

---

Johdatus tilastotieteeseen

# Yhden selittäjän lineaarinen regressiomalli

# Yhden selittäjän lineaarinen regressiomalli

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**Yhden selittäjän lineaarisen regressiomallin estimointi**

**Varianssianalyysihajotelma ja selityssaste**

**Päättely yhden selittäjän lineaarisesta regressiomallista**

**Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**2-ulotteisen normaalijakauman regressiofunktioiden estimointi**

# Yhden selittäjän lineaarinen regressiomalli: Mitä opimme? – 1/3

---

- **Yhden selittäjän lineaarinen regressiomalli** pyrkii selittämään *selitettävän muuttujan havaittujen arvojen vaihtelun yhden selittävän muuttujan havaittujen arvojen vaihtelun avulla.*
- Tässä luvussa tarkastellaan seuraavia **yhden selittävän muuttujan lineaarisen regressiomallin** soveltamiseen liittyviä kysymyksiä:
  - Miten malli **formuloidaan**?
  - Mitkä ovat mallin **osat** ja mitkä ovat osien **tulkinnat**?
  - Mitkä ovat mallia koskevat **oletukset**?
  - Miten mallin **parametrit estimoidaan**?
  - Miten mallin **parametreja koskevia hypoteeseja testataan**?
  - Miten mallin **hyvyyttä mitataan**?
  - Miten mallilla **ennustetaan**?

# Yhden selittäjän lineaarinen regressiomalli: Mitä opimme? – 2/3

---

- Regressiomallien parametrien estimointiin käytetään tavallisesti **pienimmän neliösumman menetelmää**.
- Estimoidun regressiomallin *hyvyyttä* mitataan **selitysasteella**.
- Selitysasteen määritelmä perustuu ns. **varianssianalyysihajotelmaan**.
- Varianssianalyysihajotelmassa **selitettävän muuttujan havaittujen arvojen vaihtelua kuvaava neliösumma** on jaettu kahdeksi neliösummaksi, joista toinen kuvaa mallin ja havaintojen **yhteensopivuutta** ja toinen mallin ja havaintojen **yhteensopimattomuutta**.

# Yhden selittäjän lineaarinen regressiomalli: Mitä opimme? – 3/3

---

- Lineaarisen regressiomallin perusoletuksiin kuuluu se, että selittävien muuttujien arvot ovat **ei-satunnaisia**.
- Selittävän muuttujan arvojen satunnaisuus ei kuitenkaan vaikuta mallin estimoinnissa ja testauksessa käytettäviin menetelmiin seuraavissa tilanteissa:
  - Tavanomaiset mallista tehdyt oletukset pätevät (sopivasti modifioituina), kun siirrytään tarkastelemaan selittävän muuttujan **ehdollista odotusarvoa** selittäjien suhteen.
  - Selitettävä muuttuja ja selittäjät noudattavat yhdessä **multinormaalijakaumaa**.

# Yhden selittäjän lineaarinen regressiomalli: Esitiedot

---

- Esitiedot: ks. seuraavia lukuja:
  - Tilastollinen riippuvuus ja korrelaatio**
  - Johdatus regressioanalyysiin**
- Tarvitset esitietoja myös seuraavista kalvokokoelman **Johdatus todennäköisyyslaskentaan** luvuista:
  - Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat**
  - Moniulotteisia todennäköisyysjakaumia**

# Yhden selittäjän lineaarinen regressiomalli: Lisätiedot

---

- Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään luentosarjan **Tilastollisen analyysin perusteet** luvuissa
  - Yleinen lineaarinen malli**
  - Regressiodiagnostiikka**
  - Regressiomallin valinta**
  - Regressioanalyysin erityiskysymyksiä**

# Yhden selittäjän lineaarinen regressiomalli

---

- >> **Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**
- Yhden selittäjän lineaarisen regressiomallin estimointi**
- Varianssianalyysihajotelma ja selityssaste**
- Päättely yhden selittäjän lineaarisesta regressiomallista**
- Ennustaminen yhden selittäjän lineaarisella regressiomallilla**
- Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**
- 2-ulotteisen normaalijakauman regressiofunktioiden estimointi**



# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

---

## *Avainsanat*

**Ei-satunnaisuus**

**Havainto**

**Heteroskedastisuus**

**Homoskedastisuus**

**Homoskedastisuusoletus**

**Jäännöstermi**

**Jäännösvarianssi**

**Lineaarinen regressiomalli**

**Korreloitumattomuusoletus**

**Korreloituneisuus**

**Kulmakerroin**

**Lineaarisuus**

**Normaalisuusoletus**

**Odotusarvo**

**Regressiokerroin**

**Regressiosuora**

**Satunnainen osa**

**Satunnaisuus**

**Selitettävä muuttuja**

**Selittäjä**

**Selitettävä muuttuja**

**Standardioletukset**

**Systemaattinen osa**

**Vaihtelu**

**Vakioselittäjä**

**Virhetermi**

## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

# Selitettävä muuttuja ja selittävä muuttuja

---

- Oletetaan, että **selitettävän muuttujan**  $y$  *havaintujen arvojen vaihtelu halutaan selittää selittävän muuttujan eli selittäjän*  $x$  *havaintujen arvojen vaihtelun avulla.*
- Tehdään seuraavat oletukset:
  - (i) Selitettävä muuttuja  $y$  on *suhdeasteikollinen satunnaismuuttuja.*
  - (ii) Selittävä muuttuja  $x$  on *kiinteä eli ei-satunnainen muuttuja.*

Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Selitettävä muuttuja ja selittävä muuttuja:

### Kommentti

---

- *Satunnaisen selittäjän tapausta* käsitellään tämän luvun lopussa kappaleissa **Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä ja 2-ulotteisen normaalijakauman regressiofunktioiden estimointi.**

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Havainnot

---

- Olkoot

$$y_1, y_2, \dots, y_n$$

selitettävän muuttujan  $y$  ja

$$x_1, x_2, \dots, x_n$$

selittävän muuttujan  $x$  **havaittuja arvoja**.

- Oletetaan lisäksi, että havaintoarvot  $x_j$  ja  $y_j$  liittyvät *samaan havaintoyksikköön kaikille  $j = 1, 2, \dots, n$* .
- Tällöin havaintoarvot  $x_j$  ja  $y_j$  muodostavat *pisteitä 2-ulotteisessa avaruudessa*:

$$(x_j, y_j) \in \mathbb{R}^2, \quad j = 1, 2, \dots, n$$

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Malli ja sen osat 1/2

---

- Oletetaan, että havaintoarvojen  $y_j$  ja  $x_j$  välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

$y_j$  = **selitettävän muuttujan**  $y$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$x_j$  = **selittävän muuttujan** eli **selittäjän**  $x$  *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$\varepsilon_j$  = **jäännös-** eli **virhetermin**  $\varepsilon$  *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä  $j$

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Malli ja sen osat 2/2

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat *regressiokertoimet*:

$\beta_0$  = **vakioselittäjän regressiokerroin**;

$\beta_0$  on *ei-satunnainen ja tuntematon vakio*

$\beta_1$  = **selittäjän  $x$  regressiokerroin**;

$\beta_1$  on *ei-satunnainen ja tuntematon vakio*

- Huomautus:

Regressiokertoimet  $\beta_0$  ja  $\beta_1$  on oletettu *samoiksi* kaikille havaintoyksiköille  $j$ .

## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

# Vakioselittäjä

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

kerrointa  $\beta_0$  kutsutaan *vakioselittäjän* regressio-kerroimeksi.

- Nimitys johtuu siitä, että kerrointa  $\beta_0$  vastaa *keinotekoinen selittäjä*, joka saa kaikille havaintoyksiköille  $j = 1, 2, \dots, n$  vakioarvon 1.

- Huomautus:

Jatkossa esitettävät kaavat *eivät välttämättä päde* tässä esitettävässä muodossa, jos mallissa *ei ole* vakioselittäjää.

- **Oletamme jatkossa, että mallissa on aina vakioselittäjä.**

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Standardioletukset jäännöstermeistä 1/2

---

- Tehdään yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jäännös- eli virhetermeistä  $\varepsilon_j$  ns. **standardioletukset**:

(i)  $E(\varepsilon_j) = 0, \quad j = 1, 2, \dots, n$

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

$$\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, \quad j \neq l$$



## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

# Standardioletukset jäännöstermeistä 2/2

---

- Lisäksi jäännös- eli virhetermeistä  $\varepsilon_j$  tehdään tavallisesti *normaalisuusoletus*:  
(iv)  $\varepsilon_j \sim N(0, \sigma^2)$ ,  $j = 1, 2, \dots, n$
- Huomautus:  
Oletus (iv) sisältää oletukset (i) ja (ii).

## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

# Selitettävän muuttujan ominaisuudet

---

- Jos yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös- eli virhetermejä  $\varepsilon_j$  koskevat standardioletukset*

*(i)-(iii) pätevät, mallin selitettävän muuttujan  $y$  havaituilla arvoilla  $y_j$  on seuraavat stokastiset ominaisuudet:*

(i)'  $E(y_j) = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \dots, n$

(ii)'  $\text{Var}(y_j) = \sigma^2, \quad j = 1, 2, \dots, n$

(iii)'  $\text{Cor}(y_j, y_l) = 0, \quad j \neq l$

- Jos lisäksi jäännös- eli virhetermejä  $\varepsilon_j$  koskeva *normaalisuusoletus (iv) pätee*, niin

(iv)'  $y_j \sim N(\beta_0 + \beta_1 x_j, \sigma^2), \quad j = 1, 2, \dots, n$

## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

### Mallin parametrit

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

**parametreja** ovat mallin **regressiokertoimet**  $\beta_0$  ja  $\beta_1$  sekä jäännös- eli virhetermien  $\varepsilon_j$  yhteinen *varianssi*

$$\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

jota kutsutaan **jäännösvarianssiksi**.

- Koska regressiokertoimet  $\beta_0$  ja  $\beta_1$  sekä jäännösvarianssi  $\sigma^2$  ovat tavallisesti *tuntemattomia*, ne on *estimoitava* muuttujien  $x$  ja  $y$  havaituista arvoista  $x_j$  ja  $y_j$ ,  
 $j = 1, 2, \dots, n$ .

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Mallin systemaattinen ja satunnainen osa 1/2

---

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös- eli virhetermejä  $\varepsilon_j$  koskeva standardioletus*

(i)  $E(\varepsilon_j) = 0, \quad j = 1, 2, \dots, n$

*pätee.*

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

$$y_j = E(y_j) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jossa

$$E(y_j) = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \dots, n$$

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Mallin systemaattinen ja satunnainen osa 2/2

---

- Odotusarvo

$$E(y_j) = \beta_0 + \beta_1 x_j, j = 1, 2, \dots, n$$

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

- Jäännös- eli virhetermi

$$\varepsilon_j, j = 1, 2, \dots, n$$

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

# Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

## Regressiosuora

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*systemaattinen osa*  $E(y_j) = \beta_0 + \beta_1 x_j$  määrittelee suoran

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

avaruudessa  $\mathbb{R}^2$ .

- Suoraa kutsutaan **regressiosuoraksi** ja sen yhtälössä

$\beta_0$  = regressiosuoran ja y-akselin **leikkauspiste**

$\beta_1$  = regressiosuoran **kulmakerroin**

- Jäännös- eli virhetermien  $\varepsilon_j$  varianssi  $\sigma^2$  kuvaa *havaintopisteiden*  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  *vaihtelua regressiosuoran ympärillä.*

## Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

# Regressiosuoran kulmakertoimen tulkinta

---

- Yhden selittäjän lineaarisen regressiomallin systemaattisen osan määrittelemän regressiosuoran

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

kulmakertoimella  $\beta_1$  seuraava **tulkinta**:

- Oletetaan, että *selittäjän  $x$  arvo kasvaa yhdellä yksiköllä*:

$$x \rightarrow x + 1$$

Tällöin kerroin  $\beta_1$  kertoo *paljonko selitettävän muuttujan  $y$  vastaava odotettavissa oleva arvo muuttuu*:

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

# Yhden selittäjän lineaarinen regressiomalli

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**>> Yhden selittäjän lineaarisen regressiomallin estimointi**

**Varianssianalyysihajotelma ja selityssaste**

**Päättely yhden selittäjän lineaarisesta regressiomallista**

**Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**2-ulotteisen normaalijakauman regressiofunktioiden estimointi**



# Yhden selittäjän lineaarisen regressiomallin estimointi

---

## *Avainsanat*

Aritmeettinen keskiarvo

Estimaattori

Estimointi

Harhattomuus

Jäännöstermien neliösumma

Jäännösvarianssi

Keskihajonta

Kulmakerroin

Lineaarinen regressiomalli

Minimointi

Otoskorrelaatiokerroin

Otoskovarianssi

Otostunnusluvut

Otosvarianssi

Painopiste

Pienimmän neliösumman estimaattori

Pienimmän neliösumman menetelmä

Regressiosuora

Residuaali

Sovite

Standardioletukset

Vakioselittäjä

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Estimointiongelma

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimet  $\beta_0$  ja  $\beta_1$  ovat tavallisesti *tuntemattomia*, joten *ne on estimoiva* muuttujien  $x$  ja  $y$  havaituista arvoista  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$ .

- Estimoinnissa regressiokertoimille  $\beta_0$  ja  $\beta_1$  pyritään löytämään sellaiset arvot, että niiden määräämä *regressiosuora selittäisi mahdollisimman hyvin selitettävän muuttujan  $y$  arvojen vaihtelun*.
- Regressiokertoimien  $\beta_0$  ja  $\beta_1$  estimointiin on tarjolla useita erilaisia menetelmiä, joista tavallisesti käytetään *pienimmän neliösumman menetelmää*.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Pienimmän neliösumman menetelmä

---

- **Pienimmän neliösumman menetelmässä** yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  estimaattorit määrätään *minimoimalla jäännös- eli virhetermien  $\varepsilon_j$  neliösumma*

$$\sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2$$

*regressiokertoimien  $\beta_0$  ja  $\beta_1$  suhteen.*

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Otostunnusluvut

---

- Määritellään havaintojen  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$$

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

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Regressiokertoimien PNS-estimaattorit

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  **pienimmän neliösumman (PNS-) estimaattorit** ovat

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

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

# Yhden selittäjän lineaarisen regressiomallin estimointi

## PNS-estimaattoreiden johto 1/4

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimet  $\beta_0$  ja  $\beta_1$  estimoidaan PNS-menetelmällä minimoimalla jäännöstermien  $\varepsilon_j$  neliösumma

$$S(\beta_0, \beta_1) = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2$$

kertoimien  $\beta_0$  ja  $\beta_1$  suhteen

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

# Yhden selittäjän lineaarisen regressiomallin estimointi

## PNS-estimaattoreiden johto 2/4

---

- Derivoidaan funktio

$$S(\beta_0, \beta_1) = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  suhteen ja merkitään derivaatat nolliksi:

$$(1) \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = -2 \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j) = 0$$

$$(2) \quad \frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = -2 \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j) x_j = 0$$

- Regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit saadaan normaaliyhtälöiden (1) ja (2) ratkaisuna.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## PNS-estimaattoreiden johto 3/4

---

- Kirjoitetaan normaaliyhtälöt (1) ja (2) muotoihin

$$(1)' \quad \sum_{j=1}^n y_j - n\beta_0 - \beta_1 \sum_{j=1}^n x_j = 0$$

$$(2)' \quad \sum_{j=1}^n y_j x_j - \beta_0 \sum_{j=1}^n x_j - \beta_1 \sum_{j=1}^n x_j^2 = 0$$

- Ratkaistaan  $\beta_0$  yhtälöstä (1)':

$$(3) \quad \beta_0 = \frac{1}{n} \sum_{j=1}^n y_j - \beta_1 \frac{1}{n} \sum_{j=1}^n x_j = \bar{y} - \beta_1 \bar{x}$$

ja sijoitetaan ratkaisu yhtälöön (2)':

$$(4) \quad \sum_{j=1}^n y_j x_j - n\bar{y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{j=1}^n x_j^2 = 0$$



# Yhden selittäjän lineaarisen regressiomallin estimointi

## PNS-estimaattoreiden johto 4/4

---

- Parametrin  $\beta_1$  PNS-estimaattoriksi saadaan yhtälöstä (4):

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

- Sijoittamalla  $b_1$  yhtälöön (3) saadaan parametrin  $\beta_0$  PNS-estimaattoriksi

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

- Sivuuutetaan sen osoittaminen, että saatu ääriarvo on todellakin *minimi*.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Regressiokertoimien laskeminen 1/3

---

- Oletetaan, että haluamme laskea yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaatit *käsin* tai käyttämällä *laskinta*.

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

Samasta kaaviosta voidaan laskea myös muuttujien  $x$  ja  $y$  havaittujen arvojen *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskeskihajonnat*, *otoskovarianssi* ja *otoskorrelaatio*; ks. lukua **Tilastollinen riippuvuus ja korrelaatio**.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Regressiokertoimien laskeminen 2/3

- 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	$\sum_{i=1}^n x_i$	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n x_i^2$	$\sum_{i=1}^n y_i^2$	$\sum_{i=1}^n x_i y_i$

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Regressiokertoimien laskeminen 3/3

---

- Regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaatit saadaan havaintoarvojen *summista*, *neliösummista* ja *tulosummasta* alla esitetyillä kaavoilla:

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

$$b_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{\sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2}$$

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

# Yhden selittäjän lineaarisen regressiomallin estimointi

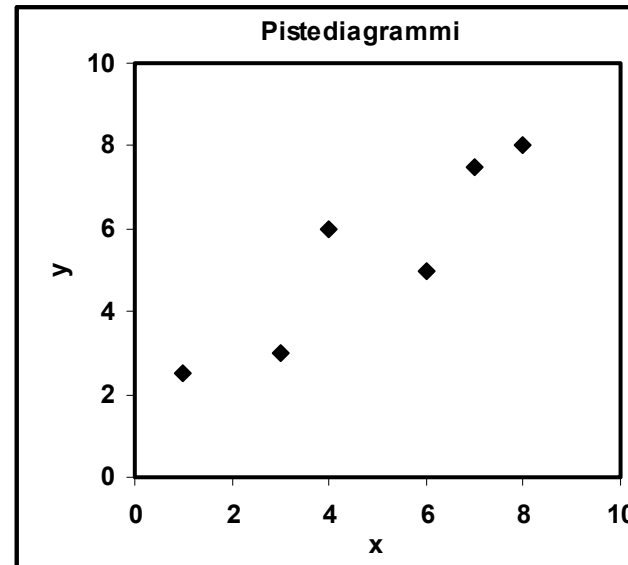
## Tunnuslukujen laskeminen:

### 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



# Yhden selittäjän lineaarisen regressiomallin estimointi

## 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
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>175</b>	<b>196.5</b>	<b>182</b>

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaatit* voidaan laskea näistä viidestä summasta; ks. seuraavaa kalvoa.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Tunnuslukujen laskeminen:

### Havainnollistava esimerkki 3/3

---

- Regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaatit:

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

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

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

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

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Estimoitu regressiosuora 1/3

---

- Yhden selittäjän lineaarinen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  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**



## Yhden selittäjän lineaarisen regressiomallin estimointi

### Estimoitu regressiosuora 2/3

---

- Sijoitetaan regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattoreiden lausekkeet

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

*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  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  painopisteen  $(\bar{x}, \bar{y})$  kautta.

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Estimoitu regressiosuora 3/3

---

- Estimoidulla regressiosuoralla

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

on seuraavat ominaisuudet:

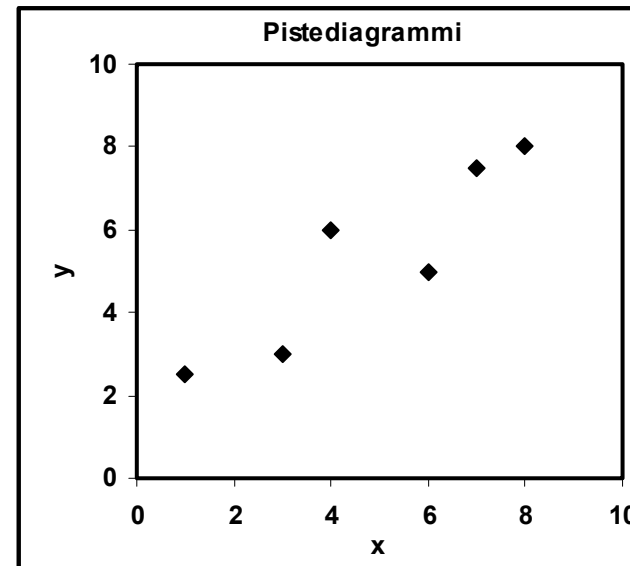
- (i) Jos  $r_{xy} > 0$ , suora on *nouseva*.
- (ii) Jos  $r_{xy} < 0$ , suora on *laskeva*.
- (iii) Jos  $r_{xy} = 0$ , suora on *vaakasuorassa*.
- (iv) Suora *jyrkkenee (loivenee)*, jos
  - korrelaation itseisarvo  $|r_{xy}|$  *kasvaa (pienenee)*
  - keskihajonta  $s_y$  *kasvaa (pienenee)*
  - keskihajonta  $s_x$  *pienenee (kasvaa)*

# Estimoitu regressiosuora: Havainnollistava esimerkki 1/2

---

- Taulukossa oikealla on keinotekoisesti 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

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

$$j = 1, 2, \dots, n$$

*regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaateiksi saatiin edellä*

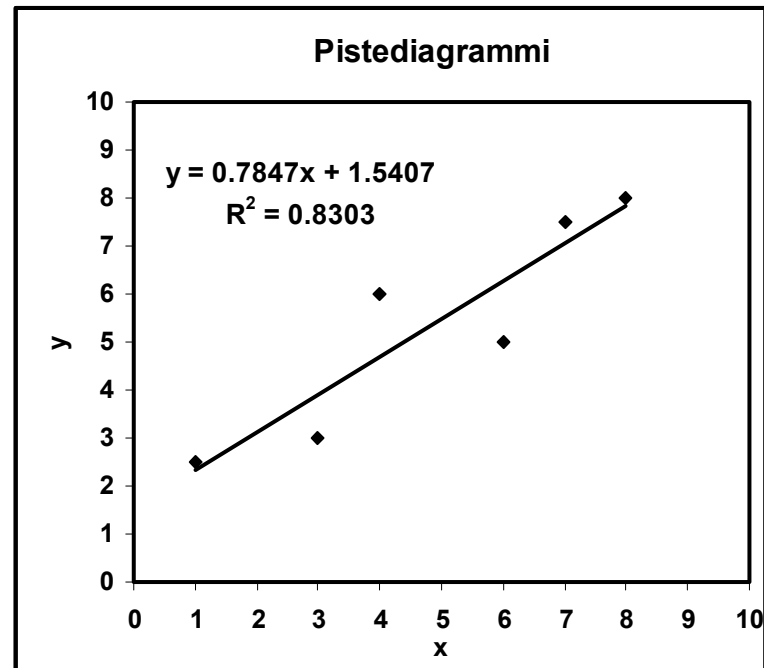
$$b_0 = 1.5407$$

$$b_1 = 0.7847$$

- *Estimoidun regressiosuoran yhtälö on siten*

$$y = 1.5407 + 0.7847x$$

ks. kuviota oikealla.



## Regressiosuoran estimointi:

### 1. esimerkki 1/2

---

- *Hooken lain* mukaan (ideaalisen) kierrejousen pituus  $y$  riippuu *lineaarisesti* jouseen ripustetusta painosta  $x$ :

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

jossa

$\alpha$  = jousen pituus ilman painoa

$\beta$  = ns. *jousivakio*

- Jousivakion määrittämiseksi jouseen ripustettiin seuraavat painot: 0, 2, 4, 6, 8, 10 kg ja jousen pituus mitattiin.
- Mittaustulokset on annettu taulukossa oikealla.

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

## Regressiosuoran estimointi:

### 1. esimerkki 2/2

- *Estimoidun regressiosuoran yhtälö* on

$$y = 43.055 + 0.2457x$$

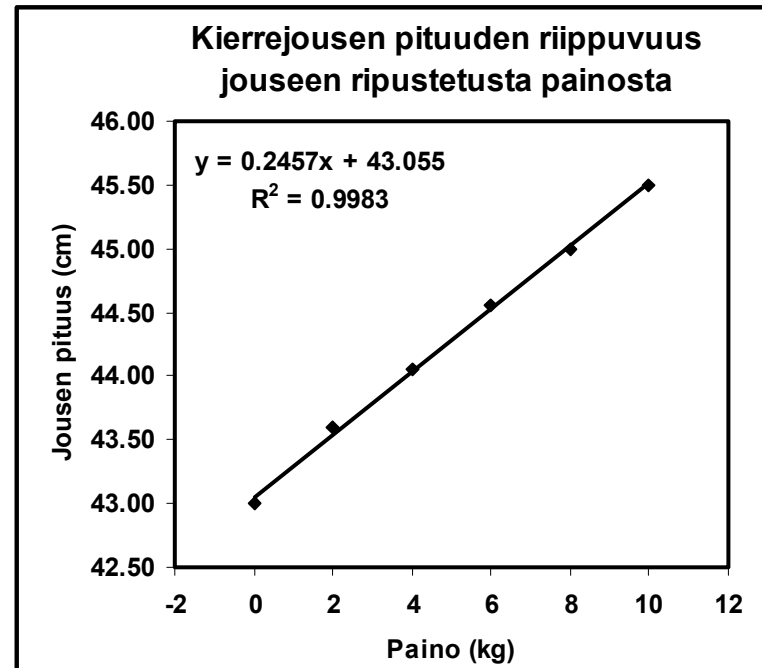
ks. kuviota oikealla.

- Suoran kulmakertoimen

$$b = 0.2457$$

tulkinta:

Jouseen ripustetun painon lisääminen 1 kg:lla pidentää jouta keskimäärin 0.2457 cm:llä.



## Regressiosuoran estimointi:

### 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

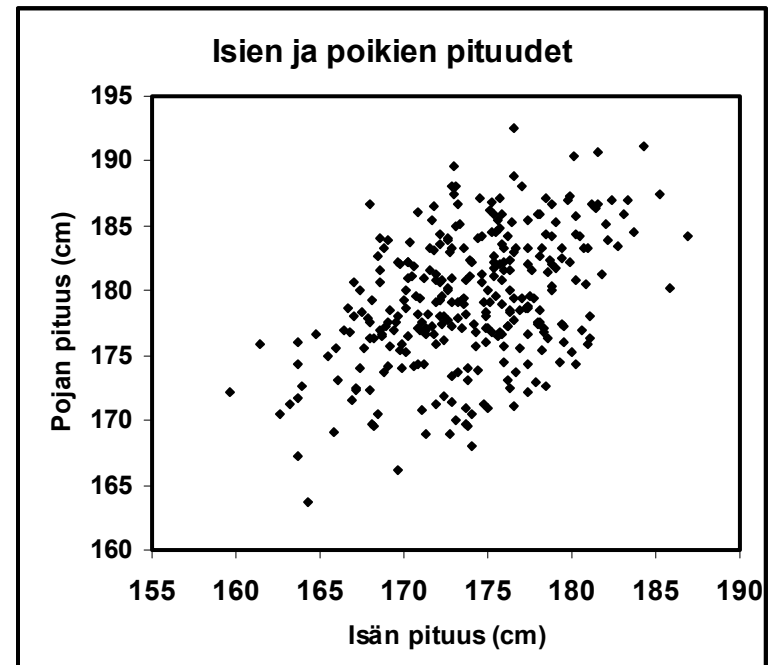
$$(x_j, y_j), j = 1, 2, \dots, 300$$

jossa

$x_j$  = isän  $j$  pituus

$y_j$  = isän  $j$  pojan pituus

- Ks. pistediagrammia oikealla.



## Regressiosuoran estimointi:

### 2. esimerkki – 2/2

- *Estimoidun regressiosuoran yhtälö on*

$$y = 97.391 + 0.4707x$$

ks. kuviota oikealla.

- Suoran kulmakertoimen

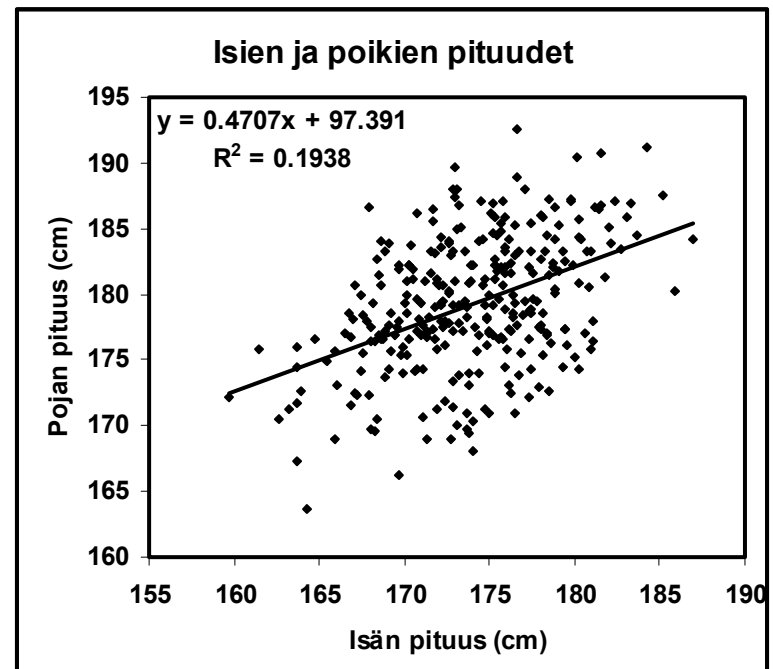
$$b = 0.4707$$

tulkinta:

Jos isä  $A$  on 1 cm pitempi kuin isä  $B$ , isä  $A$ :n poika on *keskimäärin*

$$0.4707 \text{ cm}$$

pitempi kuin isä  $B$ :n poika.





## Regressiosuoran estimointi:

### 3. esimerkki – 1/2

---

- Onko keuhkosityöpä yleisempää sellaisissa maissa, joissa tupakoidaan paljon?
- Oikealla on tiedot savukkeiden kulutuksesta ja keuhkosityövän yleisyydestä 10:ssä maassa.
- Havaintoaineisto koostuu 10:stä lukuparista

$$(x_j, y_j), j = 1, 2, \dots, 10$$

jossa

$x_j$  = savukkeiden kulutus maassa  $j$  1930

$y_j$  = sairastuvuus keuhkosityöpään maassa  $j$  1950

Maa	Savukkeiden kulutus (kpl) per capita 1930	Keuhkosityöpätapausten lkm per 1 milj. henkilöä 1950
Islanti	220	58
Norja	250	90
Ruotsi	310	115
Kanada	510	150
Tanska	380	165
Itävalta	455	170
Hollanti	460	245
Sveitsi	530	250
Suomi	1115	350
Englanti	1145	465

## Regressiosuoran estimointi:

### 3. esimerkki – 2/2

- *Estimoidun regressiosuoran* yhtälö on

$$y = 13.553 + 0.3577x$$

- Suoran kulmakertoimen

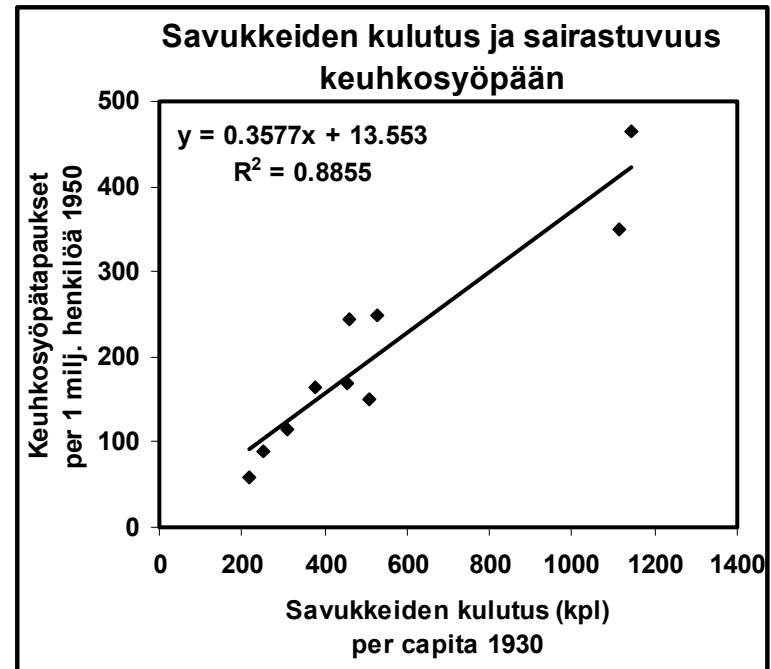
$$b = 0.3577$$

tulkinta:

Jos maassa *A* poltettiin vuonna 1930 sata savuketta enemmän per capita kuin maassa *B*, maassa *A* oli vuonna 1950 *keskimäärin*

$$100 \times 0.3577 \approx 36$$

keuhkosyöpätapausta enemmän per 1 milj. asukasta kuin maassa *B*.



# Yhden selittäjän lineaarisen regressiomallin estimointi

## Sovitteet ja residuaalit

---

- Olkoot  $b_0$  ja  $b_1$  yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Määritellään estimoidun mallin **sovitteet** kaavalla

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

- Määritellään estimoidun mallin **residuaalit** kaavalla

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

- Huomaa, että

$$y_j = \hat{y}_j + e_j, \quad j = 1, 2, \dots, n$$

## Sovitteet ja residuaalit:

### Tulkinnat 1/2

---

- *Sovite*

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

on estimoidun regressiosuoran yhtälön selitettävälle muuttujalle  $y$  antama arvo havaintopisteessä  $x_j$ .

- *Residuaali*

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

on selitettävän muuttujan  $y$  havaitun arvon  $y_j$  ja sovitteen  $\hat{y}_j$  eli estimoidun regressiosuoran yhtälön selitettävälle muuttujalle  $y$  havaintopisteessä  $x_j$  antaman arvon erotus.

## Sovitteet ja residuaalit:

### Tulkinnat 2/2

---

- Estimoitu regressiomalli selittää selitettävän muuttujan  $y$  havaittujen arvojen vaihtelun *sitä paremmin mitä lähempänä estimoidun mallin sovitteet  $\hat{y}_j$  ovat selitettävän muuttujan  $y$  havaittuja arvoja  $y_j$ .*
- Yhtäpitävästi edellisen kanssa:  
Estimoitu regressiomalli selittää selitettävän muuttujan  $y$  havaittujen arvojen  $y_j$  vaihtelun *sitä paremmin mitä pienempiä ovat estimoidun mallin residuaalit  $e_j$ .*

# Sovitteet ja residuaalit: Havainnollistus

- Kuvio oikealla havainnollistaa sovitteiden ja residuaalien geometrista tulkintaa.

- *Malli:*

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

- *PNS-suora:*

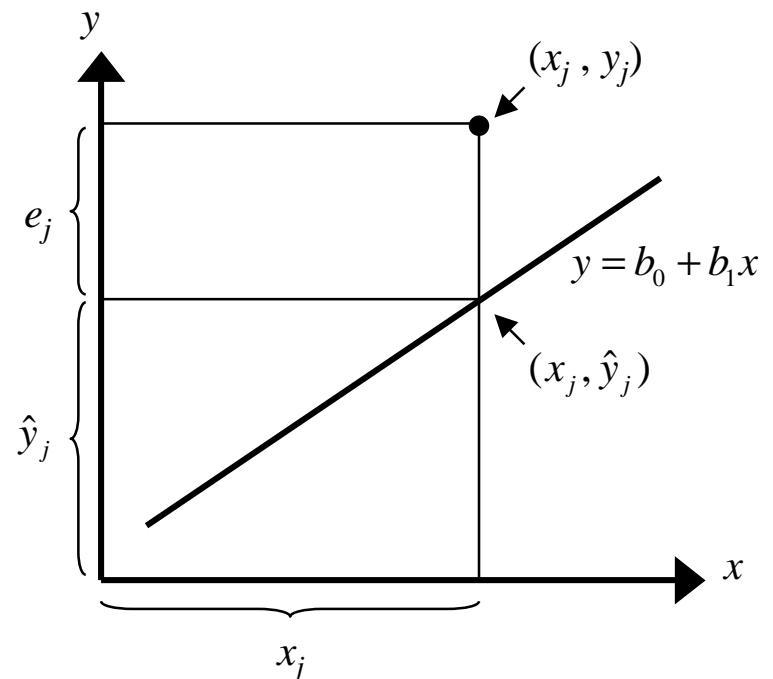
$$y = b_0 + b_1 x$$

- *Sovite:*

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

- *Residuaali:*

$$e_j = y_j - \hat{y}_j, \quad j = 1, 2, \dots, n$$



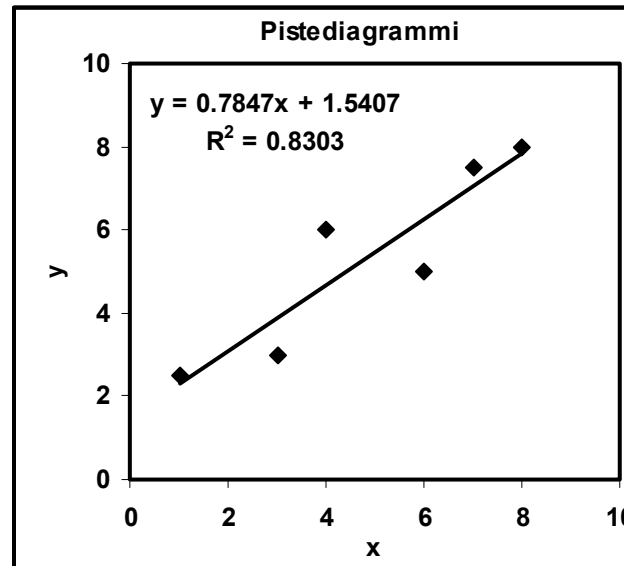
## Sovitteet ja residuaalit:

### Havainnollistava esimerkki 1/3

---

- Taulukossa oikealla on keinotekoisesti 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>i</i>	<i>x</i>	<i>y</i>	<b>Sovite</b>	<b>Residuaali</b>
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
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>32.000</b>	<b>0.000</b>

- Esimerkiksi, kun  $i = 3$ , niin

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

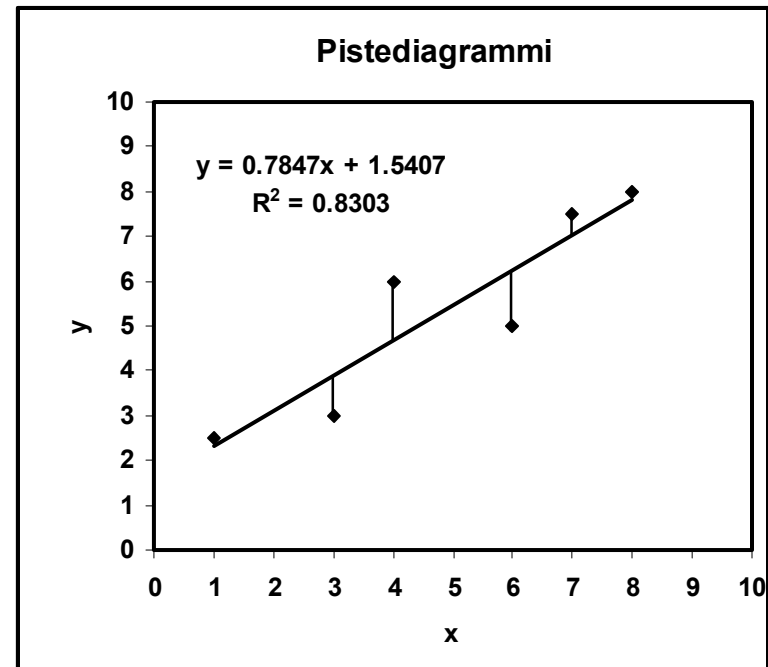
$$e_3 = y_3 - \hat{y}_3 = 6 - 4.679 = 1.321$$



## 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 mallin *residuaaleja* vastaavien janojen pituuksien neliöiden summa on pienin mahdollinen.



## Yhden selittäjän lineaarisen regressiomallin estimointi

# Jäännösvarianssin estimointi 1/2

---

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

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

jossa

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

= estimoidun mallin *residuaali*

$n$  = havaintojen lukumäärä

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Jäännösvarianssin estimointi 2/2

---

- Jäännösvarianssin  $\sigma^2$  estimaattori

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

kuvaa *havaintopisteiden*  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  *vaihtelua* *estimoidun regressiosuoran ympärillä.*

# Yhden selittäjän lineaarisen regressiomallin estimointi

## Jäännösvarianssin estimointi:

### Kommentti

---

- Estimaattori  $s^2$  on *residuaalien*  $e_j$  *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

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n (e_j - \bar{e})^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

# Yhden selittäjän lineaarisen regressiomallin estimointi

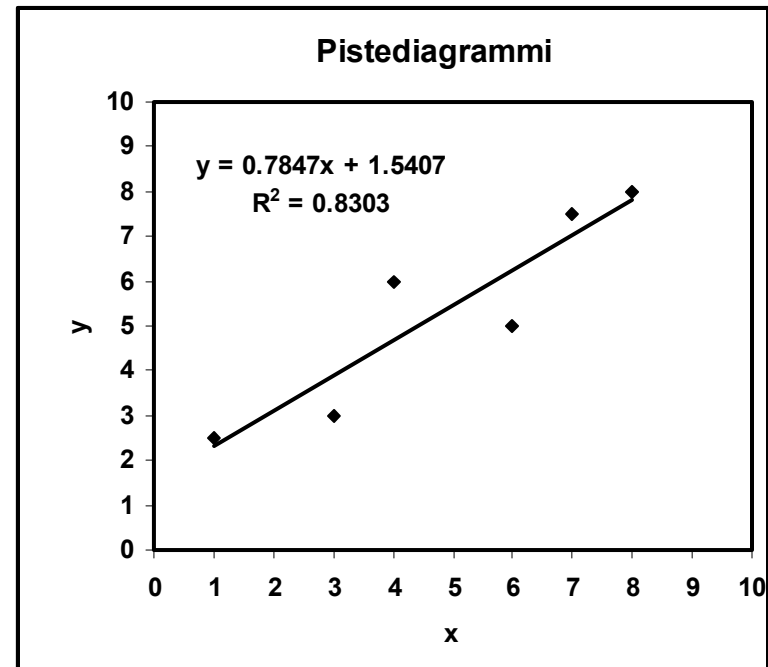
## Jäännösvarianssin estimointi:

### Havainnollistava esimerkki 1/2

- Taulukossa alla on keinotekoisien 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 regressiosuoran yhtälö*.



# Yhden selittäjän lineaarisen regressiomallin estimointi

## 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*  $e^2$ .

$i$	$x$	$y$	<i>Sovite</i>	<i>Residuaali</i>	<i>Res<sup>2</sup></i>
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
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>32.000</b>	<b>0.000</b>	<b>4.385</b>

- *Jäännösvarianssin*  $\sigma^2$  *harhaton estimaattori* on

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

# Yhden selittäjän lineaarinen regressiomalli

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**Yhden selittäjän lineaarisen regressiomallin estimointi**

**>> Varianssianalyysihajotelma ja selitysaste**

**Päättely yhden selittäjän lineaarisesta regressiomallista**

**Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**2-ulotteisen normaalijakauman regressiofunktioiden estimointi**

# Varianssianalyysihajotelma ja selitysaste

---

## *Avainsanat*

Jäännöseliösumma

Jäännösvaihtelu

Kokonaisneliösumma

Kokonaisvaihtelu

Korrelaatio

Lineaarinen regressiomalli

Mallineliosumma

Pienimmän neliösumman estimaattori

Residuaali

Selitysaste

Sovite

Standardioletukset

Varianssianalyysihajotelma



## 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*.

# Varianssianalyysihajotelma ja selitysaste

## Malli ja sen osat 1/2

---

- Oletetaan, että havaintoarvojen  $y_j$  ja  $x_j$  välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

$y_j$  = **selitettävän muuttujan**  $y$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$x_j$  = **selittävän muuttujan** eli **selittäjän**  $x$  *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$\varepsilon_j$  = **jäännös-** eli **virhetermin**  $\varepsilon$  *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä  $j$

# Varianssianalyysihajotelma ja selitysaste

## Malli ja sen osat 2/2

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat *kertoimet*:

$\beta_0$  = **vakioselittäjän regressiokerroin;**

$\beta_0$  on *ei-satunnainen ja tuntematon vakio*

$\beta_1$  = **selittäjän  $x$  regressiokerroin;**

$\beta_1$  on *ei-satunnainen ja tuntematon vakio*

# Varianssianalyysihajotelma ja selityssaste

## Oletukset

---

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jäännös- eli virhetermiä  $\varepsilon_j$  koskevat **standardioletukset** pätevät:

(i)  $E(\varepsilon_j) = 0, j = 1, 2, \dots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, \quad j \neq l$$

# Varianssianalyysihajotelma ja selityaste

## Otostunnusluvut

---

- Määritellään havaintojen  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$$

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

## Regressiokertoimien PNS-estimaattorit

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  **pienimmän neliösumman (PNS-) estimaattorit** ovat

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

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

## Varianssianalyysihajotelma ja selitysaste

# Sovitteet ja residuaalit

---

- Olkoot  $b_0$  ja  $b_1$  yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Määritellään estimoidun mallin **sovitteet** kaavalla

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

- Määritellään estimoidun mallin **residuaalit** kaavalla

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

## Varianssianalyysihajotelma ja selitysaste

# Jäännösvarianssin estimointi

---

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

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

jossa

$e_j$  = estimoidun mallin *residuaali*

$n$  = havaintojen lukumäärä



# Varianssianalyysihajotelma ja selitysaste

## Kokonaisneliösumma

---

- Neliösumma

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

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

- *Selitettävän muuttujan  $y$  havaittujen arvojen  $y_j$  varianssi* voidaan määritellä kaavalla

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

# Varianssianalyysihajotelma ja selityaste

## Jäännösneliösumma

---

- Neliösumma

$$SSE = \sum_{j=1}^n e_j^2$$

kuvaa *residuaalien*  $e_j$  *vaihtelua* ja sitä kutsutaan **jäännösneliösummaksi**.

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

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

- $s^2$  on jäännösvarianssin  $\sigma^2$  *harhaton estimaattori*.

## Kokonais- ja jäännösneliösumman yhteys 1/4

---

- Voidaan osoittaa, että yhden selittäjän lineaarisessa regressiomallissa jäännösneliösumma  $SSE$  ja kokonaisneliösumma  $SST$  toteuttavat yhtälöt

$$SSE = \sum_{j=1}^n e_j^2 = (1 - r_{xy}^2) \sum_{j=1}^n (y_j - \bar{y})^2 = (1 - r_{xy}^2) SST$$

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

$$SSE = \sum_{j=1}^n e_j^2 = (1 - r_{xy}^2) \sum_{j=1}^n (y_j - \bar{y})^2 = (1 - r_{xy}^2) SST$$

nähdään välittömästi, että

$$SSE \leq SST$$

## Kokonais- ja jäännösneliösumman yhteys 3/4

---

- Yhtälöistä

$$SSE = \sum_{j=1}^n e_j^2 = (1 - r_{xy}^2) \sum_{j=1}^n (y_j - \bar{y})^2 = (1 - r_{xy}^2) SST$$

nähdään, että seuraavat ehdot ovat yhtäpitäviä:

- (i)  $SSE = 0$
  - (ii)  $e_j = 0$  kaikille  $j = 1, 2, \dots, n$
  - (iii)  $r_{xy} = \pm 1$
- Jos ehdot (i)-(iii) pätevät, niin kaikki havaintopisteet  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  ovat samalla suoralla ja tätä suoraa vastaava *lineaarinen regressiomalli selittää täydellisesti selitettävän muuttujan  $y$  havaittujen arvojen vaihtelun.*

## Kokonais- ja jäännösneliösumman yhteys 4/4

---

- Yhtälöistä

$$SSE = \sum_{j=1}^n e_j^2 = (1 - r_{xy}^2) \sum_{j=1}^n (y_j - \bar{y})^2 = (1 - r_{xy}^2) SST$$

nähdään, että seuraavat ehdot ovat yhtäpitäviä:

- (i)'  $SSE = SST$
  - (ii)'  $e_j = y_j - \bar{y}$  kaikille  $j = 1, 2, \dots, n$
  - (iii)'  $r_{xy} = 0$
- Jos ehdot (i)'-(iii)' pätevät, niin *selitettävän muuttujan y havaittujen arvojen vaihtelua ei voida selittää lineaarisella regressiomallilla.*

# Varianssianalyysihajotelma ja selitysaste

## Mallineliösumma 1/2

---

- Määritellään suure  $SSM$  yhtälöllä

$$SSM = SST - SSE$$

- Koska

$$0 \leq SSE \leq SST$$

niin

$$SSM \geq 0$$

- Koska voidaan osoittaa, että

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

suuretta  $SSM$  kutsutaan **mallineliösummaksi**.

# Varianssianalyysihajotelma ja selityaste

## Mallineliösumma 2/2

---

- Mallineliösumma  $SSM$  voidaan esittää myös muodossa

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

jossa

$$\bar{\hat{y}} = \frac{1}{n} \sum_{j=1}^n \hat{y}_j = \frac{1}{n} \sum_{j=1}^n y_j = \bar{y}$$



# Varianssianalyysihajotelma ja selityksaste

## Varianssianalyysihajotelma 1/2

---

- Edellä esitetyn mukaan kokonaisneliösumma

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

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

$$SST = SSM + SSE$$

jossa

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

ja

$$SSE = \sum_{j=1}^n e_j^2$$

## 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) **Mallineliosumma**  $SSM$  kuvaa sitä osaa selitettävän muuttujan  $y$  havaittujen arvojen vaihtelusta, jonka *estimoitu malli on selittänyt*.
- (ii) **Jäännöseliosumma**  $SSE$  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 *jäännöseliösumman SSE osuus* kokonaisneliösummasta *SST*, sitä paremmin estimoitu malli selittää selitettävän muuttujan havaittujen arvojen vaihtelun.

# Varianssianalyysihajotelma ja selitysaste

## 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:

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

# Varianssianalyysihajotelma ja selitysaste

## Selitysaste ja korrelaatio

---

- Voidaan osoittaa, että

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

jossa

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

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

- *Yhden selittäjän lineaarisessa regressiomallissa* pätee lisäksi se, että selitysaste  $R^2$  on selitettävän ja selittävän muuttujan havaittujen arvojen *otoskorrelaatiokertoimen*  $r_{xy}$  *neliö*:

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

## Varianssianalyysihajotelma ja selitysaste

# 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:  
 $e_j = 0$ , kaikille  $j = 1, 2, \dots, n$
    - (3) Kaikki havaintopisteet  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$  asettuvat *samalle suoralle*.
    - (4)  $r_{xy} = \pm 1$
    - (5) Määritelty malli *selittää täydellisesti* selitettävän muuttujan  $y$  havaittujen arvojen vaihtelun.

## Varianssianalyysihajotelma ja selitysaste

# 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.

# Varianssianalyysihajotelma ja selitysaste

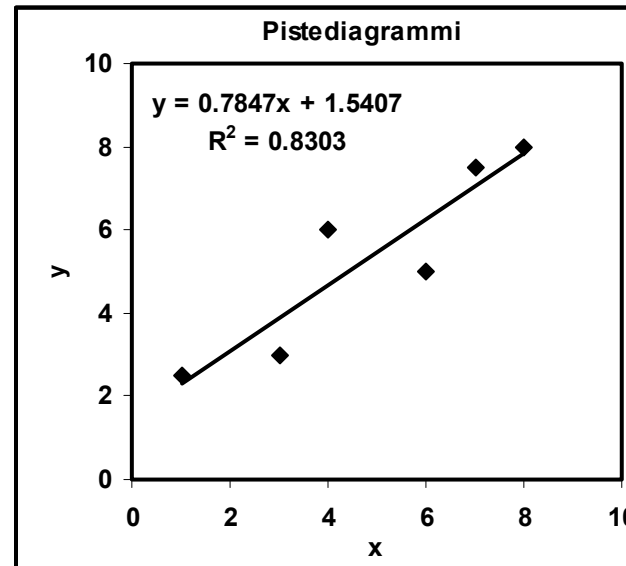
## Selitysasteen laskeminen:

### Havainnollistava esimerkki 1/3

---

- Taulukossa oikealla on keinotekoisesti 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 neliösummat 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
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>175</b>	<b>196.5</b>	<b>32</b>	<b>0.000</b>	<b>4.385</b>

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

# Varianssianalyysihajotelma ja selitysaste

## Selitysasteen laskeminen:

### Havainnollistava esimerkki 3/3

---

- *Kokonaisneliösumma:*

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

- *Jäännösneliösumma:*

$$SSE = \sum_{j=1}^n e_j^2 = 4.385$$

- *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.

# Yhden selittäjän lineaarinen regressiomalli

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**Yhden selittäjän lineaarisen regressiomallin estimointi**

**Varianssianalyysihajotelma ja selitysaste**

**>> Päättely yhden selittäjän lineaarisesta regressiomallista**

**Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**2-ulotteisen normaalijakauman regressiofunktioiden estimointi**

# Päätely yhden selittäjän lineaarisesta regressiomallista

---

## *Avainsanat*

**F-testi**

**Kulmakerroin**

**Lineaarinen regressiomalli**

**Luottamusväli**

**Otosjakauma**

**Pienimmän neliösumman estimaattori**

**Regressiokerroin**

**Selitysaste**

**Standardioletukset**

**Testaus**

**t-testi**

**Vakio**

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Mallia koskeva tilastollinen päätely

---

- Tarkastellaan seuraavia yhden selittäjän lineaarista regressiomallia koskevia päätelyn ongelmia:
  - **Regressiokertoimien estimaattoreiden odotusarvot ja varianssit**
  - **Regressiokertoimien estimaattoreiden otosjakaumat**
  - **Regressiokertoimien luottamusvälit**
  - **Testit regressiokertoimille**
  - **Testi selityksasteelle**

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Malli ja sen osat 1/3

---

- Oletetaan, että havaintoarvojen  $y_j$  ja  $x_j$  välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

$y_j$  = **selitettävän muuttujan**  $y$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$x_j$  = **selittävän muuttujan** eli **selittäjän**  $x$  *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$\varepsilon_j$  = **jäännös-** eli **virhetermin**  $\varepsilon$  *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä  $j$

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Malli ja sen osat 2/3

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat *kertoimet*:

$\beta_0$  = **vakioselittäjän regressiokerroin**;

$\beta_0$  on *ei-satunnainen ja tuntematon vakio*

$\beta_1$  = **selittäjän  $x$  regressiokerroin**;

$\beta_1$  on *ei-satunnainen ja tuntematon vakio*

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Malli ja sen osat 3/3

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

määrittelemän **regressiosuoran**

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

yhtälössä

$\beta_0$  = regressiosuoran ja  $y$ -akselin **leikkauspiste** eli  
regressiosuoran **vakio**

$\beta_1$  = regressiosuoran **kulmakerroin**



# Päätely yhden selittäjän lineaarisesta regressiomallista

## Oletukset

---

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös-* eli *virhetermiä*  $\varepsilon_j$  koskevat *standardioletukset* pätevät:

(i)  $E(\varepsilon_j) = 0, j = 1, 2, \dots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, \quad j \neq l$$

- Lisäksi oletetaan, että virhetermit  $\varepsilon_j$  ovat *normaalisia*:

(iv)  $\varepsilon_j \sim N(0, \sigma^2), j = 1, 2, \dots, n$

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Otostunnusluvut

---

- Määritellään havaintojen  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$$

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

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Regressiokertoimien PNS-estimaattorit

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  **pienimmän neliösumman (PNS-) estimaattorit** ovat

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

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

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Sovitteet ja residuaalit

---

- Olkoot  $b_0$  ja  $b_1$  yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Määritellään estimoidun mallin **sovitteet** kaavalla

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

- Määritellään estimoidun mallin **residuaalit** kaavalla

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Jäännösvarianssin estimointi

---

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

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

jossa

$e_j$  = estimoidun mallin *residuaali*

$n$  = havaintojen lukumäärä

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Regressiokertoimien estimaattorit: Odotusarvot ja varianssit

---

- *Jos standardioletukset (i)-(iii) pätevät, niin regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattoreilla  $b_0$  ja  $b_1$  on seuraavat odotusarvot ja varianssit:*

$$E(b_1) = \beta_1 \qquad \text{Var}(b_1) = D^2(b_1) = \frac{\sigma^2}{(n-1)s_x^2}$$

$$E(b_0) = \beta_0 \qquad \text{Var}(b_0) = D^2(b_0) = \frac{\sigma^2 \sum_{j=1}^n x_j^2}{n(n-1)s_x^2}$$

- Erityisesti: PNS-estimaattorit  $b_0$  ja  $b_1$  ovat *oletuksien (i)-(iii) pätiessä harhattomia*.

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Regressiokertoimien estimaattorit: Otosjakaumat

---

- *Jos standardioletuksien (i)-(iii) lisäksi normaalisuusoletus (iv) pätee, regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit  $b_0$  ja  $b_1$  ovat normaalijakautuneita:*

$$b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right)$$

$$b_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{j=1}^n x_j^2}{n(n-1)s_x^2}\right)$$

## Regressiosuoran kulmakertoimen luottamusväli

---

- *Jos standardioletuksien (i)-(iii) lisäksi normaalisuusoletus (iv) pätee*, niin regressiokertoimen  $\beta_1$  eli regressiosuoran kulmakertoimen **luottamusväli** luottamustasolla  $(1 - \alpha)$  on muotoa

$$b_1 \pm t_{\alpha/2} \frac{s}{\sqrt{n-1} s_x}$$

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  $\sigma^2$  harhaton estimaattori.



## Regressiosuoran kulmakertoimen luottamusväli: Kommentti

---

- Huomaa, että regressiokertoimen  $\beta_1$  luottamusväli on *tavanomaista muotoa*

$$b_1 \pm t_{\alpha/2} \hat{D}(b_1)$$

jossa

$$\hat{D}^2(b_1) = \frac{s^2}{(n-1)s_x^2}$$

on kertoimen  $\beta_1$  PNS-estimaattorin  $b_1$  *varianssin estimaattori*.

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Regressiosuoran vakion luottamusväli

---

- *Jos standardioletuksien (i)-(iii) lisäksi normaalisuusoletus (iv) pätee*, niin regressiokertoimen  $\beta_0$  eli regressiosuoran vakion **luottamusväli** luottamustasolla  $(1 - \alpha)$  on muotoa

$$b_0 \pm t_{\alpha/2} \frac{s \sqrt{\sum_{j=1}^n x_j^2}}{\sqrt{n(n-1)} s_x}$$

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 jäännösvariانسsin  $\sigma^2$  harhaton estimaattori.

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Regressiosuoran vakion luottamusväli:

### Kommentti

---

- Huomaa, että regressiokertoimen  $\beta_0$  luottamusväli on *tavanomaista muotoa*

$$b_0 \pm t_{\alpha/2} \hat{D}(b_0)$$

jossa

$$\hat{D}^2(b_0) = \frac{s^2 \sum_{j=1}^n x_j^2}{n(n-1) s_x^2}$$

on kertoimen  $\beta_0$  PNS-estimaattorin  $b_0$  *varianssin estimaattori*.

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Testi regressiosuoran kulmakertoimelle

---

- Oletetaan, että *standardioletuksien* (i)-(iii) lisäksi *normaalisuusoletus* (iv) pätee.

- Olkoon *nollahypoteesina*

$$H_{01} : \beta_1 = \beta_1^0$$

- Määritellään ***t*-testisuure**

$$t_1 = \frac{b_1 - \beta_1^0}{s / (\sqrt{n-1} s_x)}$$

- *Jos nollahypoteesi*  $H_{01}$  *pätee,*

$$t_1 \sim t(n-2)$$

- Itseisarvoltaan *suuret* testisuureen  $t_1$  arvot viittaavat siihen, että nollahypoteesi  $H_{01}$  *ei päde.*

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Testi regressiosuoran kulmakertoimelle:

### Kommentti

---

- Huomaa, että  $t$ -testisuure nollahypoteesille  $H_{01} : \beta_1 = \beta_1^0$  on *tavanomaista muotoa*

$$t_1 = \frac{b_1 - \beta_1^0}{\hat{D}(b_1)}$$

jossa

$$\hat{D}^2(b_1) = \frac{s^2}{(n-1)s_x^2}$$

on regressiokertoimen  $\beta_1$  PNS-estimaattorin  $b_1$  *varianssin estimaattori*, kun nollahypoteesi  $H_{01}$  pätee.

# Päätely yhden selittäjän lineaarisesta regressiomallista

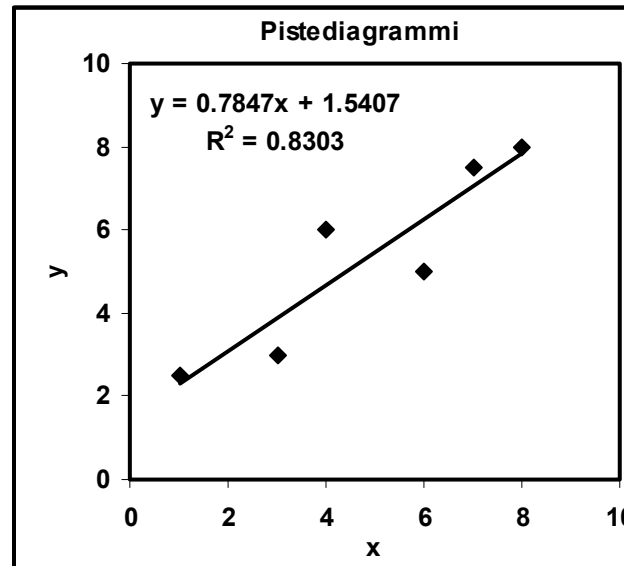
## Testi regressiosuoran kulmakertoimelle:

### Havainnollistava esimerkki 1/5

---

- Taulukossa oikealla on keinotekoisesti 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



Päätely yhden selittäjän lineaarisesta regressiomallista  
**Testi regressiosuoran kulmakertoimelle:**  
**Havainnollistava esimerkki 2/5**

---

- Alla olevassa taulukossa on laskettu havaintoarvojen summat ja neliösummat 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 <sup>2</sup>
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
<b>Summa</b>	<b>29</b>	<b>32</b>	<b>175</b>	<b>196.5</b>	<b>32</b>	<b>0.000</b>	<b>4.385</b>

- Tarkastellaan testiä mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*regressiokerrointa*  $\beta_1$  koskevalle nollahypoteesille

$$H_{01} : \beta_1 = 0$$

# Päätely yhden selittäjän lineaarisesta regressiomallista

## Testi regressiosuoran kulmakertoimelle:

### Havainnollistava esimerkki 3/5

---

- *Kertoimen  $\beta_1$  estimaatti:*

$$b_1 = 0.7847$$

- *Selittäjän  $x$  varianssi:*

$$s_x^2 = \frac{1}{n-1} \left( \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2 \right) = \frac{1}{6-1} \left( 175 - \frac{1}{6} \times 29^2 \right) = 6.967$$

- *Jäännösvarianssi:*

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

- *$t$ -testisuureen arvo:*

$$t_1 = \frac{b_1 - \beta_1^0}{s / (\sqrt{n-1} s_x)} = \frac{0.7847 - 0}{\sqrt{1.096 / ((6-1) \times 6.967)}} = 4.423$$



# Päätely yhden selittäjän lineaarisesta regressiomallista

## Testi regressiosuoran kulmakertoimelle:

### Havainnollistava esimerkki 4/5

---

- Jos nollahypoteesi  $H_{01} : \beta_1 = 0$  pätee, testisuure  $t_1$  on jakautunut *Studentin t-jakauman* mukaan vapausastein  $(n - 2) = (6 - 2) = 4$ :

$$t_1 \sim t(4)$$

- Valitaan *merkitsevyystasoksi* 0.05.
- Olkoon *vaihtoehtoinen hypoteesi* muotoa

$$H_{11} : \beta_1 \neq 0$$

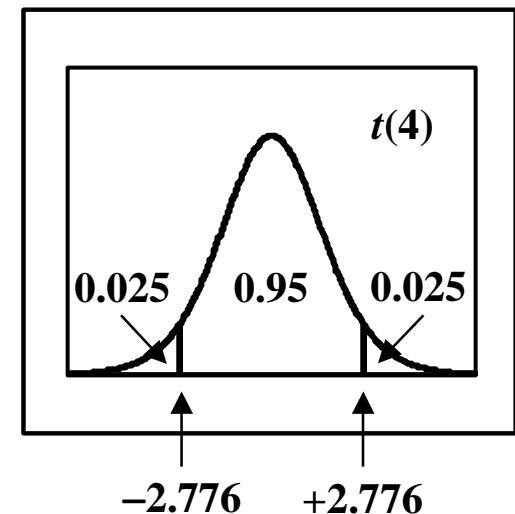
- Tällöin merkitsevyystasoa 0.05 vastaavat *kriittiset rajat* ovat

$$-2.776 \text{ ja } +2.776$$

ks. kuviota oikealla.

- Siten testin *hylkäysalue* on muotoa

$$\{t_1 \mid t_1 < -2.776\} \cup \{t_1 \mid t_1 > +2.776\}$$



# Päätely yhden selittäjän lineaarisesta regressiomallista

## Testi regressiosuoran kulmakertoimelle:

### Havainnollistava esimerkki 5/5

---

- Koska

$$t_1 = 4.423 > 2.776$$

niin testisuureen  $t_1$  arvo on hylkäysalueella ja *voimme hylätä nollahypoteesin*

$$H_{01} : \beta_1 = 0$$

ja *hyväksyä vaihtoehtoisen hypoteesin*

$$H_{11} : \beta_1 \neq 0$$

merkitsevyystasolla 0.05.

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Testi regressiosuoran vakiolle

---

- Oletetaan, että *standardioletuksien* (i)-(iii) lisäksi *normaalisuusoletus* (iv) pätee.

- Olkoon *nollahypoteesina*

$$H_{00} : \beta_0 = \beta_0^0$$

- Määritellään ***t*-testisuure**

$$t_0 = \frac{b_0 - \beta_0^0}{s \sqrt{\sum x_j^2} / (\sqrt{n(n-1)} s_x)}$$

- Jos *nollahypoteesi*  $H_{00}$  pätee,

$$t_0 \sim t(n-2)$$

- Itseisarvoltaan *suuret* testisuureen  $t_0$  arvot viittaavat siihen, että *nollahypoteesi*  $H_{00}$  *ei päde*.

## Testi regressiosuoran vakiolle:

### Kommentti

---

- Huomaa, että  $t$ -testisuure nollahypoteesille  $H_{00} : \beta_0 = \beta_0^0$  on *tavanomaista muotoa*

$$t_0 = \frac{b_0 - \beta_0^0}{\hat{D}(b_0)}$$

jossa

$$\hat{D}^2(b_0) = \frac{s^2 \sum_{j=1}^n x_j^2}{n(n-1) s_x^2}$$

on regressiokertoimen  $\beta_0$  PNS-estimaattorin  $b_0$  *varianssin estimaattori*, kun nollahypoteesi  $H_{00}$  pätee.

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Testi selityksasteelle 1/4

---

- Oletetaan, että *standardioletuksien* (i)-(iii) lisäksi *normaalisuusoletus* (iv) pätee.

- Olkoon *nollahypoteesina*

$$H_{01} : \beta_1 = 0$$

- Määritellään ***F*-testisuure**

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

jossa  $R^2$  on estimoidun mallin *selityksaste*.

## Päätely yhden selittäjän lineaarisesta regressiomallista

# Testi selitysteelle 2/4

---

- *Jos nollahypoteesi*

$$H_{01} : \beta_1 = 0$$

*pätee*, testisuure

$$F = (n - 2) \frac{R^2}{1 - R^2} \sim F(1, n - 2)$$

jossa  $F(1, n - 2)$  on Fisherin *F-jakauma* vapausastein 1 ja  $(n - 2)$ .

- *Suuret testisuureen  $F$  arvot viittaavat siihen, että nollahypoteesi  $H_{01}$  ei päde.*

## Päätely yhden selittäjän lineaarisesta regressiomallista

### Testi selitysteelle 3/4

---

- Koska  $R^2 = r_{xy}^2$ , em.  $F$ -testisuure voidaan esittää muodossa

$$F = (n - 2) \frac{r_{xy}^2}{1 - r_{xy}^2}$$

- Ottamalla tästä neliöjuuri saadaan testisuure

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

joka noudattaa *nollahypoteesin*  $H_{01}$  *pätiessä* Studentin  $t$ -*jakaumaa* vapausastein  $(n - 2)$ :

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

- Itseisarvoltaan *suuret* testisuureen  $t$  arvot viittaavat siihen, että *nollahypoteesi*  $H_{01}$  *ei päde*.

## Päätely yhden selittäjän lineaarisesta regressiomallista

### Testi selityskasteelle 4/4

---

- Voidaan osoittaa, että

$$t = \sqrt{n-2} \frac{r_{xy}}{\sqrt{1-r_{xy}^2}} = \frac{b_1}{s / \sqrt{n-1} s_x} = t_1$$

jossa testisuure  $t_1$  on tavanomainen  $t$ -testisuure nollahypoteesille

$$H_{01} : \beta_1 = 0$$

- $F$ - ja  $t$ -jakaumien yhteyden perusteella on selvää, että

$$t_1^2 = F$$

jossa  $F$  on em.  $F$ -testisuure nollahypoteesille  $H_{01}$ .

- Huomaa, että yllä esitetty  $t$ -testisuure ja  $t$ -testisuure *korreloimattomuudelle* ovat ekvivalentteja.



# Yhden selittäjän lineaarinen regressiomalli

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**Yhden selittäjän lineaarisen regressiomallin estimointi**

**Varianssianalyysihajotelma ja selityssaste**

**Päättely yhden selittäjän lineaarisesta regressiomallista**

**>> Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**2-ulotteisen normaalijakauman regressiofunktioiden estimointi**

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

---

## *Avainsanat*

**Ennustaminen**

**Ennuste**

**Lineaarinen regressiomalli**

**Luottamusväli**

**Otosjakauma**

**Pienimmän neliösumman estimaattori**

**Selitettävän muuttujan arvo**

**Selitettävän muuttujan odotusarvo**

**Standardioletukset**

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Ennustaminen

---

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

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

- 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**.

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Malli ja sen osat 1/2

---

- Oletetaan, että havaintoarvojen  $y_j$  ja  $x_j$  välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

$y_j$  = **selitettävän muuttujan**  $y$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$x_j$  = **selittävän muuttujan** eli **selittäjän**  $x$  *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$\varepsilon_j$  = **jäännös-** eli **virhetermin**  $\varepsilon$  *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä  $j$

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Malli ja sen osat 2/2

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat *kertoimet*:

$\beta_0$  = **vakioselittäjän regressiokerroin**;

$\beta_0$  on *ei-satunnainen ja tuntematon vakio*

$\beta_1$  = **selittäjän  $x$  regressiokerroin**;

$\beta_1$  on *ei-satunnainen ja tuntematon vakio*

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Oletukset

---

- Oletetaan, että yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös-* eli *virhetermiä*  $\varepsilon_j$  koskevat *standardioletukset* pätevät:

(i)  $E(\varepsilon_j) = 0, j = 1, 2, \dots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, \quad j \neq l$$

- Lisäksi oletetaan, että virhetermit  $\varepsilon_j$  ovat *normaalisia*:

(iv)  $\varepsilon_j \sim N(0, \sigma^2), j = 1, 2, \dots, n$

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Otostunnusluvut

---

- Määritellään havaintojen  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$$

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

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Regressiokertoimien PNS-estimaattorit

---

- Yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  **pienimmän neliösumman (PNS-) estimaattorit** ovat

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

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



# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## Sovitteet ja residuaalit

---

- Olkoot  $b_0$  ja  $b_1$  yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Määritellään estimoidun mallin **sovitteet** kaavalla

$$\hat{y}_j = b_0 + b_1 x_j, \quad j = 1, 2, \dots, n$$

- Määritellään estimoidun mallin **residuaalit** kaavalla

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_j, \quad j = 1, 2, \dots, n$$

## Ennustaminen yhden selittäjän lineaarisella regressiomallilla

# Jäännösvarianssin estimointi

---

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

$$s^2 = \frac{1}{n-2} \sum_{j=1}^n e_j^2$$

jossa

$e_j$  = estimoidun mallin *residuaali*

$n$  = havaintojen lukumäärä

## Ennustaminen yhden selittäjän lineaarisella regressiomallilla

### **$y$ :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*

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

kun selittäjä  $x$  saa arvon  $\tilde{x}$  ?

- Selitettävän muuttujan  $y$  ehdollinen odotusarvo  $E(\tilde{y}|\tilde{x})$  kuvaa *selitettävän muuttujan  $y$  keskimäärin saamia arvoja selittäjän  $x$  saamien arvojen funktiona.*

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **y:n odotusarvon ennustaminen:**

### **Ennuste**

---

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

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

jossa  $b_0$  ja  $b_1$  ovat regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Voidaan osoittaa, että  $\hat{y}|\tilde{x}$  on (ennustevirheen keskineliövirheen mielessä) *paras lineaarinen ja harhaton ennuste* ehdolliselle odotusarvolle  $E(\tilde{y}|\tilde{x})$ .
- Huomautus:  
Ehdollinen odotusarvo  $E(\tilde{y}|\tilde{x})$  on kiinteälle  $\tilde{x}$  vakio, kun taas ennuste  $\hat{y}|\tilde{x}$  on *satunnaismuuttuja*.

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **y:n odotusarvon ennustaminen:**

### **Otosjakauma**

---

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

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

**otosjakauma** on normaalijakauma:

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

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **$y$ :n odotusarvon ennustaminen:**

### **Luottamusväli**

---

- Odotusarvon

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

**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  $\sigma^2$  harhaton estimaattori.

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

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## $y$ :n odotusarvon ennustaminen:

### Luottamusvälin ominaisuuksia

---

- Odotusarvon

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

luottamusväli

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

*kaventuu*, jos havaintojen lukumäärä  $n$  tai selittäjän otosvarianssi  $s_x^2$  kasvaa.

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

## Ennustaminen yhden selittäjän lineaarisella regressiomallilla

### **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  $\tilde{x}$  ?



# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **y:n arvon ennustaminen:**

### **Ennuste**

---

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

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

jossa  $b_0$  ja  $b_1$  ovat regressiokertoimien  $\beta_0$  ja  $\beta_1$  PNS-estimaattorit.

- Voidaan osoittaa, että  $\hat{y}|\tilde{x}$  on (ennustevirheen keskineliövirheen mielessä) *paras lineaarinen ja harhaton ennuste* ehdolliselle odotusarvolle  $E(\tilde{y}|\tilde{x})$ .
- Huomautus:  
Sekä selitettävän muuttujan  $y$  arvo  $\tilde{y}$  että ennuste  $\hat{y}|\tilde{x}$  ovat *satunnaismuuttujia*.

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **y:n arvon ennustaminen:**

### **Otosjakauma**

---

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

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

**otosjakauma** on normaalijakauma:

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

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **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  $\sigma^2$  harhaton estimaattori.

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

# Ennustaminen yhden selittäjän lineaarisella regressiomallilla

## **y:n arvon ennustaminen:**

### **Luottamusvälin ominaisuuksia**

---

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

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

*kaventuu*, jos havaintojen lukumäärä  $n$  tai selittäjän otosvarianssi  $s_x^2$  kasvaa.

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

## Ennustaminen yhden selittäjän lineaarisella regressiomallilla

### **$y$ :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  $E(\tilde{y}|\tilde{x})$  luottamusvyö.
- Tämä johtuu siitä, että selitettävän muuttujan  $y$  *keskimääräisen arvon ennustaminen on helpompaa kuin sen yksittäisen arvon ennustaminen.*

# Yhden selittäjän lineaarinen regressiomalli

---

Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

Yhden selittäjän lineaarisen regressiomallin estimointi

Varianssianalyysihajotelma ja selityssaste

Päättely yhden selittäjän lineaarisesta regressiomallista

Ennustaminen yhden selittäjän lineaarisella regressiomallilla

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

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

---

## *Avainsanat*

Ehdollinen odotusarvo  
Harhattomuus  
Lineaarinen regressiomalli  
Lineaarisuus  
Modifioidut standardioletukset  
Pienimmän neliösumman menetelmä  
Regressiofunktio  
Regressiosuora  
Satunnaisuus  
Selittäjä  
Selittävä muuttuja  
Standardioletukset  
Tarkentuvuus

## Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

# 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*.



# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Havainnot

---

- Olkoot

$$y_1, y_2, \dots, y_n$$

selitettävän muuttujan  $y$  ja

$$x_1, x_2, \dots, x_n$$

selittävän muuttujan  $x$  **havaittuja arvoja**.

- Oletetaan lisäksi, että havaintoarvot  $x_j$  ja  $y_j$  liittyvät *samaan havaintoyksikköön kaikille  $j = 1, 2, \dots, n$* .
- Tällöin havaintoarvot  $x_j$  ja  $y_j$  muodostavat pisteitä 2-ulotteisessa avaruudessa:

$$(x_j, y_j) \in \mathbb{R}^2, \quad j = 1, 2, \dots, n$$

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Malli ja sen osat 1/2

---

- Oletetaan, että havaintojen  $y_j$  ja  $x_j$  välillä on *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

$y_j$  = **selitettävän muuttujan**  $y$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$x_j$  = **selittävän muuttujan**  $x$  *satunnainen* ja *havaittu* arvo havaintoyksikössä  $j$

$\varepsilon_j$  = **jäännös-** eli **virhetermin**  $\varepsilon$  *satunnainen* ja *ei-havaittu* arvo havaintoyksikössä  $j$

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Malli ja sen osat 2/2

---

- Yhden selittäjän lineaarisessa regressiomallissa

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat kertoimet:

$\beta_0$  = **vakioselittäjän regressiokerroin;**

$\beta_0$  on *ei-satunnainen ja tuntematon vakio*

$\beta_1$  = **selittäjän  $x$  regressiokerroin;**

$\beta_1$  on *ei-satunnainen ja tuntematon vakio*

- Huomautus:

Regressiokertoimet  $\beta_0$  ja  $\beta_1$  oletetaan *samoiksi* kaikille havaintoyksiköille  $j$ .

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Selittäjän satunnaisuuden seuraukset 1/4

---

- Yhden selittäjän lineaarisen mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

## Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

# Selittäjän satunnaisuuden seuraukset 2/4

---

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

Näin käy esimerkiksi silloin, kun virhetermi ja selittäjä *korreloivat*.

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

## Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

# Selittäjän satunnaisuuden seuraukset 3/4

---

- 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*  $\varepsilon_j$  toteuttavat kiinteälle selittäjälle esitetyt standardioletukset *ehdollisesti selittäjän  $x$  havaittujen arvojen suhteen*.**

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Selittäjän satunnaisuuden seuraukset 4/4

---

- Tässä kappaleessa tarkastellaan lähemmin yhden selittäjän lineaarisen regressiomallin määrittelemistä sellaisella tavalla, joka takaa sen, että kiinteälle selittäjälle esitetty teoria pätee.

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Modifioidut oletukset jäännöstermeistä

---

- Oletetaan, että mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös-* eli *virhetermit*  $\varepsilon_j$  toteuttavat seuraavat oletukset:

(i)  $E(\varepsilon_j | x_j) = 0, \quad j = 1, 2, \dots, n$

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

$$\text{Var}(\varepsilon_j | x_j) = \sigma^2, \quad j = 1, 2, \dots, n$$

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

$$\text{Cor}(\varepsilon_j, \varepsilon_l | x_j, x_l) = 0, \quad j \neq l$$

- Lisäksi jäännöstermeistä  $\varepsilon_j$  tehdään tavallisesti *normaalisuusoletus*:

(iv)  $\varepsilon_j | x_j \sim N(0, \sigma^2), \quad j = 1, 2, \dots, n$



## Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

# Mallin selitettävän muuttujan ominaisuudet

---

- Jos yhden selittäjän lineaarisen regressiomallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös- eli virhetermejä  $\varepsilon_j$  koskevat oletukset (i)-(iii) pätevät, mallin selitettävän muuttujan  $y$  havaituilla arvoilla  $y_j$  on seuraavat stokastiset ominaisuudet:*

(i)'  $E(y_j | x_j) = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \dots, n$

(ii)'  $\text{Var}(y_j | x_j) = \sigma^2, \quad j = 1, 2, \dots, n$

(iii)'  $\text{Cor}(y_j, y_l | x_j, x_l) = 0, \quad j \neq l$

- Jos jäännöstermejä  $\varepsilon_j$  koskeva *normaalisuusoletus (iv) pätee*, niin

(iv)'  $y_j | x_j \sim N(\beta_0 + \beta_1 x_j, \sigma^2), \quad j = 1, 2, \dots, n$

## Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

# Mallin selitettävän muuttujan ominaisuudet:

## Kommentti

---

- Jos muuttujan  $y$  arvojen  $y_j$  stokastiset ominaisuudet (i)'-(iv)' otetaan oletuksiksi, ne määrittelevät täsmälleen saman tilastollisen mallin kuin mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jäännös- eli virhetermeistä  $\varepsilon_j$  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.

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Selitettävän muuttujan ehdollisen odotusarvon tulkinta regressiofunktiona

---

- Oletuksen

$$(i)' \quad E(y_j | x_j) = \beta_0 + \beta_1 x_j, \quad j = 1, 2, \dots, n$$

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  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  noudattavat 2-ulotteista normaalijakaumaa, oletus (i)' pätee.

# Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Yhteys kiinteän selittäjän tapaukseen

---

- Oletetaan, että mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

selittävän muuttujan  $x$  arvot  $x_j$  ovat kiinteitä eli ei-satunnaisia ja mallin jäännös- eli virhetermit  $\varepsilon_j$  toteuttavat standardioletukset

- (i)  $E(\varepsilon_j) = 0, \quad j = 1, 2, \dots, n$
  - (ii)  $\text{Var}(\varepsilon_j) = \sigma^2, \quad j = 1, 2, \dots, n$
  - (iii)  $\text{Cor}(\varepsilon_j, \varepsilon_l) = 0, \quad j \neq l$
  - (iv)  $\varepsilon_j \sim N(0, \sigma_\varepsilon^2), \quad j = 1, 2, \dots, n$
- Tällöin edellä satunnaisen selittäjän tapauksessa tehdyt oletukset mallin virhetermistä pätevät triviaalisti.

Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

## Selittäjän satunnaisuuden seuraukset:

### Kommentteja

---

- **Tässä kappaleessa esitetyt modifioidutkin 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 yleensä saa 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

---

**Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset**

**Yhden selittäjän lineaarisen regressiomallin estimointi**

**Varianssianalyysihajotelma ja selityssaste**

**Päättely yhden selittäjän lineaarisesta regressiomallista**

**Ennustaminen yhden selittäjän lineaarisella regressiomallilla**

**Yhden selittäjän lineaarisen regressiomalli ja satunnainen selittäjä**

**>> 2-ulotteisen normaalijakauman regressiofunktioiden estimointi**

# 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

---

## *Avainsanat*

Ehdollinen odotusarvo  
Kaksiulotteinen normaalijakauma  
Lineaarinen regressiomalli  
Lineaarisuus  
Otostunnusluvut  
Painopiste  
Pienimmän neliösumman menetelmä  
Regressiofunktio  
Regressiosuora  
Satunnaisuus  
Selittäjä  
Selittävä muuttuja  
Standardioletukset

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

### Oletukset

---

- Oletetaan, että toisistaan *riippumattomat* havaintoparit

$$(x_j, y_j), j = 1, 2, \dots, n$$

noudattavat **2-ulotteista normaalijakaumaa**; ks. kalvokokoelman **Johdatus todennäköisyyslaskentaan** lukua **Moniulotteisia jakaumia**.

- Tällöin *ehdolliset odotusarvot* ovat muotoa

$$E(x_j | y_j) = \alpha_0 + \alpha_1 y_j, j = 1, 2, \dots, n$$

$$E(y_j | x_j) = \beta_0 + \beta_1 x_j, j = 1, 2, \dots, n$$

ja siis *lineaarisia*.



## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiomalleja on kaksi

---

- Voimme kirjoittaa

$$x_j = \alpha_0 + \alpha_1 y_j + \delta_j, \quad j = 1, 2, \dots, n$$

ja

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jossa jäännöstermit  $\varepsilon_j$  ja  $\delta_j$  ovat *keskenään korreloimattomia* satunnaismuuttujia.

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Mallien jäännöstermit 1/2

---

- Mallin

$$x_j = \alpha_0 + \alpha_1 y_j + \delta_j, \quad j = 1, 2, \dots, n$$

*jäännös-* eli *virhetermit*  $\delta_j$  toteuttavat seuraavat ehdot:

(i)  $E(\delta_j | y_j) = 0, \quad j = 1, 2, \dots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\text{Var}(\delta_j | y_j) = \sigma_\delta^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\delta_j, \delta_l | y_j, y_l) = 0, \quad j \neq l$$

(iv) Jäännöstermit ovat *normaalisia*:

$$\delta_j | y_j \sim N(0, \sigma_\delta^2), \quad j = 1, 2, \dots, n$$

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Mallien jäännöstermit 2/2

---

- Mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

*jäännös-* eli *virhetermit*  $\varepsilon_j$  toteuttavat seuraavat ehdot:

(i)  $E(\varepsilon_j | x_j) = 0, \quad j = 1, 2, \dots, n$

(ii) Jäännöstermit ovat *homoskedastisia*:

$$\text{Var}(\varepsilon_j | x_j) = \sigma_\varepsilon^2, \quad j = 1, 2, \dots, n$$

(iii) Jäännöstermit ovat *korreloimattomia*:

$$\text{Cor}(\varepsilon_j, \varepsilon_l | x_j, x_l) = 0, \quad j \neq l$$

(iv) Jäännöstermit ovat *normaalisia*:

$$\varepsilon_j | x_j \sim N(0, \sigma_\varepsilon^2), \quad j = 1, 2, \dots, n$$

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Otostunnusluvut

---

- Määritellään havaintojen  $x_j$  ja  $y_j$ ,  $j = 1, 2, \dots, n$  *aritmeettiset keskiarvot*, *otosvarianssit*, *otoskovarianssi* ja *otoskorrelaatiokerroin* tavanomaisilla kaavoillaan:

$$\bar{x} = \frac{1}{n} \sum_{j=1}^n x_j$$

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j$$

$$s_x^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y})^2$$

$$s_{xy} = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})$$

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

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Parametrien PNS-estimaattorit

---

- Mallin

$$x_j = \alpha_0 + \alpha_1 y_j + \delta_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\alpha_1$  ja  $\alpha_0$  PNS-estimaattorit ovat

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

- Mallin

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \dots, n$$

regressiokertoimien  $\beta_1$  ja  $\beta_0$  PNS-estimaattorit ovat

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

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

### 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

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

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

### Estimoidut regressiosuorat 2/3

---

- *Molemmat* estimoidut regressiosuorat voidaan esittää muuttujan  $x$  funktiona:

- (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:

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

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

### 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.
  - (2)'  $r_{xy} = 0$
- Lisäksi yhtälöistä nähdään, että *suorat leikkaavat havaintojen painopisteessä*  $(\bar{x}, \bar{y})$ .



## 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

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

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

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

*saadaan muodollisesti korvaamalla regressiofunktion parametrit  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$ ,  $\sigma_y^2$ ,  $\rho_{xy}$  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

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

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

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

*saadaan muodollisesti korvaamalla regressiofunktion parametrit  $\mu_x$ ,  $\mu_y$ ,  $\sigma_x^2$ ,  $\sigma_y^2$ ,  $\rho_{xy}$  vastaavilla otossuureilla.*

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# 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

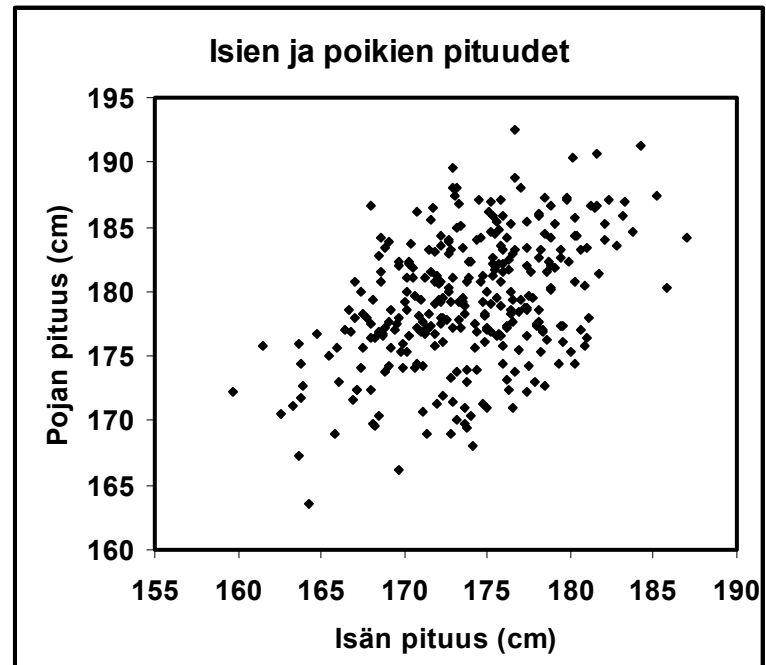
$$(x_j, y_j), j = 1, 2, \dots, 300$$

jossa

$x_j$  = isän  $j$  pituus

$y_j$  = isän  $j$  pojan pituus

- Ks. pistediagrammia oikealla.



## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 2/8

---

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

$$M_k(x|x) \text{ ja } M_k(y|x)$$

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

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

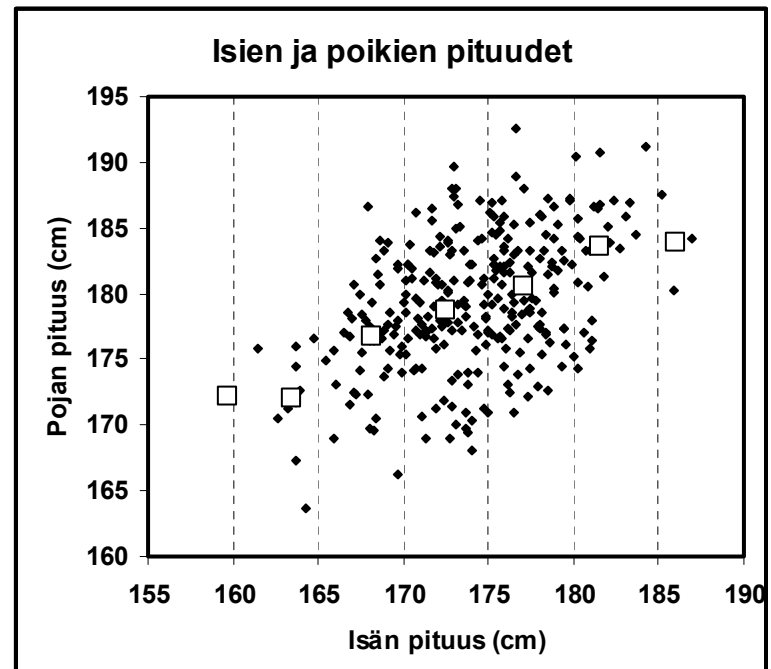
### Esimerkki 3/8

- *Ehdollisten keskiarvojen*

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

määäämiä pisteitä on merkitty kuviossa oikealla *neliöillä*.

- Havainnot on siis luokiteltu *isien* pituuden mukaan 7 luokkaan.
- Kuviossa luokkia on kuvattu katkoviivojen erottamalla pystyvöillä.
- Jokaisen *neliön koordinaatit* on saatu laskemalla keskiarvot ko. neliötä vastaavaan pystyvööhön kuuluvien havaintopisteiden koordinaateista.



## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 4/8

- Olkoon mallina

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

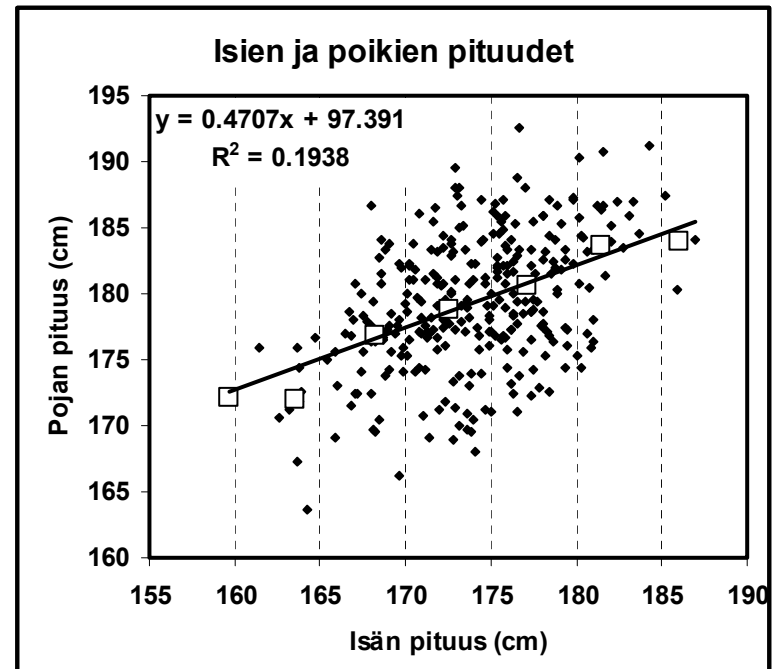
$$j = 1, 2, \dots, n$$

- Alkuperäisistä havainnoista *estimoidun regressiosuoran* yhtälö on

$$y = 97.391 + 0.4707x$$

- Selitysaste on

$$R^2 = 0.194$$



## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 5/8

---

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

$$M_k(x|y) \text{ ja } M_k(y|y)$$

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  $k$

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

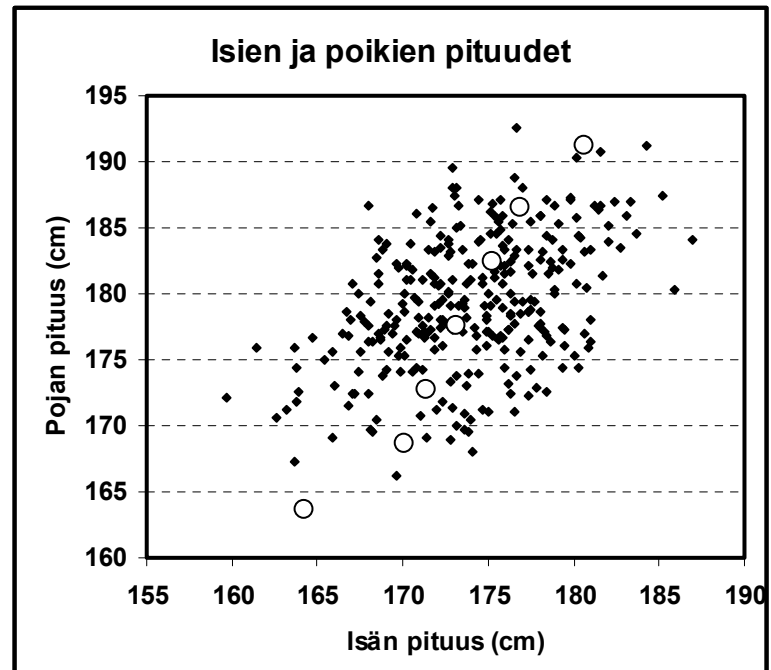
y-välin nro	y-väli	$M_k(x y)$	$M_k(y y)$
1	(160,165]	164.3	163.6
2	(165,170]	170.1	168.7
3	(170,175]	171.4	172.7
4	(175,180]	173.1	177.6
5	(180,185]	175.2	182.4
6	(185,190]	176.9	186.6
7	(190,195]	180.6	191.2

## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 6/8

- *Ehdollisten keskiarvojen*  
 $M_k(x|y)$  ja  $M_k(y|x)$   
määäämiä pisteitä on merkitty  
kuviossa oikealla *ympyröillä*.
- Havainnot on siis luokiteltu *poikien*  
pituuden mukaan 7 luokkaan.
- Kuviossa luokkia on kuvattu  
katkoviivojen erottamalla  
*vaakavoilla*.
- Jokaisen *ympyrän koordinaatit*  
on saatu laskemalla keskiarvot ko.  
ympyrää vastaavaan *vaakavyöhön*  
kuuluvien havaintopisteiden  
koordinaateista.





## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 7/8

- Olkoon mallina

$$x_j = \alpha_0 + \alpha_1 x_j + \delta_j$$
$$j = 1, 2, \dots, n$$

- Alkuperäisistä havainnoista *estimoidun regressiosuoran* yhtälö on

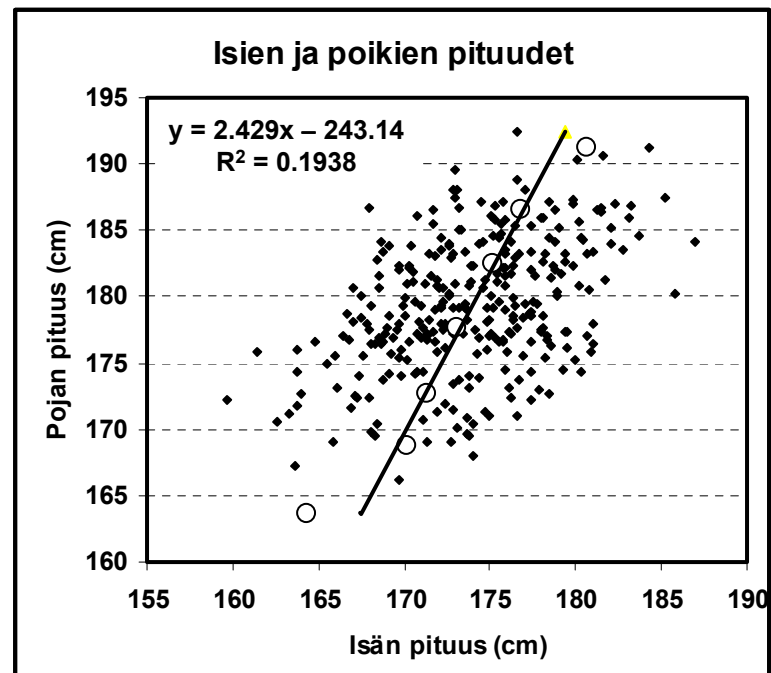
$$x = 100.10 + 0.4117 y$$

joka voidaan  $x$ :n funktiona kirjoittaa muotoon

$$y = -243.14 + 2.429x$$

- Selitysaste on

$$R^2 = 0.194$$



## 2-ulotteisen normaalijakauman regressiofunktioiden estimointi

# Regressiosuorien estimointi:

### Esimerkki 8/8

- Kuvioon oikealla on lisätty malleja

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$$

$$x_j = \alpha_0 + \alpha_1 x_j + \delta_j$$

vastaavat *estimoidut regressiosuorat*.

- Muuttujan  $y$  regressiosuora muuttujan  $x$  suhteen:

$$y = 97.391 + 0.4707x$$

- Muuttujan  $x$  regressiosuora muuttujan  $y$  suhteen muuttujan  $x$  funktiona:

$$y = -243.14 + 2.429x$$

