Tilastolliset menetelmät: Lineaarinen regressioanalyysi

Sisällys

13. Tilastollinen riippuvuus ja korrelaatio

13.1. Tilastollinen riippuvuus, korrelaatio ja regressio

13.2. Kahden muuttujan havaintoaineiston kuvaaminen

13.3. Pearsonin korrelaatiokertoimen estimointi ja testaus

13.4. Järjestyskorrelaatiokertoimet

Tarkastelemme tässä luvussa *kahden* (tai useamman) *muuttujan tilastollisten aineistojen analyysia*. Pyrimme vastaamaan seuraaviin kysymyksiin:

- Miten **kahden** (tai useamman) **muuttujan** samanaikainen tarkastelu vaikuttaa tilastollisen analyysin suorittamiseen?
- Miten kahden (tai useamman) muuttujan tilastollista aineistoa **kuvataan**?
- Mitä tarkoitetaan kahden tekijän tai muuttujan **tilastollisella riippuvuudella** ja miten tilastollinen riippuvuus eroaa **eksaktista riippuvuudesta**?
- Mitä on **korrelaatio**?
- Mikä on **korrelaation** ja **riippuvuuden** suhde?
- Miten korrelaatiot **estimoidaan**?
- Miten korrelaatioita koskevia **hypoteeseja testataan**?

Tämä kappale on johdantoa tämän tilastotiedettä käsittelevän monisteen osan pääkohteelle, mikä on **lineaariset regressiomallit**.

Avainsanat:

Aikasarjadiagrammi, Aritmeettinen keskiarvo, Eksakti riippuvuus, Estimaattori, Estimointi, Fisherin *z*-muunnos, Järjestyskorrelaatiokerroin, Kendallin järjestyskorrelaatiokerroin, Keskihajonta, Korrelaatio, Korrelaatiokerroin, Korrelaatiokertoimien vertailutesti, Korrelaation testaaminen, Korreloimattomuuden testaaminen, Kovarianssi, Keskihajonta, Kriittinen arvo, Luottamustaso, Luottamusväli, Merkitsevyys-taso, Normaalijakauma, Otos, Otostunnusluku, *p*-arvo, Pearsonin otoskorrelaatiokerroin, Piste-diagrammi, Regressioanalyysi, Regressiomalli, Riippuvuus, Spearmanin järjestyskorrelaatiokerroin, Testi, Testi korrelaatiokertoimelle, Tilastollinen riippuvuus, Usean muuttujan havaintoaineiston kuvaaminen, Varianssi

13.1. Tilastollinen riippuvuus, korrelaatio ja regressio

Tieteellisen tutkimuksen *tärkeimmät* ja *mielenkiintoisimmat kysymykset liittyvät* tavallisesti tutkimuksen kohteena olevaa ilmiötä kuvaavien **tekijöiden** tai **muuttujien välisiin riippuvuuksiin**.

Jos tilastollisen tutkimuksen kohteena olevaan ilmiöön liittyy useampia kuin yksi muuttuja, *yhden muuttujan tilastolliset menetelmät* antavat tavallisesti vain *rajoittuneen kuvan* ilmiöstä. Sovellusten kannalta ehkä merkittävin osa tilastotiedettä käsittelee kahden tai useamman muuttujan välisten *riippuvuuksien kuvaamista* ja *mallintamista*.

Esimerkki 1: Riippuvuustarkasteluja.

- Miten työttömyysaste Suomessa (% työvoimasta) *riippuu* BKT:n (bruttokansantuotteen) kasvuvauhdista Suomessa, Suomen viennin volyymista sekä BKT:n kasvuvauhdista muissa EU-maissa ja USA:ssa?
- Miten alkoholin kulutus (l *per capita* vuodessa) *riippuu* alkoholijuomien hintatasosta, ihmisten käytettävissä olevista tuloista ja alkoholin saatavuudesta?
- Miten todennäköisyys sairastua keuhkosyöpään (*p*) *riippuu* tupakoinnin määrästä ja kestosta?
- Miten vehnän hehtaarisato (t/ha) *riippuu* kesän keskilämpötilasta ja sademäärästä sekä maan muokkauksesta, lannoituksesta ja tuholaisten torjunnasta?
- Miten betonin lujuus (kg/cm^2) *riippuu* sen kuivumisajasta?
- Miten kemiallisen aineen saanto (%) *riippuu* valmistusprosessissa käytettävästä lämpötilasta?

Tarkastelemme tässä esityksessä yksinkertaisuuden vuoksi vain *kahden* muuttujan välisiä riippuvuuksia:

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

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

Kahden muuttujan välistä (lineaarista) *tilastollista riippuvuutta* kutsutaan tilastotieteessä tavallisesti **korrelaatioksi**. *Korrelaation* eli (lineaarisen) *tilastollisen riippuvuuden voimakkuutta* mittaavia tilastollisia tunnuslukuja kutsutaan **korrelaatiokertoimiksi**. Korrelaatiot muodostavat *perustan muuttujien välisten* (lineaaristen) *riippuvuuksien ymmärtämiselle*.

Vaikka korrelaatiot muodostavat perustan muuttujien välisten riippuvuuksien ymmärtämiselle, riippuvuuksia halutaan tavallisesti *analysoida* myös *tarkemmin*. **Regressioanalyysi** on tilastollinen menetelmä, jossa jonkin, ns. *selitettävän muuttujan* tilastollista riippuvuutta joistakin toisista, ns. *selittävistä muuttujista* pyritään mallintamaan **regressiomalliksi** kutsutulla tilastollisella mallilla; ks. lukua **Johdatus regressioanalyysiin**.

Huomautus:

- Tässä luvussa rajoitutaan tarkastelemaan tilastollisten riippuvuuksien *kuvaamista* ja *mittaamista*.

Kuten yhden muuttujan havaintoaineistojen tapauksessa, lähtökohdan kahden tai useamman muuttujan havaintoaineistojen kuvaamiselle muodostaa tutustuminen **havaintoarvojen jakaumaan**. Havaintoarvojen jakaumaa voidaan kuvailla ja esitellä *tiivistämällä* havaintoarvoihin sisältyvä *informaatio* sopivaan muotoon:

- Havaintoarvojen *jakaumaa kokonaisuutena* voidaan kuvata sopivasti valituilla **graafisilla esityksillä**.
- Havaintoarvojen *jakauman karakteristisia ominaisuuksia* voidaan kuvata sopivasti valituilla **otostunnusluvuilla**.

Koska useampi- kuin kaksiulotteisten kuvioiden tekeminen ei ole käytännössä mahdollista, kolmen tai useamman muuttujan havaintoaineistoja havainnollistetaan tavallisesti niin, että *muuttujia tarkastellaan pareittain*.

Kahden *järjestys-*, *välimatka-* tai *suhdeasteikoillisen* muuttujan havaittujen arvojen pareja havainnollistetaan tavallisesti graafisella esityksellä, jota kutsutaan **pistediagrammiksi**.

Huomautus:

- *Monimuuttujamenetelmissä* on kehitetty myös sellaisia tilastografiikan menetelmiä, joilla voidaan havainnollistaa *useampi- kuin kaksiulotteisia aineistoja*.

Usean muuttujan havaintoaineistojen karakteristisia ominaisuuksia voidaan kuvata *muuttuja-kohtaisilla* **otostunnusluvuilla**. Muuttujakohtaiset otostunnusluvut *eivät* kuitenkaan *voi antaa informaatiota muuttujien välisistä riippuvuuksista*. Muuttujien *pareittaisia tilastollisia riippuvuuksia voidaan kuvata* sopivasti valitulla **korrelaation mitalla**.

Tutkittavien muuttujien *mitta-asteikolliset ominaisuudet* ohjaavat korrelaation mitan valintaa:

- *Välimatka- ja suhdeasteikollisille muuttujille* käytetään tavallisesti **Pearsonin korrelaatiokerrointa**.
- *Järjestysasteikollisille muuttujille* käytetään tavallisesti **Spearmanin** tai **Kendallin järjestyskorrelaatiokerrointa**.

Satunnaismuuttujien väliseen korrelaatioon voidaan kohdistaa erilaisia *tilastollisia testejä*.

Tässä esityksessä tarkastellaan seuraavia *Pearsonin korrelaatiokertoimelle* sopivia testejä:

- **Yhden otoksen testi korrelaatiokertoimelle**
- **Korrelaatiokertoimien vertailutesti**
- **Testi korreloimattomuudelle**

Lisäksi tässä esityksessä tarkastellaan seuraavia *Spearmanin ja Kendallin järjestyskorrelaatio-kertoimille* sopivia testejä:

- **Testit korreloimattomuudelle**

13.2. Kahden muuttujan havaintoaineiston kuvaaminen

Pistediagrammi

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

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

Olkoot

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

ja

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

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

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

pisteinä avaruudessa ^{2}.

Havainnollistus:

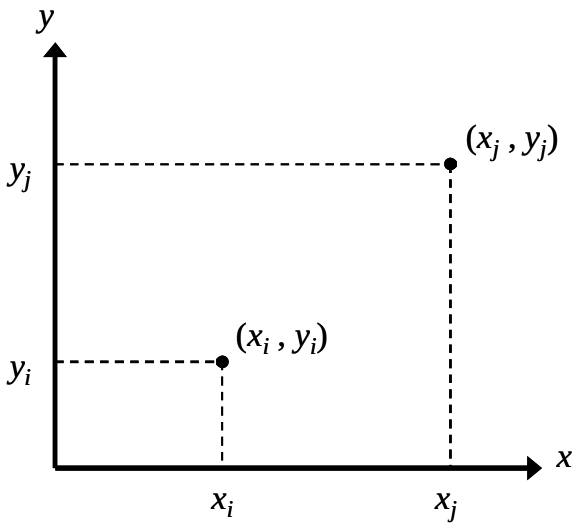
Kuvio oikealla esittää lukuparien

$$(x_i, y_i)$$

ja

$$(x_j, y_j)$$

määrittelemien pisteiden esittämistä tasokoordinaatistossa



Huomautus:

- Kahden tai useamman muuttujan havaintoaineistoja kannattaa tietysti kuvata myös soveltamalla jokaiseen muuttujaan erikseen yhden muuttujan havaintoaineistojen kuvaamiseen tarkoitettuja välineitä; ks. lukua **Tilastollisten aineistojen kuvaaminen**.

Esimerkki 1: Hooken laki.

Hooken lain mukaan kierrejousen (ns. ideaalijousen) pituus y riippuu *lineaarisesti* jouseen ripustetusta painosta x:

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

jossa

α = jousen pituus ilman painoa

β = ns. *jousivakio*

Alla olevassa taulukossa esitetään tulokset kokeesta, jossa Hooken lain pätevyyttä tutkittiin mittaamalla jousen pituus ilman painoa sekä painoilla, jotka olivat 2, 4, 6, 8 ja 10 kg.

Merkitään:

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

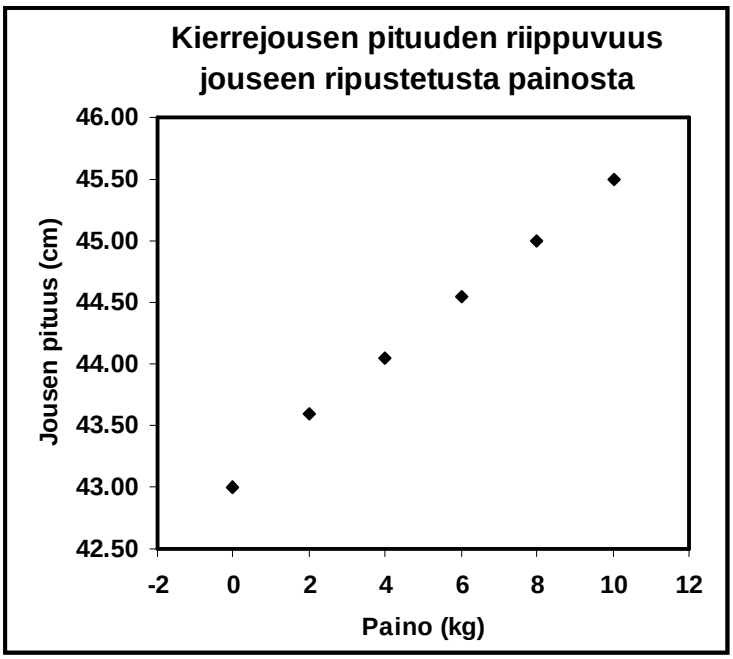
jossa

x_i = paino i

y_i = jousen pituus, kun painona on x_i

Alla oleva pistediagrammi havainnollistaa koetuloksia graafisesti.

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



Kysymys: Ovatko koetulokset *sopusoinnussa* Hooken lain kanssa?

Vastausta tähän kysymykseen tarkastellaan luvuissa **Johdatus regressioanalyysiin** ja **Yhden selittäjän lineaarinen regressiomalli**.

Esimerkki 2. Poikien pituuden riippuvuus isien pituudesta.

Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.

Kysymys: *Periytyykö isien pituus heidän pojilleen*?

Havaintoaineistona on tässä 300:n isän ja heidän poikiensa pituuksien muodostamaa lukuparia

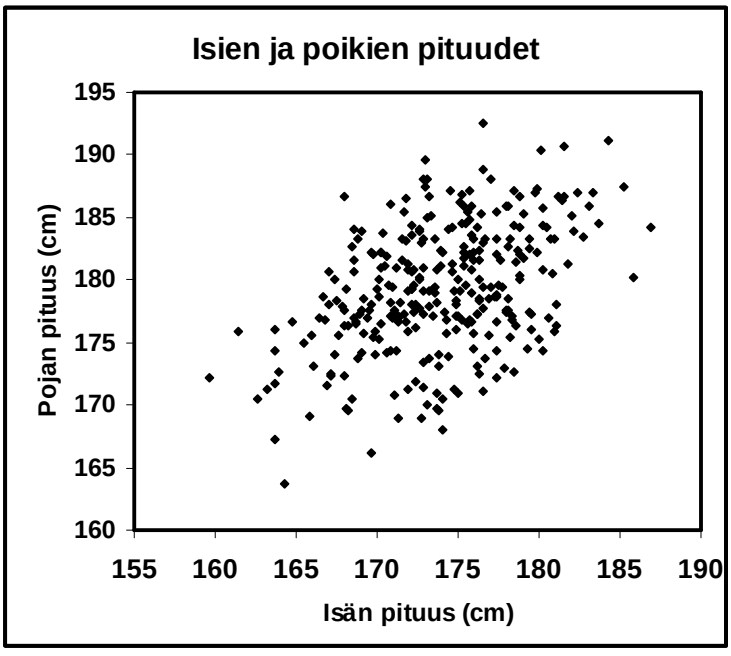
$$(x_i , y_i) , \; i = 1, 2, \ldots , 300$$

jossa siis

x_i = isän i pituus

y_i = isän i pojan pituus

Ks. pistediagrammia oikealla.



Pojan pituuden riippuvuus isän pituudesta ei selvästikään ole *eksaktia*: Saman mittaisten isien poikien pituudet näyttävät vaihtelevan paljonkin.

Kuvasta nähdään kuitenkin se, että lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkillä isillä ja vastaavasti pitkillä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.

Tällaisten *tilastollisten riippuvuuksien* analysoimista *lineaaristen regressiomallien* avulla tarkastellaan luvuissa **Johdatus regressioanalyysiin** ja **Yhden selittäjän lineaarinen regressiomalli**.

Esimerkki 3. Keuhkosyövän yleisyyden riippuvuus savukkeiden kulutuksesta.

Onko keuhkosyöpä yleisempää sellaisissa maissa, joissa tupakoidaan paljon?

Oikealla on taulukko, jossa on tiedot savukkeiden kulutuksesta ja keuhkosyövän yleisyydestä 10:ssä maailman maassa.

Huomaa, että keuhkosyövän yleisyys on mitattu 20 vuotta savukkeiden kulutuksen mittaamisen jälkeen.

Tämä johtuu tietysti siitä, että keuhkosyövän kehittyminen vaatii pitkän altistusajan.

Havaintoaineistona on tässä siis 10 lukuparia

$$(x_i , y_i) , \; i = 1, 2, \ldots , 10$$

jossa

Maa	Savukkeiden kulutus (kpl) per capita 1930	Keuhkosyöpä-tapausten 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

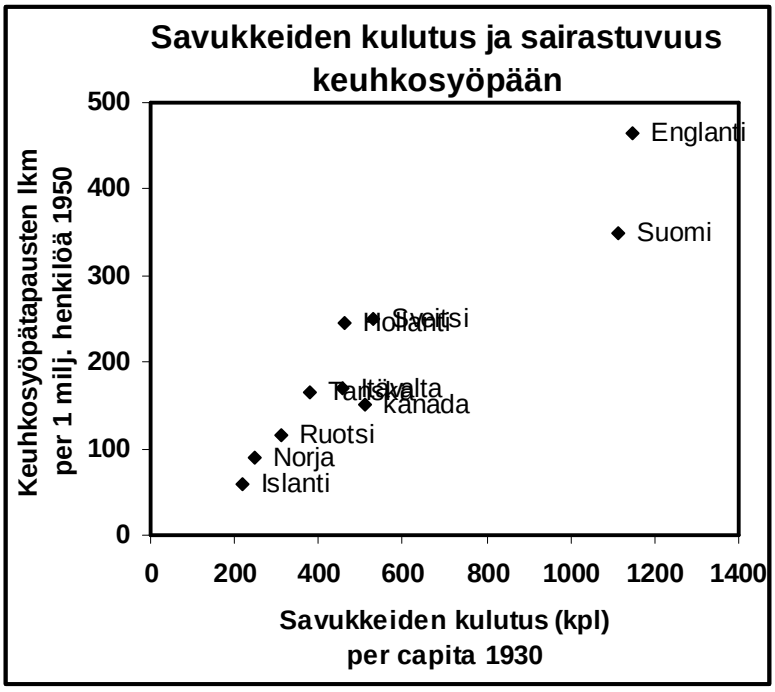
x_i = savukkeiden kulutus maassa i vuonna 1930

y_i = sairastuvuus keuhkosyöpään maassa i vuonna 1950

Oikealla oleva pistediagrammi havainnollistaa savukkeiden kulutuksen ja keuhkosyövän yleisyyden välistä yhteyttä.

Sairastuvuus keuhkosyöpään näyttää olevan *keskimäärin* korkeampaa sellaisissa maissa, joissa savukkeiden kulutus on ollut *keskimääräistä* suurempaa.

Tällaisten *tilastollisten riippuvuuksien* analysoimista *lineaaristen regressiomallien* avulla tarkastellaan luvussa **Yhden selittäjän lineaarinen regressiomalli**.



Esimerkki 4. Betonin lujuuden riippuvuus kuivumisajasta.

Kokeessa tutkittiin betonin vetolujuuden riippuvuutta betonin kuivumisajasta.

Havaintoaineistona on 21 lukuparia

(x_i , y_i) , $i = 1, 2, \ldots , 21$

jossa

x_i = betoniharkon i kuivumisaika

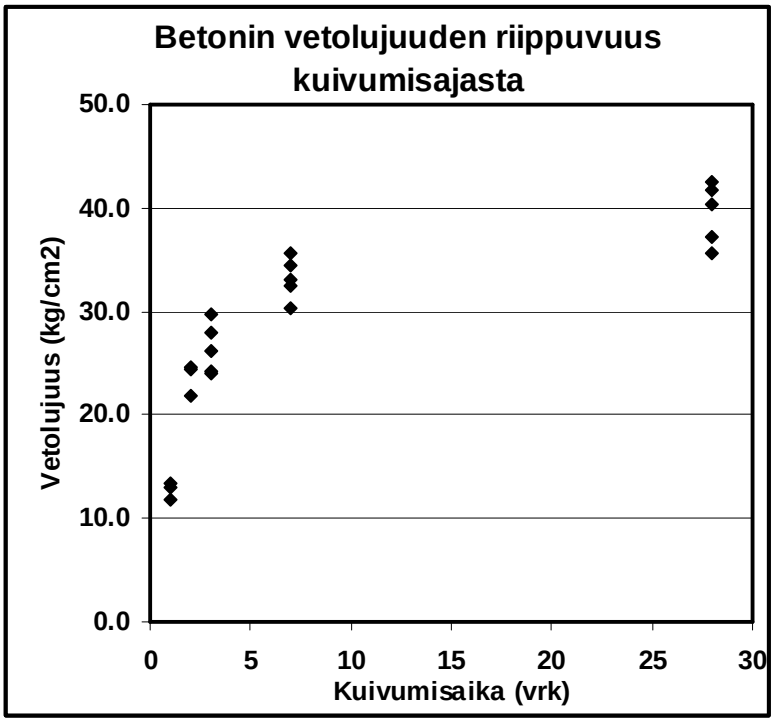
y_i = betoniharkon i vetolujuus

Ks. pistediagrammia oikealla.

Vetolujuus näyttää kuvan perusteella riippuvan *epälineaarisesti* kuivumisajasta.

Tässä tapauksessa muuttujien välinen ilmeinen epälineaarinen riippuvuus voidaan kuitenkin *linearisoida*; ks. lukua **Johdatus regressioanalyysiin**.

Linearisoinnin jälkeen riippuvuutta voidaan analysoida *lineaaristen regressiomallien* avulla.



Aikasarjadiagrammi

Oletetaan, että *järjestys-*, *välimatka-* tai *suhdeasteikollisen* muuttujan x havaitut arvot

$x_1, x_2, \ldots , x_n$

muodostavat *aikasarjan*. Tällä tarkoitetaan sitä, että havaintoarvot x_t , $t = 1, 2, \ldots , n$ on indeksoitu niin, että indeksi i viittaa *peräkkäisiin* ajanhetkiin, jolloin havainnot ovat *aikajärjestyksessä*. **Aikasarjadiagrammi** on pistediagrammi, joka saadaan esittämällä *lukuparit*

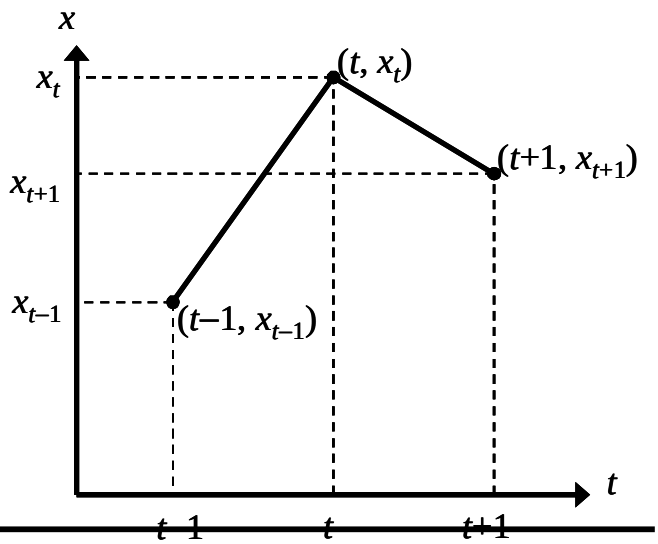
(t , x_t) , $t = 1, 2, \ldots , n$

pisteinä avaruudessa ^{2}. Lisäksi *peräkkäisiin ajanhetkiin liittyvät pisteet*

$(t–1 , x_{t–1})$ ja (t , x_t) ,

$t = 2, 3, \ldots , n$

yhdistetään aikasarjadiagrammissa tavallisesti toisiinsa *janoilla*.



Havainnollistus:

Kuvio oikealla esittää aikasarjan

x_t , $t = 1, 2, \ldots , n$

peräkkäisten havaintoarvojen

$x_{t–1}$, x_t , x_{t+1}

määrittelemien pisteiden esittämistä
tasokoordinaatistossa.

Esimerkki 5. Kuukausimyynnin arvon kehitys.

Alla on aikasarjadiagrammi, joka esittää erään tukkukaupan kk-myynnin arvon vaihtelua.

Havaintoaineistona on 144 lukuparia

$$(t\ ,\ x_t)$$

jossa

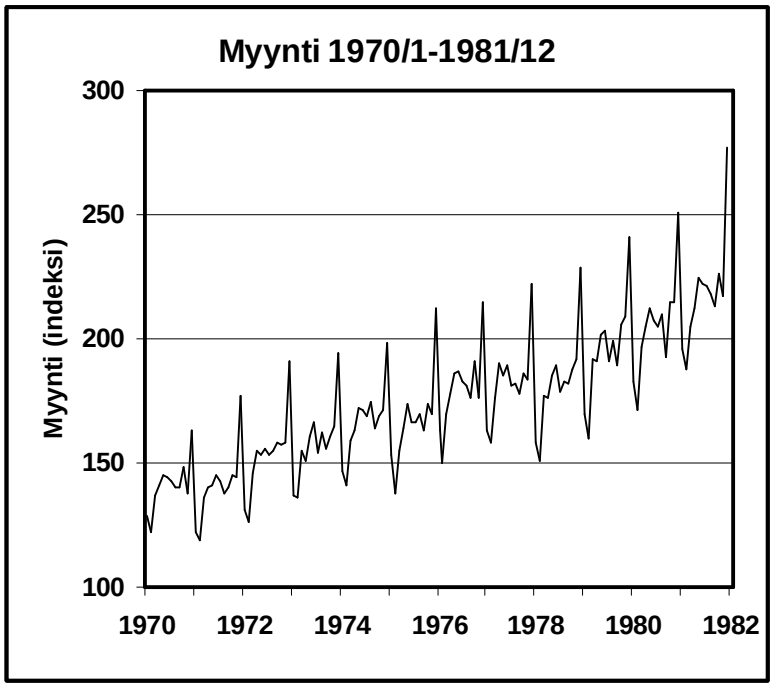
t = aika (1970/1-1981/12)

x_t = kk-myynnin arvoa kuvaava indeksi (1960/1 = 100)

Huomaa, että kk-myynnissä on ollut *nouseva trendi* ja selvää *kausivaihtelua*.

Tällaisten aikasarjojen analysoiminen vaatii menetelmiä, jotka menevät tässä monisteessa käsiteltävän alueen ulkopuolelle.

Aikasarjojen analyysia ja *ennustamista* käsitellään monisteessa **Aikasarja-analyysi**.



Aritmeettiset keskiarvot

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

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

Olkoot

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

ja

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

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

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

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

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

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Havaintoarvojen aritmeettinen keskiarvo kuvaa havaintoarvojen *keskimääräistä sijaintia*. Havainto-arvojen pareista

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

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

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

on havaintoarvojen parien muodostamien pisteiden *painopiste*. Havaintoarvojen aritmeettinen keskiarvo kuvaa havainto-arvojen *keskimääräistä sijaintia*.

Otosvarianssit ja otoskeskihajonnat

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

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

jossa $\overline{x}$ on *x*-havaintoarvojen aritmeettinen keskiarvo ja havaintoarvojen $y_1, y_2, \ldots, y_n$ (otos-) **varianssi** on

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

jossa $\overline{y}$ on *y*-havaintoarvojen aritmeettinen keskiarvo. Havaintoarvojen varianssi mittaa havainto-arvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havaintoarvojen aritmeettisen keskiarvon suhteen.

Havaintoarvojen $x_1, x_2, \ldots, x_n$ **keskihajonta** on

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

jossa $\overline{x}$ on *x*-havaintoarvojen aritmeettinen keskiarvo ja havaintoarvojen $y_1, y_2, \ldots, y_n$ **keski-hajonta** on

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

jossa $\overline{y}$ on *y*-havaintoarvojen aritmeettinen keskiarvo. Havaintoarvojen keskihajonta mittaa (kuten havaintoarvojen otosvarianssi) havaintoarvojen *hajaantuneisuutta* tai *keskittyneisyyttä* havainto-arvojen aritmeettisen keskiarvon suhteen.

Otoskovarianssi

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

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

jossa

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

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

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

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

Otoskovarianssi s_{xy} mittaa x- ja y-havaintoarvojen *yhteisvaihtelua* niiden aritmeettisten keski-arvojen ympärillä. Mitä suurempi on otoskovarianssin s_{xy} itseisarvo

$$| s_{xy} |$$

sitä voimakkaampaa on x- ja y-havaintoarvojen yhteisvaihtelu.

Tarkastellaan seuraavaksi miten otoskovarianssin s_{xy} *merkin määräytymistä*. Merkin määrää se onko summalauseke

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

negatiivinen vai positiivinen.

Todetaan ensin, että summalausekkeen (1) i. termin

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

itseisarvo

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

on sellaisen *suorakaiteen pinta-ala*, jonka sivujen pituudet ovat $| x_i - \bar{x} |$ ja $| y_i - \bar{y} |$.

Summalausekkeen (1) i. termin

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

merkki määräytyy seuraavalla tavalla:

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

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

Otoskovarianssin merkin määräytymistä voidaan *havainnollistaa geometrisesti* seuraavalla tavalla:

(i) Jaetaan xy-taso neljään osaan eli *neljännekseen* pisteen

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

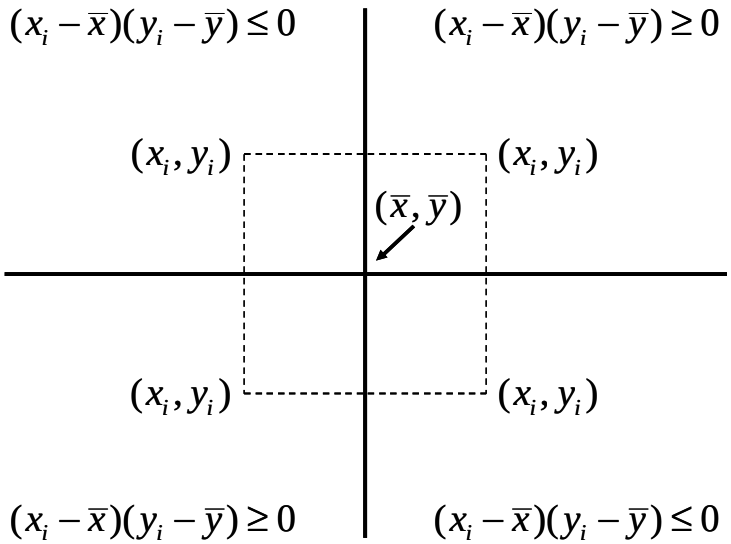
kautta piirretyillä koordinaattiakseleiden suuntaisilla suorilla.

(ii) Termin

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

merkin määrää se, mihin neljännekseen havaintopiste (x_i , y_i) sijoittuu.

Ks. alla olevaa kuvaa:



Jos positiiviset termit summalausekkeeseen

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

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

Tästä seuraa se, että otoskovarianssilla on taipumus saada *positiivisia* (*negatiivisia*) arvoja, jos havaintopisteiden muodostama pistepilvi tai -parvi *näyttää nousevalta* (*laskevalta*) *oikealle mentäessä*; ks. *pistediagrammin* ilmeen ja Pearsonin *otoskorrelaatiokertoimen* yhteyttä havainnollistavaa kuvasarjaa tässä kappaleessa.

Otoskorrelaatio

Otoskovarianssin s_{xy} avulla voidaan määritellä x- ja y-havaintoarvojen *lineaarisen tilastollisen riippuvuuden voimakkuuden mittari*, jota kutsutaan **Pearsonin otoskorrelaatiokertoimeksi**. Pearsonin otoskorrelaatiokerroin r_{xy} saadaan otoskovarianssista s_{xy} *normeerausoperaatiolla*, jossa x- ja y-havaintoarvojen otoskovarianssi s_{xy} jaetaan x- ja y-havaintoarvojen keskihajonnoilla s_x ja s_y .

Havaintoarvojen pareista (x_i , y_i) , i = 1, 2, … , n laskettu **Pearsonin otoskorrelaatiokerroin** on

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

jossa

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

s_x = x-havaintoarvojen keskihajonta

s_y = y-havaintoarvojen keskihajonta

Pearsonin otoskorrelaatiokertoimen kaava voidaan kirjoittaa myös muotoon

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

jossa

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

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

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

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

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

jos ja vain jos

$y_i = \alpha + \beta x_i$, $i = 1, 2, \ldots , n$

jossa α ja $\beta \neq 0$ ovat reaalisia vakioita.

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

Pearsonin otoskorrelaatiokerroin r_{xy} mittaa x- ja y-havaintoarvojen lineaarisen tilastollisen riippuvuuden voimakkuutta:

(i) Jos

$r_{xy} = \pm 1$

niin x- ja y-havaintoarvojen välillä *on eksakti* eli *funktionaalinen lineaarinen riippuvuus*, mikä merkitsee sitä, että kaikki havaintopisteet (x_i , y_i) , $i = 1, 2, \ldots , n$ asettuvat samalle suoralle.

(ii) Jos

$r_{xy} = 0$

niin x- ja y-havaintoarvojen välillä ei voi olla eksaktia lineaarista riippuvuutta.

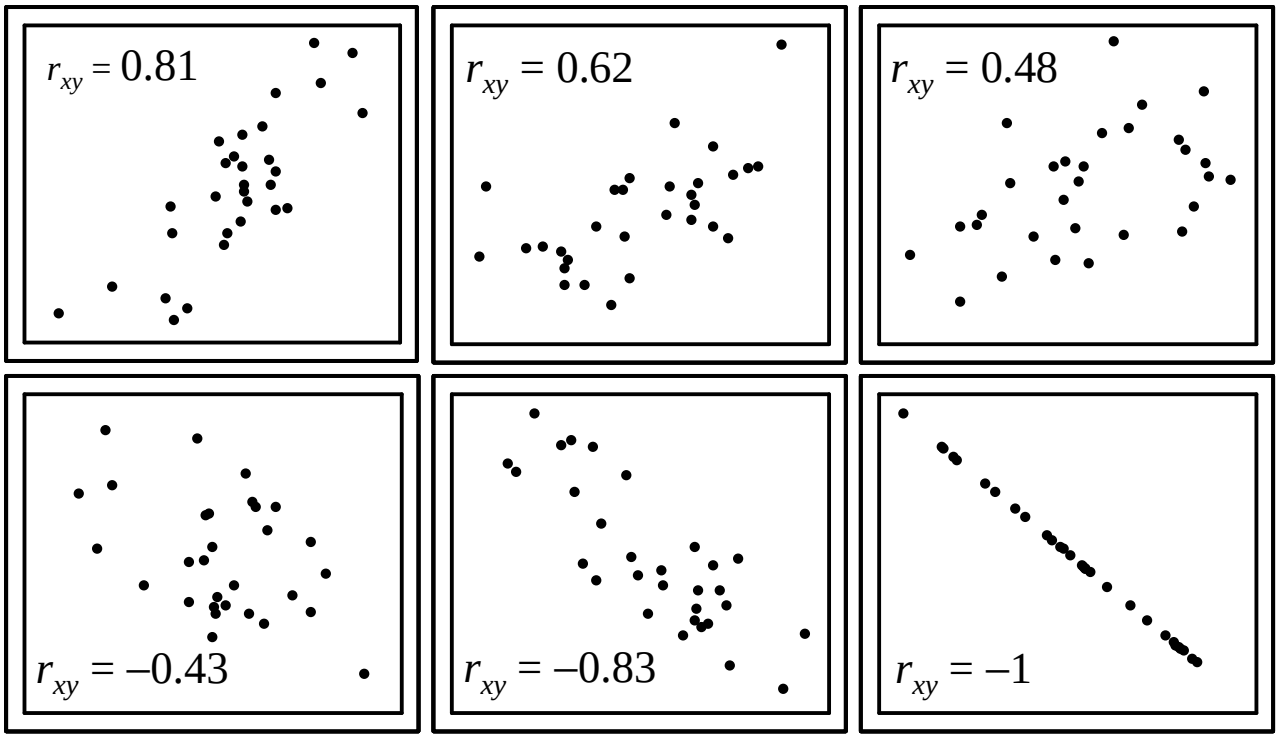
Huomautus:

- Vaikka

  $r_{xy} = 0$

  niin x- ja y-havaintoarvojen välillä saattaa olla jopa eksakti epälineaarinen riippuvuus.

Korrelaatiokertoimen *merkki* ja jopa *suuruusluokka* (jollakin tarkkuudella) voidaan melko helposti oppia arvioimaan pistediagrammin avulla. Alla olevat kuviot havainnollistavat kahden muuttujan havaittujen arvojen ($n = 30$) *pistediagrammin ilmeen* ja *korrelaation* välistä yhteyttä.



Otostunnuslukujen laskeminen

Oletetaan, että haluamme laskea havaintoarvojen pareista

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

seuraavat otostunnusluvut *käsin* tai käyttämällä *laskinta*:

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

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

(iii) *Keskihajonnat*: s_x, s_y

(iv) *Otoskovarianssi*: s_{xy}

(v) *Korrelaaatio*: r_{xy}

Tällöin tarvittavat laskutoimitukset on mukavinta järjestää alla esitettyn kaavion muotoon. 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	$x_2 y_2$
M	M	M	M	M	M
n	x_n	y_n	x_n^2	y_n^2	$x_n y_n$
Summa	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$

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

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

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

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

Havaintoarvojen *keskihajonnat* ja *Pearsonin otoskorrelaatiokerroin* saadaan havaintoarvojen *variansseista* ja *kovarianssista* alla esitetyillä kaavoilla:

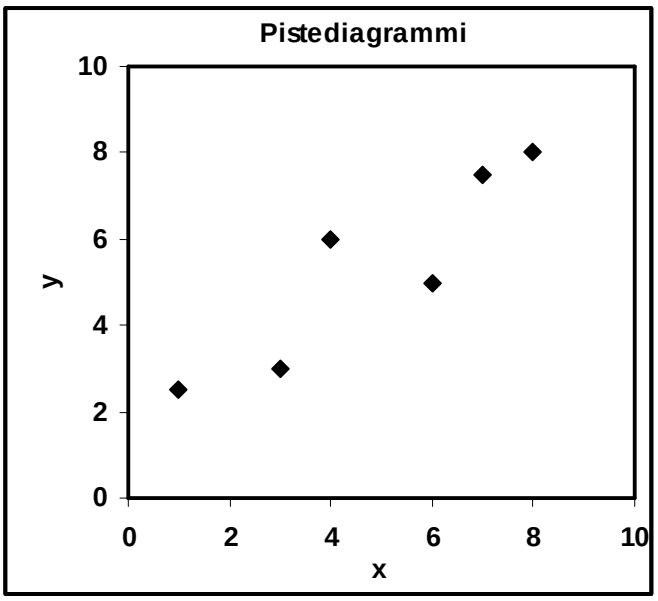
$$s_x = \sqrt{s_x^2} \qquad s_y = \sqrt{s_y^2}$$

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

Esimerkki 6: Otostunnuslukujen laskeminen.

Taulukossa alla on keinotekoisen kahden muuttujan aineiston havaintoarvot (n = 6). Myös aineistoa kuvaava *pistediagrammi* on annettu alla.

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



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

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

Muuttujien *x* ja *y* havaittujen arvojen *aritmeettiset keskiarvot, otosvarianssit, keskihajonnat, otoskovarianssi* ja *otoskorrelaatio* voidaan laskea näistä viidestä summasta.

Aritmeettiset keskiarvot, otosvarianssit ja otoskovarianssi:

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

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

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

$$s_y^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right) = \frac{1}{6-1}\left(196.5 - \frac{1}{6}\times 32^2\right) = 5.167$$

$$s_{xy} = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)\right) = \frac{1}{6-1}\left(182 - \frac{1}{6}\times 29\times 32\right) = 5.467$$

Otoskeskihajonnat ja otoskorrelaatio:

$$s_x = \sqrt{s_x^2} = \sqrt{6.967} = 2.639$$

$$s_y = \sqrt{s_y^2} = \sqrt{5.167} = 2.273$$

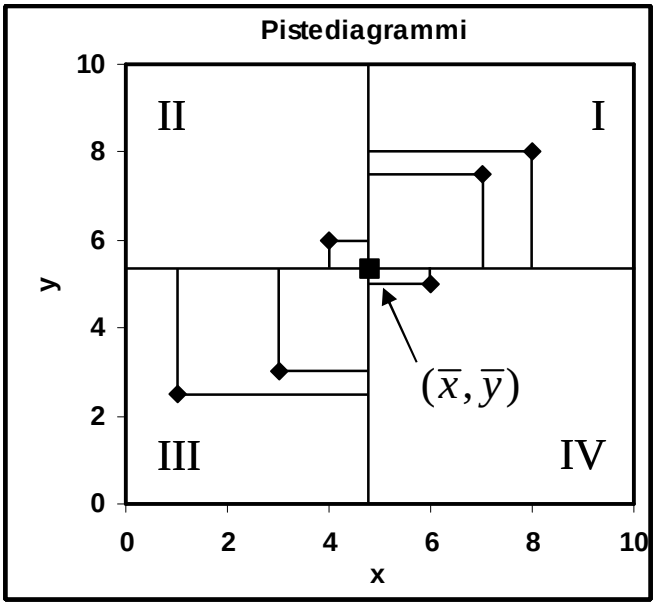
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{5.467}{2.639\times 2.273} = 0.9112$$

Alla on havaintoaineistoa kuvaava pistediagrammi, johon lisätty havaintoarvojen *painopiste*

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

Lisäksi kuvioon on lisätty painopisteen kautta kulkevat koordinaattiakseleiden suuntaiset suorat sekä *kovarianssin* ja *korrelaation merkin määräytymistä* havainnollistavat suorakaiteet; ks. tässä kappaleessa esitettyä selitystä *kovarianssin merkin* määräytymisestä..

Kovarianssi (ja siten myös korrelaatio) on *positiivinen*, koska I ja III neljänneksen suorakaiteiden yhteenlaskettu pinta-ala on selvästi *suurempi* kuin II ja IV neljännek-sen suorakaiteiden yhteenlaskettu pinta-ala.



13.3. Pearsonin korrelaatiokertoimen estimointi ja testaus

Tarkastelemme tässä kappaleessa välimatka- tai suhdeasteikollisten satunnaismuuttujien *x* ja *y* Pearsonin (tulomomentti-) korrelaatiokertoimen ρ_{xy} **estimointia** sekä seuraavia testejä korrelaatiokertoimelle ρ_{xy} :

- **Yhden otoksen testi korrelaatiokertoimelle**
- **Korrelaatiokertoimien vertailutesti**
- **Korreloimattomuuden testaaminen**

Lisätietoja moniulotteisista satunnaismuuttujista ja jakaumista: Ks. kirjan **Todennäköisyyslaskenta** lukuja **Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat** ja **Moniulotteisia jakaumia**.

Otos kaksiulotteisesta normaalijakaumasta

Oletetaan, että satunnaismuuttujien *x* ja *y* muodostama pari (*x* , *y*) noudattaa *kaksiulotteista normaalijakaumaa* parametrein $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}$:

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

Tällöin satunnaismuuttujien *x* ja *y odotusarvot* ovat

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

satunnaismuuttujien *x* ja *y varianssit* ovat

$$\sigma_x^2 = \mathrm{Var}(x) = \mathrm{E}[(x - \mu_x)^2]$$
$$\sigma_y^2 = \mathrm{Var}(y) = \mathrm{E}[(y - \mu_y)^2]$$

satunnaismuuttujien *kovarianssi* on

$$\sigma_{xy} = \mathrm{Cov}(x, y) = \mathrm{E}[(x - \mu_x)(y - \mu_y)]$$

ja satunnaismuuttujien *korrelaatio* on

$$\rho_{xy} = \mathrm{Cor}(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Korrelaatiota ρ_{xy} kutsutaan tavallisesti **Pearsonin** (**tulomomentti-**) **korrelaatiokertoimeksi** ja se mittaa *satunnaismuuttujien x ja y lineaarisen riippuvuuden voimakkuutta*.

Olkoot

$$y_1, y_2, \mathsf{K}, y_n$$

muuttujan *y* havaitut arvot ja

$$x_1, x_2, \mathsf{K}, x_n$$

muuttujan *x* havaitut arvot ja oletetaan, että havaintoarvojen x_i ja y_i parit

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

muodostavat satunnaisotoksen kaksiulotteista normaalijakaumasta

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

Tällöin

$$(x_1, y_1), (x_2, y_2), \mathrm{K}, (x_n, y_n) \perp$$
$$(x_i, y_i) \quad \mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy}), \; i = 1, 2, \mathrm{K}, n$$

Kaksiulotteisen normaalijakauman parametrien estimointi

Kaksiulotteisen normaalijakauman parametrien suurimman uskottavuuden estimaattorit tai momenttiestimaattorit ovat

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

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

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

$$\hat{\rho}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} = \frac{s_{xy}}{s_x s_y} = r_{xy}$$

jossa

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

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

Otostunnusluvut

$$\hat{\mu}_x = \bar{x} \text{ ja } \hat{\mu}_y = \bar{y}$$

ovat *x*- ja *y*-havaintojen *aritmeettiset keskiarvot*,

$$\hat{\sigma}_x^2 = ((n-1)/n)s_x^2 \text{ ja } \hat{\sigma}_y^2 = ((n-1)/n)s_y^2$$

ovat *x*- ja *y*-havaintojen *otosvarianssit*,

$$\hat{\sigma}_{xy} = ((n-1)/n)s_{xy}$$

on *x*- ja *y*-havaintojen *otoskovarianssi* ja

$$\hat{\rho}_{xy} = r_{xy}$$

on *x*- ja *y*-havaintojen *Pearsonin otoskorrelaatiokerroin*. Lisäksi

$$s_x^2 \text{ ja } s_y^2$$

ovat *x*- ja *y*-havaintojen *harhattomat otosvarianssit*.

Fisherin *z*-muunnos

Määritellään **Fisherin *z*-muunnos** kaavalla

$$z = f(u) = \frac{1}{2}\log\left(\frac{1+u}{1-u}\right)$$

Sovelletaan Fisherin *z*-muunnosta $z = f(u)$ otoskorrelaatiokertoimeen r_{xy}:

$$z_r = f(r_{xy}) = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

Satunnaismuuttuja z_r noudattaa suurissa otoksissa approksimatiivisesti normaalijakaumaa:

$$z_r \;\; {}_a \;\; \mathrm{N}\left(\mu_z, \sigma_z^2\right)$$

jossa

$$\mu_z = f(\rho_{xy}) = \frac{1}{2}\log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Approksimaatio on käytännössä riittävän hyvä, jos $n > 25$.

Soveltamalla Fisherin *z*-muunnosta **luottamusvälit** ja **testit Pearsonin korrelaatiokertoimelle** ρ_{XY} voidaan konstruoida *samalla tekniikalla* kuin luottamusvälit ja testit *normaalijakauman odotusarvolle*; ks. lukuja **Väliestimointi** ja **Testejä suhdeasteikollisille muuttujille**.

Korrelaatiokertoimen luottamusväli

Oletetaan, että satunnaismuuttujien *x* ja *y* muodostama järjestetty pari (*x* , *y*) noudattaa *kaksiulotteista normaalijakaumaa*

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

Konstruoidaan Pearsonin korrelaatiokertoimelle ρ_{XY} *approksimatiivinen luottamusväli Fisherin z-muunnoksen* avulla.

Olkoon r_{xy} satunnaisotoksesta (x_i, y_i) , $i = 1, 2, \ldots, n$ määrätty *Pearsonin otoskorrelaatiokerroin*.

Valitaan *luottamustasoksi*

$$1 - \alpha$$

Luottamustason valinta kiinnittää todennäköisyyden, jolla konstruoitava luottamusväli peittää Pearsonin korrelaatiokertoimen ρ_{XY} oikean arvon.

Määrätään piste

$$+z_{\alpha/2}$$

siten, että se erottaa standardoidun normaalijakauman N(0,1) oikealle hännälle todennäköisyysmassan $\alpha/2$. Koska normaalijakauma on symmetrinen, piste

$$-z_{\alpha/2}$$

erottaa standardoidun normaalijakauman vasemmalle hännälle todennäköisyysmassan $\alpha/2$. Siten *luottamuskertoimet* $+z_{\alpha/2}$ ja $-z_{\alpha/2}$ on määrätään siten, että

$$\Pr(Z \geq +z_{\alpha/2}) = \frac{\alpha}{2}$$

$$\Pr(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2}$$

jossa satunnaismuuttuja Z noudattaa *standardoitua normaalijakaumaa* N(0,1):

$$Z \quad \mathrm{N}(0,1)$$

Huomaa, että luottamuskertoimet $+z_{\alpha/2}$ ja $-z_{\alpha/2}$ toteuttavat ehdon

$$\Pr(-z_{\alpha/2} \leq Z \leq +z_{\alpha/2}) = 1-\alpha$$

Sovelletaan *Fisherin z-muunnosta* Pearsonin otoskorrelaatiokertoimeen r_{xy}:

$$z_r = f(r_{xy}) = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

Edellä esitetyn nojalla

$$z_r \quad {}_a \ \mathrm{N}\left(\mu_z, \sigma_z^2\right)$$

jossa

$$\mu_z = f(\rho_{xy}) = \frac{1}{2}\log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Parametrin

$$\mu_z = \frac{1}{2}\log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)$$

(approksimatiivinen) **luottamusväli luottamustasolla $(1-\alpha)$** on siten muotoa

$$\left(z_r - z_{\alpha/2}\frac{1}{\sqrt{n-3}}, z_r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right)$$

Parametrin μ_z (approksimatiivisen) luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(z_r - z_{\alpha/2}\frac{1}{\sqrt{n-3}} \leq \mu_z \leq z_r + z_{\alpha/2}\frac{1}{\sqrt{n-3}}\right) =_a 1-\alpha$$

Konstruoitu luottamusväli *peittää* parametrin μ_z oikean arvon (approksimatiivisesti) todennäköisyydellä $(1-\alpha)$ ja se *ei peitä* parametrin μ_z oikeata arvoa (approksimatiivisesti) todennäköisyydellä α.

Pearsonin korrelaatiokertoimen ρ_{XY} (approksimatiivinen) *luottamusväli luottamustasolla* $(1-\alpha)$ saadaan parametrin μ_Z luottamusvälistä ratkaisemalla ρ_{XY} epäyhtälöketjusta

$$z_r - z_{\alpha/2}\frac{1}{\sqrt{n-3}} = \frac{1}{2}\log\frac{1+r_{xy}}{1-r_{xy}} - z_{\alpha/2}\frac{1}{\sqrt{n-3}}$$
$$\leq \mu_z = \frac{1}{2}\log\frac{1+\rho_{xy}}{1-\rho_{xy}}$$
$$\leq z_r + z_{\alpha/2}\frac{1}{\sqrt{n-3}} = \frac{1}{2}\log\frac{1+r_{xy}}{1-r_{xy}} + z_{\alpha/2}\frac{1}{\sqrt{n-3}}$$

Siten **Pearsonin korrelaatiokertoimen** ρ_{xy} (approksimatiiviseksi) **luottamusväliksi luottamustasolla** $(1-\alpha)$ saadaan

$$(lb,\ ub)$$

jossa

$$lb = \frac{(1+r_{xy}) - (1-r_{xy})\exp\left(+2z_{\alpha/2}/\sqrt{n-3}\right)}{(1+r_{xy}) + (1-r_{xy})\exp\left(+2z_{\alpha/2}/\sqrt{n-3}\right)}$$

on luottamusvälin *alaraja* ja

$$ub = \frac{(1+r_{xy}) - (1-r_{xy})\exp\left(-2z_{\alpha/2}/\sqrt{n-3}\right)}{(1+r_{xy}) + (1-r_{xy})\exp\left(-2z_{\alpha/2}/\sqrt{n-3}\right)}$$

on luottamusvälin *yläraja*. Luottamusvälin konstruktiosta seuraa, että

$$\Pr\left(lb \leq \rho_{XY} \leq ub\right) =_a 1-\alpha$$

Siten konstruoitu luottamusväli *peittää* korrelaatiokertoimen ρ_{xy} oikean arvon (approksimatiivisesti) todennäköisyydellä $(1-\alpha)$ ja se *ei peitä* korrelaatiokertoimen ρ_{xy} oikeata arvoa (approksimatiivisesti) todennäköisyydellä α.

Korreloimattomuuden testaaminen

Monissa tutkimusasetelmissa ollaan kiinnostuneita siitä, ovatko satunnaismuuttujat x ja y **korreloituneita** vai **korreloimattomia**.

Yleinen hypoteesi H :

Havaintoarvojen x_i ja y_i parit

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

muodostavat satunnaisotoksen kaksiulotteista normaalijakaumasta

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

Nollahypoteesi H_0 :

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

Vaihtoehtoinen hypoteesi H_1 :

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

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

Olkoon r_{xy} otoksesta (x_i,y_i) , $i = 1, 2, \ldots, n$ määrätty *Pearsonin otoskorrelaatiokerroin*.

Määritellään ***t*-testisuure**

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

Jos nollahypoteesi

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

pätee, testisuure *t noudattaa t-jakaumaa* vapausastein $(n - 2)$:

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

Testisuureen *t normaaliarvo* = 0, koska nollahypoteesin pätiessä

$$\mathrm{E}(t) = 0$$

Siten *itseisarvoltaan suuret testisuureen t arvot viittaavat siihen, että nollahypoteesi* H_0 *ei päde*. Testin hylkäysalueen tai *p*-arvon määrääminen: ks. lukua **Tilastolliset testit**. Jos testin *p*-arvo on kyllin pieni, nollahypoteesi H_0 hylätään.

Huomautuksia:

- Satunnaismuuttujien *x* ja *y* **riippumattomuudesta *seuraa aina* niiden korreloimattomuus**.
- Satunnaismuuttujien *x* ja *y* **korreloimattomuudesta *ei yleisesti seuraa* niiden riippumattomuus**.
- Jos satunnaismuuttujat *x* ja *y* noudattavat **2-ulotteista normaalijakaumaa**, satunnaismuuttujien *x* ja *y* **korreloimattomuudesta seuraa niiden riippumattomuus**.
- Monissa tutkimusasetelmissa *toivotaan, että korreloimattomuusoletus tulee testissä hylätyksi*.

Yleinen testi korrelaatiokertoimelle

Tarkastellaan *yleistä testiä* korrelaatiokertoimelle.

Yleinen hypoteesi H :

Oletetaan, että havaintoarvojen x_i ja y_i parit

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

muodostavat satunnaisotoksen kaksiulotteista normaalijakaumasta

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

Nollahypoteesi H_0 :

$$H_0 : \rho_{xy} = \rho_0$$

Vaihtoehtoinen hypoteesi H_1 :

$$\left.\begin{array}{l} H_1 : \rho_{xy} > \rho_0 \\ H_1 : \rho_{xy} < \rho_0 \end{array}\right\} \text{ 1-suuntaiset vaihtoehtoiset hypoteesit}$$

$$H_1 : \rho_{xy} \neq \rho_0 \quad \text{2-suuntainen vaihtoehtoinen hypoteesi}$$

Olkoon r_{xy} otoksesta (x_i,y_i) , $i = 1, 2, \ldots, n$ määrätty *Pearsonin otoskorrelaatiokerroin*.

Sovelletaan *Fisherin z-muunnosta* otoskorrelaatiokertoimeen r_{xy} :

$$z_r = f(r_{xy}) = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

Satunnaismuuttuja z_r noudattaa suurissa otoksissa approksimatiivisesti normaalijakaumaa:

$$z \;\; {}_a \;\; N\left(\mu_z, \sigma_z^2\right)$$

jossa

$$\mu_z = f(\rho_{xy}) = \frac{1}{2}\log\left(\frac{1+\rho_{xy}}{1-\rho_{xy}}\right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Approksimaatio on käytännössä riittävän hyvä, jos $n > 25$.

Muodostetaan **testisuure**

$$v = \frac{z-\mu_z^0}{\sigma_z}$$

jossa siis

$$z_r = f(r_{xy}) = \frac{1}{2}\log\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$$

$$\mu_z^0 = f(\rho_0) = \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

$$\sigma_z^2 = \frac{1}{n-3}$$

Jos nollahypoteesi

$$H_0 : \rho_{xy} = \rho_0$$

pätee, testisuure v noudattaa *suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa* N(0,1):

$$v \;\; {}_a \;\; N(0,1)$$

Testisuureen v *normaaliarvo* = 0, koska nollahypoteesin pätiessä

$$E(v) = 0$$

Siten *itseisarvoltaan suuret testisuureen v arvot viittaavat siihen, että nollahypoteesi* H_0 *ei päde*. Testin hylkäysalueen tai *p*-arvon määrääminen: ks. lukua **Tilastolliset testit**. Jos testin *p*-arvo on kyllin pieni, nollahypoteesi H_0 hylätään.

Korrelaatiokertoimien vertailutesti

Tarkastellaan korrelaatiokertoimien *vertailutestiä*.

Yleinen hypoteesi H :

Oletetaan, että käytössä on kaksi toisistaan riippumatonta satunnaisotosta kaksiulotteisista normaalijakaumista, joiden korrelaatiokertoimet ovat ρ_1 ja ρ_2 .

Nollahypoteesi H_0 :

$$H_0 : \rho_1 = \rho_2 = \rho_0$$

Vaihtoehtoinen hypoteesi H_1 :

$$\left.\begin{array}{l} H_1 : \rho_1 > \rho_2 \\ H_1 : \rho_1 < \rho_2 \end{array}\right\} \quad \text{1-suuntaiset vaihtoehtoiset hypoteesit}$$

$$H_1 : \rho_1 \neq \rho_2 \quad \text{2-suuntainen vaihtoehtoinen hypoteesi}$$

Olkoot

$$n_1 \text{ ja } n_2$$

otoskoot otoksissa 1 ja 2 sekä

$$r_1 \text{ ja } r_2$$

otoksista 1 ja 2 määrätyt *Pearsonin otoskorrelaatiokertoimet*.

Sovelletaan Fisherin *z*-muunnosta otoskorrelaatiokertoimiin r_1 ja r_2 :

$$z_k = f(r_k) = \frac{1}{2}\log\left(\frac{1+r_k}{1-r_k}\right), k = 1,2$$

Jos nollahypoteesi

$$H_0 : \rho_1 = \rho_2 = \rho_0$$

pätee satunnaismuuttujat z_k noudattavat *suurissa otoksissa approksimatiivisesti normaalijakaumaa*:

$$z_k \;\; {}_a \;\; N(\mu_z^0, \sigma_k^2), k = 1,2$$

jossa

$$\mu_z^0 = f(\rho_0) = \frac{1}{2}\log\left(\frac{1+\rho_0}{1-\rho_0}\right)$$

$$\sigma_k^2 = \frac{1}{n_k - 3}, k = 1,2$$

Approksimaatio on käytännössä riittävän hyvä, jos $n_1 > 25$ ja $n_2 > 25$.

Muodostetaan **testisuure**

$$\nu = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

jossa siis

$$z_k = f(r_k) = \frac{1}{2}\log\left(\frac{1+r_k}{1-r_k}\right), k = 1,2$$

Jos nollahypoteesi

$$\mathrm{H}_0 : \rho_1 = \rho_2 = \rho_0$$

pätee, testisuure ν noudattaa *suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa* N(0,1):

$$\nu \;\; _a \; \mathrm{N}(0,1)$$

Testisuureen ν *normaaliarvo* = 0, koska nollahypoteesin pätiessä

$$\mathrm{E}(\nu) = 0$$

Siten *itseisarvoltaan suuret testisuureen* ν *arvot viittaavat siihen, että nollahypoteesi* H_0 *ei päde*. Testin hylkäysalueen tai *p*-arvon määrääminen: ks. lukua **Tilastolliset testit**. Jos testin *p*-arvo on kyllin pieni, nollahypoteesi H_0 hylätään.

13.4. Järjestyskorrelaatiokertoimet

Tarkastellaan korrelaatiokertoimen määrittelemistä ja korreloimattomuuden testaamista *järjestysasteikollisille muuttujille.* Tarkastelun kohteena ovat seuraavat **järjestyskorrelaatiokertoimet**:

- **Spearmanin järjestyskorrelaatiokerroin**
- **Kendallin järjestyskorrelaatiokerroin**

Tarkasteltavat järjestyskorrelaatiokertoimet ja testit korreloimattomuudelle sopivat myös *välimatka-* ja *suhdeasteikollisille muuttujille*.

Spearmanin järjestyskorrelaatiokerroin

Spearmanin järjestyskorrelaatiokerroin ρ_S mittaa kahden muuttujan havaintoarvojen suuruusjärjestyksien yhteensopivuutta. Spearmanin järjestyskorrelaatiokerroin sopii *järjestys-*, *välimatka-* ja *suhdeasteikollisille* muuttujille. Spearmanin järjestyskorrelaatiokertoimella on samantapaiset ominaisuudet kuin *Pearsonin otoskorrelaatiokertoimella*.

Olkoot

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

ja

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

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

Järjestetään sekä *x*- että *y*-muuttujan havaitut arvot *suuruusjärjestykseen* pienimmästä suurimpaan. Liitetään sekä *x*- että *y*-muuttujan havaittuihin arvoihin niiden *suuruusjärjestyksien* mukaiset järjestysnumerot:

$R(x_i)$ = havainnon x_i järjestysnumero parissa i

$R(y_i)$ = havainnon y_i järjestysnumero parissa i

sekä määritellään erotukset

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

Muuttujien *x* ja *y* havaituille arvoille voidaan määritellä *järjestyskorrelaatiokerroin* erotuksien D_i avulla.

Määritellään **Spearmanin järjestyskorrelaatiokerroin** ρ_S eli *Spearmanin rho* kaavalla

$$\rho_S = 1 - \frac{6\sum_{i=1}^{n} D_i^2}{n^3 - n}$$

Spearmanin järjestyskorrelaatiokerroin ρ_S voidaan laskea myös soveltamalla *Pearsonin otoskorrelaatiokertoimen* kaavaa muuttujien *x* ja *y* havaittujen arvojen pareja (x_i , y_i) vastaaviin järjestyslukujen eli *rankien* pareihin

$$(R(x_i), R(y_i))$$

Spearmanin järjestyskorrelaatiokertoimen ominaisuudet

Spearmanin järjestyskorrelaatiokertoimella ρ_S on kaikki *hyvältä korrelaation mitalta vaadittavat ominaisuudet*:

(i) $-1 \le \rho_S \le +1$

(ii) Jos muuttujien *x* ja *y* havaittujen arvojen järjestysnumerot ovat jokaisessa havaintoparissa *samat*, niin

$$\rho_S = +1$$

(iii) Jos muuttujien *x* ja *y* havaittujen arvojen järjestysnumerot liittyvät toisiinsa *täysin satunnaisesti*,

$$\rho_S \approx 0$$

Jos $\rho_S = 0$, sanomme, että muuttujat *x* ja *y* ovat *korreloimattomia*.

(iv) Jos sekä *suuret* muuttujien *x* ja *y* järjestysnumerot että *pienet* muuttujien *x* ja *y* järjestysnumerot liittyvät havaintopareissa (x_i , y_i) toisiinsa, kertoimella ρ_S on taipumus saada *positiivisia* arvoja.

(v) Jos *suuret* ja *pienet* muuttujien *x* ja *y* järjestysnumerot liittyvät havaintopareissa (x_i , y_i) toisiinsa, kertoimella ρ_S on taipumus saada *negatiivisia* arvoja.

Korreloimattomuuden testaaminen

Määritellään ***t*-testisuure**

$$z = \sqrt{n-2}\,\frac{\rho_S}{\sqrt{1-\rho_S^2}}$$

Jos nollahypoteesi

$$\mathrm{H}_0 : \mathrm{Cor}(x, y) = 0$$

pätee, testisuure z noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa N(0,1):

$$z \;\; {}_{a} \;\; \mathrm{N}(0,1)$$

Approksimaatio on melko hyvä jo, kun $n > 10$ ja riittävä lähes kaikkiin tarkoituksiin, kun $n > 30$.

Testisuureen *z normaaliarvo* = 0, koska nollahypoteesin H_0 pätiessä

$$\mathrm{E}(z) = 0$$

Siten *itseisarvoltaan suuret testisuureen z arvot viittaavat siihen, että nollahypoteesi* H_0 *ei päde*. Testin hylkäysalueen tai *p*-arvon määrääminen: ks. lukua **Tilastolliset testit**. Jos testin *p*-arvo on kyllin pieni, nollahypoteesi H_0 hylätään.

Kendallin järjestyskorrelaatiokerroin

Kendallin järjestyskorrelaatiokerroin τ mittaa kahden muuttujan havaintoarvojen suuruusjärjestyksien yhteensopivuutta. Kendallin järjestyskorrelaatiokerroin sopii *järjestys-*, *välimatka-* ja *suhdeasteikollisille* muuttujille. Kendallin järjestyskorrelaatiokertoimella on samantapaiset ominaisuudet kuin *Pearsonin otoskorrelaatiokertoimella*.

Olkoot

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

ja

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

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

Järjestetään lukuparit (x_i , y_i) muuttujan *x* havaittujen arvojen mukaan *suuruusjärjestykseen* pienimmästä suurimpaan siten, että ensimmäiseksi tulee pari, jossa muuttujan *x* arvo on pienin ja viimeiseksi pari, jossa muuttujan *x* arvo on suurin. Kendallin järjestyskorrelaatiokerroin perustuu tunnuslukuun, joka mittaa muuttujan *y* arvojen *epäjärjestystä* muuttujan *x* arvoihin nähden.

Olkoon (x_k , y_k) järjestetykseen asetetuista pareista numero *k*. Määritellään havaintoarvoon y_k liittyvät *epäjärjestyspisteet*

$$S_{kl} , \; l = k+1, k+2, \ldots, n , \; k = 1, 2, \ldots, n-1$$

seuraavalla tavalla:

$S_{kl} = +1$, jos $y_l > y_k$

$S_{kl} = -1$, jos $y_l < y_k$

Muuttujan *y* arvojen *epäjärjestysmitta S* muuttujan *x* arvojen suhteen määritellään kaavalla

$$S = \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} S_{kl}$$

Määritellään **Kendallin järjestyskorrelaatiokerroin** τ eli *Kendallin tau* kaavalla

$$\tau = \frac{S}{\frac{1}{2} n(n-1)}$$

Kendallin järjestyskorrelaatiokertoimen omininaisuudet

Kendallin järjestyskorrelaatiokertoimella τ on kaikki *hyvältä korrelaation mitalta vaadittavat ominaisuudet*:

(i) $-1 \leq \tau \leq +1$

(ii) Jos muuttujien *x* ja *y* havaittujen arvojen järjestysnumerot ovat jokaisessa havaintoparissa *samat*, niin

$$\tau = +1$$

(iii) Jos muuttujien *x* ja *y* havaittujen arvojen järjestysnumerot liittyvät toisiinsa *täysin satunnaisesti*,

$$\tau \approx 0$$

Jos $\tau = 0$, sanotaan, että muuttujat *x* ja *y* ovat *korreloimattomia*.

(iv) Jos sekä *suuret* muuttujien *x* ja *y* järjestysnumerot että *pienet* muuttujien *x* ja *y* järjestysnumerot liittyvät havaintopareissa (x_i , y_i) toisiinsa, kertoimella τ on taipumus saada *positiivisia* arvoja.

(v) Jos *suuret* ja *pienet* muuttujien *x* ja *y* järjestysnumerot liittyvät havaintopareissa (x_i , y_i) toisiinsa, kertoimella τ on taipumus saada *negatiivisia* arvoja.

Korreloimattomuuden testaaminen

Määritellään **testisuure**

$$z = \frac{\tau}{\sqrt{\frac{2(2n+5)}{9n(n+1)}}}$$

Jos nollahypoteesi

$$H_0 : \mathrm{Cor}(x, y) = 0$$

pätee, testisuure z noudattaa suurissa otoksissa approksimatiivisesti standardoitua normaalijakaumaa N(0,1):

$$z \;\; {}_{a} \;\; N(0,1)$$

Approksimaatio on melko hyvä jo, kun $n > 10$ ja riittävä lähes kaikkiin tarkoituksiin, kun $n > 30$.

Testisuureen *z normaaliarvo* = 0, koska nollahypoteesin H_0 pätiessä

$$E(z) = 0$$

Siten *itseisarvoltaan suuret testisuureen z arvot viittaavat siihen, että nollahypoteesi* H_0 *ei päde*. Testin hylkäysalueen tai *p*-arvon määrääminen: ks. lukua **Tilastolliset testit**. Jos testin *p*-arvo on kyllin pieni, nollahypoteesi H_0 hylätään.

14. Johdatus regressioanalyysiin

Regressioanalyysi on (erilaisine muunnelmineen ja johdannaisineen) ehkä *eniten sovellettu tilastotieteen menetelmä*.

Regressioanalyysin avulla voidaan analysoida *jonkin tekijän* tai *muuttujan* **riippuvuutta** *toisista tekijöistä* tai *muuttujista*, kun riippuvuus *ei ole* **eksaktia** vaan **tilastollista**. Tämä tapahtuu rakentamalla riippuvuutta kuvamaan **regressiomalliksi** kutsuttu *tilastollinen malli*. Regressiomalli pyrkii **selittämään** jonkin **selitettävän tekijän** tai **muuttujan havaittujen arvojen vaihtelun** joidenkin **selittävien tekijöiden** tai **muuttujien havaittujen arvojen vaihtelun avulla**.

Tarkastelemme tässä luvussa **regressioanalyysin lähtökohtia**, **tavoitteita** ja **tehtäviä**. Pyrimme perustelemaan myös sen, miksi tässä monisteessa rajoitutaan käsittelemään vain **lineaarisia regressiomalleja**.

Avainsanat:

Approksimointi. Deterministinen malli, Ehdollinen jakauma, Ehdollinen odotusarvo, Ehdollinen varianssi, Ei-satunnaisuus, Ennustaminen, Ennustevirhe, Epälineaarinen regressiomalli, Epälineaarisuus, Estimointi, Jäännöstermi, Kaksiulotteinen normaalijakauma, Keskineliövirhe, Lineaarinen regressiomalli, Linearisointi, Lineaarisuus, Malli, Mallin hyvyys, Minimointi, Multinormaalijakauma, Oletus, Otos, Parametri, Pienimmän neliösumman menetelmä, Rakenneosa, Regressioanalyysi,
Regressiodiagnostiikka, Regressiofunktio, Regressiomalli, Regressiosuora, Reunajakauma, Satunnainen osa, Satunnaisuus, Selitettävä muuttuja, Selittäjä, Selittäminen, Selittävä muuttuja, Systemaattinen osa, Testi, Tilastollinen malli, Tilastollinen riippuvuus, Virhetermi, Yhteisjakauma

14.1. Regressioanalyysin lähtökohdat ja tavoitteet

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.

Regressioanalyysissa selitettävän muuttujan riippuvuudelle selittävistä muuttujista pyritään rakentamaan **regressiomalliksi** kutsuttu *tilastollinen malli*. Koska riippuvuuksien analysointi on kaiken tieteellisen tutkimuksen keskeinen tavoite, *regressioanalyysi on eniten sovellettuja ja tärkeimpiä tilastotieteen menetelmiä*.

Regressioanalyysin tavoitteet

Regressioanalyysin mahdollisia tavoitteita:

(i) Selitettävän muuttujan ja selittävien muuttujien tilastollisen riippuvuuden luonteen **kuvaaminen**:

- Millainen on riippuvuuden (matemaattinen) *muoto*?
- Kuinka *voimakasta* riippuvuus on?

(ii) Selitettävän muuttujan ja selittävien muuttujien tilastollisen riippuvuuden luonteen **selittäminen**.

(iii) Selitettävän muuttujan arvojen **ennustaminen** selittävien muuttujien arvojen avulla.

(iv) Selitettävän muuttujan arvojen **kontrolli** kontrolloimalla selittävien muuttujien arvoja.

Regressiomallien luokittelu

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**

Tässä monisteessa käsitellään ainoastaan *lineaarisia yhden yhtälön regressiomalleja*; ks. lukuja **Yhden selittäjän lineaarinen regressiomalli** ja **Yleinen lineaarinen malli**. Tämä ei kuitenkaan ole kovin vakava rajoitus, koska lineaaristen yhden yhtälön regressiomallien sovellusalue on niinkin laaja kuin se on. Lisäksi lineaaristen regressiomallien teorian hyvä hallinta tekee mahdolliseksi epälineaarisiin regressiomalleihin ja moniyhtälömalleihin liittyvien erityisongelmien ymmärtämisen melko helposti.

On hyödyllistä tietää, että **varianssianalyysissa** sovellettavat tilastolliset mallit voidaan ymmärtää yleisen lineaarisen mallin erikoistapauksiksi; ks. lukuja **Yksisuuntainen varianssianalyysi**, **Kaksisuuntainen varianssianalyysi** ja **Kolmi- ja useampisuuntainen varianssianalyysi**.

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 voidaan ajatella olevan kaksi erilaista lähtökohtaa, joilla on kuitenkin myös 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**.

Käsittelemme vuorollaan kumpaakin lähtökohtaa.

14.2. Deterministiset mallit ja regressioanalyysi

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 riippuvuutta selittävien muuttujien arvoista* **deterministisen mallin** *avulla*.

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 nyt, että parametrille ei voida löytää sellaista arvoa, joka saisi mallin sopimaan *samanaikaisesti* kaikkiin havaintoihin. Voidaanko parametrille kuitenkin löytää sellainen arvo, joka saisi mallin sopimaan havaintoihin jossakin mielessä *niin hyvin kuin se on mahdollista*?

Deterministiset mallit

Oletetaan, että selitettävän muuttujan *y eksaktia* (tai *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

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

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

Oletetaan nyt, että funktion *f* muodon määräävän parametrin β arvo on *tuntematon* ja että haluamme löytää parametrille β *parhaan mahdollisen havaintoihin perustuvan estimaatin* eli *arvion*.
Regressio-ongelma syntyy deterministisen mallien yhteydessä tilanteissa, joissa *para-metrille* β *ei voida löytää sellaista arvoa, joka saisi yhtälön* $y = f(x;\beta)$ *toteutumaan saman-aikaisesti kaikille havainnoille*.

Oletetaan, että muuttujia *x* ja *y* koskevat havainnot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$. Oletamme siis, 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

$$y_i = f(x_i;\beta) + \varepsilon_i \text{ , } i = 1, 2, \mathrm{K}, n$$

jossa ε_i on *havaintoyksiköstä toiseen satunnaisesti vaihteleva* **jäännös**- eli **virhetermi**. Koska olemme olettaneet, että jäännöstermi ε_i on *satunnainen*, myös selitettävän muuttujan *y* havaittujen arvojen y_i on välttämättä oltava satunnaisia.

Yhtälö

$$y_i = f(x_i;\beta) + \varepsilon_i \text{ , } i = 1, 2, \mathrm{K}, n$$

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

Regressioanalyysissa parametrin β arvo pyritään valitsemaan tavalla, joka tekee *kaikista jäännöstermeistä* ε_j *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

$$(x_i, y_i) \in \quad ^2 \text{ , } i = 1, 2, \mathrm{K}, n$$

Erään ratkaisun tähän käyränsovitusongelmaan tarjoaa **pienimmän neliösumman menetelmä**. Siinä parametrin β arvo valitaan siten, että *jäännös- eli virhetermien* ε_i *neliösumma*

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

tulee mahdollisimman pieneksi.

Syyt regressio-ongelman syntyyn

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.

Esimerkki 1: Hooken laki.

Hooken lain mukaan kierrejousen (ns. *ideaalijousen*) pituus *y* riippuu *lineaarisesti* jouseen ripustetusta painosta *x*:

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

jossa

α = jousen pituus ilman painoa

β = ns. *jousivakio*

Alla olevassa taulukossa esitetään tulokset kokeesta, jossa Hooken lain pätevyyttä tutkittiin mittaamalla jousen pituus ilman painoa sekä painoilla, jotka olivat 2, 4, 6, 8 ja 10 kg.

Merkitään:

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

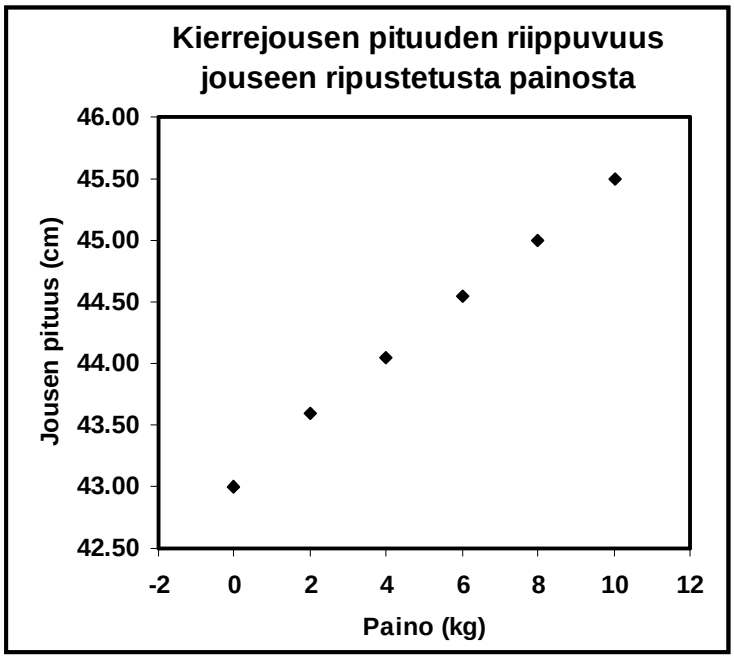
jossa

x_i = paino i

y_i = jousen pituus, kun painona on x_i

Alla oleva pistediagrammi havainnollistaa koetuloksia graafisesti.

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

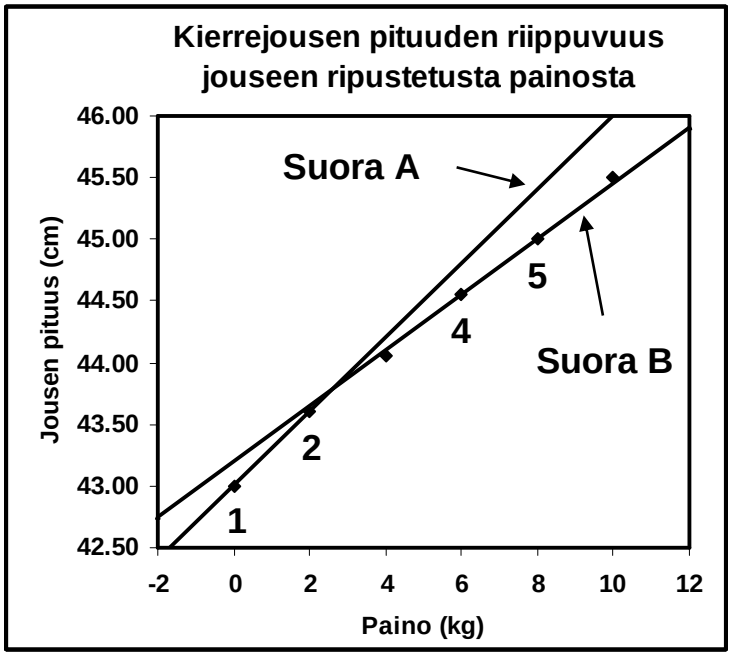


Ovatko havaintotulokset *sopusoinnussa* Hooken lain kanssa? Tämä kysymys voidaan muotoilla teknisemmin seuraavalla tavalla: Onko olemassa *yksikäsitteinen* suora, joka kulkisi *kaikkien* havaintopisteiden kautta?

Kuvio oikealla todistaa, että tällaista suoraa *ei ole olemassa*:

(i) Suora A kulkee pisteiden 1 ja 2 kautta.

(ii) Suora B kulkee pisteiden 4 ja 5 kautta.



Jos ei ole olemassa *yhtä suoraa*, joka kulkee kaikkien havaintopisteiden kautta, olisiko kuitenkin mahdollista määrätä suora, joka kulkee (jossakin mielessä) *mahdollisimman läheltä* kaikkia havaintopisteitä?

Kirjoittamalla

$$y_i = \alpha + \beta x_i + \varepsilon_i$$
$$i = 1, 2, 3, 4, 5, 6$$

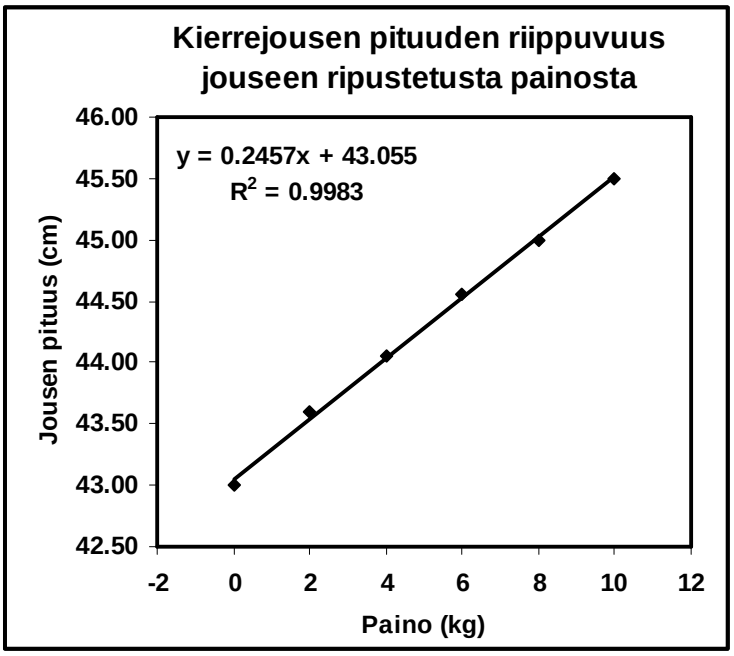
ja käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

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

kertoimet α ja β siten, että neliösumma

$$\sum_{i=1}^{n} \varepsilon_i^2$$

minimoituu.



Kuvioon oikealla on piirretty näin määrätty suora. Tarkastelemme suoran määräämistä luvussa **Yhden selittäjän lineaarinen regressiomalli**.

Regressiomalli ja kiinteät selittäjät

Olkoon

$$y_i = f(x_i; \beta) + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

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

Tehdään mallista seuraavat 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.

Kutsumme mallia tällöin **yhden selittäjän regressiomalliksi kiinteällä selittäjällä**. Mallissa

$$y_i = f(x_i; \beta) + \varepsilon_i \text{ , } i = 1, 2, \text{K}, n$$

on tällöin 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*

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*.

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

14.3. Regressiofunktiot ja regressioanalyysi

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 tarkastelemalla ko. muuttujien *yhteisjakaumaa* ja kuvaamalla *selitettävän muuttujan riippuvuutta selittävistä muuttujista selitettävän muuttujan* **regressiofunktiolla** *selittäjien suhteen*.

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*?

Ehdolliset jakaumat ja ehdolliset odotusarvot

Olkoon $f_{xy}(x, y)$ satunnaismuuttujien x ja y **yhteisjakauman** tiheysfunktio ja olkoot $f_x(x)$ ja $f_y(y)$ satunnaismuuttujien x ja y **reunajakaumien** tiheysfunktiot.

Satunnaismuuttujan y **ehdollisen jakauman** tiheysfunktio satunnaismuuttujan x suhteen on

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

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

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

jossa

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

on siis satunnaismuuttujan *y ehdollisen jakauman* tiheysfunktio satunnaismuuttujan *x* suhteen. Huomaa, että ehdollinen odotusarvo on *satunnaismuuttuja* ehtomuuttujan *x* funktiona.

Regressiofunktiot

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 yleisessä tapauksessa satunnaismuuttujan *y* ehdollisen jakauman

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

parametreista. 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 moniulotteisista satunnaismuuttujista ja niiden yhteisjakaumista, reunajakaumista, ehdollisista jakaumista, ehdollisista odotusarvoista ja regressiofunktioista: ks. kirjan **Todennäköisyyslaskenta** lukua **Moniulotteiset satunnaismuuttujat ja todennäköisyysjakaumat**.

Regressiofunktiot ja ennustaminen

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*?

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*.

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.

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

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

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

Oletetaan nyt, että funktion *f* muodon määräävän parametrin β arvo on *tuntematon* ja että haluamme löytää parametrille β *parhaan mahdollisen havaintoihin perustuvan estimaatin* eli *arvion*. **Regressio-ongelmalla** tarkoitetaan regressiofunktioiden yhteydessä *regressiofunktion muodon määräävän parametrin valintaongelmaa*.

Oletetaan, että muuttujia *x* ja *y* koskevat havainnot x_i ja y_i liittyvät *samaan havaintoyksikköön kaikille* $i = 1, 2, \ldots, n$. Kirjoitetaan

$$y_i = f(x_i;\beta) + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

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

Yhtälö

$$y_i = f(x_i;\beta) + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

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

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

$$(x_i, y_i) \in \quad ^2 \,,\; i = 1, 2, \mathrm{K}, n$$

Erään ratkaisun tähän käyränsovitusongelmaan tarjoaa **pienimmän neliösumman menetelmä**. Siinä parametrin β arvo valitaan siten, että *jäännös- eli virhetermien* ε_i *neliösumma*

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

tulee mahdollisimman pieneksi.

Esimerkki 1. Poikien pituuden riippuvuus isien pituudesta.

Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.

Kysymys: *Periytyykö isien pituus heidän pojilleen*?

Havaintoaineistona on tässä 300:n isän ja heidän poikiensa pituuksien muodostamaa lukuparia

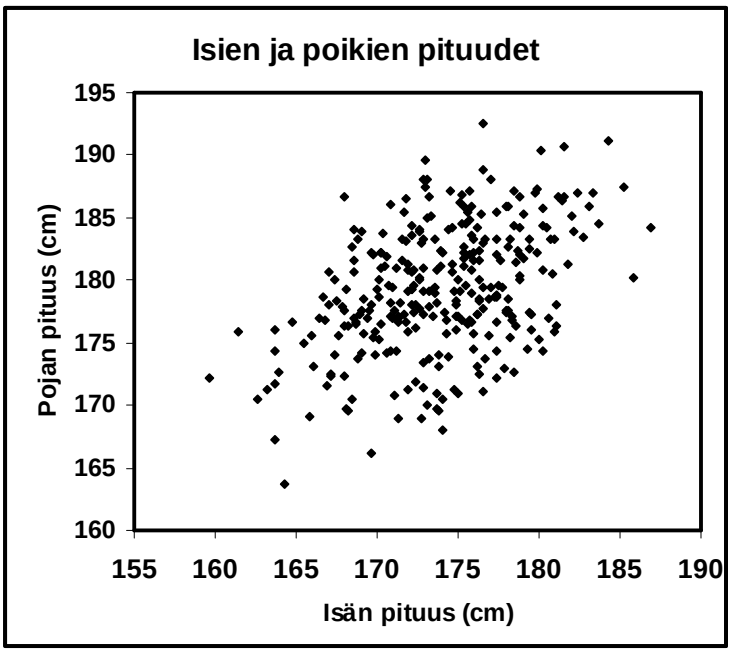
$$(x_i , y_i) , \; i = 1, 2, \ldots , 300$$

jossa siis

x_i = isän i pituus

y_i = isän i pojan pituus

Ks. pistediagrammia oikealla.



Pojan pituuden riippuvuus isän pituudesta ei selvästikään ole *eksaktia*: Saman mittaisten isien poikien pituudet näyttävät vaihtelevan paljonkin.

Kuvasta nähdään kuitenkin se, että lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkillä isillä ja vastaavasti pitkillä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.

Miten tällaista *tilastollista riippuvuutta* voidaan havainnollistaa?

Alla oleva taulukko esittää isien ja heidän poikiensa pituuksien *ehdollisia keskiarvoja*:

$M_k(x \mid x)$ = niiden *isien* pituuksien keskiarvo,
joiden *oma* pituus kuuluu x-väliin k , k = 1, 2, 3, 4, 5, 6, 7

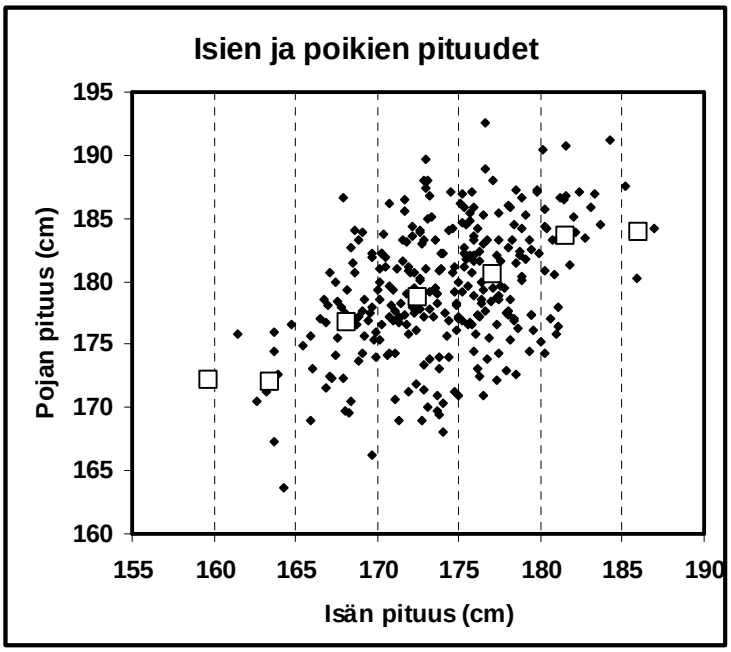
$M_k(y \mid x)$ = niiden *poikien* pituuksien keskiarvo,
joiden *isien* pituus kuuluu x-väliin k , k = 1, 2, 3, 4, 5, 6, 7

x-välin nro	x-väli	$M_k(x \mid x)$	$M_k(y \mid x)$
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

Lisätään ehdollisten keskiarvojen määräämät pisteet

$$(M_k(x \mid x)\,,\, M_k(y \mid x))\,,\; k = 1,2,3,4,5,6,7$$

edellä esitettyyn pistediagrammiin; ks. kuviota alla.



Ehdollisten keskiarvojen määräämiä pisteitä on merkitty kuviossa *neliöillä*.

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

Yo. kuvioon neliöillä merkityt, *ehdollisten keskiarvojen* määräämät pisteet

$$(M_k(x \mid x)\,,\, M_k(y \mid x))\,,\; k = 1,2,3,4,5,6,7$$

kuvaavat poikien pituuksien *keskimääräistä* tai *tilastollista riippuvuutta* heidän isiensä pituuksista. Kuvion mukaan riippuvuus näyttää olevan lähes *lineaarista*. *Regressioanalyysin tehtävänä* on juuri tällaisten *tilastollisten riippuvuuksien mallintaminen*.

Kirjoittamalla

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\; i = 1,2,\mathrm{K}\,,n = 300$$

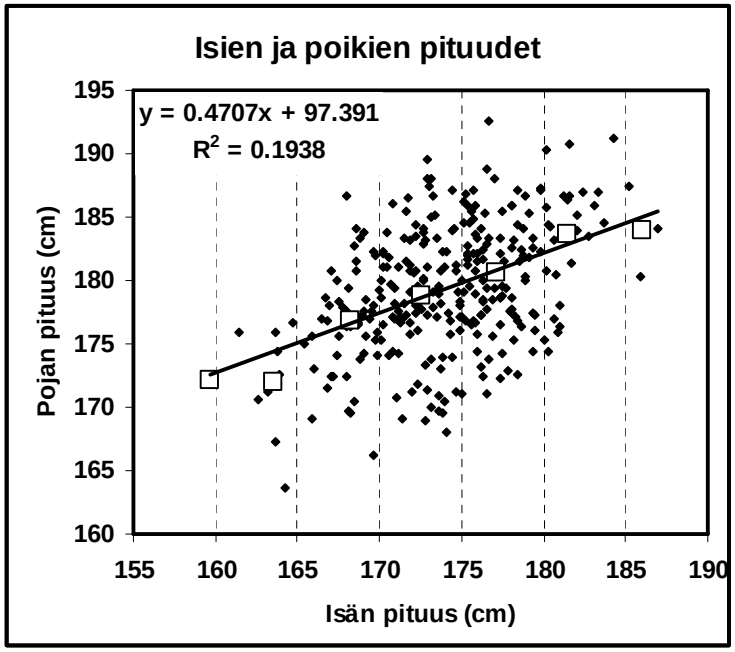
ja käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

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

kertoimet β_0 ja β_1 siten, että neliösumma

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

minimoituu. Näin määrätty suora on lisätty alla olevaan kuvioon; lisätietoja: ks. lukua **Yhden selittäjän lineaarinen regressiomalli**.



Regressiomalli ja satunnaiset selittäjät

Olkoon

$$y_i = f(x_i;\beta) + \varepsilon_i \text{ , } i = 1, 2, \text{K}, n$$

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

Tehdään mallista seuraava oletus: Selitettävän muuttujan *y* arvot y_i, selittävän muuttujan *x* arvot x_i ja jäännös- eli virhetermit ε_i ovat *satunnaisia*. Tällöin regressiomallissa

$$y_i = f(x_i;\beta) + \varepsilon_i \text{ , } i = 1, 2, \text{K}, n$$

on seuraavat osat:

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

x_i = **selittävän muuttujan** eli **selittäjän** *x 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*

Ei-satunnaisen selittävän muuttujan tapausta on käsitelty kappaleessa **Deterministiset mallit ja regressioanalyysi**.

14.4. Kaksiulotteisen normaalijakauman regressiofunktiot

Normaalijakauman yleistystä moniulotteiseen avaruuteen kutsutaan **multinormaalijakaumaksi** tai *moniulotteiseksi normaalijakaumaksi*. Multinormaalijakauman määräävät täydellisesti jakaumaan liittyvien satunnaismuuttujien *odotusarvot, varianssit* ja *korrelaatiot* (tai *kovariannssit*). Multi-normaalijakauma näyttelee *lineaaristen regressio-mallien* teoriassa keskeistä osaa, koska **multinormaali-jakauman kaikki regressiofunktiot ovat lineaarisia**.

Seuraavassa tarkastellaan lähemmin **kaksiulotteista normaalijakaumaa**; lisätietoja: ks. kirjan **Todennäköisyyslaskenta** lukua **Moniulotteisia jakaumia**.

Kaksiulotteisen normaalijakauman tiheysfunktio

Kaksiulotteisen 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

$$-\infty < \mu_x < +\infty \qquad -\infty < \mu_y < +\infty$$
$$\sigma_x > 0 \qquad \sigma_y > 0$$
$$-1 \le \rho_{xy} \le +1$$

Kaksiulotteisen normaalijakauman parametrit

Kaksiulotteisen normaalijakauman **parametreina** ovat satunnaismuuttujien *x* ja *y* **odotusarvot**, **varianssit** ja **korrelaatio**:

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

Koska satunnaismuuttujien *x* ja *y odotusarvot*, *varianssit* ja *korrelaatio* määräävät täydellisesti kaksiulotteisen normaalijakauman, merkitään

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

Kaksiulotteisen normaalijakauman parametrien tulkinta

Oletetaan, että satunnaismuuttujien *x* ja *y* muodostama pari noudattaa kaksiulotteista normaalijakaumaa $\mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$.

Satunnaismuuttujien *x* ja *y odotusarvot*

$$\mu_x = \mathrm{E}(x)$$

$$\mu_y = \mathrm{E}(y)$$

määräävät satunnaismuuttujien *x* ja *y* yhteisjakauman *todennäköisyysmassan* **painopisteen**, satunnaismuuttujien *x* ja *y varianssit*

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

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

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

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

kuvaa satunnaismuuttujien *x* ja *y lineaarisen riippuvuuden voimakkuutta*.

Koska satunnaismuuttujat *x* ja *y* noudattavat kaksiulotteista normaalijakaumaa, satunnaismuuttujat *x* ja *y ovat korreloimattomia, jos ja vain jos ne ovat riippumattomia*.

Huomaa, että yleisesti (ilman multinormaalisuusoletusta) pätee:

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

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

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

Kaksiulotteisen normaalijakauman ehdolliset jakaumat

Kaksiulotteisen normaalijakauman **ehdolliset jakaumat** ovat **normaalisia**. Satunnaismuuttujan *y* ehdollinen jakauma satunnaismuuttujan *x* suhteen on

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

jossa

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

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

ja satunnaismuuttujan *x* ehdollinen jakauma satunnaismuuttujan *y* suhteen on

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

jossa

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

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

Siten kaksiulotteisen normaalijakauman regressiofunktiot eli ehdolliset odotusarvot ovat *lineaarisia*.

Kaksiulotteisen normaalijakauman regressiofunktiot

Satunnaismuuttujan *y* **regressiofunktio** satunnaismuuttujan *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* kautta.

Satunnaismuuttujan *x* **regressiofunktio** satunnaismuuttujan *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* kautta.

Kaksiulotteisen normaalijakauman regressiofunktioiden määrittelemien **regressiosuorien** yhtälöistä nähdään seuraavaa:

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

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

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 tulee jyrkemmäksi (tulee loivemmaksi), jos

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

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 tulee jyrkemmäksi (tulee loivemmaksi), jos

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

Kaksiulotteisen normaalijakauman ehdolliset varianssit

Satunnaismuuttujan *y* **ehdollinen varianssi** satunnaismuuttujan *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. Satunnaismuuttujan *y* ehdollisella varianssilla satunnaismuuttujan *x* suhteen on seuraavat ominaisuudet:

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

(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.

Satunnaismuuttujan *x* **ehdollinen varianssi** satunnaismuuttujan *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

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

ympärillä. Satunnaismuuttujan *x* ehdollisella varianssilla satunnaismuuttujan *y* suhteen on seuraavat ominaisuudet:

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

(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.

14.5. Regressioanalyysin tehtävät

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. Mallin *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 on syytä huomata, että jäännöstermistä ε tehdyt oletukset vaikuttavat ratkaisevalla tavalla siihen tapaan, jolla regressioanalyysi tehdään.

Regressioanalyysi tarkoittaa seuraavia 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**

14.6. Regressiomallin lineaarisuus

Olkoon

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

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.

Regressiomallin soveltaminen yksinkertaistuu huomattavasti, jos mallin rakenneosa $f(x ; \beta)$ on parametrin β suhteen *lineaarinen*. Jos mallin rakenneosa $f(x ; \beta)$ on parametrin β suhteen *lineaarinen*, mallia kutsutaan **lineaariseksi regressiomalliksi**.

Huomautus:

- Epälineaaristen regressiomallien soveltaminen ei ole nykyisillä tietokoneilla ja ohjelmistoilla kovinkaan hankalaa. Emme kuitenkaan käsittele epälineaarisia regressiomalleja tässä esityksessä.

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 hyvin perusteltua, koska *kaikki multinormaalijakauman regressiofunktiot* eli ehdolliset odotusarvot *ovat lineaarisia*; ks. kappaletta **Kaksiulotteisen normaalijakauman regressiofunktiot**.

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 muuttujien y ja x *muunnoksilla*.

Esimerkki 1: Betonin lujuuden riippuvuus kuivumisajasta.

Kokeessa tutkittiin betonin vetolujuuden riippuvuutta betonin kuivumisajasta.

Havaintoaineistona on 21 lukuparia

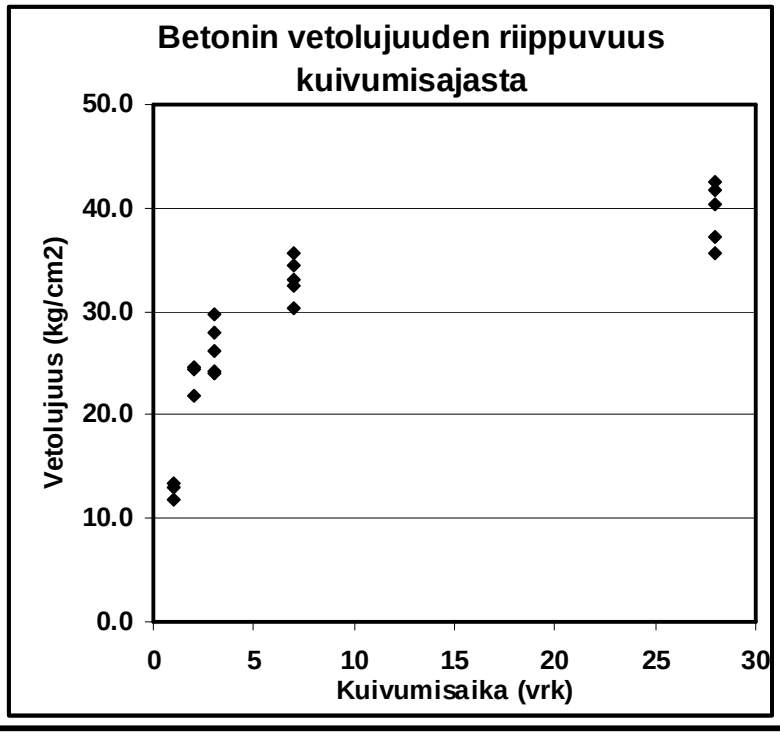
$$(x_i , y_i) , \; i = 1, 2, \ldots , 21$$

jossa

x_i = betoniharkon i kuivumisaika

y_i = betoniharkon i vetolujuus

Ks. alla olevaa pistediagrammia.

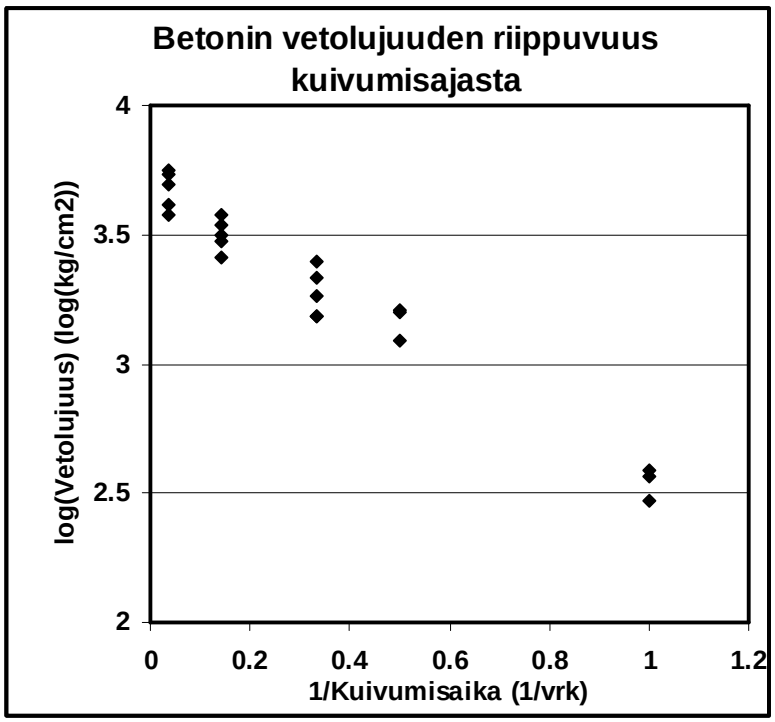


Vetolujuus näyttää kuvan perusteella riippuvan *epälineaarisesti* kuivumisajasta. Vetolujuuden epälineaarinen riippuvuus kuivumisajasta voidaan kuitenkin *linearisoida* seuraavilla muunnoksilla:

$$x_i' = 1/x_i$$

$$y_i' = \log(y_i)$$

Vrt. alla olevaa kuviota yllä olevaan kuvioon.



15. Yhden selittäjän lineaarinen regressiomalli

Regressioanalyysi on ehkä eniten sovellettu *tilastotieteen menetelmä*. **Yhden selittäjän lineaarinen regressiomalli** pyrkii *selittämään selitettävän muuttujan havaittujen arvojen vaihtelun yhden selittävän muuttujan havaittujen arvojen vaihtelun avulla*.

Tässä luvussa tarkastellaan seuraavia yhden selittäjän 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**?

Usean selittäjän lineaarisia regressiomalleja tarkastellaan luvussa **Yleinen lineaarinen malli**.

Avainsanat:

Aritmeettinen keskiarvo, Ehdollinen jakauma, Ehdollinen odotusarvo, Ehdollinen varianssi, Ei-satunnaisuus, Ennustaminen, Ennuste, Ennustusvirhe, Estimaattori, Estimointi, *F*-testi, Gaussin ja Markovin lause, Harha, Harhattomuus, Havainto, Heteroskedastisuus, Homoskedastisuus, Jäännös-neliösumma, Jäännöstermi, Jäännösvaihtelu, Jäännösvarianssi, Kaksiulotteinen normaalijakauma, Keskihajonta, Kokonaisneliösumma, Kokonaisvaihtelu, Korrelaatio, Kovarianssi, Kulmakerroin, Lineaarinen regressiomalli, Lineaarisuus, Malli, Mallin hyvyys, Mallineliösumma, Minimointi, Modifioidut standardioletukset, Neliösumma, Normaalisuusoletus, Odotusarvo, Otos, Otostunnusluku, Painopiste, Parametri, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Rakenneosa, Regressioanalyysi, Regressiodiagnostiikka, Regressiofunktio, Regressiokerroin, Regressiomalli, Regressiosuora, Residuaali, Reunajakauma, Satunnaisuus, Satunnainen osa, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Selitysaste, Sovite, Standardioletus, Systemaattinen osa, *t*-testi, Tarkentuvuus, Tehokkuus, Testi, Tyhjentävyys, Vakioselittäjä, Varianssi, Varianssianalyysihajotelma, Virhetermi, Yhden selittäjän lineaarinen regressiomalli, Yhteisjakauma

15.1. Yhden selittäjän lineaarinen regressiomalli ja sitä koskevat oletukset

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

Tehdään muuttujista *y* ja *x* seuraavat **perusoletukset**:

(i) Selitettävä muuttuja *y* on *suhdeasteikollinen* satunnaismuuttuja.

(ii) Selittävä muuttuja *x* on *kiinteä* eli *ei-satunnainen muuttuja*.

Huomautus:

- Satunnaisen selittävän muuttujan tapausta käsitellään myöhemmin erikseen.

Havainnot

Olkoot

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

selitettävän muuttujan *y* ja

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

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

$$(x_i, y_i) \in \quad ^2, i = 1, 2, \mathrm{K}, n$$

Yhden selittäjän lineaarinen regressiomalli

Tavanomaisen **yhden selittäjän lineaarinen regressiomallin** yleinen muoto on

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, \mathrm{K}, n$$

jossa

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

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

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

β_0 = *ei-satunnainen* ja *tuntematon vakio* (vakioselittäjän **regressiokerroin**)

β_1 = **selittäjän** *x* ei-satunnainen ja tuntematon **regressiokerroin**

Huomautus:

- Regressiokertoimet β_0 ja β_1 on oletettu *samoiksi* kaikille havaintoyksiköille *i*.

Huomautus:

- Kerrointa β_0 kutsutaan vakioselittäjän regressiokertoimeksi, koska sitä vastaa *keinotekoinen selittäjä*, joka saa vakioarvon = 1 jokaiselle havaintoyksiölle *i*.

Jäännöstermiä koskevat stokastiset oletukset

Mallin jäännöstermistä ε tehdään seuraavat *stokastiset oletukset*:

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

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

(iii) $\mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$

Oletuksen (ii) mukaan kaikilla jäännöstermeillä ε_i *on sama varianssi*. Jos oletus (ii) *pätee*, sanomme, että jäännöstermit ovat **homoskedastisia**. Kutsumme tällöin jäännöstermien yhteistä varianssia σ^2 mallin **jäännösvarianssiksi**. Jos oletus (ii) *ei päde*, jäännöstermeillä ε_i *ei ole samaa varianssia*. Sanomme tällöin, että jäännöstermit ovat **heteroskedastisia**.

Oletuksen (iii) mukaan jäännöstermit ε_i ovat **korreloimattomia**. Jos oletus (iii) *ei päde*, jäännöstermit ε_i ovat korreloituneita. Oletuksen (iii) sijasta tehdään usein seuraava, sitä *voimakkaampi oletus*:

(iii∗) Jäännöstermit ε_i, i = 1, 2, … , n ovat **riippumattomia**:

$$\varepsilon_1, \varepsilon_2, \mathrm{K}, \varepsilon_n \perp$$

Oletus (iii) seuraa oletuksesta (iii∗).

Lisäksi jäännöstermeistä ε_i tehdään usein **normaalisuusoletus**:

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

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

Jos yhden selittäjän lineaarinen regressiomalli

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \mathrm{K}, n$$

ja sen osat toteuttavat em. oletukset (i)-(iv), sanomme, että *malli toteuttaa ns.* **tavanomaiset** eli **standardioletukset**. Mallista tehtyjen oletuksien tarkistaminen muodostaa keskeisen osan regressioanalyysia; ks. lukua **Regressiodiagnostiikka**.

Selitettävän muuttujan ominaisuudet

Oletetaan, että yhden selittäjän lineaaristen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \; i = 1, 2, \mathrm{K}, n$$

jäännös- eli virhetermiä ε_i koskevat standardioletukset (i)-(iii) pätevät. Tällöin

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

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

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

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

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

Perustelu:

Kohdat (i)´ – (iv)´ seuraavat suoraan standardioletuksista (i) – (iv) sekä odotusarvon, varianssin, korrelaation ja normaalijakauman yleisistä ominaisuuksista sekä siitä, että regressiokertoimien β_0 ja β_1 lisäksi myös selittävän muuttujan x havaitut arvot on oletettu ei-satunnaisiksi vakioiksi.

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

(ii)´ $\mathrm{Var}(y_i) = \mathrm{E}[(y_i - \mathrm{E}(y_i))^2] = \mathrm{E}(\varepsilon_i^2) = \mathrm{Var}(\varepsilon_i) = \sigma^2$, $i = 1, 2, \mathsf{K}, n$

(iii)´ $\mathrm{Cor}(y_i, y_j) = \mathrm{E}[(y_i - \mathrm{E}(y_i))(y_i - \mathrm{E}(y_i))] = \mathrm{E}(\varepsilon_i \varepsilon_j) = \mathrm{Cor}(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$

(iv)´ Koska

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \mathsf{K}, n$$

ja

$$\varepsilon_i \quad \mathrm{N}(0, \sigma^2) \text{ , } i = 1, 2, \mathsf{K}, n$$

niin y_i on *lineaarimuunnos* normaalijakautuneesta satunnaismuuttujasta ε_i . Siten normaalijakauman yleisistä ominaisuuksista seuraa, että

$$y_i \quad \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma^2) \text{ , } i = 1, 2, \mathsf{K}, n$$

■

Mallin parametrit

Yhden selittäjän lineaaristen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \mathsf{K}, n$$

parametreja ovat *regressiokertoimet* β_0 ja β_1 ja *jäännösvarianssi* σ^2. Koska parametrit ovat tavallisesti *tuntemattomia*, ne on *estimoitava* muuttujien y ja x havaituista arvoista.

Mallin systemaattinen osa ja satunnainen osa

Oletetaan, että yhden selittäjän lineaaristen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \mathsf{K}, n$$

jäännös- eli virhetermiä ε_i koskeva standardioletus

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

pätee. Tällöin voimme esittää selitettävän muuttujan y havaitut arvot y_i kahden osatekijän summana muodossa

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

jossa

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

Sanomme, että selitettävän muuttujan y havaittujen arvojen y_i odotusarvot

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

muodostavat mallin **systemaattisen osan** ja jäännöstermi.

$$\varepsilon_i = y_i - \mathrm{E}(y_i) = y_i - \beta_0 - \beta_1 x_i \ , \ i = 1, 2, \mathrm{K}, n$$

muodostaa mallin **satunnaisen osan**. Mallin systemaattinen osa E(y_i) *riippuu* selittäjän *x* saamista arvoista, kun taas mallin satunnainen osa ε_i *ei riipu* selittäjän *x* saamista arvoista.

Regressiosuora

Tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i = 1, 2, \mathrm{K}, n$$

systemaattinen osa

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

määrittelee regressiosuoran

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

jossa

β_0 = regressiosuoran **vakiotermi**

β_1 = regressiosuoran **kulmakerroin**

Mallin jäännös- eli virhetermien ε_i varianssi σ^2 kuvaa havaintopisteiden

$$(x_i, y_i) \ , \ i = 1, 2, \mathrm{K}, n$$

vaihtelua regressiosuoran

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

ympärillä.

Regressiosuoran kulmakertoimen tulkinta

Tavanomaisen yhden selittäjän lineaarisen regressiomallin sovelluksissa on tärkeätä tuntea mallin systemaattisen osan määrittelemän regressiosuoran

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

kulmakertoimen β_1 tulkinta. Oletetaan, että selittäjän *x* arvo *kasvaa yhdellä yksiköllä*:

$$x \quad \rightarrow \quad x + 1$$

Regressiokerroin β_1 kertoo *paljonko selitettävän muuttujan y odotettavissa oleva arvo*

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

muuttuu:

$$\mathrm{E}(y) = \beta_0 + \beta_1 x \quad \rightarrow \quad \beta_0 + \beta_1 (x+1) = \beta_0 + \beta_1 x + \beta_1 = \mathrm{E}(y) + \beta_1$$

15.2. Regressiokertoimien estimointi

Koska tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ , \ i = 1, 2, \mathrm{K}, n$$

regressiokertoimet β_0 ja β_1 ovat tavallisesti *tuntemattomia*, ne on *estimoitava* muuttujien *y* ja *x* havaituista arvoista x_i ja y_i . Estimoinnissa regressiokertoimille β_0 ja β_1 pyritään löytämään sellaiset arvot, että niiden määräämä regressiosuora *selittäisi mahdollisimman hyvin selitettävän muuttujan y havaittujen arvojen vaihtelun*.

Regressiokertoimien PNS-estimointi

Regressiokertoimien β_0 ja β_1 estimointiin on tarjolla useita erilaisia menetelmiä. Menetelmistä käytettään yleisimmin *pienimmän neliösumman menetelmää*. Regressiokertoimien β_0 ja β_1 **pienimmän neliösumman** (*PNS*-) **estimaattorit** saadaan *minimoimalla neliösumma*

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

regressiokertoimien β_0 ja β_1 suhteen. **Regressiokertoimien** β_0 ja β_1 **PNS-estimaattoreiksi** saadaan

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

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

jossa

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

ovat *x*-havaintojen ja *y*-havaintojen *aritmeettiset keskiarvot*,

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

ovat *x*-havaintojen ja *y*-havaintojen *otosvarianssit*,

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

on *x*-havaintojen ja *y*-havaintojen *otoskovarianssi* ja lisäksi

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

on *x*-havaintojen ja *y*-havaintojen *otoskorrelaatiokerroin*.

Perustelu:

Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \text{K}, n$$

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

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

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

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

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

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

Kirjoitetaan normaaliyhtälöt (1) ja (2) muotoihin

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

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

Ratkaistaan β_0 yhtälöstä (1)´:

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

ja sijoitetaan ratkaisu yhtälöön (2)´:

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

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

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

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

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

Saatu ratkaisu on todellakin funktion $S(\beta_0, \beta_1)$ minimi, mikä nähdään siitä, että funktion $S(\beta_0, \beta_1)$ 2. *kertaluvun osittaisderivaattojen muodostama matriisi*

$$\mathbf{S} = \begin{bmatrix} \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0^2} & \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} \\ \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} & \frac{\partial^2 S(\beta_0, \beta_1)}{\partial \beta_1^2} \end{bmatrix} = \begin{bmatrix} 2n & 2\sum_{i=1}^{n} x_i \\ 2\sum_{i=1}^{n} x_i & 2\sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

on *positiivisesti definiitti*. Tämä seuraa esimerkiksi siitä, että

$$[\mathbf{S}]_{11} = 2n > 0$$

ja

$$\det(\mathbf{S}) = 2\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i \right)^2 \right] > 0$$

Jälkimmäinen epäyhtälö seuraa siitä, että

$$(n-1)s_x^2 = (n-1)\sum_{i=1}^{n}(x_i - \bar{x})^2 = (n-1)\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right] > 0$$

■

Huomaa, että selittäjän *x* regressiokertoimen β_1 estimaattorin b_1 lauseke voidaan edellä esitetyn mukaan kirjoittaa seuraaviin muotoihin:

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

Estimoitu regressiosuora

Tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \mathrm{K}, n$$

regressiokertoimien β_0 ja β_1 *PNS-estimaattorit* b_0 ja b_1 ja määrittelevät suoran

$$y = b_0 + b_1 x$$

jossa

b_0 = estimoidun regressiosuoran **vakiotermi**

b_1 = estimoidun regressiosuoran **kulmakerroin**

Sijoittamalla regressiokertoimien β_0 ja β_1 estimaattoreiden lausekkeet

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

$$b_1 = r_{xy}\frac{s_y}{s_x}$$

suoran $y = b_0 + b_1 x$ yhtälöön, voidaan yhtälö kirjoittaa seuraavaan muotoon:

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

Tästä muodosta nähdään, että estimoitu regressiosuura kulkee aina havaintopisteiden (x_i, y_i), $i = 1, 2, \mathrm{K}, n$ painopisteen $(\bar{x}, \bar{y})$ kautta.

Estimoidulla regressiosuoralla 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 tulee *jyrkemmäksi* (tulee *loivemmaksi*), jos

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

Regressiokertoimien PNS-estimaattoreiden ominainaisuudet

Tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

regressiokertoimien β_0 ja β_1 *PNS-estimaattorit* b_0 ja b_1 ovat **harhattomia** parametreille β_0 ja β_1 :

(i) $\mathrm{E}(b_0) = \beta_0$

(ii) $\mathrm{E}(b_1) = \beta_1$

Perustelu:

Todistetaan ensin kohta (ii).

(ii) Todetaan ensin, että regressiokertoimen β_1 *PNS-estimaattorin* b_1 kaava voidaan kirjoittaa muotoon

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Toiseksi

$$\begin{aligned}
\mathrm{E}(\bar{y}) &= \mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\beta_0 + \beta_1 x_i + \varepsilon_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i) \\
&= \beta_0 + \beta_1 \bar{x}
\end{aligned}$$

koska mallia koskevien oletusten mukaan regressiokertoimet β_0 ja β_1 sekä selittävän muuttujan x havaitut arvot x_i ovat *ei-satunnaisia vakioita* ja jäännöstermiä ε_i koskevan oletuksen (i) mukaan

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

Siten

$$\mathrm{E}(b_1) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(\beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \beta_1$$

Todistetaan kohta (i).

(i) Koska

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

kohdan (ii) todistuksesta seuraa, että

$$\mathrm{E}(b_0) = \beta_0 + \beta_1 \overline{x} - \beta_1 \overline{x} = \beta_0$$

■

Regressiokertoimien β_0 ja β_1 **PNS-estimaatoreiden** b_0 ja b_1 **varianssit** ovat

(i) $$\mathrm{Var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right)$$

(ii) $$\mathrm{Var}(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Lisäksi regressiokertoimien β_0 ja β_1 **PNS-estimaatoreiden** b_0 ja b_1 **kovarianssi** on

(iii) $$\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2 \overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Perustelu:

Todistetaan ensin kohta (ii).

(ii) Todetaan ensin, että regressiokertoimen β_1 *PNS-estimaattori* b_1 voidaan esittää selitettävän muuttujan *y* havaittujen arvojen

$$y_i \,,\ i = 1, 2, \mathrm{K}, n$$

lineaarikombinaationa, jossa painokertoimet

$$v_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \,,\ i = 1, 2, \mathrm{K}, n$$

ovat *ei-satunnaisia vakioita*:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sum_{i=1}^{n} v_i y_i$$

Regressiokerrointa b_1 koskevaa esitystä johdettaessa on käytetty hyväksi sitä, että

$$\begin{aligned}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) &= \sum_{i=1}^{n}(x_i-\bar{x})y_i - \sum_{i=1}^{n}(x_i-\bar{x})\bar{y} \\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i-\bar{x}) \\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i - \bar{y}\times 0 \\ &= \sum_{i=1}^{n}(x_i-\bar{x})y_i\end{aligned}$$

koska

$$\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x} = \sum_{i=1}^{n}x_i - n\bar{x} = 0$$

Käyttäen hyväksi regressiokertoimen b_1 esitysmuotoa satunnaismuuttujien y_i lineaarikombinaationa ja odotusarvo-operaattorin $\mathrm{E}(\cdot)$ lineaarisuutta nähdään, että

$$\mathrm{E}(b_1) = \sum_{i=1}^{n} v_i \,\mathrm{E}(y_i) = \sum_{i=1}^{n} v_i \mu_i$$

jossa

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

Siten

$$\begin{aligned}\mathrm{Var}(b_1) &= \mathrm{E}\left[(b_1 - \mathrm{E}(b_1))^2\right] \\ &= \mathrm{E}\left[\left(\sum_{i=1}^{n} v_i y_i - \sum_{i=1}^{n} v_i \mu_i\right)^2\right] \\ &= \mathrm{E}\left[\left(\sum_{i=1}^{n} v_i (y_i - \mu_i)\right)^2\right] \\ &= \mathrm{E}\left[\left(\sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j (y_i - \mu_i)(y_j - \mu_j)\right)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \,\mathrm{E}\left[(y_i - \mu_i)(y_j - \mu_j)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \,\mathrm{E}\left[(y_i - \mathrm{E}(y_i))(y_j - \mathrm{E}(y_j))\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \,\mathrm{Cov}(y_i, y_j)\end{aligned}$$

Koska edellä on todettu, että

$$\mathrm{Cov}(y_i, y_j) = \begin{cases} 0 \;,\; i \neq j \\ \sigma^2 \;,\; i = j \end{cases}$$

niin

$$\begin{aligned}
\mathrm{Var}(b_1) &= \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j \,\mathrm{Cov}(y_i, y_j) \\
&= \sum_{i=1}^{n} v_i^2 \,\mathrm{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^{n} v_i^2 \\
&= \sigma^2 \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} \\
&= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}
\end{aligned}$$

Todistetaan seuraavaksi kohta (iii).

(iii) Todetaan ensin, että selitettävän muuttujan *y* havaittujen arvojen

y_i , i = 1, 2, … , n

aritmeettinen keskiarvo $\overline{y}$ voidaan esittää havaittujen arvojen *lineaarikombinaationa*

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \sum_{i=1}^{n}\frac{1}{n} y_i = \sum_{i=1}^{n} u_i y_i$$

jossa *painokertoimet*

$$u_i = \frac{1}{n},\ i = 1, 2, \mathrm{K}, n$$

ovat ei-satunnaisia vakioita.

Kohdan (ii) todistuksessa todettiin, että myös regressiokertoimen β_1 PNS-estimaattori b_1 voidaan esittää selitettävän muuttujan *y* havaittujen arvojen

y_i , i = 1, 2, … , n

lineaarikombinaationa:

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \overline{x}) y_i}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \sum_{i=1}^{n} v_i y_i$$

jossa *painokertoimet*

$$v_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2},\ i = 1, 2, \mathrm{K}, n$$

ovat ei-satunnaisia vakioita.

Käyttäen hyväksi satunnaismuuttujien $\bar{y}$ ja b_1 esitysmuotoja selitettävän muuttujan y havaittujen arvojen lineaarikombinaatioina ja odotusarvo-operaattorin $\mathrm{E}(\cdot)$ lineaarisuutta nähdään, että

$$\mathrm{E}(\bar{y}) = \sum_{i=1}^{n} u_i \, \mathrm{E}(y_i) = \sum_{i=1}^{n} u_i \mu_i$$

ja

$$\mathrm{E}(b_1) = \sum_{i=1}^{n} v_i \, \mathrm{E}(y_i) = \sum_{i=1}^{n} v_i \mu_i$$

jossa

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

Siten

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, b_1) &= \mathrm{E}\left[(\bar{y} - \mathrm{E}(\bar{y}))(b_1 - \mathrm{E}(b_1)) \right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} u_i y_i - \sum_{i=1}^{n} u_i \mu_i \right) \left(\sum_{i=1}^{n} v_i y_i - \sum_{i=1}^{n} v_i \mu_i \right) \right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} u_i (y_i - \mu_i) \right) \left(\sum_{i=1}^{n} v_i (y_i - \mu_i) \right) \right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n} \sum_{j=1}^{n} u_i v_j (y_i - \mu_i)(y_j - \mu_j) \right) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} u_i v_j \, \mathrm{E}\left[(y_i - \mu_i)(y_j - \mu_j) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} u_i v_j \, \mathrm{E}\left[(y_i - \mathrm{E}(y_i))(y_j - \mathrm{E}(y_j)) \right] \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} u_i v_j \, \mathrm{Cov}(y_i, y_j)
\end{aligned}$$

Koska

$$\mathrm{Cov}(y_i, y_j) = \begin{cases} 0 \, , \, i \neq j \\ \sigma^2 \, , \, i = j \end{cases}$$

ja

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\bar{x} = 0$$

niin näemme, että selitettävän muuttujan y havaittujen arvojen aritmeettisen keskiarvon $\bar{y}$ ja regressiokertoimen β_1 PNS-estimaattorin b_1 *kovarianssi* = 0:

$$\begin{aligned}
\mathrm{Cov}(\bar{y}, b_1) &= \sum_{i=1}^{n}\sum_{j=1}^{n} u_i v_j \,\mathrm{Cov}(y_i, y_j) \\
&= \sum_{i=1}^{n} u_i v_i \,\mathrm{Var}(y_i) \\
&= \sigma^2 \sum_{i=1}^{n} u_i v_i \\
&= \sigma^2 \sum_{i=1}^{n} \frac{1}{n} \times \frac{x_i - \bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \sum_{i=1}^{n}(x_i - \bar{x}) \\
&= \frac{\sigma^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \times 0 = 0
\end{aligned}$$

Siten selitettävän muuttujan *y* havaittujen arvojen aritmeettinen keskiarvo $\bar{y}$ ja regressiokertoimen β_1 PNS-estimaattori b_1 ovat *korreloimattomia*.

Edellä esitetystä, kohdasta (ii) ja siitä, että

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

seuraa, että

$$\begin{aligned}
\mathrm{Cov}(b_0, b_1) &= \mathrm{Cov}(\bar{y} - b_1\bar{x}, b_1) \\
&= \mathrm{Cov}(\bar{y}, b_1) - \mathrm{Cov}(b_1\bar{x}, b_1) \\
&= 0 - \bar{x}\,\mathrm{Var}(b_1) \\
&= -\frac{\bar{x}\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}$$

Todistetaan lopuksi kohta (i).

(i) Suoraan laskemalla saamme:

$$\mathrm{Var}(b_0) = \mathrm{Var}(\bar{y} - b_1\bar{x}) = \mathrm{Var}(\bar{y}) + \bar{x}^2\,\mathrm{Var}(b_1) - 2\,\mathrm{Cov}(\bar{y}, b_1)$$

Sijoittamalla tähän selitettävän muuttujan *y* havaittujen arvojen aritmeettisen keskiarvon $\bar{y}$ varianssin lauseke

$$\mathrm{Var}(\bar{y}) = \frac{\sigma^2}{n}$$

sekä kohdissa (ii) ja (iii) saadut tulokset saamme lopulta estimaattorin b_0 varianssiksi lausekkeen

$$\mathrm{Var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

■

15.3. Sovitteet ja residuaalit

Estimoidun mallin **sovitteet** saadaan kaavalla

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

Estimoidun mallin **residuaalit** saadaan kaavalla

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

Malli selittää sitä *paremmin* selitettävän muuttujan *y* käyttäytymistä mitä *lähempänä* sovitteet ovat selitettävän muuttuja *y* havaittuja arvoja eli mitä *pienempiä* ovat residuaalit.

Sovitteiden ja residuaalien ominaisuuksia

Estimoidun mallin sovitteilla ja residuaaleilla on seuraavat ominaisuudet:

(i) Sovitteiden summa yhtyy selitettävän muuttujan havaittujen arvojen summaan:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

(ii) Residuaalien summa = 0:

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

(iii) Residuaalien ja selittävän muuttujan havaittujen arvojen tulosumma = 0:

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

(iv) Residuaalien ja sovitteiden tulosumma = 0:

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

Perustelu:

(i) Suoraan laskemalla saadaan:

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} (b_0 + b_1 x_i) = \sum_{i=1}^{n} b_0 + b_1 \sum_{i=1}^{n} x_i = nb_0 + nb_1 \bar{x} = n(\bar{y} - b_1 \bar{x}) + nb_1 \bar{x} = n\bar{y} = \sum_{i=1}^{n} y_i$$

(ii) Kohdasta (i) seuraa:

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i = 0$$

(iii) Koska

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

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

niin

$$\begin{aligned}\sum_{i=1}^{n} e_i x_i &= \sum_{i=1}^{n} e_i (x_i - \bar{x}) \\ &= \sum_{i=1}^{n} (y_i - \hat{y}_i)(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} (y_i - \bar{y} - b_1(x_i - \bar{x}))(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) - \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}) = 0\end{aligned}$$

(iv) Kohdista (ii) ja (iii) seuraa:

$$\sum_{i=1}^{n} e_i \hat{y}_i = \sum_{i=1}^{n} e_i (b_0 + b_1 x_i) = b_0 \sum_{i=1}^{n} e_i + b_1 \sum_{i=1}^{n} e_i x_i = 0$$

■

Huomautus:

- Kohdat (i)-(iv) *eivät päde*, jos regressiomallina *ei ole lineaarinen regressiomalli, jossa on mukana vakio*.

Sovitteet ja residuaalit: Havainnollistus

Alla oleva kuvio havainnollistaa sovitteiden ja residuaalien *geometrista tulkintaa*.

Malli:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,,\ i = 1, 2, \mathrm{K}, n$$

PNS-suora:

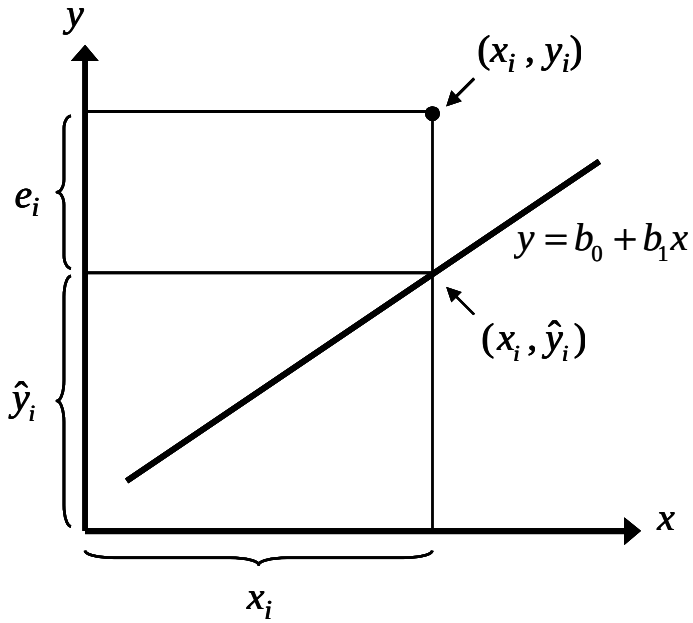
$$y = b_0 + b_1 x$$

Sovite:

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

Residuaali:

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - b_0 - b_1 x_i \,,\ i = 1, 2, \mathrm{K}, n \end{aligned}$$



15.4. Jäännösvarianssin estimointi

Tavanomaisen yhden selittäjän lineaarisen regressiomallin **jäännöstermien** ε_i , i = 1, 2, … , n **varianssin** eli **jäännösvarianssin** σ^2 **harhaton estimaattori** on

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

jossa

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

on estimoidun mallin *residuaali*.

Perustelu:

Todistetaan se, että estimaattori s^2 on *harhaton* jäännösvarianssille σ^2.

Todetaan ensin, että

$$\begin{aligned}
\mathrm{E}((n-2)s^2) &= \mathrm{E}\left(\sum_{i=1}^{n} e_i^2\right) \\
&= \mathrm{E}\left(\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2\right) \\
&= \sum_{i=1}^{n} \mathrm{E}(y_i - \overline{y} - b_1(x_i - \overline{x}))^2 \\
&= \sum_{i=1}^{n} \mathrm{Var}(y_i - \overline{y} - b_1(x_i - \overline{x}))
\end{aligned}$$

koska

$$\mathrm{E}(y_i - \overline{y} - b_1(x_i - \overline{x})) = \beta_0 + \beta_1 x_i - \beta_0 - \beta_1 \overline{x} - \beta_1(x_i - \overline{x}) = 0$$

Edelleen

$$\begin{aligned}
&\mathrm{Var}(y_i - \overline{y} - b_1(x_i - \overline{x})) \\
&\quad = \mathrm{Var}(y_i - \overline{y}) - 2(x_i - \overline{x})\,\mathrm{Cov}((y_i - \overline{y}), b_1) + (x_i - \overline{x})^2\,\mathrm{Var}(b_1)
\end{aligned}$$

Aikaisemmin todistetuista tuloksista seuraa, että

$$\mathrm{Var}(y_i - \overline{y}) = \left(1 - \frac{1}{n}\right)\sigma^2$$

ja

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

Lisäksi

$$\begin{aligned}
&\mathrm{Cov}((y_i-\bar{y}),b_1)\\
&\quad=\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sum_{j=1}^{n}(x_j-\bar{x})\,\mathrm{Cov}((y_i-\bar{y}),(y_j-\bar{y}))\\
&\quad=\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\left[\left(1-\frac{1}{n}\right)(x_i-\bar{x})-\frac{1}{n}\sum_{j\neq i}(x_j-\bar{x})\right]\\
&\quad=\frac{\sigma^2(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}
\end{aligned}$$

koska

$$\sum_{j\neq i}(x_j-\bar{x})=-(x_i-\bar{x})$$

Siten

$$\begin{aligned}
&\mathrm{Var}(y_i-\bar{y}-b_1(x_i-\bar{x}))\\
&\quad=\left(1-\frac{1}{n}\right)\sigma^2-2\frac{\sigma^2(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}+\frac{\sigma^2(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\\
&\quad=\sigma^2\left(1-\frac{1}{n}-\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)
\end{aligned}$$

ja

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

■

Huomaa, että s^2 on *residuaalien varianssi*, koska

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

mikä seuraa siitä, että residuaalien summa = 0, jos mallissa on mukana vakio.

Jäännösvarianssin σ^2 estimaattori s^2 kuvaa havaintopisteiden

$$(x_i,y_i),\ i=1,2,\mathrm{K},n$$

vaihtelua estimoidun regressiosuoran

$$y=b_0+b_1x$$

ympärillä

15.5. Varianssianalyysihajotelma ja selitysaste

Olkoon

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

selitettävän muuttujan y arvojen y_i vaihtelua kuvaava **kokonaisneliösumma** ja olkoon

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

estimoidun mallin PNS-residuaalien e_i vaihtelua kuvaava **jäännösneliösumma**. Voidaan osoittaa, että

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

missä

r_{xy} = x- ja y-havaintojen *otoskorrelaatiokerroin*

Perustelu:

Sijoitetaan estimoidun mallin residuaali

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

jäännösneliösumman *SSE* lausekkeeseen. Siten saamme lausekkeen

$$SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Sijoittamalla tähän lausekkeeseen estimaattorin b_0 lauseke

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

saamme edelleen lausekkeen

$$\begin{aligned} SSE &= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y} + b_1 \bar{x} - b_1 x_i)^2 \\ &= \sum_{i=1}^{n} [(y_i - \bar{y}) - b_1 (x_i - \bar{x})]^2 \\ &= \sum_{i=1}^{n} (y_i - \bar{y})^2 + b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2b_1 \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \end{aligned}$$

Ottamalla huomioon se, että

$$b_1 = r_{xy} \frac{s_y}{s_x}$$

ja lisäksi yhtälöt

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

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

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

saamme jäännösneliösumman *SSE* lausekkeen kehitettyä haluttuun muotoon:

$$\begin{aligned}SSE&=\sum_{i=1}^{n}e_i^2\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2+b_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2b_1\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2+r_{xy}^2\frac{s_y^2}{s_x^2}\sum_{i=1}^{n}(x_i-\bar{x})^2-2r_{xy}\frac{s_y}{s_x}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2+r_{xy}^2\frac{s_y^2}{s_x^2}(n-1)s_x^2-2r_{xy}\frac{s_y}{s_x}(n-1)r_{xy}s_xs_y\\&=\sum_{i=1}^{n}(y_i-\bar{y})^2-r_{xy}^2(n-1)s_y^2\\&=(1-r_{xy}^2)\sum_{i=1}^{n}(y_i-\bar{y})^2\\&=(1-r_{xy}^2)SST\end{aligned}$$

■

Koska aina pätee

$$|r_{xy}|<1$$

niin edellä johdetusta kaavasta nähdään, että

$$SSE\le SST$$

Määritellään erotus

$$SSM=SST-SSE$$

Voidaan osoittaa, että

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

Kutsumme neliösummaa *SSM* estimoidun mallin **malli-** (eli **regressio**-) **neliösummaksi**.

Kokonaisneliösumman *SST* hajotelmaa

$$SST = SSM + SSE$$

neliösummien *SSM* ja *SSE* summaksi kutsutaan **varianssianalyysihajotelmaksi**. Sijoittamalla neliösummien *SST*, *SSM* ja *SSE* lausekkeet varianssianalyysihajotelmaan voidaan hajotelma kirjoittaa muotoon

$$\sum_{i=1}^{n}(y_i-\overline{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

Perustelu:

Todistetaan varianssianalyysihajotelma.

Todetaan ensin, että

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\overline{y})^2&=\sum_{i=1}^{n}[(y_i-\hat{y}_i)-(\hat{y}_i-\overline{y})]^2\\&=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})\end{aligned}$$

Siten hajotelma tulee todistetuksi, jos voimme näyttää, että

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})=0$$

Olemme edellä todistaneet, että

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i=\sum_{i=1}^{n}e_i\hat{y}_i=0$$

ja

$$\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}\hat{y}_i$$

Siten

$$\begin{aligned}\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\overline{y})&=\sum_{i=1}^{n}(y_i-\hat{y}_i)\hat{y}_i-\overline{y}\sum_{i=1}^{n}(y_i-\hat{y}_i)\\&=\sum_{i=1}^{n}e_i\hat{y}_i-\overline{y}\left(\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\hat{y}_i\right)=0\\&=0+\overline{y}\times 0=0\end{aligned}$$

■

***Varianssianalyysihajotelmassa* selitettävän muuttujan *y kokonaisvaihtelua kuvaava neliösumma SST* on hajotettu kahteen osaan, joista *mallineliösumma SSM* kuvaa sitä osaa kokonaisneliösummasta, jonka *estimoitu malli on selittänyt* ja *jäännösneliösumma SSE* kuvaa sitä osaa kokonaisneliösummasta, jota *estimoitu malli ei ole selittänyt*.**

Todetaan vielä, että mallineliösumma *SSM* voidaan kirjoittaa muotoon

$$SSM=b_1^2\sum_{i=1}^{n}(x_i-\overline{x})^2$$

Perustelu:

Mallineliösumman määritelmän mukaan

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

Olemme todistaneet edellä, että

$$\sum_{i=1}^{n} \hat{y}_i = \sum_{i=1}^{n} y_i$$

Siten

$$\begin{aligned} SSM &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n} \hat{y}_i^2 + n\bar{y}^2 - 2\bar{y}\sum_{i=1}^{n} \hat{y}_i \\ &= \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 \\ &= \sum_{i=1}^{n} (\bar{y} - b_1(x_i - \bar{x}))^2 - n\bar{y}^2 \\ &= b_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

koska

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

■

Selitysaste

Varianssianalyysihajotelma motivoi määrittelemään estimoidun mallin **selitysasteen** kaavalla

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

jossa

SST = selitettävän muuttujan y arvojen vaihtelua kuvaava *kokonaisneliösumma*

SSE = estimoidun mallin residuaalien vaihtelua kuvaava *jäännösneliösumma*

SSM = estimoidun mallin *malli-* (eli *regressio-*) *neliösumma*

Varianssianalyysihajotelmasta seuraa, että aina pätee

$$0 \le R^2 \le 1$$

Selitysaste mittaa estimoidun regressiomallin *hyvyyttä*: Mitä suurempi on selitysaste, sitä suurempi on mallineliösumman (eli estimoidun mallin selittämä) osuus selitettävän muuttujan y kokonais-vaihtelua kuvaavasta neliösummasta ja sitä pienempi on jäännösneliösumman (eli estimoidun mallin selittämättä jättämä) osuus selitettävän muuttujan y kokonaisvaihtelua kuvaavasta neliösummasta.

Voidaan osoittaa, että selitysaste yhtyy selitettävän muuttujan havaittujen arvojen ja estimoidun mallin sovitteiden otoskorrelaatiokertoimeen:

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

Huomaa, että yhden selittäjän lineaarisen regressiomallin tapauksessa pätee lisäksi

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

jossa

r_{xy} = x-havaintojen ja y-havaintojen *otoskorrelaatiokerroin*

Selitysasteen ominaisuudet

Edellä esitetyistä tuloksista seuraa, että *selitysasteella* R^2 on seuraavat ominaisuudet:

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

(ii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 1$

(2) Kaikki residuaalit häviävät:

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

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

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

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

(iii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 0$

(2) $b_1 = 0$

(3) $r_{xy} = 0$

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

15.6. Laskutoimitusten järjestäminen

Jos regressiokertoimet joudutaan laskemaan *käsin* tai *laskimella*, yhden selittäjän lineaarisen regressiomallin *PNS-estimoinnin vaatimat laskutoimitukset* kannattaa järjestää seuraavan taulukon muotoon:

i	x_i	x_i^2	y_i	y_i^2	$x_i y_i$	$\hat{y}_i$	e_i	e_i^2
1	x_1	x_1^2	y_1	y_1^2	$x_1 y_1$	$\hat{y}_i$	e_1	e_1^2
2	x_2	x_2^2	y_2	y_2^2	$x_2 y_2$	$\hat{y}_i$	e_2	e_2^2
M	M	M	M	M	M	M	M	M
n	x_n	x_n^2	y_n	y_n^2	$x_n y_n$	$\hat{y}_i$	e_n	e_n^2
Summa	$\sum_{i=1}^{n} x_i$	$\sum_{i=1}^{n} x_i^2$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} y_i^2$	$\sum_{i=1}^{n} x_i y_i$	$\sum_{i=1}^{n} y_i$	$\sum_{i=1}^{n} e_i$	$\sum_{i=1}^{n} e_i^2$

Jos tarkoituksena on laskea ainoastaan *PNS-estimaatit* regressiokertoimille β_0 ja β_1, yllä olevasta taulukosta tarvitaan *x-havaintojen summa* Σx_i ja *neliösumma* Σx_i^2, *y-havaintojen summa* Σy_i sekä *x-* ja *y-havaintojen tulosumma* $\Sigma x_i y_i$.

Jos tarkoituksena on laskea myös estimoidun mallin *selitysaste,* tarvitaan edellä mainittujen suureiden lisäksi *y-havaintojen neliösumma* Σy_i^2 sekä estimoidun mallin *residuaalien neliösumma* Σe_i^2.

Havaintoarvojen **aritmeettiset keskiarvot** $\overline{x}$ ja $\overline{y}$, **otosvarianssit** s_x^2 ja s_y^2 sekä **otoskovarianssi** s_{xy} saadaan yllä olevan taulukon sarakesummista kaavoilla

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

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

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

joista **regressiokertoimien estimaatit** saadaan siis kaavoilla

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

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

Estimoidun mallin **sovitteet** saadaan kaavalla

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

ja **residuaalit** kaavalla

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

Estimoidun mallin **selitysaste** voidaan laskea kaavalla

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

jossa

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

on estimoidun mallin **jäännösneliösumma** (residuaalien neliösumma) ja

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

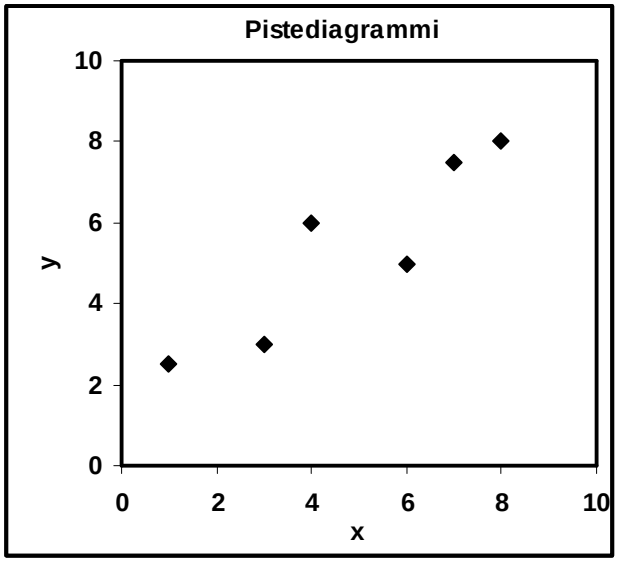
on selitettävän muuttujan arvojen vaihtelua kuvaava **kokonaisneliösumma**. Huomaa, että yhden selittäjän lineaarisen regressiomallin tapauksessa (koska mallissa on mukana vakio) pätee myös

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

Esimerkki 1: Regressiokertoimien laskeminen.

Alhaalla vasemmalla olevassa taulukossa on annettu keinotekoisen kahden muuttujan aineiston havaintoarvot (n = 6). Aineistoa kuvaava *pistediagrammi* on annettu alhaalla oikealla.

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



Samaa aineistoa on tarkasteltu luvun **Tilastollinen riippuvuus ja korrelaatio** kappaleen **Kahden muuttujan havainto-aineiston kuvaaminen** kohdassa **Otos-tunnuslukujen laskeminen**.

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

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

Yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \text{ , } i = 1, 2, \text{K}, n$$

regressiokertoimien β_0 ja β_1 *PNS-estimaatit* voidaan laskea näistä viidestä summasta:

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

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

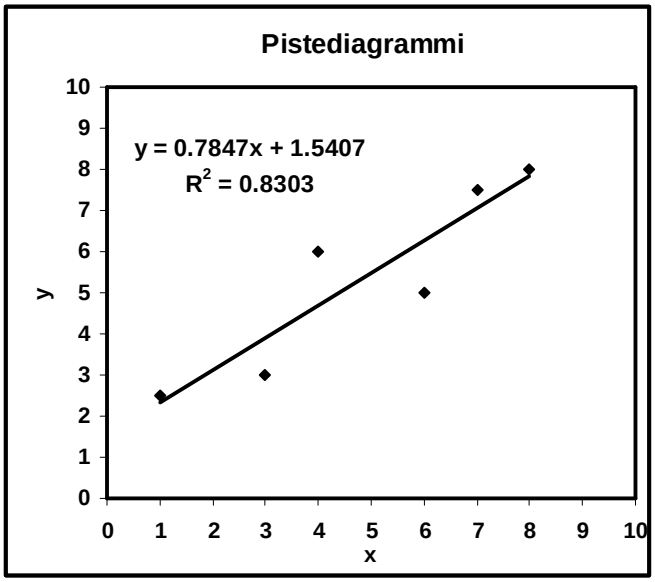
$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2} = \frac{182 - \frac{1}{6} \times 29 \times 32}{175 - \frac{1}{6} \times 29^2} = 0.785$$

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

Estimoidun regressiosuoran yhtälöksi saadaan siten

$$y = 1.5407 + 0.7847x$$

Regressiosuora on lisätty oikealla olevassa kuviossa aineistoa kuvaavaan pistediagrammiin.



Alla olevaan taulukkoon on laskettu estimoidun mallin *sovitteet* $\hat{y}_i$ ja *residuaalit* e_i :

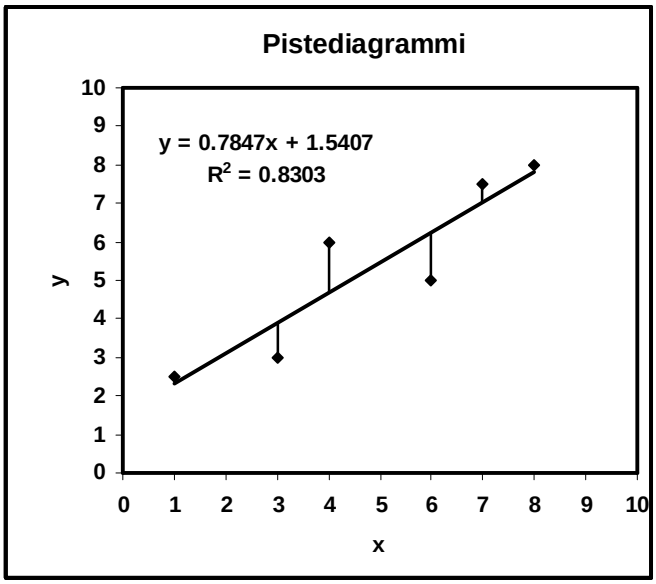
i	*x*	*y*	*Sovite*	*Residuaali*
1	1	2.5	2.325	0.175
2	3	3	3.895	-0.895
3	4	6	4.679	1.321
4	6	5	6.249	-1.249
5	7	7.5	7.033	0.467
6	8	8	7.818	0.182
Summa	29	32	32.000	0.000

Esimerkiksi, kun $i = 3$, niin

$$\begin{aligned}\hat{y}_3 &= 1.5407 + 0.7847x_3 \\ &= 1.5407 + 0.7847 \times 4 \\ &= 4.679\end{aligned}$$

$$\begin{aligned}e_3 &= y_3 - \hat{y}_3 \\ &= 6 - 4.679 \\ &= 1.321\end{aligned}$$

Oikealla olevaan kuvioon on lisätty *residuaaleja* vastaavat janat.



Huomautus:

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

Alla olevaan taulukkoon on laskettu estimoidun mallin *sovitteet* $\hat{y}_i$, *residuaalit* e_i ja *residuaalien neliöt* e_i^2.

i	*x*	*y*	*Sovite*	*Residuaali*	*Res²*
1	1	2.5	2.325	0.175	0.030
2	3	3	3.895	-0.895	0.801
3	4	6	4.679	1.321	1.744
4	6	5	6.249	-1.249	1.560
5	7	7.5	7.033	0.467	0.218
6	8	8	7.818	0.182	0.033
Summa	29	32	32.000	0.000	4.385

Mallin *jäännösvarianssin* σ^2 *harhaton estimaattori* on

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

Alla olevaan taulukkoon on laskettu havaintoarvojen summat ja neliösummat sekä estimoidun mallin *sovitteet* $\hat{y}_i$, *residuaalit* e_i ja *residuaalien neliöt* e_i^2.

i	*x*	*y*	*x²*	*y²*	*Sovite*	*Residuaali*	*Res²*
1	1	2.5	1	6.25	2.325	0.175	0.030
2	3	3	9	9	3.895	-0.895	0.801
3	4	6	16	36	4.679	1.321	1.744
4	6	5	36	25	6.249	-1.249	1.560
5	7	7.5	49	56.25	7.033	0.467	0.218
6	8	8	64	64	7.818	0.182	0.033
Summa	29	32	175	196.5	32	0.000	4.385

Estimoidun mallin *selitysaste* saadaan taulukon sarakesummista seuraavalla tavalla:

Kokonaisneliösumma:

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

Jäännösneliösumma:

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

Selitysaste:

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

Siten estimoitu malli on selittänyt

83.0 %

selitettävän muuttujan arvojen vaihtelusta.

Huomautus:

- Yllä tehdyt laskutoimitukset voidaan koota yhteen taulukkoon kuten edellä on esitetty. Esimerkissä laskutoimitukset on jaettu useaan osaan selvyyden vuoksi.

Esimerkkejä estimointitulosten tulkinnasta

Tarkastelemme alla luvuissa **Tilastollinen riippuvuus ja korrelaatio** ja **Johdatus regressioanalyysiin** käsiteltyihin esimerkkiaineistoihin sovitettuja regressiosuoria ja sekä suorien kulmakerrointen tulkintaa.

Esimerkki 2. Hooken laki.

Hooken lain mukaan kierrejousen (ns. ideaalijousen) pituus *y* riippuu *lineaarisesti* jouseen ripustetusta painosta *x*:

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

jossa

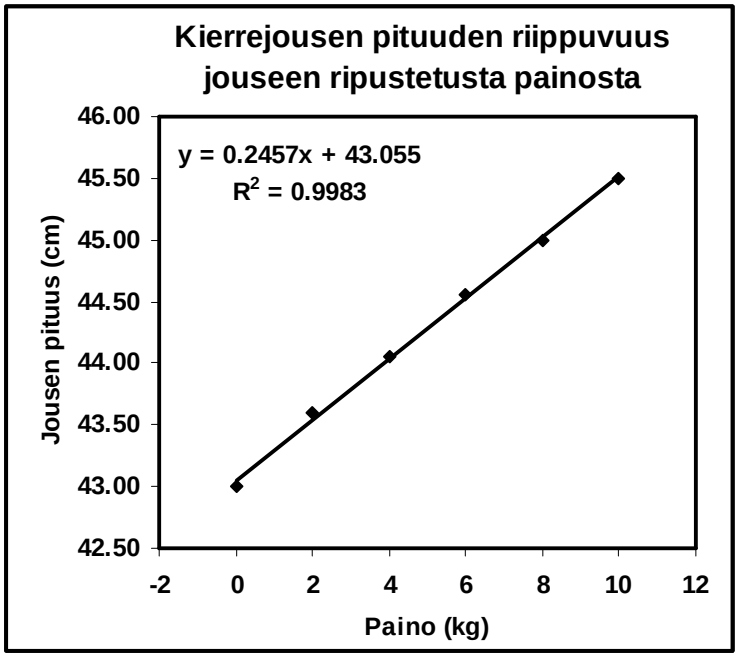
α = jousen pituus ilman painoa

β = ns. *jousivakio*

Alla olevassa taulukossa esitetään tulokset kokeesta, jossa Hooken lain pätevyyttä tutkittiin mittaamalla jousen pituus ilman painoa sekä painoilla, jotka olivat 2, 4, 6, 8 ja 10 kg.

Alla oleva pistediagrammi havainnollistaa koetuloksia graafisesti.

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



Diagrammiin on lisätty *aineistoon sovitettu regressiosuora*, jonka yhtälö on

$$y = 43.055 + 0.2457x$$

Suoran *kulmakertoimelle* (eli *jousivakiolle*)

$b = 0.2457$

voidaan antaa seuraava *tulkinta*: Jouseen ripustetun painon lisääminen 1 kg:lla pidentää jousta *keskimäärin* 0.2457 cm:llä.

Esimerkki 3. Poikien pituuden riippuvuus isien pituudesta.

Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.

Kysymys: Periytyykö isän pituus hänen pojilleen?

Havaintoaineistona on tässä 300:n isän ja heidän poikiensa pituuksien muodostamaa lukuparia

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

jossa siis

x_i = isän i pituus

y_i = isän i pojan pituus

Aineistoa on käsitelty luvun **Johdatus regressioanalyysi** kappaleessa **Regressiofunktiot ja regressioanalyysi**.

Aineistoa kuvaava pistediagrammi on oikealla. Diagrammiin on lisätty *aineistoon sovitettu regressiosuora*, jonka yhtälö on

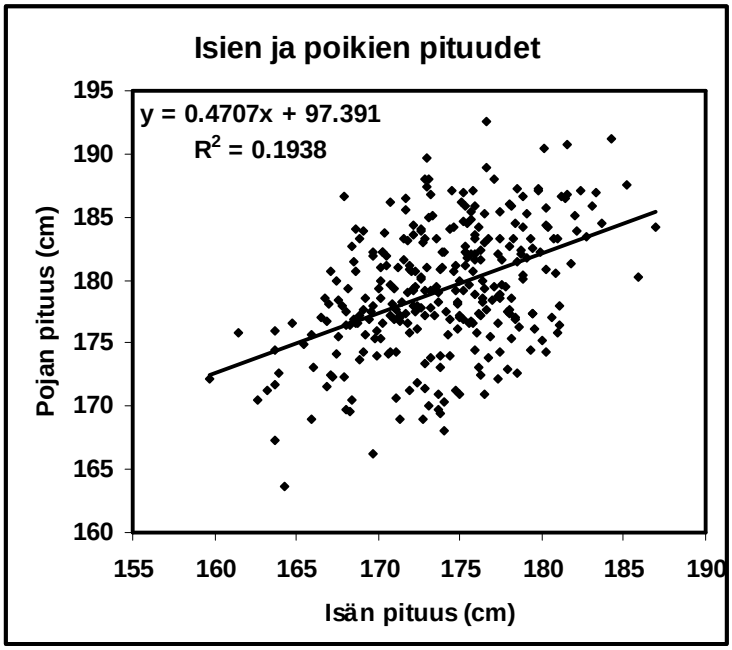
$$y = 97.391 + 0.4707x$$

Suoran *kulmakertoimelle*

$$b_1 = 0.4707$$

voidaan antaa seuraava *tulkinta*:

Jos isä A on 1 cm:n pitempi kuin isä B, isän A poika on *keskimäärin* 0.4707 cm pitempi kuin isän B poika.



Esimerkki 4. Keuhkosyövän yleisyyden riippuvuus savukkeiden kulutuksesta.

Onko keuhkosyöpä yleisempää sellaisissa maissa, joissa tupakoidaan paljon?

Alla on taulukko, jossa on tiedot savukkeiden kulutuksesta ja keuhkosyövän yleisyydestä 10:ssä maailman maassa.

Havaintoaineistona on tässä siis 10 lukuparia

$$(x_i , y_i) , \; i = 1, 2, \ldots , 10$$

jossa

x_i = savukkeiden kulutus maassa i vuonna 1930

y_i = sairastuvuus keuhkosyöpään maassa i vuonna 1950

Maa	Savukkeiden kulutus (kpl) per capita 1930	Keuhkosyöpä-tapausten 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

Aineistoa kuvaava pistediagrammi on annettu alla.

Diagrammiin on lisätty *aineistoon sovitettu regressiosuora*, jonka yhtälö on

$$y = 13.553 + 0.3577x$$

Suoran *kulmakertoimelle*

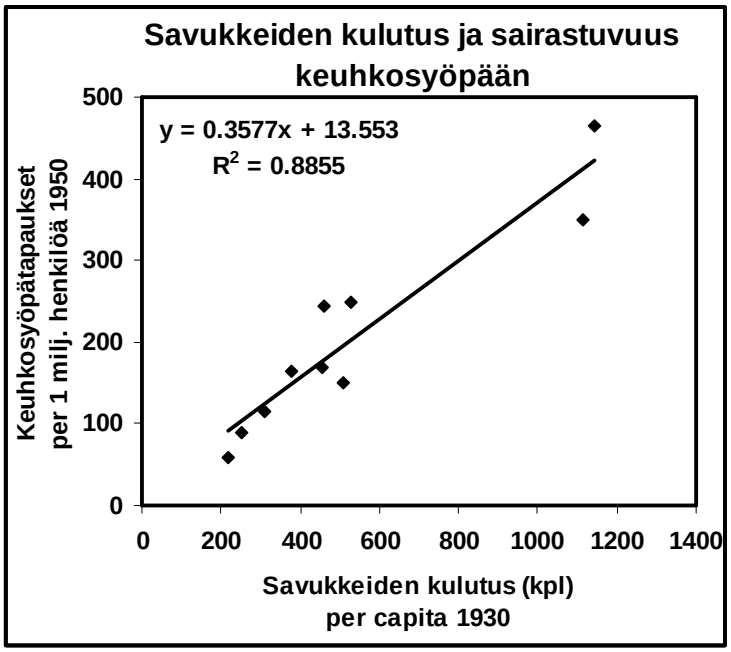
$$b_1 = 0.3577$$

voidaan antaa seuraava *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*.



15.7. Päättely yhden selittäjän lineaarisesta regressiomallista

Regressiokertoimien PNS-estimaattoreiden otosjakaumat

Jos tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät, regressiokertoimien β_0 ja β_1 **PNS-estimaattorit** b_0 ja b_1 ovat *normaalijakautuneita*:

(i) $$b_0 \quad \mathrm{N}\left(\beta_0 , \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \right)$$

(ii) $$b_1 \quad \mathrm{N}\left(\beta_1 , \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)$$

Perustelu:

Todistetaan ensin kohta (ii).

(ii) Olemme edellä todenneet kappaleen **Regressiokertoimien estimointi** kohdassa **Regressiokertoimien PNS-estimaattoreiden ominaisuudet**, että regressiokertoimen β_1 PNS-estimaattori b_1 voidaan esittää selitettävän muuttujan y havaittujen arvojen

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

lineaarikombinaationa. Tästä seuraa, että b_1 noudattaa normaalijakaumaa, koska havainnot $y_i , \; i = 1, 2, \ldots , n$ noudattavat normaalijakaumaa, ks. kohtaa **Normaalijakauma** ominaisuuksia monisteen **Todennäköisyyslaskenta** luvussa **Jatkuvia jakaumia**.

Kohta (i) tulee todistetuksi, kun toteamme vielä, että kappaleessa **Regressiokertoimien estimointi** todettiin, että

$$\mathrm{E}(b_1) = \beta_1$$

$$\mathrm{Var}(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Todistetaan kohta (i).

(i) Kohta (i) voidaan todistaa samalla tavalla kuin kohta (ii), koska myös regressiokertoimen β_0 PNS-estimaattori

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

on selitettävän muuttujan y havaittujen arvojen

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

lineaarikombinaatio.

■

Jäännösvarianssin otosjakauma

Olemme edellä todenneet, että jäännöstermien $\varepsilon_i , \; i = 1, 2, \ldots , n$ varianssin eli *jäännösvarianssin* σ^2 *harhaton estimaattori* on

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

jossa

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

on estimoidun mallin *residuaali*. Voidaan osoittaa, että s^2 on *riippumaton* estimaattoreista b_0 ja b_1 ja lisäksi

$$\frac{(n-2)s^2}{\sigma^2} \quad \chi^2(n-2)$$

Tuloksen perustelu sivuutetaan. Tulos on voidaan perustella samantapaisella argumentaatiolla kuin vastaava tulos riippumattomien normaalijakautuneiden havaintojen otosvarianssille luvussa **Otokset ja otosjakaumat**.

Yhdistämällä tämä tulos edellä johdettuihin regressiokertoimien jakaumia koskeviin tuloksiin seuraavat tärkeät jakaumatulokset:

(i) $$t_0 = \frac{b_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}} \quad t(n-2)$$

(ii) $$t_1 = \frac{b_1 - \beta_1}{s / \sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} \quad t(n-2)$$

Regressiokertoimien **luottamusvälit** ja kertoimia koskevat **testit** voidaan konstruoida näiden jakaumatulosten perusteella samalla tavalla kuin konstruoidaan luottamusväli ja yhden otoksen *t*-testi normaalijakauman odotusarvolle; ks. lukuja **Väliestimointi** ja **Testit suhdeasteikollisille muuttujille**.

Regressiokertoimien luottamusvälit

Oletetaan, että tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät.

Regressiokertoimen β_0 eli regressiosuoran **vakion luottamusväli** luottamustasolla $(1 - \alpha)$ on muotoa

$$b_0 \pm t_{\alpha/2}\, s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

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

Regressiokertoimen β_1 eli regressiosuoran **kulmakertoimen luottamusväli** luottamustasolla $(1 - \alpha)$ on muotoa

$$b_1 \pm t_{\alpha/2}\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}$$

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

Regressiokertoimia koskevat testit

Oletetaan, että tavanomaiset yhden selittäjän lineaarista regressiomallia koskevat oletukset pätevät.

Olkoon nollahypoteesina

$$\mathrm{H}_{00} : \beta_0 = \beta_0^0$$

Määritellään *t-testisuure*

$$t_0 = \frac{b_0 - \beta_0^0}{s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

Jos nollahypoteesi H_{00} pätee,

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

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

Olkoon nollahypoteesina

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

Määritellään *t-testisuure*

$$t_1 = \frac{b_1 - \beta_1^0}{s / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Jos nollahypoteesi H_{01} pätee,

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

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

Useimmissa regressioanalyysin sovellustilanteissa ollaan ensisijaisesti kiinnostuneita hypoteesin

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

testaamisesta. Tällöin yo. *t*-testisuure t_1 saa muodon

$$t_1 = \frac{b_1}{s / \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Jos nollahypoteesi H_{01} pätee, selitettävän muuttujan *y* arvojen vaihtelu *ei riipu* lineaarisesti selittävän muuttujan *x* havaittujen arvojen vaihtelusta. Useimmiten toivotaan, että nollahypoteesi H_{01} tulee testissä *hylätyksi*. Tällöin tiedetään, että selitettävän muuttujan *y* arvojen vaihtelu *riippuu* lineaarisesti selittävän muuttujan *x* havaittujen arvojen vaihtelusta.

Merkitsevyystasoa α vastaavan hylkäysalueen (kriittisten arvojen) tai testisuureiden havaittuja arvoja vastaavien *p*-arvojen määrääminen tapahtuu täsmälleen samanlaisella tekniikalla kuin normaalijakauman odotusarvoa koskevan tavanomaisen *t*-testin tapauksessa.

Voidaan osoittaa, että testi nollahypoteesille

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

voidaan perustaa myös *F-testisuureeseen*

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

jossa R^2 on estimoidun mallin *selitysaste*. Huomaa, että koska mallissa on mukana vakio, niin

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

Jos nollahypoteesi H_{01} pätee,

$$F \quad F(1, n-2)$$

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

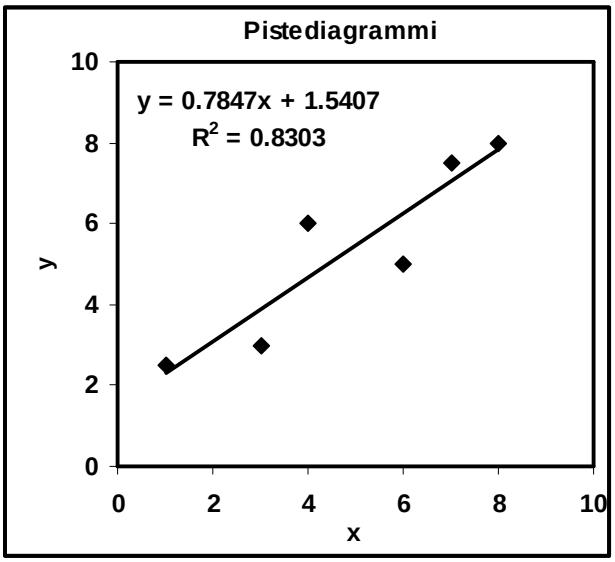
Tämä *F*-testi nollahypoteesille $H_{01} : \beta_1 = 0$ on ekvivalentti edellä esitetyn *t*-testin kanssa: Voidaan itse asiassa osoittaa, että

$$\sqrt{F} = t_1$$

Esimerkki 1: Testi regressiosuoran kulmakertoimelle.

Alhaalla vasemmalla olevassa taulukossa on annettu keinotekoisen kahden muuttujan aineiston havaintoarvot (n = 6). Aineistoa kuvaava *pistediagrammi* on annettu alhaalla oikealla.

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



Samaa aineistoa on tarkasteltu luvun **Tilastollinen riippuvuus ja korrelaatio** kappaleen **Kahden muuttujan havainto-aineiston kuvaaminen** kohdassa **Otos-tunnuslukujen laskeminen** sekä tämän luvun kappaleessa **Laskutoimitusten järjestäminen**.

Aineistoon sovitettiin kappaleessa **Laskutoimitusten järjestäminen** yhden selittäjän lineaarinen regressiomalli

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j \ , \ j = 1, 2, \mathrm{K}, n$$

ja estimoiduksi regressiosuoran yhtälöksi saatiin

$$y = 1.5407 + 0.7847x$$

Ks. yllä olevaa pistediagrammia.

Tarkastellaan testiä nollahypoteesille

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

Alla olevaan taulukkoon on laskettu havaintoarvojen summat ja neliösummat sekä estimoidun mallin *sovitteet* $\hat{y}_i$, *residuaalit* e_i ja *residuaalien neliöt* e_i^2.

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

Kulmaertoimen β_1 estimaatti:

$$b_1 = 0.7847$$

Selittäjän x havaittujen arvojen otosvarianssi:

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

Jäännösvarianssi:

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

t-testisuureen arvo:

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

Jos nollahypoteesi H_{01} pätee, testisuure t_1 on jakautunut *t-jakauman* mukaan vapausastein

$$n - 2 = 6 - 2 = 4$$

eli

$$t_1 \sim t(4)$$

jos nollahypoteesi H_{01} pätee.

Valitaan *merkitsevyystasoksi* 0.05 ja olkoon *vaihtoehtoinen hypoteesi* muotoa

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

Tällöin merkitsevyystasoa 0.05 vastaavat *kriittiset arvot* ovat

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

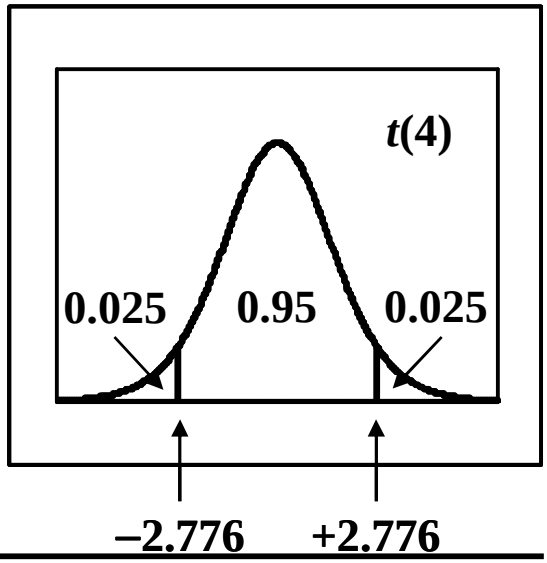
ks. kuviota oikealla. Testin *hylkäysalue* on siten muotoa

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

Koska

$$t_1 = 4.423 > 2.776$$

niin testisuureen t_1 arvo on joutunut hylkäys-



alueelle ja *voimme hylätä nollahypoteesin* H_{01} ja *hyväksyä vaihtoehtoisen hypoteesin* H_{11} merkitsevyystasolla 0.05.

15.8. Ennustaminen yhden selittäjän lineaarisella regressiomallilla

Selitettävän muuttujan odotettavissa olevan arvon ennustaminen

Oletetaan, että selitettävä muuttuja *y* saa arvon

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

kun selittäjä *x* saa arvon x. Mikä on **paras ennuste selitettävän muuttujan** *y* **odotettavissa olevalle arvolle**

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

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

Valitaan selitettävän muuttujan ehdollisen odotusarvon $\mathrm{E}(y \mid x)$ **ennusteeksi** (*estimaattoriksi*) lauseke

$$y \mid x = b_0 + b_1 x$$

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

Huomautus:

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

Selitettävän muuttujan odotettavissa olevan arvon ennusteen otosjakauma

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

Tällöin **ennusteen**

$$y \mid x = b_0 + b_1 x$$

otosjakauma on *normaalijakauma*:

$$y \mid x \sim \mathrm{N}\left(\beta_0 + \beta_1 x, \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right] \right)$$

Perustelu:

Koska

$$\mathrm{E}(\varepsilon \mid x) = 0$$

näemme suoraan, että ennusteen

$$\hat{y} \mid x = b_0 + b_1 x$$

odotusarvo on

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

Kappaleessa **Regressiokertoimien estimointi** todettiin, että

$$\mathrm{Var}(b_0) = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)$$

$$\mathrm{Var}(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\mathrm{Cov}(b_0, b_1) = -\frac{\sigma^2 \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Siten ennusteen $\hat{y} \mid x$ varianssiksi saadaan

$$\begin{aligned}
\mathrm{Var}(\hat{y} \mid x) &= \mathrm{Var}(b_0 + b_1 x \mid x) \\
&= \mathrm{Var}(b_0 \mid x) + x^2 \, \mathrm{Var}(b_1 \mid x) + 2x \, \mathrm{Cov}(b_0, b_1 \mid x) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) + \sigma^2 \frac{x^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} - \frac{2\sigma^2 x \bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{x^2 + \bar{x}^2 - 2x\bar{x}}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\end{aligned}$$

Se, että ennuste $\hat{y} \mid x$ noudattaa normaalijakaumaa todistetaan samalla tavalla kuin se, että regressiokertoimien β_0 ja β_1 PNS-estimaattorit b_0 ja b_1 noudattavat normaalijakaumaa kappaleessa **Päättely yhden selittäjän lineaarisesta regressiomallista**.

■

Selitettävän muuttujan odotettavissa olevan arvon luottamusväli

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

Tällöin ennusteen

$$\hat{y} \mid x = b_0 + b_1 x$$

luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

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

jossa $-t_{\alpha/2}$ ja $+ t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori. Väli muodostaa selittäjän *x* arvojen x funktiona *luottamusvyön* estimoidun regressiosuoran

$$y = b_0 + b_1 x$$

ympärille.

Selitettävän muuttujan odotettavissa olevan arvon luottamusvälin ominaisuudet

Ennusteen

$$y \mid x = b_0 + b_1 x$$

luottamusväli *kaventuu*, jos havaintojen lukumäärä *n* tai selittäjän otosvarianssi s_x^2 kasvaa. Toisaalta luottamusväli on sitä *leveämpi*, mitä kauempana piste x on selittäjän *x* havaittujen arvojen aritmeettisesta keskiarvosta $\bar{x}$.

Selitettävän muuttujan arvon ennustaminen

Oletetaan, että selitettävä muuttuja *y* saa arvon

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

kun selittäjä *x* saa arvon x. Mikä on **paras ennuste selitettävän muuttujan** *y* **arvolle** y, kun selittäjä *x* saa arvon x?

Valitaan *selitettävän muuttujan arvon* y **ennusteeksi** (*estimaattoriksi*) lauseke

$$y \mid x = b_0 + b_1 x$$

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

Huomautus:

- Sekä selitettävän muuttujan *y* arvo y että ennuste $y \mid x$ ovat *satunnaismuuttujia*.

Selitettävän muuttujan arvon ennusteen otosjakauma

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

Tällöin voidaan osoittaa, että **ennustevirheen**

$$y - y \mid x$$

otosjakauma on *normaalijakauma*:

$$y - y \mid x \sim \mathrm{N}\left(0\,,\, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]\right)$$

Selitettävän muuttujan arvon luottamusväli

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

Tällöin selitettävän muuttujan *y* arvon y **luottamusväli** luottamustasolla $(1 - \alpha)$ on muotoa

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

jossa $-t_{\alpha/2}$ ja $+ t_{\alpha/2}$ ovat luottamustasoon $(1 - \alpha)$ liittyvät luottamuskertoimet *t-jakaumasta*, jonka vapausasteiden luku on $(n - 2)$ ja s^2 on jäännösvarianssin σ^2 harhaton estimaattori. Väli muodostaa selittäjän *x* arvojen x funktiona *luottamusvyön* estimoidun regressiosuoran

$$y = b_0 + b_1 x$$

ympärille.

Selitettävän muuttujan arvon luottamusvälin ominaisuudet

Selitettävän muuttujan *y* arvon y luottamusväli *kaventuu*, jos havaintojen lukumäärä *n* tai selittäjän otosvarianssi s_x^2 kasvaa. Toisaalta luottamusväli on sitä *leveämpi*, mitä kauempana piste x on selittäjän *x* havaittujen arvojen aritmeettisesta keskiarvosta $\bar{x}$.

Selitettävän muuttujan odotettavissa olevan arvon luottamusväli *vs* selitettävän muuttujan arvon luottamusväli

Selitettävän muuttujan *y* arvon y luottamusvyö *on leveämpi* kuin selitettävän muuttujan *y* arvon odotusarvon $\mathrm{E}(y \mid x)$ luottamusvyö. Tämä johtuu siitä, että selitettävän muuttujan *y keskimääräisen arvon ennustaminen on helpompaa kuin sen yksittäisen arvon ennustaminen*.

15.9. Yhden selittäjän lineaarinen regressiomalli ja satunnainen selittäjä

Jos tavanomaisen yhden selittäjän lineaarisen regressiomallin

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\ i = 1, 2, \mathrm{K}, n$$

selittäjän *x* arvot ovat *satunnaisia,* mutta jäännöstermiä ε koskeva standardioletus (4) voidaan *korvata* oletuksella

$$(4)' \qquad \varepsilon_i \mid x_i \quad \mathrm{N}(0, \sigma^2)\,,\ i = 1, 2, \mathrm{K}, n$$

niin **kaikki edellä esitetty teoria pätee sopivasti modifioituna**. Oletus (4)´ tarkoittaa sitä, että satunnaismuuttujan ε_i *ehdollinen jakauma* ehdolla x_i on normaalijakauma.

15.10. Kaksiulotteisen normaalijakauman regressiofunktioiden estimointi

Kaksiulotteinen normaalijakauma ja sen tiheysfunktio

Oletetaan, että satunnaismuuttujien *x* ja *y* pari (x,y) noudattaa **kaksiulotteista normaalijakaumaa** eli

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

jossa

$$\mu_x = \mathrm{E}(x) \qquad \mu_y = \mathrm{E}(y)$$
$$\sigma_x^2 = \mathrm{Var}(x) = \mathrm{E}[(x-\mu_x)^2] \qquad \sigma_y^2 = \mathrm{Var}(y) = \mathrm{E}[(y-\mu_y)^2]$$

$$\rho_{xy} = \mathrm{Cor}(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$
$$\sigma_{xy} = \mathrm{Cov}(x,y) = \mathrm{E}[(x-\mu_x)(y-\mu_y)]$$

Kaksiulotteisen normaalijakauman **tiheysfunktio** on muotoa

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

jossa

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

Kaksiulotteisen normaalijakauman ehdolliset jakaumat

Kaksiulotteisen normaalijakauman **ehdolliset jakaumat** ovat normaalisia:

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

jossa

$$\mu_{y|x} = \mathrm{E}(y \mid x) = \mu_y + \rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\mu_x)$$
$$\sigma_{y|x}^2 = \mathrm{Var}(y \mid x) = (1-\rho_{xy}^2)\sigma_y^2$$

$$\mu_{x|y} = \mathrm{E}(x \mid y) = \mu_x + \rho_{xy}\frac{\sigma_x}{\sigma_y}(y-\mu_y)$$
$$\sigma_{x|y}^2 = \mathrm{Var}(x \mid y) = (1-\rho_{xy}^2)\sigma_x^2$$

Ehdollisten odotusarvojen E($y|x$) ja E($x|y$) kaavoista nähdään:

(i) Satunnaismuuttujan y **ehdollinen odotusarvo** satunnaismuuttujan x suhteen eli satunnaismuuttujan y **regressiofunktio** satunnaismuuttujan x suhteen *riippuu lineaarisesti* ehtomuuttujan x arvoista eli on muotoa:

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

(ii) Satunnaismuuttujan x **ehdollinen odotusarvo** satunnaismuuttujan y suhteen eli satunnaismuuttujan x **regressiofunktio** satunnaismuuttujan y suhteen *riippuu lineaarisesti* ehtomuuttujan y arvoista eli on muotoa:

$$\mathrm{E}(x \mid y) = \alpha_0 + \alpha_1 y$$

Ehdollisten varianssien Var($y|x$) ja Var($x|y$) kaavoista nähdään:

(i) Satunnaismuuttujan y **ehdollinen varianssi** satunnaismuuttujan x suhteen *ei riipu* ehtomuuttujan x arvoista.

(ii) Satunnaismuuttujan x **ehdollinen varianssi** satunnaismuuttujan y suhteen *ei riipu* ehtomuuttujan y arvoista.

Ehdollisten odotusarvojen kaavoista nähdään edelleen, että sekä satunnaismuuttujan *y regressiofunktio* satunnaismuuttujan x suhteen että satunnaismuuttujan *y regressiofunktio* satunnaismuuttujan x suhteen kulkevat satunnaismuuttujien x ja y todennäköisyysjakauman todennäköisyysmassan *painopisteen*

$$(\mu_x, \mu_y)$$

kautta.

Otos kaksiulotteisesta normaalijakaumasta

Oletetaan, että satunnaismuuttujat x ja y noudattavat **kaksiulotteista normaalijakaumaa** eli

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

Olkoot

$$y_1, y_2, \mathrm{K}, y_n$$

muuttujan y havaitut arvot ja

$$x_1, x_2, \mathrm{K}, x_n$$

muuttujan x havaitut arvot ja oletetaan, että havaintoarvojen x_i ja y_i parit

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

muodostavat **satunnaisotoksen** kaksiulotteista normaalijakaumasta

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

Tällöin

$$(x_1, y_1), (x_2, y_2), \mathrm{K}, (x_n, y_n) \perp$$
$$(x_i, y_i) \quad \mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})\ ,\ i = 1, 2, \mathrm{K}, n$$

Kaksiulotteisen normaalijakauman regressiofunktioiden PNS-estimointi

Oletetaan, että havaintoarvojen x_i ja y_i parit (x_i,y_i) , $i = 1, 2, \ldots, n$ muodostavat **satunnaisotoksen** *kaksiulotteista normaalijakaumasta* $\mathrm{N}_2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho_{xy})$.

Kaksiulotteisen normaalijakauman **regressiofunktiot** ovat muotoa

$$\mathrm{E}(y \mid x) = \beta_0 + \beta_1 x$$
$$\mathrm{E}(x \mid y) = \alpha_0 + \alpha_1 y$$

Estimoidaan regressiofunktioiden parametrit **pienimmän neliösumman menetelmällä**.

Määritellään **yhden selittäjän lineaariset regressiomallit**

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

(2) $x_i = \alpha_0 + \alpha_1 x_i + \delta_i \;,\; i = 1, 2, \mathrm{K}, n$

Muuttujan *y* PNS-suoran yhtälö muuttujan *x* suhteen on

(3) $y = b_0 + b_1 x$

jossa

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

Muuttujan *x* PNS-suoran yhtälö muuttujan *y* suhteen on

(4) $x = a_0 + a_1 y$

jossa

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

Mallien (1) ja (2) regressiokertoimien β_0, β_1, α_0, α_1 **PNS-estimaattoreiden** b_0, b_1, a_0, a_1 lausekkeissa

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

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

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

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

Muuttujan *y* PNS-suoran yhtälö muuttujan *x* suhteen voidaan kirjoittaa muotoon

(3)´ $y - \overline{y} = r_{xy} \frac{s_y}{s_x} (x - \overline{x})$

Muuttujan *x* PNS-suoran yhtälö muuttujan *y* suhteen voidaan kirjoittaa muotoon

(4)´ $x - \overline{x} = r_{xy} \frac{s_x}{s_y} (y - \overline{y})$

Yhtälöistä (3)´ ja (4)´ nähdään välittömästi, että molemmat PNS-suorat kulkevat havaintoaineiston **painopisteen**

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

kautta. Yhtälöistä (3)´ ja (4)´ nähdään edelleen, että PNS-suorien kulmakertoimien tulo on muuttujien *y* ja *x korrelaatiokertoimen* neliö:

$$a_1 b_1 = \left(r_{xy} \frac{s_x}{s_y} \right) \left(r_{xy} \frac{s_y}{s_x} \right) = r_{xy}^2$$

Voidaan osoittaa, että molempiin PNS-suoriin liittyy sama *selitysaste* R^2 ja se yhtyy muuttujien *y* ja *x* havaittujen arvojen korrelaatiokertoimen neliöön:

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

Olkoon

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

estimoidun mallin (3) **sovite** ja

$$e_{yi} = y_i - \hat{y}_i \text{ , } i = 1, 2, \text{K} , n$$

vastaava **residuaali**.

Olkoon

$$\hat{x}_i = a_0 + a_1 x_i \text{ , } i = 1, 2, \text{K} , n$$

estimoidun mallin (4) **sovite** ja

$$e_{xi} = x_i - \hat{x}_i \text{ , } i = 1, 2, \text{K} , n$$

vastaava **residuaali**.

PNS-suoraan (3) liittyvä **jäännösvarianssin** (harhaton) **estimaattori** on

$$s_{(3)}^2 = \frac{SSE_y}{n-2}$$

jossa

$SSE_y = \sum_{i=1}^{n} e_{yi}^2$ = PNS-suoraan (3) liittyvä jäännösneliösumma

Voidaan osoittaa, että

$$SSE_y = (1 - r_{xy}^2) SST_y$$

jossa

$$SST_y = \sum_{i=1}^{n} (y_i - \overline{y})^2 = (n-1) s_y^2$$

PNS-suoraan (4) liittyvä *jäännösvarianssin* (harhaton) estimaattori on

$$s_{(4)}^2 = \frac{SSE_x}{n-2}$$

jossa

$SSE_x = \sum_{i=1}^{n} e_{xi}^2$ = PNS-suoraan (4) liittyvä jäännösneliösumma

Voidaan osoittaa, että

$$SSE_x = (1 - r_{xy}^2) SST_x$$

jossa

$$SST_x = \sum_{i=1}^{n} (x_i - \overline{x})^2 = (n-1)s_x^2$$

Esimerkki 1. Poikien ja isien pituuksien riippuvuus toisistaan.

Perinnöllisyystieteen mukaan lapset perivät geneettiset ominaisuutensa vanhemmiltaan.

Kysymys 1: *Periytyykö isien pituus heidän pojilleen*?

Havaintoaineistona on tässä 300:n isän ja heidän poikiensa pituuksien muodostamaa lukuparia

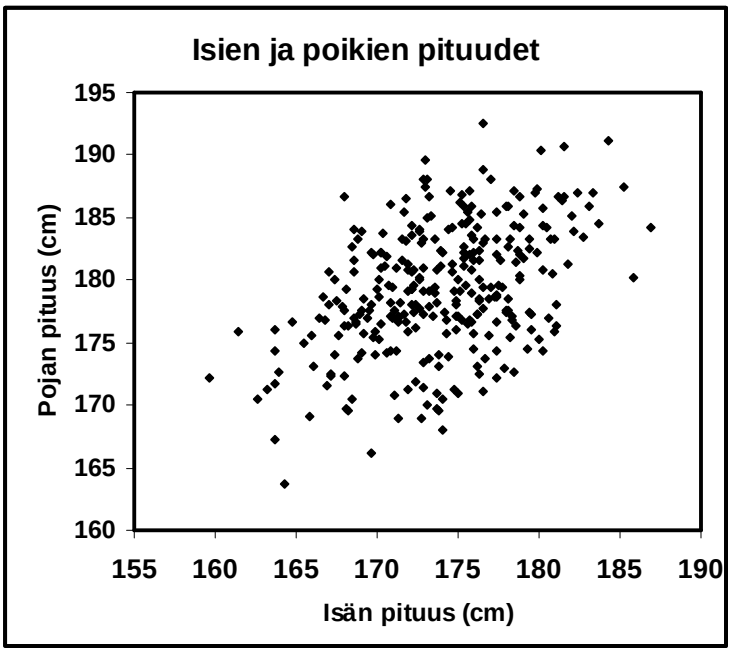
$$(x_i , y_i) , \; i = 1, 2, \ldots , 300$$

jossa siis

x_i = isän i pituus

y_i = isän i pojan pituus

Ks. pistediagrammia oikealla.



Pojan pituuden riippuvuus isän pituudesta ei selvästikään ole *eksaktia*: Saman mittaisten isien poikien pituudet näyttävät vaihtelevan paljonkin.

Kuvasta nähdään kuitenkin se, että lyhyillä isillä näyttää olevan *keskimäärin* lyhyempiä poikia kuin pitkillä isillä ja vastaavasti pitkillä isillä näyttää olevan *keskimäärin* pitempiä poikia kuin lyhyillä isillä.

Olemme luvun **Johdatus regressioanalyysiin** kappaleessa **Regressiofunktiot ja regressioanalyysi** tarkastelleet sitä, miten tällaista *tilastollista riippuvuutta* voidaan havainnollistaa.

Alla oleva taulukko esittää isien ja heidän poikiensa pituuksien *ehdollisia keskiarvoja*:

$M_k(x \mid x)$ = niiden *isien* pituuksien keskiarvo, joiden *oma* pituus kuuluu x-väliin k , k = 1, 2, 3, 4, 5, 6, 7

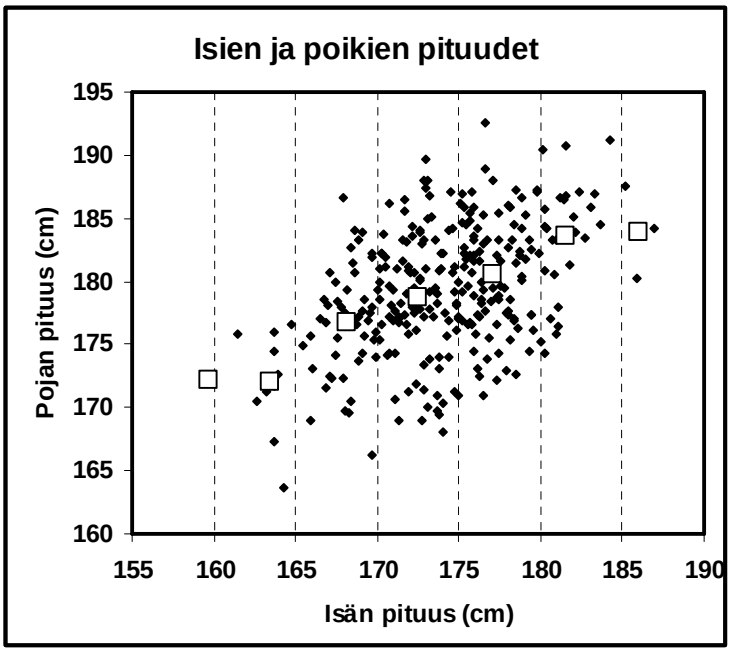
$M_k(y \mid x)$ = niiden *poikien* pituuksien keskiarvo, joiden *isien* pituus kuuluu x-väliin k , k = 1, 2, 3, 4, 5, 6, 7

x-välin nro	x-väli	$M_k(x \mid x)$	$M_k(y \mid x)$
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

Lisätään ehdollisten keskiarvojen määräämät pisteet

$$(M_k(x \mid x)\,,\, M_k(y \mid x))\,,\, k = 1,2,3,4,5,6,7$$

edellä esitettyyn pistediagrammiin; ks. kuviota alla.



Ehdollisten keskiarvojen määräämiä pisteitä on merkitty kuviossa *neliöillä*.

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

Yo. kuvioon neliöillä merkityt, *ehdollisten keskiarvojen* määräämät pisteet

$$(M_k(x \mid x)\,,\, M_k(y \mid x))\,,\, k = 1,2,3,4,5,6,7$$

kuvaavat poikien pituuksien *keskimääräistä* tai *tilastollista riippuvuutta* heidän isiensä pituuksista. Kuvion mukaan riippuvuus näyttää olevan lähes *lineaarista*. *Regressioanalyysin tehtävänä* on juuri tällaisten *tilastollisten riippuvuuksien mallintaminen*.

Kirjoittamalla

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\, i = 1,2,\mathrm{K}\,,n = 300$$

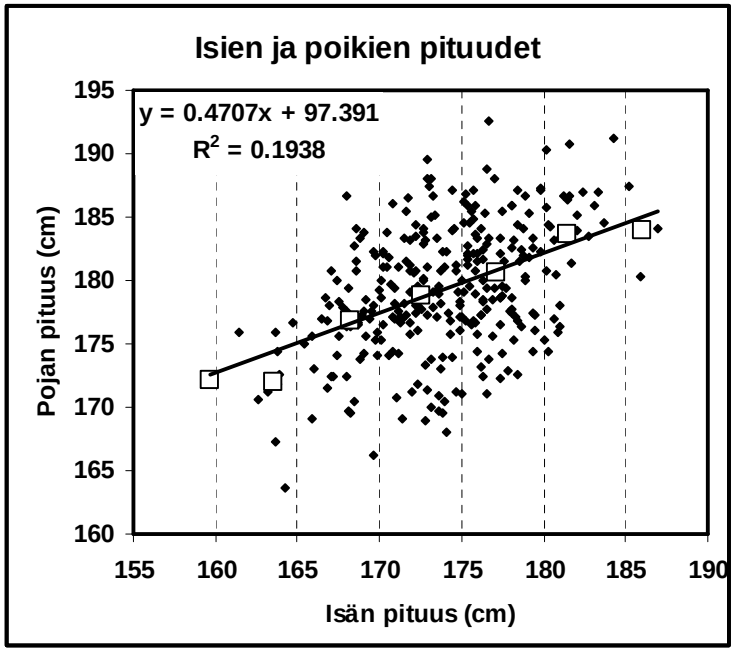
ja käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

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

kertoimet β_0 ja β_1 siten, että neliösumma

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

minimoituu. Näin määrätty suora on lisätty alla olevaan kuvioon.



Estimoidun regressiosuoran yhtälö on

$$y = 97.391 + 0.4707x$$

ja sitä vastaava selitysaste on

$$R^2 = 0.194$$

Tarkastellaan nyt seuraavaa, kysymykseen 1 nähden käänteistä kysymystä:

Kysymys 2: *Jos pojan pituus tunnetaan, voimmeko ennustaa hänen isänsä pituuden*?

Alla oleva taulukko esittää isien ja heidän poikiensa pituuksien *ehdollisia keskiarvoja*:

$M_k(x \mid y)$ = niiden *isien* pituuksien keskiarvo,
joiden *poikien* pituus kuuluu y-väliin k , k = 1, 2, 3, 4, 5, 6, 7

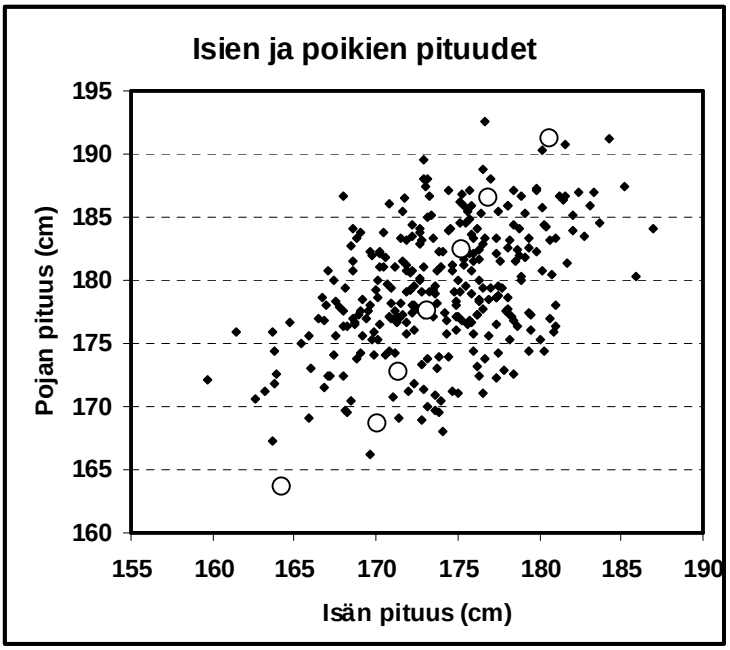
$M_k(y \mid y)$ = niiden *poikien* pituuksien keskiarvo,
joiden *oma* pituus kuuluu y-väliin k , k = 1, 2, 3, 4, 5, 6, 7

y-välin nro	y-väli	$M_k(x \mid y)$	$M_k(y \mid y)$
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

Lisätään ehdollisten keskiarvojen määräämät pisteet

$$(M_k(x \mid y)\ ,\ M_k(y \mid y))\ ,\ k = 1,2,3,4,5,6,7$$

edellä esitettyyn pistediagrammiin; ks. kuviota alla.



Ehdollisten keskiarvojen määräämiä pisteitä on merkitty kuviossa *ympyröillä*.

Havainnot on siis luokiteltu *poikien* pituuden mukaan 7 luokkaan. Kuviossa luokkia on kuvattu katkoviivojen erottamilla *vaakavöillä*. Jokaisen *ympyrän koordinaatit* on saatu laskemalla keskiarvot kaikista ko. neliötä vastaavaan *vaakavyöhön* kuuluvien havainto-pisteiden koordinaateista.

Yo. kuvioon neliöillä merkityt, *ehdollisten keskiarvojen* määräämät pisteet

$$(M_k(x \mid y)\,,\, M_k(y \mid y))\,,\; k = 1,2,3,4,5,6,7$$

kuvaavat isien pituuksien *keskimääräistä* tai *tilastollista riippuvuutta* heidän pokiensa pituuksista. Kuvion mukaan riippuvuus näyttää olevan lähes *lineaarista*. *Regressioanalyysin tehtävänä* on juuri tällaisten *tilastollisten riippuvuuksien mallintaminen*.

Kirjoittamalla

$$x_i = \alpha_0 + \alpha_1 x_i + \delta_i\,,\; i = 1,2,\mathrm{K}\,,n = 300$$

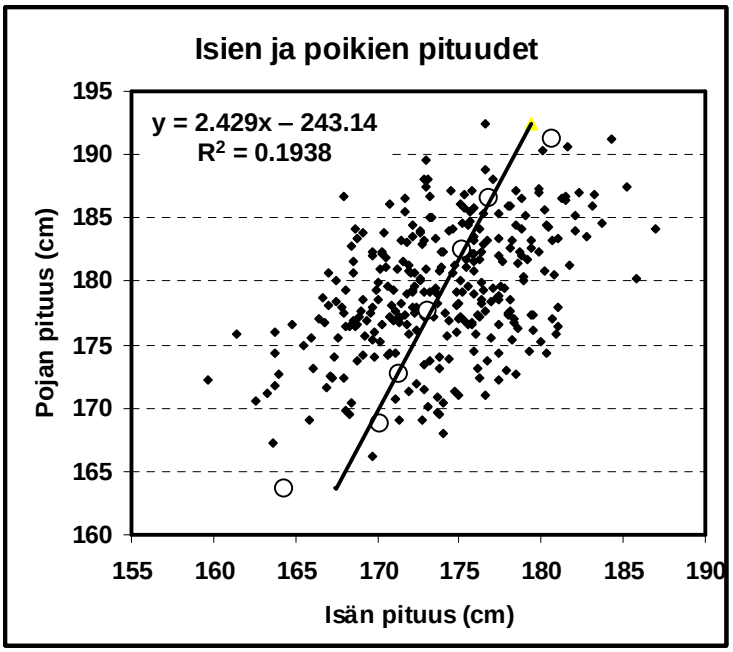
ja käyttämällä *pienimmän neliösumman keinoa* voimme määrätä suoran

$$y = \alpha_0 + \alpha_1 x$$

kertoimet α_0 ja α_1 siten, että neliösumma

$$\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} (x_i - \alpha_0 - \alpha_1 x_i)^2$$

minimoituu. Näin määrätty suora on lisätty alla olevaan kuvioon.



Estimoidun regressiosuoran yhtälö on

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

ja sitä vastaava selitysaste on

$$R^2 = 0.194$$

Yhtälö voidaan muuttujan x funktiona kirjoittaa muotoon

$$y = -243.14 + 2.429\,x$$

Alla on vielä tehtävän aineistoa kuvaava pistediagrammi, johon on piirretty sekä mallia

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i\,,\ i = 1, 2, \mathrm{K}\,, 300$$

että mallia

$$x_i = \alpha_0 + \alpha_1 x_i + \delta_i\,,\ i = 1, 2, \mathrm{K}\,, 300$$

vastaavat estimoidut regressiosuorat.

Muuttujan y (= *pojan* pituus) regressiosuora muuttujan x (= *isän* pituus) suhteen on suorista *loivempi* ja sen yhtälö on siis

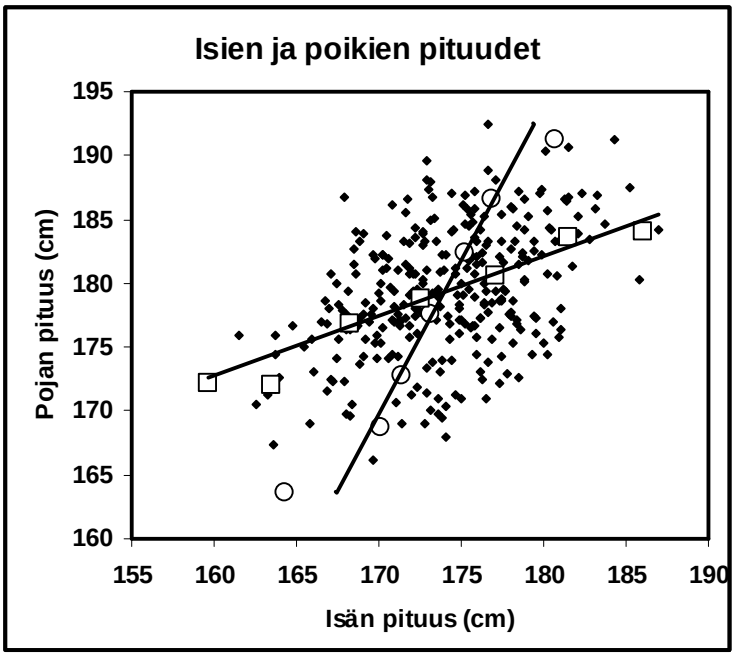
$$y = 97.391 + 0.4707\,x$$

Muuttujan x (= *isän* pituus) regressiosuora muuttujan x (= *pojan* pituus) suhteen on suorista *jyrkempi* ja sen yhtälö on siis (muuttujan x funktiona)

$$y = -243.14 + 2.429\,x$$

Huomaa, että molemmilla malleilla on *sama* selitysaste

$$R^2 = 0.194$$



Kaksiulotteisen normaalijakauman regressiofunktioiden estimointi momenttimenetelmällä ja suurimman uskottavuuden menetelmällä

Vertaamalla edellä esitettyjä kaksiulotteisen normaalijakauman regressiofunktioiden PNS-estimaattoreiden kaavoja kaksiulotteisen normaalijakauman regressiofunktioiden lausekkeisiin nähdään välittömästi, että regressiofunktioiden PNS-estimaattorit yhtyvät niiden **momentti-estimaattoreihin**.

Voidaan myös osoittaa, että regressiofunktioiden PNS-estimaattorit ovat samat kuin niiden **suurimman uskottavuuden estimaattorit**.

16. Yleinen lineaarinen malli

Regressioanalyysi on ehkä *eniten sovellettu tilastotieteen menetelmä*. **Yleinen lineaarinen malli** pyrkii *selittämään selitettävän muuttujan havaittujen arvojen vaihtelun yhden tai useamman selittävän muuttujan havaittujen arvojen vaihtelun avulla* ja on *yhden selittäjän lineaarisen regressiomallin yleistys* useamman selittäjän tapaukseen.

Tässä luvussa tarkastellaan seuraavia **usean 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 lineaarisia regressiomalleja tarkastellaan ao, luvussa.

Avainsanat:

Aritmeettinen keskiarvo, Ehdollinen jakauma, Ehdollinen odotusarvo, Ei-satunnaisuus, Ennuste, Ennustaminen, Ennustusvirhe, Estimaattori, Estimointi, *F*-testi, Gaussin ja Markovin lause, Harhattomuus, Havainto, Heteroskedastisuus, Homoskedastisuus, Jäännösneliösumma, Jäännöstermi, Jäännösvaihtelu, Jäännösvarianssi, Keskihajonta, Keskineliövirhe, Kokonaisneliösumma, Kokonaisvaihtelu, Korrelaatio, Kovarianssi, Lineaarinen regressiomalli, Lineaarisuus, Malli, Mallin hyvyys, Mallineliösumma, Minimointi, Modifioidut standardioletukset, Neliösumma, Normaalisuusoletus, Odotusarvo, Otos, Otosjakauma, Otostunnusluku, Painopiste, Parametri, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Rakenneosa, Regressioanalyysi, Regressiodiagnostiikka, Regressiokerroin, Regressiomalli, Regressiotaso, Residuaali, Satunnaisuus, Satunnainen osa, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Selitysaste, Sovite, Standardioletus, Systemaattinen osa, *t*-testi, Tarkentuvuus, Tehokkuus, Testi, Tyhjentävyys, Usean selittäjän lineaarinen regressiomalli, Vakioselittäjä, Varianssi, Varianssianalyysihajotelma, Virhetermi, Usean selittäjän lineaarinen regressiomalli, Yleinen lineaarinen malli

16.1. Yleinen lineaarinen malli ja sitä koskevat oletukset

Oletetaan, että **selitettävän muuttujan** *y havaittujen arvojen vaihtelu halutaan selittää* **selittävien muuttujien** eli **selittäjien** $x_1, x_2, \ldots, x_k$ *havaittujen arvojen vaihtelun avulla*.

Tehdään muuttujista y ja $x_1, x_2, \ldots, x_k$ seuraavat **perusoletukset**:

- Selitettävä muuttuja y on *suhdeasteikollinen satunnaismuuttuja*.
- Selittävät muuttujat $x_1, x_2, \ldots, x_k$ ovat *kiinteitä* eli *ei-satunnaisia muuttujia*.

Huomautus:

- Satunnaisten selittävien muuttujien tapausta käsitellään myöhemmin erikseen.

Havainnot

Olkoot

$$y_i, \; i = 1, 2. \ldots, n$$

selitettävän muuttujan y ja

$$x_{ij}, x_{ij}, \ldots, x_{ij}, \; i = 1, 2, \ldots, n$$

selittävän muuttujan x_j **havaittuja arvoja**. Oletetaan lisäksi, että havainnot y_j ja x_{ij} liittyvät *samaan havaintoyksikköön* $j = 1, 2, \ldots, n$ kaikille $i = 1, 2, \ldots, k$.

Tällöin voimme järjestää selitettävän muuttujaa y ja selittäjien $x_1, x_2, \ldots, x_k$ havaitut arvot havaintoyksiköittäin seuraavalla tavalla:

Havaintoyksikkö 1: $x_{11}, x_{12}, \ldots, x_{1k}, y_1$

Havaintoyksikkö 2: $x_{21}, x_{22}, \ldots, x_{2k}, y_2$

…

Havaintoyksikkö n: $x_{n1}, x_{n2}, \ldots, x_{nk}, y_n$

Havaintoyksikkökohtaisia havaintoarvoja vastaamaan voidaan asettaa *pisteet* $(k + 1)$-ulotteisessa avaruudessa:

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}, y_i) \in \mathrm{R}^{k+1}, \; i = 1, 2, \mathrm{K}, n$$

Havaintopisteen

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}, y_i), \; i = 1, 2, \mathrm{K}, n$$

koordinaateilla on seuraavat tulkinnat:

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

x_{ij} = **selitettävän muuttujan** eli **selittäjän** x_j *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä i, $j = 1, 2, \ldots, k$

k = (*aitojen*) *selittäjien* x_j lukumäärä

n = *havaintojen* lukumäärä

Yleinen lineaarinen malli

Oletetaan, että muuttujien y ja $x_1, x_2, \ldots, x_k$ havaittujen arvojen välillä vallitsee *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

jossa

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

x_{ij} = **selittävän muuttujan** eli **selittäjän** x_j *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä i, $j = 1, 2, \ldots, k$

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

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

β_j = **selittäjän** x_j **regressiokerroin**, $j = 1, 2, \ldots, k$; β_j on *ei-satunnainen* ja *tuntematon vakio*

Tällöin yhtälö määrittelee **usean selittäjän lineaarisen regressiomallin**, jota kutsutaan **yleiseksi lineaariseksi malliksi**.

Huomautuksia:

- Regressiokertoimet $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ on oletettu *samoiksi* kaikille havaintoyksiköille $i = 1, 2, \ldots, k$..
- Kerrointa β_0 kutsutaan *vakioselittäjän* regressiokertoimeksi, koska sitä vastaa *keinotekoinen selittäjä*, joka saa kaikille havaintoyksiköille $j = 1, 2, \ldots, n$ *vakioarvon* 1.
- Jatkossa esitettävät kaavat *eivät välttämättä päde* tässä esitettävässä muodossa, jos mallissa *ei ole* vakioselittäjää.
- Oletamme jatkossa, että mallissa on *aina* vakioselittäjä.

Mallia koskevat standardioletukset

Olkoon

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

yleinen lineaarinen malli.

Mallista tehdään tavallisesti seuraavat 6 oletusta, joita kutsutaan *yleistä lineaarista mallia koskeviksi* **standardioletuksiksi**. Näiden oletuksien voimassaolo takaa sen, että jatkossa esiteltäviä ns. tavanomaisia estimointi- ja testausmenetelmiä *saa* käyttää mallin analysointiin.

(i) Selittäjän x_j havaitut arvot x_{ij} ovat *kiinteitä* eli *ei-satunnaisia vakioita* kaikille

$j = 1, 2, \ldots, n$; $i = 1, 2, \ldots, k$

(ii) Selittäjien välillä ei ole *lineaarisia riippuvuuksia*.

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

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

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

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

Kommentteja standardioletuksiin

(i) Selittäjän x_j havaitut arvot x_{ij} ovat *kiinteitä* eli *ei-satunnaisia vakioita* kaikille

$j = 1, 2, \ldots, n$; $i = 1, 2, \ldots, k$

Lineaaristen regressiomallien teoria nojaa voimakkaasti oletukseen (i).
Oletus (i) on kuitenkin sangen rajoittava ja se voi toteutua käytännössä vain sellaisissa tilanteissa, joissa selittäjien arvot *valitaan*. Selittäjien arvot voidaan valita *puhtaissa koeasetelmissa*, mutta harvoin muunlaisissa tutkimusasetelmissa.

Vaikka standardioletus (i) on sangen rajoittava, tässä luvussa esitettävää lineaaristen regressiomallien teoriaa voidaan soveltaa – jos sopivat lisäehdot pätevät – myös monissa sellaisissa tilanteissa, joissa selittäjien arvot ovat satunnaisia; ks. kappaletta **Yleinen lineaarinen malli ja satunnaiset selittäjät**.

(ii) Selittäjien välillä ei ole *lineaarisia riippuvuuksia*.

Asialooginen perustelu oletukselle (ii): Jos selittäjä x_j *riippuu lineaarisesti* muista selittäjistä, muuttuja x_j on selittäjänä *redundantti* ja voidaan poistaa mallista.

Tekninen perustelu oletukselle (ii): Ehto (ii) takaa sen, että *pienimmän neliösumman menetelmä* tuottaa regressiokertoimille $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ yksikäsitteiset estimaattorit suljetussa muodossa.

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

Oletuksen (iii) mukaan *kaikilla* jäännös- eli virhetermeillä ε_i on *sama odotusarvo*.
Oletuksesta (iii) seuraa, että mallin *rakenneosan*

$$\text{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} \text{ , } i = 1, 2, \mathrm{K}, n$$

formuloinnissa ei ole tehty systemaattista virhettä.

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

Oletusta (iv) kutsutaan **homoskedastisuusoletukseksi** ja sen mukaan *kaikilla* jäännös- eli virhetermeillä ε_i on *sama varianssi*. Jos oletus (iv) *pätee*, jäännöstermejä ε_i sanotaan **homoskedastisiksi**.

Jos oletus (iv) *ei päde* ja jäännöstermien ε_i varianssi vaihtelee, jäännöstermejä ε_i sanotaan **heteroskedastisiksi**. *Heteroskedastisuus* tekee regressiokertoimien tavanomaisista estimaattoreista *tehottomia*.

Homoskedastisuusoletusta voidaan testata tilastollisesti.

Oletuksien (iii) ja (iv) mukaan jäännös- eli virhetermit ε_i vaihtelevat satunnaisesti *nollan ympärillä*.

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

Oletusta (v) kutsutaan **korreloimattomuusoletukseksi** ja sen mukaan jäännös- eli virhetermit ε_i *eivät korreloi keskenään*.

Jos oletus (v) *ei päde*, jäännöstermit ε_i ovat **korreloituneita**. *Korreloituneisuus* tekee regressiokertoimien tavanomaisista estimaattoreista *tehottomia* ja jopa *harhaisia*.

Korreloimattomuusoletusta voidaan testata tilastollisesti.

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

Oletusta (vi) kutsutaan **normaalisuusoletukseksi** ja sen mukaan jäännös- eli virhetermit ε_i ovat *normaalijakautuneita*. Oletus (vi) pitää sisällään oletukset (iii) ja (iv).

Normaalisuusoletusta voidaan *testata* tilastollisesti.

Ellei asiasta erikseen huomauteta, **oletamme jatkossa aina**, **että oletukset (i)-(vi)** ***pätevät***. Oletukset (i)-(vi) takaavat sen, että yleisen lineaarisen mallin *estimointi* ja *testaus* voidaan tehdä jatkossa esitettävällä tavalla; ks. kappaleet **Usean selittäjän lineaarisen mallin estimointi** ja **Usean selittäjän lineaarisen mallin parametreja koskevat testit**.

Mallista tehtyjen oletuksia voidaan tutkia *regressiodiagnostiikan* avulla; ks. lukua **Regressiodiagnostiikka**. *Selittäjien valintaan* liittyviä ongelmia tarkastellaan luvussa **Regressiomallin valinta**.

On syytä tietää, että osaa oletuksista (i)-(vi) voidaan lieventää tai niistä voidaan jopa luopua, mutta tällöin saattaa olla syytä käyttää muita kuin tässä esitettäviä estimointi- ja testausmenetelmiä.

Selitettävän muuttujan ominaisuudet

Jos yleistä lineaarista mallia

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,, i = 1, 2, \mathrm{K}, n$$

koskevat standardioletukset (i)-(vi) pätevät, mallin selitettävän muuttujan y havaituilla arvoilla y_i on seuraavat *stokastiset ominaisuudet*:

(iii)´ $\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik}$, $i = 1, 2, \mathrm{K}, n$

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

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

(vi)´ $y_i \sim \mathrm{N}(\mathrm{E}(y_i) , \sigma^2)$, $i = 1, 2, \ldots, n$

Mallin parametrit

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,, i = 1, 2, \mathrm{K}, n$$

parametreja ovat mallin **regressiokertoimet** β_0, β_1, β_2, … , β_k sekä jäännös- eli virhetermien ε_i yhteinen *varianssi* σ^2, jota kutsutaan **jäännösvarianssiksi**. Koska regressiokertoimet β_0, β_1, β_2, … , β_k ja jäännösvarianssi σ^2 ovat tavallisesti *tuntemattomia*, ne on *estimoitava* muuttujien x_1, x_2, … , x_k ja y havaituista arvoista.

Mallin systemaattinen osa ja satunainen osa

Oletetaan, että yleistä lineaarista mallia

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

koskevat standardioletukset (i)-(v) pätevät. Tällöin selitettävän muuttujan y havaitut arvot y_j voidaan *esittää* seuraavalla tavalla *kahden osatekijän summana*:

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

jossa osatekijää

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} \,,\; i = 1, 2, \mathrm{K}, n$$

kutsutaan mallin **systemaattiseksi** eli **rakenneosaksi** ja osatekijää ε_i kutsutaan mallin **satunnaiseksi osaksi**. Mallin rakenneosa E(y_i) *riipuu* selittäjien x_j , j = 1, 2, … , k saamista arvoista, mutta standardioletusten pätiessä mallin satunnainen osa ε_i *ei riipu* selittäjien x_j , j = 1, 2, … , k saamista arvoista.

Regressiotaso

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

systemaattinen osa

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} \,,\; i = 1, 2, \mathrm{K}, n$$

määrittelee tason

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathrm{L} + \beta_k x_k$$

avaruudessa $^{k+1}$. Tasoa kutsutaan **regressiotasoksi**. Jäännös- eli virhetermien ε_i varianssi σ^2 kuvaa *havaintopisteiden*

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}, y_i) \,,\; i = 1, 2, \mathrm{K}, n$$

vaihtelua regressiotason ympärillä.

Regressiokertoimien tulkinta

Esitetään yleisen lineaarisen mallin määrittelemän regressiotason

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \mathrm{L} + \beta_k x_k$$

kertoimien β_1, β_2, … , β_k **tulkinta**. Jos selittäjän x_j arvo kasvaa yhdellä yksiköllä:

$$x_j \;\rightarrow\; x_j + 1$$

ja *kaikkien muiden selittäjien arvot pysyvät muuttumattomina*, niin selittäjän x_j regressiokerroin β_j kertoo paljonko selitettävän muuttujan y odotettavissa oleva arvo muuttuu:

$$\mathrm{E}(y_i) \;\rightarrow\; \mathrm{E}(y_i) + \beta_j$$

16.2. Yleisen lineaarisen mallin matriisiesitys

Yleinen lineaarinen malli

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

voidaan esittää **matriisein** muodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \mathsf{M} \\ y_n \end{bmatrix}$$

on *selitettävän muuttujan* havaittujen arvojen y_i , i = 1, 2, … , n muodostama n×1-matriisi,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \mathsf{L} & x_{1k} \\ 1 & x_{21} & x_{22} & \mathsf{L} & x_{2k} \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & & \mathsf{M} \\ 1 & x_{n1} & x_{n2} & \mathsf{L} & x_{nk} \end{bmatrix}$$

on *selittävän muuttujan* havaittujen arvojen x_{ij} , i = 1, 2, … , n , j = 1, 2, … , k ja ykkösten muodostama n×(k+1)-matriisi,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \mathsf{M} \\ \beta_k \end{bmatrix}$$

on *regressiokertoimien* β_j , j = 1, 2, … , k muodostama (k + 1)×1-matriisi ja

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \mathsf{M} \\ \varepsilon_n \end{bmatrix}$$

on *jäännöstermien* ε_i , i = 1, 2, … , n muodostama n×1-matriisi.

Odotusarvovektori ja kovarianssimatriisi

Olkoon

$$\mathbf{z} = (z_1, z_2, \ldots, z_p)$$

satunnaismuuttujien z_1, z_2, … , z_p muodostama p-vektori.

Määritellään satunnaisvektorin **z odotusarvovektori** $\boldsymbol{\mu}$ kaavalla

$$\boldsymbol{\mu} = \mathrm{E}(\mathbf{z}) = (\mathrm{E}(z_1), \mathrm{E}(z_2), \mathsf{K}, \mathrm{E}(z_p))$$

p-vektorin $\boldsymbol{\mu} = \mathrm{E}(\mathbf{z})$ i. alkio μ_i on satunnaismuuttujan z_i *odotusarvo*:

$$\mu_i = \mathrm{E}(z_i)\,,\; i = 1, 2, \mathsf{K}\,, p$$

Määritellään satunnaisvektorin **z kovarianssimatriisi Σ** kaavalla

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{z}) = [\sigma_{ij}]$$

$p \times p$-matriisin $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{z})$ i. rivin ja j. sarakkeen alkio σ_{ij} on satunnaismuuttujien z_i ja z_j *kovarianssi*:

$$\sigma_{ij} = \mathrm{Cov}(z_i, z_j) = \mathrm{E}[(z_i - \mu_i)(z_j - \mu_j)]$$

Huomaa, että voimme kirjoittaa matriisimerkintöjä käyttäen

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{z}) = \mathrm{E}[(\mathbf{z} - \boldsymbol{\mu})(\mathbf{z} - \boldsymbol{\mu})']$$

Standardioletukset matriisimuodossa

Jos yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

selittäjät $x_1, x_2, \ldots, x_k$ *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** (i)-(vi) voidaan esittää matriisein seuraavassa muodossa:

(i) Matriisin **X** alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*.

(ii) Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii) $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$

(iv)&(v) $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

(vi) $\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Yllä vektori **0** on n-ulotteinen *nollavektori*:

$$\mathbf{0} = (0, 0, \ldots, 0)$$

ja matriisi **I** on $n \times n$-*yksikkömatriisi*:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & \mathsf{L} & 0 \\ 0 & 1 & 0 & \mathsf{L} & 0 \\ 0 & 0 & 1 & \mathsf{L} & 0 \\ \mathsf{M} & \mathsf{M} & \mathsf{M} & \mathsf{O} & \mathsf{M} \\ 0 & 0 & 0 & \mathsf{L} & 1 \end{bmatrix}$$

Merkintä

$$\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

tarkoittaa, että mallin jäännöstermien muodostama vektori **ε** noudattaa n-ulotteista *multinormaali-jakaumaa*, jonka *odotusarvovektori* on

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

ja kovarianssimatriisi on

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

16.3. Yleisen lineaarisen mallin parametrien estimointi

Pienimmän neliösumman estimointimenetelmä

Olkoon

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_p x_{ip} + \varepsilon_i \ , i = 1, 2, \mathrm{K}, n$$

yleinen lineaarinen malli, joka toteuttaa *standardioletukset*.

Regressiokertoimet β_0, β_1, β_2, … , β_k estimoidaan tavallisesti **pienimmän neliösumman** (**PNS**-) **menetelmällä**. PNS-menetelmässä regressiokertoimien estimaattorit määrätään *minimoimalla jäännös-* eli *virhetermien* ε_i *neliösumma*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \mathrm{L} - \beta_k x_{ik})^2$$

regressiokertoimien β_0, β_1, β_2, … , β_k suhteen.

Neliösumman

$$\sum \varepsilon_i^2$$

minimointi voidaan tehdä *derivoimalla* neliösumma regressiokertoimien suhteen ja merkitsemällä derivaatat nolliksi. Tämä johtaa regressiokertoimien suhteen *lineaariseen yhtälöryhmään*, jossa on (k + 1) yhtälöä. Yhtälöryhmällä on ratkaisu, *jos standardioletus*

(ii) $\mathrm{r}(\mathbf{X}) = k + 1$

pätee. Yhtälöryhmän ratkaisuina saadaan regressiokertoimien β_0, β_1, β_2, … , β_k **PNS-estimaattorit**. Merkitään estimaattoreita vastaavilla *latinalaisilla* kirjaimilla:

b_j = kertoimen β_j PNS-estimaattori, j = 0, 1, 2, … , k

Regressiokertoimien β_0, β_1, β_2, … , β_k PNS-estimaattoreiden b_0, b_1, b_2, … , b_k lausekkeet on mukavinta esittää *matriisimuodossa*; ks. seuraavaa kappaletta.

Regressiokertoimien vektorin PNS-etimaattori

Jos yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

matriisin $\mathbf{X}$ sarakkeet ovat *lineaarisesti riippumattomia* eli, jos standardioletus

(ii) $r(\mathbf{X}) = k + 1$

pätee, niin vektorin $\boldsymbol{\beta}$ *PNS-estimaattori* $\mathbf{b}$ voidaan esittää matriisein muodossa

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Perustelu:

Kirjoitetaan ensin

$$
\begin{aligned}
&(\mathbf{y}-\mathbf{X}(\boldsymbol{\beta}+\mathbf{h}))'(\mathbf{y}-\mathbf{X}(\boldsymbol{\beta}+\mathbf{h})) \\
&\quad=((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\mathbf{X}\mathbf{h})'((\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-\mathbf{X}\mathbf{h}) \\
&\quad=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})-2\mathbf{h}'\mathbf{X}'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+\mathbf{h}'\mathbf{X}'\mathbf{X}\mathbf{h} \\
&\quad=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})+2\mathbf{h}'(-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\boldsymbol{\beta})+\mathbf{h}'\mathbf{X}'\mathbf{X}\mathbf{h}
\end{aligned}
$$

Antamalla $\mathbf{h} \to \mathbf{0}$ nähdään, että neliömuodon

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

derivaatta on

$$\frac{\partial}{\partial\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=2(-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\boldsymbol{\beta})$$

Merkitsemällä derivaatta nollaksi saadaan *normaaliyhtälö*

$$-\mathbf{X}'\mathbf{y}+\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=0$$

jonka ratkaisuksi saadaan

$$\hat{\boldsymbol{\beta}}=\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Sama normaaliyhtälö kuin yllä saadaan myös derivoimalla neliömuoto

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}=(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$$

vektorin $\boldsymbol{\beta}$ suhteen, kun sovelletaan seuraavia *matriisien derivointisääntöjä*:

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}=2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

$$\frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\beta}'\mathbf{X}'\mathbf{y}=\mathbf{X}'\mathbf{y}$$

Ratkaisu vastaa neliösumman $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ *minimiä*, koska

$$\frac{\partial^2}{\partial\boldsymbol{\beta}'\partial\boldsymbol{\beta}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})=2\mathbf{X}'\mathbf{X}>0$$

jossa merkintä $2\mathbf{X}'\mathbf{X} > 0$ tarkoittaa sitä, että matriisi $2\mathbf{X}'\mathbf{X}$ on *positiivisesti definiitti*.

■

PNS-estimaattorin odotusarvovektori ja kovarianssimatriisi

Olkoon

$$\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori.

Jos standardioletukset (i)-(v) pätevät, niin

(i) $\mathrm{E}(\mathbf{b})=\boldsymbol{\beta}$

(ii) $\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Perustelu:

Todetaan ensin, että regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattorin $\mathbf{b}$ lauseke voidaan kirjoittaa seuraavaan muotoon:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

(i) Koska regressiokertoimien vektori ja $\boldsymbol{\beta}$ ja matriisi $\mathbf{X}$ ovat ei-satunnaisia, niin

$$\mathrm{E}(\mathbf{b}) = \mathrm{E}(\boldsymbol{\beta}) + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{0} = \boldsymbol{\beta}$$

(ii) Kohdasta (i) todistuksen mukaan

$$\mathbf{b} - \mathrm{E}(\mathbf{b}) = \mathbf{b} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

Koska matriisi $\mathbf{X}$ on ei-satunnainen, niin

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}) &= \mathrm{E}((\mathbf{b} - \mathrm{E}(\mathbf{b})(\mathbf{b} - \mathrm{E}(\mathbf{b})') \\
&= \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sigma^2\mathbf{I})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

■

Kohdan (i) mukaan PNS-estimaattori $\mathbf{b}$ on **harhaton** parametrivektorille $\boldsymbol{\beta}$.

Jos standardioletusten (i)-(v) lisäksi normaalisuusoletus (vi) pätee, niin regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori $\mathbf{b}$ noudattaa $(k + 1)$-ulotteista *multinormaalijakaumaa*, jonka *odotusarvovektori* on $\boldsymbol{\beta}$ ja *kovarianssimatriisi* on $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$:

$$\mathbf{b} \quad \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

PNS-estimaattorin $\mathbf{b}$ normaalisuus seuraa siitä multinormaalijakauman ominaisuudesta, että multinormaalijakaumaa noudattavien satunnaismuuttujien lineaarimuunnokset noudattavat multinormaalijakaumaa; lisätietoja: ks. monistetta **Monimuuttujamenetelmät**.

Gaussin ja Markovin lause

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

yleinen lineaarinen malli, joka toteuttaa standardioletukset (i)-(v).

Gaussin ja Markovin lause:

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on paras vektorin $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*.

Perustelu:

Olkoon

$$\mathbf{b}^* = \mathbf{A}^*\mathbf{y}$$

mielivaltainen regressiokertoimien vektorin *lineaarinen* ja *harhaton* estimaattori, jossa $\mathbf{A}^*$ on *ei-satunnainen* $(k + 1)\times n$-matriisi.

Määritellään $(k + 1)\times n$-matriisi $\mathbf{A}$ kaavalla

$$\mathbf{A} = \mathbf{A}^* - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Siten estimaattorin $\mathbf{b}^*$ lauseke voidaan kirjoittaa muotoon

$$\begin{aligned}\mathbf{b}^* &= \mathbf{A}^*\mathbf{y}\\ &= [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}\\ &= [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'](\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})\\ &= (\mathbf{AX} + \mathbf{I})\boldsymbol{\beta} + [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{\varepsilon}\end{aligned}$$

ja

$$\mathrm{E}(\mathbf{b}^*) = (\mathbf{AX} + \mathbf{I})\boldsymbol{\beta} + [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{AX}\boldsymbol{\beta} + \boldsymbol{\beta}$$

Siten estimaattori $\mathbf{b}^*$ voi olla harhaton parametrille $\boldsymbol{\beta}$ vain, jos

$$\mathbf{AX} = \mathbf{0}$$

jolloin siis

$$\mathrm{E}(\mathbf{b}^*) = \boldsymbol{\beta}$$

ja

$$\mathbf{b}^* - \mathrm{E}(\mathbf{b}^*) = \mathbf{b}^* - \boldsymbol{\beta} = [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{\varepsilon}$$

Siten

$$\begin{aligned}\mathrm{Cov}(\mathbf{b}^*) &= \mathrm{E}[(\mathbf{b}^* - \mathrm{E}(\mathbf{b}^*))(\mathbf{b}^* - \mathrm{E}(\mathbf{b}^*))']\\ &= \mathrm{E}[(\mathbf{b}^* - \boldsymbol{\beta})(\mathbf{b}^* - \boldsymbol{\beta})']\\ &= \mathrm{E}\{[\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'[\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'\\ &= [\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')[\mathbf{A}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\\ &= \sigma^2[\mathbf{A} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'][\mathbf{A}' + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]\\ &= \sigma^2[\mathbf{AA}' + \mathbf{AX}(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}' + (\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

Koska

$$\mathbf{AX} = \mathbf{0}$$

tämä lauseke sievenee muotoon

$$\mathrm{Cov}(\mathbf{b}^*) = \sigma^2[\mathbf{AA}' + (\mathbf{X}'\mathbf{X})^{-1}]$$

Koska matriisi $\mathbf{AA}'$ on *positiivisesti semidefiniitti matriisi* eli

$$\mathbf{AA}' \geq 0$$

olemme todistaneet, että

$$\mathrm{Cov}(\mathbf{b}^*) = \sigma^2[\mathbf{AA}' + (\mathbf{X}'\mathbf{X})^{-1}] \geq \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \mathrm{Cov}(\mathbf{b})$$

Siten olemme todistaneet Gaussin ja Markovin lauseen, koska $\mathbf{b}^*$ oli mielivaltainen.

■

Gaussin ja Markovin lauseen tulkinta

Regressiokertoimien vektorin $\boldsymbol{\beta}$ **PNS-estimaattorin b paremmuudella** tarkoitetaan Gaussin ja Markovin lauseessa seuraavaa: Jos $\mathbf{b}^*$ on mielivaltainen regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaarinen ja harhaton estimaattori*, niin tällöin

$$\mathrm{Cov}(\mathbf{b}^*) \geq \mathrm{Cov}(\mathbf{b})$$

Huomautuksia:

- Estimaattorin $\mathbf{b}^*$ *lineaarisuus*: $\mathbf{b}^*$ on muotoa

  $$\mathbf{b}^* = \mathbf{Ay}$$

  jossa $(k + 1)\times n$- matriisin $\mathbf{A}$ alkiot *eivät saa riippua* selitettävän muuttujan y havaituista arvoista.

- Estimaattorin $\mathbf{b}^*$ *harhattomuus*:

  $$\mathrm{E}(\mathbf{b}^*) = \boldsymbol{\beta}$$

- Merkinnällä

  $$\mathrm{Cov}(\mathbf{b}^*) \geq \mathrm{Cov}(\mathbf{b})$$

  tarkoitetaan sitä, että matriisi

  $$\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b})$$

  on *positiivisesti semidefiniitti* matriisi eli

  $$\mathbf{a}'(\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}))\mathbf{a} \geq 0 \text{ kaikille } \mathbf{a} \neq \mathbf{0}$$

Epäyhtälöstä

$$\mathbf{a}'(\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}))\mathbf{a} \geq 0 \text{ kaikille } \mathbf{a} \neq \mathbf{0}$$

seuraa erityisesti se, että yksittäisten regressiokertoimien PNS-estimaattoreiden b_j , j = 0, 1, 2, … , k varianssit ovat *pienimpiä mahdollisia lineaaristen ja harhattomien estimaattoreiden joukossa*:

Jos b_j^* on mikä tahansa regressiokertoimen β_j lineaarinen ja harhaton estimaattori, niin

$$\mathrm{Var}(b_j^*) \geq \mathrm{Var}(b_j)\,,\ j = 0,1,2,\mathrm{K}\,,k$$

Tämä nähdään valitsemalla vektoriksi $\mathbf{a}$ vektori, jossa ainoa nollasta poikkeava alkio 1 on paikassa j:

$$\mathbf{a} = (0,\mathrm{K}\,,0,\underset{\uparrow}{1},0,\mathrm{K}\,,0)$$
$$j.$$

PNS-estimaattorin stokastiset ominaisuudet

Yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattorilla*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on standardioletuksien (i)-(vi) pätiessä seuraavat ominaisuudet:

(1) **b** on *harhaton*.

(2) **b** on *paras* (eli *tehokkain*) *lineaaristen ja harhattomien estimaattoreiden joukossa*.

(3) **b** on *tyhjentävä*.

(4) **b** on (sopivin lisäehdoin) *tarkentuva*.

(5) **b** on *normaalinen*.

Sovitteet ja residuaalit

Olkoot yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\, i = 1, 2, \mathrm{K}, n$$

regressiokertoimien β_0, β_1, β_2, ... , β_k *PNS-estimaattorit* b_0, b_1, b_2, … , b_k .

Sovite

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \mathrm{L} + b_k x_{ik} \,,\, i = 1, 2, \mathrm{K}, n$$

on estimoidun mallin selitettävälle muuttujalle *y* antama arvo havaintopisteessä

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}) \,,\, i = 1, 2, \mathrm{K}, n$$

Jos standardioletukset (i)-(v) pätevät,

$$\mathrm{E}(\hat{y}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} \,,\, i = 1, 2, \mathrm{K}, n$$

Residuaali

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \mathrm{L} - b_k x_{ik} \,,\, i = 1, 2, \mathrm{K}, n$$

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

Jos standardioletukset (i)-(v) pätevät,

$$\mathrm{E}(e_i) = 0 \,,\, i = 1, 2, \mathrm{K}, n$$

Regressiomallin hyvyyden tutkimisessa voidaan käyttää hyväksi estimoidun mallin sovitteita ja residuaaleja :

(i) Regressiomalli selittää selitettävän muuttujan havaittujen arvojen vaihtelun sitä paremmin mitä lähempänä estimoidun mallin sovitteet $\hat{y}_i$ ovat selitettävän muuttujan y havaittuja arvoja y_i .

(ii) Regressiomalli *selittää* selitettävän muuttujan havaittujen arvojen vaihtelun *sitä paremmin mitä pienempiä ovat estimoidun mallin residuaalit* e_i .

Sovitteiden ja residuaalien matriisiesitykset

Olkoon

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori.

Määritellään **sovitteiden** $\hat{y}_i$, $i = 1, 2, \ldots, n$ muodostama n-vektori kaavalla

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b}$$

Määritellään **residuaalien** e_i , $i = 1, 2, \ldots, n$ muodostama n-vektori kaavalla

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$$

Sovitteiden muodostama n-vektori $\hat{\mathbf{y}}$ voidaan kirjoittaa seuraaviin muotoihin:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

jossa $n \times n$-matriisi

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

on *symmetrinen* ja *idempotenti* eli *projektio*:

$$\mathbf{P}' = \mathbf{P}$$

$$\mathbf{P}^2 = \mathbf{P}$$

Residuaalien muodostama n-vektori $\mathbf{e}$ voidaan kirjoittaa seuraaviin muotoihin:

$$\begin{aligned}
\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\
&= \mathbf{y} - \mathbf{X}\mathbf{b} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} \\
&= (\mathbf{I} - \mathbf{P})\mathbf{y} \\
&= \mathbf{M}\mathbf{y} \\
&= \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \qquad | \; \mathbf{M}\mathbf{X} = \mathbf{0} \text{ ; ks. todistusta alla} \\
&= \mathbf{M}\boldsymbol{\varepsilon}
\end{aligned}$$

jossa $n \times n$-matriisi

$$\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

on *symmetrinen* ja *idempotenti* eli *projektio*:

$$\mathbf{M}' = \mathbf{M}$$

$$\mathbf{M}^2 = \mathbf{M}$$

Projektiomatriisit $\mathbf{P}$ ja $\mathbf{M}$ toteuttavat yhtälön

$$\mathbf{P}\mathbf{M} = \mathbf{M}\mathbf{P} = \mathbf{0}$$

Matriisi **P** on *projektio matriisin* **X** *sarakeavaruuteen* eli matriisin **X** sarakkeiden virittämään vektorialiavaruuteen (tasoon). Tämä nähdään seuraavalla tavalla:

$$\mathbf{PX} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}$$

Matriisi **M** on *projektio matriisin* **X** *sarakeavaruuden ortogonaaliseen komplementtiin* eli vektorialiavaruuteen, joka on kohtisuorassa matriisin **X** sarakkeiden virittämää vektorialiavaruutta vastaan. Tämä nähdään seuraavalla tavalla:

$$\mathbf{MX} = (\mathbf{I} - \mathbf{P})\mathbf{X} = \mathbf{X} - \mathbf{XP} = \mathbf{X} - \mathbf{X} = \mathbf{0}$$

Matriiseja **P** ja **M** koskevilla tuloksilla on keskeinen merkitys *johdettaessa* lineaarisen mallin estimointiin ja testaukseen liittyviä *jakaumatuloksia*.

Sovitteiden ja residuaalien ominaisuudet

Sovitteilla ja residuaaleilla on seuraavat ominaisuudet:

(i) $\hat{\mathbf{y}}'\mathbf{e} = 0$

(ii) $\mathbf{1}'\mathbf{e} = 0$

(iii) $\mathbf{1}'\mathbf{y} = \mathbf{1}'\hat{\mathbf{y}}$

(iv) $\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$

Perustelu:

(i) Edellä esitetyn mukaan sovitteiden ja vastaavien residuaalien muodostamat vektorit $\hat{\mathbf{y}}$ ja **e** voidaan esittää projektiomatriisien **P** ja **M** avulla muodoissa

$$\hat{\mathbf{y}} = \mathbf{Py}$$

$$\mathbf{e} = \mathbf{My}$$

Koska edellä esitetyn mukaan $\mathbf{PM} = \mathbf{0}$, niin

$$\hat{\mathbf{y}}'\mathbf{e} = \mathbf{y}'\mathbf{P}'\mathbf{My} = \mathbf{y}'\mathbf{PMy} = 0 = \mathbf{y}'\mathbf{P}'\mathbf{My} = \mathbf{y}'\mathbf{PMy} = 0$$

(ii) Edellä esitetyn mukaan matriisi **M** on projektio matriisin **X** sarakeavaruuden ortogonaaliseen komplementtiin. Siten residuaalien muodostama vektori

$$\mathbf{e} = \mathbf{My}$$

on matriisin **X** sarakeavaruuden ortogonaalisessa komplementissa, joten vektori **e** on kohtisuorassa matriisin **X** sarakeavaruutta eli matriisin **X** sarakkeiden virittämää tasoa vastaan:

$$\mathbf{X}'\mathbf{e} = \mathbf{0}$$

Sama tulos saadaan myös suoraan laskemalla:

$$\mathbf{X}'\mathbf{e} = \mathbf{X}'(\mathbf{y} - \mathbf{Xb}) = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} = \mathbf{0}$$

Koska mallissa on mukana vakio, matriisin **X** 1. sarakkeena on vektori **1** = (1, … , 1). Siten edellä esitetystä seuraa, että

$$\mathbf{1}'\mathbf{e} = 0$$

(iii) Suoraan sovitteiden ja residuaalien muodostamien vektorien määritelmistä nähdään, että

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

Siten

$$\mathbf{1}'\mathbf{y} = \mathbf{1}'\hat{\mathbf{y}} + \mathbf{1}'\mathbf{e} = \mathbf{1}'\hat{\mathbf{y}} + 0 = \mathbf{1}'\hat{\mathbf{y}}$$

koska (ii)-kohdan mukaan $\mathbf{1}'\mathbf{e} = 0$.

(iv) Suoraan sovitteiden ja residuaalien muodostamien vektorien määritelmistä nähdään, että

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e}$$

Siten

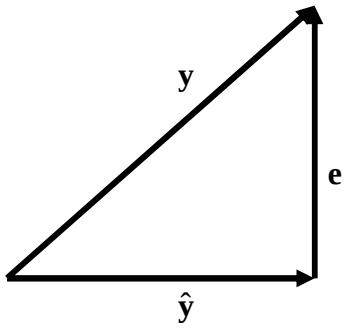
$$\mathbf{y}'\mathbf{y} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e} + 2\hat{\mathbf{y}}'\mathbf{e} = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \mathbf{e}'\mathbf{e}$$

koska (a)-kohdan mukaan $\hat{\mathbf{y}}'\mathbf{e} = 0$.

■

Huomautuksia:

- Kohdan (i) mukaan sovitteiden muodostama vektori $\hat{\mathbf{y}}$ ja residuaalien muodostama vektori **e** ovat *ortogonaalisia*.
- Kohdan (ii) mukaan residuaalien summa = 0, *jos mallissa on mukana vakio*.
- Kohdan (i) mukaan selitettävän muuttuja *y* havaituilla arvoilla ja sovitteilla on sama summa, *jos mallissa on mukana vakio*.



- Kohta (iv) on *Pythagoraan lause* (suorakulmaisessa kolmiossa kolmion hypotenuusalle piirretyn neliön pinta-ala on sama kuin kolmion kateeteille piirrettyjen neliöiden pinta-alojen summa) *n*-ulotteisessa avaruudessa.

  Ks. oikealla olevaa kuvaa.

Sovitteiden ja residuaalien stokastiset ominaisuudet

Sovitteiden muodostaman vektorin $\hat{\mathbf{y}}$ *odotusarvovektori* ja *kovarianssimatriisi*:

(i) $\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$

(ii) $\mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{P} = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

Perustelu:

(i) Koska PNS-estimaattori **b** on harhaton parametrille $\boldsymbol{\beta}$, niin

$$\mathrm{E}(\hat{\mathbf{y}}) = \mathrm{E}(\mathbf{X}\mathbf{b}) = \mathbf{X}\,\mathrm{E}(\mathbf{b}) = \mathbf{X}\boldsymbol{\beta}$$

(ii) Kohdasta (i) seuraa, että

$$\begin{aligned}
\mathrm{Cov}(\hat{\mathbf{y}}) &= \mathrm{E}[(\hat{\mathbf{y}} - \mathrm{E}(\hat{\mathbf{y}}))(\hat{\mathbf{y}} - \mathrm{E}(\hat{\mathbf{y}}))'] \\
&= \mathrm{E}[(\mathbf{Xb} - \mathbf{X\beta})(\mathbf{Xb} - \mathbf{X\beta})'] \\
&= \mathbf{X}\,\mathrm{E}[(\mathbf{b} - \boldsymbol{\beta})(\mathbf{b} - \boldsymbol{\beta})']\mathbf{X}' \\
&= \mathbf{X}\,\mathrm{Cov}(\mathbf{b})\mathbf{X}' \\
&= \mathbf{X}[\sigma^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{X}' \\
&= \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \sigma^2\mathbf{P}
\end{aligned}$$

■

Residuaalien muodostama vektorin **e** *odotusarvovektori* ja *kovarianssimatriisi*:

(i) $\mathrm{E}(\mathbf{e}) = \mathbf{0}$

(ii) $\mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{P}) = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$

Perustelu:

(i) Koska $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$, niin

$$\mathrm{E}(\mathbf{e}) = \mathbf{M}\,\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

(ii) Kohdasta (i) ja siitä, että $\mathbf{e} = \mathbf{M}\boldsymbol{\varepsilon}$ ja $\mathbf{M}$ on *projektiomatriisi* eli *symmetrinen* ja *idempotentti*, niin

$$\begin{aligned}
\mathrm{Cov}(\mathbf{e}) &= \mathrm{E}[(\mathbf{e} - \mathrm{E}(\mathbf{e}))(\mathbf{e} - \mathrm{E}(\mathbf{e}))'] \\
&= \mathrm{E}(\mathbf{ee}') \\
&= \mathrm{E}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'\mathbf{M}') \\
&= \mathbf{M}\,\mathrm{E}(\mathbf{ee}')\mathbf{M} \\
&= \mathbf{M}(\sigma^2\mathbf{I})\mathbf{M} \\
&= \sigma^2\mathbf{M}^2 \\
&= \sigma^2\mathbf{M}
\end{aligned}$$

■

Huomautus:

- Residuaalit e_i *ovat* siis (lievästi) *korreloituneita*, vaikka jäännöstermit ε_i *on oletettu korreloimattomiksi*.

Jäännösvarianssin estimointi

Jos yleisen lineaarisen mallin standardioletukset (i)-(v) pätevät, *jäännösvarianssin* σ^2 **harhaton estimaattori** on

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

jossa

e_i = estimoidun mallin *residuaali*, $i = 1, 2, \ldots, n$

n = havaintojen lukumäärä

k = (aitojen) selittäjien x_j lukumäärä

Perustelu:

Todistetaan se, että estimaattori s^2 on *harhaton* jäännösvarianssille σ^2.

Todetaan ensin, että

$$(n-k-1)s^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

jossa residuaalien muodostama n-vektorilla $\mathbf{e}$ on esitysmuodot

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\mathbf{b} = \mathbf{M}\mathbf{y} = \mathbf{M}\boldsymbol{\varepsilon}$$

jossa $n \times n$-matriisi

$$\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

on *symmetrinen* ja *idempotenti* eli *projektio*:

$$\mathbf{M}' = \mathbf{M}$$

$$\mathbf{M}^2 = \mathbf{M}$$

Koska

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

saamme suoraan laskemalla:

$$\begin{aligned}
\mathrm{E}(\mathbf{e}'\mathbf{e}) &= \mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\varepsilon}) \\
&= \mathrm{E}(\boldsymbol{\varepsilon}'\mathbf{M}\boldsymbol{\varepsilon}) \\
&= \mathrm{E}(\mathrm{trace}(\mathbf{M}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \\
&= \mathrm{trace}(\mathbf{M}\,\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')) \\
&= \mathrm{trace}(\mathbf{M}\,\mathrm{Cov}(\boldsymbol{\varepsilon})) \\
&= \sigma^2\,\mathrm{trace}(\mathbf{M})
\end{aligned}$$

Väite tulee todistetuksi, kun toteamme, että

$$\begin{aligned}
\mathrm{trace}(\mathbf{M}) &= \mathrm{trace}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= \mathrm{trace}(\mathbf{I}_n) - \mathrm{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\
&= n - \mathrm{trace}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) \\
&= n - \mathrm{trace}(\mathbf{I}_{k+1}) \\
&= n - k - 1
\end{aligned}$$

■

Estimaattori s^2 on *residuaalien* e_i , $i = 1, 2, \ldots, n$ *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

$$s_e^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n-k-1}\sum_{i=1}^{n} e_i^2$$

Estimoitu regressiotaso

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

regressiokertoimien β_0, β_1, β_2, … , β_k *PNS-estimaattorit* b_0, b_1, b_2, … , b_k *määrittelevät tason*

$$y = b_0 + b_1 x_1 + b_2 x_2 + \mathrm{L} + b_k x_k$$

avaruudessa $^{k+1}$. Tasoa kutsutaan **estimoiduksi regressiotasoksi**. Jäännösvarianssin σ^2 estimaattori s^2 kuvaa *havaintopisteiden*

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}, y_i) \in \mathrm{R}^{k+1} \,,\; i = 1, 2, \mathrm{K}, n$$

vaihtelua estimoidun regressiotason ympärillä.

16.4. Varianssianalyysihajotelma ja selitysaste

Regressiomallin tehtävänä on selittää *selitettävän muuttujan y havaittujen arvojen vaihtelu selittävien muuttujien* x_1, x_2, … , x_k *havaittujen arvojen vaihtelulla*. Tämän tehtävän onnistumista voidaan kuvata ns. **varianssianalyysihajotelman** avulla.

Hajotelmassa selitettävän muuttujan *y* havaittujen arvojen kokonaisvaihtelua kuvaava ns. kokonaisneliösumma jaetaan kahden osatekijän summaksi:

(i) Toinen osatekijä kuvaa mallin selittämää osaa kokonaisvaihtelusta.

(ii) Toinen osatekijä kuvaa mallilla selittämättä jäänyttä osaa kokonaisvaihtelusta.

Yleisen lineaarisen mallin selitettävän muuttujan *y* havaittujen arvojen y_i vaihtelun mittaaminen perustuu **kokonaisneliösummaan**

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

jossa

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

on selitettävän muuttujan *y* havaittujen arvojen y_i aritmeettinen keskiarvo. *Selitettävän muuttujan y havaittujen arvojen* y_i *varianssi* voidaan määritellä kaavalla

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

Residuaalien e_i vaihtelun mittaaminen perustuu **jäännösneliösummaan**

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

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

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

Koska

$$\mathrm{E}(s^2) = \sigma^2$$

niin estimaattori s^2 on *harhaton* jäännösvarianssille σ^2 .

Voidaan osoittaa, että jäännösneliösumma on korkeintaan yhtä suuri kuin kokonaisneliösumma:

$$SSE \leq SST$$

Määritellään erotus

$$SSM = SST - SSE$$

Koska

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

jossa

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{\hat{y}}$$

on selitettävän muuttujan *y* havaittujen arvojen y_i aritmeettinen keskiarvo, erotusta *SSM* kutsutaan **mallineliösummaksi**.

Edellä esitetyn mukaan kokonaisneliösumma

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

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

$$SST = SSM + SSE$$

jossa

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

ja

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

Perustelu:

Todistetaan varianssianalyysihajotelma *matriisilaskentaa* käyttäen.

Todetaan ensin, että kokonaisneliösumma *SST* voidaan kirjoittaa muotoon

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

jossa

$$\mathbf{y} = (y_1, y_2, \mathrm{K}, y_n)$$

on selitettävän muuttujan *y* havaittujen arvojen y_i muodostama *n*-vektori.

Aikaisemmin saatujen tulosten mukaan residuaalien e_i muodostama *n*-vektori

$$\mathbf{e} = (e_1, e_2, \mathrm{K}, e_n)$$

voidaan esittää muodossa

$$\mathbf{e} = \mathbf{M}\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y}$$

jossa matriisit

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

ja

$$\mathbf{M} = \mathbf{I} - \mathbf{P}$$

ovat *symmetrisiä* ja *idempotentteja*. Siten jäännösneliösumma *SSE* voidaan kirjoittaa muotoon

$$SSE = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{M}\mathbf{y} = \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{P}\mathbf{y}$$

Tarkastellaan lopuksi mallineliösummaa *SSM*. Jos voimme osoittaa, että

$$SSM = \mathbf{y}'\mathbf{P}\mathbf{y} - n\bar{y}^2$$

varianssianalyysihajotelma on todistettu.

Olemme todenneet aikaisemmin, että selitettävän muuttujan havaituilla arvoilla ja sovitteilla on sama summa:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$$

joten

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{\hat{y}}$$

Siten mallineliösumma *SSM* voidaan kirjoittaa muotoon

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

jossa

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \mathrm{K}, \hat{y}_n)$$

on sovitteiden $\hat{y}_i$ muodostama *n*-vektori. Olemme todenneet aikaisemmin, että

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y}$$

jossa matriisi **P** on *symmetrinen* ja *idempotentti*. Siten

$$\hat{\mathbf{y}}'\hat{\mathbf{y}} = \mathbf{y}'\mathbf{P}\mathbf{y}$$

ja

$$SSM = \hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{y}^2 = \mathbf{y}'\mathbf{P}\mathbf{y} - n\bar{y}^2$$

kuten halusimme.

■

Varianssianalyysihajotelma

$$SST = SSM + SSE$$

voidaan edellä esitetyn todistetun mukaan esittää muodossa

$$SST = (\mathbf{y} - \bar{y}\mathbf{1})'(\mathbf{y} - \bar{y}\mathbf{1}) = (\hat{\mathbf{y}} - \bar{y}\mathbf{1})'(\hat{\mathbf{y}} - \bar{y}\mathbf{1}) + \mathbf{e}'\mathbf{e} = SSM + SSE$$

Varianssianalyysihajotelman tulkinta

Varianssianalyysihajotelmassa

$$SST = SSM + SSE$$

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

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

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

Siten varianssianalyysihajotelma kuvaa estimoidun regressiomallin *hyvyyttä*:

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

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

Selitysaste

Varianssianalyysihajotelma

$$SST = SSM + SSE$$

motivoi tunnusluvun

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

käytön regressiomallin hyvyyden mittarina. Tunnuslukua R^2 kutsutaan **selitysasteeksi** ja se mittaa *regressiomallin selittämää osuutta selitettävän muuttujan havaintoarvojen kokonaisvaihtelusta*.

Koska

$$0 \le R^2 \le 1$$

selitysaste ilmaistaan tavallisesti prosentteina:

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

Voidaan osoittaa, että

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

jossa

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

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

Selitysasteen ominaisuudet

Selitysasteella R^2 on seuraavat ominaisuudet:

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

(ii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 1$

(2) Kaikki residuaalit häviävät:

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

(3) Kaikki havaintopisteet

$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$

asettuvat samalle tasolle.

(4) Malli *selittää täydellisesti* selitettävän muuttujan y havaittujen arvojen y_i vaihtelun.

(iii) Seuraavat ehdot ovat *yhtäpitäviä*:

(1) $R^2 = 0$

(2) $b_1 = b_2 = \cdots = b_k = 0$

(3) Malli *ei ollenkaan selitä* selitettävän muuttujan y havaittujen arvojen y_i vaihtelua.

16.5. Tilastollinen päättely yleisestä lineaarisesta mallista

Olkoon

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori.

Tarkastelemme tässä kappaleessa seuraavia yleistä lineaarista mallia koskevia päättelyn ongelmia:

- **Regressiokertoimien estimaattoreiden *odotusarvot, varianssit* ja *otosjakaumat***
- **Regressiokertoimien luottamusvälit**
- **Yleistesti regression olemassaololle**
- **Testit yksittäisille regressiokertoimille**

Regressiokertoimien estimaattoreiden odotusarvot, varianssit ja otosjakaumat

Olemme todenneet jo aikaisemmin, että jos standardioletukset (i)-(v) pätevät, regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattorilla

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on seuraavat *stokastiset ominaisuudet*:

(i) $\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$

(ii) $\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

Jos myös standardioletus (vi) pätee, PNS-estimaattori **b** noudattaa **normaalijakaumaa**:

$$\mathbf{b} \quad \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Perustelu:

Koska

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

ja

$$\mathbf{y} \quad \mathrm{N}_{n+1}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

PNS-estimaattorin **b** normaalisuus seuraa siitä multinormaalijakauman ominaisuudesta, että multinormaalijakaumaa noudattavien satunnaismuuttujien lineaarimuunnokset noudattavat multinormaalijakaumaa; ks. monistetta **Monimuuttujamenetelmät**.

■

Edellä esitetystä seuraa, että jos standardioletukset (i)-(v) pätevät, regressiokertoimen β_j PNS-estimaattorilla b_j on seuraavat *stokastiset ominaisuudet*:

(i) $\mathrm{E}(b_j) = \beta_j\ ,\ j = 0,1,2,\mathrm{K},k$

(ii) $\mathrm{D}^2(b_j) = \sigma_{b_j}^2 = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,\,j+1}\ ,\ j = 0,1,2,\mathrm{K},k$

jossa

$$[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,\,j+1}\ ,\ j = 0,1,2,\mathrm{K},k$$

on matriisin

$$(\mathbf{X}'\mathbf{X})^{-1}$$

(j + 1). diagonaalialkio. Jos myös standardioletus (vi) pätee, PNS-estimaattori b_j noudattaa **normaalijakaumaa**:

$$b_j \quad \mathrm{N}(\beta_j, \sigma_{b_j}^2)\ ,\ j = 0,1,2,\mathrm{K},k$$

Jos standardioletukset (i)-(v) pätevät, regressiokertoimen β_j *PNS-estimaattorin* b_j *varianssin*

$$\mathrm{D}^2(b_j) = \sigma_{b_j}^2 = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,\,j+1}\ ,\ j = 0,1,2,\mathrm{K},k$$

harhaton estimaattori on

$$\hat{\mathrm{D}}^2(b_j) = s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,\,j+1}\,,\; j = 0,1,2,\mathrm{K},k$$

jossa

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

on jäännösvarianssin σ^2 harhaton estimaattori.

Jäännösvarianssin otosjakauma

Olemme edellä todenneet, että jäännöstermien ε_i , i = 1, 2, … , n varianssin eli *jäännösvarianssin* σ^2 *harhaton estimaattori* on

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

jossa

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \mathrm{L} - b_k x_{ik}\,,\; i = 1,2,\mathrm{K},n$$

on estimoidun mallin *residuaali*. Voidaan osoittaa, että s^2 on *riippumaton* estimaattoreista b_0, b_1, … , b_k ja lisäksi

$$\frac{(n-k-1)s^2}{\sigma^2} \quad \chi^2(n-k-1)$$

Tuloksen perustelu sivuutetaan.

Yhdistämällä tämä tulos edellä johdettuihin regressiokertoimien jakaumia koskeviin tuloksiin seuraavat tärkeät jakaumatulokset:

$$t_j = \frac{b_j - \beta_j}{\hat{\mathrm{D}}^2(b_j)} = \frac{b_j - \beta_j}{s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{j+1,\,j+1}} \quad t(n-k-1)\,,\; j = 0,1,2,\mathrm{K},k$$

Regressiokertoimien **luottamusvälit** ja kertoimia koskevat **testit** voidaan konstruoida näiden jakaumatulosten perusteella samalla tavalla kuin konstruoidaan luottamusväli ja yhden otoksen *t*-testi normaalijakauman odotusarvolle tai yhden selittäjän lineaarisen regressiomallin kertoimille; ks. lukuja **Väliestimointi**, **Testit suhdeasteikollisille muuttujille** ja **Yhden selittäjän lineaarinen regressiomalli**.

Regressiokertoimien luottamusvälit

Jos standardioletukset (i)-(vi) pätevät, regressiokertoimen β_j **luottamusväli** luottamustasolla $(1 - \alpha)$ on muotoa

$$b_j \pm t_{\alpha/2}\,\hat{\mathrm{D}}(b_j)\,,\; j = 0,1,2,\mathrm{K},k$$

jossa

b_j = regressiokertoimen β_j *PNS-estimaattori*

$\pm t_{\alpha/2}$ = luottamustasoa $(1 - \alpha)$ vastaavat luottamuskertoimet *t-jakaumasta*, jonka vapausasteiden lukumäärä on $(n - k - 1)$

$\hat{\mathrm{D}}(b_j)$ = regressiokertoimen β_j *PNS-estimaattorin varianssin harhaton estimaattori*

Regressiokertoimien luottamusvälien tulkintat

Regressiokertoimen β_j luottamustasoon $(1 - \alpha)$ liittyvä *luottamusväli*

$$b_j \pm t_{\alpha/2}\,\hat{\mathrm{D}}(b_j)\,,\ j = 0,1,2,\mathrm{K},k$$

peittää regressiokertoimen β_j todennäköisyydellä $(1 - \alpha)$:

$$\Pr(b_j - t_{\alpha/2}\,\hat{\mathrm{D}}(b_j) \le \beta_j \le b_j + t_{\alpha/2}\,\hat{\mathrm{D}}(b_j)) = 1-\alpha\,,\ j = 0,1,2,\mathrm{K},k$$

Frekvenssitulkinta luottamusvälille: Jos otantaa toistetaan, otoksista konstruoiduista luottamusväleistä $100\times(1 - \alpha)$ % *peittää parametrin β_j todellisen arvon* ja $100\times\alpha$ % väleistä *ei peitä parametrin β_j todellista arvoa*.

Yleistesti regression olemassaololle

Asetetaan nollahypoteesi

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \mathrm{L} = \beta_k = 0$$

Jos nollahypoteesi H_0 *pätee*, selitettävä muuttuja *y ei riipu* lineaarisesti yhdestäkään selittäjästä x_1, x_2, … , x_k . Jos nollahypoteesi H_0 *ei päde*, selitettävä muuttuja *y riippuu* lineaarisesti ainakin yhdestä selittäjästä x_1, x_2, … , x_k .

Määritellään ***F*-testisuure**

$$F = \frac{n-k-1}{k}\cdot\frac{R^2}{1-R^2} = \frac{n-k-1}{k}\cdot\frac{SSM}{SSE}$$

jossa

R^2 = estimoidun mallin *selitysaste*

SSM = estimoidun mallin *mallineliösumma*

SSE = estimoidun mallin *jäännösneliösumma*

Testisuure F vertaa toisiinsa *residuaalivarianssia*

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

ja mallivarianssia

$$s_M^2 = \frac{1}{k}SSM$$

Oletetaan, että standardioletukset (i)-(vi) pätevät. Tällöin testisuure F noudattaa *nollahypoteesin* H_0 *pätiessä F-jakaumaa* vapausastein k ja $(n - k - 1)$:

$$F \quad F(k, n-k-1)$$

Testisuureen F *normaaliarvo* eli odotusarvo nollahypoteesin H_0 pätiessä on (suurille n)

$$\mathrm{E}(F) = \frac{n-k-1}{n-k-3} \approx 1$$

Suuret testisuureen F arvot viittaavat siihen, että nollahypoteesi H_0 *ei päde*.

Testit yksittäisille regressiokertoimille

Asetetaan nollahypoteesi

$$\mathrm{H}_{0j} : \beta_j = 0 \,,\; j = 0,1,2,\mathrm{K},k$$

Jos nollahypoteesi H_{00} *pätee*, mallissa *ei ole vakiota*. Jos nollahypoteesi H_{0j} , j = 1, 2, … , k *pätee*, selitettävä muuttuja *y ei riipu* lineaarisesti selittäjästä x_j . Jos nollahypoteesi H_{0j} , j = 1, 2, … , k *ei päde*, selitettävä muuttuja *y riippuu* lineaarisesti selittäjästä x_j .

Määritellään ***t*-testisuure**

$$t_j = \frac{b_j}{\hat{\mathrm{D}}(b_j)} \,,\; j = 0,1,2,\mathrm{K},k$$

jossa

b_j = regressiokertoimen β_j *PNS-estimaattori*

$\hat{\mathrm{D}}(b_j)$ = regressiokertoimen β_j *PNS-estimaattorin varianssin harhaton estimaattori*

Oletetaan, että standardioletukset (i)-(vi) pätevät. Tällöin testisuure t_j noudattaa *nollahypoteesin*

$$\mathrm{H}_{0j} : \beta_j = 0 \,,\; j = 0,1,2,\mathrm{K},k$$

pätiessä t-jakaumaa vapausastein $(n - k - 1)$:

$$t_j \quad t(n-k-1) \,,\; j = 0,1,2,\mathrm{K},k$$

Testisuureen t_j *normaaliarvo* eli odotusarvo nollahypoteesin H_{0j} pätiessä on

$$\mathrm{E}(t_j) = 0 \,,\; j = 0,1,2,\mathrm{K},k$$

Itseisarvoltaan *suuret* testisuureen arvot t_j viittaavat siihen, että nollahypoteesi H_{0j} *ei päde*.

Näillä *t*-testeillä on keskeinen rooli selittäjien valinnassa malliin; ks. lukua **Regressiomallin valinta**.

16.6. Ennustaminen yleisellä lineaarisella mallilla

Selitettävän muuttujan odotettavissa olevan arvon ennustaminen

Oletetaan, että selitettävä muuttuja y saa arvon

$$y_0 = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \mathrm{L} + \beta_k x_{0k} + \varepsilon_0$$

kun selittäjät x_1, x_2, … , x_k saavat arvot $x_{01}, x_{02}, \mathrm{K}, x_{0k}$. Mikä on **paras ennuste selitettävän muuttujan** y **odotettavissa olevalle arvolle**

$$\mathrm{E}(y_0 \mid x_{01}, x_{02}, \mathrm{K}, x_{0k}) = \beta_0 + \beta_1 x_{01} + \beta_2 x_{02} + \mathrm{L} + \beta_k x_{0k}$$

jos selittäjät $x_1, x_2, \ldots, x_k$ saavat arvot $x_1, x_2, \ldots, x_k$? Selitettävän muuttujan *y* ehdollinen odotusarvo $\mathrm{E}(y \mid x_1, x_2, \ldots, x_k)$ kuvaa *selitettävän muuttujan y keskimääräisiä arvoja selittäjien* $x_1, x_2, \ldots, x_k$ *saamien arvojen funktiona*.

Valitaan selitettävän muuttujan ehdollisen odotusarvon $\mathrm{E}(y \mid x_1, x_2, \ldots, x_k)$ **ennusteeksi** (*estimaattoriksi*) lauseke

$$y \mid x_1, x_2, \ldots, x_k = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

jossa $b_0, b_1, b_2, \ldots, b_k$ ovat regressiokertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ PNS-estimaattorit. Voidaan osoittaa, että $\hat{y} \mid x_1, x_2, \ldots, x_k$ on (ennustevirheen keskineliövirheen mielessä) *paras lineaarinen* ja *harhaton ennuste* ehdolliselle odotusarvolle $\mathrm{E}(y \mid x_1, x_2, \ldots, x_k)$.

Huomautus:

- Ehdollinen odotusarvo $\mathrm{E}(y \mid x_1, x_2, \ldots, x_k)$ on kiinteille $x_1, x_2, \ldots, x_k$ *vakio*, kun taas ennuste $y \mid x_1, x_2, \ldots, x_k$ on *satunnaismuuttuja*.

Selitettävän muuttujan odotettavissa olevan arvon ennusteen otosjakauma

Oletetaan, että yleistä lineaarista mallia koskevat *standardioletukset* (i)-(vi) *pätevät*. Tällöin ennusteen

$$y \mid x_1, x_2, \ldots, x_k = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

otosjakauma on normaalijakauma:

$$y \mid x_1, x_2, \ldots, x_k \sim \mathrm{N}\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k, \sigma^2 \left[\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right]\right)$$

jossa

$$\mathbf{x} = (1, x_1, x_2, \ldots, x_k)$$

on $(k + 1)$-vektori.

Selitettävän muuttujan odotettavissa olevan arvon luottamusväli

Oletetaan, että yleistä lineaarista mallia koskevat *standardioletukset* (i)-(vi) *pätevät*. Tällöin ennusteen

$$y \mid x_1, x_2, \ldots, x_k = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

luottamusväli luottamustasolla $(1 - \alpha)$ on muotoa

$$b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k \pm t_{\alpha/2}\, s \left[\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right]^{\frac{1}{2}}$$

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

Väli muodostaa selittäjien $x_1, x_2, \ldots, x_k$ arvojen $x_1, x_2, \ldots, x_k$ .funktiona *luottamusvyön* estimoidun regressiotason

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k$$

ympärille.

Selitettävän muuttujan arvon ennustaminen

Oletetaan, että selitettävä muuttuja *y* saa arvon

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

kun selittäjät $x_1, x_2, \ldots, x_k$ saavat arvot $x_1, x_2, \ldots, x_k$. Mikä on **paras ennuste selitettävän muuttujan** *y* **arvolle** y, kun selittäjät $x_1, x_2, \ldots, x_k$ saavat arvot $x_1, x_2, \ldots, x_k$?

Valitaan *selitettävän muuttujan arvon* y **ennusteeksi** (*estimaattoriksi*) lauseke

$$y \mid x_1, x_2, \ldots, x_k = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

jossa $b_0, b_1, b_2, \ldots, b_k$ ovat regressiokertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ PNS-estimaattorit. Voidaan osoittaa, että $y \mid x_1, x_2, \ldots, x_k$ on (ennustevirheen keskineliövirheen mielessä) *paras lineaarinen* ja *harhaton ennuste* ehdolliselle odotusarvolle $\mathrm{E}(y \mid x_1, x_2, \ldots, x_k)$.

Huomautus:

- Sekä selitettävän muuttujan *y* arvo y että ennuste $y \mid x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*.

Selitettävän muuttujan arvon ennusteen otosjakauma

Oletetaan, että yleistä lineaarista mallia koskevat *standardioletukset* (i)-(vi) *pätevät*. Tällöin *ennustevirheen*

$$y - y \mid x_1, x_2, \ldots, x_k$$

otosjakauma on normaalijakauma:

$$y - y \mid x_1, x_2, \ldots, x_k \sim \mathrm{N}\left(0, \sigma^2 \left[1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right]\right)$$

Selitettävän muuttujan arvon luottamusväli

Oletetaan, että yleistä lineaarista mallia koskevat *standardioletukset* (i)-(vi) *pätevät*. Tällöin selitettävän muuttujan *y* arvon y **luottamusväli** luottamustasolla $(1 - \alpha)$ on muotoa

$$b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \pm t_{\alpha/2}\, s \left[1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\right]^{\frac{1}{2}}$$

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

Väli muodostaa selittäjien $x_1, x_2, \ldots, x_k$ arvojen $x_1, x_2, \ldots, x_k$ .funktiona *luottamusvyön* estimoidun regressiotason

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

ympärille.

Selitettävän muuttujan odotettavissa olevan arvon luottamusväli *vs* selitettävän muuttujan arvon luottamusväli

Selitettävän muuttujan *y* arvon y_0 luottamusvyö *on leveämpi* kuin selitettävän muuttujan *y* arvon odotusarvon $\mathrm{E}(y_0 \mid x_1, x_2, \ldots, x_k)$ luottamusvyö. Tämä johtuu siitä, että selitettävän muuttujan *y keskimääräisen arvon ennustaminen on helpompaa kuin sen yksittäisen arvon ennustaminen*.

16.7. Yleinen lineaarinen malli ja satunnaiset selittäjät

Yleinen lineaarinen malli ja standardioletukset

Yleisessä lineaarisessa mallissa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

on seuraavat osat:

$\mathbf{y}$ = selitettävän muuttujan *y havaittujen arvojen* muodostama *satunnainen* *n*-vektori

$\mathbf{X}$ = selittäjien x_1, x_2, … , x_k *havaittujen arvojen* ja *ykkösten* muodostama $n \times (k + 1)$-matriisi

$\boldsymbol{\beta}$ = regressiokertoimien muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k + 1)$-vektori

$\boldsymbol{\varepsilon}$ = jäännöstermien muodostama *ei-havaittu* ja *satunnainen n*-vektori

Jos yleisen lineaarisen mallin selittäjät x_1, x_2, … , x_k ovat *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** voidaan esittää matriisein seuraavassa muodossa:

(i) Matriisin $\mathbf{X}$ alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*

(ii) Matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii) $$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

(iv)&(v) $$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(vi) $$\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Selittäjien satunnaisuus

Yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

matriisin $\mathbf{X}$ *satunnaisuus saattaa aiheuttaa vakavia ongelmia* mallin estimoinnille ja mallia koskevalle tilastolliselle päättelylle. Jos matriisi $\mathbf{X}$ on satunnainen, PNS-menetelmä *ei välttämättä tuota harhattomia* tai *edes tarkentuvia estimaattoreita regressiokertoimille*. Näin käy esimerkiksi silloin, kun virhetermi ja selittäjät *korreloivat*. Jos regressiokertoimien PNS-estimaattorit *eivät ole harhattomia* tai *tarkentuvia*, mallia koskevaa *tavanomaista tilastollista päättelyä ei saa soveltaa*.

Kysymys:

Milloin edellä kiinteille selittäjille esitettyä teoriaa voidaan soveltaa satunnaisille selittäjille?

Vastaus:

> Kiinteille selittäjille esitettyä teoriaa voidaan soveltaa esimerkiksi silloin, kun *jäännös-* eli *virhetermit* ε_j toteuttavat kiinteille selittäjille esitetyt standardioletukset *ehdollisesti selittäjien* $x_1, x_2, \ldots, x_k$ *havaittujen arvojen suhteen*.

Tarkastelemme tässä kappaleessa lineaarisen regressiomallin määrittelemistä sellaisella tavalla, joka takaa sen, että kiinteille selittäjille esitetty teoria pätee.

Regressiokertoimien vektorin PNS-estimaattorin harhattomuus

Olkoon

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori*.

PNS-estimaattorin **b** lauseke voidaan kirjoittaa muotoon

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

Jos matriisi **X** on *kiinteä*, estimaattori **b** on *harhaton*, koska standardioletuksen (iii) mukaan

$$\mathrm{E}(\boldsymbol{\varepsilon}) = 0$$

jolloin

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}$$

Jos matriisi **X** on *satunnainen*, **ei saa kirjoittaa**

$$\mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon})$$

Sen sijaan PNS-estimaattorin **b** *ehdollisessa odotusarvossa* matriisin **X** suhteen matriisia **X** voidaan pitää ”*kiinteänä*” ja siten

$$\mathrm{E}(\mathbf{b} \mid \mathbf{X}) = \boldsymbol{\beta} + \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X})$$

PNS-estimaattori **b** on siis **ehdollisesti harhaton** eli

$$\mathrm{E}(\mathbf{b} \mid \mathbf{X}) = \boldsymbol{\beta}$$

jos oletus

$$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

pätee. Tällöin PNS-estimaattori **b** on myös (*ehdottomasti*) **harhaton**, koska *iteroidun odotusarvon lain* mukaan

$$\mathrm{E}(\mathbf{b}) = \mathrm{E}(\mathrm{E}(\mathbf{b} \mid \mathbf{X})) = \mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}$$

Edellä esitetystä nähdään, että *ehdon*

$$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

voimassaolo ratkaisee sen, onko PNS-estimaattori **b** *harhaton* lineaarisen mallin egressiokertoimien vektorille $\boldsymbol{\beta}$.

Voidaan osoittaa, että vastaava korjaus muihin yleisen lineaarisen mallin standardioletuksiin (iii)-(vi) ”pelastaa” kiinteiden selittäjien tapauksessa esitetyn teorian.

Huomautus:

- Tilastotieteessä kohdataan tilanteita, joissa edes seuraavassa kohdassa esitettävät modifioidut ehdot *eivät päde*.

Yleinen lineaarinen malli ja modifioidut standardioletukset satunnaisten selittäjien tapaukselle

Jos yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*, mallia koskevat **standardioletukset** voidaan esittää matriisein seuraavassa muodossa:

(i)´ Matriisin **X** alkiot ovat (vakioselittäjän arvoja lukuun ottamatta) *satunnaismuuttujia*

(ii)´ Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

(iv)´&(v)´ $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$

(vi)´ $(\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Modifioidusta standardioletuksesta (iii)´ seuraa, että

$$\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

Tämä merkitsee sitä, että selitettävän muuttujan y ehdollinen odotusarvo eli regressiofunktio selittäjien havaittujen arvojen suhteen on lineaarinen. Modifioidusta standardioletuksesta (iii)´ seuraa edelleen, että

$$\mathrm{E}(\mathbf{z}_i\boldsymbol{\varepsilon}_i) = \mathbf{0}$$

jossa

$$\mathbf{z}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik})$$

Siten oletuksesta (iii)´ seuraa, että selittäjien arvot ja jäännös- eli virhetermit ovat *korreloimattomia*.

Kommentteja

Jos modifioidut standardioletukset (i)´-(vi)´ pätevät, edellä esitetty tavanomainen estimointi- ja päättelytekniikka ei-satunnaisille selittäjille pätee.

On hyvä kuitenkin tietää, että myös edellä esitetyt modifioidut ehdot jäännös- eli virhetermeille ovat melko rajoittavia ja etenkin *aikasarjojen regressiomallien soveltamisen yhteydessä kohdataan tilanteita, joissa eivät edes nämä modifioidut ehdot päde*. Sellaisessa tilanteissa **PNS-menetelmää ei saa käyttää mallin parametrien estimointiin**. Tilastotiede tuntee kuitenkin menetelmiä, joilla regressiomallin parametrit *voidaan estimoida* (*ainakin*) *tarkentuvasti* myös monissa niistä tilanteista, joissa edellä esitetyt modifioidut ehdot jäännöstermeille *eivät päde*.

17. Regressiomallin valinta

17.1. Regressiomallin valinta: Johdanto

17.2. Mallinvalintatestit

17.3. Mallinvalintakriteerit

17.4. Epälineaaristen riippuvuuksien linearisointi

Regressioanalyysin perusongelma on seuraava: Kuvaako selitettävän muuttujan ja mahdollisten selittäjäkandidaattien riippuvuudelle *spesifioitu* eli *täsmennetty* regressiomalli riippuvuutta *oikein*? Tämän yleisen ongelman ehkä kaikkein tärkein osaongelma on se, *onko malliin osattu valita oikeat muuttujat oikeassa funktionaalisessa muodossa*.

Tarkastelemme tässä luvussa **selittävien muuttujien valintaa** sekä kahta *tilastollista* ratkaisua selittäjien valintaongelmaan:

- **Mallinvalintatestit**
- **Mallinvalintakriteerit**

Lisäksi tarkastelemme lyhyesti yhden selittäjän regressiomallin selitettävän muuttujan ja selittävän muuttujan **funktionaalisten muotojen valitsemista** sellaisella tavalla, että alkuperäisten muuttujien välinen *epälineaarinen riippuvuus* saadaan **linearisoiduksi**.

Avainsanat:

Akaikein informaatiokriteeri, Askeltava regressio, Askellus alaspäin, Epälineaarisuus, Estimaattori, Estimointi, *F*-testi, Harha, Harhattomuus, Havainto, Jäännösneliösumma, Jäännöstermi, Jäännösvarianssi, Jäännösvarianssikriteeri, Keskineliövirhe, Kokonaisneliösumma, Kokonaisvaihtelu, Korjattu selitysaste, Korrelaatio, Kovarianssi, Lineaarinen regressiomalli, Linearisointi, Lineaarisuus, Malli, Mallin hyvyys, Mallin valinta, Mallineliösumma, Mallowsin kriteeri, Merkitsevyystaso, Modifioidut standardioletukset, Muunnos, Neliösumma, Odotusarvo, Otos, Otosjakauma, Otostunnusluku, Parametri, Parsimonisuus, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Rakenneosa, Regressioanalyysi, Regressiokerroin, Regressiomalli, Residuaali, Sakkofunktio, Satunnainen osa, Schwarzin kriteeri, Selittäjä, Selitettävä muuttuja, Selittävä muuttuja, Selitysaste, Sovite, Spesifikaatio, Spesifiointi, Spesifiointivirhe, Standardioletus, Systemaattinen osa, *t*-testi, Testi, Täsmentäminen, Usean selittäjän lineaarinen regressiomalli, Vakioselittäjä, Varianssianalyysihajotelma, Virhetermi, Usean selittäjän lineaarinen regressiomalli, Yleinen lineaarinen malli

17.1. Regressiomallin valinta: Johdanto

Regressiomallin selittäjiksi on usein tarjolla joukko **selittäjäkandidaatteja** tai **-ehdokkaita** ja tilastollisen analyysin tehtävänä on löytää kandidaattien joukosta *oikeat* tai *parhaat*. Selittäjien valintaa regressiomallin kutsutaan tavallisesti **mallin valinnaksi**, vaikka oikeastaan kaikkea mikä liittyy *mallin rakenneosan* ja *jäännöstermin spesifioimiseen* voidaan pitää mallin valintana.

Huomautus:

- Ihannetapauksessa regressiomalli voidaan spesifioida *asialoogisin, tutkittavan ilmiön taustateoriaan nojaavien syiden perusteella*.

17.2. Yleinen lineaarinen malli

Oletetaan, että muuttujien y ja $x_1, x_2, \ldots, x_k$ havaittujen arvojen välillä vallitsee *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

jossa

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

x_{ij} = **selittävän muuttujan** eli **selittäjän** x_j *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä i, $j = 1, 2, \ldots, k$

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

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

β_j = **selittäjän** x_j **regressiokerroin**, $j = 1, 2, \ldots, k$; β_j on *ei-satunnainen* ja *tuntematon vakio*

Tällöin yhtälö määrittelee **usean selittäjän lineaarisen regressiomallin**, jota kutsutaan **yleiseksi lineaariseksi malliksi**.

Seuraavassa kertaamme yleisen lineaarisen mallin *formuloinnin matriisein*, mallia koskevat *standardioletukset* ja pääkohdat mallin parametrien *estimoinnista*; lisätietoja: ks. lukua **Yleinen lineaarinen malli**.

Yleinen lineaarinen malli voidaan esittää matriisein muodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa

$\mathbf{y}$ = **selitettävän muuttujan** y *havaittujen arvojen* muodostama *satunnainen* n-vektori

$\mathbf{X}$ = **selittäjien** $x_1, x_2, \ldots, x_k$ *havaittujen arvojen* ja *ykkösten* muodostama $n \times (k + 1)$-matriisi

$\boldsymbol{\beta}$ = **regressiokertoimien** muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k + 1)$-vektori

$\boldsymbol{\varepsilon}$ = **jäännöstermien** muodostama *ei-havaittu* ja *satunnainen n*-vektori

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** esitetään matriisein seuraavassa muodossa:

(i) Matriisin **X** alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*

(ii) Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k+1$$

(iii) $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$

(iv)&(v) $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

(vi) $\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*, mallia koskevat **modifioidut standardioletukset** esitetään matriisein seuraavassa muodossa:

(i)´ Matriisin **X** alkiot ovat *satunnaismuuttujia*.

(ii)´ Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k+1$$

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

(iv)´&(v)´ $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$

(vi)´ $\boldsymbol{\varepsilon} \mid \mathbf{X} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Mallin rakenneosa ja jäännösosa

Oletetaan, että yleistä lineaarista mallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

koskevat standardioletukset pätevät. Tällöin selitettävä muuttujan arvojen vektori **y** voidaan esittää seuraavalla tavalla kahden osatekijän summana:

$$\mathbf{y} = \mathrm{E}(\mathbf{y} \mid \mathbf{X}) + \boldsymbol{\varepsilon}$$

Osatekijä

$$\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

muodostaa mallin **systemaattisen** eli **rakenneosan**, joka riippuu selittäjien $x_1, x_2, \ldots, x_k$ havaituista arvoista. Jäännöstermi $\boldsymbol{\varepsilon}$ muodostaa mallin **satunnaisen osan**, joka ei riipu selittäjien $x_1, x_2, \ldots, x_k$ havaituista arvoista.

Regressiokertoimien PNS-estimaattorit ja niiden ominaisuudet

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\ i = 1, 2, \mathrm{K}, n$$

regressiokertoimien

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

PNS- eli **pienimmän neliösumman estimaattorit**

$$b_0, b_1, b_2, \ldots, b_k$$

minimoivat jäännös- eli *virhetermien* ε_i *neliösumman*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots - \beta_k x_{ik})^2$$

kertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ *suhteen.*

Yleisen lineaarisen mallin $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ regressiokertoimien vektorin

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

PNS-estimaaattori voidaan esittää *matriisein* muodossa

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

PNS-estimaattorilla **b** on standardioletuksien (i)-(vi) pätiessä seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$$

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{b} \quad \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Estimoidun mallin sovitteet ja residuaalit sekä niiden ominaisuudet

Olkoon

$$\mathbf{b} = (b_0, b_1, b_2, \ldots, b_k)$$

yleisen linearegressiokertoimien vektorin

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

PNS-estimaattori.

Määritellään estimoidun mallin **sovitteet** $\hat{y}_i$ kaavalla

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ldots + b_k x_{ik}, \; i = 1, 2, \ldots, n$$

Määritellään estimoidun mallin **residuaalit** e_i kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik}, \; i = 1, 2, \ldots, n$$

Sovitteiden muodostama *n*-vektori voidaan esittää *matriisein* muodossa

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}$$

Residuaalien muodostama *n*-vektori voidaan esittää *matriisein* muodossa

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y} = \mathbf{M}\mathbf{y}$$

Huomautus:

- Koska *residuaalit kuvaavat estimoidun regressiomallin ja havaintoarvojen yhteensopivuutta*, monet regressiodiagnostiikan menetelmistä perustuvat estimoidun regressiomallin residuaaleihin tai niiden muunnoksiin.

Sovitteiden muodostamalla *n*-vektorilla $\hat{\mathbf{y}}$ on seuraavat *stokastiset ominaisuudet* :

$$\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{P} = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Residuaalien muodostamalla *n*-vektorilla **e** on seuraavat *stokastiset ominaisuudet* :

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}$$
$$\mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I}-\mathbf{P}) = \sigma^2(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$

Huomautus:

- Yllä olevan mukaan residuaalit e_j ovat yleensä sekä *heteroskedastisia* että *korreloituneita*, vaikka jäännöstermit ε_j on oletettu *homoskedastisiksi* ja *korreloimattomiksi*.

Matriisit

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathbf{M} = \mathbf{I}-\mathbf{P} = \mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

ovat *symmetrisiä* ja *idempotentteja* eli *projektioita*:

$$\mathbf{P}' = \mathbf{P} \qquad \mathbf{P}^2 = \mathbf{P}$$
$$\mathbf{M}' = \mathbf{M} \qquad \mathbf{M}^2 = \mathbf{M}$$

Lisäksi

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}$$

Matriisia **P** kutsutaan regressiodiagnostiikassa usein *hattumatriisiksi*.

Jäännösvarianssin estimointi

Yleisen lineaarisen mallin jäännöstermien ε_i varianssin eli **jäännösvarianssin σ^2 harhaton estimaattori** on

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

jossa

e_i = estimoidun mallin *residuaali*, $i = 1, 2, \ldots, n$

n = havaintojen lukumäärä

k = (aitojen) selittäjien x_j lukumäärä

Yleisen lineaarisen mallin rakenneosa ja sen spesifiointi

Yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

muotoilua ja siitä tehtävien oletusten valintaa kutsutaan mallin **spesifioinniksi** eli **täsmentämiseksi**.

Oikean spesifikaation löytäminen mallin **systemaattiselle osalle** eli **rakenneosalle**

$$\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

on regressioanalyysin päätehtävä, koska juuri mallin rakenneosa kuvaa selitettävän muuttujan y riippuvuutta selittäjistä $x_1, x_2, \ldots, x_k$.

Lineaaristen regressiomallien estimointia, testausta ja ennustamista koskevat tulokset edellyttävät, että **mallin rakenneosa on oikein spesifioitu**. Virheet regressiomallin rakenneosan spesifioinnissa saattavat johtaa karkeisiin virheellisiin johtopäätöksiin selitettävän muuttujan ja selittäjien välisestä riippuvuudesta.

Miksi oikeiden selittäjien löytäminen regressiomalliin on *tärkeätä*?

Kun regressiomallin rakenneosalle etsitään *oikeata spesifikaatiota*, keskeisenä ongelmana on löytää malliin *oikeat selittäjät*:

(i) *Jos regressiomallista puuttuu siihen kuuluvia selittäjiä,* mallin regressiokertoimien PNS-estimaattorit ovat (yleensä) **harhaisia**.

(ii) *Jos regressiomallissa on turhia selittäjiä,* mallin regressiokertoimien PNS-estimaattorit ovat (yleensä) **tehottomia**, mikä merkitsee sitä, että kertoimien varianssit ovat tarpeettoman suuria.

Huomautus:

- Estimaattorin harhaisuus on paljon vakavampi ongelma kuin estimaattorin tehottomuus.

Miksi oikeiden selittäjien löytäminen regressiomalliin on *vaikeata*?

Oikeiden selittäjien löytäminen regressiomalliin on *vaikeata*:

(i) Hyvän regressiomallin jäännösneliösumma on *pieni*, **mutta minkä tahansa selittäjän lisääminen malliin pienentää** (tai ei ainakaan kasvata) **jäännösneliösummaa** tai yhtäpitävästi hyvän regressiomallin selitysaste on *korkea*, **mutta minkä tahansa selittäjän lisääminen malliin kasvattaa** (tai ei ainakaan pienennä) **selitysastetta**.

(ii) Hyvän regressiomallin kaikki selittäjät ovat *tilastollisesti merkitseviä,* mutta minkä tahansa selittäjäkanditaatin poistaminen mallista tai lisääminen malliin saattaa muuttaa malliin jäävien tai siellä jo olevien selittäjien tilastollista merkitsevyyttä.

Puuttuvien selittäjien ongelma

Olkoon *oikea malli* selittävälle muuttujalle **y** muotoa

(1) $$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Oletetaan, että estimoimme regressiokertoimien vektorin $\boldsymbol{\beta}_1$ *väärästä* mallista

(2) $$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\delta}$$

josta siis *puuttuu osa oikean mallin* (1) *selittäjistä*. Koska *väärästä* mallista (2) puuttuu osa *oikean* mallin (1) selittäjistä, väärän mallin (2) jäännöstermi on muotoa

$$\boldsymbol{\delta} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}$$

Olkoon

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y}$$

kerroinvektorin $\boldsymbol{\beta}_1$ PNS-estimaattori *väärästä* mallista (2).

Estimaattori $\mathbf{b}_1$ on (yleensä) *harhainen*.

Perustelu:

Estimaattorin $\mathbf{b}_1$ lauseke voidaan esittää muodossa

$$\begin{aligned}\mathbf{b}_1 &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{y} \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\varepsilon}\end{aligned}$$

Estimaattori $\mathbf{b}_1$ on *harhainen*, koska

$$\mathrm{E}(\mathbf{b}_1) = \boldsymbol{\beta}_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_1$$

ellei ehto

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\boldsymbol{\beta}_2 = \mathbf{0}$$

päde. Tämä ehto *voi käytännössä toteutua* vain kahdella tavalla:

$$\boldsymbol{\beta}_2 = \mathbf{0}$$

tai

$$\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$$

Jos

$$\boldsymbol{\beta}_2 = \mathbf{0}$$

selitettävän muuttujan y havaitut arvot y_i eivät riipu lineaarisesti matriisiin $\mathbf{X}_2$ liittyvistä selittäjistä ja regressiokertoimien vektori $\boldsymbol{\beta}_1$ voidaan estimoida *harhattomasti* mallista (2).

Jos

$$\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$$

matriisin $\mathbf{X}_1$ sarakkeet ovat *kohtisuorassa* matriisin $\mathbf{X}_2$ sarakkeita vastaan ja regressiokertoimien vektori $\boldsymbol{\beta}_1$ voidaan estimoida *harhattomasti* mallista (2).

■

Huomautus:

- Edellä esitetyn nojalla *ortogonaalisten selittäjien tapauksessa* vektorin $\boldsymbol{\beta}$ komponentit voidaan estimoida harhattomasti yhden selittäjän regressiomalleista.

Selittäjien valinnan menetelmät

Regressiomallin *selittäjien valintaan* on tarjolla kaksi erilaista menetelmää:

(i) **Mallinvalintatestejä** käytettäessä malliin pyritään valitsemaan jotakin testausstrategiaa käyttäen kaikki *tilastollisesti merkitsevät selittäjät*.

(ii) **Mallinvalintakriteereitä** käytettäessä malliin valitaan selittäjiksi kaikkien tarjolla olevien selittäjien joukosta sellainen osajoukko, joka *optimoi käytetyn kriteerifunktion arvon*.

17.3. Mallinvalintatestit

Hyvässä regressiomallissa kaikki regressiokertoimet ovat *tilastollisesti merkitseviä*. Regressiokertoimen β_j **merkitsevyyttä testataan tilastollisesti** testaamalla nollahypoteesia

$$H_0 : \beta_j = 0$$

Jos nollahypoteesi H_0 *jää testissä voimaan*, selitettävä muuttuja *y ei riipu* mallin mukaan lineaarisesti kerrointa β_j vastaavasta selittäjästä x_j . Sen sijaan, *jos nollahypoteesi* H_0 *hylätään testissä,* selitettävä muuttuja *y riippuu* mallin mukaan lineaarisesti kerrointa β_j vastaavasta selittäjästä x_j, jolloin sanotaan, että regressiokerroin β_j ja myös sitä vastaava selittäjä x_j ovat **tilastollisesti merkitseviä**.

Selittäjän merkitsevyyttä testaavia tilastollisia testejä kutsutaan mallinvalinnassa **mallinvalintatesteiksi**. Tavallisesti testeinä käytetään tavanomaisella ***t*-testillä**: ks. luvun **Yleinen lineaarinen malli** kappaletta **Tilastollinen päättely yleisestä lineaarisesta mallista**.

Kun mallinvalinnassa käytetään mallinvalintatestejä, tavoitteena on ottaa malliin mukaan kaikki tilastollisesti merkitsevät selittäjät ja sulkea mallin ulkopuolelle kaikki tilastollisesti ei-merkitsevät selittäjät.

Mallinvalintatestejä käytettäessä muodostetaan tavallisesti ensin **lähtömalli**, johon *tilastollisesti merkitsevät selittäjät pyritään lisäämään* ja *josta ei-merkitsevät pyritään poistamaan*. Tilastollisesti merkitsevien selittäjien lisääminen malliin ja ei-merkitsevien selittäjien poistaminen mallista mallinvalintatestien perusteella ei kuitenkaan ole ongelmatonta, koska **selittäjän tilastolliseen merkitsevyyteen vaikuttaa yleensä se**, **mitä muita selittäjiä mallissa on testaushetkellä**. Siten **testien suoritusjärjestys saattaa vaikuttaa siihen**, **mikä malli tulee valituksi**.

Kun mallista **poistetaan** *tilastollisesti ei-merkitseviä selittäjiä* kohdataan seuraavat ongelmat:

(i) Ei-merkitseviä selittäjiä poistettaessa poistamisjärjestys saattaa vaikuttaa lopputulokseen.

(ii) Selittäjän poistaminen mallista saattaa muuttaa aikaisemmin ei-merkitsevänä poistetun selittäjäkandidaatin merkitseväksi, jos se otettaisiin takaisin malliin.

Kun malliin **lisätään** *tilastollisesti merkitseviä selittäjiä* kohdataan seuraavat ongelmat:

(i)´ Merkitseviä selittäjiä lisättäessä lisäämisjärjestys saattaa vaikuttaa lopputulokseen.

(ii)´ Selittäjän lisääminen malliin saattaa muuttaa mallissa olevan, ennen uuden selittäjän lisäämistä merkitsevän selittäjän ei-merkitseväksi.

Mallinvalintatestien soveltamisen ongelmat ovat johtaneet erilaisten *askellusstrategioiden* kehittämiseen. Esittelemme tässä 2 strategiaa:

- **Askellus alaspäin**
- **Askeltava regressio**

Huomautus:

- *Eri strategiat* saattavat johtaa *eri malleihin*!

Alapäin askellus

Alaspäin askelluksessa käytettävä mallinvalintastrategia:

(1) Otetaan *lähtömalliin* mukaan *kaikki* selittäjäkandidaatit.

(2) Valitaan mallinvalintatesteissä käytettävä *merkitsevyystaso Out.*

(3) *Estimoidaan* malli niillä selittäjillä, jotka ovat mallissa.

(4) Testataan merkitsevyystasoa *Out* käyttäen kaikkien mallissa olevien selittäjien *tilastollista merkitsevyyttä*.

(5) Jos kaikki mallissa olevat selittäjät ovat tilastollisesti merkitseviä, **malli on valmis**.

(6) *Poistetaan* mallin ei-merkitsevistä selittäjistä se, jota vastaava *p*-arvo on *suurin*.

(7) Palataan vaiheeseen (3).

Askel muodostuu vaiheista (3)-(7).

Huomautus:

- Vaihe (3) eli mallin estimointi uudelleen *on välttämätön* joka askeleessa. Tämä johtuu siitä, että – lukuun ottamatta ortogonaalisten selittäjien tapausta – estimointitulokset muuttuvat yleensä joka askeleessa.

Askeltava regressio

Askeltavassa regressiossa käytettävä mallinvalintastrategia:

(1) Muodostetaan *lähtömalli*.

(2) Valitaan *kaksi* mallinvalintatesteissä käytettävää *merkitsevyystasoa In* ja *Out*.

(3) *Estimoidaan* malli niillä selittäjillä, jotka ovat mallissa.

(4) Testataan vuorotellen merkitsevyystasoa *In* käyttäen kaikkien ko. askeleessa mallin ulkopuolella olevien selittäjäkandidaattien *tilastollista merkitsevyyttä malliin lisättyinä*.

(5) Testataan merkitsevyystasoa *Out* käyttäen kaikkien mallissa olevien selittäjien *tilastollista merkitsevyyttä*.

(6) Jos malliin liitettynä tilastollisesti merkitseviä selittäjäkandidaatteja löytyy, *lisätään* malliin kandidaateista se, jota vastaava *p*-arvo on *pienin*.

(7) Jos mallissa on tilastollisesti ei-merkityksellisiä selittäjiä, *poistetaan* niistä se, jota vastaava *p*-arvo on *suurin*.

(8) Jos malliin ei voida liittää uusia selittäjiä eikä siitä poistaa yhtään siinä olevaa selittäjää, **malli on valmis**.

(9) Palataan vaiheeseen (3).

Askel muodostuu vaiheista (3)-(9).

Huomautus:

- Vaihe (3) eli mallin estimointi uudelleen *on välttämätön* joka askeleessa. Tämä johtuu siitä, että – lukuun ottamatta ortogonaalisten selittäjien tapausta – estimointitulokset muuttuvat yleensä joka askeleessa.

17.4. Mallinvalintakriteerit

Hyvän regressiomallin **jäännösneliösumma** *SSE* on *pieni* tai – mikä on sama asia – **selitysaste** R^2 on *korkea*. Saattaisi olla houkutteleva ajatus valita tarjolla olevista selittäjäkandidaateista malliin ne, joiden muodostama joukko *minimoi jäännösneliösumman* (tai *maksimoi selitysasteen*).

Jäännösneliösumman minimointia (tai selitysasteen maksimointia) *ei* kuitenkaan *voida käyttää mallin valintaan*: Jäännösneliösumma *SSE pienenee* tai ei ainakaan kasva (selitysaste R^2 *kasvaa* tai ei ainakaan pienene) *aina, kun malliin lisätään selittäjä*. Siten jäännösneliösumman minimointi (tai selitysasteen maksimointi) johtaa *aina kaikkien tarjolla olevien selittäjien valintaan*.

Mallinvalintakriteereissä jäännösneliösummaan liitetään **sakkofunktio**, jonka arvo riippuu estimoitavien regressio-kertoimien lukumäärästä. Sakkofunktio kasvattaa kriteerifunktio arvoa, *elleivät malliin lisätyt selittäjät pienennä jäännösneliösummaa tarpeeksi paljon*. Mallinvalinta-kriteereitä voidaan pitää tieteellisen päättelyn keskeisen periaatteen – **principle of parsimony**: yksinkertainen selitys tosiasioille on parempi kuin monimutkainen – kiteytyksinä tilastollisten mallien maailmaan.

Mallivalintakriteerien yleinen muoto

Olkoon

$$\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$$

lineaarinen regressiomalli, jossa selittäjien lukumäärä on (vakioselittäjä mukaan luettuna) $p = k + 1$ ja olkoon

$$\mathbf{b}_p = (\mathbf{X}_p'\mathbf{X}_p)^{-1}\mathbf{X}_p'\mathbf{y}$$

regressiokertoimien vektorin $\boldsymbol{\beta}_p$ PNS-estimaattori sekä

$$SSE_p = (\mathbf{y} - \mathbf{X}_p\mathbf{b}_p)'(\mathbf{y} - \mathbf{X}_p\mathbf{b}_p)$$

olkoon vastaava *jäännösneliösumma*.

Useimmat **mallinvalintakriteerit** voidaan esittää muodossa

$$C(p,n) = \hat{\sigma}_p^2 + p \cdot f(n)$$

jossa

$$\hat{\sigma}_p^2 = \frac{SSE_p}{n}$$

on *jäännösvarianssin* σ^2 *suurimman uskottavuuden* (SU-) *estimaattori* mallista, jossa on p selittäjää ja $f(n)$ on *positiivinen* havaintojen ja havaintojen lukumäärän *funktio*.

Kriteerifunktiolla $C(p,n)$ on seuraavat ominaisuudet:

(i) Jäännösvarianssin σ^2 SU-estimaattorin $\hat{\sigma}_p^2$ arvo pienenee (tai ei ainakaan kasva), jos malliin lisätään selittäjä.

(ii) Termin $p \cdot f(n)$ arvo kasvaa, jos malliin lisätään selittäjä.

Kriteerifunktion $C(p,n)$ arvo *pienenee* siis vain, *jos estimaattori* $\hat{\sigma}_p^2$ *pienenee tarpeeksi paljon, kun malliin lisätään selittäjä*. Kriteerifunktion termiä $p \cdot f(n)$ kutsutaan **sakkofunktioksi**.

Mallinvalintakriteereiden soveltaminen

Oletetaan, että tarjolla olevia selittäjäkandidaattien lukumäärä on q.

(i) Määrätään kriteerifunktion arvo *kaikille mahdollisille selittäjäkandidaattien yhdistelmille* eli kaikille malleille, joissa on p selittäjää, kun $p = 1, 2, \ldots, q$.

(ii) Valitaan malliin selittäjiksi se selittäjäkandidaattien yhdistelmä, joka *optimoi kriteerifunktion arvon*.

Mallinvalintakriteereitä

Tilastotieteen kirjallisuus tuntee *useita* erilaisia mallinvalintakriteereitä. Esittelemme tässä 5 kriteeriä:

- **Jäännösvarianssikriteeri**
- **Korjattu selityaste**
- **Mallowsin C_p**
- **Akaiken informaatiokriteeri *AIC***
- **Schwarzin Bayeslainen informaatiokriteeri *SBIC***

Teoreettisesti vahvimmat perustelut on esitetty C_p-, *AIC*- ja *SBIC*-kriteereille.

Huomautus:

- *Eri kriteerit* saattavat johtaa *eri malleihin*!

Voidaan osoittaa, että kaikilla tässä esiteltävillä kriteereillä on seuraava *hyvyysominaisuus*: Kriteerit tuottavat asymptoottisesti (so. jos havaintojen lukumäärän annetaan kasvaa rajatta) mallin, joka on **harhaton** siinä mielessä, että *mallista ei jää pois malliin kuuluvia selittäjiä*. Kuitenkin *vain SBIC-kriteeri* tuottaa asymptoottisesti mallin, joka on **tehokas** siinä mielessä, että *mallissa ei ole turhia selittäjiä*.

Jäännösvarianssikriteeri

Jäännösneliösummaa SSE_p ei sellaisenaan voida käyttää mallin valinnassa, koska se pienenee (tai ei ainakaan kasva) aina, kun malliin lisätään selittäjiä.

Määritellään **jäännösvarianssikriteeri** s_p^2 kaavalla

$$s_p^2 = \frac{SSE_p}{n-p} = \hat{\sigma}_p^2 + p \cdot \frac{\hat{\sigma}_p^2}{n-p}$$

jossa

$$SSE_p = n\hat{\sigma}_p^2 = (\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)'(\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)$$

on jäännösneliösumma mallista, jossa on $p \leq q$ selittäjää. Jäännösvarianssikriteerin mukaan *paras* vertailtavista malleista on se, joka *minimoi* kriteerifunktion s_p^2 arvon.

Huomautus:

- Jäännösvarianssikriteerin s_p^2 arvo *saattaa kasvaa*, elleivät malliin lisätyt selittäjät pienennä estimoidun mallin jäännösneliösummaa SSE_p *tarpeeksi paljon*.

Korjattu selitysaste

Selitysastetta R^2 ei sellaisenaan voi käyttää mallin valinnassa, koska se kasvaa (tai ei ainakaan pienene) aina, kun malliin lisätään selittäjiä.

Määritellään **korjattu selitysaste** $\bar{R}_p^2$ kaavalla

$$\bar{R}_p^2 = 1 - \frac{n-1}{n-p} \cdot \frac{SSE_p}{SST}$$

jossa

$$SSE_p = n\hat{\sigma}_p^2 = (\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)'(\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)$$

on *jäännösneliösumma* mallista, jossa on $p \leq q$ selittäjää ja

$$SST = (n-1)s_y^2$$

on muuttujan *y* vaihtelua kuvaava *kokonaisneliösumma*. Korjatun selitysasteen mukaan *paras* vertailtavista malleista on se, joka *maksimoi* kriteerifunktion $\bar{R}_p^2$ arvon.

Huomautuksia:

- Korjatun selitysasteen $\bar{R}_p^2$ arvo *saattaa pienentyä*, elleivät malliin lisätyt selittäjät kasvata estimoidun mallin selitysastetta *tarpeeksi paljon*.
- Korjattu selitysaste $\bar{R}_p^2$ ja jäännösvarianssikriteeri ovat *ekvivalentteja*, koska ne johtavat *samaan malliin*.

Mallowsin C_p

Määritellään **Mallowsin C_p-kriteeri** kaavalla

$$C_p = \frac{SSE_p}{s_q^2} + 2p - n$$

jossa

$$SSE_p = n\hat{\sigma}_p^2 = (\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)'(\mathbf{y} - \mathbf{X}_p\boldsymbol{\beta}_p)$$

on *jäännösneliösumma* mallista, jossa on $p \leq q$ selittäjää ja

$$(n-q)s_q^2 = SSE_q$$

missä *q* on kaikkien selittäjäkandidaattien lukumäärä. Mallowsin kriteerin mukaan *paras* vertailtavista malleista on se, joka *minimoi* kriteerifunktion C_p arvon.

Mallowsin C_p-kriteeristä tunnetaan useita *ekvivalentteja muotoja*. Määritellään kriteerifunktiot C_p' ja C_p'' kaavoilla

$$C_p' = SSE_p + (2p - n)s_q^2$$

ja

$$C_p'' = \hat{\sigma}_p^2 + 2p \cdot \frac{s_q^2}{n}$$

Kriteerifunktioiden C_p, C_p' ja C_p'' minimointi johtaa täsmälleen *samaan malliin*.

Olkoon $\mathbf{b}_q^p$ vektorin $\boldsymbol{\beta}_q$ estimaattori, joka perustuu $p \le q$ selittäjäkandidaattiin, millä tarkoitetaan sitä, että ne kertoimet, joita vastaavat selittäjät on jätetty pois mallista, merkitään vektorissa $\mathbf{b}_q^p$ nolliksi. Mallowsin C_p-kriteeri on vektorin $\boldsymbol{\beta}_p$ estimaattorin $\mathbf{b}_q^p$ *prediktiivisen keskineliövirheen*

$$\mathrm{PMSE}(\mathbf{b}_q^p) = \mathrm{E}\left[(\mathbf{b}_q^p - \boldsymbol{\beta}_q)'\mathbf{X}_q'\mathbf{X}_q(\mathbf{b}_q^p - \boldsymbol{\beta}_q)\right]$$

approksimatiivisesti *harhaton estimaattori* eli

$$\mathrm{E}(C_p') \approx \mathrm{PMSE}(\mathbf{b}_q^p)$$

jos mallin $\mathbf{y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}$ *harha on pieni*.

Akaiken infornaatiokriteeri

Määritellään Akaiken informaatiokriteeri *AIC* kaavalla

$$AIC = \hat{\sigma}_p^2 + 2p \cdot \frac{\hat{\sigma}_p^2}{n}$$

jossa

$$\hat{\sigma}_p^2 = \frac{SSE_p}{n}$$

on jäännösvarianssin σ^2 SU-estimaattori mallista, jossa on $p \le q$ selittäjää. Aikaiken informaatio-kriteerin mukaan *paras* vertailtavista malleista on se, joka *minimoi* kriteerifunktion *AIC* arvon.

Akaiken informaatiokriteeri on approksimatiivisesti *harhaton estimaattori* mallin *Kullbackin* ja *Leiblerin informaatiolle*.

Schwarzin Bayeslainen informaatiokriteeri

Määritellään Schwarzin kriteeri *SBIC* kaavalla

$$SBIC = \hat{\sigma}_p^2 + 2p \cdot \frac{\log(n)\hat{\sigma}_p^2}{n}$$

Jossa

$$\hat{\sigma}_p^2 = \frac{SSE_p}{n}$$

on jäännösvarianssin σ^2 SU-estimaattori mallista, jossa on $p \le q$ selittäjää. Schwarzin kriteerin mukaan *paras* vertailtavista malleista on se, joka *minimoi* kriteerifunktion *SBIC* arvon.

Schwarzin kriteeri *maksimoi* approksimatiivisesti mallin *posteriori-todennäköisyyden* sopivasti valitulle priori-jakaumien perheelle.

17.5. Tilastolliset menetelmät tilastollisen mallin valinnassa: Kommentteja

Tilastollisen mallin valinnassa käytettävät *tilastolliset kriteerit*:

(i) Valittu malli selviää *diagnostisista tarkistuksista*; ks. lukua **Regressiodiagnostiikka**.

(ii) Valitun mallin parametrit ovat *tilastollisesti merkitseviä*; ks. kappaletta **Mallinvalintatestit**.

Tilastollisia malleja ei pidä koskaan valita *pelkästään* tilastollisin kriteerein.

Tilastollisen mallin valinnassa käytettävät *asialoogiset kriteerit*:

(i) Ovatko mallin parametrit *tulkittavissa*?

(ii) Ovatko mallin parametrit *järkevän merkkisiä* ja *järkevän kokoisia*?

(iii) Kuvaako malli todellisuutta *mielekkäällä tavalla*?

Asialoogisia kriteereitä ei voida asettaa tilastotieteestä käsin. Vain tutkimuksen kohteena olevan *ilmiön perustellinen tuntemus* ja *ilmiötä koskeva teoria* mahdollistavat asialoogisten kriteerien asettamisen.

Tilastolliset mallit pitää *aina* alistaa myös asialoogisiin tarkistuksiin.

17.6. Epälineaaristen riippuvuuksien linearisointi

Jos selitettävän muuttujan *y* tilastollinen riippuvuus selittäjistä $x_1, x_2, \ldots, x_k$ on **epälineaarinen**, riippuvuuden analysointi vaatii yleensä *epälineaarisen regressiomallin* rakentamista. Sivuutamme epälineaaristen regressiomallien käsittelyn tässä.

Joskus selitettävän muuttujan *y* ja selittävien muuttujien $x_1, x_2, \ldots, x_k$ välinen epälineaarinen tilastollinen riippuvuus voidaan kuitenkin *linearisoida selitettävän muuttujan ja selittäjien sopivilla muunnoksilla* niin, että linearisoinnin tuloksena syntynyt *transformoitu malli toteuttaa yleisen lineaarisen mallin standardioletukset*. Rajoitumme tässä linearisoivien muunnosten käytön kuvaamiseen *yhden selittäjän* tapauksessa.

Linearisointi yhden selittäjän regressiomalleissa

Olkoot

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

selitettävän muuttujan *y* havaittuja arvoja ja

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

selittävän muuttujan *x* havaittuja arvoja, jotka liittyvät kaikille $i = 1, 2, \ldots, n$ samaan havainto-yksikköön.

Oletetaan, että selitettävän muuttujan *y* tilastollinen riippuvuus selittäjästä *x* on *epälineaarista*. Sanomme, että selitettävän muuttujan *y* ja selittäjän *x* välinen epälineaarinen tilastollinen riippuvuus voidaan **linearisoida**, jos on olemassa *bijektiiviset kuvaukset f* ja *g* niin, että muunnetuille havaintoarvoille

$$(f(x_i), g(y_i)),\ i = 1, 2, \mathrm{K}, n$$

pätee regressiokertoimien β_0 ja β_1 suhteen lineaarinen esitys

$$f(y_i) = \beta_0 + \beta_1 g(x_i) + \varepsilon_i,\ i = 1, 2, \mathrm{K}, n$$

jossa jäännöstermit ε_i toteuttavat yleisen lineaarisen mallin *standardioletukset*. Tällöin *transformoituun malliin* voidaan soveltaa tavanomaisia *lineaarisen mallin estimointi-* ja *testaustekniikoita*.

Linearisoivien muunnosten etsiminen

Parhaimmillaan linearisoivat muunnokset *f* ja *g* löytyvät *taustateorian* kuten fysiikan tai taloustieteen avulla. Sopivien muunnosten etsimisissä voidaan kuitenkin usein käyttää apuna *tilastografiikkaa*:

(i) Piirretään selitettävän muuttujan *y* ja selittäjän *x* havaituista arvoista pistediagrammi

$$(x_i, y_i),\ i = 1, 2, \mathrm{K}, n$$

(ii) Piirretään selitettävän muuttujan *y* ja selittäjän *x* havaittujen arvojen muunnoksista pistediagrammit

$$(g(x_i),\ f(y_i)),\ i = 1, 2, \mathrm{K}, n$$

funktioiden *f* ja *g* kaikille mahdollisille *kandidaateille*.

Muuttujien *y* ja *x* tilastollisen riippuvuuden epälineaarisuus näkyy pistediagrammin

$$(x_i, y_i),\ i = 1, 2, \mathrm{K}, n$$

pistepilven tai -parven *käyryytenä*. Jos funktiot *f* ja *g* onnistuvat linearisoimaan muuttujien *y* ja *x* välisen epälineaarisen tilastollisen riippuvuuden, pistediagrammin

$$(g(x_i),\ f(y_i)),\ i = 1, 2, \mathrm{K}, n$$

pistepilvessä tai -parvessa *ei näy käyryyttä*.

Sopivien muunnosten *f* ja *g* etsimisessä auttaa usein myös seuraava tekniikka:

(i) Estimoidaan transformoidut mallit

$$f(y_i) = \beta_0 + \beta_1 g(x_i) + \varepsilon_i,\ i = 1, 2, \mathrm{K}, n$$

funktioiden *f* ja *g* kaikille mahdollisille *kandidaateille*.

(ii) Piirretään estimointituloksista seuraavat *residuaalikuviot*:

Standardoidut residuaalit sovitteita vastaan:

$$(\hat{f}(y_i), \mathrm{Std}(e_i)),\ i = 1, 2, \mathrm{K}, n$$

Standardoidut residuaalit selittäjän arvoja vastaan:

$$(x_i, \mathrm{Std}(e_i)),\ i = 1, 2, \mathrm{K}, n$$

Jos funktiot *f* ja *g* eivät onnistu linearisoimaan muuttujien *y* ja *x* epälineaarista tilastollista riippuvuutta, residuaalikuvioiden pistepilvissä näkyy *käyryyttä*. Sen sijaan, jos funktiot *f* ja *g* onnistuvat linearisoimaan muuttujien *y* ja *x* epälineaarisen tilastollisen riippuvuuden, residuaalikuvioiden pistepilvissä *ei näy käyryyttä*.

Linearisoivia muunnoksia

Alla oleva taulukko esittää sellaisia funktioiden *f* ja *g* kombinaatioita, joiden on monissa sovellustilanteissa havaittu tuottavan *linearisoidun esityksen*

$$f(y) = \beta_0 + \beta_1 g(x)$$

muuttujien *y* ja *x* tilastolliselle riippuvuudelle.

$f(y)$	$g(x)$		
	x	$1/x$	$\log(x)$
y	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1 / x$	$y = \beta_0 + \beta_1 \log(x)$
$1/y$	$1/y = \beta_0 + \beta_1 x$	$1/y = \beta_0 + \beta_1 / x$	$1/y = \beta_0 + \beta_1 \log(x)$
$\log(y)$	$\log(y) = \beta_0 + \beta_1 x$	$\log(y) = \beta_0 + \beta_1 / x$	$\log(y) = \beta_0 + \beta_1 \log(x)$

Olkoot funktiot *f* ja *g* kuten esityksessä

$$f(y) = \beta_0 + \beta_1 g(x)$$

edellä. Alla oleva taulukko esittää yhtälön ratkaisuja muuttujan *y* suhteen.

$f(y)$	$g(x)$		
	x	$1/x$	$\log(x)$
y	$y = \beta_0 + \beta_1 x$	$y = \beta_0 + \beta_1 / x$	$y = \beta_0 + \beta_1 \log(x)$
$1/y$	$y = \dfrac{1}{\beta_1\left(x + \dfrac{\beta_0}{\beta_1}\right)}$	$y = \dfrac{1}{\beta_0} - \dfrac{\beta_1}{\beta_0^2} \cdot \dfrac{1}{x + \dfrac{\beta_1}{\beta_0}}$	$y = \dfrac{1}{\beta_1\left(\log(x) + \dfrac{\beta_0}{\beta_1}\right)}$
$\log(y)$	$y = e^{\beta_0} e^{\beta_1 x}$	$y = e^{\beta_0} e^{\beta_1 / x}$	$y = e^{\beta_0} x^{\beta_1}$

Vaatimukset muunnoksille

On syytä huomata, että *ei riitä*, että valitut muunnokset tuottavat lineaarisen mallin, joka *sopii hyvin havaintoihin*, vaan käytettävien muunnosten *pitää toteuttaa* selitettävän muuttujan ja selittäjän käyttäytymiseen liittyvät *loogisuusehdot*:

(i) Muunnosfunktioiden *määrittely-* ja *arvoalueiden* pitää liittyä loogisella tavalla selitettävän muuttujan ja selittäjän mahdollisten arvojen alueisiin.

(ii) Muunnosfunktioiden *asymptoottisen käyttäytymisen* pitää vastata loogisella tavalla selitettävän muuttujan ja selittäjän mahdollisten arvojen käyttäytymistä niiden äärialueilla.

18. Regressiodiagnostiikka

Regressioanalyysin perusongelma on seuraava: Kuvaako selitettävän muuttujan ja selittäjien riippuvuudelle *spesifioitu* eli *täsmennetty* regressiomalli riippuvuutta *oikein*?

Tämän yleisen ongelman tärkeä osaongelma on se, *pätevätkö mallista tehdyt oletukset*. Mallista tehtyjen *oletusten tarkistamista* kutsutaan tavallisesti **regressiodiagnostiikaksi**. Oletusten tarkistaminen tapahtuu tutkimalla mallista saatavia *estimointituloksia*. Ajatuksena on se, että *spesifioinnissa tapahtuneiden virheiden pitäisi näkyä estimointituloksissa*.

Mallia pidetään *hyvänä*, jos mallista tehdyt oletukset ja estimointitulokset *ovat sopusoinnussa*. Jos estimointitulokset *eivät ole sopusoinnussa* mallista tehtyjen oletuksen kanssa, katsotaan usein, että *selitettävän muuttujan ja selittäjien riippuvuudelle spesifioitu malli ei kuvaa oikein riippuvuutta*.

Avainsanat:

Aikasarjadiagrammi, Apuregressio, Autokorrelaatio, Bowmanin ja Shentonin testi, Chow-testi, Cookin etäisyys, Diagnostinen testi, Durbinin ja Watsonin testi, Ennustaminen, Ennuste, Ennustusvirhe, *F*-testi, Harha, Harhattomuus, Hattumatriisi, Heteroskedastisuus, Homoskedastisuus, Jäännösneliösumma, Jäännöstermi, Jäännösvarianssi, χ^2-testi, Keskihajonta, Kokonaisneliösumma, Korrelaatio, Korrelaatiomatriisi, Kovarianssi, Kovarianssimatriisi, Kuntoisuusluku, Leverage, Lineaarinen regressiomalli, Lineaarinen riippuvuus, Malli, Mallin hyvyys, Mallineliösumma, Matriisin aste, Modifioidut standardioletukset, Momenttimatriisi, Multikollineaarisuus, Normaali havainto, Normaalisuusoletus, Ominaisarvo, Painotettu pienimmän neliösumman menetelmä, Parametri, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Pistediagrammi, Poikkeava havainto, Poistoresiduaali, Rakenneosa, Rankit Plot -kuvio, Regressiodiagnostiikka, Regressiografiikka, Regressiokerroin, Regressiomalli, Residuaali, Residuaalidiagrammi, Satunnainen osa, Selittäjien valinta, Selittäminen, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Sovite, Spesifikaatio, Spesifiointi, Spesifiointivirhe, Stabilointi, Standardioletus, Standardointi, Systemaattinen osa, Testi, Täsmentäminen, Usean selittäjän lineaarinen regressiomalli, Vakioparametrisuusoletus, Vakioselittäjä, Varianssi, Varianssianalyysihajotelma, Varianssin inflaatiotekijä, Vipuluku, Virhetermi, Wilkin ja Shapiron testi, Usean selittäjän lineaarinen regressiomalli, Yleinen lineaarinen malli

18.1. Regressiomallit ja regressiodiagnostiikka

Oletetaan, että tavoitteena on **selittää** *selitettävän muuttujan y havaittujen arvojen vaihtelu selittävien muuttujien* eli *selittäjien* $x_1, x_2, \ldots, x_k$ *havaittujen arvojen vaihtelun avulla*. Sitä varten selitettävän muuttujan *y* tilastolliselle riippuvuudelle selittäjistä $x_1, x_2, \ldots, x_k$ pyritään rakentamaan tilastollinen malli, jota kutsutaan *regressiomalliksi*.

Olkoon

$$y_i = f(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}; \boldsymbol{\beta}) + \varepsilon_i \,, i = 1, 2, \mathrm{K}, n$$

selitettävän muuttujan *y* **regressiomalli** selittäjien $x_1, x_2, \ldots, x_k$ suhteen. Tällöin

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

x_{ij} = **selittävän muuttujan** x_j *havaittu arvo* havaintoyksikössä $i, j = 1, 2, \ldots, k$

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

Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen funktio

$$f(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}; \boldsymbol{\beta})$$

muodostaa mallin **systemaattisen osan** eli **rakenneosan** ja jäännöstermi ε_i muodostaa mallin **satunnaisen osan**. Mallin systemaattinen osa kuvaa selitettävän muuttujan *y tilastollista riippuvuutta* selittäjistä $x_1, x_2, \ldots, x_k$. Mallin systemaattisen osan määräävä funktio *f* riippuu parametrista

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_k)$$

joka tarkemmin määrää funktion *f* muodon.

Huomautus:

- Tavallisesti parametrin $\boldsymbol{\beta}$ arvo on *tuntematon* ja se on siksi *estimoitava* havainnoista.

Regressioanalyysin peruskysymykset

(i) Kuvaako malli selitettävän muuttujan ja selittäjien välistä riippuvuutta *sisällöllisesti oikein*?

Tämä kysymys *ei ole tilastotieteellinen* ja siihen vastaaminen vaatii tutkittavaa ilmiötä kuvaavan *taustateorian* tuntemusta.

(ii) Kuvaako malli selitettävän muuttujan ja selittäjien välistä riippuvuutta *tilastollisesti oikein*?

Tämä kysymys *on tilastotieteellinen* ja siihen voidaan pyrkiä vastaamaan tilastotieteen keinoin.

Regressioanalyysin peruskysymykset ja regressiodiagnostiikka

Regressiomallia voidaan pitää *tilastollisesti oikeana*, jos mallista saadut estimointitulokset ovat sopusoinnussa mallia koskevien oletuksien kanssa. Siksi regressiomallia koskevien *oletuksien tarkistaminen* – eli **regressiodiagnostiikka** – muodostaa keskeisen osan regressioanalyysin soveltamista.

Regressiodiagnostiikassa käytetään seuraavia menetelmiä:

- Estimoinnin onnistumista *havainnollistetaan* **tilastografiikalla**.

- Estimoinnin onnistumista *kuvataan* **diagnostisilla tunnusluvuilla**.
- Mallia koskevia oletuksia *testataan* **diagnostisilla testeillä**.

Regressiomallin spesifiointi

Tilastollisen *mallin muodon* ja *mallia koskevien oletuksien* määrittelemistä kutsutaan mallin **spesifioinniksi** eli **täsmentämiseksi**. Määriteltyä mallia kutsutaan **spesifikaatioksi** tai **täsmennykseksi**.

Regressiomallin spesifioiminen tarkoittaa seuraavien valintojen tekemistä:

(i) **Selitettävän muuttujan ja selittäjien valinta**.

(ii) **Systemaattisen eli rakenneosan funktionaalisen muodon ja parametroinnin valinta**.

(iii) **Selitettävän muuttujan ja selittäjien funktionaalisen muodon valinta**.

(iv) **Jäännöstermiä koskevien stokastisten oletuksien valinta**.

Valinnat (i)-(iii) liittyvät ensisijaisesti regressiomallin *rakenneosan spesifiointiin*, kun taas valinta (iv) liittyy ensisijaisesti regressiomallin *jäännöstermin spesifiointiin*.

Huomautus:

- Valinnat (i)-(iv) eivät ole toisistaan riippumattomia.

18.2. Yleinen lineaarinen malli

Oletetaan, että muuttujien y ja $x_1, x_2, \ldots, x_k$ havaittujen arvojen välillä vallitsee *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathrm{K}, n$$

jossa

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

x_{ij} = **selittävän muuttujan** eli **selittäjän** x_j *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä i, $j = 1, 2, \ldots, k$

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

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

β_j = **selittäjän** x_j **regressiokerroin**, $j = 1, 2, \ldots, k$; β_j on *ei-satunnainen* ja *tuntematon vakio*

Tällöin yhtälö määrittelee **usean selittäjän lineaarisen regressiomallin**, jota kutsutaan **yleiseksi lineaariseksi malliksi**.

Seuraavassa kertaamme yleisen lineaarisen mallin *formuloinnin matriisein*, mallia koskevat *standardioletukset* ja pääkohdat mallin parametrien *estimoinnista*; lisätietoja: ks. lukua **Yleinen lineaarinen malli**.

Yleinen lineaarinen malli voidaan esittää matriisein muodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa

$\mathbf{y}$ = **selitettävän muuttujan** *y havaittujen arvojen* muodostama *satunnainen* *n*-vektori

$\mathbf{X}$ = **selittäjien** $x_1, x_2, \ldots, x_k$ *havaittujen arvojen* ja *ykkösten* muodostama $n\times(k+1)$-matriisi

$\boldsymbol{\beta}$ = **regressiokertoimien** muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k+1)$-vektori

$\boldsymbol{\varepsilon}$ = **jäännöstermien** muodostama *ei-havaittu* ja *satunnainen* *n*-vektori

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** esitetään matriisein seuraavassa muodossa:

(i) Matriisin $\mathbf{X}$ alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*

(ii) Matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k+1$$

(iii) $$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

(iv)&(v) *Homoskedastisuus-* ja *korreloimattomuusoletus*:

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(vi) *Normaalisuusoletus*:

$$\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*, mallia koskevat **modifioidut standardioletukset** esitetään matriisein seuraavassa muodossa:

(i)´ Matriisin $\mathbf{X}$ alkiot ovat *satunnaismuuttujia*.

(ii)´ Matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k+1$$

(iii)´ $$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

(iv)´&(v)´ *Homoskedastisuus-* ja *korreloimattomuusoletus*:

$$\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$$

(vi)´ *Normaalisuusoletus*:

$$\boldsymbol{\varepsilon} \mid \mathbf{X} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Mallin rakenneosa ja jäännösosa

Oletetaan, että yleistä lineaarista mallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

koskevat standardioletukset pätevät. Tällöin selitettävä muuttujan arvojen vektori **y** voidaan esittää seuraavalla tavalla kahden osatekijän summana:

$$\mathbf{y} = \mathrm{E}(\mathbf{y} \mid \mathbf{X}) + \boldsymbol{\varepsilon}$$

Osatekijä

$$\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

muodostaa mallin **systemaattisen** eli **rakenneosan**, joka riippuu selittäjien $x_1, x_2, \ldots, x_k$ havaituista arvoista. Jäännöstermi $\boldsymbol{\varepsilon}$ muodostaa mallin **satunnaisen osan**, joka ei riipu selittäjien $x_1, x_2, \ldots, x_k$ havaituista arvoista.

Regressiokertoimien PNS-estimaattorit ja niiden ominaisuudet

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i \,,\ i = 1, 2, \mathrm{K}\,, n$$

regressiokertoimien

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

PNS- eli **pienimmän neliösumman estimaattorit**

$$b_0, b_1, b_2, \ldots, b_k$$

minimoivat jäännös- eli *virhetermien* ε_i *neliösumman*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \mathrm{L} - \beta_k x_{ik})^2$$

kertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ *suhteen.*

Yleisen lineaarisen mallin $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ regressiokertoimien vektorin

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

PNS-estimaaattori voidaan esittää *matriisein* muodossa

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

PNS-estimaattorilla **b** on standardioletuksien (i)-(vi) pätiessä seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$$

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{b} \quad \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Estimoidun mallin sovitteet ja residuaalit sekä niiden ominaisuudet

Olkoon

$$\mathbf{b} = (b_0, b_1, b_2, \ldots, b_k)$$

yleisen linearegressiokertoimien vektorin

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

PNS-estimaattori.

Määritellään estimoidun mallin **sovitteet** $\hat{y}_i$ kaavalla

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_k x_{ik} \ , i = 1, 2, \ldots, n$$

Määritellään estimoidun mallin **residuaalit** e_i kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \cdots - b_k x_{ik} \ , i = 1, 2, \ldots, n$$

Sovitteiden muodostama *n*-vektori voidaan esittää *matriisein* muodossa

$$\hat{\mathbf{y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{Py}$$

Residuaalien muodostama *n*-vektori voidaan esittää *matriisein* muodossa

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = (\mathbf{I} - \mathbf{P})\mathbf{y} = \mathbf{My}$$

Huomautus:

- Koska *residuaalit kuvaavat estimoidun regressiomallin ja havaintoarvojen yhteensopivuutta*, monet regressiodiagnostiikan menetelmistä perustuvat estimoidun regressiomallin residuaaleihin tai niiden muunnoksiin.

Sovitteiden muodostamalla n-vektorilla $\hat{\mathbf{y}}$ on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$$
$$\mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{P} = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Residuaalien muodostamalla n-vektorilla **e** on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}$$
$$\mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{P}) = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$

Huomautus:

- Yllä olevan mukaan residuaalit e_j ovat yleensä sekä *heteroskedastisia* että *korreloituneita*, vaikka jäännöstermit ε_j on oletettu *homoskedastisiksi* ja *korreloimattomiksi*.

Matriisit

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$
$$\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

ovat *symmetrisiä* ja *idempotentteja* eli *projektioita*:

$$\mathbf{P}' = \mathbf{P} \qquad \mathbf{P}^2 = \mathbf{P}$$
$$\mathbf{M}' = \mathbf{M} \qquad \mathbf{M}^2 = \mathbf{M}$$

Lisäksi

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}$$

Matriisia **P** kutsutaan regressiodiagnostiikassa usein *hattumatriisiksi*.

Jäännösvarianssin estimointi

Yleisen lineaarisen mallin jäännöstermien ε_i varianssin eli **jäännösvarianssin σ^2 harhaton estimaattori** on

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

jossa

e_i = estimoidun mallin *residuaali*, $i = 1, 2, \ldots, n$

n = havaintojen lukumäärä

k = (aitojen) selittäjien x_j lukumäärä

Yleisen lineaarisen mallin rakenneosan spesifiointi

Yleistä lineaarista mallia

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

sovellettaessa *pääasiallinen kiinnostus kohdistuu* mallin **systemaattisen osan** eli **rakenneosan**

$$\mathrm{E}(\mathbf{y}\,|\,\mathbf{X}) = \mathbf{X\beta}$$

oikeaan spesifiointiin eli *täsmentämiseen,* koska mallin rakenneosa kuvaa selitettävän muuttujan *y* riippuvuutta selittäjistä $x_1, x_2, \ldots, x_k$. *Virheet mallin rakenneosan spesifioinnissa johtavat virheellisiin johtopäätöksiin* selitettävän muuttujan ja selittäjien välisestä riippuvuudesta.

Spesifiointivirheet mallin rakenneosassa:

(i) Sovelletaan lineaarista mallia, vaikka selitettävän muuttujan *y* riippuvuus selittäjistä $x_1, x_2, \ldots, x_k$ **ei ole lineaarista**.

(ii) Mallissa on **väärät selittäjät**:

- Mallista *puuttuu* selittäjiä.
- Mallissa *on liikaa* selittäjiä.

(iii) Selitettävä muuttuja ja/tai selittäjät ovat mallissa **väärässä funktionaalisessa muodossa**.

(iv) Oletetaan virheellisesti, että regressiokertoimet ovat **vakioita**.

Kommentteja:

- *Epälineaaristen regressiomallien* käsittely sivuutetaan tässä esityksessä.
- *Selittäjien valinta* on regressioanalyysin keskeisiä – ja vaikeimpia – ongelmia; ks. lukua **Regressiomallin valinta**.
- Sopiva selitettävän muuttujan ja/tai selittäjien muunnos saattaa *linearisoida* selitettävän muuttujan ja selittäjien epälineaarisen riippuvuuden; ks. lukua **Regressiomallin valinta**.
- Parametrien vakioisuutta on mahdollista **testata**; ks. kappaletta **Parametrien vakioisuus**.
- Vain huolellinen perehtyminen tutkittavan ilmiön **taustateoriaan** mahdollistaa regressiomallin rakenneosan spesifioinnin oikein.

- *Spesifiointivirheet* regressiomallin rakenneosassa tulevat tavallisesti esiin estimoidun mallin **residuaaleissa**.

Yleisen lineaarisen mallin jäännösosan spesifiointi

Vaikka yleistä lineaarista mallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

sovellettaessa *pääasiallinen kiinnostus kohdistuu* mallin **systemaattisen osan** eli **rakenneosan**

$$\mathrm{E}(\mathbf{y}\,|\,\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

oikeaan spesifiointiin, on syytä huomata, että mallin jäännöstermille

$$\boldsymbol{\varepsilon}$$

valittu spesifikaatio eli täsmennys vaikuttaa sekä estimointimenetelmän valintaan että mallista tehtävään tilastolliseen päättelyyn.

Spesifiointivirheet mallin jäännöstermissä:

(i) Oletetaan virheellisesti, että jäännöstermi $\boldsymbol{\varepsilon}$ on **homoskedastinen** ja **korreloimaton**.

(ii) Oletetaan virheellisesti, että jäännöstermi $\boldsymbol{\varepsilon}$ on **normaalinen**.

Kommentteja:

- Jos jäännöstermiä koskeva homoskedastisuus- tai korreloimattomuusoletus *ei päde, regressiokertoimien PNS-estimaattorit eivät ole parhaita* Gaussin ja Markovin lauseen mielessä; ks. lukua **Regressiomallin erityiskysymyksiä**.
- Jos *jäännöstermiä* koskeva normaalisuusoletus *ei päde*, *t*- ja *F*-jakaumiin perustuva *tilastolliset testit eivät välttämättä ole päteviä*.
- *Spesifiointivirheet* regressiomallin jäännöstermissä näkyvät tavallisesti estimoidun mallin **residuaaleissa**.
- Estimoidun mallin *residuaaleissa* havaittu *heteroskedastisuus, korreloituneisuus* tai *epänormaalisuus ei kuitenkaan välttämättä merkitse sitä, että mallin* **jäännöstermi** *on spesifioitu väärin*.
- Residuaalien *heteroskedastisuus, korreloituneisuus* tai *epänormaalisuus saattavat indikoida myös sitä, että mallin* **rakenneosa** *on spesifioitu väärin*.

Spesifiointivirheiden vaikutukset

Regressioanalyysissa pääkiinnostus kohdistuu oikean spesifikaation löytämiseen regressiomallin systemaattiselle osalle eli rakenneosalle, koska juuri rakenneosa kuvaa selitettävän muuttujan riippuvuutta selittäjistä. Regressiomallin *jäännöstermin spesifikaatio vaikuttaa kuitenkin voimakkaasti sekä mallin estimointiin että testaukseen*.

On syytä huomata, että rakenneosalle valittu spesifikaatio vaikuttaa tavallisesti mallin jäännöstermille valittavaan spesifikaatioon ja kääntäen jäännöstermille valittu spesifikaatio vaikuttaa mallin rakenneosalle valittavaan spesifikaatioon.

Monet regressiodiagnostiikan menetelmät perustuvat siihen, että sekä regressiomallin *rakenneosan* että *jäännöstermin virheellinen spesifiointi* näkyvät tavallisesti estimoidun mallin *residuaaleissa*.

Diagnostiset tarkistukset

Regressiomallin spesifikaation tilastollista validiteettia on aina syytä tutkia alistamalla malli seuraavien **diagnostisten tarkistusten** kohteeksi:

(i) Onko havaintojen joukossa regressioanalyysin tuloksia vääristäviä **poikkeavia havaintoja**?

(ii) Ovatko regressiokertoimet **vakioita**?

(iii) Ovatko selittäjät **itsenäisiä**?

(iv) Ovatko mallin jäännöstermit **homoskedastisia**?

(v) Ovatko mallin jäännöstermit **korreloimattomia**?

(vi) Ovatko mallin jäännöstermit **normaalisia**?

On syytä muistaa, että voimakkain testi mille tahansa tieteelliselle selitysmallille on sen kyky ennustaa. Siksi regressiomalleja sovellettaessa on aina syytä testata mallin ennustuskykyä tavanomaisten diagnostisten tarkistusten lisäksi.

18.3. Regressiografiikka

Regressiomallin hyvyyttä voidaan tutkia mallista saatuja estimointituloksia havainnollistavien *graafisten esitysten* avulla.

Regressiografiikan standardikuviot:

(i) Kuviot, joiden avulla estimoidun mallin *sovitteita* verrataan selitettävän muuttujan havaittuihin arvoihin.

(ii) Kuviot, joiden avulla havainnollistetaan estimoidun mallin *residuaalien käyttäytymistä*.

Pistediagrammit

Koska hyvällä regressiomallilla estimoidun mallin sovitteet ovat lähellä selitettävän muuttujan havaittuja arvoja, mallin hyvyyttä voidaan tutkia vertaamalla sovitteita selitettävän muuttujan havaittuihin arvoihin piirtämällä niiden riippuvuutta havainnollistava **pistediagrammi**:

Piirretään *sovitteet* $\hat{y}_i$ *selitettävän muuttujan y havaittuja arvoja* y_i *vastaan* eli esitetään lukuparit

$$(y_i, \hat{y}_i),\ i = 1, 2, \mathrm{K}, n$$

pisteinä avaruudessa ^{2}. Regressiomalli on sitä *parempi* mitä *lähempänä* pisteet suoraa, jonka kulmakerroin = 1.

Pisteiden

$$(y_i, \hat{y}_i),\ i = 1, 2, \mathrm{K}, n$$

muodostaman pistepilven tai -parven *käyryys* viittaa *mallin rakenneosan väärään spesifikaatioon* eli *täsmennykseen*. *Poikkeavat havainnot* erottuvat tavallisesti em. suorasta muita pisteitä kauempana olevina pisteinä.

Regressiomallin hyvyyden mittarina voidaan käyttää selitettävän muuttujan *y* havaittujen arvojen y_i ja estimoidun mallin sovitteiden $\hat{y}_i$ *otoskorrelaatiokerrointa*

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

Jos estimoitu regressiomalli on lineaarinen ja mallissa on vakio,

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

jossa R^2 on estimoidun mallin *selitysaste*.

Residuaalidiagrammit

Koska *hyvällä regressiomallilla estimoidun mallin residuaalit ovat pieniä*, regressiomallin hyvyyttä voidaan tutkia piirtämällä estimoidun mallin residuaaleista kuviot, joita kutsutaan **residuaali-diagrammeiksi**:

(i) Piirretään *residuaalit e_i sovitteita $\hat{y}_i$ vastaan* eli esitetään lukuparit

$$(\hat{y}_i, e_i),\ i = 1, 2, \mathrm{K}, n$$

pisteinä avaruudessa ^{2}.

(ii) Piirretään *residuaalit e_i eri selittäjien $x_1, x_2, \ldots, x_k$ havaittuja arvoja x_{ij} vastaan* eli esitetään lukuparit

$$(x_{ij}, e_i),\ i = 1, 2, \mathrm{K}, n,\ j = 1, 2, \mathrm{K}, k$$

pisteinä avaruudessa ^{2}.

Oikein täsmennetyn regressiomallin residuaalidiagrammissa pisteet muodostavat *vaakasuoran vyömäisen pistepilven* tai *-parven* eli kuvion pisteet muodostavat edettäessä vasemmalta oikealle yleisilmeeltään tasaleveän pilven, jossa ei näy poikkeavia havaintoja.

Pistepilven *käyristyminen* viittaa regressiomallin *rakenneosan väärään spesifikaatioon* eli *täsmennykseen*. Syitä:

(i) Selitettävän muuttujan riippuvuus selittäjistä *ei ole lineaarista*.

(ii) Mallissa *ei ole oikeita selittäjiä*.

(iii) Selitettävä muuttuja ja/tai selittäjät *eivät ole oikeassa funktionaalisessa muodossa*.

Jos pistepilvi *ei ole tasaleveä* (esim. pilvi levenee oikealle tai vasemmalle), regressiomallin *jäännöstermi saattaa olla* **heteroskedastinen**. On kuitenkin syytä huomata, että estimoidun mallin *residuaalien heteroskedastisuus saattaa viitata myös mallin rakenneosan väärään spesifikaatioon* eli *täsmennykseen*.

Aikasarjadiagrammit

Aikasarjojen regressiomalleissa oletetaan, että havainnot on järjestetty ajassa niin, että havaintoindeksin

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

arvot viittaavat peräkkäisiin ajanhetkiin.

Huomautus:

- Aikasarjoissa havaintoindeksinä käytetään usein kirjainta *t*:

$$t \leftarrow time$$

Aikasarjojen regressiomallien spesifikaation hyvyyttä tutkitaan tavallisesti piirtämällä seuraavat **aikasarjadiagrammit**:

(i) Piirretään selitettävän muuttujan y havaitut arvot

$$y_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

ja estimoidun mallin *sovitteet*

$$\hat{y}_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

aikasarjoina samaan kuvioon.

(ii) Piirretään estimoidun mallin *residuaalit*

$$e_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

aikasarjana.

Aikasarjadiagrammit ovat *pistediagrammeja*, joissa ko. muuttujan arvot piirretään aikaa vastaan ja lisäksi ajassa peräkkäisiin havaintoihin liittyvät pisteet yhdistetään janalla.

Regressiomalli on sitä *parempi*, mitä *lähempänä* estimoidun mallin sovitteiden muodostama aikasarja

$$\hat{y}_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

kulkee selitettävän muuttujan havaittujen arvojen muodostamaa aikasarjaa

$$y_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

tai – mikä on sama asia – mitä *pienempiä* ovat residuaalit

$$e_i\ ,\ i = 1, 2, \mathrm{K}\ , n$$

Aikasarjadiagrammeista nähdään minä ajanhetkinä malli selittää selitettävän muuttujan käyttäytymistä hyvin ja minä huonosti. Jos residuaaliaikasarjan muodostama pistepilvi *ei ole tasaleveä* (esim. pilvi levenee oikealle tai vasemmalle), regressiomallin *jäännöstermi saattaa olla* **heteroskedastinen**. On kuitenkin syytä huomata, että estimoidun mallin *residuaalien hetero-skedastisuus saattaa viitata myös mallin rakenneosan väärään spesifikaatioon* eli täsmennykseen.

Myös jäännöstermin **korreloituneisuus** tulee usein esille residuaaliaikasarjan sisäisessä rytmiikassa (autokorrelaatiorakenteessa). On kuitenkin syytä huomata, että residuaaliaikasarjan *korreloituneisuus saattaa kuitenkin viitata myös mallin rakenneosan väärään spesifikaatioon eli täsmennykseen*.

18.4. Poikkeavat havainnot

Poikkeavalla havainnolla (*engl*. outlier) tarkoitetaan havaintoa, joka *eroaa* jossakin mielessä merkitsevästi *muista havainnoista*.

Tilastollisen analyysin kannalta havaintoa voidaan pitää poikkeavana, jos se *vääristää* tilastollisen analyysin tulokset:

(i) Jos havainnon poistaminen *muuttaa olennaisesti* tilastollisen analyysin tuloksia, havainto on **poikkeava**.

(ii) Jos havainnon poistaminen *ei olennaisesti muuta* tilastollisen analyysin tuloksia, havainto on **normaali**.

Regressioanalyysissa poikkeavat havainnot *saattavat aiheuttaa* seuraavia vaikeuksia:

(i) Mallin *valinta* vaikeutuu.

(ii) Mallin *estimointi* hankaloituu.

(iii) Mallia koskeva *tilastollinen päättely* saattaa vääristyä.

Regressioanalyysissa poikkeavien havaintojen tunnistamiseen käytetään sekä *graafisia menetelmiä* että erityisesti niiden tunnistamiseen konstruoituja *tunnuslukuja*. Poikkeavat havainnot voidaan usein tunnistaa suoraan **residuaalidiagrammeista**; ks. kappaletta **Regressiografiikka**.

Tässä kappaleessa tarkastellaan seuraavia poikkeavien havaintojen tunnistamiseen tarkoitettuja tunnuslukuja:

- **Residuaalit**
- **Standardoidut residuaalit**
- **Poistoresiduaalit**
- **Standardoidut poistoresiduaalit**
- **Vipuluvut eli leverage-luvut**
- **Cookin etäisyydet**

Jos poikkeavia havaintoja havaitaan, ollaan vaikean ongelman edessä: *mitä poikkeaville havainnoille tehdään*? Joskus poikkeavat havainnot kannattaa *poistaa* aineistosta, joskus taas ne kannattaa *korjata* ”normaaleiksi” havainnoiksi; emme käsittele tätä kysymystä tässä esityksessä enempää.

Residuaalit

Olkoon

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

yleinen lineaarinen malli, jossa

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

x_{ij} = **selittävän muuttujan** eli **selittäjän** x_j *havaittu arvo* havaintoyksikössä *i* , j = 1, 2, … , k

β_0 = *vakioselittäjän tuntematon* **regressiokerroin**

β_j = *selittäjän* x_j *tuntematon* **regressiokerroin**, j = 1, 2, … , k

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

Olkoot

$$b_0, b_1, b_2, \ldots, b_k$$

regressiokertoimien

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

PNS-estimaattorit. Määritellään estimoidun mallin **sovitteet** $\hat{y}_i$ kaavalla

$$\hat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \mathsf{L} + b_k x_{ik} \,,\; i = 1, 2, \mathsf{K}, n$$

Määritellään estimoidun mallin **residuaalit** e_i kaavalla

$$e_i = y_i - \hat{y}_i = y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \mathsf{L} - b_k x_{ik} \,,\; i = 1, 2, \mathsf{K}, n$$

Estimoidun mallin residuaaleja e_i voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Voimakkaasti muista residuaaleista poikkeavat residuaalit saattavat viitata *poikkeaviin havaintoihin*.

Standardoidut residuaalit

Koska estimoidun mallin *PNS-residuaalit e_i ovat yleensä heteroskedastisia*, regressio-diagnostiikassa tarkastellaan PNS-residuaalien sijasta usein *standardoituja residuaaleja*.

Residuaalin e_i *varianssi* on

$$\mathrm{D}^2(e_i) = \sigma^2(1-h_{ii}),\ i = 1,2,\mathrm{K},n$$

jossa

$$h_{ii} = \left[\mathbf{P}\right]_{ii},\ i = 1,2,\mathrm{K},n$$

on hattumatriisin

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

i. diagonaalialkio. **Standardoidut** eli *studentisoidut* **residuaalit** Std(e_i) saadaan PNS-residuaaleista e_i kaavalla

$$\mathrm{Std}(e_i) = \frac{e_i}{\hat{\mathrm{D}}(e_i)},\ i = 1,2,\mathrm{K},n$$

Standardoidun residuaalin Std(e_i) kaavassa

$$\hat{\mathrm{D}}^2(e_i) = s^2(1-h_{ii}),\ i = 1,2,\mathrm{K},n$$

on residuaalin e_i varianssin estimaattori, jossa

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

on jäännösvarianssin σ^2 harhaton estimaattori.

Standardoituja residuaaleja Std(e_i) voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Jos estimoitu regressiomalli on riittävä kuvaamaan *kaikkia havaintoja*, standardoitujen residuaalien itseisarvot saavat *vain pienellä todennäköisyydellä suurempia arvoja kuin* 2.5-3. Lukuarvoja 2.5-3 suuremmat standardoitujen residuaalien itseisarvot saattavat viitata *poikkeaviin havaintoihin*. Standardoitujen residuaalien itseisarvoja voidaan verrata Studentin *t-jakaumasta* sopivasti valittuun kriittiseen rajaan.

Poistoresiduaalit

Poikkeavia havaintoja voidaan etsiä poistoresiduaalien avulla:

(i) Estimoidaan malli siten, että havainto *i* jätetään pois.

(ii) Määrätään havaintoa *i* vastaava *poistoresiduaali* selitettävän muuttujan *y* havaitun arvon y_i ja ilman havaintoa *i* estimoidun mallin muuttujalle *y* antaman arvon erotuksena (ennuste-virheenä).

Havaintoa *i* vastaava poistoresiduaali mittaa ilman havaintoa *i* estimoidun mallin kykyä *ennustaa* selitettävän muuttujan *y* arvo havainnossa *i*.

Poistoresiduaalit d_i saadaan PNS-residuaaleista e_i kaavalla

$$d_i = \frac{e_i}{1-h_{ii}}, i = 1, 2, \mathrm{K}, n$$

jossa

$$h_{ii} = \left[\mathbf{P}\right]_{ii}, i = 1, 2, \mathrm{K}, n$$

on hattumatriisin

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

i. diagonaalialkio.

Estimoidun mallin poistoresiduaaleja d_i voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Voimakkaasti muista poistoresiduaaleista poikkeavat poistoresiduaalit saattavat viitata *poikkeaviin havaintoihin*.

Standardoidut poistoresiduaalit

Koska estimoidun lineaarisen regressiomallin *poistoresiduaalit* d_i *ovat yleensä heteroskedastisia*, regressiodiagnostiikassa tarkastellaan poistoresiduaalien sijasta usein *standardoituja poistoresiduaaleja*.

Poistoresiduaalin d_i *varianssi* on

$$\mathrm{D}^2(d_i) = \frac{\sigma^2}{1-h_{ii}}, i = 1, 2, \mathrm{K}, n$$

jossa

$$h_{ii} = \left[\mathbf{P}\right]_{ii}, i = 1, 2, \mathrm{K}, n$$

on hattumatriisin

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

i. diagonaalialkio. **Standardoidut** eli *studentisoidut* **poistoresiduaalit** Std(d_i) saadaan poisto-residuaaleista d_i kaavalla

$$\mathrm{Std}(d_i) = \frac{d_i}{\hat{\mathrm{D}}(d_i)}, i = 1, 2, \mathrm{K}, n$$

jossa

$$\hat{\mathrm{D}}^2(d_i) = \frac{s^2_{(i)}}{1-h_{ii}}, i = 1, 2, \mathrm{K}, n$$

on poistoresiduaalin d_i varianssin estimaattori, jossa $s^2_{(i)}$ on jäännösvarianssin σ^2 harhaton estimaattori mallista, josta havainto i on jätetty pois.

Standardoituja poistoresiduaaleja Std(d_i) voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Jos estimoitu regressiomalli on riittävä kuvaamaan *kaikkia havaintoja*, standardoitujen poistoresiduaalien itseisarvot saavat *vain pienellä todennäköisyydellä suurempia arvoja* kuin 2.5-3. Lukuarvoja 2.5-3 suuremmat standardoitujen poistoresiduaalien itseisarvot saattavat viitata

poikkeaviin havaintoihin. Standardoitujen poistoresiduaalien itseisarvoja voidaan verrata Studentin *t-jakaumasta* sopivasti valittuun kriittiseen rajaan.

Vipuluvut

Poikkeavia havaintoja voidaan etsiä vipulukujen eli leverage-lukujen avulla:

Havaintoa *i* vastaava **vipuluku** (leverage) h_{ii} , $i = 1, 2, \ldots, n$ on *hattumatriisin*

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

i. diagonaalialkio:

$$h_{ii} = \left[\mathbf{P}\right]_{ii}, i = 1, 2, \mathrm{K}, n$$

Vipuluvut h_{ii} ovat verrannollisia havaintopisteiden

$$(x_{i1}, x_{i2}, \mathrm{K}, x_{ik}), i = 1, 2, \mathrm{K}, n$$

etäisyyksiin selittävien muuttujien $x_1, x_2, \ldots, x_k$ havaittujen arvojen x_{ij} aritmeettisten keskiarvojen muodostamasta pisteestä

$$(\overline{x}_1, \overline{x}_2, \mathrm{K}, \overline{x}_k)$$

Vipulukuja h_{ii} voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Jos havaintoa *i* vastaava vipuluku (leverage) h_{ii} on selvästi muita *suurempi*, havainto *i* on *syrjässä* selittävien muuttujien muihin havaintoarvoihin nähden. Syrjässä olevat havainnot saattavat *vääristää* regressioanalyysin tulokset.

Cookin etäisyydet

Poikkeavia havaintoja voidaan etsiä Cookin etäisyyksien avulla:

(i) Estimoidaan malli niin, että kaikki havainnot ovat mukana ja lasketaan estimoidulle mallille sovitteet $\hat{y}_l$, $l = 1, 2, \mathrm{K}, n$.

(ii) Estimoidaan malli jättämällä pois havainto *i* ja lasketaan ilman havaintoa *i* estimoidun mallin selitettävälle muuttujalle *y* antama arvo $\hat{y}_{l(i)}$ kaikille havaintoyksiköille $l = 1, 2, \ldots, n$.

(iii) Verrataan lukuja $\hat{y}_l$ ja $\hat{y}_{l(i)}$ toisiinsa.

Cookin etäisyydet D_i saadaan kaavalla

$$D_i = \frac{\sum_{l=1}^{n} (\hat{y}_l - \hat{y}_{l(i)})^2}{(k+1)s^2}$$

jossa

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

on jäännösvarianssin σ^2 harhaton estimaattori, joka on määrätty, kun mallin estimoinnissa on käytetty kaikkia havaintoja.

Cookin etäisyydet D_i voidaan laskea myös kaavalla

$$D_i = \frac{\mathrm{Std}(e_i)}{k+1} \cdot \frac{h_{ii}}{1-h_{ii}}$$

jossa Std(e_i) on havaintoa i vastaava *standardoitu residuaali* ja

$$h_{ii} = \left[\mathbf{P}\right]_{ii},\ i = 1, 2, \mathrm{K}, n$$

on hattumatriisin

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

i. diagonaalialkio.

Cookin etäisyyksiä D_i voidaan käyttää *poikkeavien havaintojen* tunnistamiseen:

Jos havaintoa i vastaava Cookin etäisyys

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

tai on *selvästi* muiden havaintojen Cookin etäisyyttä suurempi, havainto kannattaa ottaa erikoistarkasteluun.

Tilastografiikka ja poikkeavien havaintojen tunnistaminen

Poikkeavien havaintojen tunnistamiseen tarkoitettujen tunnuslukujen käyttöä voidaan usein helpottaa sopivilla *graafisilla esityksillä*.

Tällöin poikkeavien havaintojen tunnistamiseen käytetyn tunnusluvun havaintokohtaiset arvot

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

piirretään havaintonumeroa vastaan *pistediagrammina*

$$(i, T_i),\ i = 1, 2, \ldots, n$$

jossa tunnusluku T_i voi olla esimerkiksi mikä tahansa tunnusluvuista residuaali, standardoitu residuaali, poistoresiduaali, standardoitu poistoresiduaali, vipuluku tai Cookin etäisyys.

18.5. Regressiokertoimien vakioisuus

Kun yleinen lineaarinen malli spesifioidaan muodossa

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathrm{L} + \beta_k x_{ik} + \varepsilon_i,\ i = 1, 2, \mathrm{K}, n$$

spesifikaatioon sisältyy implisiittisesti mallin regressiokertoimia koskeva **vakioparametrisuusoletus**: Regressiokertoimet

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

ovat samat kaikille havainnoille $i = 1, 2, \ldots, n$. Lisäksi mallia koskeviin standardioletuksiin kuuluu **homoskedastisuusoletus** eli *jäännösvarianssia koskeva* **vakioparametrisuusoletus**:

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

Homoskedastisuusoletuksen testaamista käsitellään kappaleessa **Homoskedastisuus ja heteroskedastisuus**.

Testi regressiokertoimien vakioisuudelle

Jaetaan havainnot

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

kahteen osaan:

Osa 1: $i = 1, 2, \ldots, h$ h kpl

Osa 2: $i = h + 1, h + 2, \ldots, n$ $(n - h)$ kpl

Oletetaan lisäksi, että

$$h \geq k + 1$$

Muodostetaan *kaksi* lineaarista regressiomallia:

(i) Käytetään mallissa (1) havaintoja $i = 1, 2, \ldots, h$.

(ii) Käytetään mallissa (2) havaintoja $i = 1, 2, \ldots, n$.

Malli (1) voidaan esittää matriisein muodossa

$$\mathbf{y}_h = \mathbf{X}_h\boldsymbol{\beta}_h + \boldsymbol{\varepsilon}_h$$

jossa $\mathbf{X}_h$ on $h \times (k + 1)$-matriisi. Tehdään mallista (1) seuraavat oletukset:

$$r(\mathbf{X}_h) = k + 1$$

$$\boldsymbol{\varepsilon}_h \quad \mathrm{N}_h(\mathbf{0}, \sigma_h^2\mathbf{I})$$

Malli (2) voidaan esittää matriisein muodossa

$$\mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta}_n + \boldsymbol{\varepsilon}_n$$

jossa $\mathbf{X}_n$ on $n \times (k + 1)$-matriisi. Tehdään mallista (2) seuraavat oletukset:

$$r(\mathbf{X}_n) = k + 1$$

$$\boldsymbol{\varepsilon}_n \quad \mathrm{N}_n(\mathbf{0}, \sigma_n^2\mathbf{I})$$

Huomaa, että mallin (2) $n \times (k + 1)$-matriisi $\mathbf{X}_n$ voidaan esittää muodossa

$$\mathbf{X}_n = \begin{bmatrix} \mathbf{X}_h \\ \mathbf{X}_2 \end{bmatrix}$$

jossa $(n - h) \times (k + 1)$-matriisi $\mathbf{X_2}$ on liittyy havaintoihin

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

Estimoidaan molemmat mallit (1) ja (2) *PNS-menetelmällä*. Olkoon

SSE_h = jäännösneliösumma mallista (1)

SSE_n = jäännösneliösumma mallista (2)

Muodostetaan ***F*-testisuure**

$$F = \frac{n - k - 1}{n - h} \cdot \frac{SSE_n - SSE_h}{SSE_h}$$

Jos nollahypoteesi

$$\mathrm{H}_0 : \boldsymbol{\beta}_n = \boldsymbol{\beta}_h, \sigma_n^2 = \sigma_h^2$$

pätee, testisuure F noudattaa *F-jakaumaa* vapausastein $(n - h)$ ja $(n - k - 1)$:

$$F \quad F(n-h, n-k-1)$$

Suuret testisuureen arvot viittaavat siihen, että oletus parametrien vakioisuudesta ei päde.

Testi tunnetaan ekonometriassa nimellä **Chow-testi**.

Testin toinen muotoilu

Ennustetaan selitettävän muuttujan *y* arvot havaintoyksiköissä

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

edellä määritellyllä regressiomallilla (1):

$$\hat{y}_i = b_0^1 + b_1^1 x_{i1} + b_2^1 x_{i2} + \mathsf{L} + b_k^1 x_{ik}\ , i = h+1, h+2, \mathsf{K}, n$$

jossa

$$\mathbf{b}_h = \ (b_0^1, b_1^1, b_2^1, \mathsf{K}, b_k^1)$$

regressiokertoimien mallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}_h$ PNS-estimaattori.

Olkoon

$$\mathbf{u} = (u_{h+1}, u_{h+2}, \mathsf{K}, u_n)$$

ennustevirheiden

$$u_i = y_i - \hat{y}_i\ , i = h+1, h+2, \mathsf{K}, n$$

muodostama $(n - h)$-vektori. Vektorilla **u** on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{u}) = \mathbf{0}$$

$$\mathrm{Cov}(\mathbf{u}) = \sigma_h^2(\mathbf{I} + \mathbf{X}_2(\mathbf{X}_h'\mathbf{X}_h)^{-1}\mathbf{X}_2')$$

jossa $\mathbf{X}_2$ on havaintoihin $j = h + 1, h + 2, \ldots, n$ liittyvä osa matriisista $\mathbf{X}_n$. Olkoon lisäksi s_h^2 tavanomainen harhaton estimaattori mallin (1) jäännösvarianssille σ_h^2.

Tällöin matriisi

$$\hat{\mathrm{C}}\mathrm{ov}(\mathbf{u}) = s_h^2(\mathbf{I} + \mathbf{X}_2(\mathbf{X}_h'\mathbf{X}_h)^{-1}\mathbf{X}_2')$$

on ennustevirheiden vektorin **u** kovarianssimatriisin Cov(**u**) estimaattori.

Chow-testisuure nollahypoteesille

$$\mathrm{H}_0 :\ \boldsymbol{\beta}_n = \boldsymbol{\beta}_h\ , \sigma_n^2 = \sigma_h^2$$

voidaan edellä olevia merkintöjä käyttäen esittää muodossa

$$F = \frac{1}{n-h}\mathbf{u}'[\hat{\mathrm{C}}\mathrm{ov}(\mathbf{u})]^{-1}\mathbf{u}$$

Chow-testisuureella *F* on siten seuraava *tulkinta*: Chow-testisuure *F* testaa havainnoista $i = 1, 2, \ldots, h$ estimoidun mallin (1) *kykyä ennustaa* selitettävän muuttujan *y* arvoja havainnoissa $i = h + 1, h + 2, \ldots, n$.

18.6. Multikollineaarisuus

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset toteuttava **yleinen lineaarinen malli**. Standardioletuksen (ii) mukaan selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen ja ykkösten muodostama mukaan $n\times(k + 1)$-matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

Vaatimus siitä, että matriisi $\mathbf{X}$ on *täysiasteisuudesta* merkitsee sitä, että matriisin $\mathbf{X}$ sarakkeiden on oltava *lineaarisesti riippumattomia*.

Regressiokertoimien vektorin $\boldsymbol{\beta}$ **PNS-estimaattori** on

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

PNS-estimaattorin $\mathbf{b}$ *kovarianssimatriisi* on

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Yleisen lineaarinen mallin regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattorin ja sen kovarianssi-matriisin kaavoista nähdään välittömästi: **Jos matriisi $\mathbf{X}$ *on vajaa-asteinen*, PNS-estimaattoria ja sen kovarianssimatriisia *ei voida* muodostaa em. kaavoilla**.

Jos yleisen lineaarisen mallin selittävien muuttujien havaittujen arvojen muodostama $n\times(k + 1)$-matriisi $\mathbf{X}$ *on vajaa-asteinen* eli

$$r(\mathbf{X}) < k + 1$$

niin PNS-estimointi ei siis ole tavanomaisessa mielessä mahdollista. Eräs mahdollinen tapa ratkaista tämä ongelma on poistaa mallista niin monta selittäjää, että jäljelle jääneiden selittäjien havaittujen arvojen (ja ykkösten) muodostama matriisi on täysiasteinen. Emme käsittele vajaa-asteisten lineaaristen regressiomallien tapausta tässä enempää.

Multikollineaarisuus

Jos matriisi $\mathbf{X}$ *on täysiasteinen* eli

$$r(\mathbf{X}) = k + 1$$

mutta matriisin $\mathbf{X}$ sarakkeet ovat lähes lineaarisesti riippuvia, sanomme, että mallin selittäjät ovat **multikollineaarisia**. Multikollineaarisuus *saattaa hankaloittaa* sekä regressiomallin *estimointia* että mallista tehtävää *tilastollista päättelyä*. Voimakas multikollineaarisuus *saattaa hankaloittaa* myös *mallin valintaa*; lisätietoja mallin valinnasta: ks. lukua **Regressiomallin valinta**.

Koska *multikollineaarisuus* on – toisin kuin eksakti lineaarinen riippuvuus – *suhteellinen ominaisuus*, voidaan puhua *multikollineaarisuuden asteesta*. Mitä *vähäisempää* on selittäjien multi-kollineaarisuus, sitä *itsenäisempiä* ovat selittäjät selitettävän muuttujan käyttäytymisen selittäjinä. Jos selittäjät ovat *multikollineaarisia*, ne selittävät jossakin mielessä saman osan selitettävän muuttujan käyttäytymisestä.

Varianssin inflaatiotekijä

Oletetaan, että selitettävää muuttujaa y selitetään lineaarisella regressiomallilla, jonka selittäjinä ovat muuttujat $x_1, x_2, \ldots, x_k$. Olkoon b_j selittäjän x_j regressiokertoimen β_j PNS-estimaattori. Tällöin

$$\mathrm{Var}(b_j) = \frac{1}{1-R_j^2} \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2}$$

jossa R_j^2 on *selitysaste lineaarisesta regressiomallista*, jonka selitettävänä muuttujana on alkuperäisen mallin selittäjä x_j ja selittäjinä ovat kaikki muut alkuperäisen mallin selittäjistä.

Regressiokertoimen b_j varianssin kaavassa esiintyvää tekijää

$$VIF_j = \frac{1}{1-R_j^2},\ j = 1, 2, \mathrm{K}, k$$

kutsutaan selittäjää x_j vastaavaksi **varianssin inflaatiotekijäksi**.

Jos selittäjät $x_1, x_2, \ldots, x_k$ ovat *ortogonaalisia* eli *korreloimattomia*,

$$R_j^2 = 0,\ j = 1, 2, \mathrm{K}, k$$

ja

$$VIF_j = 1,\ j = 1, 2, \mathrm{K}, k$$

Jos selittäjä x_j voidaan esittää muiden selittäjien $x_1, x_2, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$ *lineaarikombinaationa*,

$$R_j^2 = 1$$

ja

$$VIF_j = +\infty$$

Kaavasta

$$\mathrm{Var}(b_j) = VIF_j \cdot \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij}-\overline{x}_j)^2}$$

nähdään edelleen seuraavaa:

(i) Estimaattorin b_j varianssi on sitä *suurempi*, mitä *suurempi* on vastaava varianssin inflaatiotekijä VIF_j .

(ii) Estimaattorin b_j varianssi on sitä *pienempi*, mitä *pienempi* on vastaava varianssin inflaatiotekijä VIF_j .

Selittäjien voimakasta multikollineaarisuutta pidetään usein *haitallisena*, kun taas selittäjien mahdollisimman suurta ortogonaalisuutta pidetään usein *tavoiteltavana*. Esimerkiksi ns. *puhtaissa koeasetelmissa*, joissa selittävien muuttujien arvot voidaan valita, arvot pyritään valitsemaan siten, että selittäjistä tulee ortogonaalisia (tai lähes ortogonaalisia).

Nyrkkisääntönä multikollineaarisuuden haitallisuuden arvioimisessa käytetään joskus seuraavaa sääntöä: Jos

$$VIF_j > 10$$

jollekin $j = 1, 2, \ldots, k$ multikollineaarisuudesta saattaa olla haittaa.

Momenttimatriisi, otoskovarianssimatriisi ja otoskorrelaatiomatriisi

Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen **momenttimatriisi** $\mathbf{A} = [\, a_{rs} \,]$ on $k \times k$-matriisi, jonka r. rivin ja s. sarakkeen alkio a_{rs} on muuttujien x_r ja x_s havaittujen arvojen **tulomomentti**

$$a_{rs} = \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)(x_{is} - \overline{x}_s) \,,\; r = 1, 2, \mathsf{K}\,, k \,,\; s = 1, 2, \mathsf{K}\,, k$$

jossa

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \,,\; j = 1, 2, \mathsf{K}\,, k$$

on selittäjän x_j havaittujen arvojen aritmeettinen keskiarvo. Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen momenttimatriisi $\mathbf{A}$ voidaan esittää matriisein muodossa

$$\mathbf{A} = (\mathbf{Z} - \mathbf{1}\overline{\mathbf{z}}')'(\mathbf{Z} - \mathbf{1}\overline{\mathbf{z}}') = \mathbf{Z}'\mathbf{Z} - n\,\overline{\mathbf{z}}\,\overline{\mathbf{z}}'$$

jossa

$\mathbf{Z}$ = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen muodostama $n \times k$-matriisi

$\overline{\mathbf{z}}$ = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen aritmeettisten keskiarvojen muodostama k-vektori

Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen **otoskovarianssimatriisi** $\mathbf{S} = [\, s_{rs} \,]$ on $k \times k$-matriisi, jonka r. rivin ja s. sarakkeen alkio s_{rs} on muuttujien x_r ja x_s havaittujen arvojen **otoskovarianssi**

$$s_{rs} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ir} - \overline{x}_r)(x_{is} - \overline{x}_s) \,,\; r = 1, 2, \mathsf{K}\,, k \,,\; s = 1, 2, \mathsf{K}\,, k$$

jossa

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \,,\; j = 1, 2, \mathsf{K}\,, k$$

on selittäjän x_j havaittujen arvojen aritmeettinen keskiarvo. Erityisesti

$$s_{jj} = s_j^2 \,,\; j = 1, 2, \mathsf{K}\,, k$$

on selittäjän x_j havaittujen arvojen otosvarianssi. Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen otoskovarianssimatriisi $\mathbf{S}$ voidaan esittää matriisein muodossa

$$\mathbf{S} = \frac{1}{n-1}(\mathbf{Z} - \mathbf{1}\overline{\mathbf{z}}')'(\mathbf{Z} - \mathbf{1}\overline{\mathbf{z}}') = \frac{1}{n-1}\mathbf{A}$$

jossa

$\mathbf{Z}$ = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen muodostama $n \times k$-matriisi

$\overline{\mathbf{z}}$ = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen aritmeettisten keskiarvojen muodostama k-vektori

$\mathbf{A}$ = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen muodostama $k \times k$-momenttimatriisi

Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen **otoskorrelaatiomatriisi R** = [r_{rs}] on $k \times k$-matriisi, jonka r. rivin ja s. sarakkeen alkio r_{rs} on muuttujien x_r ja x_s havaittujen arvojen **otoskorrelaatio**

$$r_{rs} = \frac{s_{rs}}{s_r s_s}, \; r = 1, 2, \mathsf{K}, k, \; s = 1, 2, \mathsf{K}, k$$

jossa

s_{rs} = muuttujien x_r ja x_s havaittujen arvojen otoskovarianssi

$s_r = \sqrt{s_{rr}}$ = on muuttujan x_r otoskeskihajonta

$s_s = \sqrt{s_{ss}}$ = on muuttujan x_s otoskeskihajonta

Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen otoskorrelaatiomatriisi **R** voidaan esittää *matriisein* muodossa

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}$$

jossa

S = *aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen muodostama otoskovarianssimatriisi

D = $\mathrm{diag}(s_1, s_2, \mathsf{K}, s_k)$

= *aitojen* selittäjien $x_1, x_2, \ldots, x_j$ havaittujen arvojen otoskeskihajontojen $s_1, s_2, \ldots, s_k$ muodostama diagonaalimatriisi

Multikollineaarisuuden tutkiminen

Selittäjien $x_1, x_2, \ldots, x_k$ multikollineaarisuutta voidaan tutkia – paitsi tarkastelemalla selittäjiä vastaavia varianssin inflaatiotekijöitä – tutkimalla myös seuraavien matriisien **ominaisarvoja** (ja **ominaisvektoreita**):

(i) *Aitojen* selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen $n \times k$-matriisista **Z** saatava $k \times k$-matriisi **Z´Z**

(ii) Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen *momenttimatriisi* **A**

(ii) Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen *kovarianssimatriisi* **S**

(iii) Selittäjien $x_1, x_2, \ldots, x_k$ havaittujen arvojen *korrelaatiomatriisi* **R**

Multikollineaarisuuden mittarina voidaan käyttää **matriisin kuntoisuuslukua** eli *suurimman ja pienimmän ominaisarvon suhdetta*.

18.7. Homoskedastisuus ja heteroskedastisuus

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset toteuttava **yleinen lineaarinen malli**. Standardioletuksen (iv) mukaan kaikilla mallin jäännöstermeillä ε_i on *sama varianssi*:

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

Tätä oletusta kutsutaan **homoskedastisuusoletukseksi**. Jos *homoskedastisuusoletus ei päde,* niin sanomme, että jäännöstermit ovat **heteroskedastisia** ja kirjoitamme

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

Tällöin siis on olemassa indeksit *r* ja *s* siten, että

$$\mathrm{Var}(\varepsilon_r) = \sigma_r^2 \neq \sigma_s^2 = \mathrm{Var}(\varepsilon_s)$$

Heteroskedastisuuden vaikutukset

Gaussin ja Markovin lauseen (ks. lukua **Yleinen lineaarinen malli**) mukaan yleisen lineaarisen mallin *regressiokertoimien PNS-estimaattorit ovat parhaita kertoimien lineaaristen ja harhattomien estimaattoreiden joukossa, jos mallia koskevat standardioletukset pätevät.* Erityisesti estimaattorreiden varianssit ovat pienimpiä mahdollisia.

Jos regressiomallin jäännöstermit ε_i ovat *heteroskedastisia,* niin Gaussin ja Markovin lauseen ehdot eivät toteudu, jolloin regressiokertoimien PNS-estimaattoreiden *varianssista tulee tarpeettoman suuria,* millä on seuraavat vaikutukset regressiokertoimia koskevaan tilastolliseen päättelyyn (ks. lukua **Yleinen lineaarinen malli**):

(i) Regressiokertoimien luottamusväleistä tulee tarpeettoman leveitä.

(ii) Regressiokertoimia koskevista *testisuureiden arvoista* tulee *tarpeettoman pieniä*.

Lisätietoja: Ks. luvun Erityiskysymyksiä yleisen lineaarisen mallin soveltamisessa kappaletta Yleistetty pienimmän neliösumman menetelmä.

Heteroskedastisuuden havaitseminen

Jäännöstermien heteroskedastisuus voidaan usein havaita estimoidun mallin hyvyyttä havainnollistavista *residuaalidiagrammeista*:

(i) Piirretään *standardoidut residuaalit* $\mathrm{Std}(e_i)$ sovitteita $\hat{y}_i$ vastaan:

$$(\hat{y}_i, \mathrm{Std}(e_i)) \ , i = 1, 2, \mathrm{K} \ , n$$

(ii) Aikasarjojen regressiomalleissa *residuaalit* e_i piirretään yleensä aikasarjana:

$$(i, e_i) \ , i = 1, 2, ..., n$$

Jos residuaalidiagrammin pistepilvi *ei ole tasaleveä* (esim. pilvi *levenee* siirryttäessä kuviossa oikealle tai vasemmalle), regressiomallin jäännöstermi saattaa olla *heteroskedastinen*.

Heteroskedastisuuden testaaminen

Olkoon

$$\hat{y}_i \ , i = 1, 2, \mathrm{K} \ , n$$

estimoidun lineaarisen mallin tuottama *sovite* ja

$$e_i \ , i = 1, 2, \mathrm{K} \ , n$$

vastaava *residuaali*. Määrätään selitysaste R^2 *apuregressiosta*

$$e_i^2 = \alpha_0 + \alpha_1 \hat{y}_i + \delta_i \ , i = 1, 2, \mathrm{K} \ , n$$

Jos homoskedastisuusoletus

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

pätee testisuure nR^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2-jakaumaa yhdellä vapausasteella:

$$nR^2 \;\; {}_a \;\; \chi^2(1)$$

Suuret testisuureen nR^2 arvot johtavat homoskedastisuusoletuksen hylkäämiseen.

Huomautus:

- Ym. homoskedastisuustesti saattaa reagoida myös regressiomallin rakenneosan *väärään spesifikaatioon*.
- Siten ym. **homoskedastisuustestin testisuureen merkitsevä arvo *ei saa* automaattisesti johtaa toimenpiteisiin**, **joilla *pyritään korjaamaan* jäännöstermin *heteroskedastisuus***.

Varianssin stabiloivat muunnokset

Sopiva selitettävän muuttujan arvojen muunnos saattaa stabiloida jäännöstermien varianssin.

Seuraavaan taulukkoon on koottu joukko tällaisia muunnoksia:

Heteroskedastisuuden tyyppi	Varianssin stabiloiva muunnos
$\sigma^2 \propto$ vakio	$y' = y$
$\sigma^2 \propto \mathrm{E}(y)$	$y' = \sqrt{y}$
$\sigma^2 \propto \mathrm{E}(y)[1 - \mathrm{E}(y)]$	$y' = \arcsin(\sqrt{y})$
$\sigma^2 \propto [\mathrm{E}(y)]^2$	$y' = \log(y)$

18.8. Autokorrelaatio

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset toteuttava **yleinen lineaarinen malli**. Standardioletuksen (v) mukaan mallin jäännöstermit ε_i ovat *korreloimattomia*:

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

Tätä oletusta kutsutaan **korreloimattomuusoletukseksi**. Jos *korreloimattomuusoletus ei päde*, niin on olemassa $l \neq i$ siten, että

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

ja sanomme, että jäännöstermit ovat **korreloituneita**.

Korreloituneisuuden vaikutukset

Gaussin ja Markovin lauseen (ks. lukua **Yleinen lineaarinen malli**) mukaan yleisen lineaarisen mallin *regressiokertoimien PNS-estimaattorit ovat parhaita kertoimien lineaaristen ja harhattomien estimaattoreiden joukossa, jos mallia koskevat standardioletukset pätevät*. Erityisesti estimaattorreiden varianssit ovat pienimpiä mahdollisia.

Jos regressiomallin jäännöstermit ε_i ovat *korreloituneita*, niin Gaussin ja Markovin lauseen ehdot eivät toteudu, jolloin regressiokertoimien PNS-estimaattoreiden *varianssista tulee tarpeettoman suuria*, millä on seuraavat vaikutukset regressiokertoimia koskevaan tilastolliseen päättelyyn (ks. lukua **Yleinen lineaarinen malli**):

(i) Regressiokertoimien luottamusväleistä tulee tarpeettoman leveitä.

(ii) Regressiokertoimia koskevista *testisuureiden arvoista* tulee *tarpeettoman pieniä*.

Lisätietoja: Ks. luvun **Erityiskysymyksiä yleisen lineaarisen mallin soveltamisessa** kappaletta **Yleistetty pienimmän neliösumman menetelmä**.

Jäännöstermien korreloituneisuus on lähinnä *aikasarjojen regressiomallien* ongelma.

Aikasarjojen regressiomallit ja autokorrelaatio

Aikasarjojen regressiomalleissa kiinnitetään huomio jäännöstermien korreloituneisuuden tyyppiin, jota kutsutaan *autokorrelaatioksi*. Oletetaan siis, että havainnot ovat *aikajärjestyksessä* ja olkoon ε_i lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

jäännöstermi. Koska havainnot ovat aikajärjestyksessä, jäännöstermit ε_i muodostavat *aikasarjan*. Koska lineaarista mallia koskevan standardioletuksen (i) mukaan

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

jäännöstermien ε_i muodostaman aikasarjan ***τ*. autokovarianssi** γ_τ voidaan määritellä kaavalla

$$\gamma_\tau = \mathrm{E}(\varepsilon_i \varepsilon_{i-\tau}) \,,\; i = \tau + 1, \tau + 2, \mathsf{K}, n \,,\; \tau = 0, 1, 2, \mathsf{K}, n-1$$

Erityisesti

$$\gamma_0 = \mathrm{Var}(\varepsilon_i) = \sigma^2 \,,\; i = 1, 2, \mathsf{K}, n$$

on aikasarjan ε_i *varianssi*.

Huomautus:

- Autokovarianssit γ_τ *eivät riipu* ajanhetkestä i.

Olkoot

γ_τ = jäännöstermien ε_i *τ. autokovarianssi*

γ_0 = $\mathrm{Var}(\varepsilon_i) = \sigma^2$ = jäännöstermien ε_i *varianssi*

Jäännöstermien ε_i muodostaman aikasarjan ***τ*. auto-korrelaatiokerroin** ρ_τ määritellään kaavalla

$$\rho_\tau = \frac{\gamma_\tau}{\gamma_0} \,,\; \tau = 0, 1, 2, \mathsf{K}, n-1$$

Huomautus:

- Autokorrelaatiot ρ_τ *eivät riipu* ajanhetkestä i.

Autokorrelaatiokertoimilla ρ_τ on seuraavat ominaisuudet:

(i) $\rho_0 = 1$

(ii) $\rho_{-\tau} = \rho_\tau$

(iii) $|\rho_\tau| \leq 1$

Durbinin ja Watsonin testi 1. kertaluvun autokorrelaatiolle

Tarkastelemme seuraavassa lähemmin 1. *kertaluvun autokorrelaation* testaamista.

Asetetaan nollahypoteesi

$$\mathrm{H}_0 : \rho_1 = 0$$

jossa ρ_1 on 1. *kertaluvun autokorrelaatiokerroin*. Määritellään **Durbinin ja Watsonin testisuure** kaavalla

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

jossa e_i , $i = 1, 2, \ldots, n$ on estimoidun mallin *residuaali*. Voidaan osoittaa, että on odotettavissa, että

$$DW \approx 2$$

jos nollahypoteesi H_0 pätee. *Suuret DW*-testisuureen poikkeamat sen normaaliarvosta ≈ 2 johtavat *nollahypoteesin* H_0 *hylkäämiseen*.

Durbinin ja Watsonin testisuureella on seuraavat ominaisuudet:

(i) $0 \leq DW \leq 4$

(ii) $DW \approx 0 \quad \Leftrightarrow \quad \rho_1 \approx +1$

(iii) $DW \approx 2 \quad \Leftrightarrow \quad \rho_1 \approx 0$

(iv) $DW \approx 4 \quad \Leftrightarrow \quad \rho_1 \approx -1$

Durbinin ja Watsonin testisuureen jakauma *ei ole* mitään *tavanomaista tyyppiä*, mutta *DW*-testisuureen kriittisiä arvoja on taulukoitu ja useat tilastolliset ohjelmistot tulostavat *DW*-testisuureen kriittisiä arvoja ja/tai *DW*-testisuureen arvoja vastaavia *p*-arvoja.

Huomautuksia:

- Durbinin ja Watsonin testi *on autokorrelaatiotestinä varsin rajoittunut*, koska testi kiinnittää huomiota vain 1. kertaluvun autokorrelaatioon. Korkeamman kertaluvun autokorrelaatiolle käyttökelpoisia testejä ovat esimerkiksi **Boxin ja Piercen testi** ja eräs **Lagrangen kertojatesti**; ks. kirjaa **Aikasarja-analyysi**.
- Vaikka nollahypoteesi H_0 kiinnittää huomiota vain jäännöstermien 1. kertaluvun autokorrelaatioon, *Durbinin ja Watsonin testillä on kuitenkin keskeinen rooli regressiodiagnostiikassa*.

Tämä johtuu siitä, että *testillä on voimaa monia erilaisia mallin spesifioinnissa tapahtuneita virheitä vastaan*. Siten **Durbinin ja Watsonin testisuureen merkitsevä arvo *ei saa* automaattisesti johtaa toimenpiteisiin**, **joilla *pyritään korjaamaan* jäännöstermin *autokorreloituneisuus***.

18.9. Normaalisuus

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset toteuttava **yleinen lineaarinen malli**. Standardioletuksen (vi) mukaan mallin jäännöstermit ε_i ovat *normaalisia*:

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

Tätä oletusta kutsutaan **normaalisuusoletukseksi**. Jos oletus (vi) ei päde, jäännöstermit ovat **epänormaalisia**.

Epänormaalisuuden vaikutukset

Jos regressiomallin jäännöstermit ε_i eivät ole normaalisia, t- ja F-jakaumiin perustuva tilastollinen päättely ei välttämättä ole enää pätevää. Tämä johtuu siitä, että regressiokertoimien PNS-estimaattoreiden otosjakaumat eivät tällöin ole (ainakaan eksaktisti) normaalisia.

Huomautuksia:

- Vaikka jäännöstermit ε_i eivät olisikaan normaalisia, *t*- ja *F*-jakaumiin perustuva tilastollinen päättely *on kuitenkin yleensä suuntaa-antavaa*, jos poikkeamat normaalisuudesta ovat kohtuullisia.
- Vaikka jäännöstermit ε_i eivät olisikaan normaalisia, *t*- ja *F*-jakaumiin perustuva tilastollisen päättelyn käyttöä voidaan usein perustella *asymptoottisilla argumenteilla* so. argumenteilla, joissa vedotaan *suurten otosten teoriaan*.

Regressiomallien jäännös- eli virhetermien normaalisuutta voidaan tutkia usealla eri tavalla. Monet tilastolliset ohjelmistot tarjoavat esimerkiksi toisen tai molemmat seuraavista testeistä:

- **Bowmanin ja Shentonin testi**
- **Rankit Plot -kuvio** sekä **Wilkin ja Shapiron testi**

Käsittelemme tässä vain Bowmanin ja Shentonin testiä. Rankit Plot –kuvion ja Wilkin ja Shapiron testin käyttöä normaalisuuden testaamisessa käsitellään luvussa **Yhteensopivuuden, homogeenisuuden ja riippumattomuuden testaaminen**.

Bowmanin ja Shentonin testi

Olkoon

$$m_r = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^r$$

residuaalien e_i *r. keskusmomentti*, jossa

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

on residuaalien e_i aritmeettinen keskiarvo (= 0, jos regressiomalli on lineaarinen ja mallissa on vakio). Määritellään residuaalien *vinous* kaavalla

$$c_1 = \frac{m_3}{m_2^{3/2}}$$

ja residuaalien *huipukkuus* kaavalla

$$c_2 = \frac{m_4}{m_2^2} - 3$$

Bowmanin ja Shentonin testi jäännöstermin normaalisuudelle perustuu χ^2-testisuureeseen

$$\chi^2 = \frac{n}{6}c_1^2 + \frac{n}{24}c_2^2$$

Jos nollahypoteesi jäännöstermin normaalisuudesta pätee, testisuure χ^2 noudattaa suurissa otoksissa approksimatiivisesti χ^2-jakaumaa 2:lla vapausasteella:

$$\chi^2 \sim_a \chi^2(2)$$

Testisuureen χ^2 *normaaliarvo* eli *odotusarvo nollahypoteesin* H_0 *pätiessä* on (approksimatiivisesti)

$$E(\chi^2) = 2$$

Normaaliarvoaan *merkitsevästi suuremmat* χ^2-testisuureen arvot viittaavat siihen, että nollahypoteesi H_0 *ei päde*.

Huomautus:

- Bowmanin ja Shetonin normaalisuustesti saattaa reagoida myös regressiomallin rakenneosan *väärään spesifikaatioon*.
- Siten **Bowmanin ja Shentonin testisuureen merkitsevä arvo *ei saa* automaattisesti johtaa toimenpiteisiin**, **joilla *pyritään korjaamaan* jäännöstermin *normaalisuus***.

Normaalisuustestit saattavat reagoida myös regressio-mallin rakenneosan *väärään spesifikaatioon*.

Siten normaalisuustestien testisuureiden merkitsevät arvot *eivät saa* automaattisesti johtaa toimenpiteisiin, joilla *pyritään korjaamaan* jäännöstermin *epänormaalisuus*.

18.10. Mallin ennustuskyky

Jossakin mielessä voimakkain mahdollinen testi minkä tahansa todellisuutta kuvaavan mallille on mallin kyky *ennustaa*.

Olkoon

(1) $$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

yleinen lineaarinen mallissa, jossa

y_i = *selitettävän muuttujan y* havaittu arvo havainnossa *i*

x_{ij} = *selittäjän* x_j havaittu arvo havainnossa *i*

Oletetaan, että sekä selitettävästä muuttujasta y että selittäjistä $x_1, x_2, \ldots, x_k$ on käytettävissä havaintoarvot havainnoista

$$i = 1, 2, \ldots, n, n+1, n+2, \ldots, n+h \quad (n+h) \text{ kpl}$$

Estimoidaan mallin (1) parametrit havainnoista

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

Käytetään havainnoista $i = 1, 2, \ldots, n$ estimoitua mallia selitettävän muuttujan y arvojen y_i *ennustamiseen* havainnoissa

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

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

lineaarinen malli havainnoille $i = 1, 2, \ldots, n$ matriisimuodossa. Mallissa $\mathbf{X}$ on $n \times (k+1)$-matriisi, jolle

$$r(\mathbf{X}) = k + 1$$

ja

$$\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Olkoon

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

vektorin $\boldsymbol{\beta}$ PNS-estimaattori havainnoista

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

Muodostetaan selittäjien $x_1, x_2, \ldots, x_k$ havaituista arvoista x_{ij} havainnoissa $i = n+1, n+2, \ldots, n+h$ vektori

$$\mathbf{z}_i = (1, x_{i1}, x_{i2}, \ldots, x_{ik}), \; i = n+1, n+2, \ldots, n+h$$

Muodostetaan vektoreista $\mathbf{z}_i$ $h \times (k+1)$-matriisi $\mathbf{X}_h$, jossa vektorit $\mathbf{z}_i$ ovat *riveinä*. Tällöin selitettävän muuttujan y havaittujen arvojen y_i havaintoihin

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

perustuvat **ennusteet** havainnoissa

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

saadaan kaavasta

$$\hat{y}_i = \mathbf{z}_i'\mathbf{b}, \; i = n+1, n+2, \mathrm{K}, n+h$$

ja vastaavat **ennustevirheet** saadaan kaavasta

$$u_i = y_i - \hat{y}_i, \; i = n+1, n+2, \mathrm{K}, n+h$$

Muodostetaan ennustevirheistä u_i h-vektori

$$\mathbf{u} = (u_{n+1}, u_{n+2}, \ldots, u_{n+h})$$

Ennustevirheillä

$$u_i = y_i - \hat{y}_i = \hat{y}_i - \mathbf{z}_i'\mathbf{b}, \; i = n+1, n+2, \mathrm{K}, n+h$$

on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(u_i) = 0$$
$$\mathrm{Var}(u_i) = \sigma^2(1 + \mathbf{z}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{z}_i)$$

Vastaavasti ennustevirheiden muodostamalla *h*-vektorilla

$$\mathbf{u} = (u_{n+1}, u_{n+2}, \ldots, u_{n+h})$$

on seuraavat stokastiset ominaisuudet:

$$\mathrm{E}(\mathbf{u}) = \mathbf{0}$$
$$\mathrm{Cov}(\mathbf{u}) = \sigma^2(\mathbf{I} + \mathbf{X}_h(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h')$$

Muodostetaan **χ^2-testisuure**

$$\chi^2 = \sum_{i=n+1}^{n+h} \frac{u_i^2}{s^2}$$

jossa u_i on ennustevirhe havainnossa i ja s^2 on tavanomainen havainnoista $j = 1, 2, \ldots, n$ määrätty *harhaton estimaattori jäännösvarianssille* σ^2 . Estimoitu malli *ennustaa huonosti*, jos testisuure χ^2 saa *suuria* arvoja.

Asetetaan regressiomallin (1) parametrien samuutta otos- ja ennustejaksoilla koskeva nolla-hypoteesi

$$\mathrm{H}_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \, , \sigma_1^2 = \sigma_2^2$$

jossa indeksi 1 viittaa ajanjaksoon

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

ja indeksi 2 viittaa ajanjaksoon

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

Jos nollahypoteesi H_0 *pätee*, edellä määritelty χ^2-testisuure noudattaa χ^2*-jakaumaa* vapausastein h:

$$\chi^2 \quad \chi^2(h)$$

Suuret testisuureen arvot viittaavat siihen, että oletus parametrien vakioisuudesta ei päde.

Regressiomallin ennustekykyä voidaan testata myös parametrien vakioisuutta testaavalla *Chow-testillä*. Itse asiassa tässä esitetty χ^2*-testisuure* ja *Chow-testi* ovat läheistä sukua toisilleen ja antavat asymptoottisesti eli suurilla havaintojen lukumäärillä saman tuloksen.

19. Erityiskysymyksiä yleisen lineaarisen mallin soveltamisessa

Tarkastellemme tässä luvussa seuraavia *yleisen lineaarisen mallin* soveltamisen erityiskysymyksiä:

- Jos yleisen lineaarisen mallin standardioletuksiin kuuluvat **homoskedastisuus**- tai **korreloimattomuusoletukset** *eivät päde*, niin regressiokertoimien PNS-estimaattorit *eivät ole optimaalisia*. Optimaaliset estimaattorit voidaan tällöin tuottaa **yleistetyllä pienimmän neliösumman menetelmällä**.
- Jos yleisen lineaarisen mallin regressiokertoimia sitoo **lineaarinen side-ehto** tai **rajoitus**, niin regressiokertoimien PNS-estimaattorit *eivät ole optimaalisia*. Optimaaliset estimaattorit voidaan tällöin tuottaa **rajoitetulla pienimmän neliösumman menetelmällä**.
- Jos yleisen lineaarisen mallin selittäjät ovat **satunnaismuuttujia**, **jotka korreloivat jäännöstermien kanssa**, niin regressiokertoimien PNS-estimaattorit *eivät ole harhattomia eikä edes tarkentuvia*. Tällaisessa tilanteessa PNS-menetelmää **ei saa käyttää** regressiokertoimien estimointiin. Jos selittäjien korvaamaan voidaan löytää ns. **instrumentti** eli **keinomuuttujat**, voidaan regressiokertoimille tuottaa *tarkentuvat* estimaattorit **instrumenttimuuttujamenetelmällä**.

Avainsanat:

Ehdollinen odotusarvo, Ei-satunnaisuus, Estimaattori, Estimointi, *F*-testi, Gaussin ja Markovin lause, Harha, Harhattomuus, Heteroskedastisuus, Homoskedastisuus, Instrumenttimuuttuja, Instrumentti-muuttujamenetelmä, Jäännösneliösumma, Jäännöstermi, Jäännösvarianssi, Keskihajonta, Kokonaisneliösumma, Korrelaatio, Kovarianssi, Lagrangen menetelmä, Lineaarinen regressiomalli, Lineaarisuus, Malli, Mallineliösumma, Minimointi, Modifioidut standardioletukset, Neliösumma, Normaalisuusoletus, Odotusarvo, Otos, Otosjakauma, Otostunnusluku, Painotettu pienimmän neliösumman menetelmä, Parametri, Pienimmän neliösumman estimaattori, Pienimmän neliösumman menetelmä, Rajoitettu pienimmän neliösumman estimaattori, Rajoitettu pienimmän neliösumman menetelmä, Rajoitus, Rakenneosa, Regressioanalyysi, Regressiofunktio, Regressiokerroin, Regressio- malli, Residuaali, Satunnaisuus, Satunnainen osa, Selittäjä, Selitettävä muuttuja, Selittävä muuttuja, Selitysaste, Side-ehto, Sovite, Standardioletus, Systemaattinen osa, Tarkentuvuus, Tehokkuus, Testi, Usean selittäjän lineaarinen regressiomalli, Vakioselittäjä, Varianssianalyysihajotelma, Virhetermi, Usean selittäjän lineaarinen regressiomalli, Yleinen lineaarinen malli, Yleistetty pienimmän neliö- summan estimaattori, Yleistetty pienimmän neliösumman menetelmä

19.1. Erityiskysymyksiä yleisen lineaarisen mallin soveltamisessa: Johdanto

Yleinen lineaarinen malli

Oletetaan, että muuttujien y ja $x_1, x_2, \ldots, x_k$ havaittujen arvojen välillä vallitsee *lineaarinen tilastollinen riippuvuus*, joka voidaan ilmaista yhtälöllä

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\; i = 1, 2, \mathsf{K}, n$$

jossa

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

x_{ij} = **selittävän muuttujan** eli **selittäjän** x_j *ei-satunnainen* ja *havaittu* arvo havaintoyksikössä i, $j = 1, 2, \ldots, k$

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

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

β_j = **selittäjän** x_j **regressiokerroin**, $j = 1, 2, \ldots, k$; β_j on *ei-satunnainen* ja *tuntematon vakio*

Tällöin yhtälö määrittelee **usean selittäjän lineaarisen regressiomallin**, jota kutsutaan **yleiseksi lineaariseksi malliksi**.

Seuraavassa kertaamme yleisen lineaarisen mallin *formuloinnin matriisein*, mallia koskevat *standardioletukset* ja pääkohdat mallin parametrien *estimoinnista*; lisätietoja: ks. lukua **Yleinen lineaarinen malli**.

Yleinen lineaarinen malli voidaan esittää matriisein muodossa

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa

$\mathbf{y}$ = **selitettävän muuttujan** y *havaittujen arvojen* muodostama *satunnainen* n-vektori

$\mathbf{X}$ = **selittäjien** $x_1, x_2, \ldots, x_k$ *havaittujen arvojen* ja *ykkösten* muodostama $n \times (k + 1)$-matriisi

$\boldsymbol{\beta}$ = **regressiokertoimien** muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k + 1)$-vektori

$\boldsymbol{\varepsilon}$ = **jäännöstermien** muodostama *ei-havaittu* ja *satunnainen* n-vektori

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** esitetään matriisein seuraavassa muodossa:

(i) Matriisin $\mathbf{X}$ alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*

(ii) Matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii) $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$

(iv)&(v) *Homoskedastisuus-* ja *korreloimattomuusoletus*:

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(vi) $\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia,* mallia koskevat **modifioidut standardioletukset** esitetään matriisein seuraavassa muodossa:

(i)´ Matriisin **X** alkiot ovat *satunnaismuuttujia*.

(ii)´ Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

(iv)´&(v)´ *Homoskedastisuus-* ja *korreloimattomuusoletus*:

$$\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$$

(vi)´ $\boldsymbol{\varepsilon} \mid \mathbf{X} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Regressiokertoimien PNS-estimaattorit ja niiden ominaisuudet

Yleisen lineaarisen mallin

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \mathsf{L} + \beta_k x_{ik} + \varepsilon_i \,,\ i = 1, 2, \mathsf{K}, n$$

regressiokertoimien

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$$

PNS- eli **pienimmän neliösumman estimaattorit**

$$b_0, b_1, b_2, \ldots, b_k$$

minimoivat jäännös- eli *virhetermien* ε_i *neliösumman*

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \mathsf{L} - \beta_k x_{ik})^2$$

kertoimien $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ *suhteen*.

Yleisen lineaarisen mallin $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ regressiokertoimien vektorin

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$$

PNS-estimaaattori voidaan esittää *matriisein* muodossa

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

PNS-estimaattorilla **b** on standardioletuksien (i)-(vi) pätiessä seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$$

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{b} \quad \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Gaussin ja Markovin lause

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

yleinen lineaarinen malli, joka toteuttaa standardioletukset (i)-(v). Tällöin pätee

Gaussin ja Markovin lause:

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on paras vektorin $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*.

Gaussin ja Markovin lauseen tulkinta

Regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattorin $\mathbf{b}$ *paremmuudella* tarkoitetaan Gaussin ja Markovin lauseessa seuraavaa: Jos $\mathbf{b}^*$ on mielivaltainen regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaarinen ja harhaton estimaattori*, niin tällöin

$$\mathrm{Cov}(\mathbf{b}^*) \geq \mathrm{Cov}(\mathbf{b})$$

Huomautuksia:

- Estimaattorin $\mathbf{b}^*$ *lineaarisuus*: $\mathbf{b}^*$ on muotoa

  $$\mathbf{b}^* = \mathbf{A}\mathbf{y}$$

  jossa $(k + 1)\times n$- matriisin $\mathbf{A}$ alkiot *eivät saa riippua* selitettävän muuttujan y havaituista arvoista.

- Estimaattorin $\mathbf{b}^*$ *harhattomuus*:

  $$\mathrm{E}(\mathbf{b}^*) = \boldsymbol{\beta}$$

- Merkinnällä

  $$\mathrm{Cov}(\mathbf{b}^*) \geq \mathrm{Cov}(\mathbf{b})$$

  tarkoitetaan sitä, että matriisi

  $$\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b})$$

  on positiivisesti semidefiniitti matriisi eli

  $$\mathbf{a}'(\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}))\mathbf{a} \geq 0 \text{ kaikille } \mathbf{a} \neq \mathbf{0}$$

Epäyhtälöstä

$$\mathbf{a}'(\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}))\mathbf{a} \geq 0 \text{ kaikille } \mathbf{a} \neq \mathbf{0}$$

seuraa erityisesti se, että yksittäisten regressiokertoimien PNS-estimaattoreiden b_j , $j = 0, 1, 2, \ldots, k$ varianssit ovat *pienimpiä mahdollisia lineaaristen ja harhattomien estimaattoreiden joukossa*: Jos b_j^* on mikä tahansa regressiokertoimen β_j lineaarinen ja harhaton estimaattori, niin

$$\mathrm{Var}(b_j^*) \geq \mathrm{Var}(b_j)\,,\ j = 0,1,2,\ldots,k$$

Gaussin ja Markovin lauseen mukaan tavallinen *PNS-estimaattori* **b** *on siis paras* regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*, jos standardioletukset (i)-(v) pätevät. Tämä ei tarkoita sitä, että regressiokertoimien vektorille $\boldsymbol{\beta}$ *ei voisi löytyä* (jossakin mielessä) PNS-estimaattoreita *parempia estimaattoreita*, jos *siirrytään pois lineaaristen ja harhattomien estimaattoreiden joukosta* tai *standardioletuksista* (i)-(v) *luovutaan*.

Kun PNS-estimaattori *ei ole paras*

Tarkastelemme tässä luvussa *kahta tilannetta*, joissa regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori* **b** *on kyllä harhaton, mutta* **ei ole optimaalinen**:

(a) Jos *homoskedastisuus-* ja *korreloimattomuusoletus*

(iv)&(v) $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$

ei päde, niin ns. **yleistetty PNS-estimaattori** *on parempi kuin PNS-estimaattori* **b**.

(b) Jos *lineaariset side-ehdot tai rajoitukset sitovat vektorin* $\boldsymbol{\beta}$ *alkioita*, niin ns. **rajoitettu PNS-estimaattori** *on parempi kuin PNS-estimaattori* **b**.

Kun PNS-estimaattoria *ei saa käyttää*

Tarkastelemme tässä luvussa myös erästä sellaista tilannetta, jossa yleisen lineaarisen mallin regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori* **b** *ei ole harhaton eikä edes tarkentuva, jolloin* **PNS-estimaattoria ei voi pitää kelvollisena** *estimaattorina* vektorille $\boldsymbol{\beta}$. Näin tapahtuu esim. tilanteessa, joissa *selittäjä on stokastinen* ja *korreloi jäännöstermin kanssa*. Joskus vektorille $\boldsymbol{\beta}$ voidaan tällaisessa tilanteessa kuitenkin muodostaa *tarkentuva estimaattori* ns. **instrumenttimuuttujamenetelmällä**.

19.2. Yleistetty pienimmän neliösumman menetelmä

Yleistä lineaarista mallia

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

koskevien standardioletuksien (iv) ja (v) mukaan

$$\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

Oletuksen (iv) mukaan jäännöstermit ε_i ovat **homoskedastisia** eli niillä on *sama varianssi*:

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

Oletuksen (v) mukaan jäännöstermit ε_i ovat **korreloimattomia**:

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

Korvataan oletukset (iv)&(v) ja (vi) oletuksilla

(iv)*&(v)* $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$

(vi)* $\boldsymbol{\varepsilon} \quad \text{N}(\mathbf{0}, \sigma^2\mathbf{V})$

jossa **V** on *positiivisesti definiitti* $n \times n$-matriisi. Oletuksien (iv)*&(v)* ja (vi)* mukaan jäännöstermit ε_i saavat olla sekä **heteroskedastisia** eli *erivarianssisia* että **korreloituneita**.

Huomautus:

- Koska matriisi **V** oletettiin *positiivisesti definiitiksi*, niin se on *epäsingulaarinen* ja sillä on *käänteismatriisi*.

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *yleistetty PNS-estimaattori* saadaan minimoimalla neliömuoto

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

vektorin $\boldsymbol{\beta}$ suhteen. Vektorin $\boldsymbol{\beta}$ **yleistetty PNS-estimaattoriksi** saadaan

$$\mathbf{b}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

Merkinnässä: *GLS* = Generalized Least Squares.

Perustelu:

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\ ,\ \mathbf{X}\ n\times(k+1)$$

yleinen lineaarinen malli, joka toteuttaa *tavanomaiset oletukset* paitsi, että *jäännöstermin* $\boldsymbol{\varepsilon}$ *kovarianssimatriisi* on muotoa

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

jossa **V** on *positiivisesti definiitti* $n\times n$-matriisi.

Koska matriisi **V** on oletettu *positiivisesti definiitiksi*, se on *epäsingulaarinen* ja lisäksi pätee ns. *Cholesky-hajotelma*: On olemassa *epäsingulaarinen* $n\times n$*-yläkolmiomatriisi* **U** siten, että

$$\mathbf{V} = \mathbf{U}\mathbf{U}'$$

Kerrotaan nyt regressioyhtälö

(1) $$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

vasemmalta matriisilla $\mathbf{U}^{-1}$, jolloin saamme yhtälön

(2) $$\mathbf{U}^{-1}\mathbf{y} = \mathbf{U}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{U}^{-1}\boldsymbol{\varepsilon}$$

Yhtälö (2) voidaan kirjoittaa muotoon

(3) $$\mathbf{z} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\delta}$$

jossa

$$\mathbf{z} = \mathbf{U}^{-1}\mathbf{y}$$
$$\mathbf{T} = \mathbf{U}^{-1}\mathbf{X}$$
$$\boldsymbol{\delta} = \mathbf{U}^{-1}\boldsymbol{\varepsilon}$$

Yhtälö (3) määrittelee yleisen lineaarisen mallin, jonka *jäännöstermi toteuttaa tavanomaiset oletukset*.

Todetaan ensin, että

$$\mathrm{E}(\boldsymbol{\delta}) = \mathrm{E}(\mathbf{U}^{-1}\boldsymbol{\varepsilon}) = \mathbf{U}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{U}^{-1}\mathbf{0} = \mathbf{0}$$

Lisäksi mallin (3) jäännöstermi $\boldsymbol{\delta}$ on sekä *homoskedastinen* että *korreloimaton*:

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{\delta}) &= \mathrm{E}\{[\boldsymbol{\delta} - \mathrm{E}(\boldsymbol{\delta})][\boldsymbol{\delta} - \mathrm{E}(\boldsymbol{\delta})]'\} \\ &= \mathrm{E}(\boldsymbol{\delta}\boldsymbol{\delta}')\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}[\mathbf{U}^{-1}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'(\mathbf{U}^{-1})'] \\
&= \mathbf{U}^{-1}\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')(\mathbf{U}')^{-1} \\
&= \mathbf{U}^{-1}\mathrm{Cov}(\boldsymbol{\varepsilon})(\mathbf{U}')^{-1} \\
&= \sigma^2\mathbf{U}^{-1}\mathbf{V}(\mathbf{U}')^{-1} \\
&= \sigma^2\mathbf{U}^{-1}\mathbf{U}\mathbf{U}'(\mathbf{U}')^{-1} \\
&= \sigma^2\mathbf{I}
\end{aligned}
$$

Sovelletaan tavanomaista PNS-keinoa regressioyhtälöön (3), jolloin regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattoriksi saadaan (*GLS* = Generalized Least Squares):

$$
\begin{aligned}
\mathbf{b}_{GLS} &= (\mathbf{T}\mathbf{T})^{-1}\mathbf{T}'\mathbf{z} \\
&= (\mathbf{X}'(\mathbf{U}^{-1})'\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}^{-1})'\mathbf{U}^{-1}\mathbf{y} \\
&= (\mathbf{X}'(\mathbf{U}')^{-1}\mathbf{U}^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}')^{-1}\mathbf{U}^{-1}\mathbf{y} \\
&= (\mathbf{X}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{X})^{-1}\mathbf{X}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}
\end{aligned}
$$

■

Yleistetyn PNS-estimaattorin odotusarvo ja kovarianssimatriisi

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(iii) ja *modifioidut standardioletukset*

(iv)*&(v)* $\quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$

toteuttava yleinen lineaarinen malli. Edellä todettiin, että parametrivektorin $\boldsymbol{\beta}$ *yleistetty PNS-estimaattori* on

$$\mathbf{b}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

(i) Yleistetty PNS-estimaattori $\mathbf{b}_{GLS}$ on *harhaton* parametrivektorille $\boldsymbol{\beta}$:

$$\mathrm{E}(\mathbf{b}_{GLS}) = \boldsymbol{\beta}$$

(ii) Yleistetyn PNS-estimaattorin $\mathbf{b}_{GLS}$ *kovarianssimatriisi* on

$$\mathrm{Cov}(\mathbf{b}_{GLS}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

Perustelu:

(i) Suoraan laskemalla saadaan:

$$
\begin{aligned}
\mathrm{E}(\mathbf{b}_{GLS}) &= \mathrm{E}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}] \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathrm{E}(\mathbf{y}) \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \\
&= \boldsymbol{\beta}
\end{aligned}
$$

(ii) Yleistetyn PNS-estimaattorin $\mathbf{b}_{GLS}$ kaavaa johdettaessa malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

muunnettiin malliksi

$$\mathbf{z} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\delta}$$

jossa

$$\mathbf{z} = \mathbf{U}^{-1}\mathbf{y}$$

$$\mathbf{T} = \mathbf{U}^{-1}\mathbf{X}$$

$$\boldsymbol{\delta} = \mathbf{U}^{-1}\boldsymbol{\varepsilon}$$

ja $\mathbf{U}$ on epäsingulaarinen yläkolmiomatriisi joka toteuttaa ehdon

$$\mathbf{V} = \mathbf{U}\mathbf{U}'$$

Siten

$$\mathrm{Cov}(\mathbf{b}_{GLS}) = \sigma^2(\mathbf{TT})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

■

Modifioitu Gaussin ja Markovin lause yleistetylle PNS-estimaattorille

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(iii) ja *modifioidut standardioletukset*

(iv)*&(v)* $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$

toteuttava yleinen lineaarinen malli.

Gaussin ja Markovin lause yleistetylle PNS-estimaattorille:

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *yleistetty PNS-estimaattori*

$$\mathbf{b}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

on paras (eli *tehokkain*) *vektorin* $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*.

Perustelu:

Yleistetyn PNS-estimaattorin $\mathbf{b}_{GLS}$ kaavaa johdettaessa malli

(1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

muunnettiin malliksi

(3) $\mathbf{z} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\delta}$

jossa

$$\mathbf{z} = \mathbf{U}^{-1}\mathbf{y}$$

$$\mathbf{T} = \mathbf{U}^{-1}\mathbf{X}$$

$$\boldsymbol{\delta} = \mathbf{U}^{-1}\boldsymbol{\varepsilon}$$

ja $\mathbf{U}$ on epäsingulaarinen yläkolmiomatriisi joka toteuttaa ehdon

$$\mathbf{V} = \mathbf{U}\mathbf{U}'$$

Koska malli (3) toteuttaa ns. tavanomaiset oletukset, regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori

$$\mathbf{b}_{GLS} = (\mathbf{TT})^{-1}\mathbf{T}'\mathbf{z} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

on Gaussin ja Markovin lauseen mukaan *paras* regressiokertoimien vektorin $\boldsymbol{\beta}$ lineaaristen ja harhattomien estimaattoreiden joukossa mallissa (3) ja esimerkiksi parempi kuin *tavallinen PNS-estimaattori* $\mathbf{b}$. Tämä merkitsee sitä, että matriisi

$$\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_{GLS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

on oltava *positiivisesti semidefiniitti* kaikille $\mathbf{V} > 0$.

Yleistetty PNS-estimaattori nähdään parhaaksi lineaaristen ja harhattomien estimaattoreiden joukossa myös suoraan laskemalla vetoamatta Gaussin ja Markovin lauseeseen.

Olkoon

$$\mathbf{b}^* = \mathbf{Hy}$$

jokin regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaarinen* ja *harhaton* estimaattori. Tällöin

$$\mathrm{E}(\mathbf{b}^*) = \mathrm{E}(\mathbf{Hy}) = \mathbf{H}\mathrm{E}(\mathbf{y}) = \mathbf{H}\mathrm{E}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{HX}\boldsymbol{\beta} = \boldsymbol{\beta}$$

josta seuraa, että

$$\mathbf{HX} = \mathbf{I}$$

Määritellään matriisi $\mathbf{C}$ yhtälöllä

$$\mathbf{H} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{C}$$

Koska edellä esitetyn mukaan

$$\mathbf{HX} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} + \mathbf{CX} = \mathbf{I} + \mathbf{CX} = \mathbf{I}$$

niin välttämättä

$$\mathbf{CX} = \mathbf{0}$$

Siten

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}^*) &= \mathrm{Cov}(\mathbf{Hy}) \\
&= \mathrm{E}\{[\mathbf{Hy} - \mathrm{E}(\mathbf{Hy})][\mathbf{Hy} - \mathrm{E}(\mathbf{Hy})]'\} \\
&= \mathrm{E}\{[\mathbf{Hyy}'\mathbf{H}'\} \\
&= \mathbf{H}\mathrm{E}(\mathbf{yy}')\mathbf{H}' \\
&= \sigma^2\mathbf{HVH}' \\
&= \sigma^2[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{C}]\mathbf{V}[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{C}]' \\
&= \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} + \sigma^2\mathbf{CV}^{-1}\mathbf{C}' \\
&= \mathrm{Cov}(\mathbf{b}_{GLS}) + \sigma^2\mathbf{CV}^{-1}\mathbf{C}'
\end{aligned}$$

Koska matriisi $\mathbf{CV}^{-1}\mathbf{C}'$ on *ei-negatiivisesti defiiniitti*, väite yleistetyn PNS-estimaattorin paremmuudesta on todistettu.

■

Erityisesti yleistetty PNS-estimaattori $\mathbf{b}_{GLS}$ on modifioitujen standardioletusten (i)-(iii) ja (iv)*&(v)* pätiessä *parempi* kuin tavallinen PNS-estimaattori **b**:

$$\text{Cov}(\mathbf{b}) \geq \text{Cov}(\mathbf{b}_{GLS})$$

Yleistetyn PNS-estimaattorin stokastiset ominaisuudet

Yleisen lineaarisen mallin regressiokertoimien vektorin $\boldsymbol{\beta}$ *yleistetyllä PNS-estimaattorilla* $\mathbf{b}_{GLS}$ on modifioitujen standardioletuksien (i)-(iii), (iv)*&(v)* ja (vi)* pätiessä seuraavat *stokastiset ominaisuudet*:

(1) $\mathbf{b}_{GLS}$ on *harhaton*.

(2) $\mathbf{b}_{GLS}$ *paras* (eli *tehokkain*) *lineaaristen ja harhattomien estimaattoreiden joukossa*.

(3) $\mathbf{b}_{GLS}$ on *normaalinen*.

Lisäksi voidaan osoittaa, että $\mathbf{b}_{GLS}$ on (sopivin lisäehdoin) *tarkentuva*.

Laskettava yleistetty PNS-estimaattori

Käytännössä modifioitujen standardioletuksien (iv)*&(v)* ja (vi)* jäännöstermin $\boldsymbol{\varepsilon}$ kovarianssi-matriisissa

$$\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

esiintyvä $n \times n$-matriisi **V** on *tuntematon*. Koska matriisissa **V** on

$$n(n + 1)/2$$

vapaata parametria, matriisia V ei voida sellaisenaan estimoida n:stä havainnosta.

Oletetaan siksi, että kovarianssimatriisi **V** riippuu tuntemattomista parametreista

$$\delta_1, \delta_2, \ldots, \delta_m$$

jossa

$$m < n \text{ (= havaintojen lukumäärä)}$$

Tavallisesti haluamme, että m on *selvästi pienempi* kuin n:

$$m << n$$

Oletetaan lisäksi, että parametrit $\delta_1, \delta_2, \ldots, \delta_m$ voidaan estimoida *tarkentuvasti* eli siten, että estimaattorit havaintojen lukumäärän kasvaessa lähestyvät stokastisesti parametrien oikeita arvoja.

Muodostetaan parametreista $\delta_1, \delta_2, \ldots, \delta_m$ m-vektori $\boldsymbol{\delta}$:

$$\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_m)$$

Käytetään kovarianssimatriisille **V** merkintää

$$\mathbf{V}(\boldsymbol{\delta}) = \mathbf{V}(\delta_1, \delta_2, \mathrm{K}, \delta_m)$$

korostamaan matriisin **V** riippuvuutta parametreista $\delta_1, \delta_2, \ldots, \delta_m$.

Olkoon $\hat{\mathbf{V}}(\hat{\boldsymbol{\delta}})$ matriisin $\mathbf{V}(\boldsymbol{\delta})$ estimaattori, joka saadaan sijoittamalla parametrin $\boldsymbol{\delta}$ paikalle sen *tarkentuva* estimaattori $\hat{\boldsymbol{\delta}}$. Tällöin

$$\mathbf{b}^{*}_{GLS} = (\mathbf{X}'[\hat{\mathbf{V}}(\hat{\boldsymbol{\delta}})]^{-1}\mathbf{X})^{-1}\mathbf{X}'[\hat{\mathbf{V}}(\hat{\boldsymbol{\delta}})]^{-1}\mathbf{y}$$

on regressiokertoimien vektorin $\boldsymbol{\beta}$ *tarkentuva estimaattori*, jota kutsutaan usein **laskettavaksi** (engl. *feasible*).

Matriisi $\mathbf{V}$ voidaan *spesifioida* eli *esittää parametrointi* tekemällä sopivia oletuksia mallin jäännöstermin $\boldsymbol{\varepsilon}$ *heteroskedastisuus*- tai *kovarianssirakenteesta*. Esimerkiksi tietyn tyyppisissä *aikasarjojen regressiomalleissa* matriisi $\mathbf{V}$ voidaan spesifioida jäännöstermin *autokorrelaatiorakenteen* perusteella.

Painotettu PNS-estimaattori

Oletetaan, että yleinen lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

toteuttaa modifioidut standardioletukset (i)-(iii), (iv)*&(v)* ja (vi)*, mutta jäännöstermin $\boldsymbol{\varepsilon}$ kovarianssimatriisissa

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

esiintyvä matriisi $\mathbf{V}$ on *diagonaalinen*:

$$\mathbf{V} = \mathrm{diag}(z_1^2, z_2^2, \mathrm{K}, z_n^2)$$

jossa luvut

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

ovat *tunnettuja*. Tällöin jäännöstermit ovat **korreloimattomia**, mutta ne saavat yleisessä tapauksessa olla **heteroskedastisia**.

Tällöin regressiokertoimien vektorin $\boldsymbol{\beta}$ *yleistetyn PNS-estimaattorin* kaavassa

$$\mathbf{b}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

matriisin $\mathbf{V}$ käänteismatriisi on muotoa

$$\mathbf{V}^{-1} = \mathrm{diag}(1/z_1^2, 1/z_2^2, \mathrm{K}, 1/z_n^2)$$

Tällöin estimaattoria $\mathbf{b}_{GLS}$ kutsutaan tavallisesti **painotetuksi PNS-estimaattoriksi**.

Nimitys *painotettu PNS-estimaattori* johtuu siitä, että estimaattori voidaan muodostaa *soveltamalla tavallista PNS-menetelmää* muunnettuihin havaintoarvoihin, jotka saadaan kertomalla alkuperäiset havaintoarvot

$$y_i,\ x_{i1},\ x_{i2},\ \ldots,\ x_{ik},\ i = 1, 2, \ldots, n$$

painoilla

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

19.3. Rajoitettu pienimmän neliösumman menetelmä

Oletetaan, että standardioletukset (i)-(v) toteuttavan yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa $\mathbf{X}$ on $n\times(k+1)$-matriisi, $n \geq k+1$, $r(\mathbf{X}) = k+1$, *regressiokertoimia* $\boldsymbol{\beta}$ *sitoo* **lineaarinen rajoitus** eli **side-ehto**

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

jossa $\mathbf{R}$ on $m\times(k+1)$-matriisi, $m \leq k+1$, $r(\mathbf{R}) = m$.

Huomautus:

- Jos ehto $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ pätee, regressiokertoimet $\boldsymbol{\beta}$ *eivät voi varioida vapaasti* $(k+1)$-ulotteisessa avaruudessa $^{k+1}$, vaan ainoastaan ehdon määrittelemässä m-ulotteisessa lineaarisessa aliavaruudessa (= m-ulotteisella tasolla).

Regressiokertoimien vektorin $\boldsymbol{\beta}$ rajoitettu PNS-estimaattori saadaan minimoimalla neliömuoto

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

vektorin $\boldsymbol{\beta}$ suhteen ottamalla huomioon side-ehto

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Tämä on *sidottu ääriarvotehtävä*, joka voidaan ratkaistaan *Lagrangen keinolla*. Vektorin $\boldsymbol{\beta}$ **rajoitettu PNS-estimaattori** on

$$\mathbf{b}_R = \mathbf{b} + \mathbf{U}\mathbf{R}'\mathbf{S}(\mathbf{r} - \mathbf{R}\mathbf{b})$$

jossa

$$\mathbf{U} = (\mathbf{X}'\mathbf{X})^{-1}$$
$$\mathbf{S} = (\mathbf{R}\mathbf{U}\mathbf{R}')^{-1}$$

ja

$$\mathbf{b} = \mathbf{U}\mathbf{X}'\mathbf{y}$$

on vektorin $\boldsymbol{\beta}$ tavallinen PNS-estimaattori.

Merkinnässä: R = Restricted.

Perustelu:

Oletetaan, että

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}\text{ , }\mathbf{X}\ n\times(k+1)$$

on *tavanomaiset oletukset toteuttava yleinen lineaarinen malli, jonka regressiokertoimien vektoria* $\boldsymbol{\beta}$ *sitoo lineaarinen rajoitus* tai *side-ehto*

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

jossa $\mathbf{R}$ *täysiasteinen* $m\times(k+1)$-matriisi, $m \leq k+1$.

Minimoidaan neliösumma

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

side-ehtojen

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

vallitessa. Käytetään tähän *Lagrangen keinoa*.

Minimoitava funktio on muotoa

$$f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\boldsymbol{\lambda}'(\mathbf{R}\boldsymbol{\beta} - \mathbf{r})$$
$$= \mathbf{y}'\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\boldsymbol{\lambda}'\mathbf{R}\boldsymbol{\beta} - 2\boldsymbol{\lambda}'\mathbf{r}$$

jossa $2\boldsymbol{\lambda}$ on Lagrangen kertoimien muodostama $(k + 1)$-vektori (kerroin 2 on mukana mukavuussyistä).

Derivoidaan funktio $f(\boldsymbol{\beta})$ sekä muuttujan $\boldsymbol{\beta}$ että kerroinvektorin $\boldsymbol{\lambda}$ suhteen ja merkitään derivaatat nolliksi:

(i) $$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + 2\mathbf{R}'\boldsymbol{\lambda} = \mathbf{0}$$

(ii) $$\frac{\partial f(\boldsymbol{\beta})}{\partial \boldsymbol{\lambda}} = 2\mathbf{R}\boldsymbol{\beta} - 2\mathbf{r} = \mathbf{0}$$

Yhtälöt (i) ja (ii) muodostavat yhtälösysteemin, jossa tuntemattomina ovat vektorit $\boldsymbol{\beta}$ ja $\boldsymbol{\lambda}$. Kerroinvektori $\boldsymbol{\lambda}$ voidaan eliminoida yhtälöistä, jonka jälkeen $\boldsymbol{\beta}$ saadaan ratkaistuksi.

Kerrotaan yhtälö (i) vasemmalta matriisilla $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$ (ja luvulla $-1/2$), jolloin saadaan:

$$\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - \mathbf{R}\boldsymbol{\beta} = \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\boldsymbol{\lambda}$$

Koska matriisi $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ on *täysiasteinen* $m \times m$-matriisi, vektori $\boldsymbol{\lambda}$ voidaan ratkaista tästä yhtälöstä. Ottamalla samalla huomioon yhtälö (ii), saadaan ratkaisuksi

$$\boldsymbol{\lambda} = (\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{R}\boldsymbol{\beta}) = (\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})$$

missä

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

tavanomainen PNS-estimaattori vektorille $\boldsymbol{\beta}$.

Sijoitetaan saatu vektorin $\boldsymbol{\lambda}$ lauseke yhtälöön (i), jolloin saadaan yhtälö

$$-\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r}) = \mathbf{0}$$

Ratkaisemalla $\boldsymbol{\beta}$ tästä yhtälöstä saadaan regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitettu* tai *sidottu PNS-estimaattori*:

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})$$

■

Rajoitetun PNS-estimaattorin odotusarvo ja kovarianssimatriisi

Oletetaan, että standardioletukset (i)-(v) toteuttavan yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimia $\boldsymbol{\beta}$ sitoo lineaarinen rajoitus eli side-ehto

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Edellä todettiin, että parametrivektorin $\boldsymbol{\beta}$ *rajoitettu PNS-estimaattori* on

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})$$

(i) Rajoitettu PNS-estimaattori $\mathbf{b}_R$ on *harhaton* parametrivektorille $\boldsymbol{\beta}$:

$$E(\mathbf{b}_R) = \boldsymbol{\beta}$$

(ii) Rajoitettu PNS-estimaattorin $\mathbf{b}_{GLS}$ *kovarianssimatriisi* on

$$\text{Cov}(\mathbf{b}_R) = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]$$

Perustelu:

(i) Suoraan laskemalla saadaan:

$$\begin{aligned} E(\mathbf{b}_R) &= E[\mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})] \\ &= E(\mathbf{b}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}E(\mathbf{b}) - \mathbf{r}) \\ &= \boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \\ &= \boldsymbol{\beta} \end{aligned}$$

(ii) Oletetaan, että rajoitukset $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ pätevät. Merkitsemällä

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

voidaan rajoitetun PNS-estimaattorin $\mathbf{b}_R$ lauseke kirjoittaa muotoon

$$\mathbf{b}_R = \mathbf{b} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})$$

Koska

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

saadaan yhtälö

$$\mathbf{b}_R - \boldsymbol{\beta} = [(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}']\mathbf{X}'\boldsymbol{\varepsilon}$$

Koska oletimme, että $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, jolloin $\mathbf{b}_R$ on *harhaton* parametrivektorille $\boldsymbol{\beta}$, niin

$$\begin{aligned} \text{Cov}(\mathbf{b}_R) &= E\{[(\mathbf{b}_R - E(\mathbf{b}_R)][(\mathbf{b}_R - \boldsymbol{\beta})]'\} \\ &= E[(\mathbf{b}_R - \boldsymbol{\beta})(\mathbf{b}_R - \boldsymbol{\beta})'] \\ &= [(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}']\mathbf{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}']\mathbf{X}'\mathbf{X}[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}] \end{aligned}$$

■

Modifioitu Gaussin ja Markovin lause rajoitetulle PNS-estimaattorille

Oletetaan, että standardioletukset (i)-(v) toteuttavan yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimia $\boldsymbol{\beta}$ sitoo lineaarinen rajoitus eli side-ehto

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

Edellä todettiin, että parametrivektorin $\boldsymbol{\beta}$ *rajoitettu PNS-estimaattori* on

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathbf{b} - \mathbf{r})$$

Gaussin ja Markovin lause rajoitetulle PNS-estimaattorille:

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitettu PNS-estimaattori*

$$\mathbf{b}_R = \mathbf{b} + \mathbf{UR}'\mathbf{S}(\mathbf{r} - \mathbf{Rb})$$

on paras (eli *tehokkain*) *sellaisten vektorin* $\boldsymbol{\beta}$ lineaaristen ja harhattomien estimaattoreiden joukossa, *joille side-ehto*

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

pätee.

Erityisesti rajoitettu PNS-estimaattori $\mathbf{b}_R$ on standardioletuksien (i)-(v) ja side-ehdon $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ pätiessä *parempi* kuin tavallinen PNS-estimaattori $\mathbf{b}$:

$$\text{Cov}(\mathbf{b}) \geq \text{Cov}(\mathbf{b}_R)$$

Perustelu:

Olemme aikaisemmin todenneet, että

$$\text{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

ja olemme johtaneet edellä tuloksen

$$\text{Cov}(\mathbf{b}_R) = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\,\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]$$

Siten samme suoraan laskemalla:

$$\text{Cov}(\mathbf{b}) - \text{Cov}(\mathbf{b}_R) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\,\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

mikä on selvästi *ei-negatiivisesti definiitti* matriisi.

■

Rajoitetun PNS-estimaattorin stokastiset ominaisuudet

Yleisen lineaarisen mallin regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitetulla PNS-estimaattorilla* $\mathbf{b}_R$ on standardioletuksien (i)-(vi) ja side-ehdon $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ pätiessä seuraavat *stokastiset ominaisuudet*:

(1) $\mathbf{b}_R$ on *harhaton.*

(2) $\mathbf{b}_R$ paras (eli tehokkain) lineaaristen ja harhattomien estimaattoreiden joukossa.

(3) $\mathbf{b}_R$ on *normaalinen.*

Lisäksi voidaan osoittaa, että $\mathbf{b}_R$ on (sopivin lisäehdoin) *tarkentuva*.

Rajoitusten testaus

Asetetaan nollahypoteesi

$$\text{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

ja määritellään ***F*-testisuure**

$$F = \frac{(\mathbf{r} - \mathbf{Rb})'\mathbf{S}(\mathbf{r} - \mathbf{Rb})}{ms^2}$$

jossa

$$\mathbf{b} = \mathbf{U}\mathbf{X}'\mathbf{y}$$

on vektorin $\boldsymbol{\beta}$ tavallinen PNS-estimaattori ja

$$s^2 = \frac{1}{n-k-1}(\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$$

vastaava jäännösvarianssin σ^2 harhaton estimaattori. Jos standardioletukset (i)-(vi) ja nollahypoteesi H_0 pätevät, testisuure F noudattaa *F-jakaumaa* vapausastein m ja $n - k - 1$:

$$F \quad F(m, n-k-1)$$

Suuret testisuureen F arvot merkitsevät sitä, että *nollahypoteesi* H_0 *on hylättävä*.

Edellä esitetty F-testisuure nollahypoteesille

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

voidaan kirjoittaa myös muotoon

$$F = \frac{n-k-1}{m} \cdot \frac{SSE_R - SSE}{SSE}$$

jossa

$$SSE = (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})$$

on jäännösneliösumma tavallisesta PNS-estimaattorista **b** ja

$$SSE_R = (\mathbf{y}-\mathbf{X}\mathbf{b}_R)'(\mathbf{y}-\mathbf{X}\mathbf{b}_R)$$

on jäännösneliösumma rajoitetusta PNS-estimaattorista $\mathbf{b}_R$.

Yleisen lineaarisen mallin soveltamisen yhteydessä on tapana tarkastella nollahypoteesien

$$H_{0T} : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

ja

$$H_{0j} : \beta_j = 0 \text{ , } j = 1, 2, \ldots, k$$

testaamista; ks. lukua **Yleinen lineaarinen malli**.

Nollahypoteesiin H_{0T} kohdistettu testi on *yleistesti regression olemassaololle*. Jos nollahypoteesi H_{0T} jää voimaan, selitettävä muuttujan y havaitut arvot *eivät riipu lineaarisesti* yhdenkään selittäjän x_1, x_2, … , x_k havaituista arvoista.

Nollahypoteesiin H_{0j} kohdistetulla testillä testataan selittäjän x_j , j = 1, 2, … , k vaikutusta selitettävään muuttujaan y. Jos nollahypoteesi H_{0j} jää voimaan, selitettävä muuttujan y havaitut arvot *eivät riipu lineaarisesti* selittäjän x_j , j = 1, 2, … , k havaituista arvoista.

Nollahypoteesien H_{0T} ja H_{0j} , j = 1, 2, … , k asettaminen merkitsee *lineaaristen side-ehtojen esittämistä* regressio-kertoimille β_1, β_2, … , β_k.. Siten yleisen lineaarisen mallin soveltamisen yhteydessä nollahypoteeseille H_{0T} ja H_{0j} , j = 1, 2, … , k esitetyt testit ovat *erikoistapauksia* tässä tarkastellusta yleisestä testistä lineaarisille side-ehdoille.

Rajoitusten spesifiointi

Regressiokertoimia koskevat rajoitukset seuraavat tavallisesti tutkimuksen kohteena olevaan satunnaisilmiöön liittyvästä *taustateoriasta* (kuten taloustieteestä tai ekonometriasta), mutta myös monet tilastolliset hypoteesit voidaan esittää regressiokertoimia koskevien side-ehtojen muodossa. Jos side-ehdot vastaavat jotakin *taustateorian hypoteesia*, pyritään esitetty hypoteesi *vahvistamaan* tai *kumoamaan side-ehtojen testaamisella*.

19.4. Instrumenttimuuttujamenetelmä

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

yleisen lineaarisen mallin matriisiesitys, jossa

- $\mathbf{y}$ = **selitettävän muuttujan** *y havaittujen arvojen* muodostama *satunnainen* n-vektori
- $\mathbf{X}$ = **selittäjien** $x_1, x_2, \ldots, x_k$ *havaittujen arvojen* ja *ykkösten* muodostama $n\times(k+1)$-matriisi
- $\boldsymbol{\beta}$ = **regressiokertoimien** muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k+1)$-vektori
- $\boldsymbol{\varepsilon}$ = **jäännöstermien** muodostama *ei-havaittu* ja *satunnainen* n-vektori

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *kiinteitä* eli *ei-satunnaisia* muuttujia, mallia koskevat **standardioletukset** voidaan esittää matriisein seuraavassa muodossa:

(i) Matriisin $\mathbf{X}$ alkiot ovat *kiinteitä* eli *ei-satunnaisia vakioita*

(ii) Matriisi $\mathbf{X}$ on *täysiasteinen*:

$$r(\mathbf{X}) = k+1$$

(iii) $$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

(iv)&(v) $$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(vi) $$\boldsymbol{\varepsilon} \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Yleisen lineaarisen mallin *selittäjien satunnaisuus saattaa aiheuttaa vakavia ongelmia* mallin estimoinnille ja mallia koskevalle tilastolliselle päättelylle. **Jos matriisi X on satunnainen**, **PNS-menetelmä *ei välttämättä tuota harhattomia* tai *edes tarkentuvia estimaattoreita regressiokertoimille***. Näin käy esimerkiksi silloin, kun virhetermi ja selittäjät *korreloivat*. **Jos regressiokertoimien PNS-estimaattorit *eivät ole harhattomia* tai *tarkentuvia*, mallia koskevaa *tavanomaista tilastollista päättelyä ei voida soveltaa***.

Regressiokertoimien vektorin PNS-estimaattorin harhattomuus

Olkoon

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *PNS-estimaattori*. PNS-estimaattorin **b** lauseke voidaan kirjoittaa muotoon

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

Jos matriisi **X** on *kiinteä*, estimaattori **b** on *harhaton*, koska standardioletuksen (iii) mukaan $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, jolloin

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon}) = \boldsymbol{\beta}$$

Jos matriisi **X** on satunnainen, ei saa kirjoittaa

$$\mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon})$$

Sen sijaan PNS-estimaattorin **b** *ehdollisessa odotusarvossa* matriisin **X** suhteen matriisia **X** voidaan pitää "*kiinteänä*" ja siten

$$\mathrm{E}(\mathbf{b} \mid \mathbf{X}) = \boldsymbol{\beta} + \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \mid \mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X})$$

PNS-estimaattori **b** on siis **ehdollisesti harhaton** eli

$$\mathrm{E}(\mathbf{b} \mid \mathbf{X}) = \boldsymbol{\beta}$$

jos oletus

$$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

pätee. Tällöin PNS-estimaattori **b** on myös (*ehdottomasti*) **harhaton**, koska *iteroidun odotusarvon lain* mukaan

$$\mathrm{E}(\mathbf{b}) = \mathrm{E}(\mathrm{E}(\mathbf{b} \mid \mathbf{X})) = \mathrm{E}(\boldsymbol{\beta}) = \boldsymbol{\beta}$$

Edellä esitetystä nähdään, että *ehdon*

$$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$$

voimassaolo ratkaisee sen, onko PNS-estimaattori **b** *harhaton* lineaarisen mallin regressiokertoimien vektorille $\boldsymbol{\beta}$. Voidaan osoittaa, että vastaava korjaus muihin yleisen lineaarisen mallin standardioletuksiin (iii)-(vi) pelastaa kiinteiden selittäjien tapauksessa esitetyn teorian.

Jos yleisen lineaarisen mallin selittäjät $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*, mallia koskevat **standardioletukset** voidaan esittää matriisein seuraavassa muodossa:

(i)´ Matriisin **X** alkiot ovat (vakioselittäjän arvoja lukuun ottamatta) *satunnaismuuttujia*

(ii)´ Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

(iv)´&(v)´ $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$

(vi)´ $(\boldsymbol{\varepsilon} \mid \mathbf{X}) \quad \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$

Modifioidusta standardioletuksesta

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

seuraa, että

$$\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

Tämä merkitsee sitä, että selitettävän muuttujan **y** ehdollinen odotusarvo eli regressiofunktio selittäjien havaittujen arvojen suhteen on lineaarinen.

Modifioidusta standardioletuksesta

(iii)´ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

seuraa, että

$$\mathrm{E}(\mathbf{z}_i\boldsymbol{\varepsilon}_i) = \mathbf{0}\,,\ i = 1, 2, \mathrm{K}\,, n$$

jossa

$$\mathbf{z}_i = (1, x_{i1}, x_{i2}, \mathrm{K}\,, x_{ik})\,,\ i = 1, 2, \mathrm{K}\,, n$$

Siten oletuksesta (iii)´ seuraa, että selittäjien arvot ja jäännös- eli virhetermit ovat *korreloimattomia*.

Jos modifioidut standardioletukset (i)´-(vi)´ pätevät, tavanomainen ei-satunnaisille selittäjille esitetty estimointi- ja päättelytekniikka pätee.

Myös edellä esitetyt modifioidut ehdot jäännöseli virhetermeille ovat melko rajoittavia ja etenkin aikasarjojen regressiomallien soveltamisen yhteydessä kohdataan tilanteita, joissa eivät edes nämä modifioidut ehdot päde. Tällaisissa tilanteissa PNS-menetelmää ei saa käyttää mallin parametrien estimointiin.

Tilastotiede tuntee kuitenkin menetelmiä, joilla regressiomallin parametrit *voidaan estimoida* (*ainakin*) *tarkentuvasti* myös monissa niissä tilanteissa, joissa edellä esitetyt modifioidut ehdot jäännöstermeille *eivät päde*. Eräs näistä menetelmistä on *instrumenttimuuttujamenetelmä*.

Instrumenttimuuttujamenetelmä

Tarkastellaan nyt tilannetta, jossa yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

selittäjät $x_1, x_2, \ldots, x_k$ ovat satunnaismuuttujia ja lisäksi korreloivat mallin jäännöstermin ε kanssa. Tällöin

$$\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) \neq \mathbf{0}$$

jolloin regressiokertoimien vektorin $\boldsymbol{\beta}$ PNS-estimaattori **b** on sekä **harhainen** että **ei-tarkentuva**.

Olkoon

$$w_{ij}\,,\ i = 1, 2, \ldots, n\,,\ j = 1, 2, \ldots, k$$

muuttujan w_j havaittu arvo havaintoyksikössä i ja muodostetaan havaintoarvoista w_{ij} $n\times(k + 1)$-matriisi **W**, jossa ensimmäinen sarake on ykkösten muodostama. Oletetaan, että matriisilla **W** on seuraavat ominaisuudet:

(i) $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{W}) = \mathbf{0}$

(ii) $r(\mathbf{W}'\mathbf{X}) = k + 1$

Tällöin sanomme, että muuttujat

$$w_1, w_2, \ldots, w_k$$

kelpaavat **instrumenteiksi** selittäjille

$x_1, x_2, \ldots, x_k$

Muuttujia $w_1, w_2, \ldots, w_k$ kutsutaan tavallisesti **instrumentti-** tai **välinemuuttujiksi**.

Huomautuksia:

- Ehdon (i) mukaan instrumenttimuuttujat $w_1, w_2, \ldots, w_k$ *eivät saa korreloida* mallin jäännöstermin ε kanssa.
- Ehdon (ii) mukaan instrumenttimuuttujien $w_1, w_2, \ldots, w_k$ *pitää korreloida* selittäjien $x_1, x_2, \ldots, x_k$ kanssa niin voimakkaasti, että matriisi $\mathbf{W'X}$ on *epäsingulaarinen*.

Määritellään yleisen lineaarisen mallin regressiokertoimien vektorille **β instrumenttimuuttuja-estimaattori** kaavalla

$$\mathbf{b}_{IV} = (\mathbf{W'X})^{-1}\mathbf{W'y}$$

Merkinnässä: *IV* = Instrumental Variable.

Voidaan osoittaa, että instrumenttimuuttujaestimaattori $\mathbf{b}_{IV}$ on sopivin, matriisien

$$\mathbf{W'W}\ ,\ \mathbf{W'X}\ ,\ \mathbf{W'\varepsilon}$$

asymptoottista käyttäytymistä koskevin lisäehdoin regressiokertoimien vektorin **β** *tarkentuva* estimaattori.

Voidaan osoittaa, että instrumenttimuuttujaestimaattorin $\mathbf{b}_{IV}$ **kovarianssimatriisi** on *suurissa otoksissa* approksimatiivisesti muotoa

$$\mathrm{Cov}(\mathbf{b}_{IV}) =_a \hat{\sigma}^2[\mathbf{X'W}(\mathbf{W'W})^{-1}\mathbf{W'X}]^{-1}$$

jossa

$$\hat{\sigma}^2 = \frac{1}{n-k-1}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

ja

$$\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\mathbf{b}_{IV}$$

Instrumenttimuuttujaestimaattoria koskeva *tilastollisessa päättelyssä* (esim. regressiokertoimien *luottamusväleissä* ja kertoimia koskevat *testeissä*) nojataan tähän tulokseen.

Instrumenttien spesifiointi

Sopivat instrumentit löydetään tavallisesti tutkimuksen kohteena olevaan satunnaisilmiöön liittyvästä *taustateoriasta* (kuten taloustieteestä tai ekonometriasta).