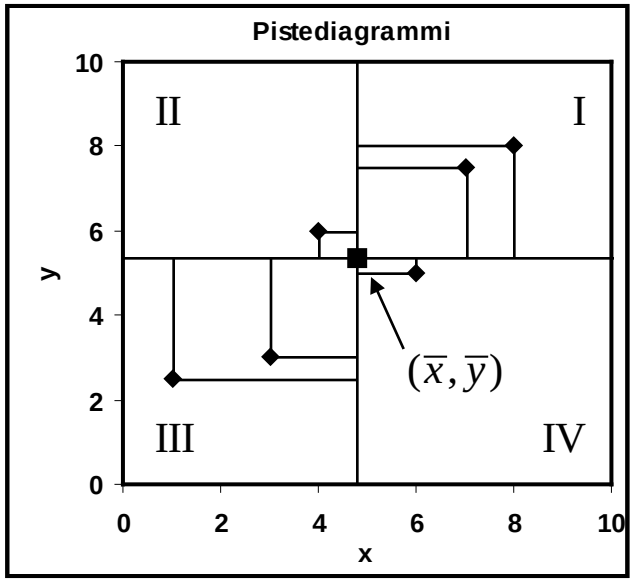


Otoskovarianssi: Merkin määräytyminen Esimerkki 2/2

- Kuvioon oikealla on lisätty havainto-pisteiden *painopiste*

$$(\overline{x}, \overline{y}) = (4.833, 5.333)$$

- Lisäksi kuvioon on piirretty painopisteen kautta kulkevat koordinaattiakseleiden suuntaiset suorat sekä *otoskovarianssin merkin määräytymistä* havainnollistavat suorakaiteet.
- Otoskovarianssi on *positiivinen*, koska I ja III neljänneksen suorakaiteiden yhteenlaskettu pinta-ala on *suurempi* kuin II ja IV neljänneksen suora-kaiteiden yhteenlaskettu pinta-ala.



Otoskovarianssi ja Pearsonin otoskorrelaatiokerroin: Määritelmä 1/2

- Määritellään otoskovarianssin avulla *x*- ja *y*-havaintoarvojen *lineaarisen tilastollisen riippuvuuden voimakkuuden mittari:*
- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ laskettu **Pearsonin otoskorrelaatiokerroin** on

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

jossa

s_{xy} = *x*- ja *y*-havaintoarvojen otoskovarianssi

s_x = *x*-havaintoarvojen keskihajonta

s_y = *y*-havaintoarvojen keskihajonta

Pearsonin otoskorrelaatiokerroin: Määritelmä 2/2

- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ laskettu *Pearsonin otoskorrelaatiokerroin* voidaan kirjoittaa myös muotoon

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

jossa

$\bar{x}$ = x-havaintoarvojen aritmeettinen keskiarvo

$\bar{y}$ = y-havaintoarvojen aritmeettinen keskiarvo

Pearsonin otoskorrelaatiokerroin: Ominaisuuksia

- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ lasketulla *Pearsonin otoskorrelaatiokertoimella* r_{xy} on seuraavat ominaisuudet:

  (i) $-1 \leq r_{xy} \leq +1$

  (ii) $r_{xy} = \pm 1$, jos ja vain jos

  $$y_i = \alpha + \beta x_i$$

  jossa α ja β ovat reaalisia *vakiota* ja $\beta \neq 0$.

  Lisäksi $\text{sgn}(\beta) = \text{sgn}(r_{xy})$

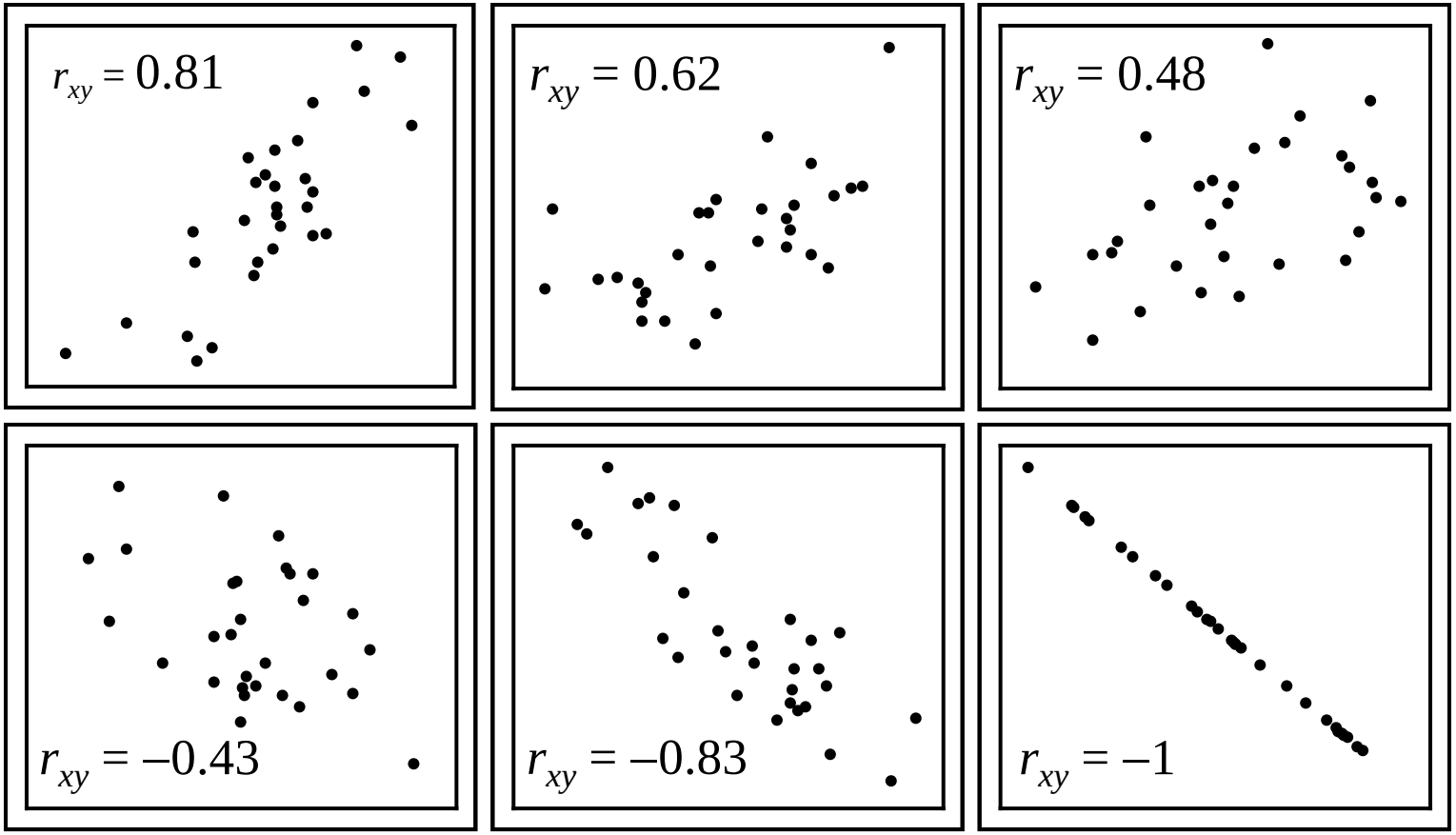
  (iii) Korrelaatiokertoimella r_{xy} ja kovarianssilla s_{xy} on aina *sama merkki*.

Pearsonin otoskorrelaatiokerroin: Tulkinta

- Havaintoarvojen pareista (x_i , y_i) , $i = 1, 2, \ldots , n$ laskettu *Pearsonin otoskorrelaatiokerroin* r_{xy} mittaa x- ja y-havaintoarvojen *lineaarisen tilastollisen riippuvuuden voimakkuutta*.
- Jos $r_{xy} = \pm 1$, niin x- ja y-havaintoarvojen välillä *on eksakti* eli *funktionaalinen lineaarinen riippuvuus*, mikä merkitsee sitä, että kaikki havaintopisteet (x_i , y_i) asettuvat samalle suoralle.
- Jos $r_{xy} = 0$, niin x- ja y-havaintoarvojen välillä *ei voi olla eksaktia lineaarista riippuvuutta*.
- Vaikka $r_{xy} = 0$, x- ja y-havaintoarvojen välillä *saattaa silti olla jopa eksakti epälineaarinen riippuvuus*.

Pearsonin otoskorrelaatiokerroin: Havainnollistus

- Kuviot alla havainnollistavat kahden muuttujan havaittujen arvojen (n = 30) *pistediagrammin ilmeen* ja *korrelaation* välistä yhteyttä.



Tunnuslukujen laskeminen 1/4

- Oletetaan, että haluamme laskea havaintoarvojen pareista

  $(x_i , y_i) , i = 1, 2, \ldots , n$

  seuraavat otostunnusluvut *käsin* tai käyttämällä *laskinta*:

  (i) *Aritmeettiset keskiarvot*: $\bar{x}$, $\bar{y}$

  (ii) *Varianssit*: s_x^2 , s_y^2

  (iii) *Keskihajonnat*: s_x , s_y

  (iv) *Kovarianssi*: s_{xy}

  (v) *Korrelaaatio*: r_{xy}

- Tällöin tarvittavat laskutoimitukset on mukavinta järjestää seuraavalla kalvolla esitettävän kaavion muotoon.

Tunnuslukujen laskeminen 2/4

- Määrätään ensin havaintoarvojen *summat*, *neliösummat* ja *tulosumma*:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	x_1	y_1	x_1^2	y_1^2	$x_1 y_1$
2	x_2	y_2	x_2^2	y_2^2	$x_2 y_2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_n	y_n	x_n^2	y_n^2	$x_n y_n$
Summa	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$

Tunnuslukujen laskeminen 3/4

- Havaintoarvojen *aritmeettiset keskiarvot*, *varianssit* ja *kovarianssi* saadaan havaintoarvojen *summista*, *neliö-summista* ja *tulosummasta* alla esitetyillä kaavoilla:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s_x^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right)$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right)$$

Tunnuslukujen laskeminen 4/4

- Havaintoarvojen *keskihajonnat* ja *Pearsonin otos-korrelaatiokerroin* saadaan havaintoarvojen *variansseista* ja *kovarianssista* alla esitetyillä kaavoilla:

$$s_x = \sqrt{s_x^2}$$

$$s_y = \sqrt{s_y^2}$$

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

Tilastollinen riippuvuus ja korrelaatio

Tilastollinen riippuvuus, korrelaatio ja regressio

Kahden muuttujan havaintoaineiston kuvaaminen

>> Pearsonin korrelaatiokertoimen estimointi ja testaus

Satunnaismuuttujien kovarianssi ja korrelaatio 1/2

- Olkoon

  (X, Y)

  satunnaismuuttujien X ja *Y* muodostama järjestetty *pari*.
- Olkoot

  $\mu_X = \mathrm{E}(X)$

  $\mu_Y = \mathrm{E}(Y)$

  satunnaismuuttujien *X* ja *Y odotusarvot* ja

  $\sigma_X^2 = \mathrm{Var}(X) = \mathrm{D}^2(X) = \mathrm{E}[(X - \mu_X)^2]$

  $\sigma_Y^2 = \mathrm{Var}(Y) = \mathrm{D}^2(Y) = \mathrm{E}[(Y - \mu_Y)^2]$

  satunnaismuuttujien *X* ja *Y varianssit*.

Satunnaismuuttujien kovarianssi ja korrelaatio 2/2

- Määritellään satunnaismuuttujien X ja Y **kovarianssi** σ_{XY} kaavalla

$$\sigma_{XY} = \mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)]$$

- Määritellään satunnaismuuttujien X ja Y **korrelaatio** ρ_{XY} kaavalla

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

jossa

$$\sigma_X = \mathrm{D}(X) = \sqrt{\sigma_X^2}$$

$$\sigma_Y = \mathrm{D}(Y) = \sqrt{\sigma_Y^2}$$

Satunnaismuuttujien korrelaatio

- Satunnaismuuttujien X ja Y korrelaatiota

  $$\rho_{XY} = \mathrm{Cor}(X, Y)$$

  kutsutaan tavallisesti **Pearsonin** (**tulomomentti-**) **korrelaatiokertoimeksi**.
- Pearsonin korrelaatiokerroin ρ_{XY} mittaa satunnaismuuttujien X ja Y *lineaarisen riippuvuuden voimakkuutta*.
- **Huomautus:**

  Tutustuimme Pearsonin korrelaatiokertoimeen todennäköisyyslaskennan luennossa **Moniulotteiset satunnaismuuttujat ja jakaumat**.

Pearsonin korrelaatiokertoimen estimointi 1/3

- **Oletetaan**, että satunnaismuuttujien X ja Y muodostama järjestetty pari (X, Y) noudattaa ***2-ulotteista normaalijakaumaa*** $\mathrm{N}_2(\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2, \rho_{XY})$, jossa

$$\mu_X = \mathrm{E}(X) \qquad \mu_Y = \mathrm{E}(Y)$$

$$\sigma_X^2 = \mathrm{Var}(X) \qquad \sigma_Y^2 = \mathrm{Var}(Y)$$

$$\rho_{XY} = \mathrm{Cor}(X, Y)$$

- Olkoon

$$(X_i, Y_i),\ i = 1, 2, \ldots, n$$

riippumaton satunnaisotos satunnaismuuttujien X ja Y muodostaman parin (X, Y) jakaumasta.

Pearsonin korrelaatiokertoimen estimointi 2/3

- Olkoot

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

tavanomaiset havaintoarvojen pareista (X_i, Y_i), $i = 1, 2, \ldots, n$ lasketut *otostunnusluvut*.

Pearsonin korrelaatiokertoimen estimointi 3/3

- Satunnaismuuttujien X ja Y *Pearsonin* (*tulomomentti-*) *korrelaatiokerroin*

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

voidaan **estimoida** vastaavalla *Pearsonin otoskorrelaatio-kertoimella*

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Pearsonin korrelaatiokerroin

- Estimaattori r_{XY} voidaan johtaa *suurimman uskottavuuden menetelmällä*.
- **Luottamusvälit** ja **testit Pearsonin tulomomenttikorrelaatiokertoimelle** ρ_{XY} voidaan konstruoida *samantapaisella tekniikalla* kuin luottamusvälit ja testit konstruoidaan *normaalijakauman odotusarvolle*.
- Tässä esityksessä tarkastellaan vain yhtä testiä Pearsonin tulomomenttikorrelaatiokertoimelle ρ_{XY}. Se on erikoistapaus *Yhden otoksen testistä korrelaatiokertoimelle:* **Korreloimattomuuden testaaminen**. (Testataan *nollahypoteesia*

$$\mathrm{H}_0 : \rho_{XY} = \rho_0$$

jossa $\rho_0 = 0$).

Korreloimattomuuden testaaminen: Testausasetelma

- Monissa tutkimustilanteissa ollaan kiinnostuneita siitä ovatko satunnaismuuttujat *X* ja *Y korreloimattomia* vai ei.
- **Huomautuksia:**
  - **Satunnaismuuttujien *X* ja *Y korreloimattomuudesta ei välttämättä seuraa niiden riippumattomuus*, vaikka satunnaismuuttujien *X* ja *Y riippumattomuudesta seuraa aina niiden korreloimattomuus*.**
  - Jos satunnaismuuttujat *X* ja *Y* noudattavat 2-*ulotteista normaalijakaumaa*, satunnaismuuttujien *X* ja *Y korreloimattomuudesta seuraa niiden riippumattomuus*.
  - Monissa tutkimusasetelmissa toivotaan, että korreloimattomuusoletus *tulee* testissä *hylätyksi*.

Korreloimattomuuden testaaminen: Yleinen hypoteesi

- *Yleinen hypoteesi* H :

  (i) Oletetaan, että satunnaismuuttujien X ja Y järjestetty pari (X, Y) noudattaa 2-*ulotteista normaalijakaumaa*, jonka parametrit ovat

  $$\mu_X = \mathrm{E}(X) \qquad \mu_Y = \mathrm{E}(Y)$$
  $$\sigma_X^2 = \mathrm{Var}(X) \qquad \sigma_Y^2 = \mathrm{Var}(Y)$$
  $$\rho_{XY} = \mathrm{Cor}(X, Y)$$

  (ii) Olkoon

  $$(X_i, Y_i),\ i = 1, 2, \ldots, n$$

  satunnaisotos satunnaismuuttujien X ja Y muodostaman parin (X, Y) jakaumasta.

Korreloimattomuuden testaaminen: Nollahypoteesi ja vaihtoehtoinen hypoteesi

- *Nollahypoteesi* H_0 :

  $$\mathrm{H}_0 : \rho_{XY} = 0$$

- *Vaihtoehtoinen hypoteesi* H_1 :

  $$\left.\begin{array}{l} \mathrm{H}_1 : \rho_{XY} > 0 \\ \mathrm{H}_1 : \rho_{XY} < 0 \end{array}\right\}$$ 1-suuntaiset vaihtoehtoiset hypoteesit

  $$\mathrm{H}_1 : \rho_{XY} \neq 0$$ 2-suuntainen vaihtoehtoinen hypoteesi

Korreloimattomuuden testaaminen: Parametrien estimointi

- Estimoidaan 2-ulotteisen normaalijakauman parametrit *tavanomaisilla estimaattoreillaan*:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

$$s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$$r_{XY} = \frac{s_{XY}}{s_X s_Y}$$

Korreloimattomuuden testaaminen: Testisuure ja sen jakauma

- Määritellään ***t*-testisuure**

$$t = \sqrt{n-2}\,\frac{r_{XY}}{\sqrt{1-r_{XY}^2}}$$

- *Jos nollahypoteesi*

$$\mathrm{H}_0 : \rho_{XY} = 0$$

pätee, testisuure *t noudattaa Studentin t-jakaumaa*, jonka vapausasteluku on $n - 2$:

$$t \sim t(n-2)$$

Korreloimattomuuden testaaminen: Testi

- Testisuureen *t normaaliarvo* = 0, koska *nollahypoteesin* $H_0 : \rho_{XY} = 0$ *pätiessä*

  $E(t) = 0$

- Siten itseisarvoltaan *suuret* testisuureen *t* arvot viittaavat siihen, että *nollahypoteesi* H_0 *ei päde*.
- Nollahypoteesi H_0 *hylätään*, jos testin *p*-arvo on *kyllin pieni*.

Osa 4: Lineaarinen regressioanalyysi

- **Yhden selittäjän lineaarinen regressiomalli (Osa 1)**

Yhden selittäjän lineaarinen regressiomalli

>> Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset
Yhden selittäjän lineaarisen regressiomallin estimointi
Varianssianalyysihajotelma ja selitysaste

Selitettävä muuttuja ja selittävä muuttuja

- Oletetaan, että **selitettävän muuttujan** *y havaittujen arvojen vaihtelu halutaan selittää* **selittävän muuttujan** eli **selittäjän** *x havaittujen arvojen vaihtelun avulla*.
- Tehdään seuraavat oletukset:
  (i) Selitettävä muuttuja *y* on *suhdeasteikollinen satunnaismuuttuja*.
  (ii) Selittävä muuttuja *x* on *kiinteä* eli *ei-satunnainen muuttuja*.
- *Satunnaisen selittäjän tapausta* käsitellään tämän luvun lopussa (Osa 2) kappaleissa **Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä** ja **2-ulotteisen normaalijakauman regressiofunktioiden estimointi.**

Havainnot

- Olkoot

  $$y_1, y_2, \ldots, y_n$$

  selitettävän muuttujan y ja

  $$x_1, x_2, \ldots, x_n$$

  selittävän muuttujan x **havaittuja arvoja**.
- Oletetaan lisäksi, että havaintoarvot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$.
- Tällöin havaintoarvot x_i ja y_i muodostavat *pisteitä* 2-ulotteisessa avaruudessa:

  $$(x_i, y_i) \in \mathbb{R}^2 \, , i = 1, 2, \ldots, n$$

Malli ja sen osat 1/2

- Oletetaan, että havaintoarvojen y_i ja x_i välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

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

- Yhtälö (1) määrittelee **yhden selittäjän lineaarisen regressiomallin**, jossa

  y_i = **selitettävän muuttujan** *y satunnainen* ja *havaittu* arvo havaintoyksikössä *i*

  x_i = **selittävän muuttujan** eli **selittäjän** *x ei-satunnainen* ja *havaittu* arvo havaintoyksikössä *i*

  ε_i = **jäännös-** eli **virhetermin** ε *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä *i*

Malli ja sen osat 2/2

- Yhden selittäjän lineaarisessa regressiomallissa

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

  on seuraavat *regressiokertoimet*:

  β_0 = **vakioselittäjän regressiokerroin**;
  β_0 on *ei-satunnainen* ja *tuntematon vakio*

  β_1 = **selittäjän** *x* **regressiokerroin**;
  β_1 on *ei-satunnainen* ja *tuntematon vakio*

- Kutsumme yhtälön (1) määrittelemää mallia **tavalliseksi** yhden selittäjän lineaariseksi regressiomalliksi.
- **Huomautus:**

  Jatkossa esitettävät kaavat *eivät välttämättä päde* tässä esitettävässä muodossa, jos mallissa *ei ole* vakioselittäjää.
- **Oletamme jatkossa, että mallissa on aina vakioselittäjä.**

Standardioletukset jäännöstermeistä 1/2

- Tehdään tavallisen yhden selittäjän lineaarisen regressiomallin

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

jäännös- eli *virhetermeistä* ε_i ns. **standardioletukset**:

(i) $\mathrm{E}(\varepsilon_i) = 0 \ , i = 1, 2, \ldots, n$

(ii) Jäännöstermeillä on *vakiovarianssi* eli ne ovat *homoskedastisia*:

$\mathrm{Var}(\varepsilon_i) = \sigma^2 \ , i = 1, 2, \ldots, n$

(iii) Jäännöstermit ovat *korreloimattomia*:

$\mathrm{Cor}(\varepsilon_i, \varepsilon_l) = 0 \ , i \neq l$

Standardioletukset jäännöstermeistä 2/2

- Lisäksi jäännös- eli virhetermeistä ε_i tehdään tavallisesti *normaalisuusoletus*:

  (iv) $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $i = 1, 2, \ldots, n$

- **Huomautus:**

  Oletus (iv) sisältää oletukset (i) ja (ii).

Selitettävän muuttujan ominaisuudet

- Jos tavallisen yhden selittäjän lineaarisen regressiomallin

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

jäännös- eli *virhetermejä* ε_i *koskevat standardioletukset* (i)-(iii) *pätevät*, mallin selitettävän muuttujan *y* havaituilla arvoilla y_i on seuraavat *stokastiset ominaisuudet*:

(i)´ $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \ , i = 1, 2, \ldots, n$

(ii)´ $\mathrm{Var}(y_i) = \sigma^2 \ , i = 1, 2, \ldots, n$

(iii)´ $\mathrm{Cor}(y_i, y_l) = 0 \ , i \neq l$

- Jos lisäksi jäännös- eli virhetermejä ε_i koskeva *normaalisuusoletus* (iv) *pätee*, niin

(iv)´ $y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i \ , \sigma^2) \ , i = 1, 2, \ldots, n$

Mallin parametrit

- Tavallisen yhden selittäjän lineaarisen regressiomallin

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

parametreja ovat mallin **regressiokertoimet** β_0 ja β_1 sekä jäännös- eli virhetermien ε_i yhteinen *varianssi*

$$\mathrm{Var}(\varepsilon_i) = \sigma^2 \ , i = 1, 2, \ldots, n$$

jota kutsutaan **jäännösvarianssiksi**.

- Koska regressiokertoimet β_0 ja β_1 sekä jäännösvarianssi σ^2 ovat tavallisesti *tuntemattomia*, ne on *estimoitava* muuttujien *x* ja *y* havaituista arvoista x_i ja y_i , $i = 1, 2, \ldots, n$.

Mallin systemaattinen ja satunnainen osa 1/2

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin

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

jäännös- eli *virhetermejä* ε_i *koskeva standardioletus*

(i) $\mathrm{E}(\varepsilon_i) = 0 \ , i = 1, 2, \ldots, n$

pätee.

- Tällöin selitettävän muuttujan *y* havaitut arvot y_i voidaan *esittää* seuraavalla tavalla *kahden osatekijän summana*:

$$y_i = \mathrm{E}(y_i) + \varepsilon_i \ , \ i = 1, 2, \ldots , n$$

jossa

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \ , \ i = 1, 2, \ldots , n$$

Mallin systemaattinen ja satunnainen osa 2/2

- Odotusarvo

  $$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i \, , \; i = 1, 2, \ldots , n$$

  muodostaa tavallisen yhden selittäjän lineaarisen regressiomallin **systemaattisen osan** eli **rakenneosan**, joka *riippuu selittäjälle x annetuista arvoista*.

- Jäännös- eli virhetermi

  $$\varepsilon_i \, , \; i = 1, 2, \ldots , n$$

  muodostaa mallin **satunnaisen osan**, joka *ei riipu selittäjälle x annetuista arvoista*.

Regressiosuora

- Tavallisen yhden selittäjän lineaarisen regressiomallin

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

  systemaattinen osa $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_i$ *määrittelee suoran*

  $$y = \beta_0 + \beta_1 x$$

  avaruudessa $\mathbb{R}^2$.
- Suoraa kutsutaan **regressiosuoraksi** ja sen yhtälössä

  β_0 = regressiosuoran ja y-akselin **leikkauspiste**

  β_1 = regressiosuoran **kulmakerroin**
- Jäännös- eli virhetermien ε_i varianssi σ^2 kuvaa *havainto-pisteiden* (x_i , y_i) , $i = 1, 2, \ldots , n$ *vaihtelua regressiosuoran ympärillä*.

Regressiosuoran kulmakertoimen tulkinta

- Tavallisen yhden selittäjän lineaarisen regressiomallin systemaattisen osan määrittelemän regressiosuoran

  $$y = \beta_0 + \beta_1 x$$

  kulmakertoimella β_1 seuraava **tulkinta**:
- Oletetaan, että *selittäjän x arvo kasvaa yhdellä yksiköllä*:

  $$x \rightarrow x + 1$$

  Tällöin kerroin β_1 kertoo *paljonko selitettävän muuttujan y vastaava odotettavissa oleva arvo muuttuu*:

$$\begin{aligned} \mathrm{E}(y) = \beta_0 + \beta_1 x &\rightarrow \beta_0 + \beta_1(x + 1) \\ &= \beta_0 + \beta_1 x + \beta_1 \\ &= \mathrm{E}(y) + \beta_1 \end{aligned}$$

Yhden selittäjän lineaarinen regressiomalli

Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

>> Yhden selittäjän lineaarisen regressiomallin estimointi

Varianssianalyysihajotelma ja selitysaste

Estimointiongelma

- Tavallisen yhden selittäjän lineaarisen regressiomallin

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

regressiokertoimet β_0 ja β_1 ovat normaalisti *tuntemattomia*, joten *ne on estimoitava* muuttujien *x* ja *y* havaituista arvoista x_i ja y_i , i = 1, 2, … , n.

- Estimoinnissa regressiokertoimille β_0 ja β_1 pyritään löytämään sellaiset arvot, että niiden määräämä *regressio-suora selittäisi mahdollisimman hyvin selitettävän muuttujan y arvojen vaihtelun*.

- Regressiokertoimien β_0 ja β_1 estimointiin on tarjolla useita erilaisia menetelmiä, joista yleisin on *pienimmän neliösumman menetelmä*.

Pienimmän neliösumman menetelmä

- **Pienimmän neliösumman menetelmässä** mallin

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

regressiokertoimien β_0 ja β_1 estimaattorit määrätään *minimoimalla jäännös-* eli *virhetermien ε_i neliösumma*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

regressiokertoimien β_0 ja *β_1 suhteen.*

Otostunnusluvut

- Määritellään havaintojen x_i ja y_i , $i = 1, 2, \ldots , n$ *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

Regressiokertoimien PNS-estimaattorit

- Tavallisen yhden selittäjän lineaarisen regressiomallin

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

regressiokertoimien β_0 ja β_1 **pienimmän neliösumman (PNS-) estimaattorit** ovat

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x}$$

PNS-estimaattoreiden johto 1/4

- Tavallisen yhden selittäjän lineaarisen regressiomallin

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

  regressiokertoimet β_0 ja β_1 estimoidaan PNS-menetelmällä *minimoimalla jäännöstermien ε_i neliösumma*

  $$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

  kertoimien β_0 ja β_1 suhteen
- Tämä tapahtuu tavanomaiseen tapaan derivoimalla funktio $S(\beta_0 , \beta_1)$ kertoimien β_0 ja β_1 suhteen ja merkitsemällä derivaatat nolliksi.

PNS-estimaattoreiden johto 2/4

- Derivoidaan funktio

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

regressiokertoimien β_0 ja β_1 suhteen ja merkitään derivaatat nolliksi:

(1) $$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

(2) $$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0$$

- Regressiokertoimien β_0 ja β_1 *PNS-estimaattorit* saadaan *normaali-yhtälöiden* (1) ja (2) ratkaisuina.

PNS-estimaattoreiden johto 3/4

- Kirjoitetaan normaaliyhtälöt (1) ja (2) muotoihin

$$(1)' \quad \sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_i = 0$$

$$(2)' \quad \sum_{i=1}^{n} y_i x_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

- Ratkaistaan β_0 yhtälöstä (1)´:

$$(3) \quad \beta_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \beta_1 \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{y} - \beta_1 \bar{x}$$

ja sijoitetaan ratkaisu yhtälöön (2)´:

$$(4) \quad \sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

PNS-estimaattoreiden johto 4/4

- Parametrin β_1 PNS-estimaattoriksi saadaan yhtälöstä (4):

$$(5) \quad b_1 = \frac{\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_j^2 - n\bar{x}^2} = \frac{s_{xy}}{s_x^2} = r_{xy}\frac{s_y}{s_x}$$

- Sijoittamalla b_1 yhtälöön (3) saadaan parametrin β_0 PNS-estimaattoriksi

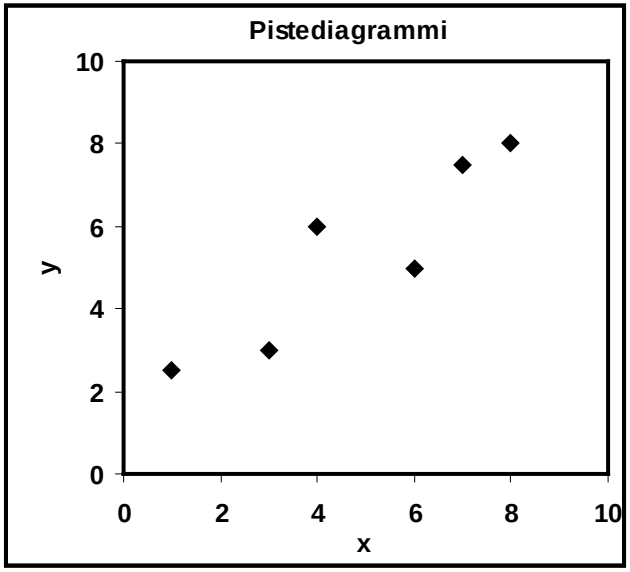
$$(6) \quad b_0 = \bar{y} - b_1\bar{x}$$

- Sivuutamme sen osoittamisen, että saatu ääriarvo on todellakin *minimi*.

Tunnuslukujen laskeminen: Havainnollistava esimerkki 1/3

- Taulukossa oikealla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$).
- Aineistoa kuvaava *pistediagrammi* on oikealla alhaalla.

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8



Tunnuslukujen laskeminen: Havainnollistava esimerkki 2/3

- Alla olevassa taulukossa on laskettu muuttujien *x* ja *y* havaittujen arvojen *summat*, *neliösummat* ja *tulosumma*.

i	*x*	*y*	x^2	y^2	*xy*
1	1	2.5	1	6.25	2.5
2	3	3	9	9	9
3	4	6	16	36	24
4	6	5	36	25	30
5	7	7.5	49	56.25	52.5
6	8	8	64	64	64
Summa	29	32	175	196.5	182

- Yhden selittäjän lineaarisen regressiomallin

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

regressiokertoimien β_0 ja β_1 *PNS-estimaatit* voidaan laskea näistä viidestä summasta; ks. seuraavaa kalvoa.

Tunnuslukujen laskeminen: Havainnollistava esimerkki 3/3

- *Regressiokertoimien* β_0 ja β_1 *PNS-estimaatit*:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{6} \times 29 = 4.833$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{6} \times 32 = 5.333$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{182 - \frac{1}{6} \times 29 \times 32}{175 - \frac{1}{6} \times 29^2} = 0.785$$

$$b_0 = \bar{y} - b_1 \bar{x} = 5.333 - 0.7847 \times 4.833 = 1.541$$

Estimoitu regressiosuora 1/3

- Tavallisen yhden selittäjän lineaarinen regressiomallin

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

regressiokertoimien β_0 ja β_1 PNS-estimaattorit b_0 ja b_1 *määrittelevät suoran* avaruudessa $\mathbb{R}^2$:

$$y = b_0 + b_1 x$$

jossa

b_0 = estimoidun regressiosuoran ja y-akselin **leikkauspiste**

b_1 = estimoidun regressiosuoran **kulmakerroin**

Estimoitu regressiosuora 2/3

- Sijoitetaan regressiokertoimien β_0 ja β_1 PNS-estimaattoreiden lausekkeet

$$b_0 = \bar{y} - b_1 \bar{x} \qquad\qquad b_1 = r_{xy} \frac{s_y}{s_x}$$

  estimoidun regressiosuoran lausekkeeseen.

- Tällöin estimoidun regressiosuoran yhtälö voidaan kirjoittaa seuraavaan muotoon:

$$y = \bar{y} + r_{xy} \frac{s_y}{s_x}(x - \bar{x})$$

- Yhtälöstä nähdään, että estimoitu regressiosuora kulkee havaintopisteiden (x_i , y_i) , $i = 1, 2, \ldots , n$ *painopisteen* $(\bar{x}, \bar{y})$ kautta.

Estimoitu regressiosuora 3/3

- Estimoidulla regressiosuoralla

$$y = \bar{y} + r_{xy} \frac{s_y}{s_x}(x - \bar{x})$$

on seuraavat ominaisuudet:

(i) Jos $r_{xy} > 0$, suora on *nouseva*.

(ii) Jos $r_{xy} < 0$, suora on *laskeva*.

(iii) Jos $r_{xy} = 0$, suora on *vaakasuorassa*.

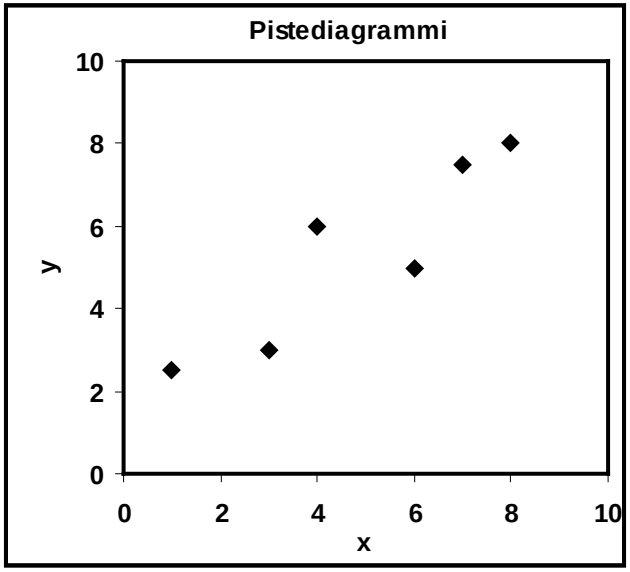
(iv) Suora *jyrkkenee* (*loivenee*), jos

- korrelaation *itseisarvo* $|r_{xy}|$ *kasvaa* (*pienenee*)
- keskihajonta s_y *kasvaa* (*pienenee*)
- keskihajonta s_x *pienenee* (*kasvaa*)

Estimoitu regressiosuora: Havainnollistava esimerkki 1/2

- Taulukossa oikealla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$).
- Aineistoa kuvaava *pistediagrammi* on oikealla alhaalla.

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8



Estimoitu regressiosuora: Havainnollistava esimerkki 2/2

- Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$i = 1, 2, \ldots, n$$

regressiokertoimien β_0 ja β_1 *PNS-estimaateiksi* saatiin edellä

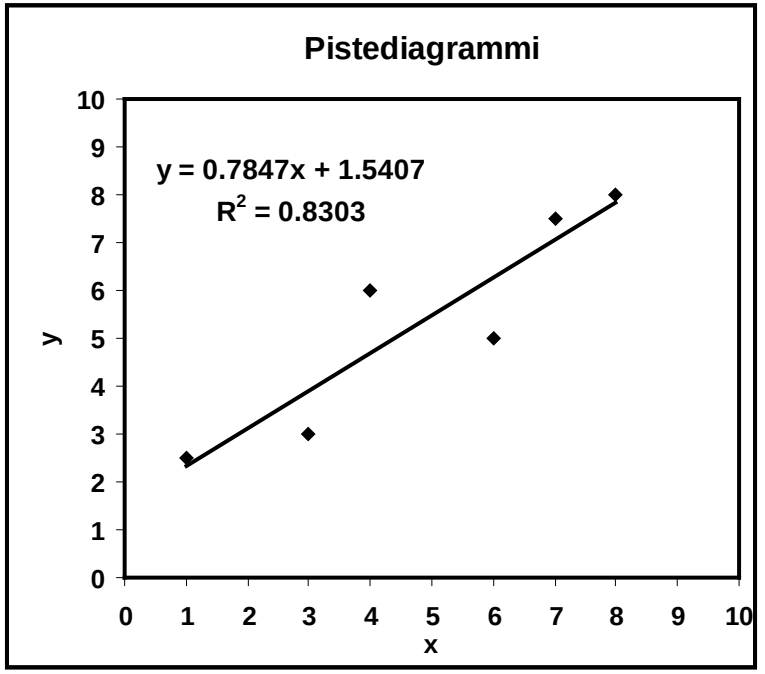
$$b_0 = 1.5407$$

$$b_1 = 0.7847$$

- *Estimoidun regressiosuoran* yhtälö on siten

$$y = 1.5407 + 0.7847x$$

ks. kuviota oikealla.



Mallia koskeva tilastollinen päättely

- Voisimme tutulla tavalla myös määrätä *tavallisen yhden selittäjän lineaarisen regressiomallin*

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

regressiokertoimien β_0 ja β_1 **pienimmän neliösumman (PNS-) estimaattoreitten** b_0 ja b_1 **otosjakaumat** ja **regressiokertoimien luottamusvälit** ja tarkastella **testejä regressiokertoimille.**

Sovitteet ja residuaalit

- Olkoot b_0 ja b_1 yhden selittäjän lineaarisen regressiomallin

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

  regressiokertoimien β_0 ja β_1 PNS-estimaattorit.
- Määritellään estimoidun mallin **sovitteet** kaavalla

$$\hat{y}_i = b_0 + b_1 x_i \ , i = 1, 2, \ldots, n$$

- Määritellään estimoidun mallin **residuaalit** kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \ , i = 1, 2, \ldots, n$$

- Huomaa, että

$$y_i = \hat{y}_i + e_i \ , i = 1, 2, \ldots, n$$

Sovitteet ja residuaalit: Tulkinnat 1/2

- *Sovite*

  $$\hat{y}_i = b_0 + b_1 x_i \ , i = 1, 2, \ldots, n$$

  on estimoidun regressiosuoran antama arvo selitettävälle muuttujalle y havaintopisteessä x_i .

- *Residuaali*

  $$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \ , i = 1, 2, \ldots, n$$

  on selitettävän muuttujan y havaitun arvon y_i ja sovitteen $\hat{y}_i$ erotus.

Sovitteet ja residuaalit: Tulkinnat 2/2

- Estimoitu regressiomalli selittää selitettävän muuttujan *y* havaittujen arvojen vaihtelun *sitä paremmin mitä lähempänä estimoidun mallin sovitteet* $\hat{y}_i$ *ovat selitettävän muuttujan y havaittuja arvoja* y_i .

- Yhtäpitävästi edellisen kanssa:

  Estimoitu regressiomalli selittää selitettävän muuttujan *y* havaittujen arvojen y_i vaihtelun *sitä paremmin mitä pienempiä ovat estimoidun mallin residuaalit* e_i .

Sovitteet ja residuaalit: Havainnollistus

- Kuvio oikealla havainnollistaa sovitteiden ja residuaalien *geometrista tulkintaa*.
- *Malli*:

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

- *PNS-suora*:

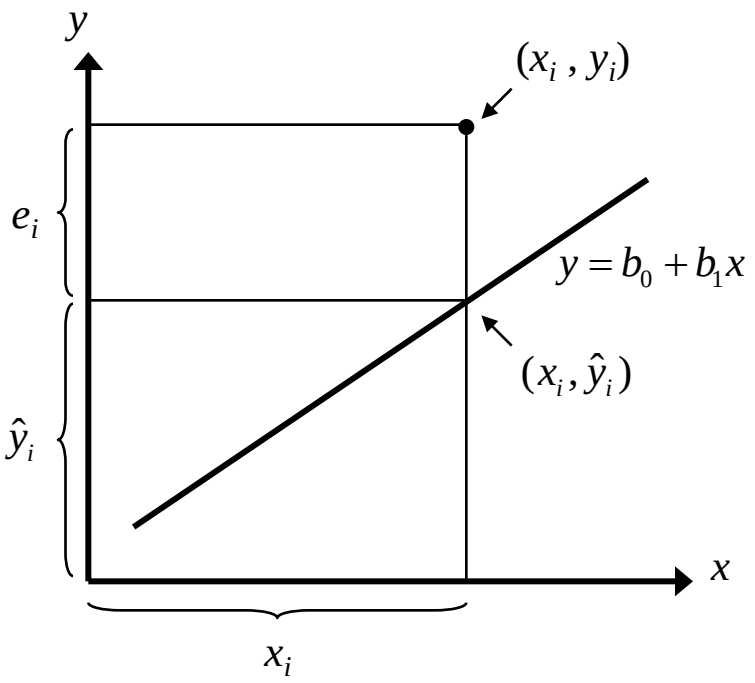
$$y = b_0 + b_1 x$$

- *Sovite*:

$$\hat{y}_i = b_0 + b_1 x_i \ , i = 1, 2, \ldots, n$$

- *Residuaali*:

$$e_i = y_i - \hat{y}_i \ , i = 1, 2, \ldots, n$$



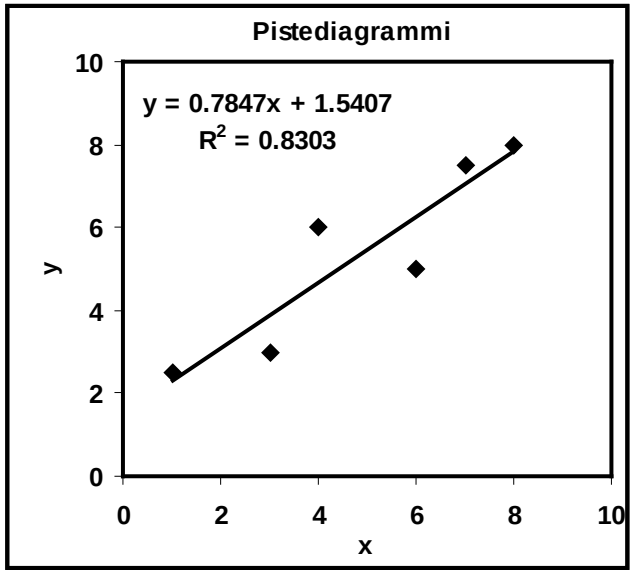
Sovitteet ja residuaalit: Havainnollistava esimerkki 1/3

- Taulukossa oikealla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$).
- *Estimoidun regressiosuoran* yhtälöksi saatiin edellä

$$y = 1.5407 + 0.7847x$$

ks. kuviota oikealla.

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8



Sovitteet ja residuaalit: Havainnollistava esimerkki 2/3

- Alla olevassa taulukossa on laskettu estimoidun mallin

$$y = 1.5407 + 0.7847x$$

sovitteet $\hat{y}$ ja *residuaalit* e:

i	*x*	*y*	*Sovite*	*Residuaali*
1	1	2.5	2.325	0.175
2	3	3	3.895	-0.895
3	4	6	4.679	1.321
4	6	5	6.249	-1.249
5	7	7.5	7.033	0.467
6	8	8	7.818	0.182
Summa	29	32	32.000	0.000

- Esimerkiksi, kun $i = 3$, niin

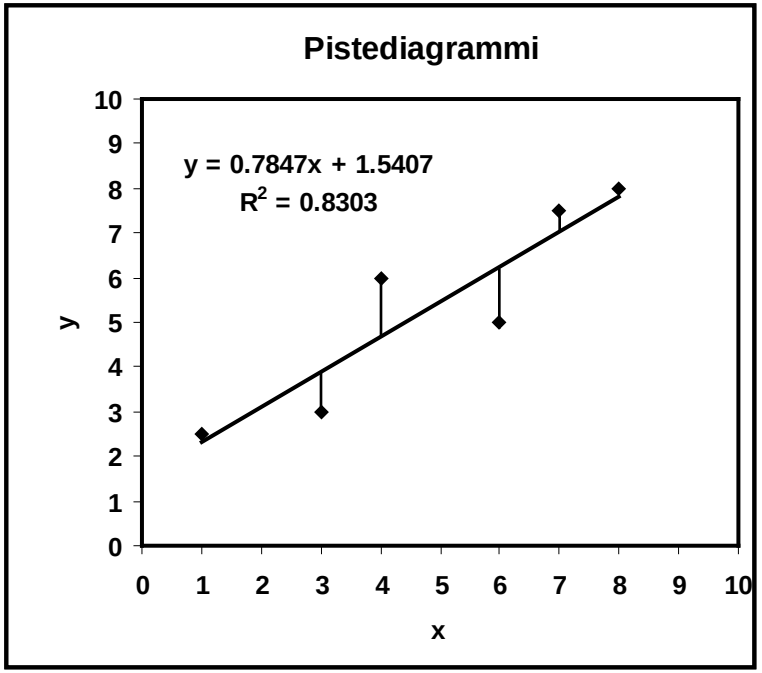
$$\hat{y}_3 = 1.5407 + 0.7847x_3 = 1.5407 + 0.7847 \times 4 = 4.679$$

$$e_3 = y_3 - \hat{y}_3 = 6 - 4.679 = 1.321$$

Sovitteet ja residuaalit: Havainnollistava esimerkki 3/3

- Kuvioon oikealla on lisätty estimoidun regressiomallin *residuaaleja* vastaavat janat.
- Huomautus:

  Pienimmän neliösumman menetelmässä regressiosuoran kertoimet tulevat valituiksi siten, että estimoidun mallin *residuaaleja vastaavien janojen pituuksien neliöiden summa on pienin mahdollinen*.



Jäännösvarianssin estimointi 1/2

- Jos tavallisen yhden selittäjän lineaarisen regressiomallin *jäännös-* eli *virhetermejä* ε_i *koskevat standardioletukset* (i)-(iii) *pätevät*, *jäännösvarianssin* $\mathrm{Var}(\varepsilon_i) = \sigma^2$ **harhaton estimaattori** on

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

jossa

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_i \ , i = 1, 2, \ldots, n$$

$$= \text{estimoidun mallin } \textit{residuaali}$$

$$n = \text{havaintojen lukumäärä}$$

Jäännösvarianssin estimointi 2/2

- Jäännösvarianssin σ^2 estimaattori

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

kuvaa *havaintopisteiden* (x_i , y_i) , $i = 1, 2, \ldots , n$ *vaihtelua estimoidun regressiosuoran ympärillä*.

Jäännösvarianssin estimointi: Kommentti

- Estimaattori s^2 on todellakin *residuaalien e_i varianssi*.
- Tämä seuraa siitä, että mallissa on *vakioselittäjä*, jolloin

$$\sum_{i=1}^{n} e_i = 0$$

ja siten myös

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$$

jolloin

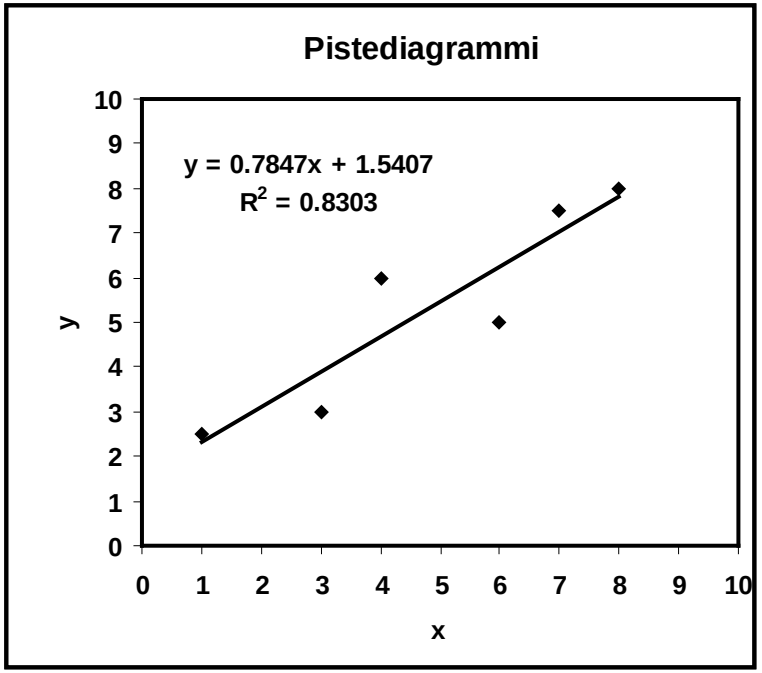
$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(e_i - \bar{e}\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

Jäännösvarianssin estimointi: Havainnollistava esimerkki 1/2

- Taulukossa alla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$):

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8

- Aineistoa kuvaava *pistediagrammi* on oikealla.
- Kuvioon on merkitty myös aineistosta *estimoidun regressio-suoran yhtälö*.



Jäännösvarianssin estimointi: Havainnollistava esimerkki 2/2

- Alla olevassa taulukossa on laskettu estimoidun mallin *sovitteet* $\hat{y}$, *residuaalit* e (sovitteiden ja residuaalien laskemista on käsitelty edellä) ja *residuaalien neliöt* e^2.

i	*x*	*y*	*Sovite*	*Residuaali*	*Res* ^{2}
1	1	2.5	2.325	0.175	0.030
2	3	3	3.895	-0.895	0.801
3	4	6	4.679	1.321	1.744
4	6	5	6.249	-1.249	1.560
5	7	7.5	7.033	0.467	0.218
6	8	8	7.818	0.182	0.033
Summa	29	32	32.000	0.000	4.385

- *Jäännösvarianssin* σ^2 *harhaton estimaattori* on

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 = \frac{1}{6-2} \times 4.385 = 1.096$$

Yhden selittäjän lineaarinen regressiomalli

- Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset
- Yhden selittäjän lineaarisen regressiomallin estimointi
- >> Varianssianalyysihajotelma ja selitysaste

Varianssianalyysihajotelman idea

- Yhden selittäjän regressiomallin tehtävänä on selittää *selitettävän muuttujan y havaittujen arvojen vaihtelu selittävän muuttujan x havaittujen arvojen vaihtelulla*.
- Onnistumista tässä tehtävässä voidaan kuvata ns. **varianssianalyysihajotelman** avulla.
- Hajotelmassa *selitettävän muuttujan y havaittujen arvojen kokonaisvaihtelua kuvaava* ns. *kokonaisneliösumma jaetaan kahden osatekijän summaksi*:
  (i) Toinen osatekijä kuvaa *estimoidun mallin selittämää osaa kokonaisvaihtelusta*.
  (ii) Toinen osatekijä kuvaa *mallilla selittämättä jäänyttä osaa kokonaisvaihtelusta*.

Kokonaisneliösumma

- Neliösumma

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

kuvaa *selitettävän muuttujan y havaittujen arvojen y_j vaihtelua* ja sitä kutsutaan **kokonaisneliösummaksi**.

- *Selitettävän muuttujan y havaittujen arvojen y_i varianssi* voidaan määritellä kaavalla

$$s_y^2 = \frac{1}{n-1} SST$$

Jäännösneliösumma

- Neliösumma

$$SSE = \sum_{i=1}^{n} e_i^2$$

kuvaa *residuaalien* e_i *vaihtelua* ja sitä kutsutaan **jäännös-neliösummaksi**.

- Koska mallissa on vakioselittäjä, jolloin $\sum e_i = 0$, *residuaalien* e_i *varianssi* voidaan määritellä kaavalla

$$s^2 = \frac{1}{n-2} SSE$$

- s^2 on jäännösvarianssin σ^2 *harhaton estimaattori*.

Kokonais- ja jäännösneliösumman yhteys 1/4

- Voidaan osoittaa, että yhden selittäjän lineaarisessa regressiomallissa jäännösneliösumma *SSE* ja kokonais-neliösumma *SST* toteuttavat yhtälöt

$$SSE = \sum_{i=1}^{n} e_i^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2) SST$$

jossa

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

= selitettävän muuttujan *y* ja selittäjän *x* havaittujen arvojen otoskorrelaatiokerroin

Kokonais- ja jäännösneliösumman yhteys 2/4

- Koska otoskorrelaatiokerroin r_{xy} toteuttaa epäyhtälöt

  $$-1 \leq r_{xy} \leq +1$$

  yhtälöistä

  $$SSE = \sum_{i=1}^{n} e_i^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2) SST$$

  nähdään välittömästi, että

  $$SSE \leq SST$$

Kokonais- ja jäännösneliösumman yhteys 3/4

- Yhtälöistä

$$SSE = \sum_{i=1}^{n} e_i^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2) SST$$

nähdään, että seuraavat ehdot ovat yhtäpitäviä:

(i) $SSE = 0$

(ii) $e_i = 0$ kaikille $i = 1, 2, \ldots, n$

(iii) $r_{xy} = \pm 1$

- Jos ehdot (i)-(iii) pätevät, niin kaikki havaintopisteet (x_i , y_i) , $i = 1, 2, \ldots, n$ ovat samalla suoralla ja tätä suoraa vastaava *lineaarinen regressiomalli selittää täydellisesti selitettävän muuttujan y havaittujen arvojen vaihtelun*.

Kokonais- ja jäännösneliösumman yhteys 4/4

- Yhtälöistä

$$SSE = \sum_{i=1}^{n} e_i^2 = (1 - r_{xy}^2) \sum_{i=1}^{n} (y_i - \bar{y})^2 = (1 - r_{xy}^2) SST$$

nähdään, että seuraavat ehdot ovat yhtäpitäviä:

(i)´ $SSE = SST$

(ii)´ $e_i = y_i - \bar{y}$ kaikille $i = 1, 2, \ldots, n$

(iii)´ $r_{xy} = 0$

- Jos ehdot (i)´-(iii)´ pätevät, niin *selitettävän muuttujan y havaittujen arvojen vaihtelua ei voida selittää mallina käytetyn lineaarisen regressiomallin avulla*.

Mallineliösumma

- Määritellään suure *SSM* yhtälöllä

$$SSM = SST - SSE$$

- Koska

$$0 \leq SSE \leq SST$$

niin

$$SSM \geq 0$$

- Koska voidaan osoittaa, että

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

suuretta *SSM* kutsutaan **mallineliösummaksi**.

Varianssianalyysihajotelma 1/2

- Edellä esitetyn mukaan kokonaisneliösumma

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

voidaan esittää kahden osatekijän *SSM* ja *SSE* summana:

$$SST = SSM + SSE$$

jossa

$$SSM = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

ja

$$SSE = \sum_{i=1}^{n} e_i^2$$

Varianssianalyysihajotelma 2/2

- **Varianssianalyysihajotelmassa**

  $$SST = SSM + SSE$$

  selitettävän muuttujan *y* havaittujen arvojen vaihtelua kuvaava **kokonaisneliösumma** *SST* on esitetty kahden osatekijän *SSM* ja *SSE* summana:

  (i) **Mallineliösumma** *SSM* kuvaa sitä osaa selitettävän muuttujan *y* havaittujen arvojen vaihtelusta, jonka *estimoitu malli on selittänyt*.

  (ii) **Jäännösneliösumma** *SSE* kuvaa sitä osaa selitettävän muuttujan *y* havaittujen arvojen vaihtelusta, jota *estimoitu malli ei ole selittänyt*.

Varianssianalyysihajotelman tulkinta

- Varianssianalyysihajotelma

  $$SST = SSM + SSE$$

  kuvaa estimoidun regressiomallin *hyvyyttä*:

  (i) Mitä *suurempi* on *mallineliösumman SSM osuus* kokonaisneliösummasta *SST*, sitä paremmin estimoitu malli selittää selitettävän muuttujan havaittujen arvojen vaihtelun.

  (ii) Mitä *pienempi* on *jäännösneliösumman SSE osuus* kokonaisneliösummasta *SST*, sitä paremmin estimoitu malli selittää selitettävän muuttujan havaittujen arvojen vaihtelun.

Selitysaste

- Varianssianalyysihajotelma

  $$SST = SSM + SSE$$

  motivoi tunnusluvun

  $$R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$$

  käytön *regressiomallin hyvyyden mittarina*.
- Tunnuslukua R^2 kutsutaan **selitysasteeksi** ja se *mittaa regressiomallin selittämää osuutta* selitettävän muuttujan *y* havaittujen arvojen kokonaisvaihtelusta.
- Selitysaste R^2 ilmaistaan tavallisesti prosentteina:

  $$100 \times R^2 \ \%$$

Selitysaste ja korrelaatio

- Voidaan osoittaa, että

$$R^2 = \left[\mathrm{Cor}(y, \hat{y})\right]^2$$

jossa

$$\mathrm{Cor}(y, \hat{y})$$

on selitettävän muuttujan y havaittujen arvojen y_i ja sovitteiden $\hat{y}_i$ *otoskorrelaatiokerroin*.

- *Yhden selittäjän lineaarisen regressiomallin* tapauksessa pätee lisäksi se, että selitysaste R^2 on selitettävän ja selittävän muuttujan havaittujen arvojen *otos-korrelaatiokertoimen* r_{xy} *neliö*:

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

Selitysasteen ominaisuudet 1/2

- *Selitysasteella* R^2 on seuraavat ominaisuudet:

  (i) $0 \leq R^2 \leq 1$

  (ii) Seuraavat ehdot ovat *yhtäpitäviä*:

  (1) $R^2 = 1$

  (2) Kaikki residuaalit häviävät:

  $e_i = 0$ kaikille $i = 1, 2, \ldots, n$

  (3) Kaikki havaintopisteet (x_i , y_i) , $i = 1, 2, \ldots, n$ asettuvat *samalle suoralle*.

  (4) $r_{xy} = \pm 1$

  (5) Määritelty malli *selittää täydellisesti* selitettävän muuttujan y havaittujen arvojen vaihtelun.

Selitysasteen ominaisuudet 2/2

(iii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 0$

(2) $b_1 = 0$

(3) $r_{xy} = 0$

(4) Määritelty malli *ei ollenkaan selitä* selitettävän muuttujan *y* havaittujen arvojen vaihtelua.

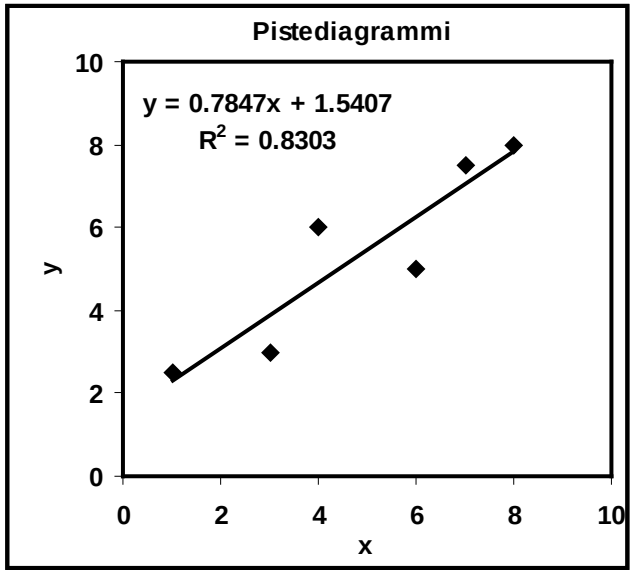
Selitysasteen laskeminen: Havainnollistava esimerkki 1/3

- Taulukossa oikealla on keinotekoisen kahden muuttujan aineiston havaintoarvot ($n = 6$).
- Aineistosta *estimoidun regressiosuoran* yhtälöksi saatiin kappaleessa **Yhden selittäjän lineaarisen regressiomallin estimointi**

$$y = 1.5407 + 0.7847x$$

ks. kuviota oikealla.

i	*x*	*y*
1	1	2.5
2	3	3
3	4	6
4	6	5
5	7	7.5
6	8	8



Selitysasteen laskeminen: Havainnollistava esimerkki 2/3

- Alla olevassa taulukossa on laskettu havaintoarvojen summat ja neliö-summat sekä estimoidun mallin *sovitteet* $\hat{y}$, *residuaalit* e (sovitteiden ja residuaalien laskemista on käsitelty em. kappaleessa) ja *residuaalien neliöt* e^2.

i	x	y	x^2	y^2	Sovite	Residuaali	Res^2
1	1	2.5	1	6.25	2.325	0.175	0.030
2	3	3	9	9	3.895	-0.895	0.801
3	4	6	16	36	4.679	1.321	1.744
4	6	5	36	25	6.249	-1.249	1.560
5	7	7.5	49	56.25	7.033	0.467	0.218
6	8	8	64	64	7.818	0.182	0.033
Summa	29	32	175	196.5	32	0.000	4.385

- Estimoidun mallin *selitysaste* saadaan taulukon sarakesummista seuraavalla kalvolla esitettävällä tavalla.

Selitysasteen laskeminen: Havainnollistava esimerkki 3/3

- *Kokonaisneliösumma*:

$$SST = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 196.5 - \frac{1}{6} \times 32^2 = 25.833$$

- *Jäännösneliösumma*:

$$SSE = \sum_{i=1}^{n} e_i^2 = 4.385$$

- *Selitysaste*:

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{4.385}{25.833} = 0.830$$

- Siten estimoitu malli on selittänyt

  83.0 %

  selitettävän muuttujan arvojen vaihtelusta.

Osa 4: Lineaarinen regressioanalyysi

- **Yhden selittäjän lineaarinen regressiomalli (Osa 2)**

Yhden selittäjän lineaarinen regressiomalli

>> Ennustaminen yhden selittäjän lineaarisella regressiomallilla
Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä
2-ulotteisen normaalijakauman regressiofunktioiden estimointi

Ennustaminen

- Oletetaan, että muuttujien x ja y havaittujen arvojen x_i ja y_i välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista muodossa

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

- Haluamme *ennustaa selitettävää muuttujaa y*, kun selittävä muuttuja x saa arvon $\tilde{x}$.
- Jaetaan tarkastelu kahteen osaan:

(i) Tavoitteena on ennustaa selitettävän muuttujan y **odotettavissa oleva** eli *keskimääräinen* **arvo**.

(ii) Tavoitteena on ennustaa selitettävän muuttujan y **arvo**.

Malli ja sen osat

- Tarkastellaan **tavallista yhden selittäjän lineaarista regressiomallia**

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

jonka *jäännöstermit* ε_i toteuttavat ns. *tavanomaiset* eli *standardioletukset*:

(i) $\mathrm{E}(\varepsilon_i) = 0 \ , i = 1, 2, \ldots, n$

(ii) $\mathrm{Var}(\varepsilon_i) = \sigma^2 \ , i = 1, 2, \ldots, n$

(iii) $\mathrm{Cor}(\varepsilon_i, \varepsilon_l) = 0 \ , i \neq l$

y:n odotusarvon ennustaminen

- Oletetaan, että selitettävä muuttuja *y* saa arvon

  $$\tilde{y} = \beta_0 + \beta_1 \tilde{x} + \tilde{\varepsilon}$$

  kun selittäjä *x* saa arvon $\tilde{x}$.
- Mikä on *paras ennuste selitettävän muuttujan y* **odotettavissa olevalle arvolle**

  $$\mathrm{E}(\tilde{y} \mid \tilde{x}) = \beta_0 + \beta_1 \tilde{x}$$

  kun selittäjä *x* saa arvon $\tilde{x}$?
- Selitettävän muuttujan *y* ehdollinen odotusarvo $\mathrm{E}(\tilde{y} \mid \tilde{x})$ kuvaa *selitettävän muuttujan y keskimäärin saamia arvoja selittäjän x saamien arvojen funktiona*.

y:n odotusarvon ennustaminen: Ennuste

- Valitaan *selitettävän muuttujan odotusarvon* $\mathrm{E}(\tilde{y}|\tilde{x})$ **ennusteeksi** (*estimaattoriksi*) lauseke

  $$\tilde{y}|\tilde{x} = b_0 + b_1\tilde{x}$$

  jossa b_0 ja b_1 ovat regressiokertoimien β_0 ja β_1 PNS-estimaattorit.
- Voidaan osoittaa, että $\tilde{y}|\tilde{x}$ on (ennustevirheen keskineliövirheen mielessä) *paras lineaarinen* ja *harhaton ennuste* ehdolliselle odotusarvolle $\mathrm{E}(\tilde{y}|\tilde{x})$.
- **Huomautus:**

  Ehdollinen odotusarvo $\mathrm{E}(\tilde{y}|\tilde{x})$ on kiinteälle $\tilde{x}$ *vakio*, kun taas ennuste $\tilde{y}|\tilde{x}$ on *satunnaismuuttuja*.

y:n odotusarvon ennustaminen: Otosjakauma

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin *jäännös*- eli *virhetermiä* ε_i koskevat *standardioletuksien* (i)-(iii) *lisäksi normaalisuusoletus* (iv) *pätee*.
- Tällöin ennusteen

$$\tilde{y}\,|\,\tilde{x} = b_0 + b_1\tilde{x}$$

otosjakauma on normaalijakauma:

$$\tilde{y}\,|\,\tilde{x} \sim \mathrm{N}\left(\beta_0 + \beta_1\tilde{x}\,,\,\sigma^2\left[\frac{1}{n} + \frac{(\tilde{x}-\overline{x})^2}{(n-1)\,s_x^2}\right]\right)$$

y:n odotusarvon ennustaminen: Luottamusväli

- Odotusarvon

$$\mathrm{E}(\tilde{y} \mid \tilde{x}) = \beta_0 + \beta_1 \tilde{x}$$

luottamusväli luottamustasolla $(1 - \alpha)$ on

$$b_0 + b_1 \tilde{x} \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{(n-1)\, s_x^2}}$$

jossa $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori.

- Väli muodostaa selittäjän x arvojen $\tilde{x}$ funktiona *luottamusvyön* estimoidun regressiosuoran $y = b_0 + b_1 x$ ympärille.

y:n odotusarvon ennustaminen: Luottamusvälin ominaisuuksia

- Odotusarvon

$$\mathrm{E}(\tilde{y} \mid \tilde{x}) = \beta_0 + \beta_1 \tilde{x}$$

luottamusväli

$$b_0 + b_1 \tilde{x} \pm t_{\alpha/2}\, s \sqrt{\frac{1}{n} + \frac{(\tilde{x} - \overline{x})^2}{(n-1)\, s_x^2}}$$

kaventuu, jos havaintojen lukumäärä n tai selittäjän otos-varianssi s_x^2 kasvaa.

- Toisaalta luottamusväli on sitä *leveämpi*, mitä kauempana piste $\tilde{x}$ on selittäjän x havaittujen arvojen aritmeettisesta keskiarvosta $\overline{x}$.

y:n arvon ennustaminen

- Oletetaan, että selitettävä muuttuja *y* saa arvon

  $$\tilde{y} = \beta_0 + \beta_1 \tilde{x} + \tilde{\varepsilon}$$

  kun selittäjä *x* saa arvon $\tilde{x}$.
- Mikä on *paras ennuste selitettävän muuttujan y* **arvolle** $\tilde{y}$, kun selittäjä *x* saa arvon $\tilde{x}$?

y:n arvon ennustaminen: Ennuste

- Valitaan *selitettävän muuttujan arvon* $\tilde{y}$ **ennusteeksi** (*estimaattoriksi*) lauseke

  $$\tilde{y}\,|\,\tilde{x} = b_0 + b_1\tilde{x}$$

  jossa b_0 ja b_1 ovat regressiokertoimien β_0 ja β_1 PNS-estimaattorit.
- $\tilde{y}\,|\,\tilde{x}$ on (ennustevirheen keskineliö-virheen mielessä) *paras lineaarinen* ja *harhaton ennuste* ehdolliselle odotusarvolle $E(\tilde{y}\,|\,\tilde{x})$
- **Huomautus:**

  Sekä selitettävän muuttujan *y* arvo $\tilde{y}$ että ennuste $\tilde{y}\,|\,\tilde{x}$ ovat *satunnaismuuttujia*.

y:n arvon ennustaminen: Otosjakauma

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin *jäännös*- eli *virhetermiä* ε_i koskevat *standardioletuksien* (i)-(iii) *lisäksi normaalisuusoletus* (iv) *pätee*.
- Tällöin *ennustevirheen*

  $$\tilde{y} - \tilde{y} \,|\, \tilde{x}$$

  otosjakauma on normaalijakauma:

$$\tilde{y} - \tilde{y} \,|\, \tilde{x} \sim \mathrm{N}\left(0 \,, \sigma^2 \left[1 + \frac{1}{n} + \frac{(\tilde{x} - \overline{x})^2}{(n-1)\, s_x^2} \right] \right)$$

y:n arvon ennustaminen: Luottamusväli

- Selitettävän muuttujan y arvon $\tilde{y}$ **luottamusväli** luottamustasolla $(1 - \alpha)$ on

$$b_0 + b_1\tilde{x} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \bar{x})^2}{(n-1)\, s_x^2}}$$

  jossa $-t_{\alpha/2}$ ja $+t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet Studentin *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori.

- Väli muodostaa selittäjän x arvojen $\tilde{x}$ funktiona *luottamusvyön* estimoidun regressiosuoran $y = b_0 + b_1x$ ympärille.

y:n arvon ennustaminen: Luottamusvälin ominaisuuksia

- Selitettävän muuttujan y arvon $\tilde{y}$ luottamusväli

$$b_0 + b_1\tilde{x} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(\tilde{x} - \overline{x})^2}{(n-1)\, s_x^2}}$$

  kaventuu, jos havaintojen lukumäärä n tai selittäjän otos-varianssi s_x^2 kasvaa.

- Toisaalta luottamusväli on sitä *leveämpi*, mitä kauempana piste $\tilde{x}$ on selittäjän x havaittujen arvojen aritmeettisesta keskiarvosta $\overline{x}$.

y:n arvon luottamusväli *vs* *y*:n odotusarvon luottamusväli

- Selitettävän muuttujan y arvon $\tilde{y}$ luottamusvyö *on leveämpi* kuin selitettävän muuttujan y arvon $\tilde{y}$ odotusarvon $\mathrm{E}(\tilde{y}|\tilde{x})$ luottamusvyö.

- Tämä seuraa olennaisesti siitä, että selitettävän muuttujan *y keskimääräisen arvon ennustaminen on helpompaa kuin sen yksittäisen arvon ennustaminen*.

Yhden selittäjän lineaarinen regressiomalli

Ennustaminen yhden selittäjän lineaarisella regressiomallilla

>> Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä
2-ulotteisen normaalijakauman regressiofunktioiden estimointi

Selitettävä muuttuja ja selittävä muuttuja

- Oletetaan, että **selitettävän muuttujan** *y havaittujen arvojen vaihtelu* halutaan **selittää selittävän muuttujan** eli **selittäjän** *x havaittujen arvojen vaihtelun avulla*.

- Tehdään seuraavat oletukset:

  (i) Sekä selitettävä muuttuja *y* että selittäjä *x* ovat *satunnaismuuttujia*.

  (ii) Selitettävä muuttuja *y* on *suhdeasteikollinen muuttuja*.

Havainnot

- Olkoot

  $$y_1, y_2, \ldots, y_n$$

  selitettävän muuttujan y ja

  $$x_1, x_2, \ldots, x_n$$

  selittävän muuttujan x **havaittuja arvoja**.
- Oletetaan lisäksi, että havaintoarvot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$.
- Tällöin havaintoarvot x_i ja y_i muodostavat pisteitä 2-ulotteisessa avaruudessa:

  $$(x_i, y_i) \in \mathbb{R}^2 \, , i = 1, 2, \ldots, n$$

Malli ja sen osat 1/2

- Oletetaan, että havaintojen y_i ja x_i välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

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

- Yhtälö määrittelee **yhden selittäjän lineaarisen regressiomallin**, jossa

y_i = **selitettävän muuttujan** *y satunnainen* ja *havaittu* arvo havaintoyksikössä *i*

x_i = **selittävän muuttujan** *x satunnainen* ja *havaittu* arvo havaintoyksikössä *i*

ε_i = **jäännös-** eli **virhetermin** ε *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä *i*

Malli ja sen osat 2/2

- Yhden selittäjän lineaarisessa regressiomallissa

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

on seuraavat kertoimet:

β_0 = **vakioselittäjän regressiokerroin**;
β_0 on *ei-satunnainen* ja *tuntematon vakio*

β_1 = **selittäjän** *x* **regressiokerroin**;
β_1 on *ei-satunnainen* ja *tuntematon vakio*

- **Huomautus:**

Regressiokertoimet β_0 ja β_1 oletetaan *samoiksi* kaikille havaintoyksiköille *i*.

Selittäjän satunnaisuuden seuraukset 1/3

- Yhden selittäjän lineaarisen mallin

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

selittäjän *x* satunnaisuus *saattaa aiheuttaa vakavia ongelmia* mallin estimoinnille ja mallia koskevalle tilastolliselle päättelylle.

Selittäjän satunnaisuuden seuraukset 2/3

- **Jos selittäjä *x* on satunnainen, PNS-menetelmä *ei välttämättä tuota harhattomia* tai *edes tarkentuvia estimaattoreita regressiokertoimille*.**

  Näin käy esimerkiksi sellaisissa tapauksissa, joissa virhetermi ja selittäjä *korreloivat*.

- **Jos regressiokertoimien PNS-estimaattorit *eivät ole harhattomia* tai *tarkentuvia*, mallia koskevaa *tavanomaista tilastollista päättelyä ei saa soveltaa*.**

Selittäjän satunnaisuuden seuraukset 3/3

- Kysymys:

  Milloin kiinteälle, ei-satunnaiselle selittäjälle esitettyä teoriaa saa soveltaa myös satunnaiselle selittäjälle?

- Vastaus:

  Kiinteälle, ei-satunnaiselle selittäjälle esitettyä teoriaa saadaan soveltaa ainakin silloin, kun *jäännös-* eli *virhetermit* ε_j toteuttavat kiinteälle selittäjälle esitetyt standardioletukset *ehdollisesti selittäjän x havaittujen arvojen suhteen*.

Modifioidut oletukset jäännöstermeistä

- Oletetaan, että mallin

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

jäännös- eli *virhetermit* ε_i toteuttavat seuraavat oletukset:

(i) $\mathrm{E}(\varepsilon_i \mid x_i) = 0 \ , i = 1, 2, \ldots, n$

(ii) Jäännöstermit ovat (ehdollisesti) *homoskedastisia*.

$\mathrm{Var}(\varepsilon_i \mid x_i) = \sigma^2 \ , i = 1, 2, \ldots, n$

(iii) Jäännöstermit ovat (ehdollisesti) *korreloimattomia*.

$\mathrm{Cor}(\varepsilon_i, \varepsilon_l \mid x_i, x_l) = 0 \ , i \neq l$

- Lisäksi jäännöstermeistä ε_i tehdään tavallisesti *normaalisuusoletus*:

(iv) $\varepsilon_i \mid x_i \sim \mathrm{N}(0, \sigma^2) \ , i = 1, 2, \ldots, n$

Mallin selitettävän muuttujan ominaisuudet

- Jos yhden selittäjän lineaarisen regressiomallin

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

jäännös- eli *virhetermejä* ε_i *koskevat modifioidut oletukset* (i)-(iii) *pätevät*, mallin selitettävän muuttujan *y* havaituilla arvoilla y_i on seuraavat *stokastiset ominaisuudet*:

(i)´ $\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \ , i = 1, 2, \ldots, n$

(ii)´ $\mathrm{Var}(y_i \mid x_i) = \sigma^2 \ , i = 1, 2, \ldots, n$

(iii)´ $\mathrm{Cor}(y_i, y_l \mid x_i, x_l) = 0 \ , i \neq l$

- Jos jäännöstermejä ε_i koskeva *normaalisuusoletus* (iv) *pätee*, niin

(iv)´ $y_i \mid x_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \ , i = 1, 2, \ldots, n$

Mallin selitettävän muuttujan ominaisuudet: Kommentti

- Jos muuttujan y arvojen y_i stokastiset ominaisuudet (i)´-(iv)´ *otetaan oletuksiksi*, ne määrittelevät täsmälleen *saman tilastollisen mallin* kuin mallin

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

jäännös- eli virhetermeistä ε_i edellä tehdyt oletukset (i)-(iv).

- Oletukset (i)-(iv) ja (i)´-(iv)´ ovat tässä mielessä *ekvivalentteja*.
- Siten myös ominaisuudet (i)´-(iv)´ voidaan ottaa yhden selittäjän lineaarisen regressiomallin määritteleviksi standardioletuksiksi.

Selitettävän muuttujan ehdollisen odotusarvon tulkinta regressiofunktiona

- Oletuksen

  (i)´ $\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \ , i = 1, 2, \ldots, n$

  mukaan selitettävän muuttujan *y ehdollinen odotusarvo* eli **regressiofunktio on selittävän muuttujan *x* havaittujen arvojen suhteen lineaarinen funktio**.
- Koska *regressiofunktiot ovat yleisessä tapauksessa epälineaarisia*, (i)´ on *hyvin voimakas oletus*.
- **Huomautus:**

  Jos havainnot x_i ja y_i , i = 1, 2, … , n noudattavat 2-*ulotteista normaalijakaumaa*, oletus (i)´ pätee.

Selittäjän satunnaisuuden seuraukset: Kommentteja

- **Myös tässä kappaleessa esitetyt modifioidut ehdot jäännös- eli virhetermeille ovat melko rajoittavia ja *etenkin aikasarjojen regressiomalleissa kohdataan sellaisia tilanteita*, *joissa eivät edes nämä modifioidut ehdot päde*.**
- **Tällaisissa tilanteissa *PNS-menetelmää ei pidä käyttää mallin parametrien estimointiin*.**
- Tilastotiede tuntee kuitenkin menetelmiä, joilla regressiomallin parametrit *voidaan estimoida* (*ainakin*) *tarkentuvasti* myös monissa sellaisissa tilanteissa, joissa tässä kappaleessa esitetyt modifioidut ehdot jäännöstermeille *eivät päde*.

Yhden selittäjän lineaarinen regressiomalli

Ennustaminen yhden selittäjän lineaarisella regressiomallilla

Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä

>> 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

Oletukset

- Oletetaan, että toisistaan *riippumattomat* havaintoparit

  $$(x_i , y_i) , i = 1, 2, \ldots, n$$

  noudattavat **2-ulotteista normaalijakaumaa**; ks. monisteen **Todennäköisyyslaskenta** lukua **Moniulotteisia jakaumia.**
- Tällöin *ehdolliset odotusarvot* $\mathrm{E}(x_i \mid y_i)$ ja $\mathrm{E}(y_i \mid x_i)$ ovat muotoa

  $$\mathrm{E}(x_i \mid y_i) = \alpha_0 + \alpha_1 y_i \ , i = 1, 2, \ldots, n$$

  $$\mathrm{E}(y_i \mid x_i) = \beta_0 + \beta_1 x_i \ , i = 1, 2, \ldots, n$$

  eli siis *lineaarisia*.

Regressiomalleja on kaksi

- Voimme kirjoittaa

$$x_i = \alpha_0 + \alpha_1 y_i + \delta_i \ , i = 1, 2, \ldots, n$$

ja

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

jossa jäännöstermit ε_i ja δ_i ovat *keskenään korreloimattomia* satunnaismuuttujia.

Mallien jäännöstermit 1/2

- Mallin

$$x_i = \alpha_0 + \alpha_1 y_i + \delta_i \ , i = 1, 2, \ldots, n$$

jäännös- eli *virhetermit* δ_i toteuttavat seuraavat ehdot:

(i) $\mathrm{E}(\delta_i \mid y_i) = 0 \ , i = 1, 2, \ldots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\mathrm{Var}(\delta_i \mid y_i) = \sigma_\delta^2 \ , i = 1, 2, \ldots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\mathrm{Cor}(\delta_i, \delta_l \mid y_i, y_l) = 0 \ , i \neq l$$

(iv) Jäännöstermit ovat *normaalisia*:

$$\delta_i \mid y_i \sim \mathrm{N}(0, \sigma_\delta^2) \ , i = 1, 2, \ldots, n$$

Mallien jäännöstermit 2/2

- Mallin

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

jäännös- eli *virhetermit* ε_i toteuttavat seuraavat ehdot:

(i) $\mathrm{E}(\varepsilon_i \mid x_i) = 0 \ , i = 1, 2, \ldots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\mathrm{Var}(\varepsilon_i \mid x_i) = \sigma_\varepsilon^2 \ , i = 1, 2, \ldots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\mathrm{Cor}(\varepsilon_i, \varepsilon_l \mid x_i, x_l) = 0 \ , i \neq l$$

(iv) Jäännöstermit ovat *normaalisia*:

$$\varepsilon_i \mid x_i \sim \mathrm{N}(0, \sigma_\varepsilon^2) \ , i = 1, 2, \ldots, n$$

Otostunnusluvut

- Määritellään havaintojen x_i ja y_i , $i = 1, 2, \ldots , n$ *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

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

Parametrien PNS-estimaattorit

- Mallin

$$x_i = \alpha_0 + \alpha_1 y_i + \delta_i \ , i = 1, 2, \ldots, n$$

regressiokertoimien α_1 ja α_0 *PNS-estimaattorit* ovat

$$a_1 = \frac{s_{xy}}{s_y^2} = r_{xy} \frac{s_x}{s_y} \qquad a_0 = \overline{x} - a_1 \overline{y}$$

- Mallin

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

regressiokertoimien β_1 ja β_0 *PNS-estimaattorit* ovat

$$b_1 = \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} \qquad b_0 = \overline{y} - b_1 \overline{x}$$

Estimoidut regressiosuorat 1/3

- Muuttujan *x estimoitu regressiosuora* muuttujan *y* suhteen voidaan kirjoittaa muotoon

$$\frac{x-\bar{x}}{s_x} = r_{xy}\left(\frac{y-\bar{y}}{s_y}\right)$$

- Muuttujan *y estimoitu regressiosuora* muuttujan *x* suhteen voidaan kirjoittaa muotoon

$$\frac{y-\bar{y}}{s_y} = r_{xy}\left(\frac{x-\bar{x}}{s_x}\right)$$

Estimoidut regressiosuorat 2/3

- *Molemmat* estimoidut regressiosuorat voidaan esittää muuttujan *x* funktioina:

  (i) Muuttujan *x estimoitu regressiosuora* muuttujan *y* suhteen:

$$\frac{y-\bar{y}}{s_y} = \frac{1}{r_{xy}}\left(\frac{x-\bar{x}}{s_x}\right)$$

  (ii) Muuttujan *y estimoitu regressiosuora* muuttujan *x* suhteen:

$$\frac{y-\bar{y}}{s_y} = r_{xy}\left(\frac{x-\bar{x}}{s_x}\right)$$

Estimoidut regressiosuorat 3/3

- Estimoitujen regressiosuorien yhtälöistä nähdään:

  Seuraavat ehdot ovat *yhtäpitäviä*:

  (1) Suorat *yhtyvät*.

  (2) $r_{xy} = \pm 1$

  Seuraavat ehdot ovat *yhtäpitäviä*:

  (1)´ Suorat ovat *kohtisuorassa* toisiaan vastaan ja koordinaattiakseleiden suuntaisia.

  (2)´ $r_{xy} = 0$

- Lisäksi yhtälöistä nähdään, että *suorat leikkaavat havaintojen painopisteessä* $(\overline{x}, \overline{y})$.

Estimoidut regressiosuorat ja 2-ulotteisen normaalijakauman regressiofunktiot 1/2

- Olkoon satunnaismuuttujien *x* ja *y* yhteisjakauma 2-*ulotteinen normaalijakauma*.
- Tällöin muuttujan *x regressiofunktion yhtälö* muuttujan *y* suhteen on

$$\mu_{x|y} = \mathrm{E}(x \mid y) = \mu_x + \rho_{xy} \frac{\sigma_x}{\sigma_y}(y - \mu_y)$$

- Siten muuttujien *x* ja *y* havaituista arvoista x_j ja y_j , $j = 1, 2, \ldots, n$ *estimoitu regressiosuora*

$$x = \bar{x} + r_{xy} \frac{s_x}{s_y}(y - \bar{y})$$

saadaan muodollisesti *korvaamalla* regressiofunktion *parametrit* μ_x, μ_y, σ_x^2, σ_y^2, ρ_{xy} *vastaavilla otossuureilla*.

Estimoidut regressiosuorat ja 2-ulotteisen normaalijakauman regressiofunktiot 2/2

- Olkoon satunnaismuuttujien *x* ja *y* yhteisjakauma 2-*ulotteinen normaalijakauma*.
- Tällöin muuttujan *y regressiofunktion yhtälö* muuttujan *x* suhteen on

$$\mu_{y|x} = \mathrm{E}(y \mid x) = \mu_y + \rho_{xy} \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

- Siten muuttujien *x* ja *y* havaituista arvoista x_j ja y_j , $j = 1, 2, \ldots, n$ *estimoitu regressiosuora*

$$y = \bar{y} + r_{xy} \frac{s_y}{s_x}(x - \bar{x})$$

saadaan muodollisesti *korvaamalla* regressiofunktion *parametrit* μ_x, μ_y, σ_x^2, σ_y^2, ρ_{xy} *vastaavilla otossuureilla*.

Regressiosuorien estimointi: Esimerkki 1/8

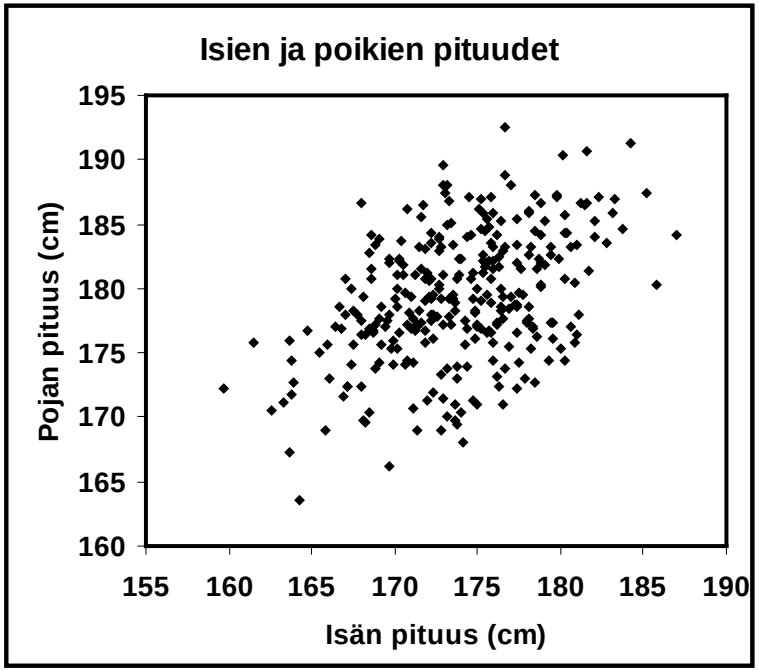
- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:n isän ja heidän poikiensa pituuksien muodostamasta lukuparista

  $(x_i , y_i) , i = 1, 2, \ldots , 300$

  jossa

  x_i = isän i pituus

  y_i = isän i pojan pituus
- Ks. pistediagrammia oikealla.



Regressiosuorien estimointi: Esimerkki 2/8

- Taulukko oikealla esittää isien ja heidän poikiensa pituuksien *ehdollisia keskiarvoja*

  $M_k(x|x)$ ja $M_k(y|x)$

  jossa

  $M_k(x|x)$ = niiden *isien* pituuksien keskiarvo, joiden pituus kuuluu x-väliin k

  $M_k(y|x)$ = niiden *poikien* pituuksien keskiarvo, joiden *isien* pituus kuuluu x-väliin k

  $k = 1, 2, 3, 4, 5, 6, 7$

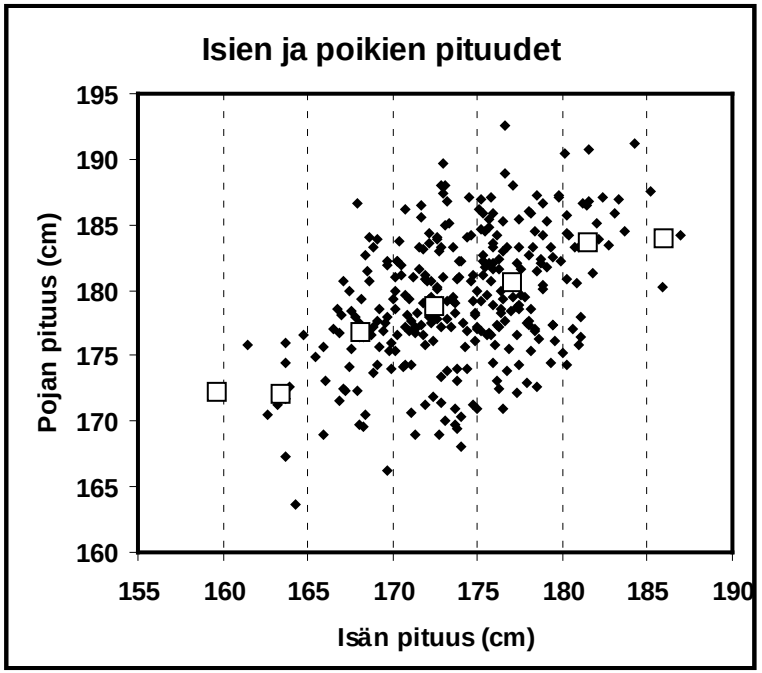
x-välin nro	x-väli	$M_k(x\|x)$	$M_k(y\|x)$
1	(155,160]	159.7	172.2
2	(160,165]	163.5	172.0
3	(165,170]	168.2	176.8
4	(170,175]	172.6	178.8
5	(175,180]	177.1	180.6
6	(180,185]	181.5	183.6
7	(185,190]	186.0	184.0

Regressiosuorien estimointi: Esimerkki 3/8

- *Ehdollisten keskiarvojen*

  $(M_k(x|x), M_k(y|x))$

  määräämiä pisteitä on merkitty kuviossa oikealla *neliöillä*.
- Havainnot on siis luokiteltu *isien* pituuden mukaan 7 luokkaan.
- Kuviossa luokkia on kuvattu katkoviivojen erottamilla pystyvöillä.
- Jokaisen *neliön koordinaatit* on saatu laskemalla keskiarvot ko. neliötä vastaavaan pystyvyöhön kuuluvien havaintopisteiden koordinaateista.



Regressiosuorien estimointi: Esimerkki 4/8

- Olkoon mallina

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

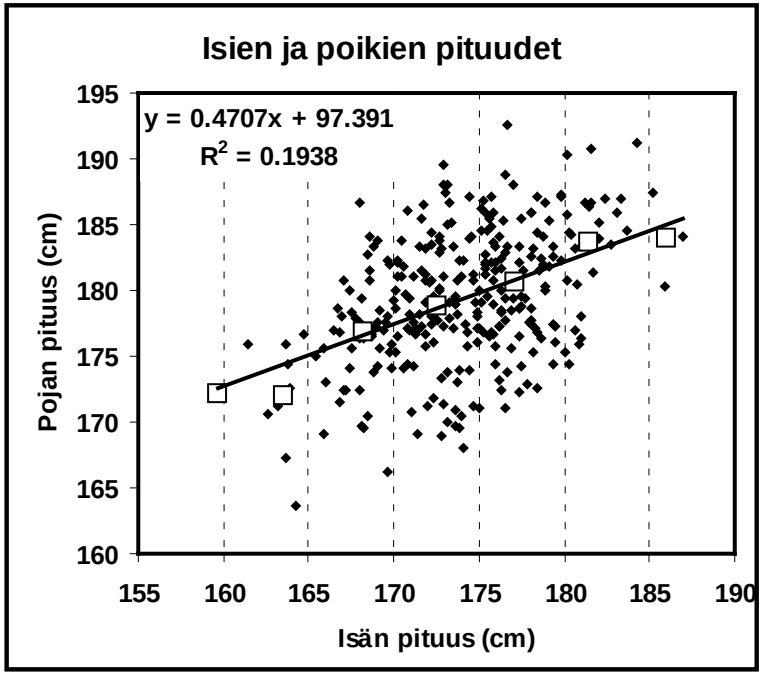
$$i = 1, 2, \ldots, n$$

- Alkuperäisistä havainnoista *estimoidun regressiosuoran* yhtälö on

$$y = 97.391 + 0.4707x$$

- Selitysaste on

$$R^2 = 0.194$$



Regressiosuorien estimointi: Esimerkki 5/8

- Taulukko oikealla esittää isien ja heidän poikiensa pituuksien *ehdollisia keskiarvoja*

  $M_k(x|y)$ ja $M_k(y|y)$

  jossa

  $M_k(x|y)$ = niiden *isien* pituuksien keskiarvo, joiden *poikien* pituus kuuluu y-väliin k

  $M_k(y|y)$ = niiden *poikien* pituuksien keskiarvo, joiden pituus kuuluu y-väliin k

  $k = 1, 2, 3, 4, 5, 6, 7$

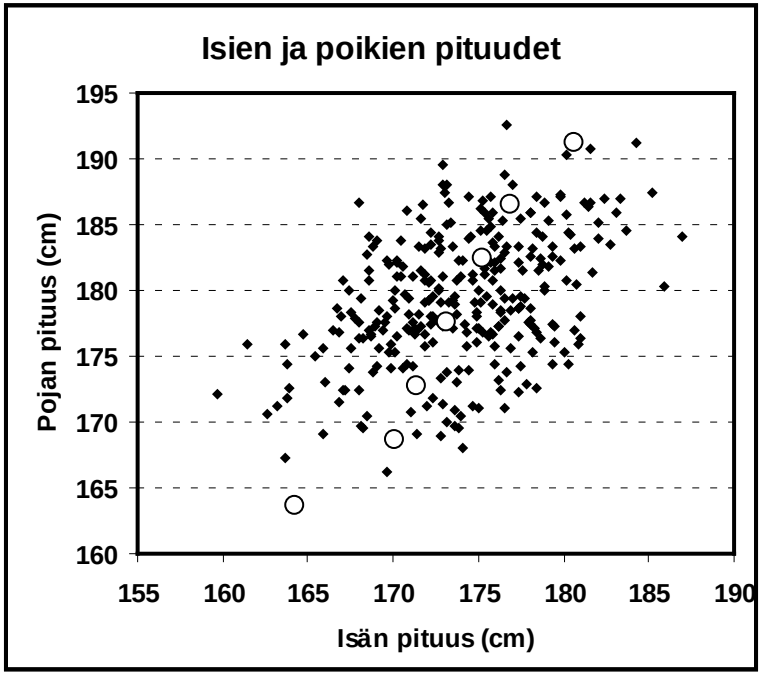
y-välin nro	y-väli	$M_k(x\|y)$	$M_k(y\|y)$
1	(160,165]	164.3	163.6
2	(165,170]	170.1	168.7
3	(170,175]	171.4	172.7
4	(175,180]	173.1	177.6
5	(180,185]	175.2	182.4
6	(185,190]	176.9	186.6
7	(190,195]	180.6	191.2

Regressiosuorien estimointi: Esimerkki 6/8

- *Ehdollisten keskiarvojen*

  $M_k(x|y)$ ja $M_k(y|y)$

  määräämiä pisteitä on merkitty kuviossa oikealla *ympyröillä*.
- Havainnot on siis luokiteltu *poikien* pituuden mukaan 7 luokkaan.
- Kuviossa luokkia on kuvattu katkoviivojen erottamilla *vaakavöillä*.
- Jokaisen *ympyrän koordinaatit* on saatu laskemalla keskiarvot ko. ympyrää vastaavaan *vaakavyöhön* kuuluvien havaintopisteiden koordinaateista.



Regressiosuorien estimointi: Esimerkki 7/8

- Olkoon mallina

$$x_i = \alpha_0 + \alpha_1 x_i + \delta_i$$

$$i = 1, 2, \ldots, n$$

- Alkuperäisistä havainnoista *estimoidun regressiosuoran* yhtälö on

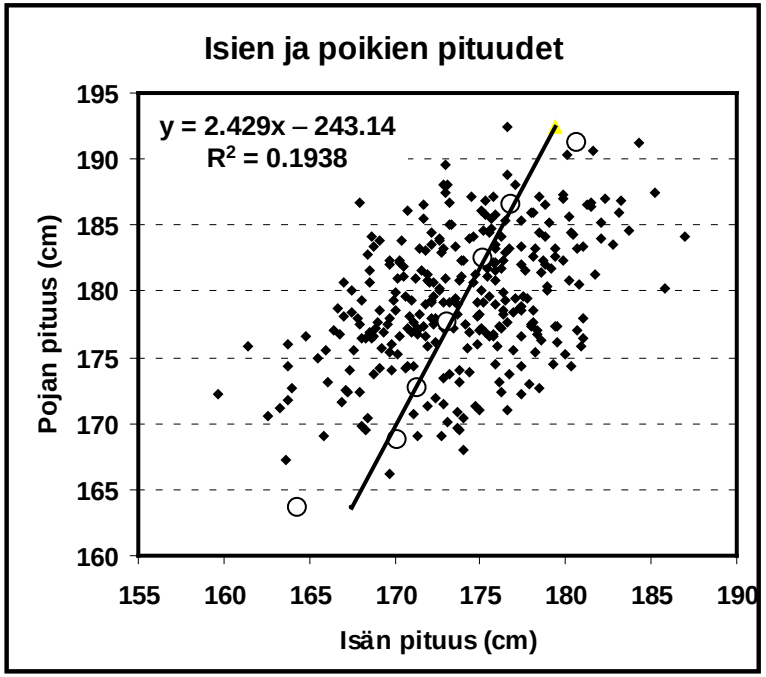
$$x = 100.10 + 0.4117\, y$$

joka voidaan x:n funktiona kirjoittaa muotoon

$$y = -243.14 + 2.429x$$

- Selitysaste on

$$R^2 = 0.194$$



Regressiosuorien estimointi: Esimerkki 8/8

- Kuvioon oikealla on lisätty malleja

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$x_i = \alpha_0 + \alpha_1 x_i + \delta_i$$

vastaavat *estimoidut regressiosuorat*.

- Muuttujan y regressiosuora muuttujan x suhteen:

$$y = 97.391 + 0.4707x$$

- Muuttujan x regressiosuora muuttujan y suhteen muuttujan x funktiona:

$$y = -243.14 + 2.429x$$