Lineaarinen malli

Pentti Saikkonen

Kevät 2007

Korjattu versio: Toukokuu 2012

Sisältö

1 Lineaarisen mallin määrittely

1.1 Yksinkertainen esimerkki

Tarkastellaan aluksi yksinkertaista esimerkkiä, joka havainnollistaa lineaarisen mallin ideaa. Oletetaan, että maaperältään homogeeninen pelto on jaettu samankokoisiin alueisiin, joita on n kappaletta ja joita on lannoitettu eri määrillä $x_1, ..., x_n$ samaa lannoitetta. Olkoot saadut satomäärät vastaavasti $y_1, ..., y_n$. Tavoitteena on rakentaa tilastollinen malli, jonka avulla voidaan selvittää lannoituksen vaikutus odotettavissa olevaan satomäärään.

Oletetaan, että havaitut satomäärät voidaan tulkita (riittävällä tarkkuudella) riippumattomien satunnaismuuttujien $Y_1, ..., Y_n$ havaituiksi arvoiksi. Mielenkiinto kohdistuu näiden satunnaismuuttujien odotusarvoihin $\mathsf{E}(Y_i) = \mu_i$, jotka riippuvat ennalta valituista ja siten ei-satunnaisista lannoitemääristä x_i $(i = 1, ..., n)$. Lineaarisessa mallissa tämä riippuvuus oletetaan lineaariseksi eli $\mu_i = \beta_1 + \beta_2 x_i$, jossa β_1 ja β_2 ovat tuntemattomia parametreja. Lisäksi oletetaan, etteivät lannoitemäärät vaikuta satunnaismuuttujien Y_i (tuntemattomiin) variansseihin eli oletetaan, että $\mathsf{Var}(Y_i) = \sigma^2$ pätee kaikilla $i = 1, ..., n$. Tilanne voidaan kuvata käyttäen yhtälöä

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \quad i = 1, ..., n, \tag{1.1}$$

jossa $\varepsilon_1, ..., \varepsilon_n$ ovat riippumattomia ei-havaittavia satunnaismuuttujia, joille pätee $\mathsf{E}(\varepsilon_i) = 0$ ja $\mathsf{Var}(\varepsilon_i) = \sigma^2$. Nämä satunnaismuuttujat voidaan tulkita havaituissa satomäärissä ilmeneväksi 'puhtaaksi satunnaisvaihteluksi', joka ei selity lannoitemäärällä.

Yhtälöä (1.1) nimitetään *yhden selittävän muuttujan lineaariseksi regressiomalliksi*. Jos asiaa tarkastellaan tiukasti tilastollisen päättelyn näkökulmasta, ei kysymyksessä ole vielä tilastollinen malli, jollainen vaatii havaintojen yhteistodennäköisyysjakauman ja parametriavaruuden spesifioinnin. Klassinen lineaarinen malli, jota tällä kurssilla tarkastellaan, olettaa normaalijakauman. Koska lannoitemäärät tulkitaan ei-satunnaisiksi, saadaan havaintojen yhteistodennäköisyysjakauma siten oletuksesta

$$Y_1, ..., Y_n \ \underline{\perp\!\!\!\perp}\ , \ Y_i \sim \mathsf{N}\left(\beta_1 + \beta_2 x_i, \sigma^2\right). \tag{1.2}$$

Vaikka (ainakin) oletus $\beta_1 > 0$ tuntuisi järkevältä, valitaan klassisen lineaarisen mallin mukaisesti parametriavaruudeksi $\beta_1, \beta_2 \in \mathbb{R}$, $\sigma^2 > 0$. Vaihtoehtoinen ja usein käytetty tapa spesifioida havaintojen yhteistodennäköisyysjakauma on yhtälön (1.1) täydentäminen oletuksella

$$\varepsilon_1, ..., \varepsilon_n \ \underline{\perp\!\!\!\perp}\ , \ \varepsilon_i \sim \mathsf{N}\left(0, \sigma^2\right). \tag{1.3}$$

Koska tässä tapauksessa satomäärää selittävän muuttujan eli lannoitemäärän arvot kiinnitettiin edeltä käsin, niiden tulkitseminen ei-satunnaisiksi on loogista. Jos kysymyksessä olisi ollut tilanne, jossa selittävän muuttujan havainnot olisi poimittu satunnaisotantaa käyttäen tai ne olisivat muuten 'satunnaisia', täytyisi tilastollinen malli periaatteessa laajentaa ja tarkastella satunnaisvektorien $[Y_i \ \ X_i]'$ $(i = 1, ..., n)$ yhteistodennäköisyysjakaumaa.[1] Myöhemmin todetaan, että sopivin oletuksin on loo-

[1] Matriisin transponointia merkitään pilkulla ja vektorit tulkitaan matriiseiksi, joissa on yksi sarake. Käytetyt matriisilaskennan merkinnät ja tulokset on koottu Liitteeseen B.

gista tarkastella selitettäviä muuttujia $Y_1, ..., Y_n$ ehdollisesti ehdolla selittävän muuttujan $X_1, ..., X_n$ saamat havaitut arvot $x_1, ..., x_n$, jolloin edellä esitetty malli soveltuu.

Käytännössä edellä tarkasteltuun malliin voisi olla aiheellista sisällyttää lannoitemäärän lisäksi myös muita satomäärää selittäviä muuttujia. Seuraavassa esitettävä yleinen lineaarinen malli ottaa tämän huomioon.

1.2 Yleinen lineaarinen malli

Yleisen lineaarisen mallin asetelma on, että analysoitavana on n:n havaintoyksikön aineisto, jonka muuttujista yksi on luonteeltaan selitettävä ja loput p sen vaihtelua selittäviä muuttujia. Kaaviona tilanne on seuraavanlainen.

Havaintoyksikkö	Selitettävä muuttuja; y	Selittävät muuttujat; $x_1, ..., x_p$
1	y_1	$x_{11}, ..., x_{1p}$
$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n1}, ..., x_{np}$

Lineaarisessa mallissa selittävän muuttujan vaikutus selitettävään muuttujaan oletetaan (tietyssä mielessä) lineaariseksi. Jos $Y_1, ..., Y_n$ ovat edellisen esimerkkitapauksen mukaisesti satunnaismuuttujia, joiden havaitut arvot ovat $y_1, ..., y_n$, niin mallin määritelmä voidaan perustaa yhtälöön

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n, \tag{1.4}$$

josta edellisen jakson yhtälö (1.1) saadaan ilmeisenä erikoistapauksena ($p = 2$, $x_{i1} = 1$ $\forall i$). Yhtälön (1.4) oikealla puolella selittävien muuttujien havaintoarvot x_{ij} ovat *ei-satunnaisia* tai *kiinteitä* lukuja, $\beta_1, ..., \beta_p$ ovat tuntemattomia parametreja ja ε_i on havaintoyksikköön i liittyvä ei-havaittava satunnaismuuttuja, joka kuvaa sitä osaa selitettävän muuttujan vaihtelusta, jota selittävät muuttujat tai niiden lineaarikombinaatio $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ ei kykene selittämään. Tästä syystä satunnaismuuttujia ε_i kutsutaan *virheiksi* tai *virhetermeiksi*. Lineaarikombinaatiota $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ kutsutaan puolestaan mallin *systemaattiseksi osaksi* tai *rakenteeksi*. Mallin lineaarisuus merkitsee sitä, että systemaattinen osa on *parametrien $\beta_1, ..., \beta_p$ lineaarinen funktio ja että virhetermi lisätään systemaattiseen osaan additiivisesti*. Mallin lineaarisuus sallii näin ollen esimerkiksi valinnan $x_{i2} = x_{i1}^2$ eli epälineaarisuuden selittävien muuttujien suhteen, kunhan lineaarisuus parametrien $\beta_1, ..., \beta_p$ suhteen säilyy.[2]

[2] Jos edellisen jakson esimerkkiin lisätään selittäjä x_i^2, voidaan mallissa ottaa huomioon se, että satomäärä pienentyy, jos lannoitetta käytetään liikaa.

Kuten edellisen jakson esimerkissäkin, täytyy yhtälöä (1.4) täydentää spesifioimalla (selitettävän muuttujan) havaintojen yhteistodennäköisyysjakauma ja parametriavaruus, jotta tilastollinen malli tulee määritellyksi. Havaintojen $Y_1, .., Y_n$ yhteistodennäköisyysjakauma tulee spesifioiduksi, kun virheisiin liitetään oletus (1.3). Jos merkitään $\mathbf{x}_i = [x_{i1} \ \cdots \ x_{ip}]'$ ja $\boldsymbol{\beta} = \left[\beta_1 \ \cdots \ \beta_p\right]'$, saadaan lineaariselle mallille siten määritelmä

$$Y_1, ..., Y_n \ \underline{\parallel} \ , \ Y_i \sim \mathsf{N}\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\right), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0. \tag{1.5}$$

Vaihtoehtoinen määritelmä saadaan liittämällä yhtälöön (1.4) oletukset (1.3) ja $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$. Kuten edellä esitetystä ilmenee, voidaan lineaarista mallia luonnehtia malliksi havaintojen odotusarvolle. Tästä johtuen on parametrivektori $\boldsymbol{\beta}$ ja sitä koskeva tilastollinen päättely ensisijaisen mielenkiinnon kohteena, kun taas parametri σ^2 on luonteeltaan kiusaparametri.

Edellä esitetty lineaarisen mallin määrittely vastaa siis tilastollisessa päättelyssä käytettyä tilastollisen mallin määrittelyä. Kuten aiemmin vihjattiin, näkee nimitystä lineaarinen malli käytettävän usein myös löyhemmässä mielessä. Erityisesti havaintojen tai virheiden yhteistodennäköisyysjakaumaa ja parametriavaruutta ei aina spesifioida (ainakaan eksplisiittisesti) ja joskus riippumattomuuden asemesta oletetaan vain korreloimattomuus.

Lineaarinen malli voidaan esittää kätevästi matriisimerkinnöin, joita tarvitaan myös mallin teorian kehittelyssä. Yhtälö (1.4) voidaan kirjoittaa

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

eli

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.6}$$

Virheitä koskeva oletus (1.3) voidaan ilmaista vaatimalla, että satunnaisvektori $\boldsymbol{\varepsilon}$ noudattaa multinormaalijakaumaa odotusarvona nolla ja kovarianssimatriisina $\sigma^2\mathbf{I}_n$ eli symbolein $\boldsymbol{\varepsilon} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$, jossa $\mathbf{I}_n$ on $(n \times n)$ yksikkömatriisi.[3] Nyt malli voidaan määritellä lyhyesti kirjoittamalla

$$\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0. \tag{1.7}$$

Ellei toisin mainita, liitetään malliin lisäksi oletus

$$\mathsf{r}\left(\mathbf{X}\right) = p \tag{1.8}$$

eli matriisin $\mathbf{X}$ $(n \times p)$ oletetaan olevan täyttä sarakeastetta, jolloin pätee erityisesti $n \geq p$ (ja käytännössä $n > p$). Tämä takaa sen, että odotusarvovektorilla $\boldsymbol{\mu} = \mathsf{E}(\mathbf{Y})$ on yksikäsitteinen esitys $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ ($\mathsf{r}\left(\mathbf{X}\right) = p \Rightarrow \mathsf{r}\left(\mathbf{X}'\mathbf{X}\right) = p \Rightarrow \boldsymbol{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\mu}$). Kyse on siten identifiointiehdosta, joka takaa parametrivektorin $\boldsymbol{\beta}$ yksikäsitteisyyden ($\mathbf{X}\boldsymbol{\beta}^{(1)} = \mathbf{X}\boldsymbol{\beta}^{(2)} \Leftrightarrow \mathbf{X}(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(2)}) = \mathbf{0} \Leftrightarrow \boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}^{(2)}$, kun $\mathsf{r}\left(\mathbf{X}\right) = p$).

[3] Jos tarkasteltavan multimormaalijakauman dimensio on aiheellista merkitä näkyviin, se osoitetaan alaindeksillä eli esimerkiksi $\mathsf{N}_n\left(\cdot, \cdot\right)$.

Jos $\mathsf{r}(\mathbf{X}) < p$, voidaan joku (tai jotkut) matriisin $\mathbf{X}$ sarakkeet lausua muiden lineaarikombinaationa ja saada odotusarvovektorille esitys $\boldsymbol{\mu} = \mathbf{X}_*\boldsymbol{\beta}_*$, jossa matriisi $\mathbf{X}_*$ on täyttä sarakeastetta. Esimerkiksi tapauksessa $p = 3$ ja $x_{i3} = x_{i1} + x_{i2}$, pätee $\mu_i = x_{i1}\beta_1 + x_{i2}\beta_2 + x_{i3}\beta_3 = x_{i1}(\beta_1 + \beta_3) + x_{i2}(\beta_2 + \beta_3)$.

Mallin parametreja $\boldsymbol{\beta}$ ja σ^2 koskevan tilastollisen päättelyn kannalta on riittävää esittää malli muodossa $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ ilman, että virhetermeistä puhutaan mitään. Joissakin tapauksissa tähän esitykseen voidaan päätyä luontevasti lähtemällä havaintojen riippumattomuudesta ja oletuksesta $Y_i \sim \mathsf{N}\left(\mu_i, \sigma^2\right)$, jossa odotusarvoille μ_i voidaan tutkittavan ilmiön taustateorian perusteella olettaa lineaarinen esitys $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ $(i = 1, ..., n)$. Usein virhetermeillä on kuitenkin luonteva tulkinta mittaus- tai selitysvirheinä, jolloin niiden käyttäminen mallin motivoinnissa ja esittämisessä on myös luontevaa.

Mallin virhetermeillä on myös toinen motivaatio, jonka tarkastelemiseksi otetaan käyttöön käsitteet sovite ja residuaali, jotka ovat mallin systemaattisen osan $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ ja virhetermin ε_i empiirisiä vastineita. Jos $\hat{\boldsymbol{\beta}}$ on parametrin $\boldsymbol{\beta}$ estimaatti, niin (havaintoyksikköön i liittyvä) *sovite* on $\hat{\mu}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ ja (havaintoyksikköön i liittyvä) *residuaali* on $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ (sovitteelle käytetään myös merkintää $\hat{y}_i$). On selvää, että residuaalit $\hat{\varepsilon}_1, ..., \hat{\varepsilon}_n$ sisältävät informaatiota parametrista σ^2 ja ovat sikäli relevantteja. Lisäksi ne sisältävät informaatiota mallin oletusten mahdollisesta paikkansapitämättömyydestä eli esimerkiksi varianssien $\mathsf{Var}(\varepsilon_i) = \mathsf{Var}(Y_i)$ vaihtelusta sekä virheiden ε_i ja siten havaintojen Y_i riippuvuudesta tai ei-normaalisuudesta. Yksinkertaisimpia malleja lukuun ottamatta näitä kysymyksiä on helpompi tutkia residuaalien kuin alkuperäisten havaintojen avulla. Tällä kurssilla näitä tärkeitä kysymyksiä ei kuitenkaan ehditä juurikaan käsitellä. Todettakoon kuitenkin, että 'kohtuullisen pieni' poikkeama normaalisuudesta ei ole tuhoisaa, sillä esitettävät teoreettiset tulokset voidaan perustella asymptoottisina approksimaatioina myös ilman normaalisuusoletusta. Selvästi ei-normaalisiin tilanteisiin tällä kurssilla tarkasteltavaa mallia ei kuitenkaan pidä mennä soveltamaan.[4]

1.3 Lineaarisen mallin erikoistapauksia

Kuten edellä todettiin, saadaan jaksossa 1.1 tarkasteltu yhden selittäjän lineaarinen regressiomalli yleisen lineaarisen mallin erikoistapauksena. Vielä yksinkertaisempi erikoistapaus on malli

$$Y_1, ..., Y_n \quad \underline{\parallel} \quad , \; Y_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$$

eli *riippumaton otos normaalijakaumasta.* Tähän malliin päädytään valitsemalla yleisessä mallissa $p = 1$, $\beta_1 = \mu$ ja $x_{i1} = 1$, $i = 1, ..., n$. Matriisiksi $\mathbf{X}$ tulee siten

$$\mathbf{X} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} = \mathbf{1}_n \quad (n \times 1).$$

[4] Esimerkiksi selitettävän muuttujan saadessa diskreettejä arvoja, tarjoavat ns. yleistetyt lineaariset mallit usein parempia vaihtoehtoja.

Tätä mallia ja sen parametrien estimointia ja testausta on tarkasteltu tilastollisen päättelyn kurssilla. Myöhemmin nähdään, miten nämä estimointi- ja testausongelmat voidaan ratkaista lineaarisen mallin avulla.

Edellisen esimerkin yleistys on *kahden odotusarvoltaan (mahdollisesti) poikkeavan riippumattoman normaalisen otoksen malli* eli

$$Y_1, ..., Y_n \quad \underline{\parallel} \ , \ Y_i \sim \begin{cases} \mathsf{N}\left(\mu_1, \sigma^2\right), & \text{kun } i = 1, ..., n_1 \\ \mathsf{N}\left(\mu_2, \sigma^2\right), & \text{kun } i = n_1 + 1, ..., n_1 + n_2 = n. \end{cases}$$

Tämä malli saadaan yleisestä mallista valitsemalla $p = 2$, $\beta_i = \mu_i$ $(i = 1, 2)$ ja

$$\mathbf{X} \quad = \quad \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{bmatrix}.$$

Mielenkiinnon kohteena on usein hypoteesi $\mu_1 = \mu_2$ ja luottamusvälin muodostaminen erotukselle $\mu_1 - \mu_2$. Konkreettisena esimerkkitilanteena voisi olla kahden vehnälajikkeen A_1 ja A_2 satoisuuden tutkiminen, kun niitä viljellään samoissa olosuhteissa. Huomaa, että tässä selittävät muuttujat ovat ryhmää osoittavia indikaattoreita. Sama pätee niiden seuraavassa tarkasteltaviin yleistyksiin.

Edellinen esimerkki voidaan yleistää koskemaan kahta useampaa otosta, jolloin tarkasteltavia vehnälajikkeita voi olla useita. Tällöin kysymyksessä on ns. *yksisuuntainen varianssianalyysimalli*, jossa havaintoina on p riippumatonta otosta jakaumista $\mathsf{N}\left(\mu_j, \sigma^2\right)$ $(j = 1, ..., p)$ ja kiinnostuksen kohteena on odotusarvoissa $\mu_1, .., \mu_p$ mahdollisesti ilmenevät erot. Toinen edellisen esimerkin yleistys on ns. *kaksisuuntainen varianssianalyysimalli*, jota voidaan käyttää tilanteessa, jossa vehnälajikkeita A_1 ja A_2 lannoitetaan kahta eri lannoitetta B_1 ja B_2 käyttäen. Tällöin havainnot ovat peräisin neljästä ryhmästä ja mallin avulla voidaan tutkia onko satomäärissä eroja eri ryhmien välillä ja johtuvatko mahdolliset erot vehnälajikkeesta, lannoitteesta vai niiden yhteisvaikutuksesta.

Edellisessä jaksossa tarkastellun yhden selittäjän lineaarisen regressiomallin ilmeinen yleistys on *usean selittäjän lineaarinen regressiomalli*

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n,$$

johon liitetään oletukset (1.3) ja $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$. Tässä mallissa matriisi $\mathbf{X}$ on kuten yleisen mallin tapauksessa lukuun ottamatta sitä, että $x_{i1} = 1$ kaikilla $i = 1, ..., n$. Parametria β_1 kutsutaan *(regressio)vakioksi* ja parametreja β_j $(j = 2, .., p)$ *regressiokertoimiksi*. Yksittäinen regressiokerroin kuvaa paljonko selitettävä muuttuja tai täsmällisemmin sen odotusarvo muuttuu, kun j:nnen selittäjän arvo muuttuu yhden yksikön muiden selittäjien arvojen pysyessä muuttumattomina. Tämä kuvastaa yhtä mallin käyttötarkoitusta, joka on selitettävän muuttujan ja selittävien muuttujien

välisen riippuvuuden tiivistetty kuvaaminen eli *selittäminen*. Mallia voidaan käyttää myös selitettävän muuttujan arvojen *ennustamiseen* ja *kontrolloimiseen* (olettaen, että selittävien muuttujien arvoihin voidaan vaikuttaa).

Samassa mallissa voi olla sekä kvantitatiivisia selittäjiä että kvalitatiivisia ryhmää osoittavia indikaattoreita. Jos esimerkiksi tutkitaan vehnälajikkeiden A_1 ja A_2 satoisuutta ja molempia lannoitetaan samaa lannoitetta käyttäen, päädytään malliin, jossa on kahden ryhmää osoittavan indikaattorin lisäksi yksi kvantitatiivinen selittäjä. Oletetaan, että n_1 ensimmäistä satomäärää liittyy lajikkeeseen A_1 ja loput $n_2 = n - n_1$ lajikkeeseen A_2 ja olkoon x_i jälleen lannoitemäärä havaintoyksikössä i. Tällöin malli on

$$Y_1, ..., Y_n \quad \underline{\parallel} \quad , \; Y_i \sim \begin{cases} \mathsf{N}\left(\beta_1 + \beta_3 x_i, \sigma^2\right), & \text{kun } i = 1, ..., n_1 \\ \mathsf{N}\left(\beta_2 + \beta_3 x_i, \sigma^2\right), & \text{kun } i = n_1 + 1, ..., n_1 + n_2 = n. \end{cases}$$

Matriisiksi $\mathbf{X}$ tulee

$$\mathbf{X} \quad = \quad \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{n_1} \\ 0 & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_n \end{bmatrix}.$$

Huomaa, että tässä mallissa lannoitteen vaikutus molempiin vehnälajikkeisiin oletetaan samaksi.

Todetaan vielä, että joissakin tapauksissa lineaarista mallia voidaan käyttää, vaikka alkuperäinen malli olisikin epälineaarinen. Tyypillisin esimerkki on

$$Y_i = e^{\beta_1} x_{i2}^{\beta_2} \cdots x_{ip}^{\beta_p} e^{\varepsilon_i}, \quad i = 1, ..., n,$$

jossa muuttujat oletetaan positiivisiksi. Ottamalla (luonnollinen) logaritmi puolittain päädytään yhtälöön

$$\log Y_i = \beta_1 + \beta_2 \log x_{i2} + \cdots + \beta_p \log x_{ip} + \varepsilon_i, \quad i = 1, ..., n,$$

josta oletuksilla (1.3) ja $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$ saadaan (merkintöjä vaille) usean selittäjän lineaarinen regressiomalli. Tällaiset ns. multiplikatiiviset mallit ovat tavallisia taloudellisissa sovelluksissa mallinnettaessa esimerkiksi jonkin tuotteen kysyntää.

2 Lineaarisen mallin parametrien estimointi

2.1 Suurimman uskottavuuden (SU) estimointi

Mallioletuksen (1.7) mukaan $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$, joten käyttäen multinormaalijakauman tiheysfunktion kaavaa nähdään suoraan, että havaintojen yhteistiheysfunktio on[5]

$$f_{\mathbf{Y}}\left(\mathbf{y}; \boldsymbol{\beta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)\right\}.$$

Parametrien $\boldsymbol{\beta}$ ja σ^2 log-uskottavuusfunktioksi saadaan siten

$$l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\right) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathsf{S}\left(\boldsymbol{\beta}\right),$$

jossa

$$\mathsf{S}\left(\boldsymbol{\beta}\right) = \left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2$$

on ns. *jäännösneliösummafunktio*. Parametrin $\boldsymbol{\beta}$ SU-estimaatti $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}\left(\mathbf{y}\right)$ löydetään minimoimalla jäännösneliösummafunktio $\mathsf{S}\left(\boldsymbol{\beta}\right)$, minkä jälkeen parametrin σ^2 SU-estimaatti $\hat{\sigma}^2 = \hat{\sigma}^2\left(\mathbf{y}\right)$ saadaan kaavalla

$$\hat{\sigma}^2 = \frac{1}{n}\mathsf{S}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})$$

tai, käyttäen residuaaleja $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Edellä sanottu voidaan perustella epäyhtälöillä

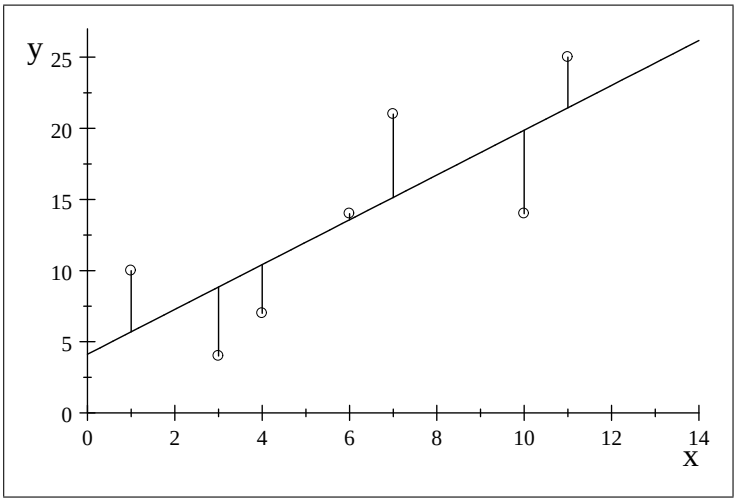
$$\begin{aligned} l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\right) &\leq -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathsf{S}(\hat{\boldsymbol{\beta}}) \\ &\leq -\frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\mathsf{S}(\hat{\boldsymbol{\beta}}) \\ &= l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2; \mathbf{y}), \end{aligned}$$

jotka pätevät kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p$ ja $\sigma^2 > 0$. Näistä ensimmäinen perustuu estimaatin $\hat{\boldsymbol{\beta}}$ määritelmään ja toinen nähdään maksimoimalla edeltävä lauseke σ^2:n suhteen (yksityiskohdat jätetään tehtäväksi).

Estimaatin $\hat{\boldsymbol{\beta}}$ lauseke voidaan johtaa joko geometrisesti tai derivoimalla jäännösneliösummafunktiota. Palataan edelliseen hieman myöhemmin ja käytetään tässä jälkimmäistä tapaa, jossa suoritettavia laskelmia voidaan käyttää parametrin $\boldsymbol{\beta}$ havaitun informaatiomatriisin johtamisessa. Suoraviivaisella derivoinnilla nähdään, että

$$\partial\mathsf{S}\left(\boldsymbol{\beta}\right)/\partial\boldsymbol{\beta} = \begin{bmatrix} \partial\mathsf{S}\left(\boldsymbol{\beta}\right)/\partial\beta_1 \\ \vdots \\ \partial\mathsf{S}\left(\boldsymbol{\beta}\right)/\partial\beta_p \end{bmatrix} = -2\sum_{i=1}^{n}\mathbf{x}_i\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right) = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$$

[5] Sama tulos voidaan johtaa helposti myös kirjoittaen oletuksen (1.5) nojalla $f_{\mathbf{Y}}\left(\mathbf{y}; \boldsymbol{\beta}, \sigma^2\right) = f_{\mathbf{y}_1}\left(\mathbf{y}_1; \boldsymbol{\beta}, \sigma^2\right)\cdots f_{\mathbf{y}_n}\left(\mathbf{y}_n; \boldsymbol{\beta}, \sigma^2\right)$, jossa $f_{\mathbf{y}_i}\left(\mathbf{y}_i; \boldsymbol{\beta}, \sigma^2\right)$ on havainnon Y_i tiheysfunktio.



Kuva 2.1. PNS-menetelmän havainnollistus yhden selittäjän regressiomallin (1.1) tapauksessa.

Välttämätön ehto minimille on $\partial \mathsf{S}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} = \mathbf{0}$, mikä johtaa ns. *normaaliyhtälöihin*

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}. \tag{2.1}$$

Oletuksesta $\mathsf{r}(\mathbf{X}) = p$ (ks. (1.8), s. 3) seuraa tunnetusti $\mathsf{r}(\mathbf{X}'\mathbf{X}) = p$, joten normaaliyhtälöillä on yksikäsitteinen ratkaisu

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

Se, että kysymyksessä todella on minimipiste, voidaan todeta eri tavoin. Differentiaali- ja integraalilaskentaan perustuvassa tavassa todetaan, että toisten derivaattojen matriisi $\partial^2 \mathsf{S}(\boldsymbol{\beta}) / \partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = \left[\partial^2 \mathsf{S}(\boldsymbol{\beta}) / \partial \beta_i \partial \beta_j\right] = 2\mathbf{X}'\mathbf{X}$ on positiivisesti definiitti, mistä haluttu tulos seuraa.[6] Seuraavassa jaksossa esitetään geometrinen perustelu.

Ilmeisistä syistä johtuen sanotaan parametrin $\boldsymbol{\beta}$ SU-estimaattia $\hat{\boldsymbol{\beta}}$ *pienimmän neliösumman (PNS) estimaatiksi*. Yhden selittäjän regressiomallin (1.1) tapauksessa PNS-estimaatti minimoi oheisen kuvan pystysuorien janojen pituuksien neliösumman.

2.1.1 PNS-estimointi geometrisesti

Eräs tapa havainnollistaa PNS-estimaattia on johtaa se geometrisesti. Olkoon $\|\mathbf{a}\| = (\mathbf{a}'\mathbf{a})^{1/2} = \left(a_1^2 + \cdots + a_n^2\right)^{1/2}$ vektorin $\mathbf{a} = [a_1 \ \cdots \ a_n]'$ (Euklidinen) normi ja $\mathcal{R}(\mathbf{X})$ matriisin $\mathbf{X}$ $(n \times p)$ sarakeavaruus.[7] Koska $\mathsf{S}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$, on PNS-estimoinnissa kysymys normin $\|\mathbf{y} - \boldsymbol{\mu}\|$ minimoinnista ehdolla $\boldsymbol{\mu} \in \mathcal{R}(\mathbf{X})$. Lineaarialgebrasta tiedetään, että minimi saavutetaan valitsemalla $\boldsymbol{\mu} = \hat{\boldsymbol{\mu}}$ siten, että erotus $\mathbf{y} - \hat{\boldsymbol{\mu}}$ tulee ortogonaaliseksi avaruutta $\mathcal{R}(\mathbf{X})$ tai yhtäpitävästi matriisin $\mathbf{X}$ sarakkeita vastaan. Toisin

[6] Määritelmän mukaan symmetrinen matriisi $\mathbf{A}$ on positiivisesti definiitti (merkintään $\mathbf{A} > \mathbf{0}$), jos $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ kaikilla (yhteensopivilla) vektoreilla $\mathbf{x} \neq \mathbf{0}$.

[7] $\mathcal{R}(\mathbf{X})$ on $\mathbb{R}^n$:n p-ulotteinen aliavaruus ja sisältää vektorit, jotka voidaan lausua $\mathbf{X}$:n sarakkeiden lineaarikombinaationa eli $\mathcal{R}(\mathbf{X}) = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z} = \mathbf{X}\mathbf{b} \text{ jollain } \mathbf{b} \in \mathbb{R}^p\}$.

sanoen,

$$\mathbf{X}'(\mathbf{y} - \hat{\boldsymbol{\mu}}) = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{X}'\hat{\boldsymbol{\mu}} = \mathbf{X}'\mathbf{y}.$$

Vektorin $\hat{\boldsymbol{\mu}}$ tiedetään olevan $\mathbf{y}$:n yksikäsitteinen ortogonaalinen projektio avaruudelle $\mathcal{R}(\mathbf{X})$. Koska matriisin $\mathbf{X}$ sarakkeet ovat oletuksen mukaan lineaarisesti riippumattomia (eli vapaita), on olemassa yksikäsitteinen vektori $\hat{\boldsymbol{\beta}}$ siten, että $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Sijoittamalla tämä edellä johdettuun yhtälöön, saadaan $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, joten $\hat{\boldsymbol{\beta}}$ on sama normaaliyhtälöiden yksikäsitteinen ratkaisu kuin aikaisemminkin ja se minimoi jäännösneliösummafunktion $\mathsf{S}(\boldsymbol{\beta})$.

PNS-estimoinnissa selitettävän muuttujan vektori $\mathbf{y}$ hajotetaan kahteen osaan:

$$\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}, \tag{2.2}$$

jossa $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ $(= \hat{\boldsymbol{\mu}})$ on sovite eli estimoitu systemaattinen osa ja $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ on residuaali eli estimoitu satunnainen osa. Sijoittamalla normaaliyhtälöihin (2.1) $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ nähdään, että

$$\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}, \tag{2.3}$$

mistä seuraa $\hat{\mathbf{y}}'\hat{\boldsymbol{\varepsilon}} = 0$ eli sovitteen ja residuaalin ortogonaalisuus. Lisäksi, jos $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, niin

$$\hat{\mathbf{y}} = \mathbf{P}\mathbf{y} \quad \text{ja} \quad \hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{P})\mathbf{y}. \tag{2.4}$$

Matriisi $\mathbf{P}$ on (ortogonaalinen) projektio, joka projisoi $\mathbb{R}^n$:n vektorit matriisin $\mathbf{X}$ p-ulotteiselle sarakeavaruudelle $\mathcal{R}(\mathbf{X})$.[8] Matriisi $\mathbf{I}_n - \mathbf{P}$ on myös projektio. Se projisoi $\mathbb{R}^n$:n vektorit avaruuden $\mathcal{R}(\mathbf{X})$ ortogonaaliselle komplementille $\mathcal{R}(\mathbf{X})^{\perp}$, joka on $\mathbb{R}^n$:n $(n-p)$-ulotteinen aliavaruus ja sisältää vektorit, jotka ovat ortogonaalisia $\mathbf{X}$:n sarakkeita vastaan. PNS-estimoinnin tuloksena saatavassa hajotelmassa $\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}$ selitettävän muuttujan vektori $\mathbf{y}$ tulee siis esitetyksi yksikäsitteisesti kahden ortogonaalisen vektorin $\hat{\mathbf{y}} \in \mathcal{R}(\mathbf{X})$ ja $\hat{\boldsymbol{\varepsilon}} \in \mathcal{R}(\mathbf{X})^{\perp}$ summana (vrt. vastaava lineaarialgebran kohtisuoria projektioita koskeva tulos).

2.1.2 PNS-estimointi ja selitysaste

Hieman toisenlainen näkökulma PNS-estimointiin saadaan hajottamalla selitettävän muuttujan vaihtelu kahteen osaan. Kun vaihtelua mitataan neliösummalla, saadaan vektorien $\hat{\mathbf{y}}$ ja $\hat{\boldsymbol{\varepsilon}}$ ortogonaalisuutta käyttäen

$$\sum_{i=1}^{n} y_i^2 = \mathbf{y}'\mathbf{y} = (\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}})'(\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}) = \hat{\mathbf{y}}'\hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$$

eli

$$\|\mathbf{y}\|^2 = \|\hat{\mathbf{y}}\|^2 + \|\hat{\boldsymbol{\varepsilon}}\|^2. \tag{2.5}$$

On intuitiivisesti selvää, että mallin antama selitys selitettävälle muuttujalle on sitä 'parempi' mitä suurempi oikean puolen ensimmäinen termi on suhteessa vasemman

[8] Projektiomatriisi on määritelmän mukaan neliömatriisi, joka on symmetrinen ja idempotentti eli $\mathbf{P}$ toteuttaa ehdot $\mathbf{P}' = \mathbf{P} = \mathbf{P}^2$ $(= \mathbf{PP})$.

puolen termiin.[9] Seuraavassa tätä ideaa tarkastellaan lähemmin mallissa, jossa on vakio.

Tarkastellaan siis malliyhtälöä

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n.$$

Koska vakiota on vaikea ajatella 'varsinaisena' selittäjänä, mitataan vaihtelua tässä tapauksessa yleensä keskistettyjä havaintoja käyttäen (eli havainnot mitataan poikkeamina keskiarvostaan). Koska matriisin $\mathbf{X}$ ensimmäinen sarake on nyt ykkösvektori $\mathbf{1}_n$, on edellä johdetun yhtälön $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ perusteella $\mathbf{1}_n'\hat{\boldsymbol{\varepsilon}} = \hat{\varepsilon}_1 + \cdots + \hat{\varepsilon}_n = 0$. Tästä ja yhtälöstä (2.2) seuraa

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

Määritellään seuraavat käsitteet:

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

Regressioneliösumma $\quad \mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

Residuaalineliösumma $\quad \mathrm{SSE} = \sum_{i=1}^{n} \hat{\varepsilon}_i^2$

Kun mallissa on vakio, näiden välillä on yhteys

$$\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}.$$

Tämän perustelemiseksi todetaan ensin, että $\mathrm{SST} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ (suora lasku) ja $y'y = \hat{y}'\hat{y}+$ SSE (ks. (2.5)). Näin ollen, $\mathrm{SST} = \hat{y}'\hat{y} - n\bar{y}^2+$ SSE, joten riittää todeta, että $\mathrm{SSR} = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2$. Tämä seuraa kuitenkin edellä todetusta seikasta $\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$.

Määritellään nyt mallin *selitysaste*

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

jossa jälkimmäinen yhtälö perustuu edellä todettuun identiteettiin. Koska SST, SSR ja SSE ovat ei-negatiivisia on selitysasteella ominaisuus

$$0 \leq R^2 \leq 1.$$

Selitysaste ilmaistaan yleensä prosentteina eli sanotaan mallin selittävän $100R^2\%$ selitettävän muuttujan havaintojen vaihtelusta. Jos SSE $= 0$, niin $\hat{\varepsilon}_i = 0$ ja $y_i = \hat{y}_i$ kaikilla $i = 1, ..., n$. Tällöin SST $=$ SSR ja $R^2 = 1$ eli selitys on 100-prosenttinen. Kun $y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{\varepsilon}_i$, tämä merkitsee, että havainnot sijaitsevat tasossa suoralla $y = \hat{\beta}_1 + \hat{\beta}_2 x$. Jos taas SSE $=$ SST, niin $R^2 = 0$ ja mallin 'varsinaisilla' selittäjillä $x_2, ..., x_p$ ei ole mitään merkitystä y:tä selitettäessä. Käytännössä on tietenkin aina $0 < R^2 < 1$.

[9] Tähän 'paremmuuteen' on syytä suhtautua varauksin, sillä mallin 'hyvyyttä' voidaan (ja on syytäkin) mitata usein eri tavoin.

Todetaan vielä, että selitysasteelle pätee

$$R^2 = r_{y\hat{y}}^2,$$

jossa $r_{y\hat{y}}$ on selitettävän muuttujan havaintojen y_i ja sovitteiden $\hat{y}_i$ $(1, ..., n)$ välinen otoskorrelaatiokerroin.[10] Tämä nähdään huomaamalla, että $r_{y\hat{y}}$ voidaan kirjoittaa

$$r_{y\hat{y}} = \frac{(\mathbf{y} - \bar{y}\mathbf{1}_n)'(\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n)}{\|\mathbf{y} - \bar{y}\mathbf{1}_n\| \, \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n\|}.$$

Sijoittamalla osoittajassa $\mathbf{y} = \hat{\mathbf{y}} + \hat{\boldsymbol{\varepsilon}}$ ja käyttämällä vektorien $\hat{\mathbf{y}}$ ja $\hat{\boldsymbol{\varepsilon}}$ sekä $\mathbf{1}_n$ ja $\hat{\boldsymbol{\varepsilon}}$ ortogonaalisuutta nähdään, että osoittaja on yhtä kuin $\|\hat{\mathbf{y}} - \bar{y}\mathbf{1}_n\|^2 = \text{SSR}$. Koska nimittäjä on määritelmien mukaan $\sqrt{\text{SST}}\sqrt{\text{SSR}}$, saadaan $r_{y\hat{y}} = \sqrt{\text{SSR}/\text{SST}}$ eli haluttu tulos.

Koska $R = \sqrt{1 - \text{SSE}/\text{SST}}$ ja $\text{SSE} = \min_\beta (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, voidaan PNS-estimointi tulkita edellä todetun mukaan siten, että sovitteeksi valitaan se selittäjien lineaarikombinaatio, jonka otoskorrelaatio selitettävän muuttujan kanssa maksimoituu. Huomaa kuitenkin tulkinnan laskennallinen luonne. Koska havainnot eivät yleensä ole otos mistään kiinteästä populaatiosta, ei otoskorrelaatiokertoimella $r_{y\hat{y}}$ ole yleensä teoreettista vastinetta. Korostettakoon myös, että tämä kuten muutkin selitysasteeseen liittyvät tarkastelut olettavat mallin, jossa on vakio.

2.2 SU-estimointi satunnaisten selittäjien tapauksessa

Joissakin tapauksissa selittävien muuttujien olettaminen ei-satunnaisiksi kiinteiksi luvuiksi saattaa tuntua rajoittavalta. Jos esimerkiksi halutaan selittää kotitalouksien sähkön kulutusta sähkön hinnalla ja kotitalouksien reaalituloilla ja käytettävissä on aikasarja-aineisto, on selittäviä muuttujia vaikea ajatella ei-satunnaisiksi. Edellä esitetty kiinteiden selittäjien malli ja siihen perustuva SU-estimointi voidaan kuitenkin perustella myös satunnaisten selittäjien tapauksessa seuraavasti.

Otetaan lähtökohdaksi malliyhtälö (ks. yhtälö (1.6) ja sitä seuraava keskustelu, s. 3)

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right),$$

ja oletetaan, että satunnainen matriisi $\mathbf{X}$ toteuttaa ehdot

(a) $\mathbf{X} \perp\!\!\!\perp \boldsymbol{\varepsilon}$

(b) $\mathbf{X}$:n todennäköisyysjakauma ei riipu parametreista $\boldsymbol{\beta}$ ja σ^2.

Tässä satunnaismatriisin $\mathbf{X}$ todennäköisyysjakaumalla tarkoitetaan sen kaikkien alkioiden yhteistodennäköisyysjakaumaa, joka voidaan samaistaa niistä muodostetun $np \times 1$ ulotteisen satunnaisvektorin todennäköisyysjakauman kanssa.

[10] Havainnoista $u_1, .., u_n$ ja $v_1, .., v_n$ laskettu otoskorrelaatiokerroin on

$$r_{uv} = \frac{\sum_{i=1}^n (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\sum_{i=1}^n (u_i - \bar{u})^2}\sqrt{\sum_{i=1}^n (v_i - \bar{v})^2}},$$

jossa $\bar{u} = (u_1 + \cdots + u_n)/n$ ja $\bar{v}$ määritellään vastaavasti.

Uskottavuusfunktio on nyt johdettava muuttujien $\mathbf{Y}$ ja $\mathbf{X}$ yhteistodennäköisyysjakaumasta, jolla yksinkertaisuuden vuoksi oletetaan seuraavassa olevan tiheysfunktio $f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{X})$. Ehdollisen tiheysfunktion määritelmän nojalla pätee

$$f_{\mathbf{Y},\mathbf{X}}(\mathbf{y},\mathbf{X}) = f_{\mathbf{X}}(\mathbf{X})\, f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{X}).$$

Tarkasteltavasta malliyhtälöstä ja ehdosta (a) seuraa, että $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{X})$ on $\mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$-jakauman tiheysfunktio. Formaali perustelu sivuutetaan, mutta intuitiivisesti tämä on varsin ilmeistä, sillä ehdollistaminen kiinnittää $\mathbf{X}$:n sen havaituksi arvoksi ja ehdon (a) riippumattomuus takaa sen, ettei $\mathbf{X}$:n kiinnittäminen vaikuta virhetermin $\boldsymbol{\varepsilon}$ satunnaisvaihteluun. Näin ollen $\mathbf{Y}$ ehdolla $\mathbf{X}$:n havaittu arvo jakautuu kuten kiinteiden selittäjien tapauksessa. Koska ehdon (b) nojalla $\mathbf{X}$:n tiheysfunktio ei riipu parametreista $\boldsymbol{\beta}$ ja σ^2, se voidaan sisällyttää uskottavuusfunktion vakioon, jolloin päädytään samaan uskottavuusfunktioon kuin aikaisemmassa kiinteiden selittäjien mallissa.

Ehtojen (a) ja (b) voimassa ollessa voidaan siis ehdollistaa satunnaisten selittäjien saamien havaintoarvojen suhteen ja tulkita ne kiinteiksi luvuiksi. Erityisesti silloin, kun mallia käytetään selitettävien muuttujien välisen riippuvuuden kuvaamiseen (eli 'selittämiseen') tai selitettävän muuttujan arvojen ennustamiseen tai kontrolloimiseen, ei selitettävien muuttujien todennäköisyysjakaumasta olla välttämättä kiinnostuneita ja ehdollisen jakauman $f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{X})$ käyttäminen on riittävää.

Pohdittaessa ehtojen (a) ja (b) paikkansapitävyyttä kannattaa kiinnittää huomiota ehtoon (a), joka on looginen silloin, kun kausaalisuuden suunta on selittävistä muuttujista selitettävään muuttujaan, mutta ei päinvastoin. Jos kausaalisuus pätisi molempiin suuntiin, voitaisiin esimerkiksi tarkastella samanaikaisesti malliyhtälöitä $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ ja $X_i = \alpha_1 + \alpha_2 Y_i + \eta_i$, jossa $\beta_2 \neq 0 \neq \alpha_2$. Tällöin ehto $X_i \perp\!\!\!\perp \varepsilon_i$ ei olisi selvästikään looginen (eikä myöskään $Y_i \perp\!\!\!\perp \eta_i$).

Ehto (a) rikkoontuu myös silloin, kun oikeiden selittäjien asemesta joudutaan käyttämään (satunnaisia) mittausvirheitä sisältäviä korvikkeita. Tarkastellaan esimerkiksi malliyhtälöä

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,$$

jossa $\beta_2 \neq 0$ ja oikean selittävän muuttujan x_i asemesta havaitaan virheellisesti

$$X_i = x_i + \eta_i,$$

jossa satunnaisella mittausvirheellä η_i on ominaisuudet $\mathsf{E}(\eta_i) = 0$, $\mathsf{Var}(\eta_i) = \sigma_\eta^2 > 0$ ja $\eta_i \perp\!\!\!\perp \varepsilon_i$. Malli, jossa selittävänä muuttujana on X_i, perustetaan yhtälöön

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i^*,$$

jossa $\varepsilon_i^* = \varepsilon_i - \beta_2\eta_i$. Tällöin oletus $X_i \perp\!\!\!\perp \varepsilon_i^*$ ei ole voimassa, sillä $\mathsf{Cov}(X_i, \varepsilon_i^*) = \mathsf{E}(X_i\varepsilon_i^*) = \mathsf{E}(\eta_i\varepsilon_i^*) = -\beta_2\mathsf{E}(\eta_i^2) = -\beta_2\sigma_\eta^2 \neq 0$.

2.3 SU-estimaattorien ominaisuudet

Tässä jaksossa tutkitaan SU-estimointia teoreettisesti, joten $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\mathbf{Y})$ ja $\hat{\sigma}^2 = \hat{\sigma}^2(\mathbf{Y})$ tulkitaan satunnaisiksi suureiksi eli ne ovat estimaattoreita. Seuraava lause, jonka

todistus esitetään jakson lopussa, selvittää näiden estimaattorien todennäköisyysjakaumat.

Lause 2.1. *Tarkastellaan lineaarista mallia* $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$, $\boldsymbol{\beta} \in \mathbb{R}^p$, $\sigma^2 > 0$, *jossa* $\mathsf{r}\left(\mathbf{X}\right) = p$. *Tällöin parametrien* $\boldsymbol{\beta}$ *ja* σ^2 *SU-estimaattoreille* $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ *ja* $\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$ *pätee*

(i) $\hat{\boldsymbol{\beta}} \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$

(ii) $n\hat{\sigma}^2/\sigma^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})/\sigma^2 \sim \chi^2_{n-p}$

(iii) $\hat{\boldsymbol{\beta}} \perp\!\!\!\perp \hat{\sigma}^2$.

Lausetta 2.1 voidaan käyttää SU-estimaattorien $\hat{\boldsymbol{\beta}}$ ja $\hat{\sigma}^2$ ominaisuuksien tutkimiseen. Kohdasta (i) nähdään heti, että PNS-estimaattori $\hat{\boldsymbol{\beta}}$ on *harhaton*. Estimaattori $\hat{\sigma}^2$ ei sen sijaan ole harhaton, sillä lauseen toisesta kohdasta seuraa $\mathsf{E}\left(\hat{\sigma}^2\right) = \mathsf{E}\left(\sigma^2\chi^2_{n-p}/n\right) = (n-p)\,\sigma^2/n$. Parametrin σ^2 harhaton estimaattori on

$$S^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{n}{n-p}\hat{\sigma}^2,$$

jota käytetään käytännössä SU-estimaattorin $\hat{\sigma}^2$ asemesta.

Suoraviivaisella laskulla nähdään, että parametrien $\boldsymbol{\beta}$ ja σ^2 Fisherin informaatiomatriisi on

$$\mathbf{i}\left(\boldsymbol{\beta}, \sigma^2\right) = \begin{bmatrix} \sigma^{-2}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0} & n/2\sigma^4 \end{bmatrix}.$$

Laskelmien yksityiskohtaiset perustelut jätetään harjoitustehtäväksi (vasemman yläkulman lohko johdettiin olennaisilta osin PNS-estimaattia $\hat{\boldsymbol{\beta}}$ johdettaessa). Tilastollisen päättelyn kurssilla todetusta informaatioepäyhtälön moniulotteisesta versiosta voidaan nyt päätellä, että parametrin $\boldsymbol{\beta}$ mille tahansa harhattomalle estimaattorille $\tilde{\boldsymbol{\beta}}$ pätee $\mathsf{Cov}(\tilde{\boldsymbol{\beta}}) - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \geq \mathbf{0}$. Tästä ja Lauseesta 2.1(i) seuraa, että PNS-estimaattori $\hat{\boldsymbol{\beta}}$ on *täystehokas*. Estimaattori S^2 ei sen sijaan ole täystehokas, sillä $\mathsf{Var}(S^2) = \mathsf{Var}(\sigma^2\chi^2_{n-p}/n - p) = 2\sigma^4/\left(n-p\right)$. (Sama pätee myös SU-estimaattorille $\hat{\sigma}^2$.)

Todetaan seuraavaksi, että estimaattorit $\hat{\boldsymbol{\beta}}$ ja S^2 ovat *tyhjentäviä*. Kirjoittamalla $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \hat{\boldsymbol{\varepsilon}}$ ja käyttämällä matriisin $\mathbf{X}$ sarakkeiden ja residuaalivektorin $\hat{\boldsymbol{\varepsilon}}$ ortogonaalisuutta (ks. (2.3), s. 9) nähdään, että jäännösneliösummafunktio

$$\begin{aligned} \mathsf{S}\left(\boldsymbol{\beta}\right) &= ((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}' + \hat{\boldsymbol{\varepsilon}}')(\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \hat{\boldsymbol{\varepsilon}}) \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (n-p)s^2. \end{aligned} \tag{2.6}$$

Havaintojen yhteistiheysfunktiolle saadaan siten esitys

$$f_{\mathbf{Y}}\left(\mathbf{y}; \boldsymbol{\beta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{1}{2\sigma^2}(n-p)s^2\right\},$$

mistä seuraa faktorointikriteerin perusteella estimaattorien $\hat{\boldsymbol{\beta}}$ ja S^2 tyhjentävyys.[11]

Edellä esitettyjä tuloksia voidaan soveltaa riippumattoman normaalisen otoksen malliin $Y_1, ..., Y_n \underline{\parallel}\ , Y_i \sim \mathsf{N}\left(\mu, \sigma^2\right)$, $(\mu \in \mathbb{R},\ \sigma^2 > 0)$. Tätä mallia on tarkasteltu tilastollisen päättelyn kurssilla ja osoitettu, että parametrien μ ja σ^2 SU-estimaatit ovat

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{ja} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2,$$

jotka saadaan myös helposti jaksossa 2.1 johdetuista yleisistä kaavoista. Tilastollisen päättelyn kurssilla todetun lisäksi voidaan Lauseen 2.1 ja edellä sanotun avulla perustella myös estimaattorin $\hat{\sigma}^2$ harhaisuus sekä otosvarianssin $S^2 = n\hat{\sigma}^2/\left(n-1\right)$ harhattomuus. Lisäksi voidaan perustella otoskeskiarvon ja otosvarianssin riippumattomuus sekä tulos $\left(n-1\right)S^2/\sigma^2 \sim \chi^2_{n-1}$, joilla on keskeinen merkitys testattaessa odotusarvoa μ koskevia hypoteeseja (tähän palataan).

Kaiken kaikkiaan voidaan todeta, että estimaattorien $\hat{\boldsymbol{\beta}}$ ja S^2 tilastolliset ominaisuudet ovat erinomaiset. Mainittakoon, että perinteisesti lineaarisen mallin teoriassa näiden estimaattorien ominaisuuksia on tutkittu olettamatta havaintojen normaalisuutta. Lauseen 2.1 todistuksesta nähdään, että olettamalla pelkästään $\mathsf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ ja $\mathsf{Cov}(\mathbf{Y}) = \sigma^2\mathbf{I}_n$ saadaan tulokset $\mathsf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ ja $\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Lisäksi voidaan osoittaa, että jos $\tilde{\boldsymbol{\beta}}$ on mikä tahansa parametrin $\boldsymbol{\beta}$ harhaton ja *lineaarinen* (eli tyyppiä $\mathbf{AY}$ oleva) estimaattori, niin $\mathsf{Cov}(\tilde{\boldsymbol{\beta}}) - \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \geq \mathbf{0}$. Tämä ns. Gaussin ja Markovin lause sanoo siis, että PNS-estimaattori on aina (varianssikriteerin mielessä) paras lineaarinen harhaton estimaattori. Normaalisessa tapauksessa PNS-estimaattori on SU-estimaattori ja paras kaikkien estimaattorien joukossa.

Lauseen 2.1 todistus: (i) Lineaarisen mallin oletuksesta seuraa, että $\mathsf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ ja $\mathsf{Cov}\left(\mathbf{Y}\right) = \sigma^2\mathbf{I}_n$, joten tunnettuja tuloksia käyttäen saadaan (ks. Liite A.1)

$$\begin{aligned}
\mathsf{E}(\hat{\boldsymbol{\beta}}) &= \mathsf{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{E}(\mathbf{Y}) \\
&= \boldsymbol{\beta}
\end{aligned}$$

ja

$$\begin{aligned}
\mathsf{Cov}(\hat{\boldsymbol{\beta}}) &= \mathsf{Cov}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathsf{Cov}\left(\mathbf{Y}\right)\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.
\end{aligned}$$

Koska ($p \times n$ matriisin $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ aste on p ja) $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$, on $\hat{\boldsymbol{\beta}}$ multinormaalisen satunnaisvektorin $\mathbf{Y}$ (täyttä riviastetta olevana) lineaarimuunnoksena multinormaalinen (ks. Liite A.2.4).

(ii) Jaksossa 2.1.1 (ks. (2.4)) todetun mukaan $\hat{\boldsymbol{\varepsilon}} = \left(\mathbf{I}_n - \mathbf{P}\right)\mathbf{Y}$, jossa $\mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ on projektiomatriisi, jolla on ominaisuus $\left(\mathbf{I}_n - \mathbf{P}\right)\mathbf{X} = \mathbf{0}$. Käyttäen

[11] Koska matriisi $\mathbf{X}'\mathbf{X}$ on positiivisesti definiitti, seuraa hajotelmasta (2.6), että $\mathsf{S}\left(\boldsymbol{\beta}\right) \geq \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ ja että $\mathsf{S}\left(\boldsymbol{\beta}\right)$ saavuttaa minimiarvonsa $\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}$ jos ja vain jos $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. Tämä on kolmas tapa nähdä, että PNS-estimaatti $\hat{\boldsymbol{\beta}}$ todella minimoi jäännösneliösummafunktion $\mathsf{S}\left(\boldsymbol{\beta}\right)$.

yhtälöä (1.6) (s. 3) voidaan näin ollen kirjoittaa $(\mathbf{I}_n - \mathbf{P})\,\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\,\boldsymbol{\varepsilon}$, joten estimaattorin $\hat{\sigma}^2$ määritelmää ja projektiomatriisien ominaisuuksia käyttäen saadaan

$$\frac{n}{\sigma^2}\hat{\sigma}^2 = \frac{1}{\sigma^2}\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{\sigma^2}\boldsymbol{\varepsilon}'\,(\mathbf{I}_n - \mathbf{P})\,\boldsymbol{\varepsilon} \sim \chi^2_{n-p}.$$

Tässä viimeinen relaatio seuraa Liitteen A Lauseesta A.2 ja siitä, että $\boldsymbol{\varepsilon} \sim \mathsf{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right)$ ja $\mathsf{r}\,(\mathbf{I}_n - \mathbf{P}) = n-p$. Viimeksi mainittu seikka nähdään seuraavasta laskelmasta, jossa $\mathsf{tr}\,(\cdot)$ on asianomaisen neliömatriisin jälki:[12]

$$\begin{aligned}
\mathsf{r}\,(\mathbf{I}_n - \mathbf{P}) &= \mathsf{tr}\,(\mathbf{I}_n - \mathbf{P}) \\
&= \mathsf{tr}\,(\mathbf{I}_n) - \mathsf{tr}\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\
&= n - \mathsf{tr}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\right) \\
&= n - \mathsf{tr}\,(\mathbf{I}_p) \\
&= n - p.
\end{aligned}$$

(iii) Koska projektiomatriisin ominaisarvot ovat nollia ja ykkösiä, on projektion $\mathbf{I}_n - \mathbf{P}$ pääakseliesitys muotoa $\mathbf{I}_n - \mathbf{P} = \mathbf{R}\mathbf{R}'$, jossa $n \times (n-p)$ matriisilla $\mathbf{R}$ on ominaisuudet $\mathsf{r}(\mathbf{R}) = n - p$ ja $\mathbf{R}'\mathbf{R} = \mathbf{I}_{n-p}$ (vrt. Lauseen A.2 todistus Liiteessä A). Näin ollen $\hat{\sigma}^2 = n^{-1}\mathbf{Y}'\,(\mathbf{I}_n - \mathbf{P})\,\mathbf{Y} = n^{-1}\mathbf{Y}'\mathbf{R}\mathbf{R}'\mathbf{Y}$, joten riittää osoittaa, että satunnaisvektorit $\mathbf{R}'\mathbf{Y}$ ja $\hat{\boldsymbol{\beta}}$ ovat riippumattomia.

Koska $\mathbf{R}\mathbf{R}'\mathbf{X} = (\mathbf{I}_n - \mathbf{P})\,\mathbf{X} = \mathbf{0}$ ja $\mathbf{R}'\mathbf{R} = \mathbf{I}_{n-p}$, on $\mathbf{R}'\mathbf{X} = \mathbf{0}$. Näin ollen (ks. Liite A.1),

$$\begin{aligned}
\mathsf{Cov}(\mathbf{R}'\mathbf{Y}, \hat{\boldsymbol{\beta}}) &= \mathsf{Cov}(\mathbf{R}'\mathbf{Y}, (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{Y}) \\
&= \mathbf{R}'\mathsf{Cov}(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \\
&= \sigma^2\mathbf{R}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-\mathbf{1}} \\
&= \mathbf{0},
\end{aligned}$$

jonka kolmannessa yhtälössä on jälleen käytetty oletusta $\mathsf{Cov}\,(\mathbf{Y}) = \sigma^2\mathbf{I}_n$. Väite seuraa tästä, koska satunnaisvektoreilla $\mathbf{R}'\mathbf{Y}$ ja $\hat{\boldsymbol{\beta}}$ on multinormaalinen yhteisjakauma ja multinormaalijakaumassa komponenttien korreloimattomuus on yhtäpitävää niiden riippumattomuuden kanssa (ks. Liite A.2.3). Edellinen seikka nähdään kirjoittamalla

$$\begin{bmatrix} \mathbf{R}'\mathbf{Y} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \mathbf{R}' \\ (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}' \end{bmatrix}\mathbf{Y}, \quad \mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$$

(ks. Liite A.2.4 ja huomaa, että yhtälön oikealla puolella olevan matriisin rivit ovat lineaarisesti riippumattomia).

2.4 SU-estimointi lineaarisin rajoittein

Tarkastellaan lineaarista mallia $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$, jossa tavanomaiseen tapaan $\mathsf{r}\,(\mathbf{X}) = p$, mutta parametriavaruus ei ole kuten aikaisemmin, vaan parametrivektorin $\boldsymbol{\beta}$ komponenttien oletetaan toteuttavan lineaariset rajoitteet

$$\mathbf{A}\boldsymbol{\beta} = \mathbf{c}, \tag{2.7}$$

[12] Jäljellä eli diagonaalialkioiden summalla on ominaisuudet $\mathsf{tr}\,(\mathbf{A}+\mathbf{B}) = \mathsf{tr}\,(\mathbf{A}) + \mathsf{tr}\,(\mathbf{B})$, $\mathsf{tr}\,(\mathbf{AB}) = \mathsf{tr}\,(\mathbf{BA})$ ja $\mathsf{tr}\,(\mathbf{P}) = \mathsf{r}\,(\mathbf{P})$, kun $\mathbf{P}$ on projektio.

jossa $\mathbf{A}$ $(q \times p)$ ja $\mathbf{c}$ $(q \times 1)$ ovat tunnettuja ja $\mathsf{r}(\mathbf{A}) = q$. Tehtävänä on estimoida parametrit $\boldsymbol{\beta}$ ja σ^2 ottaen nämä rajoitteet huomioon. Parametriavaruus on näin ollen $\{\boldsymbol{\beta} \in \mathbb{R}^p : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}\}$, $\sigma^2 > 0$.

Tämän estimointiongelman ratkaisua tarvitaan myöhemmin, kun tarkastellaan yhtälön (2.7) määrittämän hypoteesin testaamista. Tyypillinen esimerkki saadaan valitsemalla $\mathbf{A} = [\mathbf{0} \;\; \mathbf{I}_q]$ ja $\mathbf{c} = \mathbf{0}$, jolloin testattava hypoteesi on $\beta_{p-q+1} = \cdots = \beta_p = 0$ eli viimeiset q selittäjää ovat mallissa tarpeettomia. Toisena esimerkkinä mainittakoon $\beta_{p-1} = \beta_p$, joka saadaan valitsemalla $\mathbf{A} = [0 \; \cdots \; 0 \;\; 1 \;\; -1]$ ja $c = 0$. Tällainen hypoteesi voi seurata tutkittavan ilmiön taustateoriasta.

Palautetaan mieleen jäännösneliösummafunktio $\mathsf{S}(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ ja tarkastellaan log-uskottavuusfunktion

$$l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\right) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathsf{S}(\boldsymbol{\beta})$$

maksimointia edellä kuvatussa tilanteessa. Tämä johtaa parametrivektorin $\boldsymbol{\beta}$ osalta jäännösneliösummafunktion $\mathsf{S}(\boldsymbol{\beta})$ minimointiin ehdolla $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. Jos $\hat{\boldsymbol{\beta}}_H$ on saatu estimaatti (eli $\boldsymbol{\beta}$:n SU-estimaatti), nähdään kuten jaksossa 2.1 (ks. s. 7), että parametrin σ^2 SU-estimaatti on $\hat{\sigma}^2_H = \frac{1}{n}\mathsf{S}(\hat{\boldsymbol{\beta}}_H)$. Jakson lopussa osoitetaan, että[13]

$$\hat{\boldsymbol{\beta}}_H = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}), \tag{2.8}$$

jossa matriisin $\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'$ epäsingulaarisuus seuraa oletuksista $\mathsf{r}(\mathbf{A}) = q$ ja $\mathsf{r}(\mathbf{X}) = p$. Laskemalla nähdään, että vaadittu rajoite $\mathbf{A}\hat{\boldsymbol{\beta}}_H = \mathbf{c}$ toteutuu.

Usein estimaatti $\hat{\boldsymbol{\beta}}_H$ voidaan muodostaa yhtälön (2.8) yleistä kaavaa helpommin kirjoittamalla malli muotoon, jossa rajoitteet (2.7) otetaan suoraan huomioon. Tarkastellaan esimerkiksi malliyhtälöä

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n.$$

Jos nyt asetetaan rajoite $\beta_p = 0$, on $\hat{\boldsymbol{\beta}}_H = [\hat{\beta}_{H,1} \; \cdots \; \hat{\beta}_{H,p-1} \;\; 0]'$, jossa $\hat{\beta}_{H,1}, ..., \hat{\beta}_{H,p-1}$ saadaan PNS:llä malliyhtälöstä

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i, \quad i = 1, ..., n.$$

Tämä idea yleistyy seuraavasti. Tarkastellaan vaihtoehtoisia lineaarisia rajoitteita

$$\boldsymbol{\beta} = \mathbf{C}\boldsymbol{\phi} + \mathbf{d}, \tag{2.9}$$

jossa $\mathbf{C}$ on tunnettu astetta r oleva $p \times r$ matriisi, $\mathbf{d}$ on tunnettu $p \times 1$ vektori ja $\boldsymbol{\phi} \in \mathbb{R}^r$ on tuntematon parametrivektori. Sijoittamalla oikea puoli malliyhtälöön (1.6) saadaan

$$\mathbf{Y} - \mathbf{X}\mathbf{d} = (\mathbf{X}\mathbf{C})\,\boldsymbol{\phi} + \boldsymbol{\varepsilon},$$

[13] Estimaatin $\hat{\boldsymbol{\beta}}_H$ lausekkeeseen voidaan päätyä minimoimalla jäännösneliösummafunktio $\mathsf{S}(\boldsymbol{\beta})$ Lagrangen kerroinmenettelyä käyttäen. Tällöin tehtävänä on minimoida funktio $Q(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \mathsf{S}(\boldsymbol{\beta}) + \boldsymbol{\lambda}'(\mathbf{A}\boldsymbol{\beta} - \mathbf{c})$, jossa vektori $\boldsymbol{\lambda} = [\lambda_1 \; \cdots \; \lambda_q]'$ sisältää Lagrangen kertoimet. Laskemalla derivaatat $\partial Q(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\beta}$ ja $\partial Q(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\lambda}$ ja ratkaisemalla yhtälöt $\partial Q(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\beta} = \mathbf{0}$ ja $\partial Q(\boldsymbol{\beta}, \boldsymbol{\lambda})/\partial\boldsymbol{\lambda} = \mathbf{0}$ saadaan esitetty $\hat{\boldsymbol{\beta}}_H$:n lauseke.

josta PNS:ää soveltaen saadaan $\boldsymbol{\phi}$:n SU-estimaatiksi

$$\hat{\boldsymbol{\phi}} = \left(\mathbf{C}'\mathbf{X}'\mathbf{X}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{X}'\left(\mathbf{y}-\mathbf{X}\mathbf{d}\right).$$

SU-estimaatin invarianssiominaisuuden nojalla saadaan siten $\hat{\boldsymbol{\beta}}_H = \mathbf{C}\hat{\boldsymbol{\phi}} + \mathbf{d}$.

Rajoitteiden (2.9) tapauksessa voidaan PNS-estimointi suorittaa havainnollisemmin kuin rajoitteiden (2.7) tapauksessa. Jälkimmäiset ovat kuitenkin käteviä seuraavassa jaksossa tarkasteltavan testiteorian kannalta. Huomaa, että tyyppiä (2.9) olevat rajoitteet voidaan aina muuntaa tyyppiä (2.7) oleviksi rajoitteiksi, sillä lineaarialgebrasta tiedetään, että matriisin $\mathbf{C}$ ollessa annettu, voidaan aina löytää $(p-r)\times p$ matriisi $\mathbf{A}$, jolle pätee $\mathbf{AC} = \mathbf{0}$ ja $\mathsf{r}\left(\mathbf{A}\right) = p - r$. Tällöin rajoitteet (2.7) pätevät valinnoilla $q = p - r$ ja $\mathbf{c} = \mathbf{A}\mathbf{d}$.[14]

Tuloksen (2.8) perustelu. Kuten jaksossa 2.3 todettiin (ks. (2.6), s. 13), pätee

$$\mathsf{S}\left(\boldsymbol{\beta}\right) = \mathsf{S}(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}).$$

Hajotetaan oikean puolen jälkimmäinen termi osiin:

$$\begin{aligned}
(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) &= (\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H+\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H+\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}) \\
&= (\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H) + (\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}) \\
&\quad +2(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H) \\
&= (\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H) + (\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}).
\end{aligned}$$

Viimeksi esitetyn yhtälön perustelemiseksi otetaan käyttöön lyhennysmerkintä $\boldsymbol{\lambda} = (\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}}-\mathbf{c})$. Yhtälö (2.8) voidaan siten kirjoittaa $\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}'\boldsymbol{\lambda}$ ja, koska $\mathbf{A}\hat{\boldsymbol{\beta}}_H = \mathbf{c} = \mathbf{A}\boldsymbol{\beta}$,

$$(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H) = (\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{A}'\boldsymbol{\lambda} = (\mathbf{c}-\mathbf{c})'\boldsymbol{\lambda} = 0.$$

Edellä sanotusta seuraa

$$\mathsf{S}\left(\boldsymbol{\beta}\right) = \mathsf{S}(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}}_H) + (\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_H-\boldsymbol{\beta}),$$

josta nähdään, että $\mathsf{S}\left(\boldsymbol{\beta}\right)$ minimoituu täsmälleen silloin, kun oikean puolen viimeinen termi minimoituu. Koska matriisi $\mathbf{X}'\mathbf{X}$ on positiivisesti definiitti, tämä termi on aina ei-negatiivinen ja saavuttaa minimiarvonsa nolla jos ja vain jos $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_H$.

[14] Matriisin $\mathbf{A}$ riveiksi voidaan valita avaruuden $\mathcal{R}\left(\mathbf{C}\right)$ ortogonaalisen komplementin $\mathcal{R}\left(\mathbf{C}\right)^{\perp}$ (jotkin) kantavektorit. Tällöin $p = \mathsf{r}\left(\mathbf{C}\right) + \mathsf{r}\left(\mathbf{A}\right) = r + \mathsf{r}\left(\mathbf{A}\right)$, joten $\mathsf{r}\left(\mathbf{A}\right) = p - r$. Ilman perustelua mainitaan käänteinen tulos, jonka mukaan tyyppiä (2.7) olevat rajoitteet voidaan aina muuntaa tyyppiä (2.9) oleviksi rajoitteiksi.

3 Hypoteesien testaaminen

3.1 F-testi yleiselle lineaariselle hypoteesille

Oletetaan edellisen jakson tapaan $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ $(\mathsf{r}\left(\mathbf{X}\right) = p)$, ja tarkastellaan nollahypoteesia

$$H : \mathbf{A}\boldsymbol{\beta} = \mathbf{c}, \tag{3.1}$$

jossa $\mathbf{A}$ $(q \times p)$ ja $\mathbf{c}$ $(q \times 1)$ ovat tunnettuja ja $\mathsf{r}\left(\mathbf{A}\right) = q$. Testi tälle hypoteesille on luontevaa perustaa erotukseen $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}$, jossa $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$ on parametrin $\boldsymbol{\beta}$ (vapaa) PNS-estimaattori tai yhtäpitävästi (vapaa) SU-estimaattori. Koska $\hat{\boldsymbol{\beta}}$ estimoi parametria $\boldsymbol{\beta}$ tehokkaasti riippumatta siitä onko nollahypoteesi tosi vai ei, pätee aina $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} \approx \mathbf{A}\boldsymbol{\beta} - \mathbf{c}$. Erotus $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}$ saa siten tyypillisesti 'pieniä' arvoja, kun nollahypoteesi on tosi ja 'suuria' arvoja, kun nollahypoteesi ei ole tosi. Lausetta 2.1 käyttäen voidaan johtaa testisuure, jonka avulla tämän erotuksen suuruutta voidaan arvioida. Näin saatava testi perustuu tilastollisen päättelyn kurssilla esitetyn Waldin testin periaatteeseen.

Lauseen 2.1(i) nojalla $\hat{\boldsymbol{\beta}} \sim \mathsf{N}(\boldsymbol{\beta}, \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1})$, joten nollahypoteesin voimassa ollessa (ks. Liite A.2.4)

$$\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} \sim \mathsf{N}(\mathbf{0}, \sigma^2\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}').$$

Tästä ja Liitteen A Lauseesta A.1 seuraa edelleen

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})/\sigma^2 \sim \chi^2_q.$$

Lauseesta 2.1 nähdään puolestaan, että

$$(n-p)\,S^2/\sigma^2 \sim \chi^2_{n-p} \qquad \text{ja} \qquad S^2 \perp\!\!\!\perp \hat{\boldsymbol{\beta}},$$

jossa $S^2 = (n-p)^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$. Edellä mainitut χ^2-muuttujat ovat näin ollen riippumattomia, joten F-jakauman määritelmän mukaan testisuure[15]

$$F = (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})/qS^2 \overset{H}{\sim} F_{q,n-p}.$$

Tätä testisuuretta sanotaan F-testisuureeksi ja siihen perustuvaa testiä F-testiksi. Suuret testisuureen arvot ovat kriittisiä nollahypoteesin kannalta. Testin P-arvot perustetaan tulokseen

$$\text{P} = \mathsf{P}_H\left(F\left(\mathbf{Y}\right) \geq F\left(\mathbf{y}\right)\right) = \mathsf{P}\left(F_{q,n-p} \geq F\left(\mathbf{y}\right)\right),$$

jossa $F_{q,n-p}$ on $F_{q,n-p}$-jakaumaa noudattava satunnaismuuttuja.

Edellä johdetun F-testin tulkinta Waldin testinä seuraa siitä, että $s^{-2}\mathbf{X}'\mathbf{X}$ estimoi parametrin $\boldsymbol{\beta}$ Fisherin informaatiomatriisia ja että parametrien $\boldsymbol{\beta}$ ja σ^2 Fisherin informaatiomatriisi on lohkodiagonaalinen (eli $\boldsymbol{\beta}$ ja σ^2 ovat ortogonaaliset).

Testisuure F voidaan esittää myös käyttäen edellisessä jaksossa johdettua rajoitettua PNS-estimaattoria $\hat{\boldsymbol{\beta}}_H$. Huomataan ensin, että

$$\mathsf{S}\left(\boldsymbol{\beta}\right) = \mathsf{S}(\hat{\boldsymbol{\beta}}) + (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

[15] $F_{k,m}$-jakauman (vapausastein k ja m) määrittelee satunnaismuuttuja $m\chi^2_k/k\chi^2_m$, jossa $\chi^2_k \perp\!\!\!\perp \chi^2_m$.

kuten jaksossa 2.3 todettiin (ks. (2.6), s. 13). Sijoittamalla tähän $\boldsymbol{\beta}$:n paikalle $\hat{\boldsymbol{\beta}}_H$ saadaan

$$\begin{aligned} \mathsf{S}(\hat{\boldsymbol{\beta}}_H) - \mathsf{S}(\hat{\boldsymbol{\beta}}) &= (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_H) \\ &= (\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}), \end{aligned}$$

jossa jälkimmäinen yhtälö seuraa rajoitetun PNS-estimaattorin $\hat{\boldsymbol{\beta}}_H$ lausekkeesta (2.8) suoralla laskulla. Koska $S^2 = (n-p)^{-1}\,\mathsf{S}(\hat{\boldsymbol{\beta}})$, voidaan näin ollen kirjoittaa

$$F = \frac{(\mathsf{S}(\hat{\boldsymbol{\beta}}_H) - \mathsf{S}(\hat{\boldsymbol{\beta}}))/q}{\mathsf{S}(\hat{\boldsymbol{\beta}})/\,(n-p)} \overset{H}{\sim} F_{q,n-p}.$$

Tämän perusteella testisuure F asettaa nollahypoteesin epäilyksen alaiseksi, jos rajoitettu residuaalineliösumma $\mathsf{S}(\hat{\boldsymbol{\beta}}_H) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_H)$ on 'kohtuuttoman paljon' suurempi kuin vapaa residuaalineliösumma $\mathsf{S}(\hat{\boldsymbol{\beta}}) = (\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})$. Tämä on residuaalien tulkinta huomioon ottaen luonnollista. Tarkasteltavasta tapauksesta riippuu kumpi edellä esitetyistä kahdesta testisuureen lausekkeesta on kätevämpi. (Sovelluksissa tietokoneohjelma tietysti laskee testisuureen arvon automaattisesti.)

Edellä johdetusta tuloksesta voidaan myös päätellä, että F-testi on identtinen uskottavuusosamäärän testin kanssa. Koska $\hat{\sigma}^2 = n^{-1}\mathsf{S}(\hat{\boldsymbol{\beta}})$ ja $\hat{\sigma}^2_H = n^{-1}\mathsf{S}(\hat{\boldsymbol{\beta}}_H)$, on uskottavuusosamäärän testisuure

$$\begin{aligned} r\,(\mathbf{y}) &= 2\left[l(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2; \mathbf{y}) - l(\hat{\boldsymbol{\beta}}_H, \hat{\sigma}^2_H; \mathbf{y})\right] \\ &= 2\left[\frac{n}{2}\log\hat{\sigma}^2_H + \frac{n}{2} - \frac{n}{2}\log\hat{\sigma}^2 - \frac{n}{2}\right] \\ &= n\log\left(\frac{(\mathsf{S}(\hat{\boldsymbol{\beta}}_H) - \mathsf{S}(\hat{\boldsymbol{\beta}}))}{\mathsf{S}(\hat{\boldsymbol{\beta}})} + 1\right) \\ &= n\log\left(\frac{q}{n-p}F + 1\right). \end{aligned}$$

Testisuure $r\,(\mathbf{y})$ on siis monotoninen funktio F-testisuureesta, joten molemmat testisuureet määrittelevät saman testin. Ilman perustelua mainitaan, että myös Raon testisuure ja F-testisuure määrittelevät saman testin.

3.2 F-testin erikoistapauksia

Sovelletaan nyt edellä johdettua F-testiä kahteen erikoistapaukseen. Ensimmäisessä malli perustuu yhtälöön

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n,$$

jossa ns. *yleistesti* koskee nollahypoteesia

$$H : \beta_2 = \cdots = \beta_p = 0 \quad \text{eli} \quad [\mathbf{0} \;\; \mathbf{I}_{p-1}]\,\boldsymbol{\beta} = \mathbf{0}.$$

Tämän nollahypoteesin voimassa ollessa kaikki selittäjät vakiota lukuun ottamatta ovat turhia. Tässä tapauksessa testi on kätevää esittää käyttäen F-testisuureen residuaalineliösummaesitystä. Koska nollahypoteesin voimassa ollessa $\mathbf{Y} \sim \mathsf{N}\left(\beta_1\mathbf{1}_n, \sigma^2\mathbf{I}_n\right)$,

saadaan rajoitetuksi PNS-estimaattoriksi $\hat{\boldsymbol{\beta}}_H = \begin{bmatrix}\bar{Y} & 0 & \cdots & 0\end{bmatrix}'$ $\left(\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i\right)$, joten

$$\mathsf{S}(\hat{\boldsymbol{\beta}}_H) = \left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right)'\left(\mathbf{Y} - \bar{Y}\mathbf{1}_n\right) = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2.$$

Vapaa residuaalineliösumma voidaan (halutessa) kirjoittaa (ks. (2.3), s. 9)

$$\mathsf{S}(\hat{\boldsymbol{\beta}}) = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y}.$$

Testisuureeksi saadaan siis

$$F = \frac{(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2)/(p-1)}{(\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{Y})/(n-p)} \overset{H}{\sim} F_{p-1,n-p}.$$

Yksittäistä selittäjää koskevassa testissä nollahypoteesina on

$$H_j : \beta_j = 0, \quad 1 \leq j \leq p.$$

(Tässä ei välttämättä ole enää $x_{i1} = 1$, $i = 1, ..., n$.) Tämä nollahypoteesi merkitsee, että muiden selittäjien ollessa mallissa tutkitaan onko selittäjän x_j lisääminen tarpeen. Testattaessa oletetaan siis, että muut kertoimet β_k, $k \neq j$, saavat poiketa nollasta.

Merkitään $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1 \ \cdots \ \hat{\beta}_p]'$ ja

$$(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{M}^{-1} = \left[m^{ab}\right], \quad a, b = 1, ..., p.$$

Valitsemalla $\mathbf{A} = [0 \ \cdots \ 0 \ \ 1 \ \ 0 \ \cdots \ 0]$, jossa ykkönen on j. komponentti, ja $\mathbf{c} = 0$ nähdään, että nollahypoteesi on vaadittua lineaarista muotoa ja että $\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c} = \hat{\beta}_j$ sekä

$$(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c})'(\mathbf{A}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - \mathbf{c}) = \hat{\beta}_j^2/m^{jj}.$$

F-testisuureen ensimmäisestä lausekkeesta saadaan näin ollen

$$F = \hat{\beta}_j^2/S^2 m^{jj} \overset{H}{\sim} F_{1,n-p},$$

jossa jälleen $S^2 = (n-p)^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

Jos $Z \sim \mathsf{N}(0,1)$ ja $Z \perp\!\!\!\perp \chi_k^2$, niin tunnetusti $T_k = Z/\sqrt{\frac{1}{k}\chi_k^2} \sim t_k$ (t-jakauma vapausastein k) ja $T_k^2 \sim F_{1,k}$. Nollahypoteesia voidaan siis testata myös testisuureella

$$T = t(\mathbf{Y}) = \hat{\beta}_j/S\sqrt{m^{jj}} \overset{H}{\sim} t_{n-p}.$$

Tämän testin P-arvot perustetaan tulokseen

$$\text{P} = \mathsf{P}_{H_j}\left(|t(\mathbf{Y})| \geq |t(\mathbf{y})|\right) = \mathsf{P}\left(|T_{n-p}| \geq |t(\mathbf{y})|\right),$$

jossa satunnaismuuttuja $T_{n-p} \sim t_{n-p}$. Tässä vaihtoehdon ajatellaan olevan kaksisuuntainen eli $\beta_j \neq 0$. Yksisuuntaisen vaihtoehdon $\beta_j > 0$ $\left[\text{tai } \beta_j < 0\right]$ tapauksessa P-arvot lasketaan kaavalla $\mathsf{P}\left(T_{n-p} \geq t(\mathbf{y})\right)$ [tai $\mathsf{P}\left(T_{n-p} \leq t(\mathbf{y})\right)$].

Käytännössä käytetään yleensä t-testisuuretta ja laaditaan esimerkiksi seuraavanlainen taulukko, jollaisen tietokoneohjelmat tulostavat automaattisesti. (Taulukossa $\hat{\beta}_j$:n keskivirhe = $\hat{\beta}_j$:n estimoitu hajonta.)

Parametri	Estimaatti	Keskivirhe	t-suhde
β_1	$\hat{\beta}_1$	$s\sqrt{m^{11}}$	$\hat{\beta}_1/s\sqrt{m^{11}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
β_p	$\hat{\beta}_p$	$s\sqrt{m^{pp}}$	$\hat{\beta}_p/s\sqrt{m^{pp}}$

Toinen yleinen tapa esittää tulokset on kirjoittaa estimoitu malliyhtälö

$$y_i = \underset{(s.e.(\hat{\beta}_1))}{\hat{\beta}_1} x_{i1} + \cdots + \underset{(s.e.(\hat{\beta}_p))}{\hat{\beta}_p} x_{ip} + \hat{\varepsilon}_i, \quad s^2 = ...,$$

jossa $s.e.(\hat{\beta}_j)$ on $\hat{\beta}_j$:n keskivirhe (standard error). Keskivirheen paikalla näkee käytettävän myös t-suhdetta eikä havaintoyksikköä i välttämättä merkitä näkyviin.

On tärkeää huomata, että testattaessa useita hypoteeseja H_j eivät käytetyt testisuureet ole yleensä riippumattomia. Tämä vaikeuttaa näin saatavan 'yhdistetyn' testin p-arvon laskemista ja siten johtopäätösten tekoa.

Kun mallissa on vakio (eli $x_{i1} = 1$, $i = 1, .., n$), ei yksittäisiä hypoteeseja H_j kannata ilmeisestikään tutkia, ellei yleishypoteesia $\beta_2 = \cdots = \beta_p = 0$ ole hylätty (vakiotermiä ei yleensä testata tilastollisella testillä, vaan sen oletetaan olevan mallissa mukana; vrt. selitysaste ja sen tulkinnat s. 10). Yksittäisten testien riippuvuus saattaa kuitenkin aiheuttaa sen, että yleishypoteesi on hylättävä, vaikka kaikki yksittäiset hypoteesit jäävät voimaan.

Jos joitakin hypoteeseja H_j ei hylätä, niin vastaavat selittäjät ovat mallissa turhia. Selittäjien poistaminen ei ole kuitenkaan yksiselitteistä, sillä selittäjään x_j liittyvän t-suhteen saama arvo riippuu (yleensä) siitä mitä muita selittäjiä mallissa on. Jos selittäjiä poistetaan t-suhteiden perusteella, voidaan siten päätyä eri malleihin riippuen siitä, missä järjestyksessä hypoteeseja H_j testataan.

Viimeksi tarkastellusta testistä saadaan erikoistapauksena testi hypoteesille $\mu = \mu_0$ riippumattoman normaalisen otoksen mallissa $Y_1, ..., Y_n \underline{\parallel}\, Y_i \sim \mathsf{N}(\mu, \sigma^2)$, $(\mu \in \mathbb{R},\ \sigma^2 > 0)$. Koska tässä tapauksessa $\mathbf{X} = \mathbf{1}_n$, $\mathbf{A} = 1$ ja $\mathbf{c} = \mu_0$, saadaan $\mathbf{X}'\mathbf{X} = n$ ja edelleen t-testisuure

$$\sqrt{n}\left(\bar{Y} - \mu_0\right)/S \overset{H}{\sim} t_{n-1},$$

jossa nyt $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$. Tämä perustelee tilastollisen päättelyn kurssilla tähän tilanteeseen esitetyn testin. Huomaa tuloksen $\bar{Y} \underline{\parallel} S^2$ merkitys testisuureen jakauman johtamisessa.

4 Luottamusvälien ja -joukkojen muodostaminen

4.1 Luottamusvälit

Kuten edellisessä jaksossakin oletetaan, että $\mathbf{Y} \sim \mathsf{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$ $(\mathsf{r}\left(\mathbf{X}\right) = p)$. Tyypillisin esimerkki tämän jakson yleisestä otsikosta koskee luottamusvälien muodostamista parametrivektorin $\boldsymbol{\beta}$ yksittäisille komponentille β_j $(1 \leq j \leq p)$. Seuraavassa luottamusväli johdetaan hieman yleisemmin tarkastelemalla parametrivektorin $\boldsymbol{\beta}$ lineaarikombinaatiota $\mathbf{a}'\boldsymbol{\beta} = a_1\beta_1 + \cdots + a_p\beta_p$, jossa $\mathbf{a}$ $(p \times 1)$ on tunnettu (nollasta poikkeava) vektori. Valitsemalla $\mathbf{a}' = [0 \ \cdots \ 0 \ \ 1 \ \ 0 \ \cdots \ 0]$, jossa ykkönen on j. komponentti, saadaan $\mathbf{a}'\boldsymbol{\beta} = \beta_j$. Seuraavassa muita tyypillisiä erikoistapauksia.

- $\mathbf{a}' = \left[x_1^* \ \cdots \ x_p^*\right]$, jolloin $\mathbf{a}'\boldsymbol{\beta} = \beta_1 x_1^* + \cdots + \beta_p x_p^* = Y$:n odotusarvo, kun selittäville muuttujille annetaan arvot $x_1^*, ..., x_p^*$
- odotusarvojen erotus $\mu_1 - \mu_2$ kahden riippumattoman normaalisen otoksen mallissa on myös tyyppiä $\mathbf{a}'\boldsymbol{\beta}$ samoin kuin vastaavat erotukset $\mu_j - \mu_k$ $(j \neq k)$ eli ns. *kontrastit* yleisemmässä yksisuuntaisessa varianssianalyysimallissa (ks. jakso 1.3)

Kuten tilastollisen päättelyn kurssilla todetaan, voidaan luottamusvälejä muodostaa testien avulla. Tätä menettelyä käytetään seuraavassa.

Tarkastellaan ensin edellisessä jaksossa johdettua F-testiä nollahypoteesille

$$H : \mathbf{a}'\boldsymbol{\beta} = \mathbf{a}'\boldsymbol{\beta}_0 \quad \Leftrightarrow \quad \mathbf{a}'\left(\boldsymbol{\beta} - \boldsymbol{\beta}_0\right) = 0,$$

jossa $\boldsymbol{\beta}_0$ $(p \times 1)$ on tunnettu. Testisuureeksi saadaan (ks. s. 18)

$$\begin{aligned} F &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{a}(\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a})^{-1}\mathbf{a}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/S^2 \\ &= \left[\mathbf{a}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right]^2 / S^2\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a} \\ &\overset{H}{\sim} F_{1,n-p}, \end{aligned}$$

jossa aikaisempaan tapaan $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ ja $S^2 = (n-p)^{-1}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$. Kuten edellisessä jaksossa tarkastellun hypoteesin H_j tapauksessa nähdään, että F-testisuureen asemesta voidaan käyttää t-testisuuretta

$$\frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}_0}{S\sqrt{\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}} \overset{H}{\sim} t_{n-p}.$$

Kun vaihtoehtona on $\mathbf{a}'\boldsymbol{\beta} \neq \mathbf{a}'\boldsymbol{\beta}_0$, saadaan kriittiseksi alueeksi merkitsevyystasolla α

$$C_\alpha\left(\mathbf{a}'\boldsymbol{\beta}_0\right) = \left\{\mathbf{y} : \left|\frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}_0}{s\sqrt{\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}}\right| \geq t_{n-p}\left(\alpha/2\right)\right\},$$

jossa $\mathsf{P}\left(|T_{n-p}| \geq t_{n-p}\left(\alpha/2\right)\right) = \alpha$. Vastaava hyväksymisalue muodostuu aineistoista, joille pätee

$$-t_{n-p}\left(\alpha/2\right) < \frac{\mathbf{a}'\hat{\boldsymbol{\beta}} - \mathbf{a}'\boldsymbol{\beta}_0}{s\sqrt{\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}}} < t_{n-p}\left(\alpha/2\right)$$

tai yhtäpitävästi

$$\mathbf{a}'\hat{\boldsymbol{\beta}} - t_{n-p}(\alpha/2)\, s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}} < \mathbf{a}'\boldsymbol{\beta}_0 < \mathbf{a}'\hat{\boldsymbol{\beta}} + t_{n-p}(\alpha/2)\, s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

Tämä väli peittää lineaarikombinaation $\mathbf{a}'\boldsymbol{\beta}_0$ jos ja vain jos $\mathbf{y} \notin C_\alpha(\mathbf{a}'\boldsymbol{\beta}_0)$. Koska kaikilla $\boldsymbol{\beta}_0$ ja σ^2 pätee

$$\mathsf{P}_{\boldsymbol{\beta}_0,\sigma^2}\left(\mathbf{Y} \notin C_\alpha\left(\mathbf{a}'\boldsymbol{\beta}_0\right)\right) = \mathsf{P}_{\boldsymbol{\beta}_0,\sigma^2}\left(|T_{n-p}| < t_{n-p}(\alpha/2)\right) = 1-\alpha,$$

on lineaarikombinaation $\mathbf{a}'\boldsymbol{\beta}$ luottamusväli luottamustasolla $1-\alpha$

$$\mathbf{a}'\hat{\boldsymbol{\beta}} \pm t_{n-p}(\alpha/2)\, s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}.$$

Huomaa, että $\mathsf{Var}(\mathbf{a}'\hat{\boldsymbol{\beta}}) = \mathbf{a}'\mathsf{Cov}(\hat{\boldsymbol{\beta}})\mathbf{a} = \sigma^2\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}$ (Lause 2.1(i)), joten edellä $s\sqrt{\mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}}$ on estimaattorin $\mathbf{a}'\hat{\boldsymbol{\beta}}$ keskivirhe. Jos erityisesti $\mathbf{a}'\boldsymbol{\beta} = \beta_j$, saadaan luottamusväli

$$\hat{\beta}_j \pm t_{n-p}(\alpha/2)\, s\sqrt{m^{jj}},$$

jossa $m^{jj} = \left[(\mathbf{X}'\mathbf{X})^{-1}\right]_{jj}$ on matriisin $(\mathbf{X}'\mathbf{X})^{-1}$ j. diagonaalialkio.

Tapauksessa $\mathbf{a}'\boldsymbol{\beta} = \beta_1 x_1^* + \cdots + \beta_p x_p^*$ on syytä huomata, että kiinnostuksen kohteena on satunnaismuuttujan Y odotusarvo, kun selittäville muuttujille annetaan arvot $x_1^*, ..., x_p^*$. Jos tarkastellaan satunnaismuuttujaa $Y = \beta_1 x_1^* + \cdots + \beta_p x_p^* + \varepsilon^*$, jossa $\varepsilon^* \sim \mathsf{N}\left(0, \sigma^2\right)$, ja halutaan ennustaa sen arvoa, ei edellä esitettyä menettelyä soveltaen saada oikeaa ennusteen luottamusväliä, koska satunnaismuuttujan ε^* vaikutus ei tule huomioon otetuksi. Tämän ongelman ratkaisu vaatii oman menettelynsä.

On myös syytä huomata, että edellä esitetty pätee vain yksittäisen lineaarikombinaation $\mathbf{a}'\boldsymbol{\beta}$ luottamusvälille. Jos luottamusvälit muodostetaan usealle lineaarikombinaatiolle $\mathbf{a}_j'\boldsymbol{\beta}$, $j = 1, ..., k$, niin todennäköisyys, että kaikki luottamusvälit peittäisivät samanaikaisesti vastaavien lineaarikombinaatioiden todelliset arvot *ei* ole $1-\alpha$. Tämän toteamiseksi merkitään

$$E_j = \left\{j.\text{ luottamusväli peittää } \mathbf{a}_j'\boldsymbol{\beta}\text{:n}\right\}.$$

Olkoon E_j^c tämän tapahtuman komplementti $(j = 1, ..., k)$. Jos $\mathsf{P}(E_j) = 1-\alpha_j$, niin

$$\begin{aligned}
\mathsf{P}\left(\cap_{j=1}^k E_j\right) &= 1 - \mathsf{P}\left(\left(\cap_{j=1}^k E_j\right)^c\right) \\
&= 1 - \mathsf{P}\left(\cup_{j=1}^k E_j^c\right) \\
&\geq 1 - \sum_{j=1}^k \mathsf{P}\left(E_j^c\right) \\
&= 1 - \sum_{j=1}^k \alpha_j.
\end{aligned}$$

Jos erityisesti $\alpha_j = \alpha$, $j = 1, ..., k$, niin

$$\mathsf{P}\left(\cap_{j=1}^k E_j\right) \geq 1 - k\alpha$$

(vrt. edellisen jakson lopussa tehty huomautus yksittäisten t-testien P-arvojen laskemisesta).

Valitsemalla $\alpha_j = \alpha/k$ voidaan luottamusvälit muodostaa kaikille lineaarikombinaatioille $\mathbf{a}_j'\boldsymbol{\beta}$ kuten edellä esitettiin. Luottamustasoa ei kuitenkaan saada lasketuksi tarkasti, sillä edellä todetusta saadaan vain epäyhtälö

$$\mathsf{P}\left(\cap_{j=1}^{k} E_j\right) \geq 1 - k(\alpha/k) = 1 - \alpha,$$

joka on yleensä aito. Korvaamalla tämä epäyhtälö yhtälöllä saadaan ns. *Bonferronin t-välit*. Tämä on helppo ratkaisu usean samanaikaisen luottamusvälin muodostamisongelmalle, mutta johtaa hyvin leveisiin (epäinformatiivisiin) luottamusväleihin, jos k on suuri. Muita ratkaisuja ovat ns. *suurimman absoluuttisen t-suhteen menetelmä* ja *Scheffen S-menetelmä*.

4.2 Luottamusjoukot

Esimerkkinä luottamusjoukoista johdetaan luottamusjoukko parametrivektorille $\boldsymbol{\beta}$ kokonaisuudessaan. Lähtökohdaksi otetaan testi nollahypoteesille $H : \boldsymbol{\beta} = \boldsymbol{\beta}_0$, jossa $\boldsymbol{\beta}_0$ on annettu $p \times 1$ vektori. Valitsemalla $\mathbf{A} = \mathbf{I}_p$ ja $\mathbf{c} = \boldsymbol{\beta}_0$ nähdään, että tämä nollahypoteesi on tyyppiä $\mathbf{A}\boldsymbol{\beta} = \mathbf{c}$. F-testisuureeksi saadaan (ks. s. 18)

$$F = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/pS^2 \overset{H}{\sim} F_{p,n-p},$$

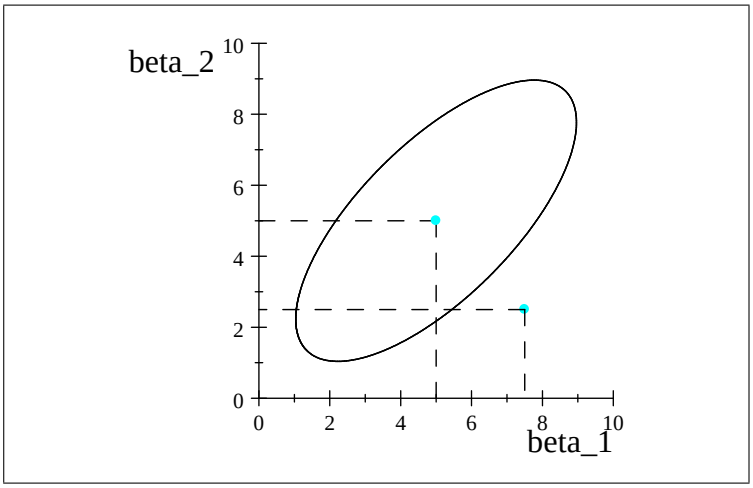
jossa merkinnät ovat kuten aikaisemmin. Jos $F_{p,n-p}(\alpha)$ on reaaliluku, joka toteuttaa $\mathsf{P}(F_{p,n-p} \geq F_{p,n-p}(\alpha)) = \alpha$, niin edellä todetun perusteella pätee kaikilla $\boldsymbol{\beta}_0$ ja σ^2

$$\mathsf{P}_{\boldsymbol{\beta}_0,\sigma^2}\left((\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)/pS^2 < F_{p,n-p}(\alpha)\right) = 1 - \alpha.$$

Parametrivektorin $\boldsymbol{\beta}$ luottamusjoukko luottamustasolla $1 - \alpha$ on näin ollen

$$\left\{\boldsymbol{\beta} \in \mathbb{R}^p : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/ps^2 < F_{p,n-p}(\alpha)\right\}.$$

Sen rajoittama pinta on $\mathbb{R}^p$:n ellipsoidi, jonka keskipiste on $\hat{\boldsymbol{\beta}}$ ja muodon määrää matriisi $\mathbf{X}'\mathbf{X}$. Tapauksessa $p = 2$ tilanne on oheisen kuvan kaltainen. Jos $p > 2$, on luottamusellipsoidien hahmottaminen hankalaa. Projektiot koordinaattiakseleille auttavat vain rajoitetusti, sillä esimerkiksi kuvan piste $(7.5, 2.5)$ kuuluu yksiulotteisiin luottamusväleihin, mutta ei luottamusjoukkoon. Tämä havainnollistaa myös sitä, mikä tekee useita parametrivektorin $\boldsymbol{\beta}$ komponentteja koskevien luottamusvälien tai t-testien muodostamisen hankalaksi.



Kuva 4.1. Parametrivektorin $\boldsymbol{\beta} = (\beta_1, \beta_2)$ luottamusellipsi. PNS-estimaatti $\hat{\boldsymbol{\beta}} = (5, 5)$.

Samaan tapaan kuin edellä (lähtemällä esimerkiksi nollahypoteesista $H : [\mathbf{0} \;\; \mathbf{I}_q]\,\boldsymbol{\beta} = \boldsymbol{\beta}_0$) voidaan muodostaa luottamusjoukkoja myös $\boldsymbol{\beta}$:n osavektoreille.

5 Empiirinen esimerkki

5.1 Aineisto ja tutkimusongelma

Eräässä kokeessa tutkittiin uuden ruokavalion vaikutusta vastasyntyneiden karitsojen painon nousuun. Koetta varten valittiin satunnaisesti 22 karitsaa. Niistä 11 muodosti koeryhmän, jota ruokittiin uutta ruokavaliota noudattaen, ja toiset 11 muodostivat kontrolliryhmän, jota ruokittiin vanhan ruokavalion mukaan. Kunkin karitsan painon nousu koeajalta mitattiin (selitettävä muuttuja, mitattu nauloissa). Selittäväksi muuttujaksi valittiin koeajan pituus, koska se ei ollut vakio ja vaikuttaa ilmeisesti koeajalla havaittavaan painon nousuun. Koeryhmässä koeajan pituus vaihteli 50 päivästä 90 päivään ja kontrolliryhmässä 54 päivästä 83 päivään. Käytettävissä oleva aineisto on esitetty graafisesti oheisessa kuvassa, johon on piirretty myös PNS-menetelmällä estimoidut regressiosuorat, jotka on saatu selittämällä kummassakin ryhmässä erikseen karitsojen painon nousua (y) koeajan pituudella (x). Vastaavien mallien estimointitulokset ovat seuraavat.

Koeryhmä: $$y_i = \underset{(4.68)}{0.64} + \underset{(0.064)}{0.53}\, x_i + \hat{\varepsilon}_i, \quad i = 1, ..., 11, \quad s_1^2 = 5.98$$

Kontrolliryhmä: $$y_i = \underset{(9.78)}{-25.80} + \underset{(0.13)}{0.75}\, x_i + \hat{\varepsilon}_i, \quad i = 12, ..., 22, \quad s_2^2 = 12.23.$$

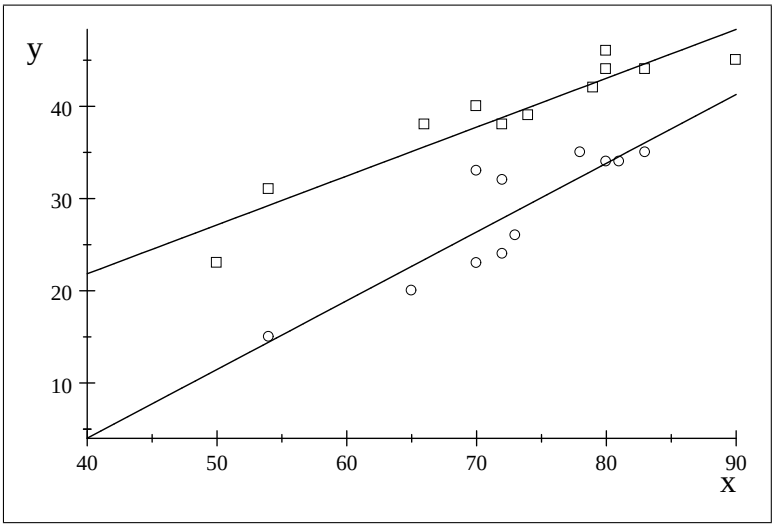
Huomaa, että tässä lineaarinen malli on ajateltava lokaalina approksimaationa, sillä se ei voi toimia, jos selittävän muuttujan arvot kasvavat kovin suuriksi. Tarkasteltavien selittävien muuttujien havaintoarvojen tapauksessa lineaarisuusoletus näyttää kuvan perusteella kuitenkin toimivan kohtuullisen hyvin.

Ensimmäinen kiinnostava kysymys koskee esitettyjen regressiosuorien yhdensuuntaisuutta eli voidaanko estimaattien 0.53 ja 0.75 välinen ero tulkita pelkästään satunnaisvaihtelusta johtuvaksi. Jos voidaan, on ruokavalion vaikutuksen mahdollinen ero riippumaton ruokinta-ajan pituudesta, jolloin regressiosuorien vakioiden ero mittaa sitä kaikilla selittävän muuttujan arvoilla. Jos regressiosuoria voidaan pitää yhdensuuntaisina, on seuraava kiinnostava kysymys näin ollen voidaanko regressiosuorien vakioiden 0.64 ja -25.80 ero tulkita pelkästään satunnaisvaihtelusta johtuvaksi.

Edellä esitettyjen kysymysten tutkiminen perustetaan lineaariseen malliin

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} & 0 \\ 0 & 1 & 0 & x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma_1 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix}, \tag{5.1}$$

jossa $\varepsilon_1, ..., \varepsilon_{22}$ ovat riippumattomia, $\varepsilon_i \sim \mathsf{N}\left(0, \sigma^2\right)$ ja parametriavaruus määritellään kuten aikaisemminkin. Koska tässä selittäjämatriisin ensimmäinen ja kolmas sarake ovat ortogonaaliset toista ja neljättä saraketta vastaan, seuraa PNS-estimaatin lausekkeesta, että PNS-estimointi tuottaa edellä esitetyt kaksi regressiosuoraa eli $\hat{\alpha}_1 = 0.64$,



Kuva 5.1. Karitsojen painonnousuun (y) ja koeajan pituuteen (x) liittyvä aineisto koeryhmän (neliöt) ja kontrolliryhmän (ympyrät) mukaan luokiteltuna sekä kummallekin ryhmälle PNS-menetelmällä estimoidut regressiosuorat

$\hat{\gamma}_1 = 0.53$, $\hat{\alpha}_2 = -25.80$ ja $\hat{\gamma}_2 = 0.75$. Virhevarianssiestimaatteja saadaan tietenkin vain yksi $s^2 = 9.10$. (Huomaa, että edellä esitetyt keskivirheet perustuvat eri virhevarianssiestimaatteihin.) $\frac{32\,195}{1248} = 25.797$ $\frac{5204}{8089} = 0.643\,34$

Ensimmäinen kiinnostava hypoteesi mallissa (5.1) on siis $\gamma_1 = \gamma_2$ ja, jos se jää voimaan, testataan hypoteesia $\alpha_1 = \alpha_2$. Ennen testaamista on kuitenkin syytä tutkia ovatko mallista tehdyt oletukset realistiset.

5.2 Mallin oletusten tarkistaminen

Koska kysymyksessä on satunnaisotanta, ei havaintojen tai virheiden riippumattomuusoletusta ole syytä epäillä, mutta normaalisuusoletuksen perusteleminen on vaikeampaa. Seuraavassa tarkastellaan ensin aivan yleisesti miten normaalisuusoletuksen realistisuutta voidaan tutkia.

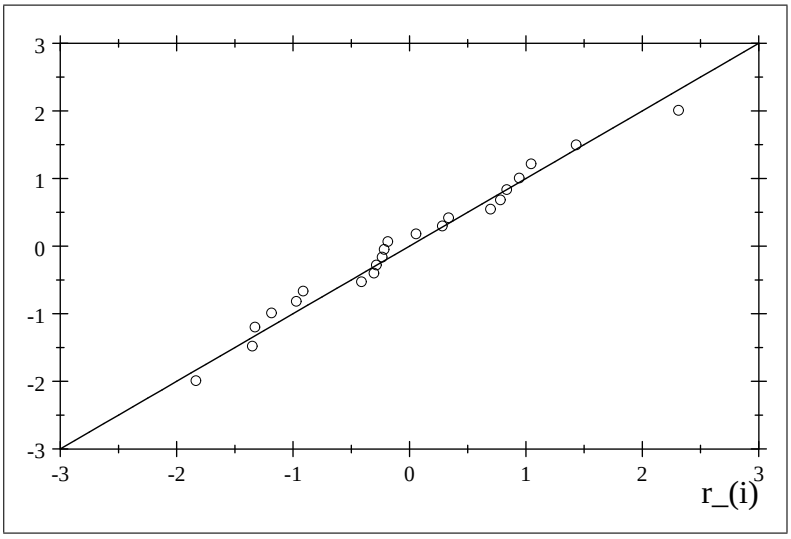
Aikaisemmin on todettu, että residuaalivektorilla $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ on esitys

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{P})\,\mathbf{y},$$

jossa $\mathbf{P} = [p_{ij}] = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$ on projektio (ks. jakso 2.1.1, yhtälö (2.4)). Tämän yhtälön ja Liitteen A.1 laskusääntöjen avulla voidaan todeta, että vastaavalle satunnaisvektorille pätee

$$\mathsf{E}\,(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0} \qquad \text{ja} \qquad \mathsf{Cov}\,(\hat{\boldsymbol{\varepsilon}}) = \sigma^2\,(\mathbf{I}_n - \mathbf{P})$$

(yksityiskohdat jätetään tehtäväksi). Koska $\mathsf{Cov}\,(\hat{\boldsymbol{\varepsilon}}) \neq \sigma^2\mathbf{I}_n\ (= \mathsf{Cov}\,(\boldsymbol{\varepsilon}))$, ovat residuaalit siis oikeankin mallin kyseessä ollessa korreloituneita ja niiden varianssit vaihtelevat havaintoyksiköstä toiseen (jälkimmäistä ominaisuutta nimitetään heteroskedastisuudeksi). Lisäksi, jos normaalisuusoletus pätee, noudattaa residuaalivektori singulaarista normaalijakaumaa ($\mathsf{r}\,(\mathsf{Cov}\,(\hat{\boldsymbol{\varepsilon}})) = \mathsf{r}\,(\mathbf{I}_n - \mathbf{P}) = n - p$, ks. Lauseen 2.1(ii)



Kuva 5.2. Mallin (5.1) järjestetyt standardoidut residuaalit arvoja $\Phi^{-1}((i-0.5)/n)$ $(i = 1, ..., 22)$ vastaan eli ns. normaalipaperipiirros (normal probability plot tai normal QQ plot).

todistus jaksossa 2.3).

Tutkittaessa virheiden ε_i normaalisuusoletuksen paikkansapitävyyttä residuaalien avulla, on residuaalit edellä sanotun perusteella järkevää standardoida varianssiltaan yhtäsuuriksi. Koska $\mathsf{Var}(\hat{\varepsilon}_i) = \sigma^2(1 - p_{ii})$, on *standardoidut residuaalit* luontevaa määritellä yhtälöllä

$$r_i = \hat{\varepsilon}_i / s\sqrt{1 - p_{ii}}.$$

Normaalijakauman ollessa oikea pätee approksimaatio $\mathsf{Var}(r_i) \approx 1$.

Virheiden normaalisuutta voidaan tutkia vertaamalla standardoitujen residuaalien jakaumaa normaalijakaumaan ns. kvantiilifunktiota käyttäen. Siinä piirretään pistekuvio, jonka toiselle akselille piirretään arvot $\Phi^{-1}((i-0.5)/n)$ $(i = 1, .., n)$, jossa Φ^{-1} on $\mathsf{N}(0, 1)$-jakauman kertymäfunktion käänteisfunktio, ja toiselle akselille suuruusjärjestykseen järjestetyt standardoidut residuaalit $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(n)}$. Jos mallin oletukset ja erityisesti normaalisuusoletus pitävät paikkansa, asettuvat kuvan pisteet suunnilleen origon kautta kulkevalle suoralle, jonka kulmakerroin on 45^o. Tarkasteltavassa esimerkissä järjestetyt standardoidut residuaalit $r_{(i)}$ piirretään vaaka-akselille ja arvot $\Phi^{-1}((i-0.5)/n)$ pystyakselille, mutta myös toisin päin piirrettyjä kuvioita näkee käytettävän.

Kuvaan 5.2 on piirretty edellä kuvattu pistekuvio, jonka mukaan selvää poikkeamaa normaalisuudesta ei ole havaittavissa. Koska havainnot ovat peräisin kahdesta eri ryhmästä, tutkitaan lisäksi ovatko havaintojen tai yhtäpitävästi virheiden varianssit samat molemmissa ryhmissä. Jos näin ei ole, ei malli (5.1) toteuta asetettuja oletuksia eikä lineaarisen mallin F-testi ole pätevä.

Koska virheiden riippumattomuusoletusta ei ole syytä epäillä, voidaan molemmille ryhmille estimoituja malleja tarkastella erikseen ja päätellä, että virhevarianssiestimaattoreille pätee $S_1^2 \underset{\text{(ii)}}{\parallel} S_2^2$. Koska normaalisuusoletuskin todettiin realistiseksi, voidaan Lauseesta 2.1(ii) päätellä edelleen, että vakiovarianssihypoteesin

voimassa ollessa $S_1^2/S_2^2 \sim \frac{1}{9}\chi_9^2/\frac{1}{9}\chi_9^2 \sim F_{9,9}$. Tämän suhteen havaituksi arvoksi saadaan $s_1^2/s_2^2 = 5.98/12.23 = 0.49$ ja, koska $\mathsf{P}(F_{9,9} \leq 0.49) = 0.15$, ei estimaattien s_1^2 ja s_2^2 ero ole hälyttävän suuri.

Edellä esitettyjen tarkastelujen perusteella voidaan mallia (5.1) käyttää hypoteesien testaamiseen.

5.3 Tilastollinen analyysi

Tarkastellaan nyt nollahypoteesin $\gamma_1 = \gamma_2$ testaamista mallissa (5.1). Tämän hypoteesin voimassa ollessa malliyhtälöstä (5.1) tulee

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{11} \\ Y_{12} \\ \vdots \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{11} \\ 0 & 1 & x_{12} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_{22} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \vdots \\ \varepsilon_{22} \end{bmatrix}, \tag{5.2}$$

jossa $\gamma = \gamma_1 = \gamma_2$. PNS-estimointi tuotti tuloksen

$$y_i = \underset{(4.98)}{-4.30}\, d_{i1} - \underset{(4.98)}{15.12} d_{i2} + \underset{(0.067)}{0.60}\, x_i + \hat{\varepsilon}_i, \quad i = 1, ..., 22, \quad s^2 = 9.76,$$

jossa d_{i1} ja d_{i2} viittaavat malliyhtälön (5.2) selittäjämatriisin ensimmäiseen ja toiseen sarakkeeseen. Mallien (5.1) ja (5.2) estimointitulosten perusteella saadaan F-testisuureen arvoksi (ks. jakso 3.1, s. 19)

$$F = \frac{(19 \times 9.76 - 18 \times 9.10)/1}{9.10} = 2.37.$$

Vastaavaksi P-arvoksi tulee

$$\text{P} = \mathsf{P}(F_{1,18} \geq 2.37) = 0.14,$$

joten painavaa näyttöä nollahypoteesia vastaan ei ole.

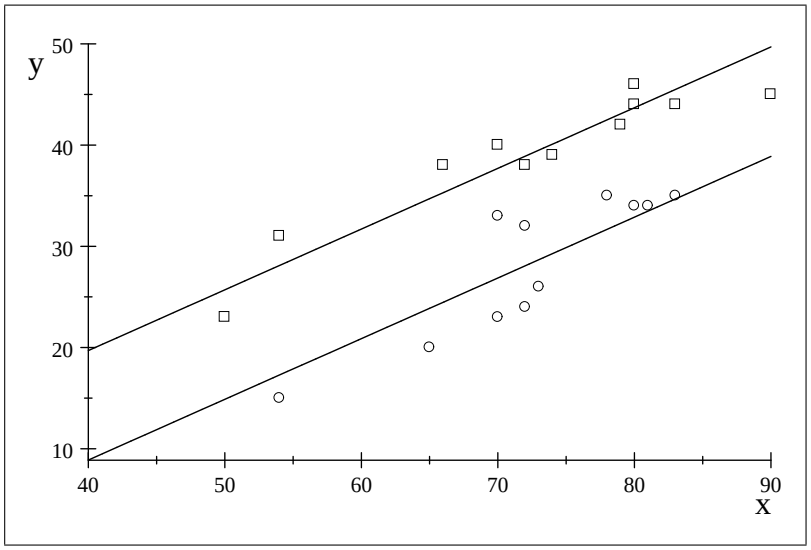
Mallin (5.2) estimointituloksista saadaan koeryhmälle ja kontrolliryhmälle regressiosuorat $y = -4.30 + 0.60x$ (koeryhmä) ja $y = -15.12 + 0.60x$ (kontrolliryhmä), jotka on piirretty Kuvaan 5.3 yhdessä aineiston kanssa.

Toisena testattavana hypoteesina on $\alpha_1 = \alpha_2$, jonka tutkiminen perustetaan malliin (5.2). Tämän hypoteesin voimassa ollessa malliyhtälöksi saadaan $Y_i = \alpha + \gamma x_i + \varepsilon_i$, $i = 1, ..., 22$, jossa $\alpha = \alpha_1 = \alpha_2$. PNS-estimointi tuotti tuloksen

$$y_i = \underset{(10.16)}{-9.71} + \underset{(0.14)}{0.60} x_i + \hat{\varepsilon}_i, \quad i = 1, ..., 22, \quad s^2 = 41.46.$$

Vastaava regressiosuora asettuu Kuvan 5.3 regressiosuorien väliin. Halutun F-testisuureen arvoksi saadaan

$$F = \frac{(20 \times 41.46 - 19 \times 9.76)/1}{9.76} = 65.97,$$



Kuva 5.3. Kuvan 5.1 aineistoon piirretyt mallin (5.2) estimointituloksiin perustuvat samansuuntaiset regressiosuorat.

joka on niin suuri, että vastaava P-arvo $\mathrm{P} = \mathsf{P}\left(F_{1,19} \geq 65.97\right) \approx 0$. Nollahypoteesi on siis syytä hylätä.

Jos karitsojen nopeaa painon nousua pidetään tavoiteltavana, osoittaa saatu tulos, että uusi ruokavalio on systemaattisesti vanhaa parempi, sillä sen paremmuus ei riipu (tarkasteltujen) ruokinta-aikojen pituuksista. Sillä, millaisia tulkinnallisia seuraamuksia regressiosuorien leikkaamisella tai yleisemmin kaltevuuksien erolla olisi, jätetään pohdittavaksi.

Verrattaessa vanhan ja uuden ruokavalion eroa on kiinnostava parametri (mallissa (5.2)) $\alpha_1 - \alpha_2$, jolle on aiheellista muodostaa luottamusväli. Jaksossa 4.1 esitetty yleinen menettely soveltuu valinnalla $\mathbf{a}' = [1 \quad -1 \quad 0]$. Luottamusväliä varten lasketaan

$$s\sqrt{\mathbf{a}'\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{a}} = 1.77 \qquad \text{ja} \qquad \hat{\alpha}_1 - \hat{\alpha}_2 = -4.30 + 15.12 = 10.82.$$

Koska $t_{19}\left(0.05/2\right) = 2.09$, saadaan 95%:n luottamusväliksi

$$10.82 - 2.09 \times 1.77 < \alpha_1 - \alpha_2 < 10.82 + 2.09 \times 1.77$$

eli

$$7.12 < \alpha_1 - \alpha_2 < 14.52.$$

Toisin sanoen, kun ruokinta-ajan pituus on n. 50 - 90 päivää, voidaan uudella ruokavaliolla odottaa päästävän 95%:n varmuudella n. 7 - 14.5 naulaa suurempaan karitsan painonnousuun kuin vanhalla ruokavaliolla.

6 Varianssianalyysia

6.1 Yksisuuntainen varianssianalyysi

Kuten jaksossa 1.3 mainittiin, tarkastellaan yksisuuntaisessa varianssianalyysissa p:tä ryhmää ja niistä satunnaisotannalla poimittuja havaintoja. Mielenkiinnon kohteena on ryhmien odotusarvojen mahdolliset erot. Asetelma on kaaviona

	Havainnot	Keskiarvot
Ryhmä 1	$y_{11}, y_{12}, ..., y_{1n_1}$	$\bar{y}_1$
Ryhmä 2	$y_{21}, y_{22}, ..., y_{2n_2}$	$\bar{y}_2$
$\vdots$	$\vdots$	$\vdots$
Ryhmä p	$y_{p1}, y_{p2}, ..., y_{pn_p}$	$\bar{y}_p$

jossa $\bar{y}_j = \left(y_{j1} + \cdots + y_{jn_j}\right)/n_j \ (j = 1, ..., p)$.

Tilastollista mallia varten oletetaan, että havaintoja vastaavat satunnaismuuttujat $Y_{11}, ..., Y_{1n_1}, Y_{21}, ..., Y_{2n_2}, ..., Y_{p1}, ..., Y_{pn_p}$ ovat riippumattomia ja $Y_{ji} \sim \mathsf{N}\left(\mu_j, \sigma^2\right)$ tai yhtäpitävästi, että

$$Y_{ji} = \mu_j + \varepsilon_{ji}, \quad i = 1, ..., n_j, \ j = 1, ..., p,$$

jossa satunnaismuuttujat $\varepsilon_{ji} \sim \mathsf{N}\left(0, \sigma^2\right)$ ovat riippumattomia ja $\mu_j \in \mathbb{R}$. Käyttäen matriisimerkintöjä saadaan yhtälö

$$\begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1n_1} \\ Y_{21} \\ \vdots \\ Y_{2n_2} \\ \vdots \\ \vdots \\ Y_{p1} \\ \vdots \\ Y_{pn_p} \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \\ \varepsilon_{21} \\ \vdots \\ \varepsilon_{2n_2} \\ \vdots \\ \vdots \\ \varepsilon_{p1} \\ \vdots \\ \varepsilon_{pn_p} \end{bmatrix},$$

josta nähdään, että kysymyksessä on lineaarisen mallin erikoistapaus. Mielenkiinnon kohteena on nollahypoteesi

$$H : \mu_1 = \cdots = \mu_p,$$

joka voidaan lausua myös muodossa $\mu_1 - \mu_p = \mu_2 - \mu_p = \cdots = \mu_{p-1} - \mu_p = 0$ tai matriisein

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & -1 \\ 0 & 1 & 0 & \cdots & \cdots & -1 \\ \vdots & \vdots & \ddots & & & \vdots \\ \vdots & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & & 1 & -1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \vdots \\ \mu_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \end{bmatrix}.$$

Nollahypoteesi on siis lineaarista muotoa $\mathbf{A}\boldsymbol{\mu} = \mathbf{0}$, jossa $\mathbf{A}$ on $(p-1) \times p$ matriisi ja $\mathsf{r}(\mathbf{A}) = p - 1$.

Edellä sanotun perusteella voidaan nollahypoteesia testata jaksossa 3.1 esitetyllä F-testillä, jota varten johdetaan vapaa residuaalineliösumma $\mathsf{S}(\boldsymbol{\mu})$ ja rajoitettu residuaalineliösumma $\mathsf{S}(\boldsymbol{\mu}_H)$. Edellinen saadaan (esimerkiksi) minimoimalla jäännösneliösumma

$$\mathsf{S}(\boldsymbol{\mu}) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \mu_j)^2 .$$

Asettamalla derivaatta μ_j:n suhteen nollaksi saadaan yhtälö

$$\partial \mathsf{S}(\boldsymbol{\mu}) / \partial \mu_j = -2 \sum_{i=1}^{n_j} (y_{ji} - \mu_j) = 0,$$

jonka ratkaisuna saadaan μ_j:n PNS-estimaatti $\hat{\mu}_j = \bar{y}_j$. Näin ollen,

$$\mathsf{S}(\hat{\boldsymbol{\mu}}) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2 .$$

Nollahypoteesin voimassa ollessa malliyhtälönä on

$$Y_{ji} = \mu_0 + \varepsilon_{ji}, \quad i = 1, ..., n_j, \ j = 1, ..., p,$$

jossa $\mu_0 = \mu_1 = \cdots = \mu_p$. Tästä nähdään, että parametrin μ_0 PNS-estimaatti on

$$\hat{\mu}_0 = \bar{y} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} y_{ji} = \frac{1}{n} \sum_{j=1}^{p} n_j \bar{y}_j \qquad (n = n_1 + \cdots + n_p),$$

joten

$$\mathsf{S}(\hat{\boldsymbol{\mu}}_H) = \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y})^2 .$$

Tämä riittäisi testisuureen muodostamiseksi, mutta esitetään rajoitettu residuaalineliösumma kuitenkin seuraavalla vaihtoehtoisella tavalla:

$$\begin{aligned} \mathsf{S}(\hat{\boldsymbol{\mu}}_H) &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j + \bar{y}_j - \bar{y})^2 \\ &= \sum_{j=1}^{p} \sum_{i=1}^{n_j} (y_{ji} - \bar{y}_j)^2 + \sum_{j=1}^{p} \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 . \end{aligned}$$

Tässä jälkimmäinen yhtälö seuraa laskelmasta

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)(\bar{y}_j-\bar{y})=\sum_{j=1}^{p}(\bar{y}_j-\bar{y})\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)=0,$$

sillä

$$\sum_{i=1}^{n_j}(y_{ji}-\bar{y}_j)=\sum_{i=1}^{n_j}y_{ji}-n_j\bar{y}_j=0.$$

Kaiken kaikkiaan edellä todetusta seuraa

$$\mathsf{S}(\hat{\boldsymbol{\mu}}_H)-\mathsf{S}(\hat{\boldsymbol{\mu}})=\sum_{j=1}^{p}\sum_{i=1}^{n_j}(\bar{y}_j-\bar{y})^2=\sum_{j=1}^{p}n_j(\bar{y}_j-\bar{y})^2.$$

F-testisuureeksi saadaan näin ollen (ks. jakso 3.1, s. 19)

$$F=\frac{\sum_{j=1}^{p}n_j\left(\bar{Y}_j-\bar{Y}\right)^2/(p-1)}{\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(Y_{ji}-\bar{Y}_j\right)^2/(n-p)}\overset{H}{\sim}F_{p-1,n-p}.$$

Usein sanotaan, että testisuure on ryhmien välisen varianssin ja ryhmien sisäisen varianssin suhde, mistä nimitys varianssianalyysi tulee. Huomaa, että testisuureen nimittäjässä on $\mathsf{S}(\hat{\boldsymbol{\mu}})/(n-p)=S^2$.

Jos nollahypoteesi hylätään, voidaan seuraavaksi tutkia miten odotusarvot $\mu_1,...,\mu_p$ poikkeavat toisistaan. Kiinnostavia hypoteeseja ovat esimerkiksi $\mu_{j_1}-\mu_{j_2}=0$ tai yleisemmin ns. *kontrastit* $\mathbf{a}'\boldsymbol{\mu}=0$, jossa $\mathbf{a}'=[a_1\ \cdots\ a_p]$ toteuttaa ehdon $a_1+\cdots+a_p=0$. Mallin matriisiesityksestä nähdään, että selittäjämatriisille pätee $\mathbf{X}'\mathbf{X}=\mathsf{diag}\,[n_1\ \cdots\ n_p]$, joten Lauseen 2.1(i) nojalla

$$\hat{\boldsymbol{\mu}}=\left[\bar{Y}_1\ \cdots\ \bar{Y}_p\right]'\sim\mathsf{N}\left(\boldsymbol{\mu},\sigma^2\mathsf{diag}\left[\frac{1}{n_1}\ \cdots\ \frac{1}{n_p}\right]\right).$$

Nollahypoteesin $\mathbf{a}'\boldsymbol{\mu}=0$ voimassa ollessa pätee näin ollen

$$\mathbf{a}'\hat{\boldsymbol{\mu}}\sim\mathsf{N}\left(0,\sigma^2\sum_{j=1}^{p}a_j^2/n_j\right)$$

(ks. Liite A.2.4). Tämä johtaa F-testisuureeseen

$$F=\frac{\left(\sum_{j=1}^{p}a_j\bar{Y}_j\right)^2}{S^2\sum_{j=1}^{p}a_j^2/n_j}\overset{H}{\sim}F_{1,n-p},$$

jossa

$$S^2=\frac{1}{n-p}\mathsf{S}(\hat{\boldsymbol{\mu}})=\frac{1}{n-p}\sum_{j=1}^{p}\sum_{i=1}^{n_j}\left(Y_{ji}-\bar{Y}_j\right)^2.$$

Vaihtoehtoisesti voidaan käyttää t-testisuuretta ja t_{n-p}-jakaumaa (vrt. s. 22).

Kontrasteille, jotka todetaan nollasta poikkeaviksi, voidaan muodostaa luottamusvälejä jaksossa 4.1 esitettyä menettelyä käyttäen. Yleisesti kontrastin $\mathbf{a}'\boldsymbol{\mu}$ luottamusväliksi luottamustasolla $1-\alpha$ saadaan

$$\sum_{j=1}^{p} a_j \bar{y}_j \pm t_{n-p}\left(\alpha/2\right) s \sqrt{\sum_{j=1}^{p} a_j^2/n_j}.$$

Useita luottamusvälejä muodostettaessa on kuitenkin syytä muistaa jaksossa 4.1 mainitut hankaluudet.

6.2 Empiirinen esimerkki

Eräässä kokeessa tutkittiin miten toisaalta veden ja toisaalta erilaisten 'urheilujuomien' nauttimisella voidaan vaikuttaa maitohapon kerääntymiseen lihaksiin pitkän matkan juoksijoilla. Kokeessa 50 satunnaisesti valittua juoksijaa jaettiin satunnaisesti viiteen 10 juoksijan ryhmään, joille annettiin eri juomia seuraavasti.

1	Vesi	
2	Urheilujuoma A1	Sama vaikuttava aine
3	Urheilujuoma A2	eri pitoisuuksina
4	Urheilujuoma B1	Sama vaikuttava aine
5	Urheilujuoma B2	eri pitoisuuksina

Juoksumatkan pituus oli 10 mailia ja juomia nautittiin kiinteät määrät ennen juoksua, juoksun puolivälissä ja juoksun päätyttyä. Tämän jälkeen mitattiin maitohappopitoisuudet, joista saatiin tiivistettynä seuraava aineisto.

	Vesi	A1	A2	B1	B2	
$\bar{y}_j$	33.3	32.6	30.9	29.0	26.1	$\bar{y} = 30.7$
s_j^2	13.1	14.2	12.2	13.9	14.2	$s^2 = 13.37$
n_j	10	7	10	8	6	$n = 41$

Vaikka juoksijoiden arvioitiinkin olevan suunnilleen yhtä hyvässä kunnossa, 9 heistä keskeytti matkan. Keskeytysten ei kuitenkaan katsottu riippuvan nautituista nesteistä tai maitohapon kertymisestä.

Olettaen, että havainnot voidaan tulkita riippumattomiksi ja että ryhmästä j saadut havainnot noudattavat $\mathsf{N}\left(\mu_j, \sigma^2\right)$-jakaumaa $(j = 1, ..., 5)$, voidaan soveltaa

varianssianalyysia ja testata nollahypoteesia $H : \mu_1 = \cdots = \mu_5$ edellisessä jaksossa esitetyllä tavalla. Koska ryhmiä koskeva vakiovarianssioletus saattaa tuntua kyseenalaiselta, todetaan ensin, että

$$\max_{1\le j\le 5} s_j^2 / \min_{1\le j\le 5} s_j^2 = 14.2/12.2 = 1.16.$$

Koska $\mathsf{P}(F_{6,9} > 1.16) \approx 0.6$, voidaan vakiovarianssioletusta pitää kohtuullisena (vrt. Lause 2.1). Huomaa kuitenkin, että toisin kuin jaksossa 5.2 F-jakauman käyttö ei tässä ole tarkkaan ottaen oikein (pohdi syytä tälle).

F-testisuure edellä esitetylle nollahypoteesille voidaan laskea sijoittamalla aineistossa esitetyt suureet testisuureen yleiseen lausekkeeseen, jolloin saadaan $F = 4.55$. Vastaava P-arvo on $\text{P} = \mathsf{P}(F_{4,36} \ge 4.55) = 0.004$, joka on niin pieni, että nollahypoteesi voidaan hylätä.

Seuraavaksi on kiinnostavaa hakea vastauksia (ainakin) seuraaviin vertailuihin.

1. Vesi vastaan urheilujuomat
2. Urheilujuoma A vastaan urheilujuoma B
3. Urheilujuoma A1 vastaan urheilujuoma A2
4. Urheilujuoma B1 vastaan urheilujuoma B2

Näihin liittyvät hypoteesit ovat kontrastien avulla esitettyinä

$$\begin{aligned}
&H_1 : \mu_1 - \tfrac{1}{4}(\mu_2 + \mu_3 + \mu_4 + \mu_5) = 0\\
&H_2 : \tfrac{1}{2}(\mu_2 + \mu_3) - \tfrac{1}{2}(\mu_4 + \mu_5) = 0\\
&H_3 : \mu_2 - \mu_3 = 0\\
&H_4 : \mu_4 - \mu_5 = 0.
\end{aligned}$$

Valitsemalla vektorin $\mathbf{a} = [a_1 \ \cdots \ a_5]'$ komponentit sopivasti voidaan näitä hypoteeseja testata testisuureella

$$F = \frac{\left(\sum_{j=1}^5 a_j \bar{y}_j\right)^2}{s^2 \sum_{j=1}^5 a_j^2/n_j},$$

jonka saamia arvoja verrataan $F_{1,36}$-jakauman prosenttipisteisiin. Tämä johtaa seuraaviin tuloksiin.

$$\begin{aligned}
&H_1 : F = 7.46, \quad \text{P} = 0.01\\
&H_2 : F = 9.87, \quad \text{P} = 0.003\\
&H_3 : F = 0.89, \quad \text{P} = 0.35\\
&H_4 : F = 2.15, \quad \text{P} = 0.15
\end{aligned}$$

Hypoteeseihin H_1 ja H_2 liittyvät P-arvot ovat niin pieniä, että nämä hypoteesit voidaan hylätä. Sen sijaan hypoteeseja H_3 ja H_4 vastaan ei ole näyttöä.

Tulosten mukaan veden ja urheilujuomien välillä on siis eroa ja samoin urheilujuomien A ja B välillä on eroa. Sen sijaan sillä ei ole merkitystä kumpaa kahdesta tutkitusta pitoisuudesta urheilujuomasta A tai B käytetään.

6.3 Kaksisuuntainen varianssianalyysi[16]

6.3.1 Ongelman asettelu

Tarkastellaan koetilannetta, jossa tutkitaan kahden tekijän A ja B vaikutusta muuttujaan y. Oletetaan, että A:lla on J eri tasoa ja B:llä K eri tasoa, jolloin eri tasokombinaatioita on kaikkiaan JK kappaletta. Jaksossa 1.3 mainitussa esimerkkitapauksessa 'tekijä' A oli vehnälajike ja 'tekijä' B on lannoitetyyppi, joita kumpaakin oli kaksi eli 'tasojen' lukumäärät olivat $J = K = 2$. Tarkasteltava muuttuja y oli satomäärä, josta saatiin havaintoja neljästä eri ryhmästä tai 'tasokombinaatiosta'.

Oletetaan yksinkertaisuuden vuoksi, että kustakin ryhmästä on käytettävissä m y:n havaintoa.[17] Olkoon y_{jki} y:n i. havainto ryhmässä (j,k) ja μ_{jk} sen odotusarvo. Tilannetta voidaan havainnollistaa seuraavan kaavion avulla.

		B-tekijän tasot			
		1	2	$\cdots\cdots$	K
A-tekijän tasot	1	μ_{11}	μ_{12}	$\cdots\cdots$	μ_{1K}
	2	μ_{21}	μ_{22}	$\cdots\cdots$	μ_{2K}
	$\vdots$	$\vdots$	$\vdots$	$\cdots\cdots$	$\vdots$
	J	μ_{J1}	μ_{J2}	$\cdots\cdots$	μ_{JK}

Olettamalla, että havaintoja vastaavat satunnaismuuttujat Y_{jki} ovat riippumattomia ja $\mathsf{N}\left(\mu_{jk}, \sigma^2\right)$-jakautuneita päädytään lineaariseen malliin, joka voidaan esittää yhtälönä

$$Y_{jki} = \mu_{jk} + \varepsilon_{jki}, \quad j = 1, ..., J, k = 1, ..., K, i = 1, ..., m, \tag{6.1}$$

jossa virheet $\varepsilon_{jki} \sim \mathsf{N}\left(0, \sigma^2\right)$ ovat riippumattomia ja $\mu_{jk} \in \mathbb{R}$.

Kiinnostava kysymys on poikkeavatko odotusarvot μ_{jk} toisistaan ja johtuvatko mahdolliset poikkeamat A-tekijästä, B-tekijästä vai niiden yhteisvaikutuksesta. Jotta nämä kysymykset voitaisiin muotoilla tilastollisina hypoteeseina, merkitään

$$\bar{\mu}_{j\cdot} = \frac{1}{K}\sum_{k=1}^{K}\mu_{jk}, \quad \bar{\mu}_{\cdot k} = \frac{1}{J}\sum_{j=1}^{J}\mu_{jk} \quad \text{ja} \quad \mu = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\mu_{jk}.$$

Tällöin

$$\mu_{jk} = \mu + \underbrace{\left(\bar{\mu}_{j\cdot} - \mu\right)}_{=\alpha_j} + \underbrace{\left(\bar{\mu}_{\cdot k} - \mu\right)}_{=\gamma_k} + \underbrace{\left(\mu_{jk} - \bar{\mu}_{j\cdot} - \bar{\mu}_{\cdot k} + \mu\right)}_{=\delta_{jk}},$$

joten malliyhtälö (6.1) voidaan kirjoittaa

$$Y_{jki} = \mu + \alpha_j + \gamma_k + \delta_{jk} + \varepsilon_{jki}, \quad j = 1, ..., J,\ k = 1, ..., K,\ i = 1, ..., m, \tag{6.2}$$

[16] Tämä jakso ei sisälly kurssiin.

[17] Tämä oletus helpottaa mallin käsittelyä ja luultavasti myös ymmärtämistä. Periaatteessa kaikki esitettävät tulokset voidaan yleistää tapaukseen, jossa havaintojen lukumäärä ryhmissä vaihtelee.

jossa

$$\begin{aligned}\alpha_j &= A\text{-tekijän vaikutus (riippuu vain } j\text{:stä)}\\ \gamma_k &= B\text{-tekijän vaikutus (riippuu vain } k\text{:sta)}\\ \delta_{jk} &= \text{yhteisvaikutus (riippuu sekä } j\text{:stä että } k\text{:sta).}\end{aligned}$$

Huomaa, että konstruktiosta seuraa ehdot

$$\sum_{j=1}^{J}\alpha_j = 0, \quad \sum_{k=1}^{K}\gamma_k = 0, \quad \sum_{j=1}^{J}\delta_{jk} = 0 \quad \forall k \quad \text{ja} \quad \sum_{k=1}^{K}\delta_{jk} = 0 \quad \forall j. \tag{6.3}$$

On selvää, että alkuperäisessä malliyhtälössä (6.1) on vähemmän parametreja kuin sen uudelleen parametroidussa muodossa (6.2). Tämä merkitsee, että jälkimmäisen parametrit eivät ole yksikäsitteisiä ilman lisäehtoja. Esimerkiksi tapauksessa $J = K = 2$ odotusarvoja μ_{jk} on neljä kappaletta, jotka esitetään yhtälössä (6.2) käyttäen yhdeksää parametria (μ, kaksi α-parametria, kaksi γ-parametria ja neljä δ-parametria). Edellä 'ylimääräiset' parametrit saadaan eliminoiduiksi summaehdoilla (6.3), joihin päädyttiin määrittelemällä parametrit μ, α_j, γ_k ja δ_{jk} sopivasti. Näiden ehtojen asemesta on mahdollista käyttää muitakin ehtoja.

Jos malliyhtälö (6.2) esitetään lineaarisen mallin matriisiesitystä käyttäen, ei sen (nollista ja ykkösistä muodostuva) selittäjämatriisi ole täyttä sarakeastetta (tämän toteaminen jätetään harjoitustehtäväksi, jossa riittää tarkastella tapausta $J = K = 2$). Tämä piirre liittyy edellä tarkasteltuun mallin (6.2) parametrien yksikäsitteisyyden puutteeseen. Kaksisuuntainen varianssianalyysimalli on esimerkki ns. vajaaasteisesta lineaarisesta mallista, jossa normaaliyhtälöillä ei ole yksikäsitteistä ratkaisua ilman parametreille asetettavia lisäehtoja. Seuraavassa tämä mallin piirre ei kuitenkaan aiheuta mitään ongelmia, koska parametrien PNS-estimointi sujuu summausehtoja (6.3) käyttäen helposti. Kuten edellä mainittiin, voidaan nämä ehdot korvata vaihtoehtoisilla ehdoilla, joiden avulla mallin teoria voidaan esittää. Tässä esityksessä on päädytty käyttämään summausehtoja (6.3) niiden havainnollisuuden vuoksi.

Tarkastellaan nyt lähemmin yhteisvaikutusta ja sen kuvaamista parametreilla δ_{jk}. On luontevaa ajatella, että yhteisvaikutusta on, jos A-tekijän vaikutus riippuu B-tekijän tasosta. Jos yhteisvaikutusta sen sijaan ei ole, riippuu erotus $\mu_{j_1k} - \mu_{j_2k}$ vain indekseistä j_1 ja j_2, mutta ei indeksistä k. Formaalisti tämä merkitsee, että jollain funktiolla ϕ pätee $\mu_{j_1k} - \mu_{j_2k} = \phi(j_1, j_2)$ kaikilla k:n arvoilla ja siten

$$\mu_{j_1k} - \mu_{j_2k} = \frac{1}{K}\sum_{k=1}^{K}\phi(j_1, j_2) = \frac{1}{K}\sum_{k=1}^{K}\left(\mu_{j_1k} - \mu_{j_2k}\right) = \bar{\mu}_{j_1\cdot} - \bar{\mu}_{j_2\cdot} \quad \forall j_1, j_2$$

eli $\mu_{j_1k} - \bar{\mu}_{j_1\cdot} = \mu_{j_2k} - \bar{\mu}_{j_2}$ kaikilla j_1, j_2. Tämä osoittaa, että $\mu_{jk} - \bar{\mu}_{j\cdot}$ ei riipu indeksistä j, joten jollain funktiolla ψ pätee $\mu_{jk} - \bar{\mu}_{j\cdot} = \psi(k)$ kaikilla j ja siten

$$\mu_{jk} - \bar{\mu}_{j\cdot} = \frac{1}{J}\sum_{j=1}^{J}\psi(k) = \frac{1}{J}\sum_{j=1}^{J}\left(\mu_{jk} - \bar{\mu}_{j\cdot}\right) = \bar{\mu}_{\cdot k} - \mu \quad \forall j, k$$

eli $\delta_{jk} = 0$ kaikilla j, k. Samaan tulokseen voidaan päätyä myös vaihtamalla indeksien j ja k roolit. Tämä osoittaa, että hypoteesiksi "ei yhteisvaikutusta" on mielekästä valita

$$H_{AB} : \delta_{jk} = 0, \quad j = 1, ..., J-1, \ k = 1, ..., K-1.$$

Indeksoinnissa voi olla $J-1$ ja $K-1$, koska ehdoista $\sum_{j=1}^{J} \delta_{jk} = \sum_{k=1}^{K} \delta_{jk} = 0$ ja hypoteesista H_{AB} seuraa, että $\delta_{jk} = 0$ pätee kaikilla $j = 1, ..., J$ ja kaikilla $k = 1, ..., K$. Hypoteesi H_{AB} on lineaarista tyyppiä $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$, jossa vektori $\boldsymbol{\beta}$ sisältää kaikki parametrit μ, α_j, γ_k ja δ_{jk} ja matriisin $\mathbf{A}$ aste on sen rivien lukumäärä $(J-1)(K-1)$. Esimerkiksi tapauksessa $J = K = 2$, $\boldsymbol{\beta}$ on muotoa $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_*' & \delta_{11} & \delta_{12} & \delta_{21} & \delta_{22} \end{bmatrix}'$ ja $\mathbf{A} = [\mathbf{0}' \ \ 1 \ \ 0 \ \ 0 \ \ 0]$, joten $\mathsf{r}(\mathbf{A}) = 1$ (tässä vektori $\boldsymbol{\beta}_*$ sisältää parametrit μ, α_1, α_2, γ_1 ja γ_2).

Jos yhteisvaikutusta ei ole, voidaan A- ja B-tekijöitä käsitellä toisistaan riippumatta ja kysyä esimerkiksi onko kaikilla A:n tasoilla sama vaikutus. Jos näin on, niin kullakin k:n arvolla μ_{jk} ei riipu indeksistä j, joten jollain funktiolla θ pätee $\mu_{jk} = \theta(k)$ kaikilla j ja edelleen

$$\mu_{jk} = \frac{1}{J}\sum_{j=1}^{J} \theta(k) = \frac{1}{J}\sum_{j=1}^{J} \mu_{jk} = \bar{\mu}_{\cdot k} \quad \forall j, k.$$

Yhdistämällä tämä ehtoon $\delta_{jk} = \mu_{jk} - \bar{\mu}_{j\cdot} - \bar{\mu}_{\cdot k} + \mu = 0$ saadaan $\bar{\mu}_{j\cdot} - \mu = \alpha_j = 0$. Tämä pätee kaikilla j:n arvoilla, joten hypoteesiksi "ei A-vaikutusta" saadaan

$$H_A : \alpha_j = 0, \quad j = 1, ..., J-1,$$

jossa indeksoinnissa on $J-1$ summausehdon $\sum_{j=1}^{J} \alpha_j = 0$ perusteella. Aivan samalla tavalla nähdään, että hypoteesiksi "ei B-vaikutusta" tulee

$$H_B : \gamma_k = 0, \quad j = 1, ..., K-1.$$

Kuten yhteisvaikutuksen tapauksessa nähdään, että hypoteesit H_A ja H_B ovat lineaarisia ja matriisin $\mathbf{A}$ aste on sen rivien lukumäärä eli $J-1$ tai $K-1$.

On syytä huomata, että hypoteesien H_A ja H_B testaaminen ei ole mielekästä, jos hypoteesi H_{AB} on hylätty, sillä tällöin esimerkiksi hypoteesit H_A ja H_B eivät ole yhtäpitäviä ehtojen $\bar{\mu}_{j\cdot} - \mu = 0$ ja $\bar{\mu}_{\cdot k} - \mu = 0$ kanssa kuten edellä. Tämä on intuitiivisesti luonnollista, sillä jos yhteisvaikutusta esiintyy, täytyy sekä A- että B-vaikutusten eli ns. päävaikutusten myös esiintyä.

6.3.2 Hypoteesien testaaminen

Tarkastellaan nyt edellisessä jaksossa esitettyjen hypotesien testaamista käyttäen F-testisuureen residuaalineliösummaesitystä (ks. jakso 3.1, s. 19). Residuaalineliösummien johtamiseksi merkitään $n = mJK$ (= kaikkien havaintojen lkm) ja

$$\begin{aligned} \bar{y} &= \frac{1}{n}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{i=1}^{m} y_{jki}, \quad \bar{y}_{j\cdot\cdot} = \frac{1}{mK}\sum_{k=1}^{K}\sum_{i=1}^{m} y_{jki} \\ \bar{y}_{\cdot k\cdot} &= \frac{1}{mJ}\sum_{j=1}^{J}\sum_{i=1}^{m} y_{jki}, \quad \bar{y}_{jk\cdot} = \frac{1}{m}\sum_{i=1}^{m} y_{jki}. \end{aligned}$$

Määritellään $\bar{\varepsilon}$, $\bar{\varepsilon}_{j\cdot\cdot}$, $\bar{\varepsilon}_{\cdot k\cdot}$ ja $\bar{\varepsilon}_{jk\cdot}$ samalla tavalla ja kirjoitetaan

$$\varepsilon_{jki} = \bar{\varepsilon} + (\bar{\varepsilon}_{j\cdot\cdot} - \bar{\varepsilon}) + (\bar{\varepsilon}_{\cdot k\cdot} - \bar{\varepsilon}) + (\bar{\varepsilon}_{jk\cdot} - \bar{\varepsilon}_{j\cdot\cdot} - \bar{\varepsilon}_{\cdot k\cdot} + \bar{\varepsilon}) + (\varepsilon_{jki} - \bar{\varepsilon}_{jk\cdot}).$$

Neliöimällä puolittain ja summaamalla yli indeksien j, k ja i nähdään, että ristitulotermit häviävät (ks. $\mathsf{S}(\boldsymbol{\mu}_H)$:n johto yksisuuntaisessa varianssianalyysissa s. 32-33). Näin ollen,

$$\begin{aligned}\sum_{j,k,i} \varepsilon_{jki}^2 &= \sum_{j,k,i} \bar{\varepsilon}^2 + \sum_{j,k,i} (\bar{\varepsilon}_{j\cdot\cdot} - \bar{\varepsilon})^2 + \sum_{j,k,i} (\bar{\varepsilon}_{\cdot k\cdot} - \bar{\varepsilon})^2 \\ &\quad + \sum_{j,k,i} (\bar{\varepsilon}_{jk\cdot} - \bar{\varepsilon}_{j\cdot\cdot} - \bar{\varepsilon}_{\cdot k\cdot} + \bar{\varepsilon})^2 + \sum_{j,k,i} (\varepsilon_{jki} - \bar{\varepsilon}_{jk\cdot})^2 .\end{aligned}$$

Sijoittamalla tähän $\varepsilon_{jki} = y_{jki} - \mu - \alpha_j - \gamma_k - \delta_{jk}$ ja käyttämällä summausehtoja (6.3) saadaan

$$\begin{aligned}\sum_{j,k,i} (y_{jki} - \mu - \alpha_j - \gamma_k - \delta_{jk})^2 &= \sum_{j,k,i} (\bar{y} - \mu)^2 + \sum_{j,k,i} (\bar{y}_{j\cdot\cdot} - \bar{y} - \alpha_j)^2 \\ &\quad + \sum_{j,k,i} (\bar{y}_{\cdot k\cdot} - \bar{y} - \gamma_k)^2 \\ &\quad + \sum_{j,k,i} (y_{jk\cdot} - \bar{y}_{j\cdot\cdot} - \bar{y}_{\cdot k\cdot} + \bar{y} - \delta_{jk})^2 \qquad (6.4) \\ &\quad + \sum_{j,k,i} (y_{jki} - \bar{y}_{jk\cdot})^2 .\end{aligned}$$

Tästä nähdään suoraan, että PNS-estimaateiksi (ehdolla summausehdot (6.3)) saadaan

$$\hat{\mu} = \bar{y}, \quad \hat{\alpha}_j = \bar{y}_{j\cdot\cdot} - \bar{y}, \quad \hat{\gamma}_k = \bar{y}_{\cdot k\cdot} - \bar{y} \quad \text{ja} \quad \hat{\delta}_{jk} = y_{jk\cdot} - \bar{y}_{j\cdot\cdot} - \bar{y}_{\cdot k\cdot} + \bar{y}.$$

Vapaaksi residuaalineliösummaksi saadaan siten

$$\mathsf{S}(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (y_{jki} - \bar{y}_{jk\cdot})^2 = (n - JK)\, s^2,$$

jossa jälkimmäinen yhtälö perustuu estimaatin s^2 määritelmään. Huomaa, että tähän tulokseen voidaan päätyä myös minimoimalla alkuperäiseen parametrointiin (6.1) liittyvä jäännösneliösumma $\sum_{j,k,i} \left(y_{jki} - \mu_{jk}\right)^2$.

Hypoteesin H_{AB} testaamiseksi tarvittava rajoitettu residuaalineliösumma saadaan minimoimalla (6.4) ehdolla $\delta_{jk} = 0$ kaikilla j, k. Minimi saavutetaan, kun $\hat{\mu} = \bar{y}$, $\alpha_j = \hat{\alpha}_j$ ja $\gamma_k = \hat{\gamma}_k$, joten tulokseksi tulee

$$\mathsf{S}(\hat{\boldsymbol{\beta}}_{AB}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (y_{jk\cdot} - \bar{y}_{j\cdot\cdot} - \bar{y}_{\cdot k\cdot} + \bar{y})^2 + \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (y_{jki} - \bar{y}_{jk\cdot})^2 .$$

Rajoitetun ja vapaan residuaalineliösumman erotus on näin ollen

$$\mathsf{S}(\hat{\boldsymbol{\beta}}_{AB}) - \mathsf{S}(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (y_{jk\cdot} - \bar{y}_{j\cdot\cdot} - \bar{y}_{\cdot k\cdot} + \bar{y})^2 = m \sum_{k=1}^{K} \sum_{i=1}^{m} \hat{\delta}_{jk}^2,$$

joten F-testisuure hypoteesille H_{AB} on

$$F = m \sum_{j=1}^{J} \sum_{k=1}^{K} \hat{\delta}_{jk}^2 \,/ (J-1)\,(K-1)\, S^2 \overset{H_{AB}}{\sim} F_{(J-1)(K-1),n-JK}.$$

Hypoteesin H_A testaamisessa tarvittava sidottu residuaalineliösumma saadaan minimoimalla (6.4) ehdolla $\alpha_j = 0$ kaikilla j. Minimi saavutetaan valitsemalla $\hat{\mu} = \bar{y}$, $\gamma_k = \hat{\gamma}_k$ ja $\delta_{jk} = \hat{\delta}_{jk}$, joten sidottu residuaalineliösumma on

$$\mathsf{S}(\hat{\boldsymbol{\beta}}_A) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (\bar{y}_{j\cdot\cdot} - \bar{y})^2 + \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (y_{jki} - \bar{y}_{jk\cdot})^2 .$$

Näin ollen,

$$\mathsf{S}(\hat{\boldsymbol{\beta}}_A) - \mathsf{S}(\hat{\boldsymbol{\beta}}) = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{m} (\bar{y}_{j\cdot\cdot} - \bar{y})^2 = mK \sum_{i=1}^{m} \hat{\alpha}_j^2$$

ja F-testisuureeksi tulee

$$F = mK \sum_{j=1}^{J} \hat{\alpha}_j^2 / \,(J-1)\, S^2 \overset{H_A}{\sim} F_{(J-1),n-JK}.$$

Vastaavasti nähdään, että F-testisuure hypoteesille H_B on

$$F = mJ \sum_{i=1}^{K} \hat{\gamma}_k^2 / \,(K-1)\, S^2 \overset{H_A}{\sim} F_{(K-1),n-JK}.$$

Liitteet

A Satunnaisvektoreista, satunnaismatriiseista ja multinormaalijakaumasta

A.1 Satunnaisvektoreista ja satunnaismatriiseista

Olkoon Z_{ij}, $i = 1, .., n$ ja $j = 1, ..., m$ reaaliarvoisia satunnaismuuttujia. Tällöin $\mathbf{Z} = [Z_{ij}]$ on $n \times m$ satunnaismatriisi ja sen odotusarvo on

$$\mathsf{E}(\mathbf{Z}) = [\mathsf{E}(Z_{ij})] \quad (n \times m).$$

Jos $m = 1$, saadaan satunnaisvektori $\mathbf{Z} = [Z_1 \ \cdots \ Z_n]'$ ja sen odotusarvo.

Käyttäen reaaliarvoisten satunnaismuuttujien tunnettua lineaarisuusominaisuutta ($\mathsf{E}(aX+bY+c) = a\mathsf{E}(X)+b\mathsf{E}(Y)+c$) ja matriisien kertolasku- ja yhteenlaskusääntöjä voidaan todeta, että

$$\mathsf{E}(\mathbf{AZB} + \mathbf{C}) = \mathbf{A}\mathsf{E}(\mathbf{Z})\mathbf{B} + \mathbf{C} \tag{A.1}$$

ja

$$\mathsf{E}(\mathbf{AX} + \mathbf{BY}) = \mathbf{A}\mathsf{E}(\mathbf{X}) + \mathbf{B}\mathsf{E}(\mathbf{Y}), \tag{A.2}$$

kun $\mathbf{A}$, $\mathbf{B}$ ja $\mathbf{C}$ ovat dimensioiltaan sopivia ei-satunnaisia matriiseja.

Satunnaisvektorien $\mathbf{X} = [X_1 \ \cdots \ X_n]'$ ja $\mathbf{Y} = [Y_1 \ \cdots \ Y_m]'$ välinen kovarianssimatriisi on

$$\mathsf{Cov}(\mathbf{X}, \mathbf{Y}) = [\mathsf{Cov}(X_i, Y_j)] \quad (n \times m),$$

jossa oikealla $\mathsf{Cov}(X_i, Y_j) = \mathsf{E}((X_i - \mathsf{E}(X_i))(Y_j - \mathsf{E}(Y_j)))$ on satunnaismuuttujien X_i ja Y_j välinen kovarianssikerroin. Määritelmästä seuraa suoraan, että

$$\mathsf{Cov}(\mathbf{X}, \mathbf{Y}) = \mathsf{E}\left((\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{Y} - \mathsf{E}(\mathbf{Y}))'\right). \tag{A.3}$$

Jos $\mathbf{X} = \mathbf{Y}$, saadaan erikoistapauksena satunnaisvektorin $\mathbf{X}$ kovarianssimatriisi

$$\mathsf{Cov}(\mathbf{X}) = [\mathsf{Cov}(X_i, X_j)] = \mathsf{E}\left((\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{X} - \mathsf{E}(\mathbf{X}))'\right) \quad (n \times n). \tag{A.4}$$

Olkoon nyt $\mathbf{A}$ ja $\mathbf{B}$ kuten tuloksessa (A.2). Tällöin

$$\begin{aligned} \mathsf{Cov}(\mathbf{AX}, \mathbf{BY}) &= \mathsf{E}\left((\mathbf{AX} - \mathbf{A}\mathsf{E}(\mathbf{X}))(\mathbf{BY} - \mathbf{B}\mathsf{E}(\mathbf{Y}))'\right) \\ &= \mathsf{E}\left(\mathbf{A}(\mathbf{X} - \mathsf{E}(\mathbf{X}))(\mathbf{Y} - \mathsf{E}(\mathbf{Y}))'\mathbf{B}'\right), \end{aligned}$$

josta saadaan tuloksen (A.1) avulla

$$\mathsf{Cov}(\mathbf{AX}, \mathbf{BY}) = \mathbf{A}\mathsf{Cov}(\mathbf{X}, \mathbf{Y})\mathbf{B}' \tag{A.5}$$

ja erikoistapauksessa $\mathbf{X} = \mathbf{Y}$

$$\mathsf{Cov}(\mathbf{AX}) = \mathbf{A}\mathsf{Cov}(\mathbf{X})\mathbf{A}'. \tag{A.6}$$

Valitsemalla $\mathbf{A} = [a_1 \ \cdots \ a_n] = \mathbf{a}'$ ja huomaamalla, että $\mathsf{Cov}(\mathbf{a}'\mathbf{X}) = \mathsf{Var}(\mathbf{a}'\mathbf{X}) \geq 0$, nähdään tästä, että

$$\mathsf{Cov}(\mathbf{X}) \geq \mathbf{0} \tag{A.7}$$

eli satunnaisvektorin kovarianssimatriisi on positiivisesti semidefiniitti. Jos se ei ole positiivisesti definiitti (eli $\mathsf{Cov}(\mathbf{X}) > \mathbf{0}$ ei päde), on olemassa vektori $\mathbf{a} \neq \mathbf{0}$ siten, että $\mathsf{Var}(\mathbf{a}'\mathbf{X}) = 0$ eli $\mathbf{a}'\mathbf{X}$ saa vakioarvon todennäköisyydellä yksi.

A.2 Multinormaalijakaumasta[18]

A.2.1 Multinormaalijakauman määritelmä

Olkoon $U_1, ..., U_k$ riippumattomia satunnaismuuttujia ja $U_i \sim \mathsf{N}(0,1)$. Tällöin satunnaisvektorille $\mathbf{U} = [U_1 \ \cdots \ U_k]'$ pätee $\mathsf{E}(\mathbf{U}) = \mathbf{0}$ ja $\mathsf{Cov}(\mathbf{U}) = \mathbf{I}_k$ ja sen tiheysfunktio on

$$f_{\mathbf{U}}(\mathbf{u}) = \prod_{i=1}^{k} (2\pi)^{-1/2} \exp\left\{-\frac{1}{2}u_i^2\right\} = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{k} u_i^2\right\}$$

tai vektorimerkinnöin

$$f_{\mathbf{U}}(\mathbf{u}) = (2\pi)^{-k/2} \exp\left\{-\frac{1}{2}\mathbf{u}'\mathbf{u}\right\},$$

jossa $\mathbf{u} = [u_1 \ \cdots \ u_k]'$. Suoritetaan nyt lineaarinen muunnos

$$\begin{aligned} X_1 &= c_{11}U_1 + \cdots + c_{1k}U_k + \mu_1 \\ &\vdots \qquad\qquad\qquad \vdots \\ X_p &= c_{p1}U_1 + \cdots + c_{pk}U_k + \mu_k \end{aligned}$$

eli

$$\mathbf{X} = \mathbf{C}\mathbf{U} + \boldsymbol{\mu}, \tag{A.8}$$

jossa matriisin $\mathbf{C}$ $(p \times k)$ oletetaan olevan astetta p eli $\mathsf{r}(\mathbf{C}) = p$ ja siten erityisesti $p \leq k$. Käyttäen kohdan A.1 tuloksia voidaan todeta, että

$$\mathsf{E}(\mathbf{X}) = \mathbf{C}\mathsf{E}(\mathbf{U}) + \boldsymbol{\mu} = \boldsymbol{\mu}$$

ja

$$\mathsf{Cov}(\mathbf{X}) = \mathsf{E}(\,(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'\,) = \mathbf{C}\mathsf{E}\left(\mathbf{U}\mathbf{U}'\right)\mathbf{C}' = \mathbf{C}\mathbf{C}'.$$

Koska $\mathsf{r}(\mathbf{C}\mathbf{C}') = \mathsf{r}(\mathbf{C}) = p$, on $\mathsf{Cov}(\mathbf{X}) > \mathbf{0}$.

Satunnaisvektorin $\mathbf{X}$ sanotaan (tässä käytettävän määritelmän mukaan) noudattavan (epäsingulaarista) p-ulotteista normaalijakaumaa tai multinormaalijakaumaa odotusarvona $\boldsymbol{\mu}$ ja kovarianssimatriisina $\boldsymbol{\Sigma} = \mathbf{C}\mathbf{C}'$, jos $\mathbf{X}$:llä on yhtälön (A.8) mukainen esitys. Tällöin merkitään $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tai $\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Valitsemalla erityisesti $p = k$, $\mathbf{C} = \mathbf{I}_k$ ja $\boldsymbol{\mu} = \mathbf{0}$ nähdään, että $\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_k)$.

Johdetaan seuraavaksi satunnaisvektorin $\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ tiheysfunktio. Aputuloksena todetaan ensin, että

$$\mathbf{V} = [V_1 \ \cdots \ V_k]' = \mathbf{Q}\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_k), \tag{A.9}$$

kun $k \times k$ matriisi $\mathbf{Q}$ on ortogonaalinen eli $\mathbf{Q}'\mathbf{Q} = \mathbf{Q}\mathbf{Q}' = \mathbf{I}_k$. Koska $\mathbf{U} = \mathbf{Q}'\mathbf{V}$, seuraa satunnaisvektorien muunnosten teoriasta, että $\mathbf{V}$:n tiheysfunktio on $f_{\mathbf{V}}(\mathbf{v}) =$

[18] Jaksot A.2.1-A.2.5 perustuvat Seppo Mustosen kirjan Tilastolliset monimuuttujamenetelmät lukuun 2 (Helsingin yliopisto, Tilastotieteen laitos, 1995)

$f_{\mathbf{U}}(\mathbf{Q}'\mathbf{v})\,|\partial\mathbf{u}/\partial\mathbf{v}|$, jossa Jakobin determinantti $|\partial\mathbf{u}/\partial\mathbf{v}| = |\det(\mathbf{Q}')| = 1$, koska $\mathbf{Q}$ on ortogonaalinen. Näin ollen,

$$f_{\mathbf{V}}(\mathbf{v}) = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}\mathbf{v}'\mathbf{Q}\mathbf{Q}'\mathbf{v}\right\} = (2\pi)^{-k/2}\exp\left\{-\frac{1}{2}\mathbf{v}'\mathbf{v}\right\}.$$

Koska $\mathbf{V}$:n tiheysfunktion lauseke on sama kuin $\mathbf{U}$:n, seuraa tulos (A.9).

Osoitetaan seuraavaksi, että määrittely-yhtälössä (A.8) voidaan aina valita $p = k$ (yleistys $k \geq p$ on kuitenkin kätevä joissakin tarkasteluissa). Käytetään matriisin $\mathbf{C}$ singulaariarvohajotelmaa $\mathbf{C} = \mathbf{O}\mathbf{D}\mathbf{R}'$, jossa $\mathbf{D}$ $(p \times p)$ on diagonaalimatriisi, jonka diagonaalialkiot ovat positiivisia (koska $\mathsf{r}(\mathbf{C}) = p$), $\mathbf{O}$ $(p \times p)$ ortogonaalinen ja $\mathbf{R}$ $(k \times p)$ on sarakkeittain ortogonaalinen eli $\mathbf{R}'\mathbf{R} = \mathbf{I}_p$. Merkitsemällä $\mathbf{A} = \mathbf{O}\mathbf{D}$ ja $\mathbf{V} = \mathbf{R}'\mathbf{U}$ voidaan yhtälö (A.8) kirjoittaa

$$\mathbf{X} = \mathbf{A}\mathbf{V} + \boldsymbol{\mu}, \tag{A.10}$$

jossa $\mathbf{A}$ $(p \times p)$ on epäsingulaarinen eli $\mathsf{r}(\mathbf{A}) = p$. Koska matriisi $\mathbf{R}$ voidaan tunnetusti täydentää ortogonaaliseksi $(k \times k)$ matriisiksi, seuraa tuloksesta (A.9), että satunnaisvektorin $\mathbf{V}$ $(p \times 1)$ komponentit ovat riippumattomia ja $\mathsf{N}(0,1)$-jakautuneita. Yhtälöä (A.10) voidaan näin ollen käyttää vaihtoehtoisena multinormaalijakauman määrittely-yhtälönä. Kuten yhtälön (A.8) tapauksessa nähdään, että $\boldsymbol{\Sigma} = \mathsf{Cov}(\mathbf{X}) = \mathbf{A}\mathbf{A}' = \mathbf{C}\mathbf{C}'$. Todetaan lisäksi, että $\det(\boldsymbol{\Sigma}) = \det(\mathbf{A}\mathbf{A}') = \det(\mathbf{A})\det(\mathbf{A}') = \det(\mathbf{A})^2$, mistä seuraa $\det(\boldsymbol{\Sigma}^{-1}) = 1/\det(\boldsymbol{\Sigma}) = 1/\det(\mathbf{A})^2 = \det(\mathbf{A}^{-1})^2$.

Yhtälöstä (A.10) seuraa $\mathbf{V} = \mathbf{A}^{-1}(\mathbf{X} - \boldsymbol{\mu})$ ja kuten tuloksen (A.9) perustelussa saadaan

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{V}}\left(\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)|\partial\mathbf{v}/\partial\mathbf{x}| \\ &= (2\pi)^{-p/2}\det(\mathbf{A}^{-1})\exp\left\{-\frac{1}{2}\left(\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)'\left(\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)\right\} \\ &= (2\pi)^{-p/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'(\mathbf{A}^{-1})'\mathbf{A}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}, \end{aligned}$$

josta saadaan multinormaalijakauman tiheysfunktioksi

$$f_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-p/2}\det(\boldsymbol{\Sigma})^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

Jos satunnaisvektorilla $\mathbf{X}$ on yhtälön (A.8) tai (A.10) mukainen esitys (tai sillä on edellä johdettu tiheysfunktio), on se multinormaalisti jakautunut. Usein lähtökohta on kuitenkin käänteinen eli oletetaan $\mathbf{X} \sim \mathsf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Tällöin voidaan todeta, että satunnaisvektorille $\mathbf{X}$ pätee yhtälön (A.8) tai (A.10) mukainen esitys, jossa matriisi $\mathbf{C}$ tai $\mathbf{A}$ voidaan valita halutulla tavalla, kunhan ehto $\mathbf{C}'\mathbf{C} = \boldsymbol{\Sigma}$ tai $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$ toteutuu. Sopivalla matriisin $\mathbf{C}$ tai $\mathbf{A}$ valinnalla voidaan helpottaa suoritettavia tarkasteluja.

Todetaan vielä, että (eräs) edellä esitettyä yleisempi multinormaalijakauman määritelmä olettaa, että $\mathsf{r}(\mathbf{C}) = k$ eikä $\mathsf{r}(\mathbf{C}) = p$, joten tällöin $k \leq p$. Jos $k < p$, ei jakauma ole aidosti p-ulotteinen, sillä $\mathsf{Cov}(\mathbf{X}) = \mathbf{C}\mathbf{C}'$ on singulaarinen, jolloin on olemassa vektori $\mathbf{a}$ siten, että $\mathbf{a}'\mathbf{X}$ saa vakioarvon $(= \mathbf{a}'\boldsymbol{\mu})$ todennäköisyydellä yksi. Tällä kurssilla tätä yleisempää singulaarista multinormaalijakaumaa ei jouduta käyttämään. Seuraavassa esitettävät multinormaalijakauman ominaisuudet pätevät kuitenkin myös singulaariselle multinormaalijakaumalle.

A.2.2 Reunajakaumat

Oletetaan $\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ja ositetaan

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix},$$

jossa vektorien $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ dimensiot ovat on $q \times 1$ ja $(p-q) \times 1$. Ositetaan odotusarvo ja kovarianssimatriisi vastaavasti

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \text{ja} \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Yhtälön (A.10) perusteella voidaan kirjoittaa

$$\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \end{bmatrix} \mathbf{V} + \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix},$$

joten $\mathbf{X}^{(1)} = \mathbf{A}_1\mathbf{V} + \boldsymbol{\mu}^{(1)}$ noudattaa määritelmän (A.8) nojalla multinormaalijakaumaa. Laskemalla nähdään, että $\mathbf{A}_1\mathbf{A}_1' = \boldsymbol{\Sigma}_{11}$, joten

$$\mathbf{X}^{(1)} \sim \mathsf{N}_q\left(\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}_{11}\right).$$

Vastaavasti nähdään, että mikä tahansa $\mathbf{X}$:n osavektori eli $\mathbf{X}$:n reunajakaumat ovat multinormaalisia.

A.2.3 Muuttujaryhmien riippumattomuus

Tarkastellaan edelleen samaa tilannetta kuin edellisessä kohdassa, mutta oletetaan, että $\mathbf{X}^{(1)}$ ja $\mathbf{X}^{(2)}$ ovat korreloimattomia eli $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21}' = \mathbf{0}$. Yhtälöksi (A.10) voidaan tällöin valita

$$\begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{V}^{(1)} \\ \mathbf{V}^{(2)} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix},$$

sillä

$$\mathbf{A}\mathbf{A}' = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{11}' & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}' \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{11}\mathbf{A}_{11}' & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_{22}\mathbf{A}_{22}' \end{bmatrix}$$

ja valitsemalla $\mathbf{A}_{11}$ ja $\mathbf{A}_{22}$ siten, että $\mathbf{A}_{ii}\mathbf{A}_{ii}' = \boldsymbol{\Sigma}_{ii}$ $(i = 1, 2)$, pätee $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$. Koska siis $\mathbf{X}^{(i)} = \mathbf{A}_{ii}\mathbf{V}^{(i)} + \boldsymbol{\mu}^{(i)}$ $(i = 1, 2)$, ja $\mathbf{V}^{(1)} \perp\!\!\!\perp \mathbf{V}^{(2)}$, seuraa tästä $\mathbf{X}^{(1)} \perp\!\!\!\perp \mathbf{X}^{(2)}$. Toisaalta, jos $\mathbf{X}^{(1)} \perp\!\!\!\perp \mathbf{X}^{(2)}$, voidaan ne esittää satunnaisvektorien $\mathbf{V}^{(1)}$ ja $\mathbf{V}^{(2)}$ avulla erikseen aivan kuten edelläkin ja todeta laskemalla, että $\mathsf{Cov}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = \boldsymbol{\Sigma}_{12} = \mathbf{0}$. Yhteenvetona voidaan siis todeta, että

$$\mathbf{X}^{(1)} \perp\!\!\!\perp \mathbf{X}^{(2)} \Leftrightarrow \mathsf{Cov}\left(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}\right) = \mathbf{0}$$

eli multinormaalijakaumassa korreloimattomuus on yhtäpitävää riippumattomuuden kanssa. Huomaa, että muilla jakaumilla pätee implikaatio yleisesti vain suuntaan "$\Rightarrow$", mutta ei päinvastoin.

A.2.4 Käyttäytyminen lineaarimuunnoksissa

Oletetaan edelleen, että $\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ ja määritellään satunnaisvektori $\mathbf{Y} = \mathbf{B}\mathbf{X} + \boldsymbol{\nu}$, jossa $\mathbf{B}$ on ei-satunnainen $q \times p$ matriisi, $\mathsf{r}(\mathbf{B}) = q$, ja $\boldsymbol{\nu}$ on ei-satunnainen $q \times 1$ vektori. Yhtälöä (A.10) käyttäen saadaan

$$\mathbf{Y} = \mathbf{B}(\mathbf{A}\mathbf{V} + \boldsymbol{\mu}) + \boldsymbol{\nu} = (\mathbf{B}\mathbf{A})\mathbf{V} + (\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu}).$$

Koska $\mathsf{r}(\mathbf{B}\mathbf{A}) = \mathsf{r}(\mathbf{B}) = q$, seuraa yhtälön (A.8) määritelmästä, että $\mathbf{Y}$ on multinormaalinen ja tulosta (A.6) käyttäen nähdään, että

$$\mathsf{Cov}(\mathbf{Y}) = (\mathbf{B}\mathbf{A})(\mathbf{B}\mathbf{A})' = \mathbf{B}\mathbf{A}\mathbf{A}'\mathbf{B}' = \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}',$$

joten

$$\mathbf{Y} \sim \mathsf{N}_q\left(\mathbf{B}\boldsymbol{\mu} + \boldsymbol{\nu}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\right)$$

eli multinormaalisuus säilyy lineaarimuunnoksissa. Huomaa, että edellä ehtoa $\mathsf{r}(\mathbf{B}) = q$ tarvittiin vain, jotta $\mathbf{Y}$:n jakaumasta tulee epäsingulaarinen. Kuten aiemmin viitattiin, pätee tulos myös ilman tätä ehtoa.

A.2.5 Ehdolliset jakaumat

Käyttäen samaa ositusta kuin kohdassa A.2.2 ja edellä esitetyn kaltaisia argumentteja voidaan johtaa myös satunnaisvektorin $\mathbf{X}^{(1)}$ ehdollinen todennäköisyysjakauma ehdolla $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$. Edellisen kohdan nojalla satunnaisvektori

$$\begin{bmatrix} \mathbf{Z} \\ \mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_q & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \\ \mathbf{0} & \mathbf{I}_{p-q} \end{bmatrix} \begin{bmatrix} \mathbf{X}^{(1)} - \boldsymbol{\mu}^{(1)} \\ \mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

on multinormaalinen. Tästä yhtälöstä seuraa

$$\mathbf{X}^{(1)} = \boldsymbol{\mu}^{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{X}^{(2)} - \boldsymbol{\mu}^{(2)}\right) + \mathbf{Z}.$$

Käyttäen kohdan A.1 yleisiä tuloksia nähdään suoraan laskemalla, että $\mathsf{E}(\mathbf{Z}) = \mathbf{0}$, $\mathsf{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{11\cdot 2}$ (merkintä) ja $\mathsf{Cov}(\mathbf{Z}, \mathbf{X}^{(2)}) = \mathbf{0}$. Kohdan A.2.3 nojalla pätee siis $\mathbf{Z} \perp\!\!\!\perp \mathbf{X}^{(2)}$. Edellä esitetystä satunnaisvektoria $\mathbf{X}^{(1)}$ koskevasta yhtälöstä ja kohdasta A.2.4 seuraa siten, että $\mathbf{X}^{(1)}$:n ehdollinen jakauma ehdolla $\mathbf{X}^{(2)} = \mathbf{x}^{(2)}$ on multinormaalinen odotusarvona $\boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right)$ ja kovarianssimatriisina $\boldsymbol{\Sigma}_{11\cdot 2}$ eli

$$\mathbf{X}^{(1)} \;\Big|\; \left(\mathbf{X}^{(2)} = \mathbf{x}^{(2)}\right) \sim \mathsf{N}_q\left(\boldsymbol{\mu}^{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{x}^{(2)} - \boldsymbol{\mu}^{(2)}\right), \boldsymbol{\Sigma}_{11\cdot 2}\right).$$

Toisin sanoen, multinormaalijakaumassa ehdolliset jakaumat ovat multinormaalisia. Huomaa myös, että ehdollisen jakauman kovarianssimatriisi $\boldsymbol{\Sigma}_{11\cdot 2}$ eli ns. osittaiskovarianssimatriisi ei riipu ehtomuuttujan $\mathbf{X}^{(2)}$ arvosta. Tämän perusteella voidaan ajatella, että $\boldsymbol{\Sigma}_{11\cdot 2}$ kertoo kovarianssimatriisina satunnaisvektorin $\mathbf{X}^{(1)}$ komponenttien välisistä riippuvuuksista, kun satunnaisvektorin $\mathbf{X}^{(2)}$ (lineaariset) vaikutukset on eliminoitu.

A.2.6 Neliömuotojen jakaumista

Jos $\mathbf{V} = [V_1 \ \cdots \ V_p]' \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$, pätee satunnaismuuttujien $V_1, \cdots, V_p$ riippumattomuuden ja χ^2-jakauman määritelmän nojalla

$$\mathbf{V}'\mathbf{V} = \sum_{i=1}^{p} V_i^2 \sim \chi_p^2.$$

Lisäksi, jos $X \sim \mathsf{N}\left(\mu, \sigma^2\right)$, niin $(X-\mu)^2/\sigma^2 \sim \chi_1^2$. Seuraavassa tämän yleistys.

Lause A.1. Jos $\mathbf{X} \sim \mathsf{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, niin $(\mathbf{X}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu}) \sim \chi_p^2$.

Todistus: Käyttäen määritelmää (A.10), jossa $\mathbf{V}$ on kuten edellä, saadaan

$$(\mathbf{X}-\boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{X}-\boldsymbol{\mu}) = \mathbf{V}'\mathbf{A}' \left(\mathbf{A}\mathbf{A}'\right)^{-1} \mathbf{A}\mathbf{V} = \mathbf{V}'\mathbf{V} \sim \chi_p^2.$$

□

Seuraava tulos on hyödyllinen lineaarisen mallin teoriassa. Neliömatriisia $\mathbf{P}$ sanotaan (ortogonaaliseksi) projektioksi, jos se on (i) symmetrinen eli $\mathbf{P} = \mathbf{P}'$ ja (ii) idempotentti eli $\mathbf{P} = \mathbf{P}\mathbf{P}$ (merkitään $\mathbf{P}\mathbf{P} = \mathbf{P}^2$).

Lause A.2. Olkoon $\mathbf{X} \sim \mathsf{N}_p\left(\boldsymbol{\mu}, \sigma^2\mathbf{I}_p\right)$ ja $\mathbf{P}$ $(p \times p)$ astetta r oleva ortogonaalinen projektio. Tällöin

$$(\mathbf{X}-\boldsymbol{\mu})' \mathbf{P} (\mathbf{X}-\boldsymbol{\mu}) / \sigma^2 \sim \chi_r^2.$$

Todistus: Koska idempotentin matriisin ominaisarvot ovat tunnetusti nollia ja ykkösiä, on $\mathbf{P}$:llä pääakseliesitys $\mathbf{P} = \mathbf{R}\mathbf{D}\mathbf{R}'$, jossa $\mathbf{R}$ $(p \times p)$ on ortogonaalinen matriisi ja diagonaalimatriisi $\mathbf{D}$ $(p \times p)$ on muotoa $\mathbf{D} = \mathsf{Diag}\,[1 \cdots 1 \ 0 \cdots 0]$ (ykkösiä r kpl). Asetetaan $\mathbf{U} = \sigma^{-1}\mathbf{R}'(\mathbf{X}-\boldsymbol{\mu})$. Koska $\mathbf{U}$ saadaan satunnaisvektorista $\sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}) \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$ ortogonaalisella muunnoksella, pätee tuloksen (A.9) nojalla $\mathbf{U} \sim \mathsf{N}(\mathbf{0}, \mathbf{I}_p)$. Näin ollen,

$$(\mathbf{X}-\boldsymbol{\mu})' \mathbf{P} (\mathbf{X}-\boldsymbol{\mu}) / \sigma^2 = \mathbf{U}'\mathbf{D}\mathbf{U} = \sum_{i=1}^{r} U_i^2 \sim \chi_r^2.$$

□

B Matriisilaskentaa

B.1 Merkintöjä ja määritelmiä

Olkoon x_{ij}, $i = 1, ..., n$, $j = 1, .., m$, reaalilukuja. Lukukaaviota

$$\mathbf{X} = [x_{ij}] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}$$

sanotaan $n \times m$ *matriisiksi*, jonka i. rivin ja j. sarakkeen *alkio* tai *elementti* on x_{ij}. Matriiseja merkitään isoilla lihavoitetuilla kirjaimilla ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{X}$, $\mathbf{Y}$,...).

Vektori tulkitaan matriisiksi, jossa on vain yksi sarake. Vektoreita merkitään pienillä lihavoitetuilla kirjaimilla ($\mathbf{a}$, $\mathbf{b}$, $\mathbf{x}$, $\mathbf{y}$,...). Esimerkiksi

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = (y_1, ..., y_n)$$

on n-vektori tai $n \times 1$ vektori ja $\mathbf{x}_{i\cdot} = (x_{i1}, ..., x_{im})$ on matriisin $\mathbf{X}$ i. *rivivektori*. Vastaavasti $\mathbf{x}_{\cdot j} = (x_{1j}, ..., x_{nj})$ on matriisin $\mathbf{X}$ j. *sarakevektori*. Vektoria $\mathbf{1}_n = (1, ..., 1)$ $(n \times 1)$ sanotaan *ykkösvektoriksi* ja, jos dimensiota n ei ole täsmennetty, käytetään merkintää $\mathbf{1}$.

Vektorien $\mathbf{x}$ ja $\mathbf{y}$ $(n \times 1)$ *sisätulo* on

$$(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} x_i y_i.$$

Jos $(\mathbf{x}, \mathbf{y}) = 0$, niin $\mathbf{x}$ ja $\mathbf{y}$ ovat *ortogonaaliset* eli kohtisuorassa toisiaan vastaan. Tällöin merkitään $\mathbf{x} \perp \mathbf{y}$. Valitsemalla sisätulossa $\mathbf{x} = \mathbf{y}$ saadaan

$$(\mathbf{x}, \mathbf{x}) = \sum_{i=1}^{n} x_i^2 = \|\mathbf{x}\|^2,$$

jossa $\|\mathbf{x}\|$ on vektorin $\mathbf{x}$ pituus tai *normi*.

Kun $n = m$, puhutaan *neliömatriisista* ja kun lisäksi $x_{ij} = 0$, $i \neq j$, puhutaan *diagonaalimatriisista* eli

$$\mathbf{X} = \begin{bmatrix} x_{11} & 0 & \cdots & 0 \\ 0 & x_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & x_{nn} \end{bmatrix} = \mathsf{diag}\,[x_{11} \ \cdots \ x_{nn}] = \lceil x_{11} \ \cdots \ x_{nn} \rfloor.$$

Nollamatriisin kaikki alkiot ovat nollia ja *yksikkömatriisi* on diagonaalimatriisi, jonka kaikki diagonaalialkiot ovat ykkösiä eli $\mathbf{I}_n = \mathsf{diag}\,[1 \ \cdots \ 1]$ ja, jos dimensiota n ei ole täsmennetty, käytetään merkintää $\mathbf{I}$.

B.2 Yhteenlasku ja skalaarilla kertominen

Olkoon $\mathbf{A}$, $\mathbf{B}$ ja $\mathbf{C}$ samaa dimensiota olevia matriiseja ja λ ja μ skalaareja (eli λ, $\mu \in \mathbb{R}$). Matriisien $\mathbf{A}$ ja $\mathbf{B}$ summa määritellään vastinalkioiden summan avulla:

$$\mathbf{A} + \mathbf{B} = [a_{ij}] + [b_{ij}] = [a_{ij} + b_{ij}] \,.$$

Skalaarilla kertomisen määritelmä on

$$\lambda \mathbf{A} = \lambda \, [a_{ij}] = [\lambda a_{ij}] \,.$$

Matriisien yhteenlaskulle ja skalaarilla kertomiselle pätevät seuraavat säännöt.

- $\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$
- $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$, joten voidaan esimerkiksi kirjoittaa $(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + \mathbf{B} + \mathbf{C}$
- $(\lambda + \mu) \, \mathbf{A} = \lambda \mathbf{A} + \mu \mathbf{A}$
- $\lambda \, (\mu \mathbf{A}) = (\lambda \mu) \, \mathbf{A}$
- $\lambda \, (\mathbf{A} + \mathbf{B}) = \lambda \mathbf{A} + \lambda \mathbf{B}$

B.3 Matriisitulo

Matriisien $\mathbf{A}$ $(n \times m)$ ja $\mathbf{B}$ $(m \times l)$ tulo on määritelmän mukaan

$$\mathbf{AB} = \left[\sum_{k=1}^{m} a_{ik} b_{kj} \right] \quad (n \times l) \,.$$

Edellyttäen, että laskutoimitukset ovat määriteltyjä, ovat seuraavat laskusäännöt voimassa.

- $(\mathbf{AB}) \, \mathbf{C} = \mathbf{A} \, (\mathbf{BC})$, joten voidaan esimerkiksi kirjoittaa $(\mathbf{AB}) \, \mathbf{C} = \mathbf{ABC}$
- $\mathbf{A} \, (\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
- $(\mathbf{A} + \mathbf{B}) \, \mathbf{C} = \mathbf{AC} + \mathbf{BC}$

Siinäkin tapauksessa, että molemmat tulot olisivat määriteltyjä, ei $\mathbf{AB}$ ole yleensä sama kuin $\mathbf{BA}$.

B.4 Transponointi

Matriisin $\mathbf{A}$ $(n \times m)$ *transpoosi* $\mathbf{A}'$ $(m \times n)$ saadaan vaihtamalla rivit ja sarakkeet keskenään eli

$$\mathbf{A}' = [a_{ij}]' = [a_{ji}] \,.$$

Neliömatriisi $\mathbf{A}$ on *symmetrinen*, jos $\mathbf{A} = \mathbf{A}'$ ja, edellyttäen, että laskutoimitukset ovat määriteltyjä, ovat seuraavat laskusäännöt voimassa.

- $(\mathbf{A}')' = \mathbf{A}$
- $(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}'$
- $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ ja yleisemmin kahta useamman matriisin tapauksessa eli esimerkiksi $(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'$
- $(\lambda\mathbf{A})' = \lambda\mathbf{A}'$

Tapauksessa $m = 1$ on transponoitava matriisi vektori. Vektorien $\mathbf{x}$ ja $\mathbf{y}$ $(n \times 1)$ sisätulo voidaan siten kirjoittaa $(\mathbf{x}, \mathbf{y}) = \mathbf{x}'\mathbf{y}$, josta erikoistapauksena saadaan vektorin $\mathbf{x}$ normi $\|\mathbf{x}\| = (\mathbf{x}'\mathbf{x})^{1/2}$.

B.5 Käänteismatriisi

Jos $\mathbf{A}$ on $n \times n$ matriisi ja on olemassa sellainen $n \times n$ matriisi $\mathbf{B}$, että $\mathbf{AB} = \mathbf{I}_n$ ja $\mathbf{BA} = \mathbf{I}_n$, niin $\mathbf{B}$ on $\mathbf{A}$:n *käänteismatriisi* ja siitä käytetään merkintää $\mathbf{B} = \mathbf{A}^{-1}$. Tällöin $\mathbf{A}$:ta sanotaan *täysiasteiseksi* tai *epäsingulaariseksi* (tai *kääntyväksi* tai *säännölliseksi*) ja muulloin *vajaa-asteiseksi* tai *singulaariseksi.*

Epäsingulaarista neliömatriisia $\mathbf{A}$ sanotaan *ortoganaaliseksi*, jos $\mathbf{A}^{-1} = \mathbf{A}'$. Tällöin $\mathbf{A}'\mathbf{A} = \mathbf{I}_n$ ja $\mathbf{AA}' = \mathbf{I}_n$, joten matriisin $\mathbf{A}$ sarakkeet ovat keskenään ortogonaaliset ja samoin rivit.

Edellyttäen, että sekä $\mathbf{A}$ että $\mathbf{B}$ ovat epäsingulaarisia pätevät seuraavat laskusäännöt.

- $\left(\mathbf{A}^{-1}\right)^{-1} = \mathbf{A}$
- $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
- $(\mathbf{A}')^{-1} = \left(\mathbf{A}^{-1}\right)'$

B.6 Ominaisarvot ja -vektorit

Neliömatriisin $\mathbf{A}$ $(n \times n)$ *ominaisarvot* $\lambda_1, ..., \lambda_n$ ovat (polynomi)yhtälön

$$\det(\mathbf{A} - \lambda\mathbf{I}_n) = 0$$

ratkaisut. Vastaavat *ominaisvektorit* $\mathbf{u}_1, ..., \mathbf{u}_n$ $(\mathbf{u}_i \neq 0)$ saadaan yhtälöistä

$$(\mathbf{A} - \lambda_i\mathbf{I}_n)\,\mathbf{u}_i = 0 \qquad \text{eli} \qquad \mathbf{A}\mathbf{u}_i = \lambda_i\mathbf{u}_i.$$

Jos $\mathbf{A}$ on symmetrinen, niin sen ominaisarvot ovat reaalisia ja $\mathbf{A}$:lla on *pääakseliesitys*

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}',$$

jossa $\mathbf{\Lambda} = \mathsf{diag}\,[\lambda_1 \ \cdots \ \lambda_n]$ ja $\mathbf{U} = [\mathbf{u}_1 \ \cdots \ \mathbf{u}_n]$ on ortogonaalinen.

Jos $\mathbf{A}$ on $n \times m$ matriisi ($n \geq m$ ja $n > m$ mahdollinen), niin sillä on *singulaariarvohajotelma*

$$\mathbf{A} = \mathbf{RDO}',$$

jossa $\mathbf{D} = \mathsf{diag}\left[d_1 \ \cdots \ d_m\right] \geq 0$, $\mathbf{R}$ $(n \times m)$ on matriisi, jonka sarakkeet ovat ortonormaalit (eli $\mathbf{R}'\mathbf{R} = \mathbf{I}_m$), ja $\mathbf{O}$ $(m \times m)$ on ortogonaalinen (eli $\mathbf{O}'\mathbf{O} = \mathbf{I}_m = \mathbf{O}\mathbf{O}'$). Täsmällisemmin, $d_1 \geq \cdots \geq d_m \geq 0$ ovat matriisin $\mathbf{A}'\mathbf{A}$ (ei-negatiivisten) ominaisarvojen neliöjuuret, $\mathbf{R}$:n sarakkeet ovat matriisin $\mathbf{A}\mathbf{A}'$ m:ää suurinta ominaisarvoa vastaavat ominaisvektorit ja $\mathbf{O}$:n sarakkeet ovat matriisin $\mathbf{A}'\mathbf{A}$ ominaisvektorit.

B.7 Matriisin jälki

Neliömatriisin $\mathbf{A}$ $(n \times n)$ *jälki* on sen diagonaalialkioiden summa eli

$$\mathsf{tr}\left(\mathbf{A}\right) = \sum_{i=1}^{n} a_{ii}.$$

Edellyttäen, että laskutoimitukset ovat määriteltyjä, ovat seuraavat laskusäännöt voimassa.

- $\mathsf{tr}\left(\mathbf{A} + \mathbf{B}\right) = \mathsf{tr}\left(\mathbf{A}\right) + \mathsf{tr}\left(\mathbf{B}\right)$
- $\mathsf{tr}\left(\mathbf{A}\mathbf{B}\right) = \mathsf{tr}\left(\mathbf{B}\mathbf{A}\right)$
- $\mathsf{tr}\left(\mathbf{A}\right) = \sum_{i=1}^{n} \lambda_i$, jossa $\lambda_1, ..., \lambda_n$ ovat $\mathbf{A}$:n ominaisarvot

B.8 Matriisin determinantti

Neliömatriisin $\mathbf{A}$ $(n \times n)$ *determinantti* määritellään yhtälöllä

$$\det\left(\mathbf{A}\right) = \sum_{\sigma \in S_n} \epsilon\left(\sigma\right) a_{\sigma(1)1} \cdots a_{\sigma(n)n},$$

jossa summaus on yli joukon $\{1, ..., n\}$ kaikkien permutaatioiden $\sigma \in S_n$ ja $\epsilon\left(\sigma\right) \in \{-1, 1\}$ on σ:n etumerkki. Vaihtoehtoinen merkintä determinantille on $\left|\mathbf{A}\right|$.

Edellyttäen, että laskutoimitukset ovat määriteltyjä, ovat seuraavat laskusäännöt voimassa.

- $\det\left(\mathbf{A}\mathbf{B}\right) = \det\left(\mathbf{A}\right)\det\left(\mathbf{B}\right)$
- $\det\left(\mathbf{A}'\right) = \det\left(\mathbf{A}\right)$
- $\det\left(\mathbf{A}^{-1}\right) = 1/\det\left(\mathbf{A}\right) \quad \left(\det\left(\mathbf{A}\right) \neq 0\right)$
- $\det\left(\mathbf{A}\right) = 1$ tai $= -1$, kun $\mathbf{A}$ on ortogonaalinen (seuraa kahdesta edellisestä)
- $\det\left(\mathsf{diag}\left[a_{11} \ \cdots \ a_{nn}\right]\right) = a_{11} \cdots a_{nn}$
- $\det\left(\mathbf{A}\right) = \lambda_1 \cdots \lambda_n$, jossa $\lambda_1, ..., \lambda_n$ ovat $\mathbf{A}$:n ominaisarvot

B.9 Vektorien lineaarinen riippuvuus ja riippumattomuus sekä matriisin aste

Vektori $\mathbf{x}_1, ..., \mathbf{x}_p$ ovat *lineaarisesti riippuvia* tai *sidottuja* jos $c_1\mathbf{x}_1 + \cdots + c_p\mathbf{x}_p = 0$ joillakin skalaareilla $c_1, ..., c_p$, joista ainakin yksi on nollasta poikkeava. Ne ovat *lineaarisesti riippumattomia* tai *vapaita*, jos ehdosta $c_1\mathbf{x}_1 + \cdots + c_p\mathbf{x}_p = 0$ seuraa $c_1 = \cdots = c_p = 0$.

Olkoon $\mathbf{x}_1, ..., \mathbf{x}_p$ $n \times 1$ vektoreita ja $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_p]$ $n \times p$ matriisi, jonka i. sarake on $\mathbf{x}_i$ $(i = 1, ..., p)$. Tällöin $\mathbf{X}$:n *sarakeavaruus* $\mathcal{R}(\mathbf{X})$ muodostuu $n \times 1$ vektoreista $\mathbf{y}$, joille pätee

$$\mathbf{y} = \sum_{i=1}^{p} \beta_i \mathbf{x}_i = \mathbf{X}\boldsymbol{\beta}$$

jollain $p \times 1$ vektorilla $\boldsymbol{\beta} = \begin{bmatrix}\beta_1 & \cdots & \beta_p\end{bmatrix}'$. Toisin sanoen,

$$\mathcal{R}(\mathbf{X}) = \{\mathbf{y} \in \mathbb{R}^n \colon \mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \quad \boldsymbol{\beta} \in \mathbb{R}^p\}.$$

$\mathcal{R}(\mathbf{X})$ on $\mathbb{R}^n$:n aliavaruus, joten $\mathbf{0} \in \mathcal{R}(\mathbf{X})$ ja, jos $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{R}(\mathbf{X})$, niin $a_1\mathbf{y}_1 + a_2\mathbf{y}_2 \in \mathcal{R}(\mathbf{X})$ kaikilla $a_1, a_2 \in \mathbb{R}$. Koska $\mathcal{R}(\mathbf{X}') = \mathcal{R}(\mathbf{X}'\mathbf{X})$, niin ehdosta $\mathbf{a} = \mathbf{X}'\mathbf{y}$ (eli $\mathbf{a} \in \mathcal{R}(\mathbf{X}')$) seuraa $\mathbf{a} = \mathbf{X}'\mathbf{X}\mathbf{b}$ jollain $\mathbf{b} \in \mathbb{R}^p$. Jos $\mathsf{r}(\mathbf{X}'\mathbf{X}) = p$, on $(\mathbf{X}'\mathbf{X})^{-1}$ olemassa ja $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{a}$.

Matriisin $\mathbf{A}$ *aste* $\mathsf{r}(\mathbf{A})$ on $\mathbf{A}$:n lineaarisesti riippumattomien sarakkeiden lukumäärä tai yhtäpitävästi $\mathbf{A}$:n lineaarisesti riippumattomien rivien lukumäärä. Määritelmän mukaan, $\mathsf{r}(\mathbf{A}') = \mathsf{r}(\mathbf{A})$. Jos $\mathbf{A}$ on $n \times n$ matriisi, niin pätee

$$\mathsf{r}(\mathbf{A}) = n \Leftrightarrow \mathbf{A}^{-1} \text{ on olemassa} \Leftrightarrow \det(\mathbf{A}) \neq 0.$$

Edelleen, matriisin aste ei muutu, jos matriisi kerrotaan (oikealta tai vasemmalta) epäsingulaarisella (neliö)matriisilla. Jos siis $\mathbf{X}$ on kuten edellä, $\mathbf{A}$ on $n \times n$ ja $\mathbf{B}$ on $p \times p$ matriisi, niin $\mathsf{r}(\mathbf{AX}) = \mathsf{r}(\mathbf{X}) = \mathsf{r}(\mathbf{XB})$, kun $\mathsf{r}(\mathbf{A}) = n$ ja $\mathsf{r}(\mathbf{B}) = p$.

Jos $\mathbf{X}$ on $n \times p$ matriisi ja $\mathsf{r}(\mathbf{X}) = p$, niin $p \leq n$ ja jokaisella vektorilla $\mathbf{y} \in \mathcal{R}(\mathbf{X})$ on yksikäsitteinen esitys $\mathbf{y} = \mathbf{X}\mathbf{a}$, jossa $\mathbf{a} \in \mathbb{R}^p$. Ositetaan $\mathbf{X} = [\mathbf{X}_1 \ \ \mathbf{X}_2]$, jossa $\mathbf{X}_1$ on $n \times m$ ja $\mathbf{X}_2$ on $n \times (p - m)$ matriisi. Oletetaan, että $\mathsf{r}(\mathbf{X}) = \mathsf{r}(\mathbf{X}_1) = m$. Tällöin matriisin $\mathbf{X}_2$ sarakkeet ovat lineaarisesti riippuvia matriisin $\mathbf{X}_1$ sarakkeista eli (matriisitulon määritelmän nojalla) $\mathbf{X}_2 = \mathbf{X}_1\mathbf{A}$ jollain $m \times (p - m)$ matriisilla $\mathbf{A}$. Lisäksi $\mathcal{R}(\mathbf{X}) = \mathcal{R}(\mathbf{X}_1)$, joten jos $\mathbf{y} \in \mathcal{R}(\mathbf{X})$, niin $\mathbf{y} = \mathbf{X}_1\mathbf{c}$ jollain $\mathbf{c} \in \mathbb{R}^m$. Toisin sanoen, $\mathcal{R}(\mathbf{X})$:n vektorit voidaan esittää $\mathbf{X}_1$:n avulla, joten tässä mielessä matriisisa $\mathbf{X}$ on 'turhia' sarakkeita.

B.10 Idempotentit matriisit ja projektiot

Neliömatriisia $\mathbf{A}$ sanotaan *idempotentiksi*, jos $\mathbf{A} = \mathbf{AA}$. Tällöin merkitään $\mathbf{AA} = \mathbf{A}^2$. Jos $\mathbf{A}$ on idempotentti, niin $\mathbf{I} - \mathbf{A}$ on myös idempotentti ja $\mathbf{A}$:n (samoin kuin $(\mathbf{I} - \mathbf{A})$:n) ominaisarvot ovat nollia ja ykkösiä.

Olkoon nyt $\mathbf{X}$ $n \times p$ matriisi. Sarakeavaruuden $\mathcal{R}(\mathbf{X})$ *ortogonaalinen komplementti* $\mathcal{R}(\mathbf{X})^{\perp}$ muodostuu niistä $n \times 1$ vektoreista, jotka ovat kohtisuorassa $\mathcal{R}(\mathbf{X})$:n vektoreita vastaan eli $\mathcal{R}(\mathbf{X})^{\perp} = \{\mathbf{y} \in \mathbb{R}^n \colon (\mathbf{y}, \mathbf{x}) = \mathbf{0}, \ \mathbf{x} \in \mathcal{R}(\mathbf{X})\}$. $\mathcal{R}(\mathbf{X})^{\perp}$ on $\mathbb{R}^n$:n

aliavaruus ja jokainen $\mathbf{y} \in \mathbb{R}^n$ voidaan esittää yksikäsitteisenä summana $\mathbf{y} = \mathbf{u} + \mathbf{v}$, jossa $\mathbf{u} \in \mathcal{R}(\mathbf{X})$ ja $\mathbf{v} \in \mathcal{R}(\mathbf{X})^{\perp}$. Kuvausta $\mathbf{P_X} : \mathbf{y} \mapsto \mathbf{u}$ sanotaan *ortogonaaliseksi (tai kohtisuoraksi) projektioksi* tai lyhyesti *projektioksi* $\mathcal{R}(\mathbf{X})$:lle (vastaavasti $\mathcal{R}(\mathbf{X})^{\perp}$:n tapauksessa). $\mathbf{P_X}$ on lineaarikuvaus ja se voidaan tulkita matriisiksi, josta käytetään myös merkitää $\mathbf{P_X}$. Koska jokaiselle $\mathbf{y} \in \mathbb{R}^n$, pätee $\mathbf{P_X^2 y} = \mathbf{P_X}(\mathbf{P_X y}) = \mathbf{P_X u} = \mathbf{u} = \mathbf{P_X y}$, on $\mathbf{P_X}$ idempotentti. Lisäksi, $\mathcal{R}(\mathbf{P_X}) = \mathcal{R}(\mathbf{X})$.

Mille tahansa $\mathbf{y} \in \mathbb{R}^n$ pätee

$$\min_{\mathbf{z} \in \mathcal{R}(\mathbf{X})} \|\mathbf{y} - \mathbf{z}\| = \|\mathbf{y} - \mathbf{P_X y}\| .$$

Toisin sanoen, $\mathbf{P_X y}$ eli $\mathbf{y}$:n (kohtisuora) projektio $\mathbf{X}$:n sarakeavaruudelle $\mathcal{R}(\mathbf{X})$ on se $\mathcal{R}(\mathbf{X})$:n yksikäsitteinen vektori, joka on lähinnä $\mathbf{y}$:tä. Vaihtoehtoisesti (kohtisuora) projektio $\mathbf{P_X y}$ voidaan määritellä (yksikäsitteisenä) vektorina, jolle pätee

$$((\mathbf{y} - \mathbf{P_X y}), \mathbf{z}) = (\mathbf{y} - \mathbf{P_X y})' \mathbf{z} = \mathbf{0} \quad \text{kaikilla } \mathbf{z} \in \mathcal{R}(\mathbf{X}) .$$

Jos edellä $\mathsf{r}(\mathbf{X}) = p$, niin $\mathbf{P_X} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$. Tällöin $\mathbf{P_X}$ on sekä idempotentti että symmetrinen. Jos matriisi $\mathbf{X}$ on asiayhteydestä selvä, merkitään $\mathbf{P_X} = \mathbf{P}$.

Yleisesti sanotaan symmetristä ja idempotenttia matriisia $\mathbf{A}$ *ortogonaaliseksi projektioksi* tai lyhyesti *projektioksi*. Koska projektion ominaisarvot ovat nollia ja ykkösiä pätee $\mathbf{A} \geq \mathbf{0}$ ja $\mathsf{r}(\mathbf{A}) = \mathsf{tr}(\mathbf{A})$. Lisäksi, projektion pääakseliesitys on muotoa $\mathbf{A} = \mathbf{RR'}$, jossa matriisin $\mathbf{R}$ sarakkeet muodostuvat ykkösominaisarvoja vastaavista (keskenään ortogonaalisista) ominaisvektoreista.

B.11 Neliömuodot

Olkoon $\mathbf{A}$ $(n \times n)$ symmetrinen matriisi ja $\mathbf{x}$ $(n \times 1)$ vektori. Tällöin

$$(\mathbf{x}, \mathbf{Ax}) = \mathbf{x'Ax} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

on ($\mathbf{A}$:n) *neliömuoto*. Neliömuodot jaotellaan seuraavasti.

- $\mathbf{x'Ax} > 0$ kaikilla $\mathbf{x} \neq \mathbf{0} \Leftrightarrow \mathbf{A}$ on positiivisesti definiitti; $\mathbf{A} > \mathbf{0}$
- $\mathbf{x'Ax} \geq 0$ kaikilla $\mathbf{x} \Leftrightarrow$ $\mathbf{A}$ on positiivisesti semidefiniitti; $\mathbf{A} \geq \mathbf{0}$

$\mathbf{A}$:n ominaisarvot ovat positiivisia (ei-negatiivisia) jos ja vain jos $\mathbf{A} > \mathbf{0}$ ($\mathbf{A} \geq \mathbf{0}$). Lisäksi $\mathbf{X'X} \geq \mathbf{0}$ ja $\mathbf{X'X} > \mathbf{0}$, jos $\mathbf{X}$:n sarakkeet ovat lineaarisesti riippumattomia ($\mathbf{X}$:n ei tarvitse olla neliömatriisi). Yleisemmin, jos $\mathbf{A} > \mathbf{0}$ ja jos $\mathbf{X}$:n sarakkeet ovat lineaarisesti riippumattomia, niin $\mathbf{X'AX} > \mathbf{0}$ (olettaen, että tulot ovat määriteltyjä).