Ilkka Mellin

Tilastolliset menetelmät

Osa 4: Lineaarinen regressioanalyysi

Johdatus regressioanalyysiin

Johdatus regressioanalyysiin

>> **Regressioanalyysin lähtökohdat ja tavoitteet**

Deterministiset mallit ja regressioanalyysi

Regressiofunktiot ja regressioanalyysi

Kaksiulotteisen normaalijakauman regressiofunktiot

Regressioanalyysin tehtävät

Regressiomallin lineaarisuus

Regressioanalyysin idea 1/2

- Oletetaan, että haluamme **selittää** jonkin **selitettävän tekijän** tai **muuttujan** *havaittujen arvojen vaihtelun* joidenkin **selittävien tekijöiden** tai **muuttujien** *havaittujen arvojen vaihtelun avulla*.
- Jos *tilastollisesti merkitsevä osa* selitettävän muuttujan havaittujen arvojen vaihtelusta *voidaan selittää* selittävien muuttujien havaittujen arvojen vaihtelun avulla, sanomme, että selitettävä muuttuja **riippuu tilastollisesti** selittäjinä käytetyistä muuttujista.

Regressioanalyysin idea 2/2

- **Regressioanalyysissa** selitettävän muuttujan *tilastolliselle riippuvuudelle* selittävistä muuttujista *pyritään rakentamaan tilastollinen malli*, jota kutsutaan **regressiomalliksi**.
- Koska *riippuvuuksien analysointi* on tavallisesti tieteellisen tutkimuksen keskeinen tavoite, **regressio-analyysi on ehkä eniten sovellettu ja tärkein tilasto-tieteen menetelmä**.

Regressioanalyysin tavoitteet

- Regressioanalyysin mahdollisia *tavoitteita*:
  (i) Selitettävän muuttujan ja selittävien muuttujien tilastollisen riippuvuuden luonteen **kuvaaminen**:
    – Millainen on riippuvuuden *muoto*?
    – Kuinka *voimakasta* riippuvuus on?
  (ii) Selitettävän muuttujan ja selittävien muuttujien tilastollisen riippuvuuden luonteen **selittäminen**.
  (iii) Selitettävän muuttujan arvojen **ennustaminen**.
  (iv) Selitettävän muuttujan arvojen **kontrolli**.

Regressiomallien luokittelu 1/2

- Regressioanalyysissa sovellettavat tilastolliset mallit voidaan *luokitella* usealla eri periaatteella.
- Luokittelu regressiomallin *funktionaalisen muodon* mukaan:
  - **Lineaariset regressiomallit**
  - **Epälineaariset regressiomallit**
- Luokittelu regressiomallin *yhtälöiden lukumäärän* mukaan:
  - **Yhden yhtälön regressiomallit**
  - **Moniyhtälömallit**

Regressiomallien luokittelu 2/2

- Tässä johdatuksessa tilastotieteeseen käsitellään pääasiassa **lineaarisia yhden yhtälön regressiomalleja**; ks. lukuja **Yhden selittäjän lineaarinen regressiomalli** ja **Yleinen lineaarinen malli.**
- On hyödyllistä tietää, että **varianssianalyysin** tilastolliset mallit voidaan ymmärtää *yleisen lineaarisen mallin* erikoistapauksiksi.

Regressioanalyysin sovellukset tilastotieteessä

- Regressiomalleja käytetään *apuvälineinä* monilla tilastotieteen osa-alueilla.
- Esimerkkejä regressiomallien käyttökohteista tilastotieteessä:
  - **Varianssianalyysi**
  - **Koesuunnittelu**
  - **Monimuuttujamenetelmät**
  - **Kalibrointi**
  - **Biometria** tai **-statistiikka**
  - **Aikasarjojen analyysi** ja **ennustaminen**
  - **Ekonometria**

Regressioanalyysin lähtökohdat

- Regressioanalyysilla on kaksi erilaista *lähtökohtaa*, joilla on kuitenkin monia yhtymäkohtia:
  (i) Ongelmat **deterministisen mallien** sovittamisessa havaintoihin; ks. kappaletta **Deterministiset mallit ja regressioanalyysi.**
  (ii) Moniulotteisten todennäköisyysjakaumien **ehdollisten odotusarvojen** eli **regressiofunktioiden** parametrien estimointi; ks. kappaletta **Regressiofunktiot ja regressioanalyysi.**

Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet

>> Deterministiset mallit ja regressioanalyysi

Regressiofunktiot ja regressioanalyysi

Kaksiulotteisen normaalijakauman regressiofunktiot

Regressioanalyysin tehtävät

Regressiomallin lineaarisuus

Deterministiset mallit regressio-analyysin lähtökohtana 1/2

- Oletetaan, että haluamme **selittää** jonkin **selitettävän tekijän** tai **muuttujan** käyttäytymisen joidenkin **selittävien tekijöiden** tai **muuttujien** avulla.
- Oletetaan, että sekä selitettävä muuttuja että selittäjät ovat *ei-satunnaisia* muuttujia.
- Tällöin tavoitteeseen voidaan pyrkiä kuvaamalla *selitettävän muuttujan arvojen riippuvuus selittävien muuttujien arvoista* **deterministisen mallin** *avulla*.

Deterministiset mallit regressio-analyysin lähtökohtana 2/2

- Oletetaan, että selitettävän muuttujan riippuvuutta selittävistä muuttujista kuvaavan *deterministisen mallin muoto riippuu* tuntemattomasta **parametrista** (vakiosta).
- Tällöin parametrin arvo voidaan pyrkiä **estimoimaan** eli *arvioimaan havaintojen avulla*.
- Oletetaan, että **parametrille ei ole mahdollista löytää sellaista arvoa**, **joka saisi mallin sopimaan *saman-aikaisesti* kaikkiin havaintoihin**.
- **Voidaanko parametrille löytää kuitenkin sellainen arvo, joka saisi mallin sopimaan havaintoihin jossakin mielessä niin hyvin kuin se on mahdollista**?

Deterministiset mallit

- Oletetaan, että selitettävän muuttujan *y eksaktia* (*kausaalista*) *riippuvuutta* selittäjästä *x halutaan mallintaa yhtälöllä*

  $$y = f(x ; \beta)$$

  jossa funktion *f* muoto riippuu *parametrista* eli *vakiosta* β.
- Yhtälö määrittelee **deterministisen mallin** selitettävän muuttujan *y* ja selittäjän *x* riippuvuudelle:

  Jos selittäjän *x* ja parametrin β arvot *tunnetaan*, niin selitettävän muuttujan *y* arvo on *täysin määrätty*.

Deterministiset mallit ja regressio-ongelma 1/4

- Oletetaan, että selitettävän muuttujan *y riippuvuutta* selittäjästä *x halutaan mallintaa deterministisellä yhtälöllä*

$$y = f(x\,;\beta)$$

- Oletetaan, että funktion *f* muodon määräävän parametrin β arvo on *tuntematon*.
- Haluamme löytää parametrille β *parhaan mahdollisen havaintoihin perustuvan estimaatin* eli *arvion*.
- **Regressio-ongelma** syntyy deterministisen mallien soveltamisen yhteydessä tilanteissa, joissa *parametrille* β *ei voida löytää sellaista arvoa*, *joka saisi ym. yhtälön toteutumaan samanaikaisesti kaikille havainnoille*.

Deterministiset mallit ja regressio-ongelma 2/4

- Oletetaan, että muuttujia x ja y koskevat havainnot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$.
- Oletetaan, että ei ole olemassa *yhtä* parametrin β arvoa, joka saa yhtälön

  $$y = f(x\,;\beta)$$

  toteutumaan *samanaikaisesti kaikille havainnoille* x_i ja y_i .
- Kirjoitetaan

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

  jossa ε_i on havaintoyksiköstä toiseen *vaihteleva jäännös-* eli *virhetermi*.

Deterministiset mallit ja regressio-ongelma 3/4

- Oletetaan, että *jäännös*- eli *virhetermit* ε_i yhtälössä

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

  vaihtelevat satunnaisesti yhtälöstä toiseen.

  Huomaa, että oletuksesta seuraa, että *selitettävän muuttujan y havaittujen arvot* y_i *ovat satunnaisia*.

- Yhtälö

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

  kuvaa selitettävän muuttujan *y* **tilastollista riippuvuutta** selittävän muuttujan *x* saamista arvoista.

- Sanomme, että yhtälö määrittelee selitettävän muuttujan *y* **regressiomallin** selittävän muuttujan *x* suhteen.

Deterministiset mallit ja regressio-ongelma 4/4

- **Regressioanalyysissa** parametrin β arvo pyritään valitsemaan tavalla, joka tekee *kaikista jäännöstermeistä* ε_i *samanaikaisesti mahdollisimman pieniä*.
- Tämä on *käyränsovitusongelma*:

  Miten parametrin β arvo pitää valita, jotta käyrä

  $$y = f(x\,;\beta)$$

  kulkisi jossakin mielessä *mahdollisimman läheltä jokaista havaintopistettä*

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

- Erään ratkaisun tähän käyränsovitusongelmaan tarjoaa *pienimmän neliösumman menetelmä*.

Deterministiset mallit ja regressio-ongelma: Esimerkki 1/4

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

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

  jossa

  α = jousen pituus ilman painoa

  β = ns. *jousivakio*

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

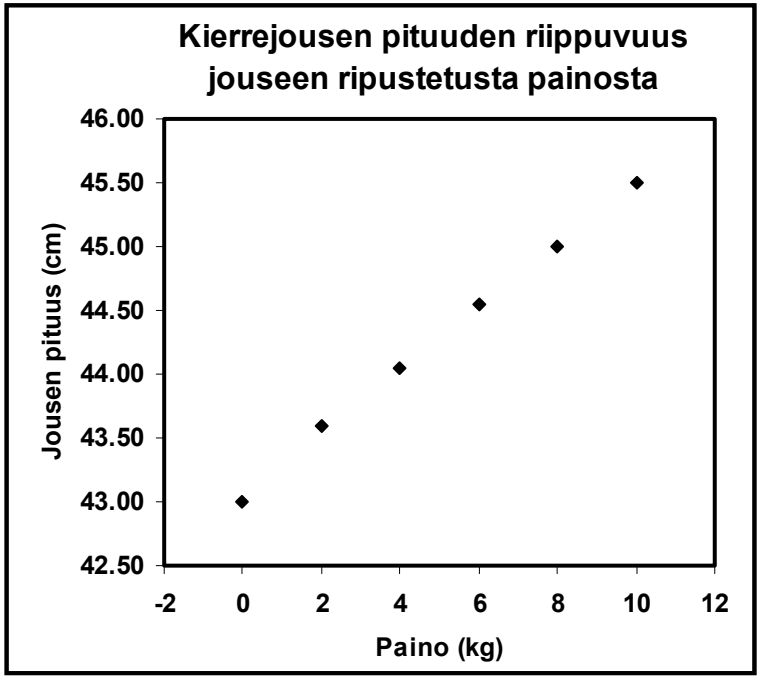
Paino (kg)	Pituus (cm)
0	43.00
2	43.60
4	44.05
6	44.55
8	45.00
10	45.50

Deterministiset mallit ja regressio-ongelma: Esimerkki 2/4

- Pistediagrammi oikealla havainnollistaa koetuloksia.
- Kysymys 1:

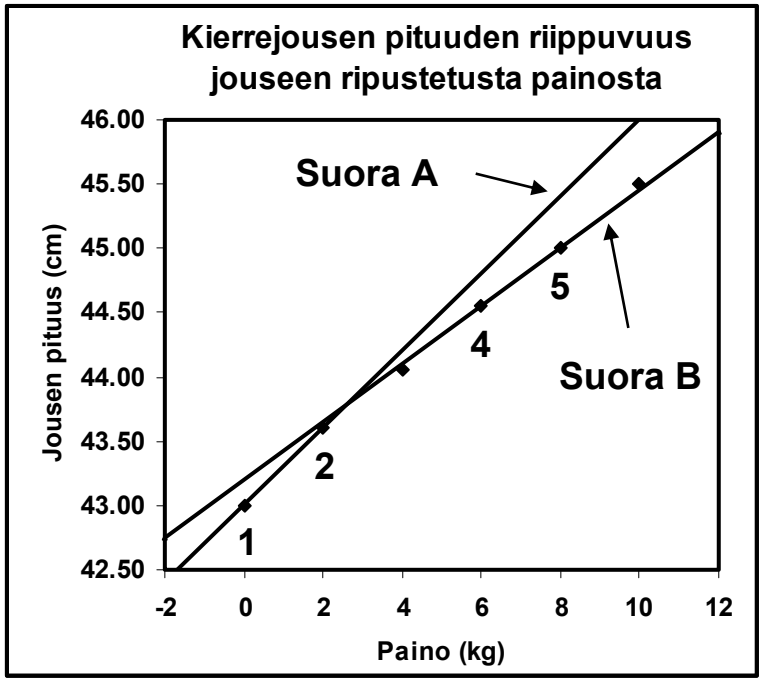
  Ovatko havaintotulokset *sopusoinnussa* Hooken lain kanssa?
- Kysymys 2:

  Onko olemassa *yksikäsitteinen* suora, joka kulkee *kaikkien* havaintopisteiden kautta?



Deterministiset mallit ja regressio-ongelma: Esimerkki 3/4

- Kuvio oikealla todistaa, että ei ole olemassa *yhtä suoraa*, joka kulkisi kaikkien havainto-pisteiden kautta:
  (i) Suora A kulkee pisteiden 1 ja 2 kautta.
  (ii) Suora B kulkee pisteiden 4 ja 5 kautta.
- Onko mahdollista määrätä yksikäsitteisellä tavalla suora, joka kulkisi jossakin mielessä *mahdollisimman läheltä* jokaista havaintopistettä?



Deterministiset mallit ja regressio-ongelma: Esimerkki 4/4

- Käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

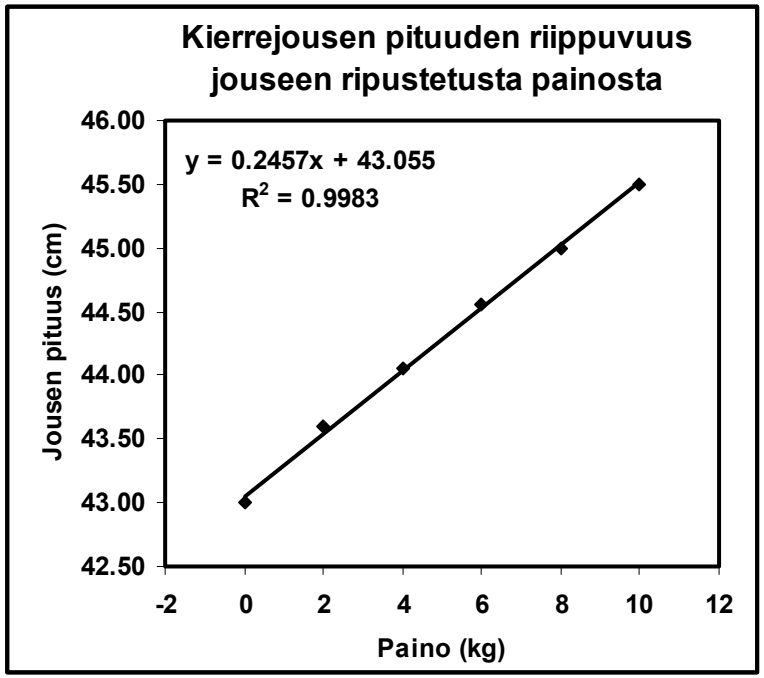
$$y = \alpha + \beta x$$

kertoimet niin, että neliösumma

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

minimoituu.

- Kuvioon oikealla on piirretty näin määrätty suora; ks. tarkemmin lukua **Yhden selittäjän lineaarinen regressiomalli**.



Syyt regressio-ongelman syntymiseen

- Mitkä *syyt* johtavat regressio-ongelman syntymiseen deterministisen mallien yhteydessä?
- Syitä regressio-ongelman syntymiseen:

  (i) *Havaintovirheet* selitettävän muuttujan y havaituissa arvoissa.

  (ii) *Yhtälö*

  $$y = f(x\,;\beta)$$

  on idealisointi:

  Osaa selitettävän muuttujan y käyttäytymiseen vaikuttavista tekijöistä *ei haluta* tai *ei pystytä* ottamaan huomioon.

Regressiomalli ja kiinteät selittäjät 1/2

- Olkoon

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

selitettävän muuttujan *y tilastollista riippuvuutta* selittävän muuttujan *x* saamista arvoista kuvaava *regressiomalli*.

- Oletukset:
  - (i) Selittävän muuttujan *x* arvot x_i voidaan *valita*, jolloin ne ovat *kiinteitä* eli *ei-satunnaisia*.
  - (ii) Jäännös- eli virhetermit ε_i ovat *satunnaisia*, jolloin myös selitettävän muuttujan *y* havaitut arvot y_i pitää olettaa satunnaisiksi.

Regressiomalli ja kiinteät selittäjät 2/2

- Regressiomallissa

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

on seuraavat osat:

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

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

β = *tuntematon* ja *kiinteä* eli *ei-satunnainen* **parametri** (vakiokerroin)

ε_i = *satunnainen* ja *ei-havaittu* **jäännös-** eli **virhetermi** havaintoyksikössä *i*

Regressiomallit ja kiinteät selittäjät: Kommentteja

- Kun regressiomalleja sovelletaan *luonnontieteissä* tai *tekniikassa*, oletus selittävien muuttujien ei-satunnaisuudesta on usein hyvin perusteltu.

  Tämä johtuu siitä, että monissa luonnontieteiden tai tekniikan sovelluksissa regressiomallien *selittäjien arvot voidaan valita* eli selittäjät ovat muuttujia, joiden *arvoja voidaan kontrolloida*.

  Esimerkki: *Puhtaat koeasetelmat*.

- Monissa tilastotieteen sovelluksissa kohdataan kuitenkin sellaisia tilanteita, joissa ainakin osa selittäjistä on sellaisia, joiden arvot määräytyvät *satunnaisesti*; ks. kappaletta **Regressiofunktiot ja regressioanalyysi.**

Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet
Deterministiset mallit ja regressioanalyysi
>> Regressiofunktiot ja regressioanalyysi
Kaksiulotteisen normaalijakauman regressiofunktiot
Regressioanalyysin tehtävät
Regressiomallin lineaarisuus

Regressiofunktiot regressio-ongelman lähtökohtana 1/2

- Oletetaan, että haluamme **selittää** jonkin **selitettävän tekijän** tai **muuttujan** käyttäytymisen joidenkin **selittävien tekijöiden** tai **muuttujien** avulla.
- Oletetaan, että sekä selitettävä muuttuja että selittäjät ovat *satunnaismuuttujia*.
- Tällöin tavoitteeseen voidaan pyrkiä kuvaamalla *selitettävän muuttujan riippuvuutta selittävistä muuttujista* selitettävän muuttujan **regressiofunktiolla** selittäjien suhteen.

Regressiofunktiot regressio-ongelman lähtökohtana 2/2

- Oletetaan, että selitettävän muuttujan riippuvuutta selittävistä muuttujista kuvaavan *regressiofunktion muoto* riippuu tuntemattomasta **parametrista** (vakiosta).
- Tällöin parametrin arvo voidaan pyrkiä **estimoimaan** eli *arvioimaan havaintojen avulla*.
- **Miten parametrille löydetään jossakin mielessä mahdollisimman hyvä estimaatti eli arvio**?

Ehdollinen jakauma

- Olkoon $f_{xy}(x, y)$ satunnaismuuttujien x ja y **yhteis-jakauman** tiheysfunktio.
- Olkoot $f_x(x)$ ja $f_y(y)$ satunnaismuuttujien x ja y **reuna-jakaumien** tiheysfunktiot.
- Satunnaismuuttujan y **ehdollisen jakauman** tiheysfunktio satunnaismuuttujan x suhteen on

$$f_{y|x}(y \mid x) = \frac{f_{xy}(x, y)}{f_x(x)}, \text{ jos } f_x(x) > 0$$

Ehdollinen odotusarvo

- Satunnaismuuttujan y **ehdollinen odotusarvo** satunnaismuuttujan x suhteen on

$$\mathrm{E}(y \mid x) = \int_{-\infty}^{+\infty} y f_{y|x}(y \mid x) dy$$

jossa

$$f_{y|x}(y \mid x)$$

on satunnaismuuttujan y *ehdollisen jakauman* tiheysfunktio satunnaismuuttujan x suhteen

- Huomaa, että ehdollinen odotusarvo on ehtomuuttujan x funktiona *satunnaismuuttuja*.

Regressiofunktio 1/2

- Tarkastellaan satunnaismuuttujan *y ehdollista odotusarvoa ehtomuuttujan x arvojen funktiona*.
- Ehdollista odotusarvoa

  $$\mathrm{E}(y \mid x)$$

  kutsutaan ehtomuuttujan *x* arvojen funktiona *satunnaismuuttujan y* **regressiofunktioksi** *muuttujan x suhteen*.
- Regressiofunktion $\mathrm{E}(y \mid x)$ muoto riippuu satunnaismuuttujan *y* ehdollisen jakauman

  $$f_{y|x}(y \mid x)$$

  parametreista.

Regressiofunktio 2/2

- Olkoon

$$\mathrm{E}(y \mid x)$$

satunnaismuuttujan *y regressiofunktio* satunnaismuuttujan *x* suhteen.

- Koska haluamme korostaa regressiofunktion arvojen *riippuvuutta ehtomuuttujan x arvoista*, kirjoitamme

$$\mathrm{E}(y \mid x) = f(x\,;\beta)$$

jossa β on satunnaismuuttujan *y* ehdollisen jakauman

$$f_{y|x}(y \mid x)$$

muodon määräävä *parametri*.

Lisätietoja

- Lisätietoja **moniulotteisista satunnaismuuttujista** ja niiden **yhteisjakaumista**, **reunajakaumista**, **ehdollisista jakaumista**, **ehdollisista odotusarvoista** ja **regressio-funktioista**:

  Ks. monisteen **Todennäköisyyslaskenta** lukua **Moniulotteiset satunnaismuuttujat ja jakaumat.**

Regressiofunktio ja ennustaminen 1/3

- Olkoon $f_{xy}(x, y)$ satunnaismuuttujien x ja y *yhteisjakauman tiheysfunktio*.
- Oletetaan, että satunnaismuuttujan x arvo *tunnetaan*.
- Kysymys:

  Miten tietoa satunnaismuuttujan x saamasta arvosta voidaan käyttää hyväksi satunnaismuuttujan y arvon *ennustamisessa*?
- Olkoon $d(y \mid x)$ muuttujan x saamaan arvoon perustuva **ennuste** muuttujan y arvolle.
- Miten ennuste $d(y \mid x)$ valitaan *optimaalisella tavalla*?

Regressiofunktio ja ennustaminen 2/3

- Valitaan ennuste $d(y \mid x)$ siten, että *ennusteen* **keskineliövirhe**

$$\mathrm{MSE}[d(y \mid x)] = \mathrm{E}[y - d(y \mid x)]^2$$

minimoituu.

- Voidaan osoittaa, että keskineliövirhe $\mathrm{MSE}(d(y \mid x))$ minimoituu valinnalla

$$d(y \mid x) = \mathrm{E}(y \mid x)$$

- Siten satunnaismuuttujan y *regressiofunktio* $\mathrm{E}(y \mid x)$ *satunnaismuuttujan x suhteen* tuottaa muuttujan x saamiin arvoihin perustuvat, *keskineliövirheen mielessä optimaaliset ennusteet* muuttujalle y.

Regressiofunktio ja ennustaminen 3/3

- Olkoon

$$y - \mathrm{E}(y \mid x) = \varepsilon$$

optimaalisen ennusteen $\mathrm{E}(y \mid x)$ **ennustevirhe**.

- Tällöin voimme kirjoittaa

$$y = \mathrm{E}(y \mid x) + \varepsilon$$
$$= f(x;\beta) + \varepsilon$$

jossa

$$\mathrm{E}(y \mid x) = f(x;\beta)$$

on satunnaismuuttujan y *regressiofunktio* satunnaismuuttujan x suhteen.

Regressiofunktio regressiomallina

- Edellisen nojalla muuttujan x arvoihin perustuva optimaalinen *ennuste* satunnaismuuttujan y arvolle määrittelee **regressiomallin**

$$y = \mathrm{E}(y \mid x) + \varepsilon$$
$$= f(x;\beta) + \varepsilon$$

jossa y on mallin **selitettävä muuttuja** ja x on mallin **selittävä muuttuja**.

Regressiofunktiot ja regressio-ongelma 1/3

- Oletetaan, että selitettävän muuttujan *y riippuvuutta* selittäjästä *x halutaan mallintaa regressiofunktiolla*

$$\mathrm{E}(y \mid x) = f(x\,;\beta)$$

- Oletetaan, että regressiofunktion *f* muodon määräävän parametrin β arvo on *tuntematon*.
- Parametrille β halutaan löytää *paras mahdollinen estimaatti* eli *arvio havaintojen perusteella*.
- **Regressio-ongelmalla** tarkoittaa tässä *regressiofunktion muodon määräävän parametrin β valintaongelmaa*.

Regressiofunktiot ja regressio-ongelma 2/3

- Oletetaan, että satunnaismuuttujia x ja y koskevat havainnot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$.
- Edellä esitetyn nojalla voimme kirjoittaa yhtälön

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

jossa ε_i on havaintoyksiköstä toiseen *satunnaisesti vaihteleva jäännös-* eli *virhetermi*.

- Yhtälö kuvaa muuttujan y **tilastollista riippuvuutta** muuttujan x saamista arvoista.
- Sanomme, että yhtälö määrittelee selitettävän muuttujan y **regressiomallin** selittävän muuttujan x suhteen.

Regressiofunktiot ja regressio-ongelma 3/3

- **Regressioanalyysissa** parametrin β arvo pyritään valitsemaan sellaisella tavalla, joka tekee *kaikista* jäännös-termeistä ε_i *samanaikaisesti mahdollisimman pieniä*.
- Tämä on *käyränsovitusongelma*:

  Miten parametrin β arvo on valittava niin, että käyrä

  $$y = f(x\,;\beta)$$

  kulkisi *mahdollisimman läheltä jokaista havaintopistettä*

  $$(x_i, y_i) \in \mathbb{R}^2\,,\, i = 1,2,\ldots,n\,?$$
- Erään ratkaisun tähän käyränsovitusongelmaan tarjoaa *pienimmän neliösumman menetelmä*.

Mitä regressiofunktio mallintaa? Esimerkki 1/6

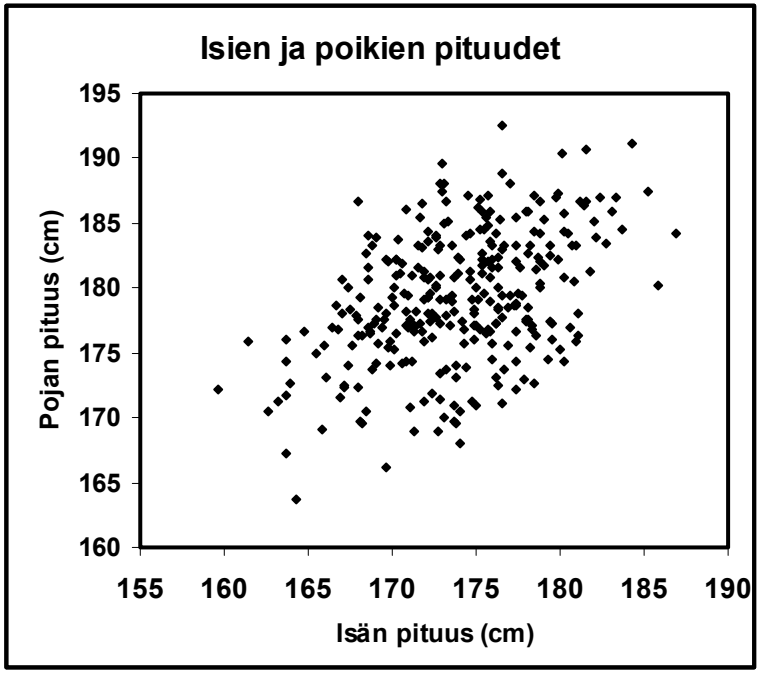
- Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.
- Periytyykö isän pituus heidän pojilleen?
- Havaintoaineisto koostuu 300:n isän ja heidän poikiensa pituuksien muodostamasta lukuparista

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

  jossa

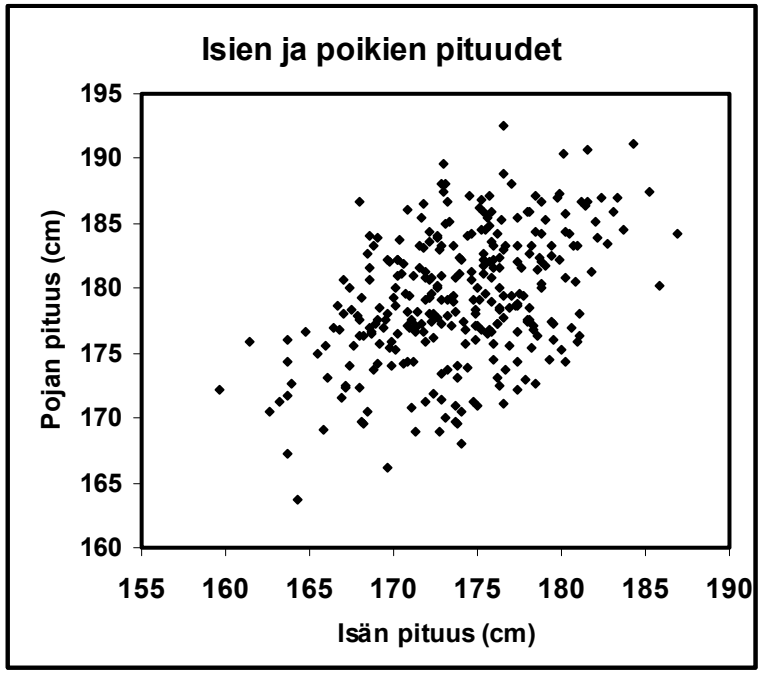
  x_i = isän i pituus

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



Mitä regressiofunktio mallintaa? Esimerkki 2/6

- Pojan pituuden riippuvuus isän pituudesta ei selvästikään ole *eksaktia*.
- Mutta: Lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkillä isillä ja pitkillä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.
- Miten tällaista *tilastollista riippuvuutta* voidaan havainnollistaa?



Mitä regressiofunktio mallintaa? Esimerkki 3/6

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

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

  jossa

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

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

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

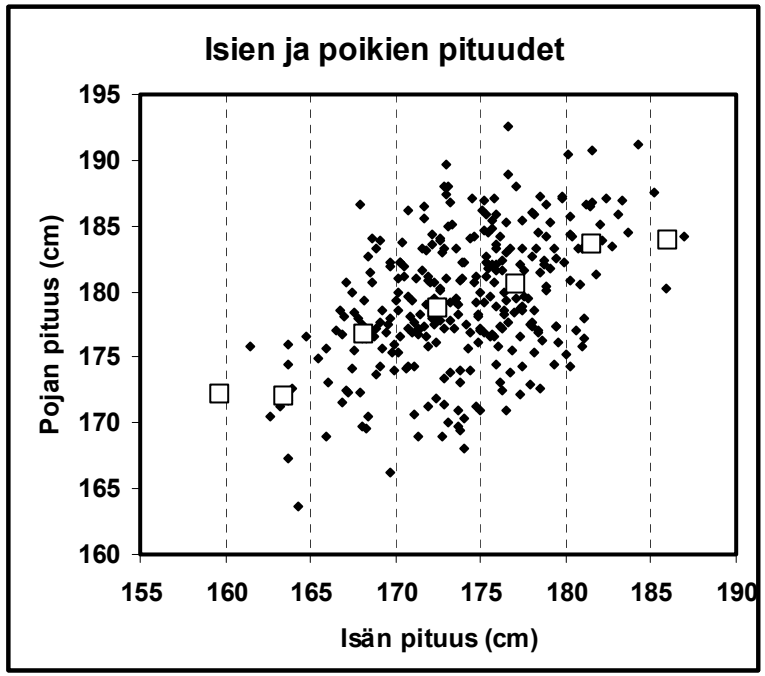
x-välin nro	x-väli	$M_k(x\|x)$	$M_k(y\|x)$
1	(155,160]	159.7	172.2
2	(160,165]	163.5	172.0
3	(165,170]	168.2	176.8
4	(170,175]	172.6	178.8
5	(175,180]	177.1	180.6
6	(180,185]	181.5	183.6
7	(185,190]	186.0	184.0

Mitä regressiofunktio mallintaa? Esimerkki 4/6

- *Ehdollisten keskiarvojen*

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

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

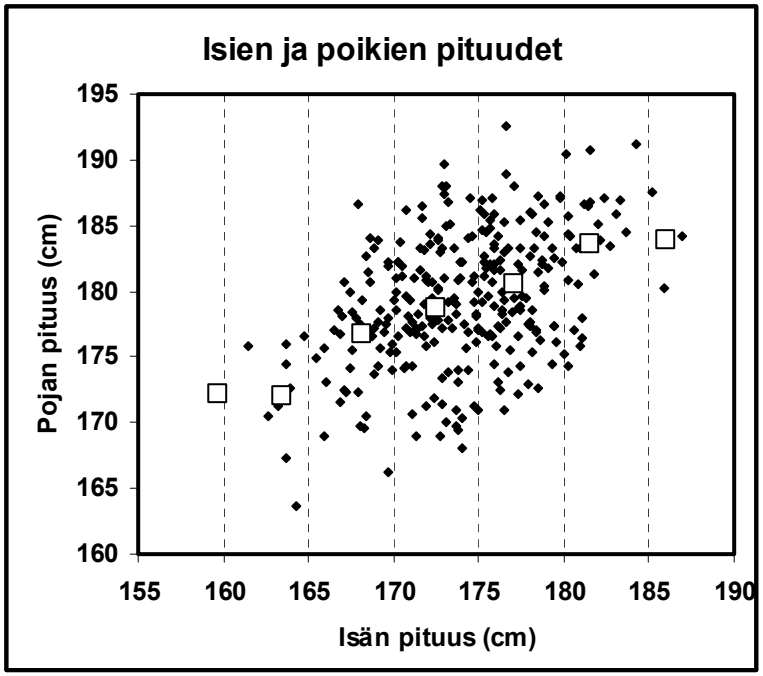


Mitä regressiofunktio mallintaa? Esimerkki 5/6

- Oikealla olevaan kuvioon neliöillä merkityt *ehdollisten keskiarvojen* määräämät pisteet

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

  kuvaavat poikien pituuksien *keskimääräistä* tai *tilastollista riippuvuutta* heidän isiensä pituuksista.
- Riippuvuus näyttää olevan lähes *lineaarista*.
- *Regressioanalyysin tehtävänä* on juuri tällaisen *tilastollisen riippuvuuden mallintaminen*.



Mitä regressiofunktio mallintaa? Esimerkki 6/6

- Käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

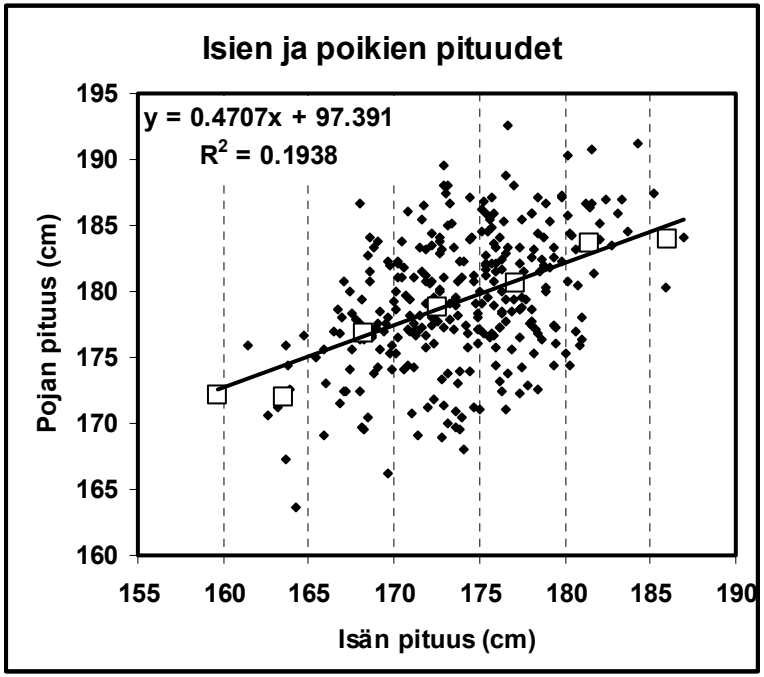
$$y = \alpha + \beta x$$

kertoimet niin, että neliösumma

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

minimoituu.

- Kuvioon oikealla on piirretty näin määrätty suora; ks. tarkemmin lukua **Yhden selittäjän lineaarinen regressiomalli**.



Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet
Deterministiset mallit ja regressioanalyysi
Regressiofunktiot ja regressioanalyysi
>> Kaksiulotteisen normaalijakauman regressiofunktiot
Regressioanalyysin tehtävät
Regressiomallin lineaarisuus

Multinormaalijakauma

- Normaalijakauman yleistystä moniulotteiseen avaruuteen kutsutaan **multinormaalijakaumaksi** tai *moniulotteiseksi normaalijakaumaksi*.
- Multinormaalijakauman määräävät täydellisesti jakaumaan liittyvien satunnaismuuttujien *odotusarvot*, *varianssit* ja *korrelaatiot*.
- Multinormaalijakauma näyttelee *lineaaristen regressiomallien* teoriassa keskeistä osaa, koska **multinormaalijakauman kaikki regressiofunktiot ovat lineaarisia**.
- Seuraavassa tarkastellaan lähemmin **2-ulotteista normaalijakaumaa**; lisätietoja: ks. monisteen **Todennäköisyyslaskenta** lukua **Moniulotteisia jakaumia.**

2-ulotteinen normaalijakauma: Tiheysfunktio 1/2

- **2-ulotteisen normaalijakauman** *tiheysfunktio* on

$$f_{xy}(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}} \exp\left\{-\frac{1}{2(1-\rho_{xy}^2)}Q(x,y)\right\}$$

jossa

$$Q(x,y) = \left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho_{xy}\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right) + \left(\frac{y-\mu_y}{\sigma_y}\right)^2$$

ja

$$-\infty < \mu_x < +\infty\ , -\infty < \mu_y < +\infty$$

$$\sigma_x > 0, \sigma_y > 0$$

$$-1 \le \rho_{xy} \le +1$$

2-ulotteinen normaalijakauma: Tiheysfunktio 2/2

- 2-ulotteisen normaalijakauman *parametreina* ovat satunnaismuuttujien x ja y *odotusarvot*, *varianssit* ja *korrelaatio*:

$$\mu_x = \mathrm{E}(x) = \text{muuttujan } x \textit{ odotusarvo}$$
$$\mu_y = \mathrm{E}(y) = \text{muuttujan } y \textit{ odotusarvo}$$
$$\sigma_x^2 = \mathrm{Var}(x) = \text{muuttujan } x \textit{ varianssi}$$
$$\sigma_y^2 = \mathrm{Var}(y) = \text{muuttujan } y \textit{ varianssi}$$
$$\rho_{xy} = \mathrm{Cor}(x, y) = \text{muuttujien } x \text{ ja } y \textit{ korrelaatio}$$

2-ulotteinen normaalijakauma: Jakauman parametrit

- Oletetaan, että satunnaismuuttujien x ja y muodostama pari (x, y) noudattaa 2-*ulotteista normaalijakaumaa*.
- Koska satunnaismuuttujien x ja y *odotusarvot*, *varianssit* ja *korrelaatio*

$$\mathrm{E}(x) = \mu_x \qquad \mathrm{E}(y) = \mu_y$$

$$\mathrm{Var}(x) = \sigma_x^2 \qquad \mathrm{Var}(y) = \sigma_y^2$$

$$\mathrm{Cor}(x, y) = \rho_{xy}$$

määräävät täydellisesti 2-ulotteisen normaalijakauman, merkitään

$$(x, y) \sim \mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

2-ulotteinen normaalijakauma: Parametrien tulkinta 1/2

- Oletetaan, että satunnaismuuttujien x ja y muodostama pari (x, y) noudattaa 2-*ulotteista normaalijakaumaa*.
- Satunnaismuuttujien x ja y *odotusarvot*

  $$\mathrm{E}(x) = \mu_x \qquad \mathrm{E}(y) = \mu_y$$

  määräävät satunnaismuuttujien x ja y yhteisjakauman *todennäköisyysmassan painopisteen*.
- Satunnaismuuttujien x ja y *varianssit*

  $$\mathrm{Var}(x) = \sigma_x^2 \qquad \mathrm{Var}(y) = \sigma_y^2$$

  kuvaavat satunnaismuuttujien x ja y *todennäköisyys-massojen hajaantuneisuutta* niiden odotusarvojen μ_x ja μ_y ympärillä.

2-ulotteinen normaalijakauma: Parametrien tulkinta 2/2

- Satunnaismuuttujien x ja y *korrelaatio*

  $$\mathrm{Cor}(x, y) = \rho_{xy}$$

  kuvaa satunnaismuuttujien x ja y *lineaarisen riippuvuuden voimakkuutta*.
- Koska pari (x, y) noudattaa 2-ulotteista normaalijakaumaa, satunnaismuuttujat x ja y *ovat korreloimattomia*, *jos ja vain jos ne ovat riippumattomia*.
- Yleisesti pätee:

  $$\mathrm{Cor}(x, y) = \pm 1$$

  jos ja vain jos on olemassa vakiot α ja $\beta \neq 0$ siten, että

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

2-ulotteinen normaalijakauma: Ehdolliset jakaumat 1/2

- 2-ulotteisen normaalijakauman **ehdolliset jakaumat** ovat *normaalisia*.
- *Satunnaismuuttujan y ehdollinen jakauma satunnais-muuttujan x suhteen* on

$$y \mid x \sim \mathrm{N}\left(\mu_{y|x}, \sigma^2_{y|x}\right)$$

jossa

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

$$\sigma^2_{y|x} = \mathrm{Var}(y \mid x) = (1 - \rho^2_{xy})\sigma^2_y$$

2-ulotteinen normaalijakauma: Ehdolliset jakaumat 2/2

- 2-ulotteisen normaalijakauman **ehdolliset jakaumat** ovat *normaalisia*.
- *Satunnaismuuttujan x ehdollinen jakauma satunnais-muuttujan y suhteen* on

$$x \mid y \sim \mathrm{N}\left(\mu_{x|y}, \sigma^2_{x|y}\right)$$

jossa

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

$$\sigma^2_{x|y} = \mathrm{Var}(x \mid y) = (1 - \rho^2_{xy})\sigma^2_x$$

2-ulotteinen normaalijakauma: Regressiofunktiot 1/2

- 2-ulotteisen normaalijakauman *regressiofunktiot* eli *ehdolliset odotusarvot* ovat *lineaarisia*.
- *Satunnaismuuttujan y* **regressiofunktio** *satunnais-muuttujan x suhteen*

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

määrittelee *xy*-koordinaatistossa *suoran*

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

- Suora kulkee satunnaismuuttujien *x* ja *y yhteisjakauman todennäköisyysmassan painopisteen* (μ_x, μ_y) kautta.

2-ulotteinen normaalijakauma: Regressiofunktiot 2/2

- 2-ulotteisen normaalijakauman *regressiofunktiot* eli *ehdolliset odotusarvot* ovat *lineaarisia*.
- *Satunnaismuuttujan x* **regressiofunktio** *satunnais-muuttujan y suhteen*

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

määrittelee *xy*-koordinaatistossa *suoran*

$$y = \mu_y + \frac{1}{\rho_{xy}} \times \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

- Suora kulkee satunnaismuuttujien *x* ja *y yhteisjakauman todennäköisyysmassan painopisteen* (μ_x, μ_y) kautta.

2-ulotteinen normaalijakauma: Regressiosuorat

- 2-ulotteisen normaalijakauman regressiofunktioiden määrittelemien **regressiosuorien** yhtälöistä

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

$$y = \mu_y + \frac{1}{\rho_{xy}} \times \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

nähdään seuraavaa:

(i) Jos $\rho_{xy} = 0$, suorat ovat *kohtisuorassa* toisiaan vastaan.

(ii) Jos $\rho_{xy} = \pm 1$, suorat *yhtyvät*.

2-ulotteinen normaalijakauma: Regressiosuorien ominaisuudet 1/2

- Muuttujan y regressiosuoralla muuttujan x suhteen

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

on seuraavat ominaisuudet:

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

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

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

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

- korrelaation itseisarvo $|\rho_{xy}|$ *kasvaa* (*pienenee*)
- standardipoikkeama σ_y *kasvaa* (*pienenee*)
- standardipoikkeama σ_x *pienenee* (*kasvaa*)

2-ulotteinen normaalijakauma: Regressiosuorien ominaisuudet 2/2

- Muuttujan x regressiosuoralla muuttujan y suhteen

$$y = \mu_y + \frac{1}{\rho_{xy}} \times \frac{\sigma_y}{\sigma_x}(x - \mu_x)$$

on seuraavat ominaisuudet:

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

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

(iii) Jos $\rho_{xy} = 0$, suora on *pystysuorassa*.

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

- korrelaation itseisarvo $|\rho_{xy}|$ *pienenee* (*kasvaa*)
- standardipoikkeama σ_y *kasvaa* (*pienenee*)
- standardipoikkeama σ_x *pienenee* (*kasvaa*)

2-ulotteinen normaalijakauma: Ehdolliset varianssit 1/2

- *Satunnaismuuttujan y* **ehdollinen varianssi** *satunnais-muuttujan x suhteen* on

$$\sigma^2_{y|x} = \mathrm{Var}(y \mid x) = (1 - \rho^2_{xy})\sigma^2_y$$

ja se kuvaa *satunnaismuuttujan y ehdollisen jakauman* (*satunnaismuuttujan x suhteen*) *todennäköisyysmassan hajaantuneisuutta regressiosuoran*

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

ympärillä.

2-ulotteinen normaalijakauma: Ehdolliset varianssit 2/2

- *Satunnaismuuttujan x* **ehdollinen varianssi** *satunnais-muuttujan y suhteen* on

$$\sigma^2_{x|y} = \mathrm{Var}(x \mid y) = (1 - \rho^2_{xy})\sigma^2_x$$

ja se kuvaa *satunnaismuuttujan x ehdollisen jakauman* (*satunnaismuuttujan y suhteen*) *todennäköisyysmassan hajaantuneisuutta regressiosuoran*

$$y = \mu_y + \frac{1}{\rho_{xy}} \times \frac{\sigma_x}{\sigma_y}(x - \mu_x)$$

ympärillä.

2-ulotteinen normaalijakauma: Ehdollisten varianssien ominaisuudet 1/2

- Satunnaismuuttujan y ehdollisella varianssilla satunnaismuuttujan x suhteen

$$\sigma^2_{y|x} = \mathrm{Var}(y \mid x) = (1-\rho^2_{xy})\sigma^2_y$$

on seuraavat ominaisuudet:

(i) $\sigma^2_{y|x} \leq \sigma^2_y$

(ii) Jos $\rho_{xy} = 0$, niin $\sigma^2_{y|x} = \sigma^2_y$.

(iii) Jos $\rho_{xy} = \pm 1$, niin $\sigma^2_{y|x} = 0$ ja satunnaismuuttujien x ja y *yhteisjakauman todennäköisyysmassa* keskittyy muuttujien x ja y yhteiselle regressiosuoralle.

(iv) Ehdollinen varianssi ei riipu ehtomuuttujasta.

2-ulotteinen normaalijakauma: Ehdollisten varianssien ominaisuudet 2/2

- Satunnaismuuttujan x ehdollisella varianssilla satunnaismuuttujan y suhteen

$$\sigma^2_{x|y} = \mathrm{Var}(x \mid y) = (1-\rho^2_{xy})\sigma^2_x$$

on seuraavat ominaisuudet:

(i) $\sigma^2_{x|y} \leq \sigma^2_x$

(ii) Jos $\rho_{xy} = 0$, niin $\sigma^2_{x|y} = \sigma^2_x$.

(iii) Jos $\rho_{xy} = \pm 1$, niin $\sigma^2_{x|y} = 0$ ja satunnaismuuttujien x ja y *yhteisjakauman todennäköisyysmassa* keskittyy muuttujien x ja y yhteiselle regressiosuoralle.

(iv) Ehdollinen varianssi ei riipu ehtomuuttujasta.

Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet
Deterministiset mallit ja regressioanalyysi
Regressiofunktiot ja regressioanalyysi
Kaksiulotteisen normaalijakauman regressiofunktiot
>> Regressioanalyysin tehtävät
Regressiomallin lineaarisuus

Regressiomalli ja sen osat 1/2

- *Yhden yhtälön* **regressiomallin** yleinen muoto on

$$y = f(x;\beta) + \varepsilon$$

jossa

y = **selitettävä muuttuja**

$f(x;\beta)$ = mallin **systemaattinen** eli **rakenneosa**

ε = mallin **satunnainen osa**

- Mallin *systemaattinen osa* $f(x;\beta)$ on **selittävän muuttujan** x funktio, joka riippuu funktion f muodon määräävästä **parametrista** β.
- Mallin *satunnainen osa* ε on **jäännöstermi**, joka tavallisesti *ei riipu* selittäjästä x.

Regressiomalli ja sen osat 2/2

- Regressiomallin

$$y = f(x;\beta) + \varepsilon$$

systemaattinen osa $f(x;\beta)$ *kuvaa selitettävän muuttujan y riippuvuutta* selittävästä *muuttujasta x*.

- Regressioanalyysissa *pääasiallinen kiinnostus* kohdistuu regressiomallin *systemaattiseen osaan* $f(x;\beta)$ *ja sen muotoon*.
- Regressiomallin *jäännöstermiä* ε pidetään usein pelkkänä *virheterminä*, mutta *jäännöstermistä* ε *tehdyt oletukset vaikuttavat ratkaisevalla tavalla siihen tapaan, jolla regressioanalyysi tehdään.*

Regressioanalyysi

- **Regressioanalyysi** tarkoittaa seuraavia malliin

  $$y = f(x;\beta) + \varepsilon$$

  liittyvien tehtävien suorittamista:
  - Funktion f **valinta**
  - Parametrin β **estimointi**
  - Parametria β koskevien **hypoteesien testaaminen**
  - Estimoidun mallin **hyvyyden arviointi**
  - Mallista tehtyjen **oletusten tarkistaminen**
  - Selitettävän muuttujan käyttäytymisen **ennustaminen** ja **ennusteiden epävarmuuden arviointi**

Johdatus regressioanalyysiin

- Regressioanalyysin lähtökohdat ja tavoitteet
- Deterministiset mallit ja regressioanalyysi
- Regressiofunktiot ja regressioanalyysi
- Kaksiulotteisen normaalijakauman regressiofunktiot
- Regressioanalyysin tehtävät
- >> Regressiomallin lineaarisuus

Regressiomalli

- Olkoon

$$y = f(x;\beta) + \varepsilon$$

yhden yhtälön **regressiomalli**, jossa

y = **selitettävä muuttuja**

$f(x;\beta)$ = mallin **systemaattinen** eli **rakenneosa**

ε = mallin **satunnainen osa**

- Mallin *systemaattinen osa* $f(x;\beta)$ on **selittävän muuttujan** x funktio, joka riippuu funktion f muodon määräävästä **parametrista** β.
- Mallin *satunnainen osa* ε on **jäännöstermi**, joka tavallisesti *ei riipu* selittäjästä x.

Lineaarinen regressiomalli – miksi?

- Regressiomallin

$$y = f(x;\beta) + \varepsilon$$

soveltaminen *yksinkertaistuu huomattavasti*, jos mallin *rakenneosa* $f(x;\beta)$ on *parametrin* β *suhteen lineaarinen funktio*.

- Jos mallin *rakenneosa* $f(x;\beta)$ on *parametrin* β *suhteen lineaarinen funktio*, mallia kutsutaan **lineaariseksi regressiomalliksi**.
- Huomautus:

  Epälineaaristen regressiomallien soveltaminen ei ole nykyisillä tietokoneilla ja ohjelmistoilla kovinkaan hankalaa.

Lineaarinen regressiomalli – milloin? – 1/2

- Vaikka oletus regressiomallin lineaarisuudesta saattaa tuntua rajoittavalta, oletus *on käytännössä osoittautunut* monissa regressioanalyysin sovellustilanteissa *erittäin hyvin toimivaksi*.
- Erityisesti, jos muuttujat x ja y ovat satunnaismuuttujia, joiden yhteisjakauma on **multinormaalinen**, lineaarisen regressiomallin soveltaminen on perusteltua, koska *kaikki multinormaalijakauman regressiofunktiot* eli ehdolliset odotusarvot *ovat lineaarisia*; ks. kappaletta **Kaksiulotteisen normaalijakauman regressiofunktiot.**

Lineaarinen regressiomalli – milloin? – 2/2

- Lineaarisen regressiomallin soveltaminen saattaa olla perusteltua myös monissa sellaisissa tilanteissa, joissa selitettävän muuttujan *y* riippuvuus selittäjästä *x* on **epälineaarista**:
  - (i) Muuttujien *y* ja *x* riippuvuutta voidaan usein **approksimoida** ainakin *lokaalisti* lineaarisella mallilla.
  - (ii) Muuttujien *y* ja *x* epälineaarinen riippuvuus voidaan usein **linearisoida** sopivilla *muunnoksilla*.

Epälineaarisen riippuvuuden linearisointi: Esimerkki 1/2

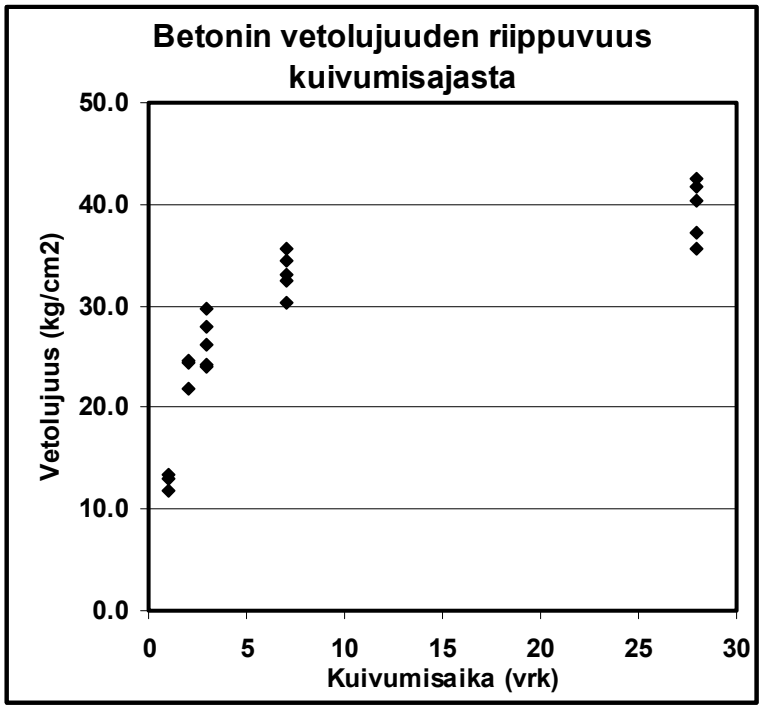
- Betonin vetolujuus riippuu betonin kuivumisajasta.
- Havaintoaineisto koostuu 21:stä lukuparista

  (x_j, y_j) , $j = 1, 2, \ldots, 21$

  jossa

  x_j = betoniharkon j kuivumisaika

  y_j = betoniharkon j vetolujuus
- Vetolujuus riippuu selvästi *epälineaarisesti* kuivumisajasta; ks. kuviota oikealla.



Epälineaarisen riippuvuuden linearisointi: Esimerkki 2/2

- Vetolujuuden epälineaarinen riippuvuus kuivumisajasta voidaan *linearisoida* seuraavilla muunnoksilla:

$$x'_j = 1/x_j$$
$$y'_j = \log(y_j)$$

jossa

x_j = betoniharkon j kuivumisaika

y_j = betoniharkon j vetolujuus

- Vrt. kuviota oikealla edellisen kalvon kuvioon.