
Johdatus tilastotieteeseen
Johdatus regressioanalyysiin

Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet

Deterministiset mallit ja regressioanalyysi

Regressiofunktiot ja regressioanalyysi

Kaksiulotteisen normaalijakauman regressiofunktiot

Regressioanalyysin tehtävät

Regressiomallin lineaarisuus

Johdatus regressioanalyysiin:

Mitä opimme? – 1/3

- Pyrimme tässä luvussa vastaamaan seuraavaan kysymykseen:
Miten jonkin, **selitettäväksi muuttujaksi** sanotun muuttujan **tilastollista riippuvuutta** joistakin toisista, **selittäviksi muuttujiksi** sanotuista muuttujista voidaan mallintaa **regressiomalliksi** sanotulla tilastollisella mallilla?
- Regressiomallin tehtävänä on **selittää selitettävän muuttujan havaittujen arvojen vaihtelu** selittävien muuttujien havaittujen arvojen vaihtelun avulla.
- **Regressioanalyysin** tavoitteet:
 - Muuttujien välisten **riippuvuuksien kuvaaminen.**
 - Muuttujien välisten **riippuvuuksien selittäminen.**
 - Selitettävän muuttujan käyttäytymisen **ennustaminen.**
 - Selitettävän muuttujan käyttäytymisen **kontrolli.**

Johdatus regressioanalyysiin:

Mitä opimme? – 2/3

- Regressioanalyysille voidaan esittää kaksi asialoogisesti varsin erilaista lähtökohtaa, joilla on kuitenkin myös monia yhtymäkohtia:
 - (i) Ongelmat **determinististen mallien** sovittamisessa havaintoihin: Havainnoille postuloitu malli ei sovi täsmällisesti kaikkiin havaintoihin.
 - (ii) Tavoitteena on moniulotteisen todennäköisyysjakauman **regressiofunktion** parametrien estimointi.
- Vaikka moniulotteisten todennäköisyysjakaumien regressiofunktiot ovat yleisesti **epälineaarisia**, **lineaariset regressiomallit** muodostavat tärkeän ja paljon sovelletun malliluokan.

Johdatus regressioanalyysiin:

Mitä opimme? – 3/3

- Lineaaristen regressiomallien suuri käyttökelpoisuus muuttujien välisten riippuvuuksien tilastollisissa analyysissä perustuu seuraaviin seikkoihin:
 - Jos havainnot noudattavat **multinormaalijakaumaa**, lineaarisen regressiomallin soveltaminen on perusteltua, koska *kaikki moniulotteisen normaalijakauman regressiofunktiot ovat lineaarisia*.
 - Lineaarisella regressiomallilla voidaan usein riittävällä tarkkuudella **approksimoida** muuttujien välisiä epälineaarisia riippuvuuksia.
 - Muuttujien välinen epälineaarinen riippuvuus voidaan usein **linearisoida** sopivilla muunnoksilla.

Johdatus regressioanalyysiin: Esitiedot

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

Johdatus regressioanalyysiin: Lisätiedot

- Regressioanalyysia *yhden selittäjän lineaarisen regressiomallin* tapauksessa käsitellään luvussa

Yhden selittäjän lineaarinen regressiomalli

- Pitemmälle meneviä regressioanalyysin kysymyksiä käsitellään luvuissa

Yleinen lineaarinen malli

Regressiodiagnostiikka

Regressiomallin valinta

Regressioanalyysin erityiskysymyksiä

Johdatus regressioanalyysiin

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

Regressioanalyysin lähtökohdat ja tavoitteet

Avainsanat

Deterministinen malli
Lineaarinen regressiomalli
Regressioanalyysi
Regressiofunktio
Regressiomalli
Selitettävä muuttuja
Selittäminen
Selittävä muuttuja
Tilastollinen riippuvuus

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

- **Regressioanalyysissä** selitettävän muuttujan *tilastolliselle riippuvuudelle* selittävistä muuttujista *pyritään rakentamaan tilastollinen malli*, jota kutsutaan **regressiomalliksi**.
- Koska *riippuvuuksien analysointi* on tavallisesti tieteellisen tutkimuksen keskeinen tavoite, **regressioanalyysi on eniten sovellettuja ja tärkeimpiä tilastotieteen menetelmiä**.

Regressioanalyysin lähtökohdat ja tavoitteet

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**.

Regressioanalyysin lähtökohdat ja tavoitteet

Regressiomallien luokittelu 1/2

- Regressioanalyysissä 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**

Regressioanalyysin lähtökohdat ja tavoitteet

Regressiomallien luokittelu 2/2

- Tässä johdatuksessa tilastotieteeseen käsitellään pääasiassa **lineaarisia yhden yhtälön regressiomalleja**; ks. lukua **Yhden selittäjän lineaarinen regressiomalli**.
- On hyödyllistä tietää, että **variانسianalyysissa** sovellettavat tilastolliset mallit voidaan ymmärtää ns. *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

- Regressioanalyysillä on kaksi erilaista *lähtökohtaa*, joilla on kuitenkin monia yhtymäkohtia:
 - (i) Ongelmat **determinististen 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 ja regressioanalyysi

Avainsanat

Deterministinen malli

Estimointi

Parametri

Regressioanalyysi

Regressiomalli

Selitettävä muuttuja

Selittäminen

Selittävä muuttuja

Tilastollinen riippuvuus

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 **parametrasta** (vakioista).
- Tällöin parametrin arvo voidaan pyrkiä **estimoimaan** eli *arvioimaan havaintojen avulla*.
- Oletetaan, että **parametrille ei ole mahdollista löytää sellaista arvoa, joka saisi mallin sopimaan samanaikaisesti** 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 *vakioista* 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 determinististen 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_j ja y_j liittyvät *samaan havaintoyksikköön kaikille $j = 1, 2, \dots, n$.*
- Oletetaan, että ei ole olemassa *yhtä* parametrin β arvoa, joka saa yhtälön

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

toteutumaan *samanaikaisesti kaikille havainnoille x_j ja y_j .*

- Kirjoitetaan

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

Deterministiset mallit ja regressio-ongelma 3/4

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

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

vaihtelevat satunnaisesti yhtälöstä toiseen.

Huomaa, että oletuksesta seuraa, että *selitettävän muuttujan y havaittujen arvojen y_j on oltava satunnaisia.*

- Yhtälö

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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_j samanaikaisesti mahdollisimman pieniä.*

- Tämä on *käyränsovitusongelma:*

Miten parametrin β arvo on valittava, jotta käyrä

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

kulkisi jossakin mielessä *mahdollisimman läheltä jokaista havaintopistettä*

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

- 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

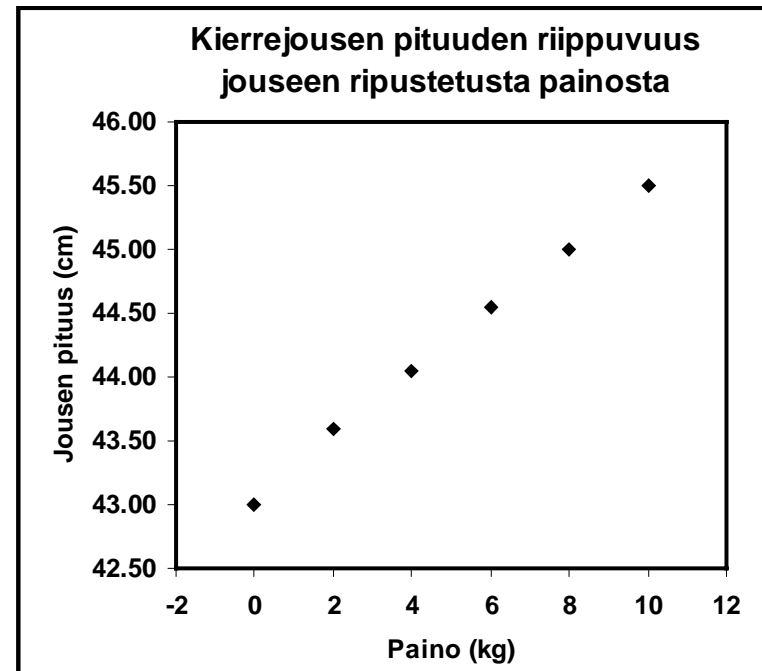
β = ns. *jousivakio*

- Jousivakion 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

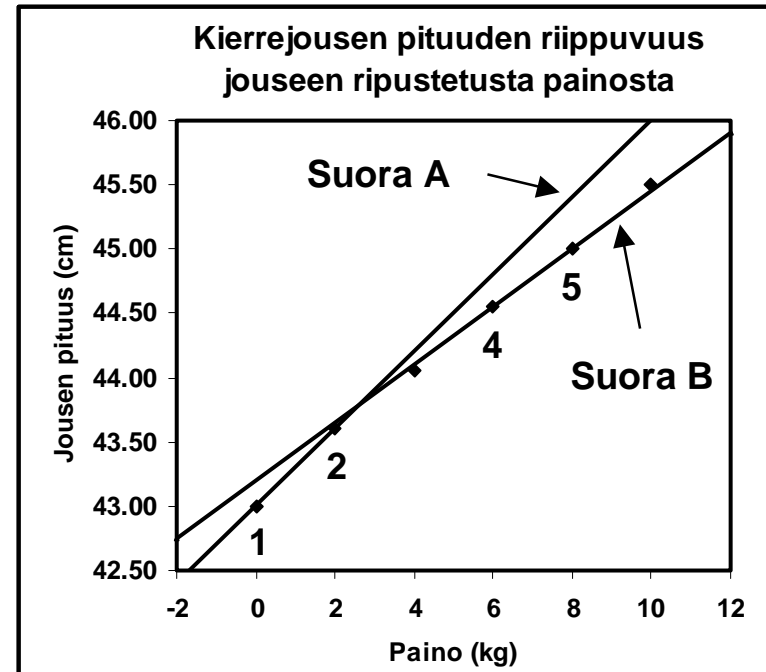
Deterministiset mallit ja regressio-ongelma: Esimerkki 2/4

- 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 havaintopisteiden 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 kulkee jossakin mielessä *mahdollisimman läheltä* jokaista havaintopistettä?



Deterministiset mallit ja regressio-ongelma: Esimerkki 4/4

- Käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

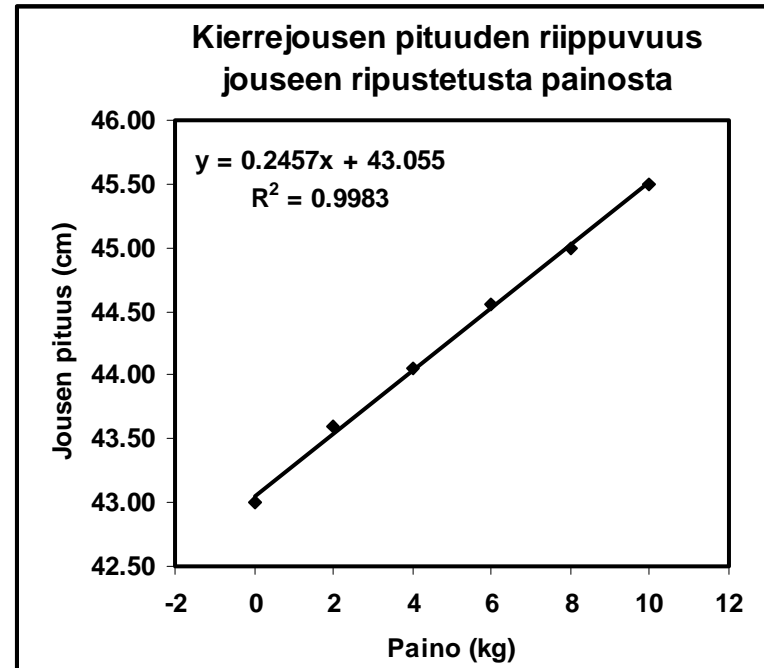
$$y = \alpha + \beta x$$

kertoimet niin, että neliösumma

$$\sum_{j=1}^n (y_j - \alpha - \beta x_j)^2$$

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 determinististen mallien yhteydessä?
- Syitä regressio-ongelman syntymiseen:
 - (i) *Havaintovirheet* selitettävän muuttujan y havaituissa arvoissa.
 - (ii) *Yhtälö*

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

on idealisointi:

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

Regressiomalli ja kiinteät selittäjät 1/2

- Olkoon

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

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

- Oletukset:

- (i) Selittävän muuttujan x arvot x_j voidaan *valita*, jolloin ne ovat *kiinteitä* eli *ei-satunnaisia*.
- (ii) Jäännös- eli virhetermit ε_j ovat *satunnaisia*, jolloin myös selitettävän muuttujan y havaitut arvot y_j pitää olettaa satunnaisiksi.

Regressiomalli ja kiinteät selittäjät 2/2

- Regressiomallissa

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

on seuraavat osat:

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

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

ε_j = *satunnainen* ja *ei-havaittu* **jäännös-** eli **virhetermi** havaintoyksikössä j

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 ja regressioanalyysi

Avainsanat

Ehdollinen jakauma

Ehdollinen odotusarvo

Ennustaminen

Ennustevirhe

Estimointi

Keskineliövirhe

Parametri

Regressioanalyysi

Regressiofunktio

Regressiomalli

Reunajakauma

Selitettävä muuttuja

Selittäminen

Selittävä muuttuja

Yhteisjakauma

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 **parametrasta** (vakioista).
- Tällöin parametrin arvo voidaan pyrkiä **estimoimaan** eli *arvioimaan havaintojen avulla*.
- **Miten parametrille löydetään jossakin mielessä mahdollisimman hyvä estimaatti eli arvio?**

Regressiofunktiot ja regressioanalyysi

Ehdollinen jakauma

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

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

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

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

jossa

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

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

$$E(y | x)$$

kutsutaan ehtomuuttujan x arvojen funktiona *satunnaismuuttujan y regressiofunktioksi* muuttujan x suhteen.

- Regressiofunktion $E(y | x)$ muoto riippuu satunnaismuuttujan y ehdollisen jakauman

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

parametreista.

Regressiofunktio 2/2

- Olkoon

$$E(y | x)$$

satunnaismuuttujan y regressiofunktio satunnaismuuttujan x suhteen.

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

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

jossa β on satunnaismuuttujan y ehdollisen jakauman

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

muodon määräävä *parametri*.

Lisätietoja

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

Ks. kalvokokoelman **Johdatus todennäköisyyslaskentaan** lukua **Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat.**

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 $d(y | x)$ muuttujan x saamaan arvoon perustuva **ennuste** muuttujan y arvolle.
- Miten ennuste $d(y | x)$ valitaan *optimaalisella tavalla*?

Regressiofunktio ja ennustaminen 2/3

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

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

minimoituu.

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

$$d(y | x) = \text{E}(y | x)$$

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

Regressiofunktio ja ennustaminen 3/3

- Olkoon

$$y - E(y | x) = \varepsilon$$

optimaalisen ennusteen $E(y | x)$ **ennustevirhe**.

- Tällöin voimme kirjoittaa

$$\begin{aligned} y &= E(y | x) + \varepsilon \\ &= f(x; \beta) + \varepsilon \end{aligned}$$

jossa

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

on satunnaismuuttujan y *regressiofunktio* satunnaismuuttujan x suhteen.

Regressiofunktio regressiomallina

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

$$\begin{aligned}y &= E(y | x) + \varepsilon \\ &= f(x; \beta) + \varepsilon\end{aligned}$$

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

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

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

Regressiofunktiot ja regressio-ongelma 2/3

- Oletetaan, että satunnaismuuttujia x ja y koskevat havainnot x_j ja y_j liittyvät *samaan havaintoyksikköön kaikille* $j = 1, 2, \dots, n$.

- Edellä esitetyn nojalla voimme kirjoittaa yhtälön

$$y_j = f(x_j; \beta) + \varepsilon_j, \quad j = 1, 2, \dots, n$$

jossa ε_j 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* jäännöstermeistä ε_j *samanaikaisesti mahdollisimman pieniä*.
- Tämä on *käyränsovitusongelma*:

Miten parametrin β arvo on valittava niin, että käyrä

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

kulkisi *mahdollisimman läheltä jokaista havaintopistettä*

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

- 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

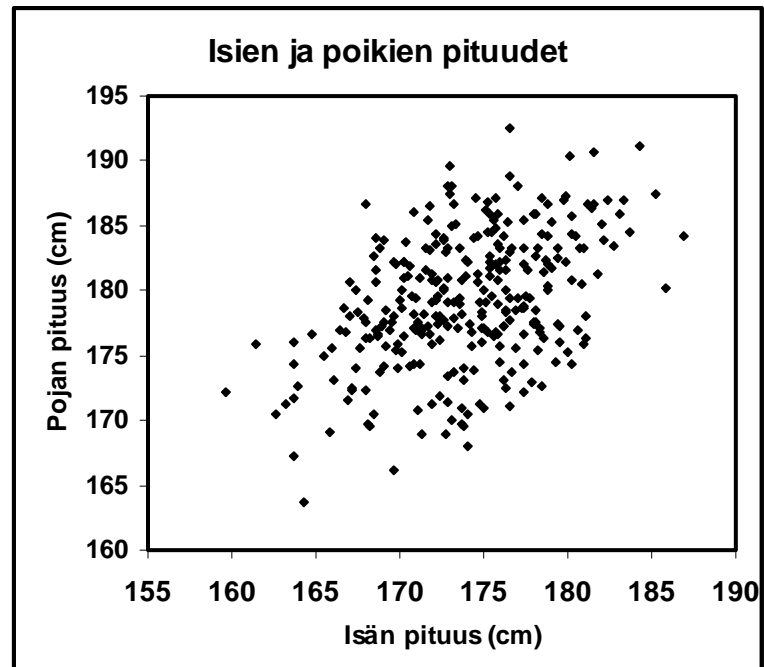
$$(x_j, y_j), j = 1, 2, \dots, 300$$

jossa

$$x_j = \text{isän } j \text{ pituus}$$

$$y_j = \text{isän } j \text{ pojan pituus}$$

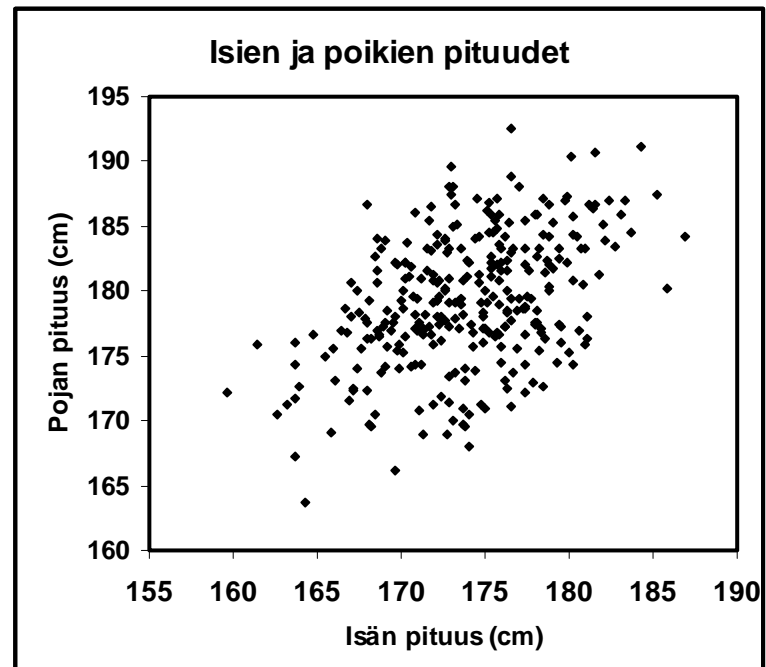
- Ks. pistediagrammia oikealla.



Mitä regressiofunktio mallintaa?

Esimerkki 2/6

- Pojan pituuden riippuvuus isän pituudesta ei 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*

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

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

Mitä regressiofunktio mallintaa?

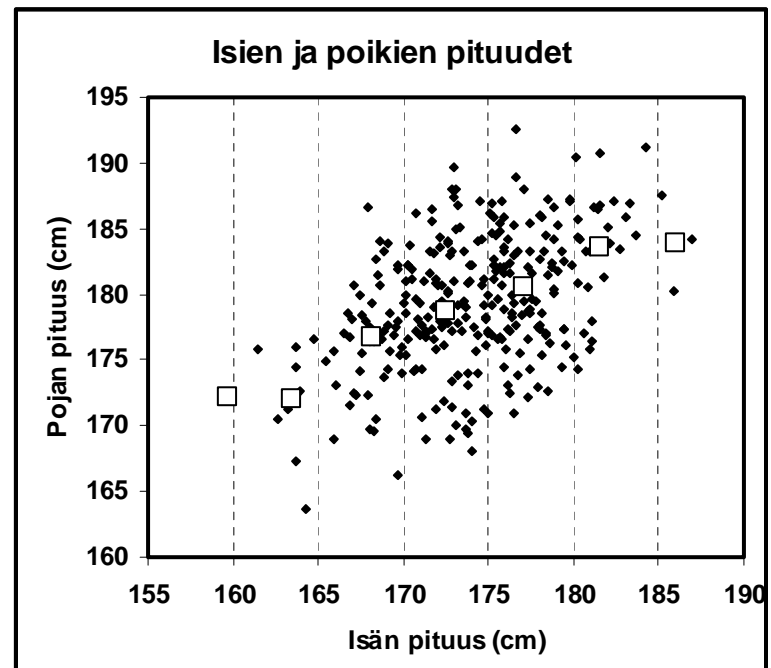
Esimerkki 4/6

- *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.



Mitä regressiofunktio mallintaa?

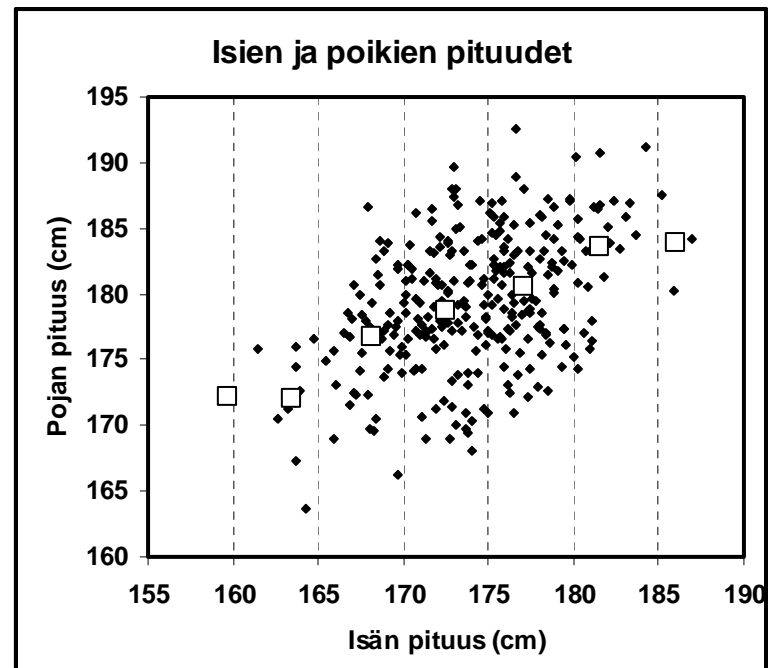
Esimerkki 5/6

- 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

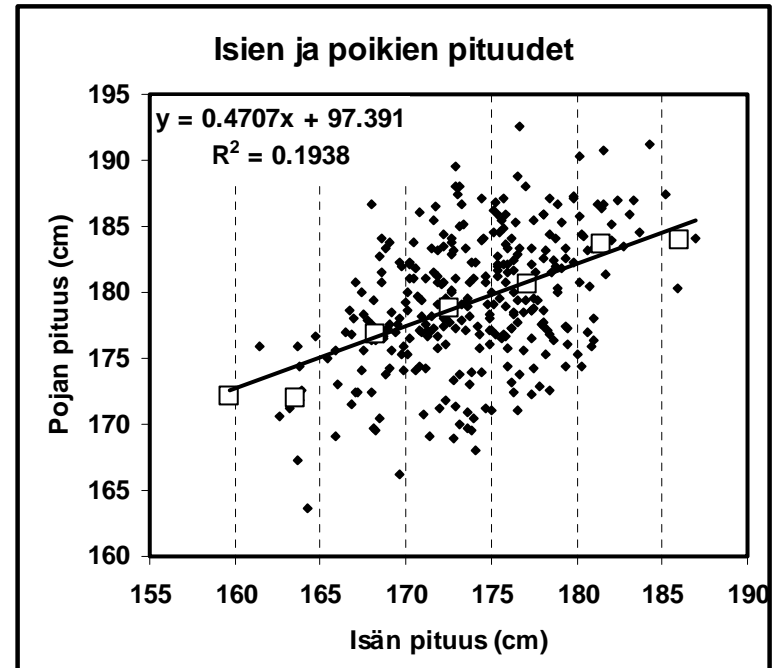
$$y = \alpha + \beta x$$

kertoimet niin, että neliösumma

$$\sum_{j=1}^n (y_j - \alpha - \beta x_j)^2$$

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

Kaksiulotteisen normaalijakauman regressiofunktiot

Avainsanat

Ehdollinen jakauma
Ehdollinen odotusarvo
Ehdollinen varianssi
Kaksiulotteinen normaalijakauma
Multinormaalijakauma
Regressiofunktio
Regressiosuora
Reunajakauma

Kaksiulotteisen normaalijakauman regressiofunktiot

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 normaali-jakaumaa**; lisätietoja: ks. kalvokokoelman Johdatus todennäköisyyslaskentaan 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

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

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

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

2-ulotteinen normaalijakauma:

Tiheysfunktio 2/2

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

$$\mu_x = E(x) = \text{muuttujan } x \text{ odotusarvo}$$

$$\mu_y = E(y) = \text{muuttujan } y \text{ odotusarvo}$$

$$\sigma_x^2 = \text{Var}(x) = \text{muuttujan } x \text{ varianssi}$$

$$\sigma_y^2 = \text{Var}(y) = \text{muuttujan } y \text{ varianssi}$$

$$\rho_{xy} = \text{Cor}(x, y) = \text{muuttujien } x \text{ ja } y \text{ korrelaatio}$$

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*

$$E(x) = \mu_x$$

$$E(y) = \mu_y$$

$$\text{Var}(x) = \sigma_x^2$$

$$\text{Var}(y) = \sigma_y^2$$

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

määräävät täydellisesti 2-ulotteisen normaalijakauman, merkitään

$$(x, y) \sim N_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$$

Kaksiulotteisen normaalijakauman regressiofunktiot

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*

$$E(x) = \mu_x \qquad E(y) = \mu_y$$

määräävät satunnaismuuttujien x ja y yhteisjakauman *todennäköisyysmassan painopisteen*.

- Satunnaismuuttujien x ja y *variانسsit*

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

kuvaavat satunnaismuuttujien x ja y *todennäköisyysmassojen hajaantuneisuutta* niiden odotusarvojen μ_x ja μ_y ympärillä.

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma:

Parametrien tulkinta 2/2

- Satunnaismuuttujien x ja y *korrelaatio*

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

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:

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

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

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

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma: Ehdolliset jakaumat 1/2

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

$$y | x \sim N(\mu_{y|x}, \sigma_{y|x}^2)$$

jossa

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

$$\sigma_{y|x}^2 = \text{Var}(y | x) = (1 - \rho_{xy}^2) \sigma_y^2$$

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma: Ehdolliset jakaumat 2/2

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

$$x | y \sim \mathbf{N}(\mu_{x|y}, \sigma_{x|y}^2)$$

jossa

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

$$\sigma_{x|y}^2 = \mathbf{Var}(x | y) = (1 - \rho_{xy}^2) \sigma_x^2$$

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma: Regressiofunktiot 1/2

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

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

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 satunnaismuuttujan y suhteen*

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

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.

Kaksiulotteisen normaalijakauman regressiofunktiot

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*.

Kaksiulotteisen normaalijakauman regressiofunktiot

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)*

Kaksiulotteisen normaalijakauman regressiofunktiot

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)*

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma: Ehdolliset varianssit 1/2

- *Satunnaismuuttujan y ehdollinen varianssi satunnaismuuttujan x suhteen on*

$$\sigma_{y|x}^2 = \text{Var}(y | x) = (1 - \rho_{xy}^2) \sigma_y^2$$

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.

Kaksiulotteisen normaalijakauman regressiofunktiot

2-ulotteinen normaalijakauma:

Ehdolliset varianssit 2/2

- *Satunnaismuuttujan x ehdollinen varianssi satunnaismuuttujan y suhteen on*

$$\sigma_{x|y}^2 = \text{Var}(x | y) = (1 - \rho_{xy}^2) \sigma_x^2$$

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

$$\sigma_{y|x}^2 = \text{Var}(y | x) = (1 - \rho_{xy}^2) \sigma_y^2$$

on seuraavat ominaisuudet:

(i) $\sigma_{y|x}^2 \leq \sigma_y^2$

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

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

2-ulotteinen normaalijakauma:

Ehdollisten varianssien ominaisuudet 2/2

- Satunnaismuuttujan x ehdollisella varianssilla satunnaismuuttujan y suhteen

$$\sigma_{x|y}^2 = \text{Var}(x | y) = (1 - \rho_{xy}^2) \sigma_x^2$$

on seuraavat ominaisuudet:

(i) $\sigma_{x|y}^2 \leq \sigma_x^2$

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

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

Johdatus regressioanalyysiin

Regressioanalyysin lähtökohdat ja tavoitteet

Deterministiset mallit ja regressioanalyysi

Regressiofunktiot ja regressioanalyysi

Kaksiulotteisen normaalijakauman regressiofunktiot

>> Regressioanalyysin tehtävät

Regressiomallin lineaarisuus

Regressioanalyysin tehtävät

Avainsanat

Ennustaminen

Estimointi

Jäännöstermi

Mallin rakenneosa eli

systemaattinen osa

Mallin satunnainen osa

Oletusten tarkistaminen

Regressioanalyysi

Regressiomalli

Regressiomallin hyvyys

Satunnainen osa

Selitettävä muuttuja

Selittäminen

Selittävä muuttuja

Testaus

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 **parametrasta** 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 .

- Regressioanalyysissä *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.*

Regressioanalyysin tehtävät

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

Regressiomallin lineaarisuus

Avainsanat

Approksimointi

Lineaarinen regressiomalli

Linearisointi

Multinormaalijakauma

Regressiofunktio

Regressiomallin lineaarisuus

Regressiomalli

- Olkoon

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

*y*hden 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 rakenneosaa $f(x; \beta)$ on parametrin β suhteen lineaarinen funktio.

- Jos mallin rakenneosaa $f(x; \beta)$ 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

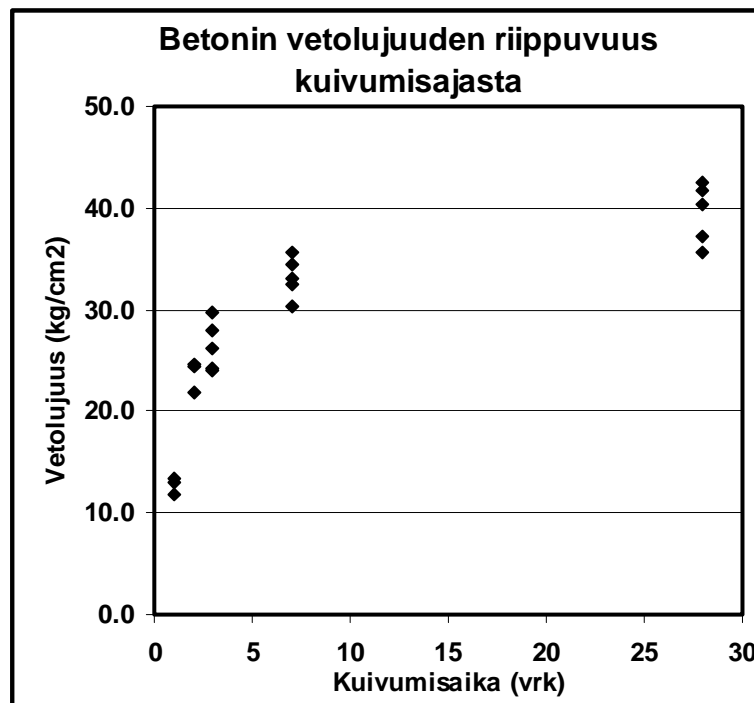
$$(x_j, y_j), j = 1, 2, \dots, 21$$

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.

