Tilastollinen päättely

8. Yleinen lineaarinen malli

8.1. Yleinen lineaarinen malli ja sen parametrien estimointi

Estimointi, Estimaattorin hyvyys, Gaussin ja Markovin lause, Harhattomuus, Homoskedastisuus, Jakaumaoletus, Jäännösneliösumma, Jäännöstermi, Jäännösvarianssi, Kokonaisneliösumma, Korrelaatio, Korreloimattomuus, Kovarianssi, Lineaarinen regressiomalli, Mallineliösumma, Normaalijakauma, Optimaalisuus, Otos, Pienimmän neliösumman menetelmä, Regressiofunktio, Regressiokerroin, Regressiomalli, Regressiotaso, Residuaali, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Selitysaste, Sovite, Suurimman uskottavuuden menetelmä, Tehokkuus, Tyhjentävyys, Vakiotermi, Varianssianalyysihajotelma, Yleinen lineaarinen malli

8.2. Päättely yleisestä lineaarisesta mallista

F-jakauma, *F*-testi, Harhattomuus, Jäännösvarianssi, Kriittinen arvo, Lagrangen kertojatesti, Luottamuskerroin, Luottamustaso, Luottamusväli, Merkitsevyystaso, Nollahypoteesi, Normaalijakauma, Osamäärätesti, Parametri, Regressiokerroin, Residuaali, Selitettävä muuttuja, Selittäjä, Selittävä muuttuja, Selitysaste, Sovite, *t*-jakauma, *t*-testi, Testi, Testisuure, Vakiotermi, Vapausasteet, Waldin testi, Yleinen lineaarinen malli

8.3. Yleinen lineaarinen malli ja yleistetty pienimmän neliösumman menetelmä

Kovarianssimatriisi, Estimaattorin hyvyys, Gaussin ja Markovin lause, Pienimmän neliösumman menetelmä, Tehokkuus, Yleinen lineaarinen malli, Yleistetty pienimmän neliösumman estimaattori, Yleistetty pienimmän neliösumman menetelmä

8.4. Yleinen lineaarinen malli ja lineaariset rajoitukset

Estimaattorin hyvyys, Gaussin ja Markovin lause, Lagrangen menetelmä, Lineaariset rajoitukset, Lineaariset side-ehdot, Pienimmän neliösumman menetelmä, Rajoitettu pienimmän neliösumman estimaattori, Sidottu pienimmän neliösumman estimaattori, Tehokkuus, Yleinen lineaarinen malli

8.5. Bayeslainen yleinen lineaarinen malli

Bayesin kaava, Huonosti asetettu ongelma. Priorijakauma, Posteriorijakauma, Rajoitettu pienimmän neliösumman estimaattori, Regularisointi, Sekaestimointi, Yleistetty pienimmän neliösumman menetelmä

8.6. Yleinen lineaarinen malli ja stokastiset selittäjät

Ehdollistaminen, Kiinteät selittäjät, Satunnaiset selittäjät

Liite 1: Odotusarvovektori ja kovarianssimatriisi

Liite 2: Multinormaalijakauma

8.1. Yleinen lineaarinen malli ja sen parametrien estimointi

Yleinen lineaarinen malli ja sen parametrointi

Olkoon

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \varepsilon_j \ , j = 1, 2, \ldots, n$$

yleinen lineaarinen malli, jossa

y_j = **selitettävän muuttujan** *y havaittu* ja *satunnainen arvo* havaintoyksikössä $j = 1, 2, \ldots , n$

x_{ji} = **selittävän muuttujan** eli **selittäjän** x_i *havaittu* ja *kiinteä* eli *ei-satunnainen arvo* havaintoyksikössä $j = 1, 2, \ldots , n$; $i = 1, 2, \ldots , k$

β_0 = vakioselittäjän *regressiokerroin*, *tuntematon* ja *kiinteä* eli *ei-satunnainen vakio*

β_i = selittäjän x_i *regressiokerroin*, *tuntematon* ja *kiinteä* eli *ei-satunnainen vakio*, $i = 1, 2, \ldots , k$

ε_j = *ei-havaittava* ja *satunnainen* **jäännöstermi**, $j = 1, 2, \ldots , n$

Yleistä lineaarista mallia koskevat standardioletukset

(i) Selittäjän x_i havaitut arvot x_{ji} ovat *kiinteitä* eli *ei-satunnaisia vakioita*, $j = 1, 2, \ldots , n$; $i = 1, 2, \ldots , k$

(ii) Selittäjien välillä ei ole *lineaarisia riippuvuuksia*.

(iii) $\mathrm{E}(\varepsilon_j) = 0 \ , j = 1, 2, \ldots , n$

(iv) $\mathrm{Var}(\varepsilon_j) = \sigma^2 \ , j = 1, 2, \ldots , n$

(v) $\mathrm{Cor}(\varepsilon_j, \varepsilon_l) = 0 \ , j \neq l$

(vi) $\varepsilon_j \sim \mathrm{N}(0,\sigma^2) \ , j = 1, 2, \ldots , n$

Oletuksen (iv) mukaan kaikilla jäännöstermeillä ε_j , $j = 1, 2, \ldots , n$ on sama varianssi, jota kutsutaan *jäännösvarianssiksi*. Oletusta (iv) kutsutaan *homoskedastisuusoletukseksi*. Oletuksen (v) mukaan jäännöstermit eivät korreloi keskenään. Oletusta (v) kutsutaan *korreloimattomuusoletukseksi*.

Huomaa, että normaalisuusoletus (vi) sisältää oletukset (iv) ja (v).

Regressiotaso

Yhtälö

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

määrittelee *tason* avaruudessa $\mathbb{R}^{k+1}$. Tätä tasoa kutsutaan **regressiotasoksi**. Jäännösvarianssi σ^2 kuvaa havaintopisteiden

$$(x_{j1}, x_{j2}, \ldots, x_{jk}, y_j) \in \mathbb{R}^{k+1} \ , \ j = 1, 2, \ldots, n$$

vaihtelua regressiotason ympärillä.

Yleisen lineaarisen mallin matriisiesitys

Muodostetaan *selitettävän muuttujan y* havaituista arvoista y_j ($j = 1, 2, \ldots , n$) n-vektori

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Muodostetaan *selittäjien* x_i havaituista arvoista x_{ji} ($j = 1, 2, \ldots , n$; $i = 1, 2, \ldots , k$) ja ykkösistä $n\times(k + 1)$-matriisi

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

Muodostetaan *regressiokertoimista* β_i ($i = 0, 1, 2, \ldots , k$) $(k + 1)$-vektori

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

Muodostetaan *jäännöstermeistä* ε_j ($j = 0, 1, 2, \ldots , n$) n-vektori

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Tällöin *yleinen lineaarinen malli* voidaan kirjoittaa *matriisein* muotoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Mallissa on seuraavat osat:

$\mathbf{y}$ = *selitettävän muuttujan y havaittujen arvojen* y_j , $j = 1, 2, \ldots , n$ muodostama *satunnainen n*-vektori

$\mathbf{X}$ = *selittävien muuttujien* eli *selittäjien* $x_1, x_2, \ldots , x_k$ *havaittujen arvojen* x_{ji} , $j = 1, 2, \ldots , n$, $i = 1, 2, \ldots , k$ ja *ykkösten* muodostama *ei-satunnainen* $n\times(k + 1)$-matriisi

$\boldsymbol{\beta}$ = *regressiokertoimien* β_i , $i = 0, 1, 2, \ldots , k$ muodostama *tuntematon* ja *kiinteä* eli *ei-satunnainen* $(k + 1)$-vektori

$\boldsymbol{\varepsilon}$ = *jäännöstermien* ε_j , $j = 1, 2, \ldots , n$ muodostama *ei-havaittava* ja *satunnainen* n-vektori

Yleistä lineaarista mallia koskevat standardioletukset matriisimuodossa

(i) Matriisin **X** alkiot ovat *ei-satunnaisia vakioita*.

(ii) Matriisi **X** on *täysiasteinen*:

$$r(\mathbf{X}) = k + 1$$

(iii) $\mathrm{E}(\boldsymbol{\varepsilon}) = 0$

(iv)&(v) *Homoskedastisuus*- ja *korreloimattomuusoletus*:

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

(vi) *Normaalisuusoletus*:

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0},\sigma^2\mathbf{I})$$

Lisätietoja *odotusarvovektorista* $\mathrm{E}(\,\cdot\,)$, *kovarianssimatriisista* $\mathrm{Cov}(\,\cdot\,)$ ja *multinormaalijakaumasta* $\mathrm{N}_n(\,\cdot\,,\cdot\,)$: ks. liitteitä 1 ja 2.

Kohdissa (iv), (v) ja (vi) esiintyvä matriisi **I** on $n\times n$-*yksikkömatriisi*.

Standardioletuksista (i) ja (iii) seuraa, että

$$\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$$

Standardioletuksista (i), (iii), (iv) ja (v) seuraa, että

$$\begin{aligned}\mathrm{Cov}(\mathbf{y}) &= \mathrm{E}[(\mathbf{y}-\mathrm{E}(\mathbf{y}))(\mathbf{y}-\mathrm{E}(\mathbf{y}))'] \\ &= \mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] \\ &= \mathrm{Cov}(\boldsymbol{\varepsilon}) \\ &= \sigma^2\mathbf{I}\end{aligned}$$

Normaalisuusoletuksesta (vi) seuraa, että jäännöstermin $\boldsymbol{\varepsilon}$ tiheysfunktio on muotoa

$$f(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n;\sigma^2) = (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}\right\}$$

ja edelleen, että selitettävän muuttujan **y** tiheysfunktio on

$$f(y_1,y_2,\ldots,y_n;\boldsymbol{\beta},\sigma^2) = (2\pi)^{-n/2}\sigma^{-n}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

Yleisen lineaarisen mallin uskottavuusfunktio ja logaritminen uskottavuusfunktio

Standardioletuksesta (vi) seuraa, että havaintojen $y_1, y_2, \ldots, y_n$ *uskottavuusfunktio* on muotoa

$$L(\boldsymbol{\beta},\sigma^2;y_1,y_2,\ldots,y_n) = (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right\}$$

Vastaava *logaritminen uskottavuusfunktio* on muotoa

$$\begin{aligned}l(\boldsymbol{\beta},\sigma^2;y_1,y_2,\ldots,y_n) &= \log L(\boldsymbol{\beta},\sigma^2;y_1,y_2,\ldots,y_n) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\end{aligned}$$

Yleisen lineaarisen mallin parametrien suurimman uskottavuuden estimointi

Yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *suurimman uskottavuuden estimaattori* yhtyy standardioletuksien (i)-(vi) pätiessä vektorin $\boldsymbol{\beta}$ *pienimmän neliösumman estimaattoriin*

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Jäännösvarianssin σ^2 SU-estimaattori on

$$\hat{\sigma}^2 = \frac{1}{n} SSE$$

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

Perustelu:

Maksimoidaan logaritminen uskottavuusfunktio

$$l(\boldsymbol{\beta}, \sigma^2; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

parametrien $\boldsymbol{\beta}$ ja σ^2 suhteen.

Derivoidaan funktio $l(\boldsymbol{\beta}, \sigma^2; \mathbf{y})$ regressiokertoimien vektorin $\boldsymbol{\beta}$ suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X}'\mathbf{y} + \mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = \mathbf{0}$$

Vektori $\boldsymbol{\beta}$ voidaan ratkaista tästä yhtälöstä, koska matriisi $\mathbf{X}$ oletettiin täysiasteiseksi, jolloin matriisi $\mathbf{X}'\mathbf{X}$ on *epäsingulaarinen*. Ratkaisuksi saadaan

$$\hat{\boldsymbol{\beta}} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Ratkaisu vastaa uskottavuusfunktion maksimia, koska matriisi

$$-\frac{\partial l(\boldsymbol{\beta}, \sigma^2; \mathbf{y})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'} = \frac{1}{2\sigma^2}\mathbf{X}'\mathbf{X}$$

on *positiivisesti definiitti*.

Vektorin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori $\mathbf{b}$ yhtyy regressiokertoimien vektorin $\boldsymbol{\beta}$ *pienimmän neliösumman estimaattoriin*, koska logaritmisen uskottavuusfunktion maksimointi parametrin $\boldsymbol{\beta}$ suhteen on selvästi ekvivalenttia neliömuodon (neliösumman)

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

minimoinnin kanssa.

Sijoitetaan ratkaisu $\hat{\boldsymbol{\beta}} = \mathbf{b}$ logaritmiseen uskottavuusfunktion lausekkeeseen, derivoidaan saatava lauseke parametrin σ^2 suhteen ja merkitään derivaatta nollaksi:

$$\frac{\partial l(\mathbf{b}, \sigma^2; \mathbf{y})}{\partial \sigma^2} = -\frac{n}{2} \cdot \frac{1}{\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = 0$$

Parametri σ^2 saadaan ratkaistuksi tästä yhtälöstä. Ratkaisuksi saadaan

$$\hat{\sigma}^2 = \frac{1}{n} SSE$$

jossa

$$SSE = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb})$$

■

Estimoitu regressiotaso

Olkoon

$$\mathbf{b} = (b_0, b_1, b_2, \ldots, b_k)$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattori*.

Yhtälö

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

määrittelee *tason* avaruudessa $\mathbb{R}^{k+1}$. Tätä tasoa kutsutaan **estimoiduksi regressiotasoksi**.

Yleisen lineaarisen mallin suurimman uskottavuuden estimaattoreiden ominaisuudet

Jos yleinen lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

toteuttaa standardioletukset (i)-(vi), niin regressiokertoimien vektorin $\boldsymbol{\beta}$ SU-estimaattorilla

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on seuraavat ominaisuudet:

(i) **b** on *harhaton* parametrille $\boldsymbol{\beta}$:

$$\mathrm{E}(\mathbf{b}) = \boldsymbol{\beta}$$

(ii) $\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$

(iii) **b** ja $\hat{\sigma}^2$ ovat yhdessä *tyhjentäviä* parametreille $\boldsymbol{\beta}$ ja σ^2.

(iv) **b** on *tehokas* eli *minimivarianssinen* estimaattori.

(v) **b** on *tarkentuva*.

(vi) **b** noudattaa *normaalijakaumaa*:

$$\mathbf{b} \sim \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Jos yleinen lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

toteuttaa standardioletukset (i)-(vi), niin jäännösvarianssin σ^2 SU-estimaattorilla

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$

on seuraavat ominaisuudet:

(i) $\hat{\sigma}^2$ on *harhainen*, mutta estimaattori

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

on *harhaton* parametrille σ^2.

(ii) $\mathbf{b}$ ja $\hat{\sigma}^2$ ovat yhdessä *tyhjentäviä* parametreille $\boldsymbol{\beta}$ ja σ^2.

(iii) $\hat{\sigma}^2$ ei ole *tehokas* eli *minimivarianssinen* estimaattori.

(iv) $\hat{\sigma}^2$ on *tarkentuva*.

(v) $(n-k-1)s^2/\sigma^2$ noudattaa χ^2-*jakaumaa* vapaustein $(n-k-1)$:

$$\frac{(n-k-1)s^2}{\sigma^2} \sim \chi^2(n-k-1)$$

Gaussin ja Markovin lause

Olkoot $\hat{\boldsymbol{\theta}}_1$ ja $\hat{\boldsymbol{\theta}}_2$ kaksi *harhatonta* estimaattoria parametrille $\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \ldots, \theta_p)$. Sanomme, että estimaattori $\hat{\boldsymbol{\theta}}_1$ on **parempi** eli **tehokkaampi** kuin estimaattori $\hat{\boldsymbol{\theta}}_2$, jos

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}_1) \le \mathrm{Cov}(\hat{\boldsymbol{\theta}}_2)$$

jolla tarkoitetaan sitä, että

$$\mathbf{t}'\mathrm{Cov}(\hat{\boldsymbol{\theta}}_1)\mathbf{t} \le \mathbf{t}'\mathrm{Cov}(\hat{\boldsymbol{\theta}}_2)\mathbf{t}$$

kaikille $\mathbf{t} \in \mathbb{R}^p$.

Keskeisen perustelun *suurimman uskottavuuden* (ja *pienimmän neliösumman*) *menetelmän* käytölle yleisen lineaarisen mallin regressiokertoimien estimoimisessa antaa seuraava lause:

Gaussin ja Markovin lause:

Oletetaan, että yleinen lineaarinen malli

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

toteuttaa standardioletukset (i)-(v). Tällöin regressiokertoimien vektorin $\boldsymbol{\beta}$ *suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattori*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on *paras* regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*.

Perustelu:

Olkoon

$$\mathbf{c} = \mathbf{D}^*\mathbf{y}$$

jokin standardioletukset toteuttavan yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$$

regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaarinen* ja *harhaton* estimaattori. Matriisi $\mathbf{D}^*$ on $(k+1)\times n$-matriisi, joka ei riipu vektorista $\mathbf{y}$.

Olkoon

$$\mathbf{D} = \mathbf{D}^* - (\mathbf{X'X})^{-1}\mathbf{X'}$$

Siten

$$\begin{aligned}
\mathbf{c} &= \mathbf{D}^*\mathbf{y} \\
&= (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\mathbf{y} \\
&= (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})(\mathbf{X\beta} + \boldsymbol{\varepsilon}) \\
&= (\mathbf{DX} + \mathbf{I})\boldsymbol{\beta} + (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\boldsymbol{\varepsilon}
\end{aligned}$$

Koska

$$\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$$

niin

$$\mathrm{E}(\mathbf{c}) = (\mathbf{DX} + \mathbf{I})\boldsymbol{\beta} + (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\,\mathrm{E}(\boldsymbol{\varepsilon}) = (\mathbf{DX} + \mathbf{I})\boldsymbol{\beta}$$

Siten estimaattori $\mathbf{c}$ on *harhaton* vain, jos

$$\mathbf{DX} = \mathbf{0}$$

jolloin

$$\mathbf{c} = \mathbf{D}^*\mathbf{y} = \boldsymbol{\beta} + (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\boldsymbol{\varepsilon}$$

Tarkastellaan nyt estimaattorin $\mathbf{c}$ *kovarianssimatriisia*:

$$\begin{aligned}
\mathrm{Cov}(\mathbf{c}) &= \mathrm{E}[(\mathbf{c} - \mathrm{E}(\mathbf{c}))(\mathbf{c} - \mathrm{E}(\mathbf{c}))'] \\
&= \mathrm{E}[(\mathbf{c} - \boldsymbol{\beta})(\mathbf{c} - \boldsymbol{\beta})'] \\
&= \mathrm{E}[(\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\boldsymbol{\varepsilon\varepsilon}'(\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})'] \\
&= (\mathbf{D} + (\mathbf{X'X})^{-1}\mathbf{X'})\,\mathrm{E}(\boldsymbol{\varepsilon\varepsilon}')(\mathbf{D}' + \mathbf{X}(\mathbf{X'X})^{-1}) \\
&= \sigma^2[\mathbf{DD}' + \mathbf{DX}(\mathbf{X'X})^{-1} + (\mathbf{X'X})^{-1}\mathbf{X'D}' + (\mathbf{X'X})^{-1}] \qquad |\ \mathrm{E}(\boldsymbol{\varepsilon\varepsilon}') = \sigma^2\mathbf{I} \\
&= \sigma^2[\mathbf{DD}' + (\mathbf{X'X})^{-1}] \qquad |\ \mathbf{DX} = \mathbf{0}
\end{aligned}$$

Koska

$$\begin{aligned}
\mathbf{t}'\,\mathrm{Cov}(\mathbf{c})\mathbf{t} &= \sigma^2[t'\mathbf{DD}'t + t'(\mathbf{X'X})^{-1}t] \\
&\geq \sigma^2[t'(\mathbf{X'X})^{-1}t] \qquad |\ t'\mathbf{DD}'t \geq 0 \\
&= \mathbf{t}'\,\mathrm{Cov}(\mathbf{b})\mathbf{t}
\end{aligned}$$

niin

$$\mathrm{Cov}(\mathbf{c}) \geq \mathrm{Cov}(\mathbf{b})$$

ja siten suurimman uskottavuuden (pienimmän neliösumman) estimaattori **b** on *paras* regressiokertoimien vektorin $\boldsymbol{\beta}$ *lineaaristen ja harhattomien estimaattoreiden joukossa*.

■

Siten yleisen lineaarisen mallin regressiokertoimien *suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattori* on *optimaalinen estimaattori* Gaussin ja Markovin lauseen määrittelemässä mielessä.

Siirtymällä pois lineaaristen ja *harhattomien estimaattoreiden luokasta* tai *käyttämällä jotakin muuta optimaalisuuden kriteeriä* kuin estimaattorin varianssia *saattaa olla mahdollista löytää suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattoria parempia estimaattoreita*.

Sovite ja residuaali

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 $\boldsymbol{\beta}$ SU-estimaattori.

Määritellään estimoidun mallin **sovite** $\hat{\mathbf{y}}$ kaavalla

$$\begin{aligned}\hat{\mathbf{y}} &= \mathbf{X}\mathbf{b} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= \mathbf{P}\mathbf{y}\end{aligned}$$

n-vektorin $\hat{\mathbf{y}}$ *j*. alkio on

$$\hat{y}_j = b_0 + b_1 x_{j1} + b_2 x_{j2} + \cdots + b_p x_{jp} \ , \ j = 1, 2, \ldots, n$$

Matriisi

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

on *projektio* (*symmetrinen* ja *idempotentti*):

$$\mathbf{P}' = \mathbf{P}$$

$$\mathbf{P}^2 = \mathbf{P}$$

Määritellään estimoidun mallin **residuaali e** kaavalla

$$\begin{aligned}\mathbf{e} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{X}\mathbf{b} \\ &= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{I} - \mathbf{P})\mathbf{y} \\ &= \mathbf{M}\mathbf{y}\end{aligned}$$

n-vektorin **e** *j*. alkio on

$$e_j = y_j - \hat{y}_j = y_j - b_0 - b_1 x_{j1} - b_2 x_{j2} - \cdots - b_p x_{jp}\ ,\ j = 1, 2, \ldots, n$$

Matriisi

$$\mathbf{M} = \mathbf{I} - \mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

on *projektio* (*symmetrinen* ja *idempotentti*):

$$\mathbf{M}' = \mathbf{M}$$

$$\mathbf{M}^2 = \mathbf{M}$$

Lisäksi matriisin **P** rivit ovat *kohtisuorassa* matriisin **M** sarakkeita vastaan ja matriisin **M** rivit ovat *kohtisuorassa* matriisin **P** sarakkeita vastaan:

$$\mathbf{PM} = \mathbf{MP} = \mathbf{0}$$

Sovitteen ja residuaalin ominaisuudet

Sovitteella $\hat{\mathbf{y}}$ on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\hat{\mathbf{y}}) = \mathbf{X}\boldsymbol{\beta}$$

$$\mathrm{Cov}(\hat{\mathbf{y}}) = \sigma^2\mathbf{P} = \sigma^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Lisäksi, jos jäännöstermi $\boldsymbol{\varepsilon}$ on normaalijakautunut (standardioletus (vi)), niin

$$\hat{\mathbf{y}} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{P})$$

Sovitteella **e** on seuraavat *stokastiset ominaisuudet*:

$$\mathrm{E}(\mathbf{e}) = \mathbf{0}$$

$$\mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I} - \mathbf{P}) = \sigma^2(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$$

Lisäksi, jos jäännöstermi $\boldsymbol{\varepsilon}$ on normaalijakautunut (standardioletus (vi)), niin

$$\mathbf{e} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{M})$$

Huomaa, että molemmat multinormaalijakaumat ovat *singulaarisia*.

Varianssianalyysihajotelma

Olkoon

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

selitettävän muuttujan y havaittujen arvojen $y_1, y_2, \ldots, y_n$ **kokonaisneliösumma**, jossa

$$\mathbf{y} = (y_1, y_2, \ldots, y_n)$$

on havaintoarvojen $y_1, y_2, \ldots, y_n$ muodostama n-vektori,

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

on havaintoarvojen $y_1, y_2, \ldots, y_n$ aritmeettinen keskiarvo ja

$$\mathbf{1} = (1, 1, \ldots, 1)$$

on ykkösten muodostama n-vektori. Kokonaisneliösumma SST kuvaa selitettävän muuttujan y havaittujen arvojen $y_1, y_2, \ldots, y_n$ vaihtelua.

Huomaa, että selitettävän muuttujan y havaittujen arvojen $y_1, y_2, \ldots, y_n$ otosvarianssi saadaan kaavalla

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

Määritellään **residuaalien neliösumma** eli **jäännösneliösumma** SSE kaavalla

$$SSE = \mathbf{e}'\mathbf{e} = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = \sum_{i=1}^{n} e_i^2$$

jossa

$$\mathbf{e} = (e_1, e_2, \ldots, e_n)$$

on residuaalien muodostama n-vektori. Jäännösneliösumma SSE kuvaa residuaalien $e_1, e_2, \ldots, e_n$ vaihtelua estimoidun regressiotason ympärillä.

Huomaa, että jäännösvarianssin σ^2 harhaton estimaattori saadaan kaavalla

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

Voidaan osoittaa, että aina pätee

$$SSE \le SST$$

Tarkastellaan erotusta

$$SSM = SST - SSE \ge 0$$

Voidaan osoittaa, että

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

jossa

$$\hat{\mathbf{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n)$$

on sovitteiden $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ muodostama n-vektori,

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

on havaintoarvojen $y_1, y_2, \ldots, y_n$ aritmeettinen keskiarvo ja

$$\mathbf{1} = (1, 1, \ldots, 1)$$

on ykkösten muodostama n-vektori. Koska erotus $SSM = SST - SSE$ voidaan esittää neliösummana, erotusta SSM kutsutaan **mallineliösummaksi**.

Huomaa, että

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

Edellä esitetyn nojalla kokonaisneliösumma *SST* voidaan hajottaa neliösummien *SSM* ja *SSE* summaksi:

$$SST = SSM + SSE$$

Kokonaisneliösumman *SST* hajotelmaa mallineliösumman *SSM* ja jäännösneliösumman *SSE* summaksi kutsutaan **varianssianalyysihajotelmaksi**.

On ilmeistä, että mitä *pienempi* on jäännösneliösumman $SSE = \mathbf{e}'\mathbf{e}$ osuus kokonaisneliösummasta *SST*, sitä *suurempi* on mallineliösumman *SSM* osuus ja *sitä paremmin malli selittää selitettävän muuttujan havaittujen arvojen vaihtelun*.

Selitysaste ja sen ominaisuudet

Määritellään estimoidun lineaarisen regressiomallin **selitysaste** R^2 kaavalla

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

Koska *varianssianalyysihajotelman* mukaan

$$SST = SSM + SSE$$

niin

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

Voidaan osoittaa, että seuraavat ehdot ovat *yhtäpitäviä*:

(i) $R^2 = 1$

(ii) Kaikki residuaalit häviävät:

$e_j = 0$ kaikille $j = 1, 2, \ldots , n$

(iii) Kaikki havaintopisteet

$$(x_{j1}, x_{j2}, \ldots, x_{jk}, y_j)\,,\; j = 1, 2, \ldots, n$$

asettuvat *samalle tasolle*.

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

Edelleen voidaan osoittaa, että myös seuraavat ehdot ovat *yhtäpitäviä*:

(i) $R^2 = 0$

(ii) $b_1 = b_2 = \cdots = b_k = 0$

(iii) Malli *ei ollenkaan selitä* selitettävän muuttujan arvojen vaihtelua.

Siten selitysaste R^2 kuvaa *estimoidun mallin selittämää osuutta selitettävän muuttujan y arvojen vaihtelusta*.

Selitysaste ilmoitetaan tavallisesti prosentteina, jolloin sanotaan, että *estimoitu malli on selittänyt*

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

selitettävän muuttujan y arvojen vaihtelusta.

Voidaan osoittaa, että

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

jossa

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

on selitettävän muuttujan y havaittujen arvojen $y_1, y_2, \ldots, y_n$ ja estimoidun mallin sovitteiden $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ tavanomainen *otoskorrelaatiokerroin*.

8.2. Päättely yleisestä lineaarisesta mallista

Oletukset

Olkoon

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \cdots + \beta_k x_{jk} + \varepsilon_j \ , j = 1, 2, \ldots, n$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jonka matriisiesitysmuoto on

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

jossa $\mathbf{X}$ on $n\times(k + 1)$-matriisi ja

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

PNS-estimaattoreiden varianssien estimointi

Jos yleistä lineaarista mallia koskevat standardioletukset (i)-(vi) pätevät, niin regressiokertoimen

$$\beta_i \ , i = 0, 1, 2, \ldots, k$$

suurimman uskottavuuden estimaattorin b_i varianssin

$$\mathrm{D}^2(b_i) = \sigma_{b_i}^2 = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1}]_{i+1,i+1}$$

harhaton estimaattori on

$$\hat{\mathrm{D}}^2(b_i) = s^2[(\mathbf{X}'\mathbf{X})^{-1}]_{i+1,i+1}$$

jossa

$$s^2 = \frac{1}{n-1}SSE = \frac{1}{n-1}(\mathbf{y}-\mathbf{Xb})'(\mathbf{y}-\mathbf{Xb}) = \frac{1}{n-1}\mathbf{e}'\mathbf{e} = \frac{1}{n-1}\sum_{i=1}^{n} e_i^2$$

on *jäännösvarianssin σ^2 harhaton estimaattori*.

Regressiokertoimien luottamusvälit

Valitaan luottamustasoksi $(1 - \alpha)$. Jos standardioletukset (i)-(vi) pätevät, niin regressiokertoimen

$$\beta_i \ , i = 0, 1, 2, \ldots, k$$

luottamusvälit ovat muotoa

$$b_i \pm t_{\alpha/2}\,\hat{\mathrm{D}}(b_i) \ , i = 0,1,2,\ldots,k$$

jossa

b_i = regressiokertoimen β_i *SU-estimaattori*

$\pm t_{\alpha/2}$ = luottamustasoa $(1 - \alpha)$ vastaavat *luottamuskertoimet t*-jakaumasta, jonka vapausasteiden lukumäärä on $(n - k - 1)$

$\hat{\mathrm{D}}^2(b_i)$ = regressiokertoimen β_i SU-estimaattorin *varianssin harhaton estimaattori*

Regressiokertoimen β_i luottamusväli voidaan helposti johtaa käyttämällä hyväksi sitä, että satunnaismuuttuja

$$t_i = \frac{b_i - \beta_i}{\hat{\mathrm{D}}(b_i)}, \; i = 0,1,2,\ldots,k$$

noudattaa t-jakaumaa vapausastein $(n - k - 1)$:

$$t_i \sim t(n-1), \; i = 1,2,\ldots,n$$

mikä merkitsee sitä, että satunnaismuuttujan t_i jakauma ei riipu estimoinnin kohteena olevista parametreista ja siten satunnaismuuttuja t_i on *saranasuure*.

Regressiokertoimen β_i *luottamusväli*

$$b_i \pm t_{\alpha/2}\,\hat{\mathrm{D}}(b_i), \; i = 0,1,2,\ldots,k$$

luottamustasolla $(1 - \alpha)$ *peittää regressiokertoimen* β_i *todellisen arvon todennäköisyydellä* $(1 - \alpha)$:

$$\Pr(b_i - t_{\alpha/2}\,\hat{\mathrm{D}}(b_i) \le \beta_i \le b_i + t_{\alpha/2}\,\hat{\mathrm{D}}(b_i)) = 1 - \alpha$$

Jos otantaa toistetaan, niin luottamustason *frekvenssitulkinnan* mukaan otoksista konstruoiduista luottamusväleistä

$$100\times(1 - \alpha)\ \%$$

peittää regressiokertoimen β_i *todellisen arvon* ja

$$100\times\alpha\ \%$$

väleistä *ei peitä regressiokertoimen* β_i *todellista arvoa*.

Lineaaristen rajoitusten testaus

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jossa $\mathbf{X}$ on ei-satunnainen $n\times(k + 1)$-matriisi, $n \ge k + 1$, $r(\mathbf{X}) = k + 1$.

Oletetaan, että regressiokertoimien vektorille $\boldsymbol{\beta}$ on asetettu *nollahypoteesi*

$$\mathrm{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

jossa $\mathbf{R}$ on ei-satunnainen $m\times(k + 1)$-matriisi, $m \le k + 1$, $r(\mathbf{R}) = m$. Nollahypoteesin H_0 mukaan regressiokertoimia *sitoo m lineaarisia rajoitusta* eli *side-ehtoa*. Side-ehtoa $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ kutsutaan usein *yleiseksi lineaariseksi hypoteesiksi*. Yhtälö $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ määrittelee m-ulotteisen tason $(k + 1)$-ulotteisessa avaruudessa $\mathbb{R}^{k+1}$.

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoittamaton suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattori* on

$$\mathbf{b} = \mathbf{U}\mathbf{X}'\mathbf{y}$$

jossa

$$\mathbf{U} = (\mathbf{X}'\mathbf{X})^{-1}$$

Regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitettu* eli *sidottu suurimman uskottavuuden* (*pienimmän neliösumman*) *estimaattori* on (ks. kappaletta 8.4)

$$\mathbf{b}_R = \mathbf{b} + \mathbf{U}\mathbf{R}'\mathbf{S}(\mathbf{r} - \mathbf{R}\mathbf{b})$$

jossa

$$\mathbf{b} = \mathbf{U}\mathbf{X}'\mathbf{y}$$

on vektorin $\boldsymbol{\beta}$ *rajoitettu suurimman uskottavuuden estimaattori*, matriisi $\mathbf{U}$ on kuten edellä ja

$$\mathbf{S} = (\mathbf{R}\mathbf{U}\mathbf{R}')^{-1}$$

Rajoitettu suurimman uskottavuuden estimaattori $\mathbf{b}_R$ voidaan johtaa sidottujen ääriarvojen etsimiseen tarkoitetulla *Lagrangen menetelmällä* (ks. kappaletta 8.4).

Määritellään *jäännösneliösummat*

$$SSE = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})$$
$$SSE_R = (\mathbf{y} - \mathbf{X}\mathbf{b}_R)'(\mathbf{y} - \mathbf{X}\mathbf{b}_R)$$

Tällöin **osamäärätestisuure** *nollahypoteesille*

$$\mathrm{H}_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

on muotoa

$$\lambda = \left(\frac{SSE}{SSE_R}\right)^{-\frac{n}{2}}$$

Osamäärätestisuureen λ jakaumaa ei tunneta, mutta osamäärätestisuureen *asymptoottista jakaumaa* koskevasta yleisestä tuloksesta seuraa, että *nollahypoteesin pätiessä*

$$-2\log\lambda = -2\log\left(\frac{SSE}{SSE_R}\right)^{-\frac{n}{2}} = n\log(SSE_R) - n\log(SSE) \sim_a \chi^2(m)$$

jossa

$m =$ rajoitusten eli side-ehtojen lukumäärä

$=$ rivien lukumäärä matriisissa $\mathbf{R}$

Suuret testisuureen $-2\log\lambda$ arvot merkitsevät sitä, että *nollahypoteesi on asetettava kyseenalaiseksi*.

Osamäärätestisuureen λ jakaumaa ei siis tunneta, kun otoskoko on *äärellinen*. Sen sijaan testisuure

$$F = \frac{n-k-1}{m}\left(\frac{1}{\lambda^{\frac{2}{n}}} - 1\right)$$

noudattaa eksaktisti *F-jakaumaa* vapausastein m ja $(n - k - 1)$ nollahypoteesin

$$\mathrm{H}_0 : \mathbf{R\beta} = \mathbf{r}$$

pätiessä:

$$F \sim_{\mathrm{H}_0} F(m, n-k-1)$$

Testisuure F voidaan kirjoittaa seuraaviin muotoihin:

$$\begin{aligned} F &= \frac{n-k-1}{m} \cdot \frac{SSE_R - SSE}{SSE} \\ &= \frac{(\mathbf{r} - \mathbf{Rb})'\mathbf{S}(\mathbf{r} - \mathbf{Rb})}{ms^2} \end{aligned}$$

jossa

$$(n-k-1)s^2 = (\mathbf{y} - \mathbf{Xb})'(\mathbf{y} - \mathbf{Xb}) = SSE$$

Testisuure F voidaan johtaa myös *Waldin testinä* tai *Lagrangen kertojatestinä*.

Testi regression olemassaololle

Olkoon testattavana nollahypoteesina

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

Jos nollahypoteesi H_0 *pätee*, selitettävä muuttuja *y ei riipu* lineaarisesti yhdestäkään selittäjästä x_1, x_2, … , x_k. Sen sijaan, jos nollahypoteesi H_0 *ei päde*, selitettävä muuttuja *y riippuu* lineaarisesti ainakin yhdestä selittäjästä x_1, x_2, … , x_k .

Määritellään ***F*-testisuure**

$$\begin{aligned} F &= \frac{n-k-1}{k} \cdot \frac{SSM}{SSE} \\ &= \frac{n-k-1}{k} \cdot \frac{SST - SSE}{SSE} \\ &= \frac{n-k-1}{k} \cdot \frac{R^2}{1 - R^2} \end{aligned}$$

jossa

SST = selitettävän muuttujan y havaittujen arvojen kokonaisvaihtelua kuvaava neliösumma

SSM = estimoidun mallin *mallineliösumma*

SSE = estimoidun mallin *jäännösneliösumma*

R^2 = estimoidun mallin *selitysaste*

Testisuure F vertaa toisiinsa *residuaalivarianssia*

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

ja *mallivarianssia*

$$s_M^2 = \frac{1}{k} SSM$$

Oletetaan, että standardioletukset (i)-(vi) pätevät. Tällöin testisuure F noudattaa *nollahypoteesin*

$$\mathrm{H}_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

pätiessä F-jakaumaa vapausastein k ja $(n - k - 1)$:

$$F \sim_{\mathrm{H}_0} F(k, n-k-1)$$

Testisuureen F *normaaliarvo* eli odotusarvo nollahypoteesin H_0 pätiessä on approksimatiivisesti (suurille n) = 1:

$$\mathrm{E}(F) \approx_{\mathrm{H}_0} 1$$

Suuret testisuureen F arvot viittaavat siihen, että nollahypoteesi H_0 *ei päde*.

Testi on erikoistapaus edellä esitetystä testistä *yleiselle lineaarisille hypoteesille*. Testisuure F voidaan johtaa myös suoraan *osamäärätestinä*, *Waldin testinä* tai *Lagrangen kertojatestinä* nollahypoteesille H_0 .

Testit regressiokertoimille

Olkoon nollahypoteesina

$$\mathrm{H}_{0i} : \beta_i = 0 \,,\, i = 0,1,2,\ldots,k$$

Jos nollahypoteesi H_{00} *pätee*, mallissa *ei ole vakiota*. Jos nollahypoteesi H_{0i} *pätee*, selitettävä muuttuja *y ei riipu* lineaarisesti selittäjästä x_i , i = 1, 2, … , k. Jos nollahypoteesi H_{0i} *ei päde*, selitettävä muuttuja *y riippuu* lineaarisesti selittäjästä x_i , i = 1, 2, … , k.

Määritellään ***t*-testisuureet**

$$t_i = \frac{b_i}{\hat{\mathrm{D}}(b_i)}$$

jossa

b_i = regressiokertoimen β_i *SU-estimaattori*

$\hat{\mathrm{D}}^2(b_i)$ = regressiokertoimen β_i SU-estimaattorin *varianssin harhaton estimaattori*

Oletetaan, että standardioletukset (i)-(vi) pätevät. Tällöin testisuure t_i noudattaa *nollahypoteesin*

$$\mathrm{H}_{0i} : \beta_i = 0 \,,\, i = 0,1,2,\ldots,k$$

pätiessä *t-jakaumaa* vapausastein $(n - k - 1)$:

$$t_i \sim_{H_0} t(n-k-1)\ ,\ i = 1,2,\ldots,n$$

Testisuureen t_i normaaliarvo eli odotusarvo nollahypoteesin H_{0i} pätiessä on

$$E(t_i) =_{H_{0i}} 0\ ,\ i = 0,1,2,\ldots,k$$

Itseisarvoltaan suuret testisuureen t_i arvot viittaavat siihen, että nollahypoteesi H_{0i} ei päde.

Testit ovat erikoistapauksia edellä esitetystä testistä *yleiselle lineaarisille hypoteesille*. Testisuureet t_i , $i = 0, 1, 2, \ldots , k$ voidaan johtaa myös suoraan *osamäärätesteinä*, *Waldin testeinä* tai *Lagrangen kertojatesteinä* nollahypoteesille H_{0i} , $i = 0, 1, 2, \ldots , k$.

8.3. Yleinen lineaarinen malli ja yleistetty pienimmän neliösumman menetelmä

Oletukset

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jossa $\mathbf{X}$ on $n\times(k + 1)$-matriisi.

Standardioletuksien (iv) ja (v) mukaan

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$$

Korvataan oletukset (iv) ja (v) nyt seuraavilla oletuksilla:

(iv)′-(v)′ $\quad \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$

jossa $\mathbf{V}$ on *positiivisesti definiitti* $n\times n$-matriisi. Tällöin jäännöstermiä $\boldsymbol{\varepsilon}$ koskeva normaalisuusoletus

$$\boldsymbol{\varepsilon} \sim N_n(\mathbf{0},\sigma^2\mathbf{I})$$

on korvattava oletuksella

(vi)′ $\quad \boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$

Huomautus:

Matriisi $\mathbf{V}$ on monissa käytännön sovelluksissa *tuntematon*. Tällöin matriisissa $\mathbf{V}$ on

$$\tfrac{1}{2}n(n+1)$$

vapaata, *estimoitavaa parametria* ja yleinen lineaarinen malli on voimakkaasti *yli-parametroitu* ja malli ei ole *estimoituva*. Tällöin matriisi $\mathbf{V}$ on *parametroitava uudelleen* sellaisella tavalla, joka vähentää estimoitavien parametrien lukumäärää niin paljon, että mallista tulee *estimoituva*.

Yleistetty pienimmän neliösumman estimaattori

Koska matriisi

$$\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{V}$$

on oletettu *positiivisesti definiitiksi*, matriisilla $\mathbf{V}$ on *Cholesky-hajotelma*

$$\mathbf{V} = \mathbf{U}\mathbf{U}'$$

missä $n\times n$-matriisi $\mathbf{U}$ on *epäsingulaarinen* yläkolmiomatriisi.

Kerrotaan regressioyhtälö

(1) $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

vasemmalta matriisilla $\mathbf{U}^{-1}$, jolloin saadaan regressioyhtälö

(2) $\mathbf{U}^{-1}\mathbf{y} = \mathbf{U}^{-1}\mathbf{X}\boldsymbol{\beta} + \mathbf{U}^{-1}\boldsymbol{\varepsilon}$

Regressioyhtälö (2) voidaan kirjoittaa muotoon

(3) $\mathbf{z} = \mathbf{T}\boldsymbol{\beta} + \boldsymbol{\delta}$

jossa

$$\begin{aligned}
\mathbf{z} &= \mathbf{U}^{-1}\mathbf{y} \\
\mathbf{T} &= \mathbf{U}^{-1}\mathbf{X} \\
\boldsymbol{\delta} &= \mathbf{U}^{-1}\boldsymbol{\varepsilon}
\end{aligned}$$

Regressioyhtälön (3) jäännöstermi $\boldsymbol{\delta}$ on *korreloimaton*:

$$\mathrm{Cov}(\boldsymbol{\delta}) = \mathbf{U}^{-1}\mathrm{Cov}(\boldsymbol{\varepsilon})(\mathbf{U}^{-1})' = \sigma^2\,\mathbf{U}^{-1}\mathbf{V}(\mathbf{U}')^{-1} = \sigma^2\,\mathbf{U}^{-1}\mathbf{U}\mathbf{U}'(\mathbf{U}')^{-1} = \sigma^2\mathbf{I}$$

joten standardioletukset (i)-(v) pätevät regressiomallille (3).

Soveltamalla pienimmän neliösumman menetelmää regressioyhtälöön (3) vektorin $\boldsymbol{\beta}$ pienimmän neliösumman estimaattoriksi saadaan

$$\begin{aligned}
\mathbf{b}_{GLS} &= (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{z} \\
&= (\mathbf{X}'(\mathbf{U}')^{-1}\mathbf{U}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{U}')^{-1}\mathbf{U}^{-1}\mathbf{y} \\
&= (\mathbf{X}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{X})^{-1}\,\mathbf{X}'(\mathbf{U}\mathbf{U}')^{-1}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}
\end{aligned}$$

Estimaattoria $\mathbf{b}_{GLS}$ kutsutaan mallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}$ **yleistetyksi pienimmän neliösumman** (**PNS-**) **estimaattoriksi**.

Yleistetyn PNS-estimaattorin ominaisuudet

Yleisen lineaarisen mallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}$ yleistetyn PNS-estimaattorin $\mathbf{b}_{GLS}$ keskeiset *stokastiset ominaisuudet* on esitetty seuraavassa lauseessa:

Lause 8.3.1.

Oletetaan, että yleisen lineaarisen mallin (1) oletukset (i)-(iii) ja (iv)′-(v)′ pätevät. Tällöin

(i) $\mathrm{E}(\mathbf{b}_{GLS}) = \boldsymbol{\beta}$

(ii) $\mathrm{Cov}(\mathbf{b}_{GLS}) = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$

(iii) Erityisesti

$$\mathrm{Var}(b_i^G) = \sigma^2[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}]_{i+1,i+1}\,,\ i = 0,1,2,\ldots,k$$

missä

$$\mathbf{b}_{GLS} = (b_0^G, b_1^G, b_2^G, \ldots, b_k^G)$$

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

$$\begin{aligned}
\mathbf{z} &= \mathbf{U}^{-1}\mathbf{y} \\
\mathbf{T} &= \mathbf{U}^{-1}\mathbf{X} \\
\boldsymbol{\delta} &= \mathbf{U}^{-1}\boldsymbol{\varepsilon}
\end{aligned}$$

ja $\mathbf{U}$ on epäsingulaarinen yläkolmiomatriisi joka toteuttaa ehdon

$$\mathbf{V} = \mathbf{U}\mathbf{U}'$$

Siten

$$\mathrm{Cov}(\mathbf{b}_{GLS}) = \sigma^2(\mathbf{T}\mathbf{T})^{-1} = \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

(iii) Kohta (iii) on suora seuraus kohdasta (ii).

■

Huomautus:

Lauseen 8.3.1. kohdan (i) mukaan yleistetty PNS-estimaattori $\mathbf{b}_{GLS}$ on regressiokertoimien vektorin $\boldsymbol{\beta}$ *harhaton estimaattori*.

Lause 8.3.2.

Oletetaan, että yleisen lineaarisen mallin (1) oletuksien (i)-(iii) ja (iv)′-(v)′ lisäksi normaalisuusoletus (vi)′ pätee. Tällöin

$$\mathbf{b}_{GLS} \sim \mathrm{N}_{k+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1})$$

Erityisesti

$$b_i^G \sim \mathrm{N}\left(\beta_i, \sigma^2[(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}]_{i+1,i+1}\right), \; i = 0, 1, 2, \ldots, k$$

missä

$$\mathbf{b}_{GLS} = (b_0^G, b_1^G, b_2^G, \ldots, b_k^G)$$

Perustelu:

Lause 8.3.2. seuraa suoraan lauseesta 8.3.1., koska yleistetty PNS-estimaattori

$$\mathbf{b}_{GLS} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

on multinormaalisen satunnaismuuttujan **y** lineaarimuunnoksena multinormaalinen.

■

Yleistetyn PNS-estimaattorin hyvyys

Koska malli (3) toteuttaa ns. standardioletukset (i)-(v), kerroinvektorin $\boldsymbol{\beta}$ yleistetty PNS-estimaattori

$$\mathbf{b}_{GLS} = (\mathbf{T}'\mathbf{T})^{-1}\mathbf{T}'\mathbf{z} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

on Gaussin ja Markovin lauseen (ks. kappale 8.1.) mukaan *paras lineaaristen ja harhattomien estimaattoreiden joukossa*.

Jos siis yleisen lineaarisen mallin (1) standardioletukset (i)-(iii) ja oletukset (iv)′-(v)′ pätevät, niin *yleistetty PNS-estimaattori* $\mathbf{b}_{GLS}$ *on parempi kuin tavallinen PNS-estimaattori*

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

mikä 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 *ei-negatiivisesti definiitti* kaikille positiivisesti definiiteille $n\times n$-matriiseille **V**.

Yleistetty PNS-estimaattori nähdään parhaaksi lineaaristen ja harhattomien estimaattoreiden joukossa myös seuraavalla tavalla:

Olkoon

$$\mathbf{b}^* = \mathbf{H}\mathbf{y}$$

jokin kerroinvektorin $\boldsymbol{\beta}$ lineaarinen ja harhaton estimaattori. Tällöin

$$\mathrm{E}(\mathbf{b}^*) = \mathbf{H}\mathrm{E}(\mathbf{y}) = \mathbf{H}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$$

josta seuraa, että

$$\mathbf{H}\mathbf{X} = \mathbf{I}$$

Määritellään matriisi **C** yhtälöllä

$$\mathbf{H} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{C}$$

Koska välttämättä $\mathbf{C}\mathbf{X} = \mathbf{0}$,

$$\mathrm{Cov}(\mathbf{b}^*) = \mathrm{Cov}(\mathbf{H}\mathbf{y}) = \mathrm{Cov}(\mathbf{b}_{GLS}) + \mathbf{C}\mathbf{V}^{-1}\mathbf{C}'$$

Koska matriisi

$$\mathrm{Cov}(\mathbf{b}^*) - \mathrm{Cov}(\mathbf{b}_{\mathrm{GLS}}) = \mathbf{C}\mathbf{V}^{-1}\mathbf{C}'$$

on ei-negatiivisesti definiitti, niin *yleistetty PNS-estimaattori* $\mathbf{b}_{GLS}$ *on parempi kuin mikä tahansa muu lineaarinen ja harhaton estimaattori* $\mathbf{b}^*$.

8.4. Yleinen lineaarinen malli ja lineaariset rajoitukset

Oletukset

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jossa $\mathbf{X}$ on $n\times(k + 1)$-matriisi ja

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0},\sigma^2\mathbf{I})$$

Oletetaan, että yleisen lineaarisen mallin (1) standardioletukset (i)-(vi) pätevät, mutta oletetaan lisäksi, että *regressiokertoimia sitoo* **lineaarinen rajoitus** eli **side-ehto**

(2) $\quad \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$

jossa $\mathbf{R}$ *täysiasteinen* $m\times(k + 1)$-matriisi, $m \le k + 1$.

Huomautus:

Lineaarisen mallin (1) regressiokertoimien vektori $\boldsymbol{\beta}$ *voi varioida vapaasti avaruudessa* $\mathbb{R}^{k+1}$. Jos lineaarinen rajoitus (2) pätee, vektori $\boldsymbol{\beta}$ *varioi siinä m-ulotteisessa vektorialiavaruudessa, jonka lineaarinen rajoitus* (2) *määrittelee*. Tämä aliavaruus on m-ulotteinen *taso* avaruudessa $\mathbb{R}^{k+1}$.

Rajoitettu pienimmän neliösumman estimaattori

Minimoidaan neliösumma

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

vektorin $\boldsymbol{\beta}$ suhteen, kun lineaarinen rajoitus

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

pätee. Käyttämällä *sidottujen ääriarvojen* etsimiseen tarkoitettua *Lagrangen kertojien menetelmää* saadaan regressiokertoimien vektorin $\boldsymbol{\beta}$ estimaattoriksi (joka siis toteuttaa rajoitukset $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$) estimaattori

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})$$

Estimaattoria $\mathbf{b}_R$ kutsutaan mallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}$ **rajoitetuksi** tai **sidotuksi pienimmän neliösumman (PNS-) estimaattoriksi**.

Rajoitetun PNS-estimaattorin ominaisuudet

Lineaarisen regressiomallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitetun PNS-estimaattorin* $\mathbf{b}_R$ keskeiset *stokastiset ominaisuudet* on esitetty seuraavassa lauseessa:

Lause 8.4.1.

Oletetaan, että yleisen lineaarisen mallin (1) standardioletukset (i)-(v) pätevät. Tällöin

(i) $\quad \mathrm{E}(\mathbf{b}_R) = \boldsymbol{\beta}$

(ii) $\quad \mathrm{Cov}(\mathbf{b}_R) = \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]$

jos lineaarinen rajoitus $\mathbf{R\beta} = \mathbf{r}$ pätee.

Perustelu:

(i) Suoraan laskemalla saadaan:

$$\begin{aligned}
\mathrm{E}(\mathbf{b}_R) &= \mathrm{E}[\mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})] \\
&= \mathrm{E}(\mathbf{b}) - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\mathrm{E}(\mathbf{b}) - \mathbf{r}) \\
&= \boldsymbol{\beta} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\boldsymbol{\beta} - \mathbf{r}) \\
&= \boldsymbol{\beta}
\end{aligned}$$

(ii) Oletetaan, että rajoitukset $\mathbf{R\beta} = \mathbf{r}$ pätevät. Merkitsemällä

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

voidaan rajoitetun PNS-estimaattorin $\mathbf{b}_R$ lauseke kirjoittaa muotoon

$$\mathbf{b}_R = \mathbf{b} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}(\mathbf{Rb} - \mathbf{r})$$

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 $\mathbf{R\beta} = \mathbf{r}$, jolloin $\mathbf{b}_R$ on *harhaton* parametrivektorille $\boldsymbol{\beta}$, niin

$$\begin{aligned}
\mathrm{Cov}(\mathbf{b}_R) &= \mathrm{E}\{[(\mathbf{b}_R - \mathrm{E}(\mathbf{b}_R)][(\mathbf{b}_R - \boldsymbol{\beta})]'\} \\
&= \mathrm{E}[(\mathbf{b}_R - \boldsymbol{\beta})(\mathbf{b}_R - \boldsymbol{\beta})'] \\
&= [(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}']\mathbf{X}'\mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\mathbf{X} \\
&\qquad \times[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}']\mathbf{X}'\mathbf{X} \\
&\qquad \times[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - \mathbf{C}(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C})^{-1}\mathbf{C}'] \\
&= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\,\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]
\end{aligned}$$

■

Huomautus:

Lauseen 8.4.1. kohdan (i) mukaan rajoitettu PNS-estimaattori $\mathbf{b}_R$ on lineaarisen rajoituksen $\mathbf{R\beta} = \mathbf{r}$ pätiessä regressiokertoimien vektorin $\boldsymbol{\beta}$ *harhaton estimaattori*.

Lause 8.4.2.

Oletetaan, että yleisen lineaarisen mallin (1) standardioletuksien (i)-(v) lisäksi normaalisuusoletus (vi) pätee. Tällöin

$$\mathbf{b}_R \sim \mathrm{N}_{k+1}(\boldsymbol{\beta},\sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}])$$

jos lineaarinen rajoitus $\mathbf{R\beta} = \mathbf{r}$ pätee.

Perustelu:

Lause 8.4.2. seuraa suoraan lauseesta 8.4.1., koska rajoitettu PNS-estimaattori

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})$$

on multinormaalisen satunnaismuuttujan $\mathbf{y}$ lineaarimuunnoksena multinormaalinen.

■

Rajoitetun PNS-estimaattorin hyvyys

Olkoon

$$\mathbf{b}_R = \mathbf{b} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{Rb} - \mathbf{r})$$

lineaarisen regressiomallin (1) regressiokertoimien vektorin $\boldsymbol{\beta}$ *rajoitettu PNS-estimaattori*, missä

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

on vektorin $\boldsymbol{\beta}$ *tavallinen PNS-estimaattori*.

Koska

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon}$$

saadaan yhtälö

$$\mathbf{b}_R - \boldsymbol{\beta} = [(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{X}'\boldsymbol{\varepsilon}$$

jos lineaarinen rajoitus

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

pätee. Tällöin

$$\begin{aligned}\mathrm{Cov}(\mathbf{b}_R) &= \mathrm{E}[(\mathbf{b}_R - \boldsymbol{\beta})(\mathbf{b}_R - \boldsymbol{\beta})'] \\ &= \sigma^2[(\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}]\end{aligned}$$

Tästä nähdään välittömästi, että *rajoitettu PNS-estimaattori* $\mathbf{b}_R$ *on lineaarisen rajoituksen* $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ *pätiessä parempi kuin tavallinen PNS-estimaattori* $\mathbf{b}$, koska

$$\mathrm{Cov}(\mathbf{b}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

ja matriisi

$$\mathrm{Cov}(\mathbf{b}) - \mathrm{Cov}(\mathbf{b}_R) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

on *ei-negatiivisesti definiitti*.

8.5. Bayeslainen yleinen lineaarinen malli

Oletukset

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jossa $\mathbf{X}$ on $n\times(k + 1)$-matriisi ja

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0},\sigma^2\mathbf{I})$$

Oletetaan seuraavassa yksinkertaisuuden vuoksi, että jäännösvarianssi,σ^2 on tunnettu.

Lineaarisen mallin ilmaisema *otosinformaatio* voidaan esittää muodossa

$$\mathbf{y}|\boldsymbol{\beta} \sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I})$$

Oletetaan, että parametrista $\boldsymbol{\beta}$ on *prioritietoa*, joka voidaan esittää muodossa

$$\mathbf{R}\boldsymbol{\beta} \sim \mathrm{N}_m(\mathbf{r},\sigma^2/c\mathbf{W})$$

jossa $\mathbf{R}$ on täysiasteinen $m\times(k + 1)$-matriisi, $m \le k + 1$. Parametri c kontrolloi *otostiedon ja prioritiedon suhdetta*.

Soveltamalla *Bayesin kaavaa* saadaan parametrin $\boldsymbol{\beta}$ *posteriorijakaumaksi* normaalijakauma:

$$\boldsymbol{\beta}|\mathbf{y} \sim \mathrm{N}_{k+1}(\mathbf{b}_R(c),\sigma^2\mathbf{Z}_R(c))$$

jossa

$$\mathbf{b}_R(c) = \mathbf{Z}_R(c)(\mathbf{X'y} + c\mathbf{R'W}^{-1}\mathbf{r})$$

ja

$$\mathbf{Z}_R(c) = (\mathbf{X'X} + c\mathbf{R'W}^{-1}\,\mathbf{R})^{-1}$$

Parametrin $\boldsymbol{\beta}$ **Bayes-estimaattoriksi** voidaan valita *posteriorijakauman odotusarvo*

$$\mathbf{b}_R(c) = (\mathbf{X'X} + c\mathbf{R'W}^{-1}\,\mathbf{R})^{-1}(\mathbf{X'y} + c\mathbf{R'W}^{-1}\mathbf{r}) \tag{1}$$

Estimaattori (1) on samaa muotoa kuin ns. **sekaestimaattori**, joka saadaan liittämällä *stokastinen prioritieto*

$$\mathbf{r} = \mathbf{R}\boldsymbol{\beta} + \boldsymbol{\delta}$$

$$\boldsymbol{\delta} \sim \mathrm{N}_m(\mathbf{r},\sigma^2/c\mathbf{W})$$

otosinfomaatioon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0},\sigma^2\mathbf{I})$$

lisähavaintoina ja soveltamalla näin saatuun lineaariseen malliin *yleistettyä pienimmän neliösumman menetelmää* (ks. kappale 8.3).

Estimaattori (1) saadaan myös *minimoimalla*

$$f(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + c(\mathbf{r} - \mathrm{R}\boldsymbol{\beta})'\mathbf{W}^{-1}(\mathbf{r} - \mathbf{R}\boldsymbol{\beta})$$

parametrin $\boldsymbol{\beta}$ *suhteen*. Tämä minimointitehtävä on *sakotetunpienimmän neliösumman menetelmän* muoto, joka voidaan tulkita *huonosti asetettujen ongelmien* (*engl*. ill-posed problem) ratkaisemisessa sovellettavan *regularisaatiomenetelmän* diskretisoinniksi.

Jos $c \to 0$, niin parametrin $\boldsymbol{\beta}$ *posteriorijakauman* rajajakaumaksi saadaan normaalijakauma

$$\mathrm{N}_{k+1}(\mathbf{b},\sigma^2\mathbf{U})$$

jossa

$$\mathbf{b} = \mathbf{UX'y}$$

$$\mathbf{U} = (\mathbf{X'X})^{-1}$$

Siten yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pararmetrin $\boldsymbol{\beta}$ *tavanomainen pienimmän neliösumman estimaattori* saadaan pararmetrin $\boldsymbol{\beta}$ Bayes-estimaattorina, kun priorijakaumana käytetään *epäinformatiivista priorijakaumaa*.

Jos $c \to +\infty$, niin parametrin $\boldsymbol{\beta}$ *posteriorijakauman* rajajakaumaksi saadaan normaalijakauma

$$\mathrm{N}_{k+1}(\mathbf{b}_R, \sigma^2\mathbf{Z}_R)$$

jossa

$$\mathbf{b}_R = \mathbf{b} - \mathbf{UR}'(\mathbf{RUR}')^{-1}(\mathbf{Rb} - \mathbf{r})$$

$$\mathbf{Z}_R = \mathbf{U} - \mathbf{UR}'(\mathbf{RUR}')^{-1}\mathbf{RU}$$

$$\mathbf{b} = \mathbf{UX}'\mathbf{y}$$

$$\mathbf{U} = (\mathbf{X}'\mathbf{X})^{-1}$$

Siten yleisen lineaarisen mallin

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

pararmetrin $\boldsymbol{\beta}$ *rajoitettu pienimmän neliösumman estimaattori* $\mathbf{b}_R$ (ks. kapaletta 8.4.) saadaan Bayes-estimaattorina, kun priorijakaumana käytetyn normaalijakauman

$$\mathrm{N}_m(\mathbf{r}, \sigma^2/c\mathbf{W})$$

annetaan konvergoida kohden yhden pisteen jakaumaa, jossa jakauman koko todennäkösyysmassa on keskittynyt pisteeseen $\mathbf{r}$.

8.6. Yleinen lineaarinen malli ja stokastiset selittäjät

Oletukset

Olkoon

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

standardioletukset (i)-(vi) toteuttava yleinen lineaarinen malli, jossa $\mathbf{X}$ on $n\times(k + 1)$-matriisi ja

$$\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$$

Oletuksen (i) mukaan matriisi $\mathbf{X}$ on *ei-satunnainen*. Korvataan oletus (i) nyt oletuksella

(i)´ Matriisi $\mathbf{X}$ on *satunnainen*

Huomautus:

Oletus (i)´ merkitsee sitä, että selittäjät $x_1, x_2, \ldots, x_k$ oletetaan *satunnaismuuttujiksi*.

Kiinteät ja satunnaiset selittäjät

Lineaarista regressiomallia (1) koskevissa standardioletuksissa selittäjien havaittujen arvojen muodostama matriisi $\mathbf{X}$ on oletettu *kiinteäksi* eli *ei-satunnaiseksi*. Tiukasti ottaen tämä *oletus voi päteä vain sellaisissa tilanteissa, joissa selittäjien arvot päästään valitsemaan*. Selittäjien arvot päästään valitsemaan *puhtaissa koeasetelmissa*, mutta muulloin oletus on vaikeasti perusteltavissa.

Tarkastelemme nyt siis tilannetta, jossa selittäjät ovat *stokastisia muuttujia* eli *satunnaismuuttujia*. Miten tämä vaikuttaa kappaleessa 1.esitettyihin lineaarisen regressiomallin estimointia koskeviin tuloksiin? Täydellisen vastauksen antaminen tähän kysymykseen on monimutkainen tehtävä eikä siihen tässä edes pyritä.

Jos sekä *selitettävä muuttuja y* että *selittäjät* $x_1, x_2, \ldots, x_k$ ovat *satunnaismuuttujia*, täydellisen kuvauksen niiden käyttäytymisestä antaa niiden *yhteisjakauma*. Muuttujan *y riippuvuutta* muuttujista $x_1, x_2, \ldots, x_k$ voidaan tutkia yhteisjakauman muodostamassa kehikossa tarkastelemalla muuttujan *y regressiofunktiota* eli *ehdollista odotusarvoa muuttujien* $x_1, x_2, \ldots, x_k$ suhteen.

Koska regressiofunktiot ovat yleensä *epälineaarisia*, joudutaan tällaisissa tilanteissa tavallisesti soveltamaan **epälineaarista regressioanalyysia**; epälineaarisen regressioanalyysin käsittely sivuutetaan tässä esityksessä.

Ehdollistaminen

Voidaan osoittaa, että *kaikki* kappaleissa 1. ja 2. *esitetyt lineaarisen regressiomallin estimointia ja testausta koskevat tulokset pätevät*, *jos seuraavat oletukset pätevät*:

(i)′ $\mathrm{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \mathbf{0}$

(ii)′ $\mathrm{Cov}(\boldsymbol{\varepsilon} \mid \mathbf{X}) = \sigma^2\mathbf{I}$

Näistä oletuksista seuraa:

(i)′′ $\mathrm{E}(\mathbf{y} \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$

(ii)′′ $\mathrm{Cov}(\mathbf{y} \mid \mathbf{X}) = \sigma^2\mathbf{I}$

Ehdon (i)′′ mukaan selitettävän muuttujan arvojen *ehdollinen odotusarvo* eli *regressiofunktio* on *lineaarinen*, kun ehdollistus tapahtuu selittävien muuttujien *havaittujen arvojen suhteen*.

Huomautus 1:

Koska moniulotteisten satunnaismuuttujien ehdolliset odotusarvot ovat yleisessä tapauksessa ehtomuuttujien *epälineaarisia funktioita*, *oletus regressiofunktion lineaarisuudesta on* stokastisten selittäjien tapauksessa *hyvin voimakas oletus*.

Huomautus 2:

Jos selitettävän muuttujan y ja selittäjien $x_1, x_2, \ldots, x_k$ yhteisjakauma on **multinormaalinen**, niin satunnaismuuttujan *y ehdollinen jakauma* satunnaismuuttujien $x_1, x_2, \ldots, x_k$ suhteen on *normaalinen*.

Lisäksi tällöin satunnaismuuttujan *y ehdollinen odotusarvo* satunnaismuuttujien $x_1, x_2, \ldots, x_k$ suhteen on *lineaarinen* ja satunnaismuuttujan *y ehdollinen varianssi* satunnaismuuttujien $x_1, x_2, \ldots, x_k$ suhteen on *vakio*.

Tällöin oletukset (i)′ ja (ii)′ pätevät ja voimme soveltaa kappaleissa 1. ja 2. esitettyä yleisen lineaarisen mallin tavanomaista estimointi- ja testiteoria. Tämä merkitsee sitä, että *stokastisten selittäjien tapauksessa multinormaalijakauman regressiofunktiot ja lineaariset regressiomallit kytkeytyvät toisiinsa*.

Lisätietoja multinormaalijakaumasta, sen ehdollisista jakaumista ja ehdollisista odotusarvoista sekä ehdollisten odotusarvojen estimoinnista: ks. **Monimuuttujamenetelmät: Multinormaalijakauma**.

Huomautus 3:

Aikasarjojen analyysissa ja *ekonometriassa* joudutaan soveltamaan myös sellaisia regressiomalleja, joissa selittäjät ovat *stokastisia* ja oletukset (i)′ ja (ii)′ *eivät päde*.

Tällaisissa tilanteissa PNS-menetelmä *ei välttämättä tuota harhattomia eikä edes tarkentuvia estimaattoreita regressiokertoimille*. Tällöin *PNS-menetelmää ei saa käyttää regressiokertoimien estimointiin*.

Sen sijaan *suurimman uskottavuuden menetelmä* tuottaa tavallisesti kelvolliset estimaattorit regressiokertoimille myös niissä tilanteissa, joissa PNS-menetelmää ei saa soveltaa.

Liite 1: Odotusarvovektori ja kovarianssimatriisi

Olkoot

$$x_1, x_2, \dots, x_p$$

satunnaismuuttujia, joiden *odotusarvot* ovat

$$\mathrm{E}(x_i) = \mu_i \,, \; i = 1, 2, \dots, p$$

ja *kovarianssit* ovat

$$\begin{aligned}\mathrm{Cov}(x_i, x_j) &= \mathrm{E}[(x_i - \mathrm{E}(x_i))(x_j - \mathrm{E}(x_j))] \\ &= \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \\ &= \sigma_{ij} \,, \; i = 1, 2, \dots, p, j = 1, 2, \dots, p\end{aligned}$$

Huomaa, että

$$\mathrm{Cov}(x_i, x_j) = \mathrm{E}(x_i x_j) - \mu_i \mu_j$$

ja

$$\mathrm{Cov}(x_i, x_i) = \sigma_{ii} = \mathrm{Var}(x_i) = \sigma_i^2$$

Olkoon

$$\mathbf{x} = (x_1, x_2, \dots, x_p) = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

satunnaismuuttujien $x_1, x_2, \dots, x_p$ muodostama p-vektori.

Satunnaisvektorin $\mathbf{x}$ *odotusarvovektori* on p-vektori

$$\mathrm{E}(\mathbf{x}) = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

jonka i. alkio

$$\mu_i = \mathrm{E}(x_i) \,, \; i = 1, 2, \dots, p.$$

on satunnaismuuttujan x_i *odotusarvo*.

Satunnaisvektorin **x** *kovarianssimatriisi* on (symmetrinen) $p\times p$-matriisi

$$\mathrm{Cov}(\mathbf{x}) = [\sigma_{ij}] = \mathbf{\Sigma}$$

jonka *i*. rivin ja *j* sarakkeen alkio

$$\sigma_{ij} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)] \text{ , } i, j = 1, 2, \ldots , p$$

on satunnaismuuttujien x_i ja x_j *kovarianssi*.

Huomaa, että

$$\begin{aligned}\mathrm{Cov}(\mathbf{x}) &= \mathrm{E}[(\mathbf{x}-\mathrm{E}(\mathbf{x}))(\mathbf{x}-\mathrm{E}(\mathbf{x}))'] \\ &= \mathrm{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})'] \\ &= \mathrm{E}(\mathbf{x}\mathbf{x}') - \boldsymbol{\mu}\boldsymbol{\mu}'\end{aligned}$$

Liite 2: Multinormaalijakauma

Määritelmä

Satunnaisvektori $\mathbf{x} = (x_1, x_2, \ldots , x_p)$ noudattaa *p-ulotteista multinormaalijakaumaa* parametrein

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$$

ja

$$\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma} > 0$$

jos sen *tiheysfunktio* on

$$f(\mathbf{x}) = (2\pi)^{-\frac{1}{2}p} |\mathbf{\Sigma}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})' \mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Merkintä:

$$\mathbf{x} \sim \mathrm{N}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$$

Satunnaisvektorin **x** *odotusarvovektorin*

$$\mathrm{E}(\mathbf{x}) = \boldsymbol{\mu}$$

i. alkio μ_i on satunnaismuuttujan x_i *odotusarvo*:

$$[\boldsymbol{\mu}]_i = \mu_i = \mathrm{E}(x_i) \text{ , } i = 1, 2, \ldots , p$$

Satunnaisvektorin **x** *kovarianssimatriisin*

$$\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma} = \mathrm{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})']$$

i. rivin ja *j*. sarakkeen alkio σ_{ij} on satunnaismuuttujien x_i ja x_j *kovarianssi*:

$$[\mathbf{\Sigma}]_{ij} = \sigma_{ij} = \mathrm{E}[(x_i - \mu_i)(x_j - \mu_j)]$$

Tiheysfunktion ominaisuudet

(i) p-ulotteisen multinormaalijakauman tiheysfunktio $f_{\mathbf{X}}(\mathbf{x})$ määrittelee *pinnan*

$$y = f_{\mathbf{X}}(\mathbf{x})$$

$(p + 1)$-ulotteisessa avaruudessa.

(ii) Multiplikatiivinen tekijä

$$(2\pi)^{-p/2} |\,\mathbf{\Sigma}\,|^{-1/2}$$

multinormaalijakauman tiheysfunktion lausekkeessa on *skaalaus-* tai *normeeraustekijä*, jonka tehtävänä on pitää huolta siitä, että pinnan

$$y = f_{\mathbf{X}}(\mathbf{x})$$

ja tason $y = 0$ rajoittaman kappaleen tilavuus = 1.

(iii) Pinnalla $y = f_{\mathbf{X}}(\mathbf{x})$ on yksikäsitteinen *maksimi* pisteessä $\boldsymbol{\mu}$.

(iv) Pinnan $y = f_{\mathbf{X}}(\mathbf{x})$ muodon määrää matriisin $\mathbf{\Sigma}^{-1}$ *neliömuoto*

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

(v) Neliömuoto

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

määrittelee pinnalle $y = f_{\mathbf{X}}(\mathbf{x})$ *tasa-arvoellipsoidit*

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

(vi) Ellipsoidien

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

yhteisenä *keskipisteenä* on piste $\boldsymbol{\mu}$.

(vii) Ellipsoidien

$$(\mathbf{x}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

pääakselit saadaan kovarianssimatriisin $\mathbf{\Sigma}$ *pääakseliesityksestä*

$$\mathbf{B\Lambda B}'$$

jossa $\mathbf{\Lambda}$ on matriisin $\mathbf{\Sigma}$ *ominaisarvojen* muodostama *diagonaalimatriisi* ja $\mathbf{B}$ on vastaavien *ominaisvektoreiden* muodostama *ortogonaalinen matriisi*, jossa ominaisvektorit ovat sarakkeina.

Matriisin $\mathbf{\Sigma}$ *ominaisvektorit* määräävät tasa-arvoellipsoidien pääakseleiden *suunnat* ja niiden *pituudet* ovat verrannollisia matriisin $\mathbf{\Sigma}$ *ominaisarvojen* neliöjuuriin.

Multinormaalijakauman karakterisointi 1

Jokainen multinormaalijakaumaa noudattava satunnaismuuttuja saadaan *lineaarimuunnoksella* riippumattomista *standardoitua normaalijakaumaa* noudattavista satunnaismuuttujista.

Kääntäen: jokainen multinormaalijakaumaa noudattava satunnaismuuttuja voidaan muuntaa *lineaarimuunnoksella* riippumattomiksi *standardoitua normaalijakaumaa* noudattaviksi satunnaismuuttujiksi.

Multinormaalijakauman karakterisointi 2

Jos *kaikki* satunnaisvektorin $\mathbf{X}$ *lineaarikombinaatiot* noudattavat (yksiulotteista) *normaalijakaumaa*, satunnaisvektori $\mathbf{X}$ noudattaa *multinormaalijakaumaa*.

Kääntäen: *multinormaalijakaumaa noudattavan satunnaisvektorin kaikki lineaarikombinaatiot ovat normaalisia*.

Muuttujien vaihto

Multinormaalisen satunnaismuuttujan *epäsingulaariset lineaarimuunnokset* noudattavat multinormaalijakaumaa.

Korreloimattomuus ja riippumattomuus

Satunnaismuuttujien *riippumattomuudesta seuraa aina* niiden *korreloimattomuus*. Sen sijaan *korreloimattomat satunnaismuuttujat eivät välttämättä ole riippumattomia*. Korreloimattomien satunnaismuuttujien välillä saattaa olla jopa eksakti (epälineaarinen) riippuvuus.

Mutta *jos satunnaismuuttujien yhteisjakaumana on multinormaalijakauma*, *niin satunnaismuuttujat ovat riippumattomia*, *jos ja vain jos ne ovat korreloimattomia*.

Reunajakaumat

Multinormaalijakaumaa noudattavan satunnaismuuttujan *kaikki reunajakaumat* ovat multinormaalijakaumia (tai normaalijakaumia).

Ehdolliset jakaumat

Multinormaalijakaumaa noudattavan satunnaismuuttujan *kaikki ehdolliset jakaumat* ovat multinormaalijakaumia (tai normaalijakaumia).