Säännöllistämismenetelmät regressioanalyysissä

Jaana Jokiniemi

Tilastotieteen pro gradu -tutkielma

29. huhtikuuta 1997

Jyväskylän yliopisto

tilastotieteen laitos

Tiivistelmä

Jaana Jokiniemi: *Säännöllistämismenetelmät regressioanalyysissä.*
Tilastotieteen pro gradu -tutkielma, Jyväskylän yliopisto, 29. huhtikuuta 1997.
Sivuja 88, liitteitä 7.

Säännöllistämismenetelmiä käytetään regressioanalyysissä, kun selittävien muuttujien $x_1,\ldots,x_p$ välillä on lähes lineaarisia tai lineaarisia riippuvuuksia ts. selittävät muuttujat $x_1,\ldots,x_p$ ovat multikollineaarisia, jolloin rakennematriisin $X = (x_1,\ldots,x_p)$ neliömatriisi $X'X$ ei ole oikein kääntyvä. Vaikka regressiomatriisin B pienimmän neliösumman estimaatilla on minimivarianssi harhattomien estimaattoreiden joukossa, niin selittävien muuttujien ollessa multikollineaarisia sen varianssi ja täten myös sen keskineliövirhe on saattanut kasvaa kohtuuttoman suureksi, jolloin sen pienimmän neliösumman estimaattorin $B_{PNS} = (X'X)^{-1}X'Y$ harhattomuudella ei ole merkitystä. Tällöin säännöllistämismenetelmillä voidaan approksimoida käänteismatriisia $(X'X)^{-1}$ matriisilla G, joka ei ole yhtä herkkä selittävien muuttujien multikollineaarisuuden vaikutukselle kuin matriisi $(X'X)^{-1}$. Matriisi G pyritään valitsemaan siten, että säännöllistämismenetelmillä saadaan harhainen regressiomatriisin B estimaattori $B_S = GX'Y$, joka on keskimäärin tarkempi kuin estimaattori B_{PNS}, tai matriisin G käänteismatriisi sisältää suurimman osan rakennematriisin X vaihtelusta, jolla voidaan parhaiten ennustaa havaintomatriisia Y.

Tässä tutkielmassa esitetyt säännöllistämismenetelmät regressioanalyysissä ovat harjaregressio, pääkomponenttiregressio ja osittainen pienimmän neliösumman regressio eli PLS-regressio. Kustakin menetelmästä on kuvaus, tarvittavien parametrien estimointi, selittävien muuttujien tai komponenttien valinta malliin ja yhteys toisiin menetelmiin. Menetelmiä sovelletaan teräksen karkenevuuden mallittamiseen.

Avainsanat: lineaarinen regressio; säännöllistämismenetelmät; harjaregressio; pääkomponentti-regressio; osittainen pienimmän neliösumman regressio

Sisällysluettelo

1 Johdanto

Teräksen karkenevuus on eräs sen laatuominaisuus, joka voidaan määritellä sen kykynä muuttua kiderakenteeltaan kokonaan tai osittain austeniitista martensiitiksi, kun se karkaisuhehkutuksen jälkeen sammutetaan tietyllä tavalla. Näin määriteltyyn teräksen karkenevuuteen vaikuttavat teräksen seostus eli seosaineiden (C, Si,...) pitoisuudet, kappaleen koko ja muoto sekä hehkutuslämpötila ja jäähdytysnopeus. Teräksen karkenevuutta voidaan mitata Jominy-kokeella, jossa standardin SFS 2375 (tai iso 642-1979) mukaan terässauva kuumennetaan austenitointilämpötilaan, siirretään se nopeasti erityiseen jäähdytyslaitteeseen ja sammutetaan suuntaamalla vesisuihku sen alapäähän. Näin terässauva jäähtyy sitä hitaammin mitä ylemmäs sen alapäästä noustaan ja erityisesti koesauvan alapää jäähtyy äärimmäisen nopeasti muuttuen kiderakeenteeltaan täysin marteensiittiseksi. Terässauvan jäähdyttyä täysin hiotaan sen sivuille mittaustasot, joista sen kovuus mitataan joko Rockwellin C-yksikkönä tai Vickers-yksikkönä. Kovuusmittaukset tehdään esimerkiksi 1.5, 3, 5, 7, 9, 11, 13, 15, 20, 25,...,50 mm etäisyyksiltä terässauvan alapäästä. Mittaukset suoritetaan erikseen kummastakin tasosta ja mikäli tulokset poikkeavat toisistaan alle 2 HRC (tai alle 20 HV), vastaavat lukemat keskiarvostetaan. Saadut etäisyys-kovuus -parit muodostavat Jominy-käyrän.

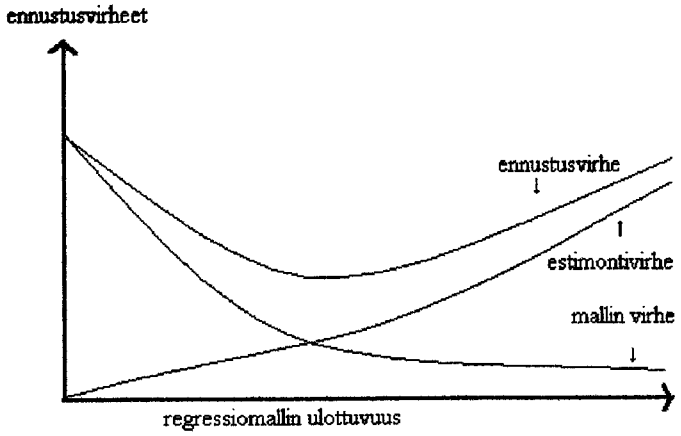
Tehtävänä on muodostaa lineaarinen regressiomalli teräksen karkenevuudelle. Koska Jominy-koetta varten on vakioitu kaikki muut teräksen karkaisutulokseen vaikuttavat tekijät paitsi seosaineiden pitoisuudet, niin muodostettavassa teräksen karkenevuuden regressiomallissa selitettävinä muuttujina ovat eri etäisyyksiltä terässauvan alapäästä mitatut teräksen kovuudet ja selittävinä muuttujina ovat teräksen seosaineiden pitoisuudet. (Huhtala, 1991.)

Kun X on selittävien muuttujien rakennematriisi, Y on selitettävien muuttujien havaintomatriisi ja muodostettava teräksen karkenevuuden regressiomalli ratkaistaan pienimmän neliösumman menetelmällä, niin tällöin joudutaan kääntämään neliömatriisi $X'X$. Koska teräksen seosaineiden pitoisuudet ovat multikollineaarisia, niin neliömatriisi $X'X$ ei ole oikein kääntyvä. Teräksen seosaineiden pitoisuuksien multikollineaarisuuden ongelmaa ei voida ratkaista poistamalla multikollineaarisia teräksen seosaineiden pitoisuuksia, sillä muodostettavan regressiomallin käytön kannalta kaikkien teräksen seosaineiden mukanaolo on tärkeätä. Tässä tutkielmassa teräksen karkenevuuden regressiomalli estimoidaan harjaregressiolla, pääkomponenttiregressiolla ja osittaisella pienimmän neliösumman regressiolla eli PLS-regressiolla. Nämä menetelmät ovat säännöllistämismenetelmiä, joilla saadaan käänteismatriisille $(X'X)^{-1}$ approksimaatio G, joka ei ole

yhtä herkkä selittävien muuttujien multikollineaarisuuden vaikutukselle kuin matriisi $(X'X)^{-1}$.

Vaikka regressiomatriisin B pienimmän neliösumman estimaatilla on minimivarianssi harhattomien estimaattoreiden joukossa, niin selittävien muuttujien ollessa multikollineaarisia sen varianssi on saattanut kasvaa kohtuuttoman suureksi, koska pienimmän neliösumman estimaattorissa $B_{PNS} = (X'X)^{-1}X'Y$ joudutaan kääntämään singulaarinen tai lähes singulaarinen neliömatriisi $X'X$. Jos selittävät muuttujat ovat multikollineaarisia, niin säännöllistämismenetelmillä voidaan saada regressiomatriisin B estimaattori $B_s = GX'Y$, joka on hiukan harhainen, mutta jonka varianssi on huomattavasti pienempi kuin harhattoman estimaattorin B_{PNS} varianssi. Estimaattorin hyvyyttä voidaan mitata estimaattorin keskineliövirheellä, joka on estimaattorin harhan neliön ja varianssin summa. Säännöllistämismenetelmillä pyritään valitsemaan matriisi G siten, että estimaattorin B_s keskineliövirhe on pienempi kuin estimaattorin B_{PNS} keskineliövirhe.

Säännöllistämismenetelmiä on käytetty usein ennustamisessa, koska säännöllistämismenetelmillä voidaan vähentää regressiomallin ulottuvuutta. Jotta saadaan ennustevirheeltään pienin mahdollinen ennuste, niin regressiomallin ulottuvuuden tulee olla sopiva (ks. kuva 1).



Kuva 1: Ennustevirhe regressiomallin ulottuvuuden funktiona (Naes & Irgens, 1986)

Kemiassa ja monissa teollisissa sovelluksissa on havaittu todellisten selittävien piilomuuttujien ja havaittujen muuttujien välinen ero, sillä kemiallisen prosessin ulottuvuus saattaa muuttua, jos sitä tutkitaan uudesta näkökulmasta. Sosiologiassa ja psykologiassa vastaava ongelma on jo kauan

tiedostettu, minkä seurauksena faktorimalleista on tullut yksi tilastotieteen keskeinen osa-alue. Selittävien muuttujien valinnan ongelma on usein helpommin ratkaistavissa kemiallisissa sovelluksissa kuin psykologissa ja sosiologisissa sovelluksissa, koska kemiallisissa sovelluksissa kuten Jominy-kokeessa selittävät muuttujat ovat kokeellisesti tulkittavissa ja kiistattomasti mitattavissa.

Tässä tutkielmassa muodostettavan teräksen karkenevuuden regressiomallin ulottuvuutta approksimoidaan säännöllistämismenetelmillä. PLS-regressiossa ja pääkomponenttiregressiossa regressoidaan havaintomatriisi Y rakennematriisia X vasten, mutta lisäksi niissä mallitetaan rakennematriisia X ja PLS-regressiossa myös havaintomatriisia Y. Pääkomponenttiregressiossa approksimoitu regressiomallin ulottuvuus on rakennematriisin X valittujen pääkomponenttien lukumäärä, joka on pienempi tai yhtä pieni kuin selittävien muuttujien lukumäärä. PLS-regressiossa muodostetaan PLS-komponentit siten, että ne selittävät mahdollisimman vähällä komponenttien lukumäärällä mahdollisimman paljon rakennematriisin X kokonaisvaihtelusta, jolla voidaan parhaiten ennustaa havaintomatriisia Y. Harjaregressiossa, kuten myös pienimmän neliösumman regressiossa approksimoitu regressiomallin ulottuvuus on regressiomallissa olevien selittävien muuttujien lukumäärä ts. harjaregressiossa ja pienimmän neliösumman regressiossa ei regresiomallin ulottuvuutta vähennetä.

1.2 Tutkielman sisältö

Tässä tutkielmassa perehdytään harjaregressioon, pääkomponenttiregressioon ja osittaisen pienimmän neliösumman regressioon. Näitä menetelmiä käytetään siis silloin, kun selittävät muuttujat ovat multikollineaarisia.

Tämän tutkielman luvussa 2 on regressioanalyysin perusteita. Kappaleessa 2.1 esitetään kanoninen regressiomalli, joka on yleinen regressiomallin muoto, kun säännöllistämismenetelmiä käytetään. Kappaleessa 2.2 esitetään pienimmän neliösumman menetelmä, jota verrataan säännöllistämismenetelmiin. Koska selittävien muuttujien multikollineaarisuutta ei voida nähdä suoraan selittävien muuttujien korrelaatiomatriisista, niin kappaleessa 2.4 on menetelmiä selittävien muuttujien multikollineaarisuuden havaitsemiseen. Vaikka regressiomatriisin $\mathbf{B}$ pienimmän neliösumman estimaatilla on minimivarianssi harhattomien estimaattorien joukossa, niin kappaleessa 2.5 esitetään, miksi pienimmän neliösumman menetelmän hyvyys voidaan kyseenalaistaa, kun selittävät muuttujat ovat multikollineaarisia.

Luvussa 3 on esitetty harjaregressio, luvussa 4 pääkomponenttiregressio ja luvussa 5

osittainen pienimmän neliösumman regressio. Kustakin menetelmästä on kuvaus, tarvittavien parametrien estimointi ja selittävien muuttujien tai komponenttien valinta malliin. Lisäksi näissä luvuissa osoitetaan menetelmien yhteys toisiin säännöllistämismenetelmiin.

Säännöllistämismenetelmiä sovelletaan teräksen karkenevuusaineistoon luvussa 6. Muodostettavassa regressiomallissa on selitettävinä muuttujina eri etäisyyksiltä terässauvan alapäästä mitatut teräksen kovuudet, joten muodostettava regressiomalli on monimuuttujainen. Koska ennalta jo tiedetään, että teräksen seosaineiden pitoisuudet korreloivat voimakkaasti keskenään, niin muodostettavaa teräksen karkenevuuden regressiomallia estimoidaan säännöllistämismenetelmillä. Eri menetelmillä estimoituja teräksen karkenevuuden regressiomallien riittävyyttä arvioidaan selityskertoimien avulla. Muodostettujen mallien selityskertoimet lasketaan mallin sovitteelle eli havainnoille, joista regressiokertoimien estimaattit on laskettu, ja ennusteelle eli havainnoille, jotka eivät olleet mukana estimoitaessa regressiomallin regressiokertoimien estimaatteja.

2 Regressioanalyysi

Regressioanalyysissä tutkitaan satunnaismuuttujien $Y^{(1)},...,Y^{(q)}$ riippuvuutta eli regressiota annetuista muuttujista $X = (x_1,...,x_p)$. Regressioanalyysissä satunnaismuuttujia $Y^{(1)},...,Y^{(q)}$ kutsutaan selitettäviksi muuttujiksi (engl. explanated variables), riippuviksi muuttujiksi (engl. dependent variables) tai vastemuuttujiksi (engl. response variables) ja muuttujia $x_1,...,x_p$ selittäviksi muuttujiksi (engl. explanatory variables), ennustaviksi muuttujiksi (engl. predictor variables) tai riippumattomiksi muutujiksi (engl. independent variables). Kun selittäviä muuttujia on enemmän kuin yksi eli $p > 1$, niin regressioanalyysiä kutsutaan usean selittäjän regressioanalyysiksi tai moniulotteiseksi regressioanalyysiksi (engl. multiple regression). Kun selitettäviä muuttujia on enemmän kuin yksi eli $q > 1$, niin regressioanalyysiä kutsutaan tässä monimuuttujaiseksi regressioanalyysiksi (engl. multivariate regression). (Wang & Chow, 1994.)

Oletetaan, että satunnaismuuttujien $Y^{(1)},...,Y^{(q)}$ ehdollinen odotusarvo ehdolla $x_1,...,x_p$ on

$$E(Y|X) = I\beta_0 + X\mathbf{B} ,$$

missä $Y = (Y^{(1)},...,Y^{(q)})$, $X = (x_1,...,x_p)$, vektori β_0 on ns. vakiovektori, jonka alkiot ovat vakiota (engl. constant term), I on ykkösvektori ja B on regressiomatriisi (engl. regression coeffients matrix). Satunnaismuuttujat $Y^{(1)},...,Y^{(q)}$ ja annetut muuttujat $x_1,...,x_p$ ovat mitattavissa olevia suureita tai joitain niiden funktiota. Vakiovektori β_0 ja regressiomatriisi B ovat tuntemattomia parametreja, jotka on kiinnitetty, ja ne tulee estimoida havaintoaineistosta. Jos jokaisessa havainnossa $i = 1,...,n$ on havaittu sekä $y_i = y_i^{(1)},...,y_i^{(q)}$ että $x_i = x_{i1},...,x_{ip}$ ja ehdollinen odotusarvo $E(Y|X)$ on lineaarinen annetulla X, niin

$$Y = E(Y|X) + \mathbf{E} = I\beta_0 + X\mathbf{B} + \mathbf{E} , \tag{2.1}$$

missä siis Y on $(n \times q)$ -havaintomatriisi, vakiovektorin β_0 ulottuvuus on $1 \times q$, I on $(n \times 1)$-ykkösvektori, X on $(n \times p)$-rakennematriisi, regressiomatriisin B ulottuvuus on $p \times q$ ja E on $(n \times q)$-jäännösmatriisi. Kun merkitään $\beta_0 = (\beta_0^{(1)},...,\beta_0^{(q)})$, $\mathbf{B} = (\beta^{(1)},...,\beta^{(q)})$ ja $\mathbf{E} = (\epsilon^{(1)},...,\epsilon^{(q)})$, niin regressiomalli (2.1) voidaan esittää myös muodossa

$$y_i^{(j)} = I\beta_0^{(j)} + \beta_1^{(j)}x_{i1} + ... + \beta_p^{(j)}x_{ip} + \epsilon_i^{(j)} , \quad i = 1,...,n \text{ ja } j = 1,...,q . \tag{2.2}$$

(Brown, 1993.)

Regressiomallissa (2.1) tuntemattomat parametrit β_0 ja **B** ovat lineaarisia satunnaismuuttujan $\boldsymbol{Y}$ ehdolliselle odotusarvolle $E(\boldsymbol{Y}|\boldsymbol{X})$, minkä seurauksena regressiomallia (2.1) voidaan pitää lineaarisena. Odotusarvon $E(\boldsymbol{Y}|\boldsymbol{X})$ lineaarisuus selittävien muuttujien $x_1,\ldots,x_p$ suhteen ei siis ole oleellinen vaatimus lineaarisessa regressiomallissa. (Brown, 1993.)

Regressiomallissa (2.1) oletetaan, että jäännösmatriisin **E** odotusarvo on nollamatriisi. Merkitään jäännösmatriisin **E** kovarianssimatriisia merkinällä $\boldsymbol{\Sigma}_E$. Kovarianssimatriisin $\boldsymbol{\Sigma}_E$ ulottuvuus on $nq \times nq$ ja se on positiivisesti definiitti matriisi, mitä merkitään $\boldsymbol{\Sigma}_E > \boldsymbol{0}$. Jos muuttujien $\boldsymbol{X}$ ja $\boldsymbol{Y}$ välinen lineaarinen riippuvuus on täydellistä, niin $\mathbf{E} = \boldsymbol{0}$. Tuskin koskaan **E** on nollamatriisi, sillä yleensä kokeessa on aina satunnaisvaihtelua, joka poikkeuttaa arvoja muuttujien $\boldsymbol{X}$ ja $\boldsymbol{Y}$ välisestä täydellistä lineaarisesta riippuvuudesta. Täten havaintomatriisin $\boldsymbol{Y}$ ja rakennematriisin $\boldsymbol{X}$ riippuvuussuhteen kuvaaminen regressiomallin avulla on käytännössä likimääräistä.

2.1 Kanoninen regressiomalli

Siirrytään käyttämään standardoituja muuttujia, joille on tehty muunnokset

$$x_{ij} = \frac{x_{ij}^* - \bar{x}_j^*}{\sqrt{\mathrm{var}(x_j^*)(n-1)}}, \quad j = 1,\ldots,p \tag{2.3}$$

missä $\bar{x}_j^*$ on alkuperäisen selittävän muuttujan x_j^* keskiarvo, $\mathrm{var}(x_j^*)$ on alkuperäisen selittävän muuttujan x_j^* varianssi, n on havaintoaineiston otoskoko ja $i = 1,\ldots,n$, ja

$$y_i^{(j)} = \frac{y_i^{*(j)} - \bar{y}^{*(j)}}{\sqrt{\mathrm{var}(y^{*(j)})(n-1)}}, \; j = 1,\ldots,q, \tag{2.4}$$

missä $\bar{y}^{*(j)}$ on alkuperäisen selitettävän muuttujan $y^{*(j)}$ keskiarvo, $\mathrm{var}(y^{*(j)})$ on alkuperäisen selitettävän muuttujan $y^{*(j)}$ varianssi, n on havaintoaineiston otoskoko, ja $i = 1,\ldots,n$. Nyt regressiomalissa (2.1) vakiovektori β_0 on nollavektori. Täten regressiomalli (2.1) voidaan kirjoittaa muotoon

$$\boldsymbol{Y} = E(\boldsymbol{Y}|\boldsymbol{X}) + \mathbf{E} = \boldsymbol{1}\beta_0 + \boldsymbol{X}\mathbf{B} + \mathbf{E} = \boldsymbol{X}\mathbf{B} + \mathbf{E} \ . \tag{2.5}$$

Määritellään seuraavaksi regressiomallin (2.5) kanoninen muoto. Olkoon matriisit $\boldsymbol{T}$, jonka ulottuvuus on $n \times n$, ja $\boldsymbol{V}$, jonka ulottuvuus on $p \times p$, ortogonalisia siten, että

$$\boldsymbol{U} = \boldsymbol{T}'\boldsymbol{Y} = \boldsymbol{T}'\boldsymbol{X}\boldsymbol{V}\boldsymbol{V}'\mathbf{B} + \boldsymbol{T}'\mathbf{E} \ , \tag{2.6}$$

missä matriisin $\boldsymbol{T}$ sarakkeina ovat neliömatriisin $\boldsymbol{X}\boldsymbol{X}'$ ominaisvektorit ja matriisin $\boldsymbol{V}$ sarakkeina neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisvektorit. Oletetaan tästä lähtien, että neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvot on järjestetty seuraavaan suuruusjärjestykseen $\lambda_1 \geq \ldots \geq \lambda_p$, minkä mukaan on järjestetty matriiseissa $\boldsymbol{V}$ ja $\boldsymbol{T}$ ominaisvektorit $\boldsymbol{v}_j, j = 1,\ldots,p$, ja $\boldsymbol{t}_j, j = 1,\ldots,n$. Tulo $\boldsymbol{T}'\boldsymbol{X}\boldsymbol{V}$ on

$$\boldsymbol{T}'\boldsymbol{X}\boldsymbol{V} = \begin{pmatrix} \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_p}) \\ \mathbf{0} \end{pmatrix} , \tag{2.7}$$

missä nollamatriisin $\mathbf{0}$ ulottuvuus on $n - p \times p$ ja neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvojen $\lambda_1,\ldots,\lambda_p$ neliöjuuret ovat matriisin $\boldsymbol{X}$ singulaariarvoja. Matriisi $\boldsymbol{X}$ voidaan kirjoittaa muotoon

$$\boldsymbol{X} = \boldsymbol{T}\begin{pmatrix} \operatorname{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_p}) \\ \mathbf{0} \end{pmatrix} \boldsymbol{V}' \ , \tag{2.8}$$

jota kutsutaan matriisin $\boldsymbol{X}$ singulaariarvohajotelmaksi. Matriisin $\boldsymbol{X}$ asteelle r on voimassa epäyhtälö $r \leq \min(n,p)$ ja se on neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ nollasta eroavien ominaisarvojen lukumäärä. Täten matriisi $\boldsymbol{X}$ voidaan esittää muodossa

$$\boldsymbol{X} = \boldsymbol{T}_r\begin{pmatrix} \operatorname{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_r}) \\ \mathbf{0} \end{pmatrix} \boldsymbol{V}_r' \tag{2.9}$$

ja tulo (2.7) on muotoa

$$\boldsymbol{T}_r'\boldsymbol{X}\boldsymbol{V}_r = \begin{pmatrix} \operatorname{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_r}) \\ \mathbf{0} \end{pmatrix} , \tag{2.10}$$

missä nollamatriisin $\mathbf{0}$ ulottuvuus on $n - r \times r$.

Kun merkitään

$$\boldsymbol{V}'\mathbf{B} = \mathbf{A} \ , \tag{2.11}$$

niin regressiomalli (2.6) voidaan kirjoittaa muotoon

$$u_i^{(j)} = \sqrt{\lambda_i}\alpha_i^{(j)} + \epsilon_i^{(j)}, \quad i = 1,...,r, \quad \text{ja} \quad u_i^{(j)} = \epsilon_i^{(j)}, \quad i = r+1,...,n, \tag{2.12}$$

missä $\epsilon^{(j)} = T\epsilon^{(j)}$, r on matriisin X aste ja $j = 1,...,q$. Tätä regressiomallia kutsutaan kanoniseksi regressiomalliksi. (Brown, 1993.)

2.2 Pienimmän neliösumman menetelmä, kun selitettäviä muuttujia on yksi

Oletetaan, että estimoitavana on regressiomalli (2.5), missä selitettäviä muuttujia on yksi eli $q = 1$. Pienimmän neliösumman menetelmässä määrätään parametrin β estimaatti siten, että estimoidun mallin antamien sovitteiden $\hat{y}$ ja havaintojen y välinen jäännösneliösumma

$$S(\beta) = \sum_{i=n}^{n} \epsilon_i^2 = (y - \hat{y})'(y - \hat{y}) = \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2 \tag{2.13}$$

on mahdollisimman pieni. Asettamalla neliösumman osittaisderivaatat

$$\frac{\partial S(\beta)}{\partial \beta} = -2X'y + 2X'X\beta$$

nolliksi saadaan normaaliyhtälöt

$$(X'X)b_{PNS} = X'y, \tag{2.14}$$

missä b_{PNS} on parametrin β pienimän neliösumman ratkaisu. Kun $r = p$, niin neliömatriisi $X'X$ on epäsingulaarinen, jolloin pienimmän neliösumman estimaattori b_{PNS} voidaan kirjoittaa muotoon

$$b_{PNS} = (X'X)^{-1}X'y. \tag{2.15}$$

(Liski & Puntanen, 1976.)

Pienimmän neliösumman menetelmässä sovite $\hat{y}_{PNS}$ on

$$\hat{y}_{PNS} = X b_{PNS} = X(X'X)^{-1}X'y = Hy, \tag{2.16}$$

missä $(n \times n)$-projektiomatriisi $H = X(X'X)^{-1}X'$ on ns. hattumatriisi. Koska pienimmän neliösumman menetelmässä jäännös e voidaan kirjoittaa muotoon

$$\boldsymbol{e} = (\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{y} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y} \ , \tag{2.17}$$

missä matriisi $\boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ on $(n \times n)$-projektiomatriisi, niin jäännöksen $\boldsymbol{e}$ riippuvuus selitettävästä muuttujasta $\boldsymbol{y}$ on lineaarista. Sovitetut jäännökset $\boldsymbol{e}$ ja rakennematriisin $\boldsymbol{X}$ sarakevektorit ovat keskenään ortogonaalisia, sillä

$$\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_{PNS}) = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{y} = 0 \tag{2.18}$$

ja jäännöksen $\boldsymbol{e}$ kovarianssimatriisi on

$$\operatorname{cov}(\boldsymbol{e},\boldsymbol{e}) = E\boldsymbol{e}\boldsymbol{e}' = E(\boldsymbol{P}\boldsymbol{y}\boldsymbol{y}'\boldsymbol{P}') = \boldsymbol{P}E(\boldsymbol{y}\boldsymbol{y}')\boldsymbol{P}' = \boldsymbol{P}\operatorname{cov}(\boldsymbol{y})\boldsymbol{P}' \ . \tag{2.19}$$

Koska $E(\boldsymbol{e}'\boldsymbol{e}) = E(\boldsymbol{y}'\boldsymbol{P}'\boldsymbol{P}\boldsymbol{y}) = \sigma_y^2(n - p)$, niin virhevarianssin σ_y^2 pienimmän neliösumman estimaattori on

$$\hat{\sigma}_y^2 = \frac{\boldsymbol{e}'\boldsymbol{e}}{n - p} \ . \tag{2.20}$$

Estimaatorin $\boldsymbol{b}_{PNS}$ kovarianssimatriisi on

$$\operatorname{cov}(\boldsymbol{b}_{PNS}) = \sigma_y^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \ . \tag{2.21}$$

Koska estimaattorin $\mathbf{b}$ ja sen todellisen arvon β poikkeuman neliön odotusarvo on estimaattorin keskineliövirhe

$$\operatorname{MSE}(\boldsymbol{b}) = E(\boldsymbol{b} - \beta)'(\boldsymbol{b} - \beta) = E(\boldsymbol{b} - E\boldsymbol{b})^2 + (E\boldsymbol{b} - \beta)^2 = \operatorname{var}(\boldsymbol{b}) + B^2(\boldsymbol{b}) \ ,$$

missä $\operatorname{var}(\boldsymbol{b})$ on estimaattorin varianssi ja $B^2(\boldsymbol{b})$ on estimaattorin harhan neliö, ja estimaattori $\boldsymbol{b}_{PNS}$ on harhaton, niin estimaattorin $\boldsymbol{b}_{PNS}$ kokonaisvarianssi ja samalla sen keskineliövirhe on

$$\operatorname{var}(\boldsymbol{b}_{PNS}) = \sigma_y^2 \operatorname{tr}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma_y^2 \sum_{j=1}^{p} \lambda_j^{-1} = \operatorname{MSE}(\boldsymbol{b}_{PNS}) \ . \tag{2.22}$$

Kun lineaarisessa regressiomallissa (2.5) $q = 1$, $r = p$, $E(\epsilon) = 0$ ja $\operatorname{cov}(\epsilon) = \sigma_y^2\boldsymbol{I}_n$, niin Gauss-Markovin lauseen mukaan pienimmän neliösumman estimaattori $\boldsymbol{b}_{PNS}$ on tarkentuva ja harhaton

estimaattori, jolla on pienin mahdollinen varianssi, eli estimaattori $\boldsymbol{b}_{PNS}$ on BLUE (engl. Best Linear Unbiased Estimator) (Wang & Chow, 1994).

Oletetaan, että tehtävänä on ennustaa annetulla selittävän muuttujan $\boldsymbol{x}_0$ arvolla tuntematonta selitettävän muuttujan y_0 arvoa käyttäen mallia

$$y_0 = E(y_0|\boldsymbol{x}_0) + \epsilon = \boldsymbol{x}_0'\boldsymbol{\beta} + \epsilon \ , \quad E(\epsilon) = 0 \ , \quad \text{var}(\epsilon) = \sigma_y^2 \ . \tag{2.23}$$

Pienimmän neliösumman menetelmässä tuntemattoman selitettävän muuttujan y_0 ennuste on $\hat{y}_{0PNS}$ $= \boldsymbol{x}_0'\boldsymbol{b}_{PNS}$. Sen varianssi ja keskineliövirhe on

$$\text{var}(\hat{y}_{0PNS}) = \sigma_y^2(1 + 1/n) + \sigma_y^2 \text{tr}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0) = \text{MSE}(\hat{y}_{0PNS}) \ . \tag{2.24}$$

(Brown, 1993.)

Yllä olevassa ennusteen keskineliövirheessä annettu selittävän muuttujan arvo $\boldsymbol{x}_0$ on siis keskistetty. Tällöin mitä enemmän alkuperäinen selittävän muuttujan arvo $\boldsymbol{x}^*_0$ eroaa arvosta $\bar{\boldsymbol{x}} = (\bar{x}_1^*,\ldots,\bar{x}_p^*)$, missä $\bar{x}_j^*$ on alkuperäisen selittävän muuttujan x_j^* keskiarvo, $j = 1,\ldots,p$, niin sitä suurempia ovat annetun $\boldsymbol{x}_0$ alkiot ja sitä suurempia ovat var($\hat{y}_{0PNS}$) ja MSE($\hat{y}_{0PNS}$). Geometrisesti tulkittuna regressiokertoimien estimaateissa olevat virheet aiheuttavat sen, että estimoitu regressiosuora on ikään kuin kiertynyt suhteessa todelliseen regressiosuoraan. Koska regressiosuora kulkee keskiarvopisteen kautta, niin kiertäminen tapahtuu tämän pisteen ympäri. Tämän vuoksi ennusteen epävarmuus on sitä suurempi mitä kauempana selittävän muuttujan arvo on keskiarvostaan. (Ranta, 1991.)

Jos regressiomalli on esitetty kanonisessa muodossa (2.6), missä selitettäviä muuttujia on yksi, niin pienimmän neliösumman estimaattori $\boldsymbol{a}_{PNS}$ on

$$a_{jPNS} = u_j/\sqrt{\lambda_j} \ , \quad j = 1,\ldots,r \ , \qquad \text{ja} \qquad a_{jPNS} = 0 \ , \quad j = r+1,\ldots,n \ , \tag{2.25}$$

missä r on rakennematriisin $\boldsymbol{X}$ aste, sillä normaaliyhtälöt (2.14) ovat nyt muotoa

$$\boldsymbol{\Lambda}_r \boldsymbol{a}_{PNS} = \left(\sqrt{\boldsymbol{\Lambda}_r} \ \ \boldsymbol{0}\right)\boldsymbol{u} \ ,$$

joten

$$\boldsymbol{a}_{PNS} = \Lambda_r^{-1}(\sqrt{\Lambda}_r \ \boldsymbol{0})\boldsymbol{u} = (\sqrt{\Lambda}_r^{-1} \ \boldsymbol{0})\boldsymbol{u} \ ,$$

missä $\Lambda_r = \mathrm{diag}(\lambda_1,...,\lambda_r)$ ja nollamatriisin $\boldsymbol{0}$ ulottuvuus on $n - r \times r$.

Pienimmän neliösumman estimaattorin $\boldsymbol{a}_{PNS}$ kovarianssimatriisi on

$$\mathrm{cov}(\boldsymbol{a}_{PNS}) = \sigma_y^2 \Lambda_r^{-1} = \sigma_y^2 \mathrm{diag}(1/\lambda_1,...,1/\lambda_r) \ . \tag{2.26}$$

Huomattavaa on, että estimaattorit a_{jPNS}, $j = 1,...,r$, ovat korreloimattomia ja estimaattorit b_{jPNS}, $j = 1,...,r$, korreloivat. Koska estimaattori $\boldsymbol{a}_{PNS}$ on kanonisen regressiomallin (2.6) parametrin $\boldsymbol{\alpha}$ pienimmän neliösumman estimaattori, niin se on Gauss-Markovin lauseen mukaan BLUE. Koska estimaattori $\boldsymbol{a}_{PNS}$ on harhaton, sen kokonaisvarianssi ja samalla keskineliövirhe on

$$\mathrm{var}(\boldsymbol{a}_{PNS}) = \mathrm{MSE}(\boldsymbol{a}_{PNS}) = \sigma_y^2 \mathrm{tr}(\Lambda_r^{-1}) = \sigma_y^2 \sum_{j=1}^{r} \lambda_j^{-1} \ . \tag{2.27}$$

Harhattomilla estimaattoreilla $\boldsymbol{b}_{PNS}$ ja $\boldsymbol{a}_{PNS}$ on siis sama kokonaisvarianssi ja siten myös yhtäsuuret keskineliövirheet.

Kun matriisilla $\Lambda_p^{1/2}$, joka kuuluu havaintoaineistoon, ennustetaan tuntematonta $\boldsymbol{u}$, niin ennusteen $\hat{\boldsymbol{u}}_{PNS}$ keskineliövirhe on

$$\mathrm{MSE}(\hat{\boldsymbol{u}}_{PNS}) = \sigma_y^2(n+1) + E((\boldsymbol{a}_{PNS} - \boldsymbol{\alpha})'\Lambda_r(\boldsymbol{a}_{PNS} - \boldsymbol{\alpha})) \tag{2.28}$$

$$= \sigma_y^2(n+1) + \sum_{j=1}^{p} \lambda_j \mathrm{var}(a_{jPNS}) = (n+1+p)\sigma_y^2 \ ,$$

missä $\mathrm{var}(a_{jPNS}) = \sigma_y^2/\lambda_j$.

2.3 Pienimmän neliösumman menetelmä, kun selitettäviä muutujia on useita

Oletetaan nyt, että estimoitavana on regressiomalli (2.5), missä selitettäviä muuttujia on useita eli $q > 1$. Pienimmän neliösumman menetelmässä valitaan parametrin B estimaatti siten, että kaikkien havaintojen poikkeamien neliösumma

$$S(\mathbf{B}) = (\boldsymbol{Y} - \hat{\boldsymbol{Y}})'(\boldsymbol{Y} - \hat{\boldsymbol{Y}}) = \sum_{i=1}^{n} \sum_{j=1}^{q} (y_i^{(j)} - \beta_1^{(j)} x_{i1} - \ldots - \beta_p^{(j)} x_{ip})^2$$

on mahdollisimman pieni. Tällöin parametrin B pienimmän neliösumman estimaattori $\boldsymbol{B}_{PNS}$ ratkaistaan normaaliyhtälöistä

$$(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{B}_{PNS} = \boldsymbol{X}'\boldsymbol{Y} \tag{2.29}$$

(vrt. (2.14)). Kun nelioatriisi $\boldsymbol{X}'\boldsymbol{X}$ on epäsingulaarinen, niin pienimmän neliösumman estimaattori $\boldsymbol{B}_{PNS}$ voidaan kirjoittaa muotoon

$$\boldsymbol{B}_{PNS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \, . \tag{2.30}$$

Havaintomatriisin $\boldsymbol{Y}$ pienimmän neliösumman sovite on

$$\hat{\boldsymbol{Y}}_{PNS} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y} \, , \tag{2.31}$$

missä $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ on projektiomatriisi. Koska

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_{PNS} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} \, , \tag{2.32}$$

niin $\mathrm{E}(\boldsymbol{E}'\boldsymbol{E}) = \mathrm{E}(\boldsymbol{Y}'(\boldsymbol{I}_p - \boldsymbol{H})\boldsymbol{Y}) = \Sigma_Y \operatorname{tr}(\boldsymbol{I}_p - \boldsymbol{H}) = \Sigma_Y (n - p)$ ja täten

$$\hat{\Sigma}_Y = \frac{\boldsymbol{Y}'(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}}{n - p} \tag{2.33}$$

(vrt. (2.20)) (Wang & Chow, 1994) .

Estimaattorin vek($\boldsymbol{B}_{PNS}$) kovarianssimatriisi on

$$\operatorname{cov}(\operatorname{vek}(\boldsymbol{B}_{PNS})) = \Sigma_Y \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1} \, , \tag{2.34}$$

missä $\otimes$ on Kroneckerin tulo, eli $\operatorname{cov}(\boldsymbol{b}^{(i)}_{PNS}, \boldsymbol{b}^{(j)}_{PNS}) = \sigma^{(ij)}(\boldsymbol{X}'\boldsymbol{X})^{-1}$, $i, j = 1,\ldots,q$ (Wang & Chow, 1994). Selitettävien muuttujien $\boldsymbol{y}^{(i)}$ ja $\boldsymbol{y}^{(j)}$ korkeat korrelaatiot kasvattavat siis kovarianssia $\operatorname{cov}(\boldsymbol{b}^{(i)}_{PNS}, \boldsymbol{b}^{(j)}_{PNS})$, $i \neq j$, $i, j = 1,\ldots,q$. Koska Brownin ja Paynen (1975) mukaan

$$\mathrm{MSE}(\boldsymbol{B}) = E(\sum_{i=1}^{p} \sum_{j=1}^{q} (b_i^{(j)} - \beta_i^{(j)})^2) \, , \tag{2.35}$$

missä $\boldsymbol{B} = (\boldsymbol{b}^{(1)},\ldots,\boldsymbol{b}^{(q)})$ on jokin paramerin $\mathrm{B} = (\beta^{(1)},\ldots,\beta^{(q)})$ estimaattori, niin estimaattorin $\boldsymbol{B}_{PNS}$ varianssi ja keskineliövirhe on varianssien $\mathrm{var}(\boldsymbol{b}_{PNS}^{(j)})$, $j = 1,\ldots,q$, summa.

Huomattavaa on, että koska monimuuttujainen pienimmän neliösumman estimaattori $\boldsymbol{B}_{PNS}$ voidaan kirjoittaa muotoon $\boldsymbol{B}_{PNS} = (\boldsymbol{b}^{(1)}{}_{PNS},\ldots,\boldsymbol{b}^{(q)}{}_{PNS})$, missä

$$\boldsymbol{b}_{PNS}^{(j)} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}^{(j)}, \qquad j = 1,\ldots,q, \tag{2.36}$$

niin estimaattorit $\boldsymbol{b}^{(j)}{}_{PNS}$, $j = 1,\ldots,q$, ovat q yksimuuttujaisen regressiomallin

$$\boldsymbol{y}^{(j)} = \boldsymbol{X}\beta^{(j)} + \epsilon^{(j)}, \qquad j = 1,\ldots,q \tag{2.37}$$

parametrien $\beta^{(j)}$ $j = 1,\ldots,q$, pienimmän neliösumman estimaattoreita. Koska ratkaistaessa pienimmän neliösumman estimaattoria $\boldsymbol{B}_{PNS}$ ei käytetä matriisia Σ_Y, niin määrättäessä estimaattoria $\boldsymbol{B}_{PNS}$ ei hyödynnetä q normaaliyhtälöiden yhteisinformaatiota.

2.4 Multikollineaarisuus

Yleisesti multikollineaarisuus termiä käytetään, kun lineaarisen mallin selittävistä muuttujista $\boldsymbol{x}_j$, $j = 1,\ldots,p$, kaikki tai osajoukko $\boldsymbol{x}_j$, $j = A + 1,\ldots,p$, on lähes lineaarisesti riippuvia. Vektorit $\boldsymbol{x}_j$, $j = 1,\ldots,p$, ovat multikollineaarisia, jos on olemassa vakiot $\boldsymbol{c}_j \neq 0$, siten että

$$\sum_{j=1}^{p} c_j \boldsymbol{x}_j \approx 0, \tag{2.38}$$

missä vektorit $\boldsymbol{x}_j$, $j = 1,\ldots,p$, on keskistetty. Täten selittävien muuttujien $\boldsymbol{x}_j$, $j = 1,\ldots,p$, multikollineaarisuudella tarkoitetaan sitä, että selittävien muuttujien $\boldsymbol{x}_j$, $j = 1,\ldots,p$, välillä on lähes suoria tai lineaarikombinaatioiden välisiä riippuvuuksia. Mitä lähempänä summa (2.38) on nollaa sitä suurempi on selittävien muuttujien $\boldsymbol{x}_j$, $j = 1,\ldots,p$, multikollineaarisuuden aste. Harvoin selittävät muuttujat $\boldsymbol{x}_j$, $j = 1,\ldots,p$, ovat täysin lineaarisesti riippuvia, jolloin summa (2.38) on nolla ja rakennematriisin $\boldsymbol{X}$ aste $r < p$. Yleensä regressioanalyysissä ongelmana onkin selittävien muuttujien multikolineaarisuuden asteen arvioinen eikä täydellisen lineaarisen riippuvuuden tunnistaminen. (Gunst & Mason, 1980; Leskinen, 1977.)

Selittävien muuttujien multikollineaarisuus samaistetaan usein selittävien muuttujien voimakkaaseen korreloitumiseen ja multikollineaarisuutta arvioidaan selittävien muuttujien korrelaatiokertoimien avulla. Jos multikollineaarisuutta on useamman kuin kahden selittävän

muuttujan välillä, niin multikollineaarisuutta ei voida päätellä selittävien muuttujien korrelaatiomatriisista, koska selittävät muuttujat voivat olla lähes lineaarisesti riippuvia, vaikka niiden väliset korrelaatiokertoimet olisivatkin pieniä (Leskinen, 1977).

Seuraavassa on muutamia tapoja arvioida selittävien muuttujien $x_j, j = 1,...,p$, multikollineaarisuutta:

1) Jos selittävien muuttujien $x_j, j = 1,...,p$ korrelaatiomatriisissa on korrelaatiokertoimia, jotka ovat suurempia kuin 0.7, niin selittävien muuttujien multikolineaarisuuden astetta voidaan pitää korkeana (Gunst & Mason, 1980).

2) Oletetaan, että neliömatriisin $X'X$ pienimmät ominaisarvot $\lambda_j, j = A + 1,...,p$, ovat lähes nollia. Ominaisarvojen määritelmästä seuraa, että

$$X'Xv_j - \lambda_j v_j \approx 0, \qquad j = A + 1,...,p \ . \tag{2.39}$$

missä v_j on ominaisarvoa λ_j vastaava neliömatriisin $X'X$ ominaisvektori. Kun kerrotaan yllä oleva yhtälö ominaisvektorilla v_j, niin saadaan

$$(Xv_j)'(Xv_j) - \lambda_j v_j v_j' - \lambda_j \approx 0 \ . \tag{2.40}$$

Kun $Xv_j = z_j$, niin $z_j'z_j = (z_{j1}^2 + .. + z_{jp}^2)$, missä alkiot $z_{ji}, i = 1,...,p$, eivät ole negatiivisia, niin yllä olevassa yhtälössä $z_j = Xv_j \approx 0, j = A + 1,...,p$. Jos $\lambda_j \approx 0$, niin

$$Xv_j - \sum_{i=1}^{p} v_{ij}x_i \approx 0 \tag{2.41}$$

(vrt. (2.38)). Täten kun neliömatriisilla $X'X$ on pieni ominaisarvo $\lambda_j \approx 0$, niin selittävät muuttujat $x_j, j = 1,...,p$, ovat multikollineaarisia. Kun $\lambda_j \approx 0$, niin ominaisvektorin v_j suuret alkiot määräävät, mitkä selittävät muuttujat $x_j, j = 1,...,p$, ovat eniten multikollineaarisia. (Wang & Chow, 1994.)

3) Neliömatriisin $X'X$ determinantti on

$$\det(X'X) - \prod_{j=1}^{p} \lambda_j \ . \tag{2.42}$$

Koska det($X'X$) on matriisin $X'X$ ominaisarvojen tulo, niin selittävien muuttujien multikollineaarisuuden asteen ollessa korkea det($X'X$) $\approx$ 0. Jos selittävien muuttujien välillä on täydellisiä

lineaarisia riippuvuuksia, niin $\lambda_j = 0$, jolloin $\det(X'X) = 0$. (Leskinen, 1977.)

4) Matriisin X sanotaan olevan kollineaarinen, jos

$$f_1 = \sqrt{\lambda_1 / \lambda_p}\,, \tag{2.43}$$

missä λ_1 on suurin ja λ_p on pienin neliömatriisin $X'X$ ominaisarvo, on suuri (Jackson, 1991). Kun rakennematriisi X on kollineaarinen, niin se on myös multikollineaarinen.

5) Jos rakennematriisin X korrelaatiomatriisi $X'X$ on ortogonalinen eli selittävät muuttujat x_j, $j = 1,...,p$, ovat riippumattomia, niin $\lambda_1 = ... = \lambda_p$, jolloin $1/\lambda_1 = ... = 1/\lambda_p = 1$. Mitä enemmän lukujen $1/\lambda_j$, $j = 1,...,p$, summa eroaa selittävien muuttujien lukumäärästä p sitä enemmän selittävien muuttujien x_j, $j = 1,...,p$, välillä on multikollineaarisuutta.

6) Kun oletetaan, että selittävät muuttujat ovat satunnaismuuttujia ja ne noudattavat p-ulotteista normaalijakaumaa, niin multikollineaarisuuden havaitseminen voidaan perustaa testisuureeseen

$$\chi^2 = -(n - 1 - (1/6))(2p + 5)\ln|X'X|\,, \tag{2.44}$$

joka noudattaa suurilla otoskoon n arvoilla selittävien muuttujien ollessa korreloimattomia χ^2 -jakaumaa vapausastein $p(p - 1)/2$ (Leskinen, 1977).

Huomattavaa on, että kaikkia edellä olevien testien tuloksia on käsiteltävä lähinnä suuntaa antavina.

Kun on todettu, että selittävien muuttujien välillä on multikollineaarisuutta, niin ratkaistaessa selittävien muuttujien multikollineaarisuuden ongelmaa on hyödyllistä yrittää selvittää selittävien muuttujien multikollineaarisuuden lähde, jotta tiedettäisiin, miten tätä ongelmaa tulisi käsitellä regressioanalyysissä. Seuraavassa on kolme selittävien muuttujien multikollineaarisuuden päälähdettä Leskisen (1977) mukaan:

1) Multikollineaarisuutta saattaa esiintyä mallissa, jos aineisto on kerätty virheellisesti. Jos selittävien muuttujien multikollineaarisuus johtuu otannasta, niin on kerättävä joko uusi aineisto tai riittävä määrä uutta aineistoa. Multikollineaarisuutta saattaa aiheuttaa myös vieraat (outliers)

havainnot.

2) Regressiomallin rakennematriisi X on väärin määritelty. Yleismmin regressiomalliin valitaan liian paljon selittäviä muuttujia, joissa 'ylimääräiset selittävät muuttujat' korreloivat muiden muuttujien kanssa.

3) Regressiomalliin sisäiset rajoitukset voivat aiheuttaa selittävien muuttujien välille multikollineaarisuutta. Tällä tarkoitetaan sitä, että selittävien muuttujien multikollineaarisuus on nyt tutkittavan ilmiön luontainen ominaisuus, jota ei voida muuttaa. Tällöin jos multikollineaarisuuden poistamiseksi poistettaisiin regressiomallista multikollineaarisia selittäviä muuttujia, niin tutkittavaa ilmiötä kuvattaisiin puutteellisesti. Jos selittävien muuttujien multikollineaarisuus ei ole vain otoksesta johtuvaa, vaan se osa tutkittavan ilmiön teoreettista ominaisuutta, niin tällöin on turvauduttava harhaisiin estimointimenetelmiin eli säännöllistämismenetelmiin.

2.5 Selittävien muuttujien multikollineaarisuuden vaikutus pienimmän neliösumman menetelmässä

Selittävien muuttujien x_j, $j = 1,...,p$, korkea multikollineaarisuuden aste aiheuttaa epätarkkuutta pienimmän neliösumman estimaattorissa b_{PNS}, koska siinä joudutaan kääntämään neliömatriisi $X'X$. Neliömatriisin $X'X$ käänteismatriisi $(X'X)^{-1}$ on

$$(X'X)^{-1} = V_p \Lambda_p^{-1} V_p = \sum_{j=1}^{p} (1/\lambda_j) v_j v_j', \qquad (2.45)$$

missä $\Lambda_p = \mathrm{diag}(\lambda_1,...,\lambda_p)$, $V_p = (v_1,...,v_p)$, λ_j on neliömatriisin $X'X$ ominaisarvo ja v_j sitä vastaava ominaisvektori, $j = 1,...,p$. Jos selittävät muuttujat x_j, $j = 1,...,p$, ovat täysin lineaarisesti riippuvia, niin neliömatriisin $X'X$ pienin ominaisarvo $\lambda_p = 0$ ja $\det(X'X) = 0$. Jos käänteismatriisissa joudutaan jakamaan nollalla, niin käänteismatriisia $(X'X)^{-1}$ ei ole olemassa ja matriisin $X'X$ sanotaan olevan singulaarinen. Jos selittävät muuttujat eivät ole täysin lineaarisesti riippuvia, mutta niiden multikollineaarisuuden aste on korkea, niin $\lambda_p \approx 0$ ja $\det(X'X) \approx 0$. Tällöin neliömatriisia $X'X$ on epäsingulaarinen eli on olemassa yksikäsitteinen käänteismatriisi $(X'X)^{-1}$, mutta käänteismatriisi $(X'X)^{-1}$ saattaa olla epävakaa ja sen alkiot saattavat olla kohtuuttoman suuria, koska käänteismatriisissa $(X'X)^{-1}$ joudutaan jakamaan hyvin pienellä positiivisella luvulla. Kun $\lambda_p \approx 0$, niin $\det(X'X)$

≈ 0 ja neliömatriisin $X'X$ sanotaan olevan lähes singulaarinen.

Oletetaan nyt, että selittävät muuttujat x_j, $j = 1,...,p$, ovat lineaariset riippuvia, jolloin rakennematriisin $X = (x_1,...,x_p)$ aste r on pienempi kuin p. Tällöin pienimmän neliösumman estimaattori b_{PNS} ei ole yksikäsitteinen, koska käänteismatriisia $(X'X)^{-1}$ ei ole olemassa. Tällöin estimaattori b_{PNS} ei ole parametrin β harhaton estimaattori, sillä jos estimaattori b_{PNS} on parametrin β harhaton estimaattori, niin $E(b_{jPNS}) = \beta_j$, kaikilla $j = 1,...,p$, jolloin $(X'X)^{-1}(X'X) = I_p$, mikä on nyt mahdotonta, koska rakennematriisin X aste r on pienempi kuin p. (Wang & Chow, 1994.)

Lause 2.1 Jos rakennematriisin $X = (x_1,...,x_p)$ aste r on pienempi kuin p, niin parametrilla β ei ole olemassa harhatonta estimaattoria (Wang & Chow, 1994).

Todistus: Oletetaan, että C on $(p \times n)$ -matriisi ja Cy on parametrin β harhaton estimaattori, jolloin $CX = I_p$. Koska $p = \text{rank}(I_p) = \text{rank}(CX) \leq \text{rank}(X) < p$, niin matriisia C ei ole olemassa. Täten parametrilla β ei ole olemassa harhatonta estimaattoria. □

Koska pienimmän neliösumman estimaattorin b_{PNS} kovarianssimatriisi on

$$\text{cov}(b_{PNS}) = \sigma_y^2 (X'X)^{-1},$$

niin selittävien muuttujien multikolineaarisuuden asteen ollessa korkea neliömatriisi $X'X$ on lähes singulaarinen, mikä kasvattaa estimaattoreiden b_{jPNS}, $j = 1,...,p$, välisiä kovariansseja. Täten, kun selittävät muuttujat ovat multikollineaarisia, niin kovarianssit $\text{cov}(b_{jPNS}, b_{iPNS})$ ja varianssit $\text{var}(b_{jPNS})$ saattavat olla kohtuuttoman suuria, $i, j = 1,...,p$.

Pienimmän neliösumman estimaattorin b_{PNS} neliöidun pituuden odotusarvo on

$$E(b'_{PNS} b_{PNS}) = \beta'\beta + \text{var}(b_{PNS}) = \beta'\beta + \sigma_y^2 \,\text{tr}(X'X)^{-1} \qquad (2.46)$$

(Brown, 1993). Koska yllä olevassa yhtälössä on käänteismatriisi $(X'X)^{-1}$, niin korkea selittävien muuttujien multikollineaarisuuden aste lisää estimaattorin b_{PNS} etäisyyttä sen teoreettisesta arvosta β, jolloin estimaattorit b_{jPNS}, $j = 1,...,p$, saattavat olla itseisarvoltaan kohtuuttoman suuria. Koska selittävien muuttujien korkea multikollineaarisuuden aste epävakauttaa matriisia $(X'X)^{-1}$, niin se epävakauttaa myös estimaattoria b_{PNS} ja estimaattoreilla b_{jPNS}, j = 1,...,p, saattaa olla väärät etumerkit.

Estimaattorin b_{PNS} varianssi ja samalla sen keskineliövirhe on

$$\text{var}(\boldsymbol{b}_{PNS}) = \text{MSE}(\boldsymbol{b}_{PNS}) = \sigma_y^2 \,\text{tr}(\boldsymbol{X}'\boldsymbol{X})^{-1},$$

missä selitettävien muuttujien ollessa multikollineaarisia joudutaan kääntämään lähes singulaarinen matriisi $\boldsymbol{X}'\boldsymbol{X}$. Tällöin estimaattorin $\boldsymbol{b}_{PNS}$ keskineliövirhe saattaa olla kohtuuttoman suuri, koska selittävien muuttujien multikollineaarisuus on kasvattanut estimaattorin $\boldsymbol{b}_{PNS}$ kokonaisvarianssia kohtuuttomasti.

Sovitteen $\hat{\boldsymbol{y}}_{PNS} = \boldsymbol{X}\boldsymbol{b}_{PNS}$ korrelaatiomatriisi on $\boldsymbol{b}_{PNS}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_{PNS}$. Mitä enemmän selittävät muuttujat korreloivat sitä enemmän korreloivat myös sovitteet $\hat{y}_{jPNS}, j = 1,...,n$.

Kun y_0 on määritelty yhtälöllä (2.23), niin ennusteen $\hat{y}_{0PNS}$ varianssi ja keskineliövirhe on

$$\text{var}(\hat{y}_{0PNS}) = \text{MSE}(\hat{y}_{0PNS}) = \sigma_y^2(1 + 1/n) + \sigma_y^2 \,\text{tr}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0).$$

Jos annetun $\boldsymbol{x}_0$ multikollineaarisuusyhtälö (2.38) on vastaava kuin matriisin $\boldsymbol{X}$ multikollineaarisuusyhtälö (2.38), niin annetulla $\boldsymbol{x}_0$ voidaan melko tarkasti ennustaa tuntematonta y_0 selittävien muuttujien multikollineaarisuudesta huolimatta. Kun annettujen $x_{01},...,x_{0p}$ multikollineaarisuuden aste on pienempi kuin muuttujien $\boldsymbol{x}_1,...,\boldsymbol{x}_p$ multikollineaarisuuden aste, niin tämä kasvattaa ennusteen $\hat{y}_{0PNS}$ keskineliövirhettä.

Selittävien muuttujien multikollineaarisuuden asteen ollessa korkea pienimmän neliösumman estimaattoreiden $b_{jPNS}, j = 1,...,p$, varianssit ja kovarianssit saattavat olla siis kohtuuttoman suuria. Tällöin regressiokertoimien $\beta_j, j = 1,...,p$, pienimmän neliösumman estimaatit ovat voineet liukua kauaksi pois niiden todellista arvoista ja estimaattorin $\boldsymbol{b}_{PNS}$ harhattomuudella ei ole enää merkitystä. Täten kun selittävät muuttujat ovat multikollineaarisia, niin yksittäisten regressiokertoimen $\beta_1,....,\beta_p$ pienimmän neliösumman estimaateilla ei voida tehdä mitään sisällöllisiä tulkintoja regressiomallin rakenteesta (Leskinen, 1977). Geometrisesti tulkittuna selittävien muuttujien multikolineaarisuuden vaikutus parametrien estimointiin nähdään selvästi: jos kaksi selittävää muuttujaa korreloivat voimakkaasti, niin niiden sirontakuvio on litteä ellipsi (Ranta, 1991). Regressiotason sovittaminen tapahtuu tällöin suppealle alueelle keskittyneiden havaintojen perusteella. Tällainen kovin kapealle pohjautuva estimointi on epätarkkaa. Lisäksi nämä korreloituneet muuttujat liikkuvat yhdessä. Säännöllistämismenetelmillä voidaan saada parametrin β estimaattori, joka on keskimäärin tarkempi kuin estimaattori $\boldsymbol{b}_{PNS}$ selittävien muuttujien ollessa multikollineaarisia. (Leskinen, 1977; Ranta, 1991.)

3 Harjaregressio

Harjaregressiolla (engl. ridge regression) approksimoidaan rakennematriisin X neliömatriisin käänteismatriisia $(X'X)^{-1}$ matriisilla

$$G_H = (X'X + kI_p)^{-1} , \tag{3.1}$$

missä harjaparametri (engl. ridge parameter) $k \geq 0$ ja neliömatriisi $X'X$ on korrelaatiomuodossa (Hoerl & Kennard, 1970a). Kun harjaregressiossa lisätään tarpeeksi pieni positiivinen luku k neliömatriisin $X'X$ diagonaalille, niin selittävien muuttujien ollessa multikollineaarisia ainoastaan hyvin pienet neliömatriisin $X'X$ ominaisarvot kasvavat ja neliömatriisin $X'X$ kohtuullisilla ja suurilla ominaisarvoilla tulee olemaan lähes sama approksimaatio $\lambda_j \approx \lambda_j + k$. Koska neliömatriisin $X'X + kI_p$ pienimmät ominaisarvot ovat suurempia kuin neliömatriisin $X'X$ pienimmät ominaisarvot, niin matriisi G_H on vakaampi kuin käänteismatriisi $(X'X)^{-1}$ (ks. (2.45)). Täten approksimaatiossa (3.1) harjaparametrin k avulla pyritään vähentämään selittävien muuttujien multikollineaarisuuden vaikutusta neliömatriisissa $X'X$.

Kun regressiomallissa selittävät muuttujat ovat multikollineaarsia ja harjaparametri k on valittu sopivasti approksimaatiossa (3.1), niin harjaregressiossa saadaan regressiokertoimien estimaatit, jotka eivät ole itseisarvoltaan liian suuria, niillä on oikeat etumerkit ja ne ovat vakaampia havaintoaineiston vaihtelun suhteen kuin vastaavat pienimmän neliösumman estimaatit (Leskinen, 1977). Harjaparametri k pyritään valitsemaan approksimaatiossa (3.1) siten, että selittävien muuttujien multikollineaarisuuden asteen ollessa korkea harjaregressiolla saadaan regressiokertoimien β harhainen estimaattori, jonka keskineliövirhe on pienempi kuin estimaattorin b_{PNS} keskineliövirhe.

3.1 Yksimuuttujaiset harjaestimaattorit $b_H(k)$ ja a_H(k)

Tarkastellaan regressiomallia (2.5), missä $q = 1$ ja rakennematriisin X neliömatriisi $X'X$ on korrelaatiomatriisi. Ratkaistaan parametrin β harjaestimaattori $b_H(k)$ (engl. ordinary ridge estimator) normaaliyhtälöistä

$$(X'X + kI_p)b_H(k) = X'y \tag{3.2}$$

missä harjaparametri $k \geq 0$. Harjaestimaattori $\boldsymbol{b}_H(k)$ on

$$\boldsymbol{b}_H(k) = (\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{X}'\boldsymbol{y} \ , \tag{3.3}$$

kun $(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)$ on epäsingulaarinen. Yleistetyksi harjaestimaattoriksi (engl. generalized ridge estimator) kutsutaan estimaattoria, joka ratkaistaan normaaliyhtälöistä

$$(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{K})\boldsymbol{b}_H(K) = \boldsymbol{X}'\boldsymbol{y} \ , \tag{3.4}$$

missä $\boldsymbol{K} > \boldsymbol{0}$ ja $\boldsymbol{K} = \text{diag}(k_1,...,k_p)$. (Hoerl & Kennard, 1970a & 1970b.)

Harjaestimaattori esitetään usein kanonisessa muodossa. Oletetaan, että regressiomalli on määritelty kanonisessa muodossa (2.6), missä $q = 1$. Tällöin normaaliyhtälöt (3.2) ovat

$$(\boldsymbol{V}'\boldsymbol{X}'\boldsymbol{T}\boldsymbol{T}'\boldsymbol{X}\boldsymbol{V} + k\boldsymbol{I}_p)\boldsymbol{V}'\boldsymbol{b}_H(k) = \boldsymbol{V}'\boldsymbol{X}'\boldsymbol{T}\boldsymbol{T}'\boldsymbol{y} \ . \tag{3.5}$$

Kun käytetään matriisin $\boldsymbol{X}$ singulaariarvohajotelmaa (2.8), missä $r = p$, ja lisäksi merkintätapaa $\boldsymbol{T}'_p\boldsymbol{y} = \boldsymbol{u}$, niin yllä olevat normaaliyhtälöt ovat

$$(\boldsymbol{\Lambda}_p + k\boldsymbol{I}_p)\boldsymbol{V}'\boldsymbol{b}_H(k) = \sqrt{\boldsymbol{\Lambda}_p}\boldsymbol{u} \ , \tag{3.6}$$

$$\boldsymbol{V}'\boldsymbol{b}_H(k) = \boldsymbol{a}_H(k) = \frac{\sqrt{\boldsymbol{\Lambda}_p}\boldsymbol{u}}{\boldsymbol{\Lambda}_p + k\boldsymbol{I}_p} \ . \tag{3.7}$$

Saatu kanoninen harjaestimaattori $\boldsymbol{a}_H(k)$ voidaan kirjoittaa muotoon

$$a_{jH}(k) = \frac{u_j\sqrt{\lambda_j}}{(\lambda_j + k)} \ , \quad j = 1,...,p \ . \tag{3.8}$$

Kun on laskettu kanoninen harjaestimaattori $\boldsymbol{a}_H(k)$, niin aina voidaan tehdä muunnos

$$\boldsymbol{V}\boldsymbol{a}_H(k) = \boldsymbol{b}_H(k) \ , \tag{3.9}$$

jos on tarvetta esittää harjaestimaattori muodossa $\boldsymbol{b}_H(k)$.

Harjaestimaattorin $\boldsymbol{a}_H(k)$ ja pienimmän neliösumman estimaattorin $\boldsymbol{a}_{PNS}$ välisestä yhteydestä saadaan harjaestimaattorille $\boldsymbol{a}_H(k)$ vaihtoehtoinen esitys

$$a_{jH}(k) = \frac{\sqrt{\lambda_j}\sqrt{\lambda_j}u_j}{\sqrt{\lambda_j}(\lambda_j + k)} = \frac{\lambda_j}{\lambda_j + k} a_{jPNS} = f_j a_{jPNS}, \qquad j = 1,...,p, \tag{3.10}$$

missä $a_{jPNS} = u_j/\lambda_j^{1/2}$ ja $f_j = \lambda_j/(\lambda_j + k)$, kun $r = p$. Harjaestimaattori $a_{jH}(k)$ on eräänlainen estimaattorin a_{jPNS} konveksi kombinaatio

$$a_{jH}(k) = w_j 0 + (1 - w_j)a_{jPNS}, \tag{3.11}$$

missä w_j on

$$w_j = \frac{k}{\lambda_j + k}. \tag{3.12}$$

(Brown, 1993). Jos $r < p$, niin $\lambda_{r+1} = ... = \lambda_p = 0$ ja täten harjaestimaattoria $a_{jH}(k)$ ei voida esittää muodossa (3.10) ja (3.11), koska estimaattoreissa a_{jPNS} jouduttaisiin jakamaan nollalla, $j = r + 1,...,p$. Kun harjaparametri $k > 0$ on tarpeeksi pieni, niin harjaregressiossa vain hyvin pienet neliömatriisin $X'X$ ominaisarvot kasvavat ja suurilla ominaisarvoilla tulee olemaan lähes sama approksimaatio $\lambda_j \approx \lambda_j + k$. Kun $k > 0$, niin $|a_{jH}(k)| < |a_{jPNS}|$, $j = 1,...,p$. Mitä pienempi on ominaisarvo λ_j, sitä pienempi on f_j ja sitä suurempi on w_j, missä f_j ja w_j ovat välillä [0,1], ja sitä enemmän tulee estimaattori a_{jPNS} kutistetuksi kohti origoa estimaatorissa $a_{jH}(k)$. Täten mitä suurempi on pienimmän neliösumman estimaattorin varianssi $\text{var}(a_{jPNS}) = 1/\lambda_j$, sitä enemmän harjaestimaattori $a_{jH}(k)$ kutistaa sitä. Niinpä harjaregressiota kutsutaan kutsitavaksi menetelmäksi (engl. shrinking method) ja harjaestimaattori $a_H(k)$ on kutistavaksi estimaattoriksi (engl. shrinking estimator) (Brown, 1993).

3.2 Harjaestimaattoreiden $b_H(k)$ ja $a_H(k)$ keskineliövirhe

Harjaregressiossa tehtävänä on määrätä harhainen estimaattori $a_H(k)$, joka on MSE-kriteerin mukaan parempi estimaattori kuin harhaton estimaattori a_{PNS}, jolloin $\text{MSE}(a_H(k)) < \text{MSE}(a_{PNS})$. Harjaestimaattorin $a_{jH}(k)$ harha on

$$E(a_{jH}(k) - \alpha_j) = \frac{\lambda_j \alpha_j}{\lambda_j + k} - \alpha_j = \frac{\lambda_j \alpha_j - (\lambda_j \alpha_j + k\alpha_j)}{\lambda_j + k} = -\frac{k}{\lambda_j + k}\alpha_j \tag{3.13}$$

eli estimaattorin $a_{jH}(k)$ harha on painon w_j vastaluku yhtälössä (3.11). Harjaestimaattorin $a_{jH}(k)$ varianssi on

$$\text{var}(a_{jH}(k)) = \text{var}\left(\frac{\lambda_j a_{jPNS}}{\lambda_j + k}\right) = \frac{\lambda_j^2 \text{var}(a_{jPNS})}{(\lambda_j + k)^2} = \frac{\lambda_j^2 \sigma_y^2}{(\lambda_j + k)^2 \lambda_j} = \sigma_y^2 \frac{\lambda_j}{(\lambda_j + k)^2}, \tag{3.14}$$

missä $\text{var}(a_{jPNS}) = \sigma_y^2/\lambda_j$ (vrt. yhtälö (2.27)). Harjaestimaattorin $a_{jH}(k)$ varianssi on pienempi kuin estimaattorin a_{jPNS} varianssi, kun harjaparametri $k > 0$ ja

$$\lambda_j(\lambda_j + k)^{-2} < \lambda_j^{-1} \tag{3.15}$$

(Gunst & Mason, 1980).

Harjaestimaattoreiden $\boldsymbol{b}_H(k)$ ja $\boldsymbol{a}_H(k)$ keskineliövirhe on

$$\text{MSE}(\boldsymbol{b}_H(k)) = \sigma_y^2(\text{tr}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-1} - k\,\text{tr}(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-2}) + k^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + k\boldsymbol{I}_p)^{-2}\boldsymbol{\beta}$$

$$= \gamma_1(k) + \gamma_2(k) = \tag{3.16}$$

$$\text{MSE}(\boldsymbol{a}_H(k)) = \sigma_y^2 \sum_{j=1}^{p} \lambda_j(\lambda_j + k)^{-2} + k^2 \sum_{j=1}^{p} \alpha_j^2(\lambda_j + k)^{-2},$$

missä komponentti $\gamma_1(k)$ on estimaattoreiden $\boldsymbol{b}_H(k)$ ja $\boldsymbol{a}_H(k)$ kokonaisvarianssi ja komponentti $\gamma_2(k)$ on estimaattoreiden $\boldsymbol{b}_H(k)$ ja $\boldsymbol{a}_H(k)$ harhan neliö. Komponentti $\gamma_1(k)$ on harjaparametrin k funktiona monotonisesti jatkuva ja vähenevä funktio. Komponentti $\gamma_2(k)$ on harjaparametrin k funktiona monotonisesti jatkuva ja kasvava funktio. Komponenttien derivaatat harjaparametrin k lähestyessä nollaa ovat

$$\lim_{k \to 0+} \left(\frac{d\gamma_1}{d\,k} \right) = -2\sigma_y^2 \sum_{j=1}^{r} \left(\frac{1}{\lambda_j} \right) \tag{3.17}$$

ja

$$\lim_{k \to 0+} \left(\frac{d\gamma_2}{dk} \right) = 0 \,. \tag{3.18}$$

Kun selittävät muuttujat ovat riippumattomia ja harjaparametri k lähestyy nollaa, niin komponentilla γ_1 on negatiivinen derivaatta $-2p\sigma_y^2$. Kun selittävät muuttujat ovat multikollineaarisia ja harjaparametri k lähestyy nollaa, niin komponentin γ_1 derivaatta lähestyy negatiivista ääretöntä. Kun $k \to 0^+$, niin komponentti γ_2 on vakaa ja nolla origossa. Nämä edellä esitetyt komponenttien γ_1 ja γ_2 ominaisuudet mahdollistavat, että kun

$$0 < k < \frac{\sigma_y^2}{\max\limits_{(1 \le j \le p)} \alpha_j^2} \,, \tag{3.19}$$

niin harjaregressiolla saadaan harhainen estimaattori $\boldsymbol{a}_H(k)$, jolla on pienempi keskineliövirhe kuin harhattomalla estimaattorilla $\boldsymbol{a}_{PNS}$. (Hoerl & Kennard, 1970a & 1970b.)

Kun

$$k_j = \sigma_y^2 / \alpha_j^2 \,, \qquad j = 1,\ldots,p \,, \tag{3.20}$$

niin harjaestimaattorilla $\boldsymbol{a}_H(k)$ on pienin mahdollinen keskineliövirhe. Yleensä parametrit $\boldsymbol{\alpha}$ ja σ_y^2 ovat tuntemattomia, joten harjaparametrin k yläraja on teoriassa tuntematon. Kuitenkin käytännössä harjaparametrin k yläraja on äärellinen, sillä on aina olemassa harjaparametri k, jolla

$$\mathrm{MSE}\,(\boldsymbol{b}_H(k)) = \mathrm{MSE}\,(\boldsymbol{a}_H(k)) < \mathrm{MSE}\,(\boldsymbol{b}_{PNS}) = \mathrm{MSE}\,(\boldsymbol{a}_{PNS}) \,,$$

kun harjaparametri k on

$$0 < k < \frac{\hat{\sigma}_y^2}{\max_{(1 \le j \le p)} a_{jPNS}^2} . \tag{3.21}$$

(Hoerl & Kennard, 1970a &1970b.)

3.3 Harjaregression geometrinen tulkinta

Olkoon $\boldsymbol{a}$ mikä tahansa parametrin $\boldsymbol{\alpha}$ estimaateista ja tehtävänä on minimoida estimaattorin $\boldsymbol{a}$ neliöity pituus $\boldsymbol{a'a}$ ehdolla

$$\sum_{j=1}^{r} (a_j - a_{jPNS})^2 \lambda_j = \text{vakio} . \tag{3.22}$$

Langrangen lauseke tälle minimointitehtävälle on

$$\sum_{j=1}^{r} a_j^2 + (1/k) \sum_{j=1}^{r} (a_j - a_{jPNS})^2 \lambda_j . \tag{3.23}$$

Kun derivoidaan tämä lauseke estimaattorin a_j suhteen ja asetetaan derivoitu lauseke nollaksi, niin saadaan

$$a_j = \frac{\lambda_j a_{jPNS}}{\lambda_j + k} , \tag{3.24}$$

joka on harjaestimaattori $a_{jH}(k)$. Lisäksi

$$\sum_{j=1}^{r} a_{jH}^2(k) = \frac{\sum_{j=1}^{r} a_{jPNS}^2 \lambda_j^2}{(\lambda_j + k)^2} = f(k) , \tag{3.25}$$

missä $f(k)$ on harjaparametrin k funktiona monotoninen ja

$$f(0) = \sum_{j=1}^{r} a_{jPNS}^2 \quad \text{ja} \quad f(\infty) = 0 . \tag{3.26}$$

Harjaregression ratkaisu $\boldsymbol{a}_H(k)$ voidaan pelkistää säteeksi origosta pisteessä, jossa se kohtaa pienimmän neliösumman ratkaisun. Pienimmän neliösumman estimaattorin jäännösneliösumma on parametrin α funktiona ellipsoidi, jonka keskipiste on $\boldsymbol{a}_{PNS}$ ja pääakseleina ovat neliömatriisin $\boldsymbol{X'X}$ ominaisarvojen neliöjuuret. Tämä ellipsoidi tulee olemaan hyvin kapea multikollineaarisessa tapauksessa, koska tällöin neliömatriisin $\boldsymbol{X'X}$ pienin ominaisarvo $\lambda_p \approx 0$. Kun harjaparametri k on nolla, niin estimaattorin $\boldsymbol{a}_H(k)$ etäisyys origosta on sama kuin estimaattrorin $\boldsymbol{a}_{PNS}$ etäisyys origosta ja harjaregression jäännösneliösumman ellipsoidi ei ole pyöristynyt pienimmän neliösumman jäännösneliösumman ellipsoidista, eli harjaregressio on sama kuin pienimmän neliösumman regressio. Kun harjaparametri k lähestyy ääretöntä, niin estimaattorin $\boldsymbol{a}_H(k)$ etäisyys origosta on rajattu hyvin pieneksi, jolloin estimaattori $\boldsymbol{a}_H(k) = 0$. Koska parametri α sijaitsee säteellä $[0, \boldsymbol{a}_{PNS}]$, niin harjaregressiossa tehtävänä on harjaparametrin k avulla kutistaa estimaattoria $\boldsymbol{a}_{PNS}$ kohti origoa ja pyöristää tämän jäännösneliösumman ellipsoidia siten, että sen keskipiste, joka on nyt siis harjaestimaattori $\boldsymbol{a}_H(k)$, on mahdollisimman lähellä parametria α. Kun selittäviä muuttujia on kaksi, niin pienimmän neliösumman jäännösneliösumman muoto on harjanne, jonka mukaan harjaregressio on saanut nimensä.

3.4 Ennusteet harjaregressiossa

Oletetaan, että tuntematon y_0 voidaan kirjoittaa muotoon (2.23). Tällöin ennusteen $\hat{y}_{0H} = \boldsymbol{x}_0'\boldsymbol{b}_H(k)$ varianssi on

$$\text{var}(\hat{y}_{0H}) = \sigma_y^2(1 + 1/n + \boldsymbol{x}_0' \boldsymbol{G}_H \boldsymbol{x}_0) = \sigma_y^2(1 + 1/n + \boldsymbol{x}_0'(\boldsymbol{X'X} + k\boldsymbol{I}_p)^{-1}\boldsymbol{x}_0), \qquad (3.27)$$

missä $\boldsymbol{G}_H = (\boldsymbol{X'X} + k\boldsymbol{I}_p)^{-1}$ (Brown, 1993).

Oletetaan, että regressiomalli on määritelty kanonisen yhtälön (2.6) mukaan, ja tehtävänä on matriisilla $\Lambda_p^{-1/2}$ ennustaa tuntematonta $\boldsymbol{u}$. Tällöin enusteen $\hat{\boldsymbol{u}}_H$ keskineliövirhe on

$$\text{MSE}(\hat{\boldsymbol{u}}_H(k)) = \sigma_y^2(n + 1) + E\sum_{j=1}^{p} \lambda_j(a_{jH}(k) - \alpha_j)^2 \qquad (3.28)$$

$$= \sigma_y^2(n+1) + \sum_{j=1}^{p} \lambda_j(\mathrm{var}(a_{jH}(k)) + B^2(a_{jH}(k)))$$

$$= \sigma_y^2(n+1) + \sigma_y^2\sum_{j=1}^{p} \lambda_j^2/(\lambda_j+k)^2 + k^2\sum_{j=1}^{p} \lambda_j\alpha_j^2/(\lambda_j+k)^2 ,$$

missä $B^2(a_{jH}(k))$ on estimaattorin $a_{jH}(k)$ harhan neliö. Wang ja Chow (1994) ovat osoittaneet, että aina on olemassa harjaparametrin k arvo, jolla $GMSE(\boldsymbol{a}_H(k)) < GMSE(\boldsymbol{a}_{PNS})$, kun

$$0 < k < 2\sigma_y^2/\alpha'\alpha \ . \tag{3.29}$$

Täten $MSE(\boldsymbol{a}_H(k)) < MSE(\boldsymbol{a}_{PNS})$ ja $MSE(\hat{\boldsymbol{u}}_H) < MSE(\hat{\boldsymbol{u}}_{PNS})$, kun harjaparametrille $\boldsymbol{k}$ on voimassa yllä oleva epäyhtälö.

3.5 Harjaparametrin k valintamenetelmiä yksimuuttujaisessa tapauksessa

Kun parametrit α ja σ_y^2 ovat tuntemattomia, niin ei ole mahdollista määrätä selvää automaattista estimointimenetelmää harjaparametrille k. Seuraavassa on esitetty muutamia menetelmiä harjaparametrin k valinnalle.

1) Harjajälki: Piirretään harjaestimaattori $\boldsymbol{a}_H(k)$ harjaparametrin $k \geq 0$ funktiona. Kyseistä kuvaa kutsutaan harjajäljeksi (engl. ridge trace). Piirretään kuvaan myös sovitteen jäännösneliösumma harjaparametrin k funktiona. Kuvan perusteella valitaan se harjaparametrin k arvo, jolla estimaattoreiden $a_{jH}(k)$, $j = 1,...,r$, arvot näyttävät vakiintuneen ja sovitteen jäännösneliösumma ei ole kasvanut kohtuuttoman suureksi. Jos matriisin X aste r on pienempi kuin selittävien muuttujien lukumäärä p, niin $\lambda_{r+1} = ... = \lambda_p = 0$, jolloin estimaattoria $a_{jH}(k)$, missä $k = 0$, ei ole määritelty, $j = r + 1 ,...,p$. Kun harjaparametri k määrätään graafisesti, niin sopivan harjaparamtrin k arvon määrääminen perustuu tutkijan subjektiiviseen käsitykseen sopivasta harjaparametrin k arvosta, mikä vaatii tämän menetelmän huolellista ja ammattitaitoista käyttöä. (Hoerl & Kennard, 1970a & 1970b.)

2) *Cross-validation*: Kullakin annetulla harjaparametrin k arvolla jätetään harjaestimaattorin $\boldsymbol{b}_H(k)$ laskennassa vuorollaan kukin havainto $i = 1,...,n$ pois ja lasketaan jokaiselle poistetulle havainnolle i poistetun havainnon $y_{(i)}$ ennuste. Lasketaan lopuksi

$$\text{PRESS}(k) = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 , \tag{3.30}$$

missä siis ennuste $\hat{y}_{(i)}$ on laskettu harjaregressiolla ilman havaittua havainto i. Valitaan se harjaparametrin k estimaattori k_{cross}, jolla ennusteen virheneliösumma PRESS(k) on pienin. (Brown, 1993.)

3) *Generalized cross-validition:* Valitaan harjaparametrin k estimaattoriksi k_{gcross}, joka minimoi lausekkeen

$$\|(I - H(k))y\|^2 / (\text{tr}(I - H(k)))^2 , \tag{3.31}$$

missä $H(k) = X(X'X + kI_p)^{-1}X'$ (Brown, 1993).

4) Hoerlin, Kennardin ja Baldwin säännön mukaan:

$$k_{HKB} = (r - 2)\hat{\sigma}_y^2 / a'_{PNS} a_{PNS} . \tag{3.32}$$

Tämä harjaparametrin k valintamentelmä k_{HKB} perustuu epäyhtälöön (3.19), missä parametreja σ_y^2 ja α on estimoitu niiden pienimmän neliösuman estimaattoreilla. Estimaattori k_{HKB} saattaa olla liian pieni, kun selittävien muuttujien x_j, $j = 1,...,p$, kollineaarisuuden indeksi (2.43) on suuri ja neliömatriisin $X'X$ pienimpiä ominaisarvoja vastaavat neliömatriisin $X'X$ ominaisvektorit ovat korreloimattomia havaintovektorin y kanssa. Jos

$$h_p^2(p + 2) < 2p\sum_{j=1}^{p} h_j^2 / p , \tag{3.34}$$

missä $h_j = \lambda_j^{-1}$, niin MSE($a_H(k_{HKB})$) < MSE(a_{PNS}) ja jos

$$h_p(p + 2) < 2p\sum_{j=1}^{p} h_j / p , \tag{3.35}$$

niin MSE($\hat{a}_H(k_{HKB})$) < MSE($\hat{a}_{PNS}$). (Brown, 1993.)

5) Brown (1993):

$$k_{MLW} = (r - 2)\hat{\sigma}_y^2 \text{tr}(X'X) / (r b'_{PNS} X'X b_{PNS}) . \tag{3.36}$$

Kun selittävien muuttujien x_j, $j = 1,...,p$, kollineaarisuudenindeksi on suuri ja neliömatriisin $X'X$ pienimpiä ominaisarvoja vastaavat neliömatriisin $X'X$ ominaisvektorit ovat korreloimattomia havaintovektorin y kanssa, niin k_{MLW} voidaan pitää luotettavampana kuin k_{HKB}. Jos ominaisarvo λ_j on pieni, niin estimaattoria a_{jPNS}, joka voi olla itseisarvoltaan kohtuuttoman suuri, painotetaan pienellä ominaisarvolla λ_j harjaestimaattorissa $a_H(k_{MLW})$. (Brown, 1993.)

6) Brown (1993): Lauseke

$$\hat{\sigma}_y^2 \sum_{j=1}^{p} (\lambda_j - k)/(\lambda_j(\lambda_j + k)) + k^2 \sum_{j=1}^{p} a_{jPNS}^2/(\lambda_j + k)^2 \qquad (3.37)$$

eli keskineliövirheen harhaton estimaattori minimoidaan harjaparametrin k suhteen, mistä saadaan harjaparametrin k valinta k_{MUR}.

7) Wang ja Chow (1994): Yleistetyyn harjaestimaattoriin $a_H(k)$ harjaparametrivektori k voidaan valinta

$$k_{j\,ad} = \hat{\sigma}_y^2/a_{jPNS}^2 \qquad j = 1,...,r\ . \qquad (3.38)$$

Tämä harjaparametrivektorin k_{ad} valintamenetelmä perustuu epäyhtälöön (3.20), missä parametreja σ_y^2 ja α_j on estimoitu niiden pienimmän neliösumman estimaattoreilla. Jos tällä harjaparametrin k valintamenetelmällä halutaan valita harjaestimaattorin $b_H(k)$ harjaparametrivektori k, niin korvataan tässä valintamenetelmässä estimaattori a_{jPNS} estimaattorilla b_{jPNS}.

3.6 Monimuuttujainen harjaregressio

Olkoon monimuuttujainen regressiomalli määritelty yhtälön (2.5) mukaan. Kirjoitetaan regressiomatriisi B vektorimuotoon β^*

$$\beta^* = (\beta_1^*,....,\beta_p^*)'\ , \qquad (3.39)$$

missä β^*_j on $(1 \times q)$-vektori ja se on regressioparametrimatriisin B j:nnes rivivektori. Brownin ja Zidekin (1980) ehdottama parametrin β^* harjaestimaattori $b_H^*(K)$ on

$$b^*(K) = (X'X \otimes I_q + I_p \otimes K)^{-1}(X'X \otimes I_q)b^*_{PNS}\,, \tag{3.40}$$

missä $\otimes$ on Kroneckerin tulo, $K > 0$ on $(q \times q)$ -harjaparametrimatriisi, b^*_{PNS} on parametrin β^* pienimmän neliösumman estimaattori. Kun regressiomalli on määritelty kanonisen yhtälön (2.6) mukaan, niin monimuuttujaisen harjaestimaattorin $b^*_H(K)$ kanoninen muoto $a^*_H(K)$ on

$$a^*(K) = (\Lambda_p \otimes I_q + I_p \otimes K)^{-1}(\Lambda_p \otimes I_q)a^*_{PNS}\,, \tag{3.41}$$

missä a^*_{PNS} on estimaattorin b^*_{PNS} kanoninen muoto. Harjaestimaattori $a^*_H(K)$ voidaan esittää myös muodossa

$$a^*_{jH}(K) = a^*_{jPNS}\lambda_j(\lambda_j I_q + K)^{-1}\,, \quad j = 1,\ldots,p\,. \tag{3.42}$$

Kuten yksimuuttujainen harjaestimaattori $a_{jH}(k)$, niin myös monimuuttujainen harjaestimaattori $a^*_{jH}(K)$ on painotettu a^*_{jPNS}

$$a^*_{jH}(K) = W_j + (I_q - W_j(K))a^*_{jPNS}\,, \quad j = 1,\ldots,p\,, \tag{3.43}$$

missä W_j on

$$W_j(K) = K(\lambda_j I_q + K)^{-1}\,. \tag{3.44}$$

Monimuuttujaisessa harjaregressiossa kutistetaan estimaattoria a^*_{PNS} kohti origoa ja sitä enemmän estimaattori a^*_{jPNS} kutistuu, mitä suurempi on paino $W_j(K)$ eli mitä pienempi on neliömatriisin $X'X$ ominaisarvo λ_j. Harjaestimaattorin $a_H^*(K)$, missä $K = kI_q$, keskineliövirhe on

$$\mathrm{MSE}(a^*_H) = \mathrm{tr}(\Sigma_Y)\sum_{j=1}^{p}\lambda_j/(\lambda_j + k)^2 + \sum_{j=1}^{p}k^2/(\lambda_j + k)^2\sum_{i=1}^{q}\alpha_j^{(i)2}\,, \tag{3.45}$$

missä $\Sigma_Y = (\sigma_Y^{(ij)})$, $i,j = 1,\ldots,q$. Kun yksimuuttujaista harjaestimaattoria $a_H(k)$ ja sen keskivirhettä $\mathrm{MSE}(a_H(k))$ verrataan monimuuttujaisen harjaestimaattoriin $a_H^*(K)$ ja sen keskivirheeseen $\mathrm{MSE}(a_H^*(K))$, niin estimaattoreilla $a^*_H(K)$ ja $a_H(k)$ ja täten myös keskivirheillä $\mathrm{MSE}(a_H^*(K))$ ja

MSE($a_H(k)$) ovat samat matemaattiset ominaisuudet. Siis kun $q = 1$ ja $\boldsymbol{K} = k$, niin $a_H^*(\boldsymbol{K}) = a_H(k)$ ja MSE($a_H^*(\boldsymbol{K})$) = MSE($a_H(k)$). (Brown & Payne, 1975; Broewn & Zidek, 1980.)

3.7 Harjaparametrimatriisin $\boldsymbol{K}$ valinta monimuuttujaisessa tapauksessa

Parametrin B monimuuttujainen harjaestimaattori voidaan aina hajottaa q yksimuuttujaiseen harjaestimaattoriin

$$b_H^{(j)}(k^{(j)}) = (X'X + k^{(j)})^{-1}X'y^{(j)}, \qquad j = 1,...,q, \tag{3.46}$$

missä harjaparametrin $k^{(j)}$ ratkaisu riippuu ainoastaan yhdestä selitettävästä muuttujasta $y^{(j)}$. Tällöin harjaparametrit $k^{(j)}, j = 1,...,q$, voidaan ratkaista kappaleessa 3.5 esitettyjen harjaparametrin k valintamenetelmien mukaan. Tämä saattaa kuitenkin olla epätyydyttävä monimuuttujaisesa tapauksessa, koska siinä ei hyödynnetä kovarianssimatriisia Σ_Y eli kaikkien q normaaliyhtälöiden yhteisinformaatiota.

Monimuuttujaisessa harjaregressiossa voidaan hyödyntää kovarianssimatriisia Σ_Y valittaessa harjaparametrimatriisia $\boldsymbol{K}$. Seuraavassa on eri tapoja valita harjaparametrimatriisi $\boldsymbol{K}$ monimuuttujisessa harjaregressiossa.

3.7.1 Harjaparametrimatriisin $\boldsymbol{K}$ valinta Brownin ja Paynen (1975) mukaan

Oletetaan, että regressiomallina on malli (2.6), missä Σ_Y on tunnettu ja $\alpha_i^{(j)}$ noudattaa jakaumaa $N(0,\sigma_\alpha^2)$, $i = 1,...,p, j = 1,...,q$. Kun harjaparametrimatriisi $\boldsymbol{K}$ on valittu Brownin ja Paynen (1975) mukaan, niin harjaestimaattori $A_H(\boldsymbol{K}_{BP})$ on

$$a_{iH}^{(j)}(k_{BP}^{(j)}) = \frac{\lambda_i a_{iPNS}^{(j)}}{\lambda_i + h^{(j)}\sigma_\alpha^2} = \frac{\lambda_i a_{iPNS}^{(j)}}{\lambda_i + k_{BP}^{(j)}}, \qquad i = 1,....,p, \quad j = 1,...,q, \tag{3.47}$$

missä $h^{(j)}$ on kovarianssimatriisin Σ_Y ominaisarvo ja harjaparametri $k^{(j)} = h^{(j)}\sigma_\alpha^2, j = 1,...,q$,. Jos ominaisarvo λ_i on pieni, niin erityisesti tällöin estimaattori $a_{iH}^{(j)}(k^{(j)})$ on kutistanut enemmän estimaattoria $a^{(j)}_{iPNS}$, missä kovarianssimatriisin Σ_Y ominaisarvo $h^{(j)}$ on suuri, kohti origoa kuin estimaattoria $a^{(k)}_{iPNS}$, missä kovarianssimatriisin Σ_Y ominaisarvo $h^{(k)}$ on pieni. Kun kovarianssimatriisi Σ_Y on diagonaalinen, niin tällöin kovarianssimatriisilla Σ_Y on vain yksi nollasta eroava

ominaisarvo ja $\boldsymbol{K} = k\boldsymbol{I}_q$, missä harjaparametri k on

$$k = \frac{h}{\sigma_\beta^2} \tag{3.48}$$

ja h on ainoa matriisin Σ_Y nollasta eroava ominaisarvo. Jos oletetaan, että matriisi Σ_Y on diagonaalinen, niin tällöin monimuuttujaisessa harjaregressiossa ei hyödynnetä q normaaliyhtälöiden yhteisinformaatiota. (Brown & Payne, 1975.)

3.7.2 Harjaparametrimatriisin $\boldsymbol{K}$ valinta Brownin ja Zidekin (1982) mukaan

Oletetaan, että regressiomalli on määritelty kanonisella yhtälöllä (2.6), missä $r = p$, $\mathrm{E}(\epsilon^{(j)}) = 0$, $\mathrm{cov}(\epsilon^{(j)},\epsilon^{(i)}) = \sigma_Y^{(ij)}\boldsymbol{I}_n$, $i, j = 1,\ldots,q$. Kun parametri $\boldsymbol{A}$ noudattaa jakaumaa $\mathbb{N}(0,\Sigma_\alpha)$, niin yleinen muoto harjaparametrimatriisille $\boldsymbol{K}$ on

$$\boldsymbol{K} = \Sigma_Y \Sigma_\alpha^{-1} . \tag{3.49}$$

Jos parametrit Σ_Y ja Σ_α ovat tuntemattomia, niin käytetään estimaattoreita

$$\hat{\Sigma}_Y^{-1} = \boldsymbol{R} \quad \text{ja} \quad \hat{\Sigma}_\alpha = \sum_{j=1}^{p} w_j \boldsymbol{a}_{jPNS}^{*\prime} \boldsymbol{a}_{jPNS}^{*} , \tag{3.50}$$

jossa $\boldsymbol{R}^{-1}$ on matriisin $\boldsymbol{\Sigma}_Y$ suurimman uskottavuuden estimaattori tai pienimmän neliösumman estimaattori.

Merkitään $v_j = \lambda_j^{-1}$, $\lambda_j, j = 1,\ldots,p$, ovat matriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvoja, jotka ovat suuruusjärjestyksessä $\lambda_1 \geq \ldots \geq \lambda_p, f = p - q - 1$,

$$\bar{v}^{-1} = \sum_{j=1}^{p} v_j^{-1}/p \quad \text{ja} \quad \bar{v}^{2} = \sum_{j=1}^{p} v_j^{2}/p . \tag{3.51}$$

Kun harjaparametrimatriisi $\boldsymbol{K}$ valitaan Scloven säännön mukaan, niin monimuuttujainen harjaestimaattori (3.42) on

$$\boldsymbol{a}_{jH}^{*}(\boldsymbol{K}_S) = \boldsymbol{a}_{jPNS}^{*}(\boldsymbol{I}_q - v_j f \bar{v}^{-1}(v_j f \bar{v}^{-1}\boldsymbol{I}_q + \boldsymbol{R}\sum_{i=1}^{p} v_i^{-1}\boldsymbol{a}_{iPNS}^{*\prime}\boldsymbol{a}_{iPNS}^{*})^{-1}) , \tag{3.52}$$

$j = 1,\ldots,p$. Monimuuttujaisella harjaestimaattorilla $\boldsymbol{a}^*_H(\boldsymbol{K}_S)$ on pienempi keskineliövirhe kuin pienimmän neliösumman estimaattorilla $\boldsymbol{a}^*_{PNS}$, kun

$$(n - p)(p - q - 1)\bar{v}^{-1}v_p^3 - 2(n - p - 2(q + 1))(p\bar{v}^2 - (q + 1)v_p^2) < 0 \qquad (3.53)$$

ja

$$n - p - 2(q + 1) > 0. \qquad (3.54)$$

Kun harjaparametrimatriisi $\boldsymbol{K}$ valitaan Hoerlin, Kennardin ja Baldwinin säännön mukaan, niin monimuuttujainen harjaestimaattori (3.42) on muotoa

$$\boldsymbol{a}^*_{jH}(\boldsymbol{K}_{HKB}) = \boldsymbol{a}^*_{jPNS}(\boldsymbol{I}_q - fv_j(fv_j\boldsymbol{I}_q + \boldsymbol{R}\textstyle\sum_{i=1}^{p}\boldsymbol{a}^*_{iPNS}{}'\boldsymbol{a}^*_{iPNS})^{-1}), \qquad (3.55)$$

$j = 1,\ldots,p$. Monimuuttujaisella harjaestimaattorilla $\boldsymbol{a}^*_H(\boldsymbol{K}_{HKB})$ on pienempi keskineliövirhe kuin pienimmän neliösumman estimaattorilla $\boldsymbol{a}^*_{PNS}$, kun

$$n - p - 2q - 2 > 0 \qquad (3.56)$$

ja

$$(n - p)(p - q - 1)v_p^2 - 2(n - p - 2q - 2)(p\bar{v}^2 - (q + 1)v_p^2) < 0\ . \qquad (3.57)$$

3.7.3 Harjaparametrimatriisin K valinta Frankin ja Friedmanin (1993) mukaan

Kun harjaparametrimatriisi $\boldsymbol{K}$ valitaan Frankin ja Friedmanin (1993) mukaan, niin kanonisen monimuuttujaisen regressiomallin (2.12) harjaestimaattori on

$$a^{(j)}_{iH}(k^{(j)}_{FF}) = a^{(j)}_{iPNS}\lambda_i/(\lambda_i + k^{(j)}_{FF})\ , \qquad i = 1,\ldots,p \quad \text{ja} \quad j = 1,\ldots,q\ , \qquad (3.58)$$

missä λ_i, $i = 1,\ldots,p$, ovat neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvoja ja $k_{FF}{}^{(j)}$ on

$$k_{FF}^{(j)} = p\,\sigma_Y^2 / n\,(h^{(j)} - \sigma_Y^2)\,, \tag{3.59}$$

$h^{(j)}$, $j = 1,\ldots,q$, ovat neliömatriisin $\boldsymbol{Y'Y}$ ominaisarvoja, n on otoskoko ja $\boldsymbol{\Sigma}_Y = \sigma^2{}_Y\boldsymbol{I}_q$. Jos $\sigma_y{}^2$ on tuntematon, niin käytetään tässä sen estimaattoria

$$\hat{\sigma}_Y^2 = \frac{\sum_{j=1}^{q} (y^{(j)} - b_{PNS}^{(j)})'(y^{(j)} - b_{PNS}^{(j)})}{q(n-p)}\,. \tag{3.60}$$

Mitä pienemmät ovat ominaisarvot λ_i ja $h^{(j)}$ sitä enemmän kutistetaan estimaattoria $a_i^{(j)}{}_{PNS}$ harjaestimaattorissa $a_i^{(j)}{}_H(k_{FF}^{(j)})$ kohti origoa. (Frank & Friedman, 1993.)

3.7.4 Harjaparametrimatriisin $\boldsymbol{K}$ valinta *cross-validation* -menetelmällä

Sovelletaan yksimuuttujaisen harjaestimaattorin $\boldsymbol{a}_H(k)$ harjaparametrin k *cross-validation* valintamenetelmää monimuuttujaisen harjaestimaattorin $\boldsymbol{a}^*{}_H(\boldsymbol{K})$ harjaparametrimatriisin $\boldsymbol{K}$ valintamenetelmäksi. Oletetaan, että $\boldsymbol{\Sigma}_Y = \sigma_Y{}^2\boldsymbol{I}_q$, jolloin $\boldsymbol{K} = k\boldsymbol{I}_q$, missä $k \geq 0$. Lasketaan ennusteen virhevarianssi

$$\text{PRESS}\,(k) = \sum_{i=1}^{n} \sum_{j=1}^{q} (u_i^{(j)} - \hat{u}_{(i)}^{(j)})^2\,, \tag{3.61}$$

missä $\hat{\boldsymbol{u}}_{(i)}^{(j)}$ on ennuste, joka on saatu harjaregressiolla ilman havaintoa i, $i = 1,\ldots,n$. Valitaan se harjaparametrin k estimaattori k_{cross}, jolla PRESS(k) = PRESS(k_{cross}) on pienin.

3.7.5 Harjaparametrimatriisin $\boldsymbol{K}$ valinta Fülen (1995) mukaan

Kirjoitetaan regressiomalli (2.5) tässä muotoon

$$\text{vek}(\boldsymbol{Y}) = (\boldsymbol{I}_q \otimes \boldsymbol{X})\text{vek}(\mathbf{B}) + \text{vek}(\mathbf{E})\,, \tag{3.62}$$

missä $\otimes$ on Kroneckerin tulo. Fülen (1995) mukaan parametrin vek(B) harjaestimaattori on

$$\begin{aligned}\text{vek}(\boldsymbol{B}(\boldsymbol{K})) &= (\boldsymbol{D'D} + \boldsymbol{K})^{-1}(\boldsymbol{D'}\text{vek}(\boldsymbol{Y}) + \boldsymbol{K}\text{vek}(\boldsymbol{B}_0)) \\ &= (\boldsymbol{D'D} + \boldsymbol{K})^{-1}(\boldsymbol{D'D}\,\text{vek}(\boldsymbol{B}_{PNS}) + \boldsymbol{K}\,\text{vek}(\boldsymbol{B}_0))\,,\end{aligned} \tag{3.63}$$

missä vek($\boldsymbol{B}_0$) on valittu ($pq \times 1$) -vektori, $\boldsymbol{K}$ on ($pq \times pq$) -harjaparametrimatriisi ja $\boldsymbol{D} = \boldsymbol{I}_q \otimes \boldsymbol{X}$. Tämä harjaestimaattori vek($\boldsymbol{B}_H(\boldsymbol{K})$) kutistaa pienimmän neliösumman estimaattoria vek($\boldsymbol{B}_{PNS}$) kohti valittua vektoria vek($\boldsymbol{B}_0$), joka on oletettu estimoitavan parametrin vek(B) todellinen arvo. Koska estimoitava parametri vek(B) on välillä [0,vek($\boldsymbol{B}_{PNS}$)], niin

$$\text{vek}(B_{j0}) = c\,\text{vek}(B_{jPNS}), \quad j = 1,\ldots,pq, \tag{3.64}$$

missä $0 < c < 1$ eli täten valitaan vek($\boldsymbol{B}_0$) väliltä [0,vek($\boldsymbol{B}_{PNS}$)]. Fülen (1995) ehdottama harjaparametrimatriisin $\boldsymbol{K}$ valinta on $\boldsymbol{K}_F = \text{diag}(k_{jF})$, missä

$$k_{jF} = \frac{pq\hat{\sigma}_Y^2}{(\text{vek}(B_{jPNS}) - \text{vek}(B_{j0}))^2}, \quad j = 1,\ldots,pq, \tag{3.65}$$

ja $\hat{\sigma}_Y$ on (3.60). Täten valittuun $\boldsymbol{K}_F$ vaikuttaa valittu vek($\boldsymbol{B}_0$).

Tässä harjaparametrimatriisin $\boldsymbol{K}$ valintamenetelmässä on ongelmana valita sopiva vek($\boldsymbol{B}_0$). Koska tehtävänä oli muodostaa harhainen estimmaatori vek($\boldsymbol{B}_H(\boldsymbol{K}_{HMSE})$), jonka keskineliövirhe on pienempi kuin estimaattorin vek($\boldsymbol{B}_{PNS}$) keskineliövirhe, niin nollahypoteesina on

$$\text{MSE}(\text{vek}(\boldsymbol{B}_H(\boldsymbol{K}_F))) < \text{MSE}(\text{vek}(\boldsymbol{B}_{PNS}))$$

ja sitä voidaan testata testisuureella

$$F = \frac{(\text{vek}(\boldsymbol{B}_{PNS}) - (\boldsymbol{B}_0))'((\boldsymbol{D}'\boldsymbol{D})^{-1} + \boldsymbol{K}_F)^{-1}(\text{vek}(\boldsymbol{B}_{PNS}) - \text{vek}(\boldsymbol{B}_0))/v_1}{(\text{vek}(\boldsymbol{Y}) - \boldsymbol{D}\,\text{vek}(\boldsymbol{B}_{PNS}))'(\text{vek}(\boldsymbol{Y}) - \boldsymbol{D}\,\text{vek}(\boldsymbol{B}_{PNS}))/v_2}, \tag{3.66}$$

joka noudattaa F-jakumaa vapausastein $v_1 = qp$ ja $v_2 = q(n - p)$. (Füle, 1995.)

4 Pääkomponenttiregressio

Neliömatriisin $X'X$ käänteismatriisi on

$$(X'X)^{-1} = (V_p \Lambda_p V_p')^{-1} = \sum_{j \in S_p} (1/\lambda_j) v_j v_j' , \qquad (4.1)$$

missä $\Lambda_p = \mathrm{diag}(\lambda_1,...,\lambda_p)$, $V_p = (v_1,...,v_p)$, $S_p = \{1,...,p\}$, λ_j, $j \in S_p$, ovat neliömatriisin $X'X$ ominaisarvot ja v_j, $j \in S_p$, ovat vastaavat ominaisvektorit. Pääkomponenttiregressiossa (engl. principal components regression) tätä approksimoidaan matriisilla

$$G_{PK} = (V_A \Lambda_A V_A')^{-1} = \sum_{j \in S_A} (1/\lambda_j) v_j v_j' , \qquad (4.2)$$

missä $\Lambda_A = \mathrm{diag}(\lambda_j)$, $V = (v_j)$, $j \in S_A$, S_A on valittujen pääkomponenttien indeksijoukko, A on valittujen pääkompenttein lukumäärä ja $S_A \in S_p$ (vrt. (3.1)) (Brown, 1993). Kun selittävien muuttujien multikollineaarisuuden aste on korkea, $\lambda_j \neq 0$ ja $j \in S_A$, niin erityisesti tällöin matriisi $V_A \Lambda_A V_A'$ on paremmin kääntyvä kuin matriisimatriisi $V_p \Lambda_p V_p'$.

Kirjoitetaan regressiomalli (2.5), missä selitettäviä muuttujia on yksi eli $q = 1$, kanoniseen muotoon

$$y = X\beta + \epsilon = XVV'\beta + \epsilon = Z\alpha + \epsilon = \alpha_1 z_1 + ... + \alpha_p z_p + \epsilon , \qquad (4.3)$$

missä $V'\beta = \alpha = (\alpha_1,...,\alpha_p)'$, $XV = Z = (z_1,...,z_p)$, matriisi $V = (v_1,...,v_p)$ on ortogonalinen ja v_j, $j \in S_p$, ovat neliömatriisin $X'X$ ominaisvektorit, joiden pituus on skaalattu ykköseksi. Vektoreita $z_1,...,z_p$ kutsutaan rakennematriisin X pääkomponenteiksi (engl. principal components) ja ne ovat selittäviä muuttujia pääkomponenttiregressiossa (Jackson, 1991).

Pääkomponenttiregressiossa tehdään siis rakennematriisille X ortogonaalinen muunnos $XV = Z = (z_1,...,z_p)$. Pääkomponentit z_j, $j = 1,...,p$, ovat selittävien muuttujien x_j, $j = 1,...,p$, lineaarikombinaatiota

$$z_j = v_{1j} x_1 + v_{2j} x_2 + ... + v_{pj} x_p , \qquad j = 1,...,p . \qquad (4.4)$$

Niiden kovarianssimatriisi $\mathrm{cov}(Z,Z)$ on

$$\text{cov}(\boldsymbol{Z}'\boldsymbol{Z}) = E((\boldsymbol{XV})'(\boldsymbol{XV})) = \boldsymbol{V}'E(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{V} = \boldsymbol{V}'\boldsymbol{\Sigma}_{\boldsymbol{X}}\boldsymbol{V} = \boldsymbol{\Lambda}_p = \text{diag}(\lambda_1,\ldots,\lambda_p)\,, \tag{4.5}$$

missä siis $\text{cov}(z_i,z_j) = 0$, kun $i \neq j$, ja $\text{var}(z_j) = \lambda_j$, $i,j = 1,\ldots,p$, eli pääkomponentit z_j, $j = 1,\ldots,p$, ovat korreloimattomia. Pääkomponentit z_j, $j = 1,\ldots,p$, selittävät rakennematriisin $\boldsymbol{X}$ kokonaisvaihtelun, sillä

$$\text{var}(\boldsymbol{Z}) = \text{tr}(\boldsymbol{Z}'\boldsymbol{Z}) = \sum_{j=1}^{p} \lambda_j = r = \text{tr}(\boldsymbol{X}'\boldsymbol{X}) = \text{var}(\boldsymbol{X})\,, \tag{4.6}$$

missä r on rakennematriisin $\boldsymbol{X}$ aste. Nämä kaikki edellä esitetyt pääkomponenttien z_j, $j = 1,\ldots,p$, ominaisuudet ovat seurausta ortogonaalisesta muunnoksesta $\boldsymbol{XV} = \boldsymbol{Z}$.

Pääkomponenttiregressiossa selitettävä muuttuja $\boldsymbol{y}$ regressoidaan pääkomponentteja z_j, $j \in S_A$, vasten pienimmän neliösumman menetelmällä. Koska matriisi $\boldsymbol{Z}$ on ortogonaalinen, niin neliömatriisi $\boldsymbol{Z}_A'\boldsymbol{Z}_A$ on diagonaalinen ja sen käänteismatriisi on $(\boldsymbol{Z}_A'\boldsymbol{Z}_A)^{-1} = \text{diag}(1/\lambda_j)$, $j \in S_A$, joka ei ole yhtä herkkä mahdollisille aineiston pyöristys- ja mittausvirheelle kuin käänteismatriisi $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. Koska pääkomponentit z_j, $j = 1,\ldots,p$, ovat korreloimattomia, niin pääkomponentteja z_j, $j \in S_A$, , vastaavat regressiokertoimet ovat korreloimattomia (Jackson, 1991).

Pääkomponenttiregressiossa valitaan kaikista rakennematriisin $\boldsymbol{X}$ pääkomponenteista z_j, $j \in S_p$, sopiva osajoukko z_j, $j \in S_A$, selittäviksi muuttujiksi pääkomponenttiregressiomalliin. Usein pääkomponenttiregressiomalliin valitaan selittäviksi muuttujiksi pääkomponentit z_j, $j \in S_A$, jotka selittävät suurimman osan rakennematriisin $\boldsymbol{X}$ vaihtelusta, jolloin pääkomponenttiregressiossa joukon S_A valinta perustuu pelkästään rakennematriisin $\boldsymbol{X}$ pääkomponenttianalyysiin. Tällöin yritetään selittää mahdollisimman paljon rakennematriisin $\boldsymbol{X}$ kokonaisvaihtelusta mahdollisimman pienellä selittävien pääkomponenttien lukumäärällä, jolloin indeksijoukkoon S_A eivät tule valituksi pääkomponentit, joiden $\text{var}(z_j) = \lambda_j \approx 0$. Indeksijoukko S_A voidaan valita myös siten, että valitaan pääkomponentit z_j, $j \in S_A$, jotka korreloivat voimakkaasti selitettävän muuttujan $\boldsymbol{y}$ kanssa. Täten pääkomponenttiregressiossa samassa tapauksessa indeksijoukon S_A alkiot saattavat vaihdella sen mukaan, mitä valintamenetelmää on käytetty joukkon S_A määräämiseen ja mikä on tämän valintamenetelmän tilastollinen peruste. (Brown, 1991.)

Kun joukko S_A on valittu, niin pääkomponenttiregressiossa rakennematriisin $\boldsymbol{X}$ approksimaatio on matriisi $\boldsymbol{Z}_A\boldsymbol{V}_A'$, missä $\boldsymbol{Z}_A = (z_j)$, $\boldsymbol{V}_A = (v_j)$ ja $j \in S_A$, jolloin neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ approksimaatio on matriisi (4.2). Pääkomponenttiregressiossa rakennematriisia $\boldsymbol{X}$ voidaan approksimoida myös singulaariarvohajotelmalla (2.10), missä nyt $\boldsymbol{T}_r = (t_j)$, $\boldsymbol{V}_r = (v_j)$, $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_j)$, $j \in S_A$, jolloin neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ approksimaatio on matriisi (2.10). Pääkomponenttiregressiossa

rakennematriisin X approksimaatioden $Z_A V_A'$ ja $T_A \Lambda_A V_A'$ aste on $A \leq p$.

4.1 Pääkomponenttiestimaattori a_{APK}

Oletetaan, että regressiomalli on määritelty yhtälön (4.3) mukaan. Oletetaan nyt, että lopulliseen pääkomponenttiregressiomalliin on valittu selittäviksi muuttujiksi pääkomponentit z_j, missä $j \in S_A$ ja $S_A \in S_p$. Kun käytetään merkintätapaa $Z_A = (z_j)$, $V_A = (v_j)$ ja $\Lambda_A = \mathrm{diag}(\lambda_j)$, $j \in S_A$, niin pääkomponenttiregressiossa estimoidaan regressiomallia (4.3) malilla $y = Z_A a_{APK} + e$, missä pääkomponenttiestimaattori $a_{APK} = (a_{jAPK})$, $j \in S_A$, on normaaliyhtälöiden

$$(Z_A' Z_A)^{-1} a_{APK} = Z_A' y \tag{4.7}$$

ratkaisu, joka on

$$a_{APK} = (Z_A' Z_A)^{-1} Z_A' y = (V_A' X' X V_A)^{-1} V_A' X' y = (V_A' V_p \Lambda_p V_p' V_A)^{-1} V_A' X' y$$

$$= \left((I_A \; 0) \Lambda_p (I_A \; 0)'\right)^{-1} V_A' X' y = \Lambda_A^{-1} V_A' X' y = \Lambda_A^{-1} Z_A' y \; . \tag{4.8}$$

Kun regressiomalli on esitetty kanonisessa muodossa (2.6), missä $q = 1$, niin pääkomponenttiestimaattori a_{APK} on

$$a_{jAPK} = u_j / \sqrt{\lambda_j} \; , \quad j \in S_A \; , \qquad \text{ja} \qquad a_{jAPK} = 0 \; , \quad j \notin S_A \; . \tag{4.9}$$

Pääkomponenttiestimaattorin a_{APK} ja pienimmän neliösumman estimaattorin a_{PNS} välisestä yhteydestä saadaan pääkomponenttiestimaattorille a_{APK} vaihtoehtoinen esitys

$$a_{jAPK} = f_j a_{jPNS} \; ,$$

missä $f_j = 1$, kun $j \in S_A$, ja $f_j = 0$, kun $j \notin S_A$, eli

$$a_{jAPK} = a_{jPNS} \; , \quad j \in S_A \; , \qquad \text{ja} \qquad a_{jAPK} = 0 \; , \quad j \notin S_A \tag{4.10}$$

(vrt. (3.10)). Pääkomponenttiestimaattori a_{APK} ja harjaestimaattori $a_H(k)$ ovat estimaattorin a_{PNS} eräänlaisia konveksikombinaatiota. Pääkomponenttiestimaattori a_{APK} on

$$a_{jAPK} = w_j 0 + (1 - w_j) a_{jPNS} , \qquad j = 1,\ldots,p , \tag{4.11}$$

missä $w_j = 0$, kun $j \in S_A$, ja $w_j = 1$, kun $j \notin S_A$ (vrt. (3.11)). Jos $S_A \neq S_p$, niin pääkomponenttiestimaattorin $\boldsymbol{a}_{APK}$ etäisyys origosta on pienempi kuin estimaattorin $\boldsymbol{a}_{PNS}$ etäisyys origosta ts.

$$\|\boldsymbol{a}_{APK}\| < \|\boldsymbol{a}_{PNS}\| . \tag{4.12}$$

Tämän vuoksi pääkomponenttiregressiota kutsutaan kutistavaksi menetelmäksi (engl. shrinking method) ja pääkomponenttiestimaattoria kutistavaksi estimaattoriksi (engl. shrinking estimator) (Brown, 1993).

4.2 Pääkomponenttiestimaattorin a_{APK} keskineliövirhe

Pääkomponenttiregressioestimaattorin a_{jAPK} harha on

$$E(a_{jAPK} - \alpha_j) = E(a_{jPNS}) - \alpha_j = 0 , \quad j \in S_A \tag{4.13}$$

tai

$$E(a_{jAPK} - \alpha) = 0 - \alpha_j = -\alpha_j , \quad j \notin S_A .$$

Pääkomponettiestimaattorin $\boldsymbol{a}_{APK}$ varianssi on

$$\mathrm{var}(a_{jAPK}) = \mathrm{var}(a_{jPNS}) = \sigma_y^2/\lambda_j , \qquad j \in S_A \tag{4.14}$$

tai

$$\mathrm{var}(a_{jAPK}) = 0 , \qquad j \notin S_A .$$

Koska estimaattorin a_{APK} kokonaisvarianssi on

$$\mathrm{var}(\boldsymbol{a}_{APK}) = \sum_{j \in S_A} \mathrm{var}(a_{jAPK}) = \sigma_y^2 \sum_{j \in S_A} 1/\lambda_j , \tag{4.15}$$

niin mitä pienempi on var(z_j) = λ_j, $j \notin S_A$, sitä pienempi on var($\boldsymbol{a}_{APK}$) kuin var($\boldsymbol{a}_{PNS}$).

Koska estimaattorin keskineliövirhe on sen harhan neliön ja kokonaisvarianssin summa, niin pääkomponenttiestimaattorin $\boldsymbol{a}_{APK}$ keskineliövirhe on

$$\text{MSE}(\boldsymbol{a}_{APK}) = \sum_{j \in S_A} \alpha_j^2 + \sigma_y^2 \sum_{j \in S_A} 1/\lambda_j$$

$$= \sum_{j \in S_A} \alpha_j^2 + \text{MSE}(\boldsymbol{a}_{PNS}) - \sigma_y^2 \sum_{j \in S_A} 1/\lambda_j . \qquad (4.16)$$

Pääkomponenttiestimaatorilla $\boldsymbol{a}_{APK}$ on pienempi keskineliövirhe kuin pienimmän neliösumman estimaattorilla $\boldsymbol{a}_{PNS}$ eli MSE($\boldsymbol{a}_{APK}$) < MSE($\boldsymbol{a}_{PNS}$), kun

$$\sigma_y^2 \sum_{j \in S_A} 1/\lambda_j + \sum_{j \in S_A} \alpha_j^2 < \sigma_y^2 \sum_{j \in S_p} 1/\lambda_j \qquad (4.17)$$

eli

$$\sum_{j \in S_A}^{p} \alpha_j^2 < \sigma_y^2 \sum_{j \in S_A}^{p} 1/\lambda_j$$

eli

$$\sum_{j \in S_A}^{p} \lambda_j \alpha_j^2 / \sigma_y^2 < 1 .$$

Estimaattorin $\boldsymbol{a}_{APK}$ keskineliövirheeseen vaikuttavat siis neliömatriisin $\boldsymbol{X'X}$ ominaisarvot ja parametrit α ja σ_y^2. (Leskinen, 1981.)

4.3 Ennusteet pääkomponenttiregressiossa

Ennustetaan estimoidulla pääkomponenttiregressiomallilla tuntematonta selitettävän muuttujan arvoa y_0, joka voidaan kirjoittaa muotoon (2.23). Ennusteen $\hat{y}_{0APK} = \boldsymbol{x}_0'\boldsymbol{b}_{APK}$ varianssi on

$$\text{var}(\hat{y}_{0APK}) = \sigma_y^2(1 + 1/n + \boldsymbol{x}_0' \boldsymbol{G}_{PK} \boldsymbol{x}_0) \qquad (4.18)$$

$$= \sigma_y^2(1 + 1/n + \boldsymbol{x}_0'(\sum_{j \in S_A} (1/\lambda_j) \boldsymbol{v}_j \boldsymbol{v}_j') \boldsymbol{x}_0) ,$$

missä G_{PK} on (4.2) (Brown, 1993). Oletetaan, että estimointiaineiston rakennematriisin X pääkomponenteilla $z_j, j \in S_A$, ennustetaan tuntematonta havaintovektoria y. Tällöin ennusteen $\hat{y}_{APK} = Za_{APK}$ keskineliövirhe MSE($\hat{y}_{APK}$) on

$$\text{MSE}(\hat{y}_{APK}) = \sigma_y^2(n+1) + \sigma_y^2 \sum_{j \in S_p} \lambda_j (\text{var}(a_{jAPK}) + \text{B}^2(a_{jAPK}))$$

$$= \sigma_y^2(n+1+A) + \sum_{j \notin S_A} \lambda_j \alpha_j^2 , \qquad (4.19)$$

missä $\text{B}^2(a_{jAPK})$ on estimaattorin a_{jAPK} harhan neliö.

4.4 Yksimuuttujaisessa tapauksessa valittujen pääkomponenttien indeksijoukon S_A valintamenetelmiä

Pääkomponenttiregressiossa on ongelmana valita kaikista pääkomponenteista $z_j, j \in S_p$ sopiva osajoukko $z_j, j \in S_A$, selittäväksi pääkomponenteiksi pääkomponenttiregressiomalliin. Seuraavassa on muutamia valittujen pääkomponenttien indeksijoukon S_A valintamenetelmiä.

1) Oletetaan, että $\epsilon \sim \text{N}(0, \sigma_y^2 I_n)$. Pääkomponenttiestimaattori a_{APK} voidaan nyt muodostaa seuraavasti

$$a_{jAPK} = \begin{cases} 0, & \text{jos } \lambda_j a_{jPNS}^2 / \hat{\sigma}_y < c \\ a_{jPNS}, & \text{jos } \lambda_j a_{jPNS}^2 / \hat{\sigma}_y \geq c \end{cases} \qquad j = 1,\ldots,p , \qquad (4.20)$$

missä c on 4, 2 tai 1, jolloin se vastaa likimain merkitsevyystasoja 5%, 16% ja 35%. Jos epäyhtälössä (4.17) parametreja α ja σ_y^2 approksimoidaan niiden pienimmän neliösumman estimaattoreilla, niin $c = 1$. (Leskinen, 1981.)

2) Rakennematriisin X pääkomponentit ovat seuraavassa suuruusjärjestyksessä $\text{var}(z_1) = \lambda_1 \geq \ldots \geq \text{var}(z_p) = \lambda_p$, missä $\lambda_j, j = 1,\ldots,p$, ovat neliömatriisin $X'X$ ominaisarvot. Valitaan indeksijoukkoon S_A ne A ensimmäistä rakennematriisin X pääkomponenttia, joiden varianssien summa on selittänyt

esimerkiksi 75% neliömatriisin $X'X$ kokonaisvarianssista. Jos neliömatriisi $X'X$ on korrelaatiomatriisi, niin tällöin tulee valituksi ne A ensimmäistä pääkomponenttia, jotka toteuttavat ensimmäisenä ehdon

$$\sum_{j=1}^{A} \mathrm{var}(z_j)/r = \sum_{j=1}^{A} \lambda_j/r \geq 0.75 \tag{4.21}$$

(Draper & Smith, 1966). Tämän valintamenetelmän mukaan valittujen pääkomponenttien indeksijoukkoon S_A eivät kuulu ne pääkomponentit, joiden varianssi on pieni eli $\mathrm{var}(z_j) = \lambda_j \approx 0$, sillä nämä pääkomponentit selittävät heikosti rakennematriisin X kokonaisvaihtelua. Tämä valintamenetelmä takaa, että pääkomponentit, jotka selittävät hyvin vähintään yhden selittävän muuttujan varianssin, tulevat mukaan joukkoon S_A.

3) Oletetaan, että rakennematriisi X on täysiasteinen eli $r = p$, mutta sen pienimmät ominaisarvot ovat lähes nollia. Jos oletetaan, että matriisin X osittaisaste (engl. fractional rank) s on välillä $A - 1 < s < A$, missä niin tällöin voidaan muodostaa osittainen pääkomponenttiestimaattori (engl. fractional principal component estimator)

$$a_{jPK} = \begin{cases} 0, & j = A+1,\ldots,p \\ a_A a_{jPNS}, & j = A \\ a_{jPNS}, & j = 1,\ldots,A-1, \end{cases} \tag{4.22}$$

missä $0 \leq a_A \leq 1$. Kun

$$\hat{a}_A = \hat{\tau}_A^2/(1+\hat{\tau}_A^2) \qquad \text{ja} \qquad \hat{\tau}_A^2 = \lambda_j a_{jPNS}^2/\sigma_y^2, \tag{4.23}$$

niin saadaan seuraava osittainen pääkomponenttiestimaattori (engl. modified fractional rank estimator)

$$a_{jPK} = \begin{cases} 0, & \text{jos } \hat{\tau}_j^2 \leq 1,\ j = A+1,\ldots,p \\ \hat{a}_A a_{jPNS}, & j = A \\ a_{jPNS}, & j = 1,\ldots,A-1, \end{cases} \tag{4.24}$$

missä osittaisaste $s = A + \hat{a}_A - 1$. (Leskinen, 1981.)

4) Estimoidaan indeksijoukko S_A *cross-validation* -menetelmällä. Muodostetaan selitysasteen mielessä paras yhden selittävän pääkomponentin regressiomalli ja tähän malliin lisätään uusia selittäviä pääkomponentteja, kunnes ennusteen virheneliösumma

$$PRESS\ (S_A) = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 , \tag{4.25}$$

missä y_i on selitettävän muuttujan havaittu arvo ja $\hat{y}_{(i)}$ on tämän ennuste, joka on saatu pääkomponenttiregressiolla ilman havaittua havaintoa i, on saavuttanut miniminsä. Indeksijoukkoon S_A lisätään uusia pääkomponentteja kuten etenevässä (engl. forward) valinnassa. (Stone & Brown, 1990.)

5) Lasketaan testisuure

$$W = \frac{(\text{PRESS}\ (S_A) - \text{PRESS}\ (S_{A+1}))/D_m}{\text{PRESS}\ (S_{A+1})/D_r} , \tag{4.26}$$

missä PRESS(S_A) on ennusteen virheneliösumma (4.25), kun valittujen indeksien joukko on S_A, PRESS(S_{A+1}) on ennusteen virheneliösumma (4.25), kun valittujen indeksien joukkoon S_A on lisätty uusi selittävä pääkomponentti kuten etenevässä (engl. forward) valinassa, $D_m = (n + p - 2A)$ ja

$$D_r = p(n - 1) - \sum_{j=1}^{A} (n + p - 2j) . \tag{4.27}$$

Oletetaan aluksi, että $S_A = \{\}$, jolloin $\hat{y}_{(i)} = \bar{y}_{(i)}$, missä $\bar{y}_{(i)}$ selitettävän muuttujan y keskiarvo, joka on laskettu ilman havaintoa i. Jos testisuure $W > 1$, niin valittujen indeksien joukko S_A kasvaa yhdellä indeksillä, jolloin $S_A = S_{A+1}$, minkä jälkeen testataan uutta indeksijoukkoa S_{A+1}, kunnes testisuure $W \leq 1$. (Jackson, 1991.)

4.5 Monimuuttujainen pääkomponenttiestimaattori A_{PK}

Oletetaan, että regressiomalli on määritelty yhtälön (2.5) mukaan, missä selitettäviä muuttujia on useita. Kun pääkomponenttiregressiomalliin valitut pääkomponentit ovat z_j, $j \in S_A$, niin monimuuttujainen pääkomponenttiestimaattori A_{PK} on

$$A_{PK} = \Lambda_A^{-1} Z_A' Y = \Lambda_A^{-1} V_A' X' Y , \tag{4.28}$$

missä $S_A \in S_p$, $\Lambda_A^{-1} = \mathrm{diag}(1/\lambda_j)$, $\boldsymbol{Z}_A = \boldsymbol{X}\boldsymbol{V}_A$ ja $\boldsymbol{V}_A = (\boldsymbol{v}_j)$, $\boldsymbol{v}_j$ on neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvoa λ_j vastaava ominaisvektori ja S_A on valittujen pääkomponenttien indeksijoukko, $j \in S_A$ (vrt. kaava (4.8)).

Kun regressiomalli on määritelty kanonisen yhtälön (2.6) mukaan, niin monimuuttujainen pääkomponenttiestimaattori $\boldsymbol{A}_{PK}$ on

$$\boldsymbol{A}_{PK} = (\sqrt{\Lambda_A^{-1}}\ \boldsymbol{0})\boldsymbol{U}\ , \tag{4.29}$$

missä $\Lambda_A^{-1} = \mathrm{diag}(1/\lambda_A)$, $j \in S_A$, S_A on valittujen pääkomponenttien indeksijoukko, nollamatriisin $\boldsymbol{0}$ ulottuvuus on $(n - A \times A)$ ja $\boldsymbol{U} = \boldsymbol{T}\boldsymbol{Y}$ (vert. (4.9)).

Monimuuttujainen pääkomponenttiestimaattori $\boldsymbol{A}_{PK}$ voidaan kirjoittaa muotoon

$$a_{iPK}^{(j)} = u_i^{(j)}/\sqrt{\lambda_i}\ , \quad i \in S_A^{(j)}\ , \quad \text{ja} \quad a_{iPK}^{(j)} = 0\ , \quad i \notin S_A \quad j = 1,\ldots,q\ , \tag{4.30}$$

eli pääkomponenttiestimaattori $\boldsymbol{A}_{PK}$ voidaan hajottaa q yksimuuttujaiseen pääkomponenttiestimaattoriin $\boldsymbol{a}^{(j)}{}_{APK}$, $j = 1,\ldots,q$ (vrt. (4.9)). Jos pääkomponenttiregressiossa indeksijoukkojen $S_A^{(j)}$, $j = 1,\ldots,q$, estimointi perustuu pelkästään selittävien muuttujien $\boldsymbol{x}_1,\ldots,\boldsymbol{x}_p$, pääkomponenttianalyysiin eikä siinä huomioida pääkomponenttien $\boldsymbol{z}_j$, $j = 1,\ldots,p$ korrelaatiota selitettävien muuttujien $\boldsymbol{y}^{(j)}$, $j = 1,\ldots,q$, kanssa, niin tällöin indeksijoukot $S_A^{(j)}$, $j = 1,\ldots,q$, ovat samoja ja monimuuttujainen pääkomponenttiestimaattori voidaan esittää muodossa (4.29). Jos indeksijoukkojen $S_A^{(j)}$, $j = 1,\ldots,q$, yhtenä valintaperustana on pääkomponenttien $\boldsymbol{z}_j$, $j = 1,\ldots,p$, korrelaatiot selitettävien muuttujien $\boldsymbol{y}^{(j)}$, $j = 1,\ldots,q$, kanssa, niin tällöin indeksijoukot $S_A^{(j)}$, $j = 1,\ldots,q$, saattavat erota toisistaan ja monimuuttujaista pääkomponenttiestimaattoria ei voida esittää muodossa (4.29), vaan se esitetään muodossa (4.30).

Monimuuttujainen pääkomponenttiestimaattori $\boldsymbol{A}_{PK}$ on harhainen estimaattori, jonka harha on yksimuuttujaisten estimaattoreiden $\boldsymbol{a}^{(j)}{}_{APK}$, $j = 1,\ldots,q$, harhojen summa ja jonka kokonaisvarianssi on yksimuuttujaisten estimaattoreiden $\boldsymbol{a}^{(j)}{}_{APK}$, $j = 1,\ldots,q$, varianssien summa. Täten monimuuttujaisen pääkomponenttiestimaattorin A_{PK} keskineliövirhe on yksimuuttujaisten estimaattoreiden $\boldsymbol{a}^{(j)}{}_{APK}$, $j = 1,\ldots,q$, keskineliövirheiden summa.

4.6 Monimuuttujaisessa tapauksessa valittujen pääkomponenttien indeksijoukon S_A valintamenetelmiä

Monimuuttujaiselle pääkomponenttiestimaattorille $A_{PK} = (a^{(1)}_{APK},...,a^{(q)}_{APK})$ voidaan valita indeksijoukot $S_A^{(j)}, j = 1,...,q$, toisistaan riippumatta kappaleessa 4.4. estitetyin valintamenetelmin. Koska kappaleessa 4.4 indeksijoukkojen $S_A^{(j)}$, $j = 1,...,q$, valintamenetelmä (4.21) perustuu pelkästään selittävien muuttujien pääkomponenttianalyysiin, niin niissä indeksijoukot $S_A^{(j)}$, $j = 1,...,q$, tulevat olemaan samoja. Jos indeksijoukot $S_A^{(j)}, j = 1,...,q$, valitaan toisistaan riippumatta, niin tällöin ei hyödynnetä q normaaliyhtälöiden yhteisinformaatiota.

4.6.1 *Cross-validation*

Valitaan indeksijoukko $S_A = S_A^{(j)}$, $j = 1,...,q$, *cross-validation* -menetelmällä. Muodostetaan selitysasteen mielessä paras yhden selittävän pääkomponentin regressiomalli ja lasketaan ennustteen virheneliösuma

$$\text{PRESS}\ (S_A) = \sum_{i=1}^{n} \sum_{j=1}^{q} (y_i^{(j)} - \hat{y}_{(i)}^{(j)})^2 , \qquad (4.31)$$

missä $y_i^{(j)}$ on selitettävän muuttujan havaittu arvo ja $\hat{y}_{(i)}^{(j)}$ on tämän ennuste, joka on saatu pääkomponenttiregressiolla ilman havaintoa i. Valitaan indeksijoukkoon S_A uusia pääkomponentteja kuten etenevässä menetelmässä (vrt. (4.25)). Kun joukkoon S_A on lisätty uusi pääkomponentti, niin lasketaan tämän jälkeen aina ennusteen virheneliösumma. Valitaan se valittujen pääkomponenttien indeksijoukko S_A, joka antaa pienimmän ennusteen virheneliösumman.

5 Osittainen pienimmän neliösumman regressio

Osittaista pienimmän neliösumman regressiota (engl. partial least squares regression) eli PLS-regressiota käytetään, kun joko selittävät tai selitettävät muuttujat tai kumpikin ovat multikollineaarisia. PLS-regressiota voidaan käyttää jopa silloin, kun havaintoja on vähemmän kuin selittäviä muuttujia tai kun selitettäviä muuttujia on enemmän kuin selittäviä muuttujia. PLS-algoritmi on laskuteknisesti helpohko menetelmä, jolla pystytään melko vaivattomasti mallittamaan suuriakin aineistoja. Selittävien muuttujien multikollineaarisuuden ongelma ratkaistaan pääkomponenttiregressiossa ja PLS-regressiossa samalla periaatteella, mutta lisäksi PLS-regressiossa mallitetaan havaintomatriisia Y. Koska PLS-regressiossa regressoidaan havaintomatriisi Y rakennematriisia X vasten ja päin vastoin, niin siinä voidaan hylätä selittävän ja selitettävän muuttujan roolit. PLS-regressiolla on saatu tarkkoja selitettävien muuttujien ennusteita ja sitä on käytetty usein kalibrointimenetelmänä esim. Garthwaite (1994), Otto & Wegscheider (1985) ja Haaland & Thomas (1988). PLS-regressio on tässä tutkielmassa esitetyistä säännöllistämismenetelmistä nuorin. Sen teoriaa ovat kehittäneet mm. Helland (1988) ja Höskuldsson (1988).

PLS-algoritmissa matriisi X, jonka ulottuvuus on $n \times p$, kirjoitetaan rekursiiviseen muotoon

$$X = TP' + X_{A+1} = t_1 p_1' + t_2 p_2' + \ldots + t_A p_A' + X_{A+1}, \tag{5.1}$$

missä matriisin T ulottuvuus on $n \times A$, matriisin P ulottuvuus on $p \times A$ ja X_{A+1} on $(n \times p)$-jäännösmatriisi. PLS-regressiossa matriisi Y, jonka ulottuvuus on $n \times q$, kirjoitetaan myös rekursiiviseen muotoon

$$Y = TC' + Y_{A+1} = t_1 c_1' + t_2 c_2' + \ldots + t_A c_A' + Y_{A+1}, \tag{5.2}$$

missä matriisin T ulottuvuus on $n \times A$, matriisin C ulottuvuus on $q \times A$ ja jäännösmatriisi Y_{A+1} ulottuvuus on $n \times q$. PLS-regressiossa matriisin X ja matriisin Y yhteyden ajatellaan välittyvän PLS-komponenttien kautta. PLS-regressiossa muodostetaan PLS-komponentit siten, että niitä on mahdollisimman vähän, mutta ne selittävät suurimman osan rakennematriisin X vaihtelusta, jolla voidaan parhaiten ennustaa havaintomatriisia Y.

Kun matriisi X kirjoitetaan rekursiiviseen muotoon (5.1), niin joko matriisin T sarakevekto-

reiden tai matriisin C sarakevektoreiden tulee olla keskenään ortogonalisia. Tässä tutkielmassa esitetyssä PLS-algoritmissa, joka on Höskuldssonin (1988) mukaan, on vaadittu matriisin T sarakevektoreiden keskinäinen ortogonalisuus. Riippumatta siitä, onko PLS-algoritmissa vaadittu matriisin T tai matriisin C sarakevektoreiden keskenäinen ortogonalisuus, se antaa samat sovitteet (Helland, 1988).

Kun pääkomponenttianalyysiä voidaan pitää menetelmänä, jossa mallitetaan havainto-aineiston selitettävien muuttujien kokonaisvarianssin rakennetta, niin PLS-menetelmää voidaan pitää menetelmänä, jossa mallitetaan havaintoaineiston selitettävien ja selittävien muuttujien kovarianssirakennetta. Tällöin yksi havainto ei ole olennainen osa mallin teoriaa vaan yksittäisellä havainnolla estimoidaan ainoastaan mallin parametreja.

Tässä tutkielmassa on esitetty vain monimuuttujainen PLS-regressio. Yksimuuttujaisen PLS-algoritmin on esittänyt mm. Helland (1988). Tässä tutkielmassa osittaisen pienimmän neliösumman regression teoria ja algoritmi on esitetty Höskuldssonin (1988) mukaan.

5.1 Osittaisen pienimmän neliösumman algoritmi ja sen geometrinen tulkinta

Asetetaan vektorin u_0 arvoksi matriisin Y ensimmäinen sarake. Asetetaan $X_1 = X$ ja $Y_1 = Y$. PLS-algoritmin i:nnes kierros on

1) $w_i = X_i'u_{i-1}/(u_{i-1}'u_{i-1})$
2) Skaalataan vektorin w_i pituudeksi yksi.
3) $t_i = X_i w_i$
4) $c_i = Y_i't_i/(t_i't_i)$
5) Skaalataan vektorin c_i pituudeksi yksi
6) $u_i = Y_i c_i/(c_i'c_i)$
7) Jos u_i ei konvergoi askeleessa 6, niin siirrytään askeleeseen 2, muutteen siirrytään askeleeseen 8.
8) $p_i = X_i't_i/(t_i't_i)$
9) $q_i = Y_i'u_i/(u_i'u_i)$
10) $b_i = u_i't_i/(t_i't_i)$
11) $X_{i+1} = X_i - t_i p_i'$
12) $Y_{i+1} = Y_i - b_i t_i c_i'$

13) Jos matriisi $X \neq 0$ tai käytetyn pysähdyssäännön kriteerin mukaan jatketaan iterointia, niin siirrytään askeleeseen 1, muuten lopetetaan iterointi.

PLS-algoritmissa on useita pienimmän neliösumman estimaattoreita. Alla ovat regressiomallit, joissa numero regressiomallin edessä viittaa PLS-algoritmin askeleeseen, jossa on esiteyn regressiomallin pienimmän neliösumman estimaattori:

1) $X_i = u_{i-1}w_i' + \epsilon$

4) $Y_i = t_i c_i' + \epsilon$

8) $X_i = t_i p_i' + \epsilon$

9) $Y_i = u_i q_i' + \epsilon$

10) $u_i = t_i b_i' + \epsilon$

PLS-regressiomallin ulottuvuuden i vektorit voidaan kirjoittaa edellisen ulottuvuuden i - 1 vektoreiden funktiona. Vektori u_i on vektorin u_{i-1} funktiona

$$u_i = Yc_i/(c_i'c_i)$$

$$= YY't_i/(c_i'c_i)(t_i't_i)$$

$$= YY'Xw_i/((c_i'c_i)(t_i't_i)(w_i'w_i))$$

$$= YY'XX'u_{i-1}/((c_i'c_i)(t_i't)(w_i'w_i)(u_{i-1}'u_{i-1})) .$$

Vektorit c_i, t_i ja w_i ovat

$$c_i = Y'XXYc_{i-1}/((t_i't_i)(w_i'w_i)(u_{i-1}'u_{i-1})(c_{i-1}'c_{i-1}))$$

$$t_i = XX'YY't_{i-1}/((w_i'w_i)(u_{i-1}'u_{i-1})(c_{i-1}'c_{i-1})(t_{i-1}'t_{i-1}))$$

$$w_i = X'YY'Xw_{i-1}/((u_{i-1}'u_{i-1})(c_{i-1}'c_{i-1})(t_{i-1}'t_{i-1})(w_{i-1}'w_{i-1})) .$$

$$XX'YY't = at \tag{5.3}$$

Jos PLS-algoritmin vektorit u, t, w ja c konvergoivat, niin

$$YY'XX'u = au \tag{5.4}$$

$$Y'XX'Yc = ac \tag{5.5}$$

$$X'YY'Xw = aw\ , \tag{5.6}$$

missä a on kyseisen ominaisarvotehtävän suurin ominaisarvo. Täten vektorit u, c, t ja w ovat approksimoitujen matriisien suurinta ominaisarvoa vastaavat ominaisvektorit.

Tässä tutkielmassa matriisien U ja T sarakevektoreita kutsutaan PLS-komponenteiksi ja vektorit u_i ja t_i muodostavat ns. PLS-komponenttiparin.

Tarkastellaan PLS-algoritmissa olevien vektoreiden ja matriisien geometrisia ominaisuuksia. Merkitään

$$W = (w_1, w_2, \ldots, w_A)\ , \tag{5.7}$$

missä W on $(n \times A)$-matriisi, jonka sarakkeina ovat vektorit w. Käytetään tätä merkintätapaa kaikille PLS-algoritmin tuottamille matriiseille.

PLS-algoritmin ideana on, että jäännösmatriisi X_i lasketaan aina edellisestä jäännösmatriisista X_{i-1}:

$$\begin{aligned} X_i &= X_{i-1} - t_{i-1}p'_{i-1} \\ &= X_{i-1} - t_{i-1}t'_{i-1}X_{i-1}/(t'_{i-1}t_{i-1}) \\ &= (I - t_{i-1}t'_{i-1}/(t'_{i-1}t_{i-1}))X_{i-1} \end{aligned} \tag{5.8}$$

$$= (I - t_{i-1}t'_{i-1}/(t'_{i-1}t_{i-1}))(X_{i-2} - t_{i-2}t'_{i-2}X_{i-2}/(t'_{i-2}t_{i-2})) .$$

PLS-algoritmissa olevilla vektoreilla w, t, u ja p ovat seuraavat ortogonalisuusominaisuudet:

Ominaisuus 1

Painovektorit w ovat keskenään ortogonalisia:

$$w'_i w_j = 0 , \qquad \text{kun } i \neq j . \tag{5.9}$$

Todistus : Oletetaan, että $i < j$. Kirjoitetaan jäännösmatriisi X_j muotoon

$$X_j = Z(X_i - t_i t'_i/(t'_i t_i)X_i) , \tag{5.10}$$

missä Z on jokin matriisi. Osoitetaan, että

$$X_j w_i = 0 , \tag{5.11}$$

kun $j > i$. Kun määritellään jäännösmatriisi X_j yhtälön (5.11) mukaan, niin saadaan

$$X_j w_i = (X_i - t_i t'_i/(t'_i t_i)X_i) w_i = t_i - t_i t'_i/(t'_i t_i) t_i = t_i - t_i = 0 . \tag{5.12}$$

Kun w_j on määritelty yhtälön (5.7) mukaan ja käytetään apuna yhtälöä (5.12), niin saadaan

$$w'_j w_i = w'_j X'_j Y_j Y'_j X_j w_i / a_j = 0 . \square \tag{5.13}$$

Ominaisuus 2

Vektorit t_i ovat keskenään ortogonalisia:

$$t'_i t_j = 0 , \qquad \text{kun } i \neq j . \tag{5.14}$$

Todistus: Oletetaan $i < j$. Jäännösmatriisi X_j on jäännösmatriisin X_i funktiona

$$\begin{aligned} X_j &= X_{j-1} - X_{j-1} w_{j-1} t_{j-1}' X_{j-1} / (t_{j-1}' t_{j-1}) \\ &= X_{j-1}(I - w_{j-1} t_{j-1}' X_{j-1} / (t_{j-1}' t_{j-1})) \\ &= X_{i+1} Z \\ &= (X_i - t_i t_i' X_i / (t_i' t_i)) Z \ , \end{aligned} \tag{5.15}$$

missä Z on jokin matriisi. Nyt saadaan

$$t_i' X_j = (t_i' X_i - t_i' t_i t_i' X_i / (t_i' t_i))) = (t_i' X_i - t_i' X_i) = 0 \tag{5.16}$$

kun $i < j$ ja tällöin

$$t_i' t_j = t_i' X_j w_j = 0 \ , \qquad \text{kun } i < j \ . \ \square \tag{5.17}$$

Täten avaruuden C(X), joka on matriisin X sarakkeiden virittämä avaruus, ortogonalisena kantana ovat vektorit t. Kun oletetaan, että matriisin X aste on r, niin matriisi X voidaan kirjoittaa rekursiiviseen muotoon

$$X = \sum_{i=1}^{r} t_i p_i' + X_{r+1} \ , \tag{5.18}$$

missä jäännösmatriisi X_{r+1} on ortogonalinen matriisin Y kanssa. Jos oletetaan, että matriisi X_{r+1} on nollamatriisi, niin matriisi X_0 ei sisällä informaatiota matriisista Y.

Ominaisuus 3

Vektorit w_i ja p_j ovat keskenään ortogonalisia:

$$w_i' p_j = 0 \ , \qquad \text{kun } i < j \ . \tag{5.19}$$

Todistus: Yhtälön (5.11) mukaan saadaan

$$w_i' p_j = w_i' X_j' t_j (t_j' t_j) = 0 , \qquad \text{kun } i < j . \ \square \tag{5.20}$$

Ominaisuus 4

Vektorit p_i ovat ortogonalisia avaruudessa $\ker(X) = \{p_i \,|Xp_i = 0\}$:

$$(p_i, p_j)_X = p_i'(X'X)^- p_j = 0 \tag{5.21}$$

kaikilla $i \neq j$, missä $(X'X)^-$ on matriisin $X'X$ yleistetty käänteismatriisi.

Todistus: Yhtälöstä (5.18) saadaan neliömatriisille $X'X$ esitys

$$X'X = \sum (t_i' t_i) p_i p_i' = PLP' , \tag{5.22}$$

missä L diagonaalimatriisi, jossa on kertoimet $t_i' t_i$. $\square$

Edellä on esitetty vektoreiden w, t, u ja p ortogonaliset ominaisuudet. Yleisesti kuitenkin on

$$p_i' w_j \neq 0 \tag{5.23}$$

kaikilla $i < j$. Neljästä edellä mainitusta PLS-algoritmin ominaisuudesta seuraa, että jäännösmatriisi X_i ei riipu tavasta, jolla uusi vektori t_i saadaan, vaan vektorit t määritellään siten, että ne muodostavat avaruuden $C(X)$ aliavaruuden $C(T)$ ortogonalisen kannan.

Kun vektorin c_i pituus on skaalattu ykköseksi, niin jäännösmatriisit voidaan kirjoittaa muotoon

$$X_{i+1} = X_i - t_i p_i' \quad \text{ja} \quad Y_{i+1} = Y_i - b_i t_i c_i' \tag{5.24}$$

$$Y_{i+1}' Y_{i+1} = Y_i' Y_i - b_i^2 (t_i' t_i) c_i c_i' . \tag{5.25}$$

$$Y'_{i+1}X_{i+1} = Y'_iX_i - b_i(t'_it_i)c_ip'_i . \tag{5.26}$$

$$X'_{i+1}X_{i+1} = X'_iX_i - (t'_it_i)p_ip'_i \tag{5.27}$$

Oletetaan nyt, että vektorin c pituutta ei ole skaalattu ykköseksi PLS-algoritmissa. Tällöin

$$u't = c'Y't/(c'c) = c'(Y't)/(c'c) = c'c(t't)/(c'c) = t't . \tag{5.28}$$

Yllä olevasta yhtälöstä seuraa, että

$$b_i = u'_it_i/(t'_it_i) = 1 \tag{5.29}$$

kaikilla i. Tästä lähtien oletetaan, että vektorin c pituus ei ole skaalattu ykköseksi. Jos kuitenkin halutaan saada kerroin b_i, niin se saadaan yhtälöstä

$$b_i = |c_i| . \tag{5.30}$$

5.2 Muita osittaisen pienimmän neliösumman tulkintoja

Rotatoidaan matriisi X eli tehdään ortogonaalinen muunnos

$$S = XO_X , \tag{5.31}$$

missä O_x on ortogonalinen matriisi ja S on rotatoitu matriisi X. Koska

$$\sum s_{ij}^2 = \mathrm{tr}(S'S) = \mathrm{tr}(X'X) = \sum x_{ij}^2 , \tag{5.32}$$

niin matriisin X rotatointi ei muuta matriisin X kokonaisvaihtelua (vrt. (4.6)). Rotatoidaan myös matriisi Y

$$Z = YO_Y, \tag{5.33}$$

missä Z on rotatoitu matriisi Y ja O_Y ortogonalinen matriisi. Merkitään matriisin S sarakkeita merkinnällä s_i ja matriisin Z sarakkeita merkinnällä z_i ja oletetaan lisäksi, että selittäviä muuttujia on enemmän kuin selitettäviä muuttujia eli $p > q$. Mitataan vektorin s etäisyyttä vektorista z niiden neliöidyllä etäisyydellä I. Vektorin s neliöity etäisyys I vektorista z on

$$I = \sum_{i=1}^{q} |s_i - z_i|^2 + \sum_{i=q+1}^{p} |s_i|^2 . \tag{5.34}$$

Tulkinta 1

Oletetaan, että ortogonaliset matriisit O_X ja O_Y ovat määritelty siten, että neliöity etäisyys I mimoituu, mikä intuitiivisesti merkitsee, että rotatoidaan matriisin X sarakkeiden virittämä avaruus $C(X)$ ja matriisin Y sarakkeiden virittämä avaruus $C(Y)$ siten, että komponentit, jotka kuuluvat avaruuteen $C(Y)$, ovat mahdollisimman lähellä komponenttejaan, jotka kuuluvat avaruuteen $C(X)$.

Oletetaan, että kaksi ortogonalista matriisia on määritelty siten, että neliöity etäisyys I minimoituu, jolloin PLS-algoritmin vektorit t ja u eli PLS-komponenttipari saa arvon

$$t = s_1 \quad \text{ja} \quad u = z_1 . \tag{5.35}$$

Todistus: Yhtälö (5.35) voidaan kirjoittaa muotoon

$$I = \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\,\mathrm{tr}(X'YO_Y O_X') . \tag{5.36}$$

Kuten myöhemmin tullaan osoittamaan, niin ortogonaliset matriisit O_Y ja O_X, jotka ovat matriisin $X'Y$ singulaariarvohajotelmasta, minimoivat neliöidyn etäisyyden I

$$X'Y = \sum e_i g_i f_i' = GEF' , \tag{5.37}$$

missä matriisin G ensimmäisenä sarakeena on vektori t ja matriisin F ensimmäisenä sarakeena on vektori u. □

Neliöidyn etäisyyden I minimi on

$$\min I = \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\,\mathrm{tr}((X'YY'X)^{1/2}) \tag{5.38}$$

$$= \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\sum e_i .$$

PLS-algoritmissa valitaan yksi PLS-komponenttipari kerallaan. Etäisyys I vähenee jokaisella iterointikierroksella i $2e_{(i)1}$ verran, missä $e_{(i)1}$ on suurin matriisin $X_i'Y_i$ singulaariarvo. Koska iterointikierroksella i matriisin $X_i'Y_i$ toiseksi suurin singulaariarvo on pienempi kuin interointikierroksella $i+1$ matriisin $X_{i+1}'Y_{i+1}$ suurin singulaariarvo, niin vain yksi PLS-komponenttipari tulee valituksi jokaisella PLS-algoritmin iterointikierroksella. Täten etäisyyttä I voidaan pienentää eniten, kun valitaan yksi PLS-komponenttipari kerrallaan, lasketaan matriisin X ja Y jäännösmatriisit ja lasketaan seuraavan ulottuvuuden PLS-komponenttipari saaduista jäännösmatriiseista. PLS-algoritmia voidaan pitää siis askeltavana menetelmänä (engl. stepwise procedure), jossa valitaan yksitellen PLS-komponenttipari, joiden välinen etäisyys on pienin. PLS-algoritmia jatketaan niin kauan kuin avaruudessa $C(X_i)$ tai avaruudessa $C(Y_i)$ ei ole PLS-komponenttipareja, jotka ovat riittävän lähellä toisiaan, ts. tilastollisesti merkitseviä PLS-komponenttipareja ei ole jäljellä. Kun avaruudessa $C(X_i)$ ei ole jäljellä tilastollisesti merkitseviä PLS-komponentteja, niin avaruudessa $C(X_i)$ ei ole mitään, millä voitaisiin ennustaa havaintomatriisia Y.

Olkoon komponentit $f \in C(X)$ ja $g \in C(Y)$:

$$f = Xd \quad \text{ja} \quad g = Ye , \tag{5.39}$$

missä $|d| = 1$ ja $|e| = 1$. Komponenttien f ja g otoskovarianssi on

$$\mathrm{cov}(f,g) = f'g/n . \tag{5.40}$$

PLS-algoritmi valitsee komponentit f ja g siten, että niiden välinen otoskovarianssi on suurin mahdollinen otoskokovarianssi komponenteilla $f \in C(X)$ ja $g \in C(Y)$.

Tulkinta 2

Vektoreilla w ja c saadaan komponentit $f \in C(X)$ ja $g \in C(Y)$ siten, että

$$(\mathrm{cov}(t,u))^2 = (\mathrm{cov}(Xw,Yc))^2 = \max(\mathrm{cov}(f,q))^2 , \tag{5.41}$$

kun $|d| = |e| = 1$.

Todistus: Oletetaan, että matriisin $X'Y$ singulaariarvohajotelma on

$$X'Y = \sum a_i f_i g_i' . \tag{5.42}$$

Matriisin $X'Y$ suurin singulaariarvon a_1 neliö toteuttaa ehdon

$$(a_1)^2 = \max (d'X'Ye)^2 , \tag{5.43}$$

kun $|d| = |e| = 1$, $d = f_1$ ja $e = g_1$. Vektorit w ja c saavat PLS-algoritmissa arvot

$$w = f_1 \quad \text{ja} \quad c = g_1 , \tag{5.44}$$

jolloin väite (5.41) on tosi. □

Tulkinta 3

Monissa käytännön sovelluksissa korvataan kovarianssimatriisi $X'X$ matriisilla $X'VX$, missä matriisi V on positiivisesti definiitimatrisiisi. PLS-algoritmissa matriisi V määritellään

$$V = YY' , \tag{5.45}$$

jolloin

$$X'VX = X'YY'X . \tag{5.46}$$

Tällöin PLS-algoritmissa havainnot, joita vastaavat selitettävien muuttujien arvot ovat lähellä nollaa, saa alhaisen painon ja havainot, joita vastaavat selitettävien muuttujien arvot ovat itseisarvoltaan suuria, saa vastaavasti suuremman painon. Koska matriisi Y on keskitetty, niin matriisin Y keskiarvoa lähellä olevat arvot ovat lähes nollia. PLS-algoritmissa lähellä keskiarvoaan olevat havaintomatriisin Y arvot katsotaan olevan kohinaa, jota ei PLS-regressiolla hyödytä mallitaa. PLS-algoritmissa painotetaan havaintoja, joita vastaavilla $y^{(j)}$, $j = 1,...,q$, on suurimmat arvot, koska niissä katsotaan olevan eniten informaatiota käytetystä datasta. Tämä kuitenkin edellyttää, että matriisissa Y arvot, jotka eivät ole lähellä nollaa, ovat virheettömästi mitattu.

PLS-algoritmissa lasketaan ensiksi ominaisarvot ja -vektorit painotetusta kovarianssimatriisista

$$X'YY'X = ODO' , \tag{5.47}$$

jonka jälkeen rotatoidaan matriisi X eli tehdään muunnos

$$F = XO . \tag{5.48}$$

PLS-algoritmissa regressoidaan Y aina yhtä muunnoksen (5.48) komponenttia vasten iterointikierroksittain. Huomattavaa on, että pääkomponenttiregressiossa ominaisarvot ja -vektorit lasketaan kovarianssimatriisista

$$X'X = ODO' , \tag{5.49}$$

jonka jälkeen rotatoidaan matriisi X eli tehdään muunnos

$$F = XO$$

ja regressiodaan havaintomatriisi Y valittuja f_j, $j \in S_A$, vasten.

Oletetaan, että iteroidaan PLS-algoritmin iterointikierrosta i. Tällöin w_i on suurinta ominaisarvoa vastaava ominaisarvovektori yhtälössä (5.48), missä X on korvattu jäännösmatriisilla X_i ja Y on korvattu jäännösmatriisilla Y_i. Vektori $t_i = X_i w_i$ eli vektori t_i vastaa ortogonalista muunnosta (5.49). Regressoidaan matriisi Y_i muunnosta t_i vasten. Jäännösmatriisin Y_i projektio on

$$t_i c_i' \tag{5.50}$$

ja täten jäännösmatriisi Y_{i+1} voidaan kirjoittaa muotoon

$$Y_{i+1} = Y_i - t_i c_i' . \tag{5.51}$$

Jäännösmatriisin Y_{i+1} kovarianssimatriisi on

$$Y_{i+1}' Y_{i+1} = Y_i' Y_i - c_i c_i' (t_i' t_i) . \tag{5.52}$$

Projektio (5.50) ei muutu, jos jäännösmatriisi Y_i korvataan matriisilla Y.

Todistus: Olkoon w_i matriisin $X'YY'X$ suurinta ominaiarvoa vastaava ominaisarvovektori ja $t_i =$

$x_i w_i$. Matriisin Y_i projektio vektorille t_i on

$$t_i' Y_i / (t_i' t_i) t_i = t_i c_i' \tag{5.53}$$

ja jäännöskovarianssimatriisi $Y_{i+1}' Y_{i+1}$ on

$$Y_{i+1}' Y_{i+1} = (Y_i - t_i c_i')' (Y_i - t_i c_i') \tag{5.54}$$

$$= Y_i' Y_i - c_i t_i' Y_i - Y_i' t_i c_i' + c_i c_i' (t_i' t_i)$$

$$= Y_i' Y_i - c_i c_i' (t_i' t_i) .$$

Kun $Y_i' t_i = Y' t_i$, niin jäännösmatriisin Y_i projektio vektorille t_i on sama kuin matriisin Y projektio vektorille t_i. □

PLS-algoritmia voidaan täten pitää ortogonalisten komponenttien askeltavana regressiomenetelmänä, jossa jokaisessa askeleessa määrätään PLS-komponenttipari, jolla saadaan painotetun kovarianssimatriisin $X'VX = X'YY'X$ suurin 'varianssi'. Tämä kovarianssimatriisi on painotettu siten, että x_{ij}, jota vastaava $y_i^{(k)}$ on lähellä kohinaa eli $y_i^{(k)} \approx 0$, saa pienen painon ja x_{ij}, jota vastaava $y_i^{(k)}$ ei ole lähellä kohinaa, saa suuremman painon $i = 1,...,n, j = 1,...,p, k = 1,...,q$.

Matriisin Y regressio matriisille $T = (t_1, t_2, ..., t_i)$ voidaan kirjoittaa muotoon

$$Y = \sum_{j=1}^{i} t_j c_j' + Y_{i+1} , \tag{5.55}$$

Täten jäännösmatriisin Y_{i+1} kovarianssimatriisi voidaan kirjoittaa muotoon

$$Y_{i+1}' Y_{i+1} = Y'Y - \sum_{j=1}^{i} c_j c_j' (t_j' t_j) . \tag{5.56}$$

5.3 PLS-algoritmin matemaattinen perusta

PLS-menetelmä perustuu matriisin $X'Y$ singulaariarvohajotelmaan

$$X'Y = \sum a_i f_i g_i' , \tag{5.57}$$

missä f_i ja g_i ovat approksimoidun ulottuvuuden i ortonormaalit vektorit ja a_i on matriisin $X'Y$ i:nneksi suurin singulaariarvo. Matriisin $X'Y$ suurimmalle singulaariarvolle a_1 pätee

$$a_1^2 = \max (f'X'Yg)^2 , \tag{5.58}$$

missä $| f | = | g | = 1$.Tämä on PLS-algoritmin tulkinta 2. Yhtälön (5.58) todistus perustuu Cauchy-Schwartz -epäyhtälöön.

Tulkintaa 1 todistettaessa oletettiin, että matriisilla Y on vähemmän sarakkeita kuin matriisilla X. Matriisit Y ja O_Y saadaan samankokoiseksi kuin matriisi X, kun korvataan matriisien Y ja O_Y puuttuvat sarakkeet nollavektoreilla. Nyt neliöity etäisyys I voidaan kirjoittaa muotoon

$$I = |XO_X - YO_Y|^2 = \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\,\mathrm{tr}(X'YO_YO_X') . \tag{5.59}$$

Neliöidyn etäisyyden I minimi on sama kuin

$$\max \mathrm{tr}(X'YV) , \tag{5.60}$$

missä $X'YV$ maksimoidaan ortogonalisten matriisien V suhteen. Tällöin matriisin V ratkaisu on

$$V = GF' , \tag{5.61}$$

missä G ja F ovat matriisin $X'Y$ singulaariarvohajotelman matriiseja. Täten yksi neliöidyn etäisyyden I mahdollinen minimi on neliöity etäisyys I (5.59), missä

$$O_X = F \quad \text{ja} \quad O_Y = G , \tag{5.62}$$

minkä todistaa tulkinta 1, eli

$$I_{\min} = \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\,\mathrm{tr}(X'YY'X)^{1/2} \tag{5.63}$$

$$= \mathrm{tr}(X'X) + \mathrm{tr}(Y'Y) - 2\sum a_i .$$

Pääperiaate, jolla PLS-algoritmi toimii, saadaan epäyhtälöstä

$$s_i^2(A - B) \geq s_{i+k}^2(A) - a_{i+k}^2 , \tag{5.64}$$

missä $s_i(A)$ on matriisin A i:nneksi suurin singulaariarvo ja k on matriisin B aste. Kun tätä epäyhtälöä sovelletaan PLS-algoritmiin, niin saadaan

$$s_1^2(X_{i+1}'Y) = s_1^2(X_i'Y - p_i t_i'Y) \geq s_2^2(X_i'Y) . \tag{5.65}$$

Täten matriisin $X_i'Y$ suurin singulaariarvo iterointikierroksella $i + 1$ on suurempi kuin toiseksi suurin singulaariarvo iterointikierroksella i. Tämän vuoksi vain yksi PLS-komponenttipari kerrallaan tulee valituksi PLS-algoritmissa.

Matriisin $X_i'Y$ suurimmalla singulaariarvolla s_1 on myös tulkinta

$$(s_1(X_i'Y))^2 = \max(f'X_i'YY'X_i f) , \tag{5.66}$$

kun $|f| = 1$. Tarkastellaan seuraavaa epäyhtälöä

$$(s_1(X_{i+1}'Y))^2 = \max(f'X_{i+1}'YY'X_{i+1}f) \tag{5.67}$$

$$= \max(f'(I - p_i w_i')X_i'YY'X_i(I - w_i p_i')f)$$

$$\leq \max(f'X_i'YY'X_i f)$$

$$= ((s_1(X_i'Y))^2 ,$$

kun $|f| = 1$. Yllä olevasta epäyhtälöstä seuraa, että koska valituttu ehto on $X_i'Y_i = X_i'Y$, niin jokaisella PLS-algoritmin kierroksella matriisin $X_i'Y_i$ arvot ovat pienempiä kuin edellisen kierroksen vastaavat arvot. Tämän vuoksi PLS-algoritmissa voidaan valita PLS-komponenttipareja niin kauan, kunnes matriisin $X_i'Y$ suurin singulaariarvo $s_1(X_i'Y)$ on kyllin pieni PLS-regressiossa käytetyn kriteerin mukaan.

5.4 Ennusteet PLS-regressiossa

Höskuldsson (1988) on myös esittänyt, kuinka PLS-regressiossa annetulla x_0 voidaan ennustaa tuntematonta y_0. Koska kappaleessa 6.6 annetulla testiaineiston rakennematriisilla X_t on ennustettu 'tuntematonta' testiaineiston havaintomatriisia Y_t Jacksonin (1991) mukaan, niin tässä esitetään PLS-regressiolla ennustaminen Höskuldssonin (1988) ja Jacksonin (1991) mukaan.

Ennustettaessa annetulla X_0 tuntematonta Y_0 annettu X_0 pitää esittää PLS-regressiossa vastaavassa rekurisiivisessa muodossa kuin rakennematriisin X approksimaatio on PLS-algoritmin askeleessa A, jotta voidaan laskea ennuste

$$Y_{0(A)} = t_{01}c_1' + \ldots + t_{0A}c_A' , \tag{6.68}$$

missä $Y_{0(A)}$ on tutemattoman havaintomatriisin Y_0 ennuste, A on valittujen PLS-komponenttiparien lukumäärä, matriisi C on saatu jo PLS-regressiomallia sovitettaessa ja vektoreiden c pituutta ei ole skaalattu ykköseksi. Ennusteessa (6.67) vektorit t_{0i}, $i = 1,\ldots,A$, saadaan Jacksonin (1991) mukaan

$$t_{0i} = X_{0i-1}w_i , \qquad i = 1,\ldots,A , \tag{6.69}$$

missä

$$X_{0i} = X_{0i-1} - t_{0i}p_i' , \qquad i = 2,\ldots,A , \tag{6.70}$$

vektorit w ja p on jo saatu PLS-regressiomallia sovitettaessa ja $X_{01} = X_0$.

Kun PLS-regressiossa rakennematriisi X aproksimoidaan matriisilla

$$X_{(A)} = \sum_{i=1}^{A} t_i p_i' , \tag{5.71}$$

missä A on valittujen PLS-komponenttiparien lukumäärä, niin PLS-regressiolla saatu havaintomatriisin Y sovite $Y_{(A)}$ voidaan kirjoittaa muotoon

$$Y_{(A)} = H_T Y = (\sum_{i=1}^{A} t_i t_i'/(t_i't_i))Y , \tag{5.72}$$

missä projektiomatriisi $\boldsymbol{H}_T$ on

$$\boldsymbol{H}_T = \sum_{i=1}^{A} \boldsymbol{t}_i \boldsymbol{t}_i' / (\boldsymbol{t}_i' \boldsymbol{t}_i) = \sum_{i=1}^{A} \boldsymbol{X}_i \boldsymbol{w}_i \boldsymbol{t}_i' / (\boldsymbol{t}_i' \boldsymbol{t}_i) \,. \tag{5.73}$$

Projektiomatriisi $\boldsymbol{H}_T$ on symmetrinen ja idempotentti ja siten

$$\sum_i h_{ij}^2 = h_{ij} \,, \quad 0 \le h_{ij} \le 1 \quad \text{ja} \quad \sum h_{ij} = A \,, \tag{5.74}$$

missä h_{ij} on matriisin $\boldsymbol{H}_T$ alkio ja A on PLS-komponenttiparien lukumäärä. Projektiomatriisilla $\boldsymbol{H}_T$ on siis vastaavat ominaisuudet kuin matriisilla $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, joten projektiomatriisia $\boldsymbol{H}_T$ voidaan analysoida vastaavasti kuin matriisia $\boldsymbol{H}$ lineaarisessa regressiotehtävässä.

5.5 PLS-regressiomallin ulottuvuuden valinta

Seuraavassa on muutamia valintamenetelmiä, joilla voidaan valita PLS-regressiomallin PLS-komponenttiparien lukumäärä A, joka on PLS-regressiomallin ulottuvuus.

5.5.1 PLS-regressiomallin jäännösten graafinen tarkastelu

Piirretään jäännösmatriisin $\boldsymbol{Y}_{A+1}$ normi $\|\boldsymbol{Y}_{A+1}\|$ PLS-regressiomallin ulottuvuuden A funktiona. Kuvan perusteella yritetään paikallistaa mallin luonnollinen PLS-komponenttiparien lukumäärä. (Jackson, 1991.)

5.5.2 *Cross-validation*

Lasketaan PLS-algoritmin jokaisella kierroksella

$$\text{PRESS}(A) = \sum_{i=1}^{n} \sum_{j=1}^{q} (y_i^{(j)} - \hat{y}_{(i)}^{(j)})^2 \,, \tag{5.75}$$

missä A on mallissa olevien PLS-komponenttiparien lukumäärä, $y_i^{(j)}$ on selitettävän muuttujan havaittu arvo ja $\hat{y}_{(i)}^{(j)}$ on tämän ennuste, joka on laskettu ilman havaintoa i. Valitaan se PLS-komponenttien lukumäärä A, jolla saadaan pienin PRESS(A). Yleensä *cross-validation* kriteeri

'pysäyttää' ennen kuin kaikki mahdolliset PLS-komponentit ovat mukana mallisissa, sillä 'loput' PLS-komponenttiparit eivät paranna mallin kykyä toimia ennusteena. (Höskuldsson, 1988.)

6 Teräksen karkenevuuden mallittamisesta

Tehtävänä on muodostaa monimuuttujainen regressiomalli teräksen karkenevuudelle käyttäen estimointimenetelminä säännöllistämismenetelmiä ja vertailla käytettyjen menetelmien "luotettavuutta" keskenään. Copas (1983) ehdottaa kokonaisaineiston jaettavaksi retrospektiiviseen aineistoon (engl. retrospective sample), josta estimoidaan regressiomallin parametrit, ja prospektiiviseen tai vahvistavaan aineistoon (engl. prospective or validation sample), josta ei estimoida regressiomallin parametreja. Tässä kokonaisaineiston eli teräksen karkenevuusaineiston otoskoko on 220 havaintoa. Jotta estimodun regressiomallin sopivuutta voidaan testata myös havaitsemattomiin otospisteisiin, niin kokonaiaineisto on jaettu prospektiivisen aineistoon, jota kutsutaan tässä testiaineistoksi, ja retrospektiivisen aineistoon, jota kutsutaan tässä estimointiaineistoksi. Testiaineiston otoskoko on 44 ja se on valittu kokonaisaineistosta systemaattisella otannalla. Loput kokonaisaineiston havainnot, joita on 176, muodostavat estimointiaineiston.

Mallitettaessa teräksen karkenevuutta on tässä valittu selitettäviksi muuttujiksi teräksen kovuudet 1.5, 13 ja 25 mm etäisyyksiltä koesauvan alapäästä. Merkitään näitä mitattuja kovuuksia seuraavasti: $y^{*(j)}$ ja $y_t^{*(j)}$, j = 1.5, 13, 25 mm, missä muuttujat $y^{*(j)}$ ja $y_t^{*(j)}$ ovat mitatut teräksen kovuudet etäisyyksiltä j mm koesauvan alapäästä, muuttuja $y^{*(j)}$ kuuluu estimointiaineistoon ja muuttuja $y_t^{*(j)}$ kuuluu testiaineistoon. Muuttujien $y^{*(j)}$ ja $y_t^{*(j)}$, j = 1.5, 13, 25, keskiarvot ja -hajonnat ovat taulukossa 6.1.

Taulukko 6.1: Selitettävien muuttujien $y^{*(j)}$ ja $y_t^{*(j)}$, j = 1.5, 13, 25, keskiarvot ja -hajonnat

Tunnusluku	Estimointiaineisto			Testiaineisto		
	$y^{*(1.5)}$	$y^{*(13)}$	$y^{*(25)}$	$y_t^{*(1.5)}$	$y_t^{*(13)}$	$y_t^{*(25)}$
Keskiarvo	48.05	40.27	29.27	48.43	40.76	29.27
Keskihajonta	6.0259	9.5228	9.1767	5.9621	9.6699	8.8026

Kuten taulukosta 6.1 nähdään, niin testiaineiston selitettävien muuttujien keskiarvot ja -hajonnat eivät juurikaan eroa estimointiaineiston selitettävien muuttujien keskiarvoista ja -hajonnoista. Tehdään selitettäville muuttujille $y^{*(j)}$ ja $y_t^{*(j)}$ muunnos (2.4), missä $\text{var}(y^{*(j)})^{1/2}$, $\bar{y}^{*(j)}$ ja n on laskettu estimointiaineistosta, j = 1.5, 13, 25. Tämän muunnoksen jälkeen estimointiaineiston havaintomatriisin $\boldsymbol{Y} = (y^{(1.5)}, y^{(13)}, y^{(25)})$ neliömatriisi $\boldsymbol{Y}'\boldsymbol{Y}$ on korrelaatiomatriisi

$$\boldsymbol{Y'Y} = \begin{pmatrix} 1.000 & .922 & .810 \\ .922 & 1.000 & .924 \\ .810 & .924 & 1.000 \end{pmatrix}. \qquad (6.1)$$

Kuten yllä olevasta neliömatriisista $\boldsymbol{Y'Y}$ nähdään, niin selitettävät muuttujat korreloivat keskenään erittäin voimakkaasti. Neliömatriisin $\boldsymbol{Y'Y}$ determinantti on 0.020 ja neliömatriisin $\boldsymbol{Y'Y}$ ominaisarvot ovat 2.771, 0.190 ja 0.038. Jotta monimuuttujainen regressiomalli olisi hyödyllinen, niin selitettävien muuttujien tulee korreloida melko voimakkaasti keskenään, mikä pätee tässä aineistossa.

Teräksen karkenevuuteen vaikuttavat teräksen seosaineiden C, Si, Mn, P, S, Cr, Ni, Mo, V, Ti, Cu, Als, B, Nb ja N pitoisuudet. Edellä mainittujen seosaineiden pitoisuuksien lisäksi selittävänä muuttujana on ns. vapaa boori Bjab = B + 0.212Ti - 0.715N. Koska seosaine Bjab on seosaineiden B, Ti ja N lineaarikombinaatio, niin selittävät muuttujat ovat multikollineaarisia. Estimointi- ja testiaineiston selittävien muuttujien keskiarvot ja -hajonnat ovat liitteessä A taulukossa A.1. Testiaineiston selittävien muuttujien keskiarvot ja -hajonnat eivät juurikaan eroa estimointiaineiston selittävien muuttujien keskiarvoista ja -hajonnoista. Tehdään estimointiaineiston selittäville muuttujille x_j^* ja testiaineiston selittäville muuttujille x_{tj}^*, j = C,...,Bjab, muunnos (2.3), missä $\bar{x}_j^*$, var(x_j^*) ja n on laskettu estimointiaineistosta. Näiden muunnosten jälkeen rakennematriisin $\boldsymbol{X} = (x_C,\ldots,x_{Bjab})$ neliömatriisi $\boldsymbol{X'X}$ on korrelaatiomatriisi, joka on liitteessä A taulukossa A.2.

Mitä enemmän selitettävien ja selittävien muuttujien välillä on lineaarista riippuvuutta, sitä luotettavampaa on muuttujilla x_j, j = C,...,Bjab, selittää muuttujien $y^{(j)}$, j = 1.5, 13, 25, vaihtelua. Estimointiaineiston selitettävien ja selittävien muuttujien korrelaatiokertoimet ovat liitteessä A taulukossa A.3. Kuten estimointiaineston selitettävien ja selittävien muuttujien korrelaatiomatriisista nähdään, varsinkin muuttuja C korreloi voimakkaasti selitettävien muuttujien kanssa.

Tässä estimoitava monimuuttujainen regressiomalli teräksen karkenevuudelle on

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}, \qquad (6.2)$$

missä selitettävien muuttujien (176 × 3) -havaintomatriisi $\boldsymbol{Y}$ ja selittävien muuttujien (176 × 16) -rakennematriisi $\boldsymbol{X}$ ovat estimointiaineistosta, B on (16 × 3) -regressiomatriisi ja E on (176 × 3) -jäännösmatriisi. Regressiomalli (6.2) voidaan kirjoittaa myös muotoon

$$y^{(j)} = X\beta^{(j)} + \epsilon^{(j)} = Z\alpha^{(j)} + \epsilon^{(j)}, \qquad j = 1.5,\ 13,\ 25\ , \tag{6.3}$$

missä vektori $\beta^{(j)}$ on selitettävää muuttujaa $y^{(j)}$ vastaava (16 × 1) -regressiovektori, $Z = XV$, $\alpha^{(j)} = V'\beta^{(j)}$, V on neliömatriisin $X'X$ ominaisvektorimatriisi ja $\epsilon^{(j)}$ on (176 × 1) -jäännösvektori. Vastaava kanoninen muoto (2.12) on

$$u_i^{(j)} = \sqrt{\lambda_i}\alpha_i^{(j)} + \epsilon_i^{(j)}\ , \quad i = 1,...,16\ , \quad \text{ja} \quad u_i^{(j)} = \epsilon_i^{(j)}\ , \quad i = 17,...,176\ , \tag{6.4}$$

$j = 1.5, 13, 25$.

Kun regressiomalli on muodossa (6.3), niin alkuperäisten havaintojen $y^{*(j)}, j = 1.5, 13, 25$ sovitteet, saadaan muunnoksella

$$\hat{y}^{*(j)} = \sqrt{n-1}\sqrt{\mathrm{var}(y^{*(j)})}\,\hat{y}^{(j)} + \bar{y}^{*(j)}\ , \qquad j = 1.5,\ 13,\ 25, \tag{6.5}$$

missä n on estimointiaineiston otoskoko, $(\mathrm{var}(y^{*(j)}))^{1/2}$ ja $\bar{y}^{*(j)}$ ovat estimointiaineiston muuttujan $y^{*(j)}$ keskihajonta ja keskiarvo, $\hat{y}^{(j)} = Xb^{(j)}$, $b^{(j)}$ on parametrin $\beta^{(j)}$ estimaattori tai $\hat{y}^{(j)} = Za^{(j)}$ ja $a^{(j)}$ on parametrin $\alpha^{(j)}$ estimaattori, $j = 1.5, 13, 25$. Jos regressiomalli on määritelty kanonisen yhtälön (6.4) mukaan, niin alkuperäisten havaintojen $y^{*(j)}, j = 1.5, 13, 25$, sovitteet saadaan muunnoksella

$$\hat{y}^{*(j)} = \sqrt{n-1}\sqrt{\mathrm{var}(y^{*(j)})}\,T\hat{u}^{(j)} + \bar{y}^{*(j)}\ , \qquad j = 1.5,\ 13,\ 25\ , \tag{6.6}$$

missä n, $\mathrm{var}(y^{*(j)})$ ja $\bar{y}^{*}$ ovat kuten yhtälössä (6.5), matriisi T on estimointiaineiston neliömatriisin XX' ominaisvektorimatriisi, $\hat{u}^{(j)} = \Lambda_p^{1/2}a^{(j)}$, $\Lambda_p = \mathrm{diag}(\lambda_1,...,\lambda_p)$, λ_i, $i = 1,...,p$, ovat estimointiaineiston neliömatriisin $X'X$ ominaisarvot ja $a^{(j)}$ on estimoitu parametrin $\alpha^{(j)}$ estimaattori, $j = 1.5, 13, 25$.

Jos oletetaan, että regressiomalli on muotoa (6.2), niin testiaineiston havaintomatriisin Y_t ennuste on

$$\hat{Y}_t = X_tB = Z_tA\ , \tag{6.7}$$

missä X_t on testiaineiston rakennematriisi, $Z_t = X_tV$, matriisin V sarakevektoreina ovat estimointiaineiston neliömatriisin $X'X$ ominaisvektorit, B ja A ovat parametrien B ja A estimaattorit, jotka on laskettu estimointiaineistosta. Ennuste (6.7) voidaan esittää myös muodossa

$$\hat{y}_t^{(j)} = X_t b^{(j)} + Z_t a^{(j)}, \quad j = 1.5, 13, 25, \tag{6.8}$$

missä X_t ja $Z_t = X_t V$ ovat kuten yhtälössä (6.7) ja $b^{(j)}$ ja $a^{(j)}$ ovat parametrien $\beta^{(j)}$ ja $\alpha^{(j)}$ estimaattorit, jotka on laskettu estimointiaineistosta, $j = 1.5, 13, 25$, (vrt. (6.3)). Jotta testiaineistosta saatua ennustetta voidaan verrata alkuperäiseen havaintomatriisiin Y_t, niin ennusteelle (6.7) tai (6.8) tehdään muunnos (6.5), missä $\hat{y}^{(j)}$ on ennuste $\hat{y}_t^{(j)}$ ja muut tarvittavat tunnusluvut ja matriisit saadaan estimointiaineistosta.

6.1 Sovitteen ja ennusteen selityskertoimet

Koska regressiomalleissa (6.3) ja (6.4) ovat parametrit α ja β tuntemattomia, niin parametrien α ja β säännöllistämismenetelmien estimaattoreiden keskineliövirheitä ei voida laskea eikä niitä siten voida verrata vastaavaan pienimmän neliösumman estimaattorin keskineliövirheeseen.

Koska tavoitteena on, että muodostettu teräksen karkenevuuden regressiomalli sopisi hyvin myös havaitsemattomiin otospisteisiin ts. regressiomallin validiteetti olisi hyvä, niin valittaessa lopullista regressiomallia sekä regressiomallin sovitteen että ennusteen tulisi olla riittäviä. Jos muodostetun regressiomallin parametrien estimaattoreilla ennustetaisiisn vain estimointiaineiston havaintoja, niin mallin ennustamiskykyä olisi vaikea arvioda täsmällisesti, koska estimointiaineisto olisi käytetty kahdesti (Helland, 1988). Ennusteen ja sovitteen riittävyyttä arvioidaan tässä tutkielmassa selityskertoimien avulla.

Yksimuuttujaisten regressiomallin (6.3) selityskertoimet, joita tässä kutsutaan sovitteiden selityskertoimeksi, saadaan Heikan, Minkkisen ja Taavitsaisen (1994) mukaan

$$R^{2(j)} = 1 - \frac{\sum_{i=1}^{n} (y_i^{(j)} - \hat{y}_i^{(j)})^2 / n}{\sum_{i=1}^{n_{total}} (y_{itotal}^{(j)} - \bar{y}_{total}^{(j)})^2 / n_{total}} \quad j = 1.5, 13, 25, \tag{6.9}$$

missä n_{total} on kokonaisaineiston otoskoko, $y^{(j)}_{total}$ on kokonaisaineiston selitettävä muuttuja ja $\bar{y}^{(j)}_{total}$ sen keskiarvo, $j = 1.5, 13, 25$. Selityskertoimessa (6.9) huomiodaan myös testianeistossa olevien selitettävien muuttujien kokonaisvaihtelu. Koska sovitteen keskineliövirhe on

$$\text{MSE}(\hat{y}) = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n}, \tag{6.10}$$

niin mitä pienempi on sovitteen keskineliövirhe, sitä suurempi on sovitteen selityskerroin. Monimuuttujaisen regressiomallin selityskerroin, jota tässä kutsutaan sovitteen selityskertoimeksi, on Höskuldssonin (1988) mukaan

$$R^2 = 1 - \mathrm{tr}((\hat{Y} - Y)'(\hat{Y} - Y))/\mathrm{tr}(Y'Y) , \tag{6.11}$$

missä siis selitettävät muuttujat on keskistetty ja niiden varianssit ovat yhtäsuuret. Regressiomallin selityskerroin ilmaisee, kuinka suuren osan estimointiaineiston selitettävän muuttujan vaihtelusta estimoitu sovite selittää.

Koska ennusteen $\hat{y}_t$ keskineliövirheessä on lauseke

$$E((\beta - b)'X_t'X_t(\beta - b)) , \tag{6.12}$$

missä parametri β on tuntematon, niin ennusteen keskineliövirhettä ei voida laskea, kun yllä oleva lauseke on ennusteen keskineliövirheessä. Koska tässä ennusteen $\hat{y}_t$ todellinen arvo y_t on tunnettu, niin ennusteiden (6.8) keskineliövirhe saadaan yhtälöstä

$$\mathrm{MSE}\,(\hat{y}_t^{(j)}) = \sqrt{\sum_{i=1}^{n_t} (y_{ti}^{(j)} - \hat{y}_{ti}^{(j)})^2/n_t} , \qquad j = 1.5,\ 13,\ 25 , \tag{6.13}$$

ja täten ennusteiden (6.8) selityskertoimet saadaan Heikan ym. (1994) mukaan

$$Q^{2(j)} = 1 - \frac{\sum_{i=1}^{n_t} (y_{it}^{(j)} - \hat{y}_{it}^{(j)})^2/n_t}{\sum_{i=1}^{n_{total}} (y_{itotal}^{(j)} - \bar{y}_{total}^{(j)})^2/n_{total}} \qquad j = 1.5,\ 13,\ 25, \tag{6.14}$$

missä n_t on testianeiston otoskoko, $y_t^{(j)}$ on testiaineiston selitettävä muuttuja ja $\hat{y}^{(j)}{}_t$ on tämän ennuste, n_{total} on kokonaisaineiston otoskoko, $y^{(j)}{}_{total}$ on kokonaisaineiston selitettävä muuttuja ja $\bar{y}^{(j)}{}_{total}$ on tämän keskiarvo, $j = 1.5,\ 13,\ 25$. Kun testiaineiston selitettäville muuttujille $y_t^{(j)}$, $j = 1.5,\ 13,\ 25$, on tehty muunnokset (2.3) ja (2.4), niin $\bar{y}_t^{(1.5)} = 0.0048$, $\bar{y}_t^{(13)} = 0.0042$, $\bar{y}_t^{(25)} = 0.0006$, $\mathrm{var}(y^{(1.5)}) = 0.0748$, $\mathrm{var}(y^{(13)}) = 0.0768$ ja $\mathrm{var}(y^{(25)}) = 0.0725$. Laskettaessa sovitteen selityskerrointa (6.11) Höskuldssonin (1988) mukaan oletetaan, että selitettävät muuttujat on keskistetty ja niiden varianssit ovat yhtäsuuria. Tämän vuoksi tehdään vielä testiaineiston selitettävälle muuttujalle $y_t^{(j)}$ ja sen ennusteelle $\hat{y}^{(j)}$ muunnos (2.4), missä nyt tarvittavat tunnusluvut ja matriisit saadaan testiaineistosta, $j = 1.5,\ 13,\ 25$. Tämän jälkeen lasketaan monimuuttujainen ennusteen

roin

$$Q^2 = 1 - \mathrm{tr}((\boldsymbol{Y}_t - \hat{\boldsymbol{Y}}_t)'(\boldsymbol{Y}_t - \boldsymbol{Y}_t))/\mathrm{tr}(\boldsymbol{Y}_t'\boldsymbol{Y}_t)\,, \tag{6.15}$$

joka siis ilmaisee, kuinka suuri osa havaintomatriisin $\boldsymbol{Y}_t$ vaihtelusta on onnistuttu estimointiaineistosta estimoidulla regressiomallilla ennustamaan.

6.2 Selittävien muuttujien multikollineaarisuus

Kuten aiemmin on jo todettu, niin säännöllistämismenetelmiä käytetään, kun selittävät muuttujat ovat multikollineaarisia. Vaikka ennalta jo tiedetään, että selittävien muuttujien välillä on lineaarikombinaatioden välisiä riippuvuuksia, niin tutkitaan estimointiaineiston selittävien muuttujien x_j, j = C,...,Bjab, multikollineaarisuuden astetta kappaleessa 2.4 esitetyin menetelmin.

1) Selittävien muuttujien neliömatriisi $\boldsymbol{X}'\boldsymbol{X}$, joka on siis korrelaatiomatriisi, on liitteessä A taulukossa A.2. Neliömatriisista $\boldsymbol{X}'\boldsymbol{X}$ nähdään, että selittävien muuttujien välillä on melko voimakkaita kahden muuttujan välisiä korrelaatiota. Näistä korrelaatiokertoimista kaksi eri korrelaatiokerrointa ovat suurempi kuin 0.7. Täten rakennematriisissa $\boldsymbol{X} = (x_C,\ldots,x_{Bjab})$ on selittäviä muuttujia, joiden multikollineaarisuuden aste on korkea. Huomattavaa on, että vaikka selittävä muuttuja x_{Bjab} on muuttujien x_B, x_{TI} ja x_N lineaarikombinaatio, niin niiden korrelaatiokertoimet ovat $\mathrm{cor}(x_{Bjab},x_B) = 0.0271$, $\mathrm{cor}(x_{TI},x_{Bjab}) = 0.7346$ ja $\mathrm{cor}(x_{Bjab},x_N) = -0.6724$. Täten selittävät muuttujat voivat olla multikollineaarisia, vaikka niiden korrelaatiokertoimet ovat pieniä.

2) Neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvot ovat liitteessä A taulukossa A.4. Vertailun vuoksi on myös laskettu vastaavat luvut testiaineistosta ja ne ovat myös taulukossa A.4. Vaikka selittävien muuttujien välillä on lineaarikombinaatioiden välisiä riippuvuuksia, niin tässä neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ pienin ominaisarvo ei ole nolla, vaan se on 0.005, mikä johtuu neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvoja approksimoitaessa käytetystä laskentatarkkuudesta. Kun tarkastellaan neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ pienintä ominaisarvoa $\lambda_p = 0.005$ vastaavan ominaisvektorin v_p alkiota, niin alkiot v_{pj}, j = Bjab, B,Ti ja N ovat suurimmat, jolloin Wangin ja Chowin (1994) mukaan selittävät muuttujat x_{Bjab}, x_{Bjab}, x_{Ti}, ja x_N, ovat eniten multikollineaarisia, kuten olla pitääkin.

3) Neliömatriisin $X'X$ suurin ominaisarvo on 4.484 ja sen pienin ominaisarvo on 0.005. Rakennematriisin X kollineaarisuuden indeksi (2.43) on 954. Tämän perusteella voidaan sanoa, että selittävät muuttujat x_j , j = C,...,Bjab, ovat kollineaarisia. Koska selittävät muuttujat ovat multikollineaarisia, niin ne ovat myös kollineaarisia.

4) Neliömatriisin $X'X$ ominaisarvojen käänteislukujen summa on 243.63, joka on 15 kertaa suurempi kuin ortogonalisen neliömatriisin $X'X$ käänteislukujen summa. Myös tämän perusteella voidaan sanoa, että selittävien muuttujien multikollineaarisuuden aste on korkea.

5) Neliömatriisin $X'X$ determinantti on 2.855×10^{-5}. Koska käytetystä laskentatarkkuudesta johtuen tämä determinantti ei ole nolla, vaikka todellisuudessa selittävien muuttujien välillä on lineaarikombinaatioiden välisiä riippuvuuksia, niin neliömatriisi $X'X$ on epäsingulaarinen eli sille on olemassa käänteismatriisi $(X'X)^{-1}$. Koska neliömatriisin $X'X$ determinantti on melko pieni, niin neliömatriisi $X'X$ on lähes singulaarinen.

6) Testisuure χ^2 (2.44) on 238.5 ja sen vapausasteet ovat 120. Koska tässä testisuureen χ^2 vapausasteet > 100, niin tehdään testisuureelle χ^2 muunnos $Z = (2\chi^2)^{½} - (2df - 1)^{½}$, missä df on testisuureen χ^2 vapausasteet (Jackson, 1991). Tämä muunnos Z noudattaa normaalijakaumaa N(0,1), jos nollahypoteesi eli selittävien muuttujien korreloimattomuus on tosi. Koska tässä Z = 6.38, niin nollahypoteesi hylätään 5% riskitasolla. Täten selittävät muuttujat x_j, j = C,...,Bjab, eivät ole korreloimattomia, vaan niiden välillä on korreloituneisuutta.

Vaikka selittävien muuttujien x_j, j = C,...,Bjab, välillä on lineaarisia kombinaatioiden välisiä riippuvuuksia, niin ominaisarvoja laskennassa käytetystä laskentatarkkuudesta johtuen neliömatriisi $X'X$ on kääntyvä. Tällöin voidaan laskea regressiomalleissa (6.3) ja (6.4) parametrien $\alpha^{(j)}$ ja $\beta^{(j)}$ pienimmän neliösumman estimaatit, mutta ne eivät ole luotettavia, koska neliömatriisin $X'X$ käänteismatriisia ei todellisesti ole olemassa, j = 1.5, 13, 25.

6.3 Pienimmän neliösumman regressio

Regressiomallin (6.3) regressiomatriisin B pienimmän neliösumman estimaattori $\boldsymbol{B}_{PNS} = (\boldsymbol{b}_{PNS}^{(1.5)}, \boldsymbol{b}_{PNS}^{(13)}, \boldsymbol{b}_{PNS}^{(25)})$ ja sen varianssi var($\boldsymbol{B}_{PNS}$) ovat liitteessä A taulukossa A.5. Regressiomalli-

en (6.3) ja (6.4) parametrin $A = (\alpha^{(1.5)}, \alpha^{(13)}, \alpha^{(25)})$ pienimmän neliösumman estimaattori $A_{PNS} = (a_{PNS}^{(1.5)}, a_{PNS}^{(13)}, a_{PNS}^{(25)})$ ja sen varianssi var(A_{PNS}) ovat liitteessä A taulukossa A.6.

Kun regressiomalli (6.3) estimoidaan pienimmän neliösumman menetelmällä, niin virhevarianssin Σ_Y estimaattori (2.31) on

$$\begin{pmatrix} 0.000162 & 0.000118 & 0.000140 \\ 0.000118 & 0.000507 & 0.000363 \\ 0.000140 & 0.000363 & 0.000506 \end{pmatrix}, \tag{6.16}$$

joka on korrelaatiomuodossa

$$\begin{pmatrix} 1.0000 & 0.4123 & 0.4909 \\ 0.4123 & 1.0000 & 0.7170 \\ 0.4909 & 0.7170 & 1.0000 \end{pmatrix}.$$

Kuten yllä olevasta matriisista nähdään, niin regressiomallien $\hat{y}_{PNS}^{(j)} = Xa_{PNS}^{(j)} + e^{(j)}$, $j = 1.5, 13, 25$, jäännökset korreloivat melko voimakkaasti. Pienimmän neliösuman menetelmällä saatujen estimointiaineiston sovitteiden ja testiaineiston ennusteiden selityskertoimien arvot ovat taulukossa 6.2.

Taulukko 6.2: Pienimmän neliösumman menetelmällä saatujen sovitteiden ja ennusteiden selityskertoimien arvot

$R^{2(1.5)}$	$R^{2(13)}$	$R^{2(25)}$	R^2	$Q^{2(1.5)}$	$Q^{2(13)}$	$Q^{2(25)}$	Q^2
0.974	0.919	0.918	0.937	0.970	0.921	0.910	0.931

Kuten taulukosta 6.2 nähdään, niin pienimmän neliösumman menetelmällä saadaan melko korkeat selityskertoimeien arvot sekä estimointiaineiston sovitteelle että testiaineiston ennusteelle. Erikoista on, että $Q^{2(13)} > R^{2(13)}$.

Koska selittävien muuttujien x_j, j = C,...,Bjab, välillä on lineaarikombinaatioden välisiä riippuvuuksia, niin todellisuudessa parametreilla $\alpha^{(j)}$ ja $\beta^{(j)}$ ei ole olemassa yksikäsitteisiä pienimmän neliösumman estimaatteja, $j = 1.5, 13, 25$. Koska käytännössä johtuen laskennassa käytetystä laskentatarkkuudesta matriisi $X'X$ on epäsingulaarinen, niin parametreille $\alpha^{(j)}$ ja $\beta^{(j)}$ voidaan laskea yksikäsitteiset pienimmän neliösumman estimaatit, $j = 1.5, 13, 25$. Koska todellisesti

käänteismatriisia $(X'X)^{-1}$ ei ole olemassa, niin parametrien $\alpha^{(j)}$ ja $\beta^{(j)}$, $j = 1.5, 13, 25$, pienimmän neliösumman estimaattien perusteella ei voida tehdä luotettavia sisällöllisiä tulkintoja estimoitavan regressiomallin rakenteesta.

Vaikka selittävien muuttujien ollessa multikollineaarisia ennusteen keskineliövirheessä $\mathrm{var}(a_{PNS})$ saattaa olla kasvanut kohtuuttoman suureksi, niin ennustaminen voidaan suorittaa kohtalaisen tarkasti, jos selittävien muuttujien multikollineaarisuusyhtälö pysyy ennustejakson aikana voimassa (Leskinen, 1977). Myös tässä esimerkkiaineistossa on saatu melko hyviä ennusteen selityskertoimia selittävien muuttujien multikolineaarisuudesta huolimatta.

6.4 Harjaregressio

Mallitettaessa teräksen karkenevuutta harjaregressiolla estimoidaan tässä regressiomalli (6.3) harjaregressiomallilla

$$\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{A}_H(K) + \boldsymbol{E}\ , \tag{6.17}$$

missä matriist $\boldsymbol{Y}$, $\boldsymbol{E}$ ja $\boldsymbol{Z}$ ovat kuten yhtälössä (6.3) ja harjaestimaattori $\boldsymbol{A}_H(\boldsymbol{K}) = (\boldsymbol{a}_H^{(1.5)}(k^{(1.5)}), \boldsymbol{a}_H^{(13)}(k^{(13)}), \boldsymbol{a}_H^{(25)}(k^{(25)}))$ on

$$\boldsymbol{a}_H^{(j)}(k^{(j)}) = \frac{u^{(j)}\sqrt{\Lambda_p}}{\Lambda_p + k^{(j)}I_p}, \qquad j = 1.5,\ 13,\ 25\ , \tag{6.18}$$

ja $\boldsymbol{u}^{(j)}$ ja Λ_p on määritelty kuten kanonisessa yhtälössä (6.4). Jotta harjaestimaattori (6.18) voidaan laskea, niin on valittava harjaparametrin $k^{(j)}$, $j = 1.5, 13, 25$, arvot. Valitaan harjaestimaattoriin (6.18) harjaparametrien $k^{(j)}$, $j = 1.5, 13, 25$, arvot kappaleissa 3.5 ja 3.7 esitettyjen valintamenetelmien mukaan. Valitut harjaparametrien arvot ovat taulukossa 6.3. Harjaestimaattori $\boldsymbol{A}_H(\boldsymbol{K}_{ad})$ ja sen varianssi ovat liitteessä A taulukossa A.6. Harjaestimaattorit $\boldsymbol{A}_H(\boldsymbol{K}_{FF})$ ja $\boldsymbol{A}_H(\boldsymbol{K}_{cross})$ ja niiden variansit ovat liitteessä A taulukossa A.7.

Taulukossa 6.3 kaikissa muissa harjaparametrien $k^{(j)}$, $j = 1.5, 13, 25$, valintamenetelmissä kuin valintamenentelmässä $\boldsymbol{K}_{cross}$ valitaan harjaparametrien $k^{(j)}$, $j = 1.5, 13, 25$, arvot toisistaan riippumatta. Tällöin ei hydynnetä selitettävien muuttujien $y^{(1.5)}$, $y^{(13)}$ ja $y^{(25)}$ yhteisinformaatiota, vaan on estimoitu kolme yksimuuttujaista harjaestimaattoria $a_H^{(j)}(k^{(j)})$, $j = 1.5, 13, 25$.

Harjaregressiolla saadut estimointiaineiston sovitteen ja testiaineiston ennusteen selityskertoimien arvot ovat taulukossa 6.4.

Taulukko 6.3: Harjaestimaattoriin (6.18) valitut harjaparametrien $k^{(j)}$, j = 1.5, 13, 25, arvot

k	$k^{(1.5)}$	$k^{(13)}$	$k^{(25)}$
$k_{jälki}$	0.03	0.07	0.09
k_{cross}	0.0046	0.0106	0.0068
k_{gross}	0.0037	0.0045	0.0037
k_{HKB}	0.0023	0.0050	0.0035
k_{MUR}	0.0013	0.0050	0.0103
k_{MLW}	0.0023	0.0077	0.0077
$\boldsymbol{K}_{cross}$	0.0072	0.0072	0.0072

Taulukko 6.4: Harjaregressiolla saatujen sovitteiden ja ennusteiden selityskertoimien arvot

k	$R^{2\,(1.5)}$	$R^{2\,(13)}$	$R^{2\,(25)}$	R^2	$Q^{2\,(1.5)}$	$Q^{2\,(13)}$	$Q^{2\,(25)}$	Q^2
$k_{jälki}$	0.971	0.900	0.883	0.920	0.971	0.901	0.880	0.916
k_{cross}	0.974	0.919	0.917	0.937	0.971	0.922	0.911	0.932
k_{gcross}	0.974	0.919	0.917	0.937	0.971	0.923	0.911	0.932
k_{HKB}	0.974	0.919	0.917	0.937	0.971	0.923	0.911	0.932
k_{MUR}	0.974	0.919	0.916	0.937	0.971	0.923	0.911	0.932
k_{MLW}	0.974	0.919	0.917	0.937	0.971	0.922	0.911	0.932
k_{ad}	0.974	0.918	0.917	0.937	0.971	0.921	0.908	0.931
$\boldsymbol{K}_{FF}$	0.974	0.919	0.918	0.937	0.970	0.922	0.910	0.931
$\boldsymbol{K}_{cross}$	0.974	0.919	0.917	0.937	0.971	0.923	0.911	0.932
$\boldsymbol{K} = \boldsymbol{0}$[a]	0.974	0.919	0.918	0.937	0.970	0.921	0.910	0.931

[a] Valinta $\boldsymbol{K} = \boldsymbol{0}$ vastaa regressiomallin (6.3) pienimmän neliösumman menetelmän ratkaisua.

Kun harjaparametrien $k^{(j)}$, j = 1.5, 13, 25, arvot ovat $k^{(1.5)} = 0.03$, $k^{(13)} = 0.07$ ja $k^{(25)} = 0.09$, niin harjaestimaattoreiden $a_H^{(j)}(k^{(j)})$, j = 1.5, 13, 25, arvot näyttivät vakiintuneen ilman, että sovitteiden jäännösneliösummat olisivat kasvaneet kohtuuttomasti. Kun verrataan valintoja $k_{jälki}^{(j)}$,

$j = 1.5, 13, 25$, muihin harjaparametrien $k^{(j)}, j = 1.5, 13, 25$, valintoihin, niin harjajäljellä valitut harjaparmetrien arvot ovat suuria. Kun verrataan taulukossa 6.4 olevien sovitteiden ja ennusteiden selityskertoimien arvoja, niin harjajäljellä valitut harjaparametrien $k^{(13)}$ ja $k^{(25)}$ arvot antavat huonoimmat sovitteiden ja ennusteiden selityskertoimien arvot, minkä perusteella voidaan olettaa, että harjajäljellä valitut harjaprametrien $k^{(13)}$ ja $k^{(25)}$ arvot ovat epäonnistuneet.

Kun harjaparametrien $k^{(j)}, j = 1.5, 13, 25$, arvot valitaan *cross-validation* -menetelmällä, niin laskettaessa ennusteen virheneliösummaa PRESS käytetään apuna Wangin ja Chowin (1994) esittämää lauseketta

$$\text{PRESS} = \sum_{i=1}^{n} \frac{e_i^2}{(1 - h_{ii})^2}, \tag{6.19}$$

missä $\text{diag}(\boldsymbol{H}) = \text{diag}(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}') = \text{diag}(h_{ii})$, $i = 1,...,n$, jolloin ennusteen virheneliösumma $\text{PRESS}(k^{(j)})$ voidaan kirjoittaa muotoon

$$\text{PRESS}(k^{(j)}) = \sum_{1=1}^{n} \frac{(e_i^{(j)})2}{(1 - h_{ii}^{(j)})^2} \qquad j = 1.5, 13, 25, \tag{6.20}$$

missä $\boldsymbol{e}^{(j)} = \boldsymbol{y}^{(j)} - \boldsymbol{Z}\boldsymbol{a}_H^{(j)}(k^{(j)})$, $\text{diag}(\boldsymbol{H}) = \text{diag}(\boldsymbol{Z}(\boldsymbol{\Lambda}_p + k\boldsymbol{I}_p)^{-1}\boldsymbol{Z}') = \text{diag}(h_{ii})$, $i = 1,...,n$, ja matriisit $\boldsymbol{Z}$ ja $\boldsymbol{\Lambda}_p$ on määritelty kuten yhtälössä (6.3). Vaikka valittaessa harjaparametrin k arvoa *cross-validation* -menetelmällä valitaan harjaparametrin k arvo siten, että sillä saadaan pienin ennusteen virheneliösumma (6.20), niin harjaparametrin $k^{(13)}$ valinta $k_{cross}^{(13)} = 0.0106$ ei anna parasta ennusteen selityskertoimen $Q^{2(13)}$ arvoa. Tämä saattaa johtua siitä, että kun $0.0068 \leq k^{(13)} \leq 0.0106$, niin $\text{PRESS}(k^{(13)}) \approx 0.1035$ ja täten valinta $k^{(13)} = 0.0106$ saattaa olla liian suuri.

Estimointiaineistossa neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ ominaisarvojen välillä on suuria eroja ja täten kollineaarisuuden indeksi on suuri. Koska epäyhtälö (3.34) on epätosi, niin $\text{MSE}(\boldsymbol{a}_H^{(j)}(k_{HKB}^{(j)})) \geq \text{MSE}(\boldsymbol{a}_{PNS}^{(j)}), j = 1.5, 13, 25$. Koska epyhtälö (3.35) on epätosi, niin $\text{MSE}(\hat{y}_{0HKB}^{(j)}) \geq \text{MSE}(\hat{y}_{0PNS}^{(j)})$, $j = 1.5, 13, 25$. Koska kollineaarisuuden indeksi on suuri, niin voidaan olettaa, että valinnat $k_{HKB}^{(j)}$, $j = 1.5, 13, 25$, ovat liian pieniä ja harjaparametrin k valintamenetelmä k_{HKB} ei sovi käytettyyn estimointiaineistoon. Koska $k_{MUR}^{(j)} \leq k_{HKB}^{(j)}, j = 1.5, 13$ ja $k_{gcross}^{(13)} < k_{HKB}^{(13)}$, niin voidaan olettaa, että myös valinnat $k_{MUR}^{(1.5)}$, $k_{MUR}^{(1.5)}$ ja $k_{gcross}^{(13)}$ ovat liian pieniä. Koska tässä aineistossa kollineaarisuuden indeksi on suuri, niin Brownin (1993) mukaan harjaparametrin k valintaa k_{MLW} tulee pitää parempana kuin valintaa k_{HKB}. Valinnat $k_{MLW}^{(j)}, j = 13, 25$, ovatkin suurempia kuin $k_{HKB}^{(j)}, j = 13, 25$, ja $k_{HKB}^{(1.5)} = k_{MLW}^{(1.5)}$, mutta $Q^{2(13)}_{MLW} = 0.922 < Q^{2(13)}_{HKB} = 0.923$.

parempana kuin valintaa k_{HKB}. Valinnat $k_{MLW}^{(j)}$, $j = 13, 25$, ovatkin suurempia kuin $k_{HKB}^{(j)}$, $j = 13, 25$, ja $k_{HKB}^{(1.5)} = k_{MLW}^{(1.5)}$, mutta $Q^{2(13)}{}_{MLW} = 0.922 < Q^{2(13)}{}_{HKB} = 0.923$.

Kun harjaparametrimatriisin $\boldsymbol{K}$ estimaattori valitaan Brownin ja Paynen (1975) mukaan, niin ongelmana on, että parametri $\sigma_\beta{}^2$ on tässä tuntematon. Parametrin $\sigma_\beta{}^2$ estimaattori on $\sigma_Y{}^2/k$ (Brown & Payne, 1975). Koska tässä on myös harjaparametri k tuntematon, niin tässä tutkielmassa tätä harjaparametrimatriisin $\boldsymbol{K}$ valintamenetelmää ei käytetä.

Kun harjaparametrimatriisin $\boldsymbol{K}$ valitaan Scloven säännön (3.52) mukaan, niin tässä epäyhtälö (3.53) on epätosi. Tällöin $\mathrm{MSE}(\boldsymbol{A}_{PNS}) < \mathrm{MSE}(\boldsymbol{A}_H(\boldsymbol{K}_{Sclove}))$. Täten harjaparametrimatriisin $\boldsymbol{K}$ valintamentelmä Scloven säännön mukaan ei sovi tässä käytettyyn estimointiaineistoon.

Kun harjaparametrimatriisin $\boldsymbol{K}$ valitaan Hoerlin, Kennardin ja Baldwinin säännön mukaan, niin tässä epäyhtälö (3.57) on epätosi, minkä mukaan $\mathrm{MSE}(\boldsymbol{A}_{PNS}) < \mathrm{MSE}(\boldsymbol{A}_H(\boldsymbol{K}_{HKB}))$. Täten harjaparametrimatriisin $\boldsymbol{K}$ valintamenetelmä Hoerlin, Kennardin ja Baldwinin säännön mukaan ei myöskään sovi käytettyyn estimointiaineistoon.

Kun verrataan pienimmän neliösumman menetelmällä ja harjaregressiolla saatuja sovitteiden selityskertoimien (6.9) ja (6.11) arvoja keskenään, niin harjaregressiolla saadut sovitteiden selityskertoimien arvot eivät ole juurikaan pienempiä kuin pienimmän neliösumman menetelmällä saadut sovitteiden selityskertoimien arvot. Täten harjaregressiossa pienimmän neliösumman estimaattoreissa $\boldsymbol{b}_{PNS}^{(j)}$, $j = 1.5, 13, 25$, harjaparametrin k lisääminen neliömatriisin $\boldsymbol{X}'\boldsymbol{X}$ diagonaalille ei juurikaan kasvata sovitteiden jäännösneliösummia, sillä valitut harjaparametrit olivat hyvin pieniä. Toisaalta harjaregressiolla on saatu keskimäärin hiukan parempia ennusteen selityskertoimien (6.14) ja (6.15) arvoja kuin pienimmän neliösumman menetelmällä, joten kun harjaparametri k on valittu sopivasti, niin harjaregressiolla saadaan hieman tarkempia ennusteita kuin pienimmän neliösumman menetelmällä.

Valitaan harjaestimaattoriin (6.17) sopivimpana harjaparametrimatriisin $\boldsymbol{K}$ valintamenetelmä valintamenetelmä $\boldsymbol{K}_{CROSS}$. Valintamenetelmällä $\boldsymbol{K}_{CROSS}$ valittu harjaparametrin $k^{(j)}$ arvo on suurempi kuin $k_{HKB}^{(j)}$, joka saattoi olla liian pieni, koska selittävien muuttujien kollineaarisuuden indeksi on suuri, $j = 1.5, 13, 25$. Kun valitaan harjaestimaattoriin (6.17) harjaparametrimatriisi $\boldsymbol{K}$ valintamenetelmällä $\boldsymbol{K}_{CROSS}$, niin tällöin saadut sovitteiden ja ennusteiden selityskertoimien arvot ovat hyviä, kun niitä verrataan muihin sovitteiden ja ennusteiden selityskertoimien arvoihin taulukossa 6.4

6.5 Pääkomponenttiregressio

Kun teräksen karkenevuutta mallitetaan pääkomponenttiregressiolla, niin approksimoidaan regressiomalleissa (6.3) ja (6.4) parametreja $\alpha^{(j)}$, $j = 1.5, 13, 25$, pääkomponenttiestimaattoreilla

$$a_{iAPK}^{(j)} = f_i^{(j)} a_{iPNS}^{(j)}, \qquad i = 1,...,16, \quad j = 1.5, 13, 25, \tag{6.21}$$

missä $f_i^{(j)} = 1$, kun $i \in S_A^{(j)}$, $f_i^{(j)} = 0$, kun $i \notin S_A^{(j)}$, $j = 1.5, 13, 25$. Pienimmän neliösumman estimaattorit $a^{(j)}_{PNS}$, $j = 1.5, 13, 25$, ovat liitteessä A taulukossa A.6. Pääkomponenttiestimaattorin $a^{(j)}_{APK}$ kertoimet $f^{(j)}$, $j = 1.5, 13, 25$, kun valittujen pääkomponenttien indeksijoukot $S_A^{(j)}$, $j = 1.5, 13, 25$, on valittu kappaleissa 4.4 ja 4.6 olevien valintamenetelmien mukaan, ovat taulukossa 6.5

Kun pääkomponenttiregressio perustuu rakennematriisin X korrelaatiomatriisiin, niin rakennematriisin X selityskerroin R^2_X saadaan yhtälöstä

$$R_X^2 = \sum_{j \in S_A} \lambda_j / r, \tag{6.22}$$

missä r on rakennematriisin X aste. Pääkomponenttiregressiossa rakennematriisin selityskerroin R^2_X kertoo, kuinka paljon rakennematriisin X kokonaisvaihtelusta valitut pääkomponentit z_j, $j \in S_A$, selittävät eli kuinka paljon pääkomponenttiregressiossa rakennematriisin X vaihtelusta on käytetty selittämään selitettävää muuttujaa y. Rakennematriisin X selityskertoimien arvot ovat taulukossa 6.6.

Kun valittujen pääkomponenttien indeksijoukko S_A valitaan *cross-validation* menetelmällä, niin laskettaessa ennusteen virheneliösummia (4.25) ja (4.31) käytetään apuna Wangin ja Chowin (1994) lauseketta (6.19), jolloin ennusteen virheneliösumma PRESS($S_A^{(j)}$) voidaan kirjoittaa muotoon

$$\text{PRESS}\,(S_A^{(j)}) = \sum_{i=1}^{n} \frac{(e_i^{(j)})^2}{(1 - h_{ii}^{(j)})^2}, \qquad j = 1.5, 13, 25, \tag{6.23}$$

missä $e^{(j)} = y^{(j)} - Z_A^{(j)} a^{(j)}_{APK}$, $H^{(j)} = \text{diag}(Z_A^{(j)}(Z_A^{(j)\prime} Z_A^{(j)})^{-1} Z_A^{(j)\prime}) = \text{diag}(h^{(j)}_{11},...,h^{(j)}_{nn})$, $Z_A^{(j)} = (z_i^{(j)},)$ ja $i \in S_A^{(j)}$, $j = 1.5, 13, 25$. Kuvassa 6.1 on lasketut ennusteen virheneliösummat PRESS($S_A^{(j)}$), $j = 1.5, 13, 25$, ja PRESS($S_A^{(I)}$)/3, missä $S_A^{(I)} = S_A^{(1,5)} = S_A^{(13)} = S_A^{(25)}$, valittujen pääkomponenttien lukumäärän funktiona.

Taulukko 6.5: Pääkomponenttiestimaattorin (6.12) kertoimet $f^{(j)}$, $j = 1.5, 13, 25$

S_A	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
$S_{A(4.20)}^{(1.5)}$	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0
$S_{A(4.20)}^{(13)}$	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1
$S_{A(4.20)}^{(25)}$	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1
$S_{A(4.21)}^{(1.5)}$	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
$S_{A(4.21)}^{(13)}$	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
$S_{A(4.21)}^{(25)}$	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
$S_{A(4.24)}^{(1.5)}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$S_{A(4.24)}^{(13)}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$S_{A(4.24)}^{(25)}$	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
$S_{A(4.25)}^{(1.5)}$	1	1	1	1	1	1	0	1	1	1	0	0	1	1	1	0
$S_{A(4.25)}^{(13)}$	1	1	1	1	1	1	0	1	1	1	0	1	1	1	0	1
$S_{A(4.25)}^{(25)}$	1	1	1	1	1	1	1	0	0	1	0	0	1	1	1	1
$S_{A(4.26)}^{(1.5)}$	1	1	1	1	1	1	0	0	0	1	0	0	1	1	1	0
$S_{A(4.26)}^{(13)}$	1	1	1	0	1	0	0	0	0	1	0	0	1	1	0	0
$S_{A(4.26)}^{(25)}$	1	1	1	0	1	0	0	0	0	1	0	0	1	1	1	0
$S_{A(4.31)}$	1	1	1	1	1	1	0	0	1	1	0	0	1	1	1	0

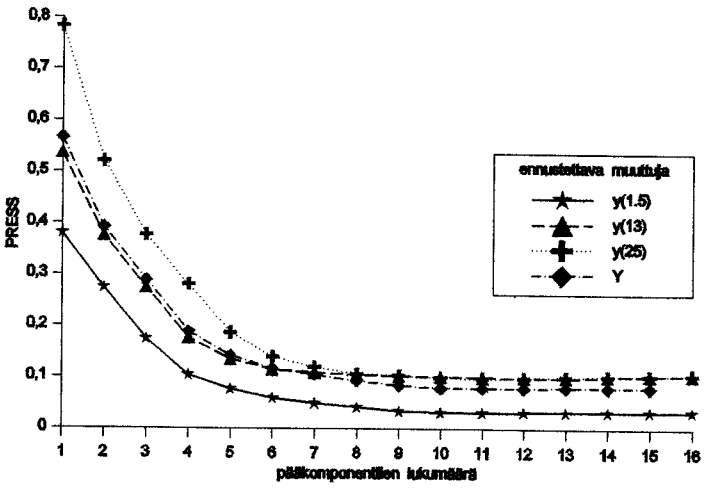
Valittaessa pääkomponenettien indeksijoukkoa S_A *cross-validation* -menetelmällä vaarana voi olla pääkomponenttiregressiomallin ylisovittaminen ts. liian monen pääkomponentin valitseminen indeksijoukkoon S_A (Jackson, 1991). Tämän vuoksi indeksijoukko S_A on valittu myös valintamenetelmän (4.36) mukaan. Kun verrataan indeksijoukkoja $S_{A(4.35)}$ ja $S_{A(4.36)}$, niin valittujen pääkomponenttien lukumäärä on indeksijoukoissa $S_{A(4.25)}^{(1.5)} = 12$, $S_{A(4.26)}^{(1.5)} = 10$, $S_{A(4.25)}^{(13)} = 13$, $S_{A(4.26)}^{(13)} = 7$, $S_{A(4.25)}^{(25)} = 12$ ja $S_{A(4.26)}^{(25)} = 8$.

Pääkomponenttiregressiolla saatujen estimointiaineiston sovitteiden ja testiaineston ennusteiden selityskertoimien arvot ovat taulukossa 6.7

Taulukko 6.6: Pääkomponenttiregressiossa rakennematriisin X selityskertoimien arvot ja valittujen pääkomponenttien lukumäärä A

S_A	$R^2_X{}^{(1.5)}$	$A^{(1.5)}$	$R^2_X{}^{(13)}$	$A^{(13)}$	$R^2_X{}^{(25)}$	$A^{(25)}$
$S_{A(4.20)}$	0.913	13	0.906	13	0.852	12
$S_{A(4.21)}$	0.779	7	0.779	7	0.779	7
$S_{A(4.24)}$	1.000	16	1.000	16	1.000	16
$S_{A(4.25)}$	0.885	12	0.913	13	0.852	12
$S_{A(4.26)}$	0.800	10	0.636	7	0.646	8
$S_{A(4.31)}$	0.842	11	0.842	11	0.842	11
$S_p{}^A$	1.000	16	1.000	16	1.000	16

[a] Pienimmän neliösumman menetelmä



Kuva 6.1: *Cross-validation* -menetelmässä lasketut ennusteen virheneliösummat PRESS($S_A^{(j)}$), j = 1.5, 13, 25, ja PRESS($S_A^{(Y)}$)/3 valittujen pääkomponentien z_j, $j \in S_A^{(j)}$, lukumäärän funktiona

Taulukko 6.7: Pääkomponenttiregressiolla saatujen sovitteiden ja ennusteiden selityskertoimien arvot

S_A	$R^{2(1.5)}$	$R^{2(13)}$	$R^{2(25)}$	R^2	$Q^{2(1.5)}$	$Q^{2(13)}$	$Q^{2(25)}$	Q^2
$S_{A(4.20)}$	0.974	0.918	0.917	0.936	0.970	0.921	0.910	0.931
$S_{A(4.21)}$	0.781	0.665	0.553	0.669	0.830	0.679	0.593	0.689
$S_{A(4.24)}$	0.974	0.919	0.918	0.937	0.970	0.921	0.910	0.931
$S_{A(4.25)}$	0.973	0.918	0.917	0.935	0.972	0.915	0.905	0.928
$S_{A(4.26)}$	0.973	0.902	0.905	0.927	0.972	0.909	0.899	0.924
$S_{A(4.31)}$	0.973	0.912	0.914	0.934	0.973	0.923	0.904	0.931
S_P [a]	0.974	0.919	0.918	0.937	0.970	0.921	0.910	0.931

[a] Pienimmän neliösumman menetelmä

Kun verrataan pääkomponenttiregressiossa saatuja sovitteiden selityskertoimien $R^{2(j)}$, j = 1.5, 13, 25, ja R^2 ja ennusteiden selityskertoimien $Q^{2(j)}$, j = 1.5, 13, 25, ja Q^2 arvoja taulukossa 6.7, niin näiden selityskertoimien arvot erovat melkoisesti eri indeksijoukon S_A valintamenetelmien mukaan. Teräksen karkenevuusaineistossa saadaan pääkomponenttiregressiolla parhaimmat selityskertoimien R^2, Q^2, $R^{2(j)}$ ja $Q^{2(j)}$ j = 1.5, 13, 25, arvot, kun pääkomponenttiestimaattorissa (6.21) valitaan indeksijoukko S_A valintamenetelmillä $S_{A(4.20)}$, $S_{A(4.24)}$, $S_{A(4.31)}$ ja pienimmän neliösummman menetelmällä eli indeksijoukolla S_p. Kun näistä valittujen pääkomponenttien indeksijoukon S_A valintamenetelmistä valitaan sopivin pääkomponenttiestimaattoriin (6.21), niin ennusteiden seltyskertoimien arvojen perusteella valitaan valintamenetelmä $S_{A(4.20)}$. Valintamenetelmä $S_{A(4.31)}$ ei valittu, koska ennusteen selityskertoimen $Q^{(25)}$ arvo on pieni verratuna valinatamenetelmällä $S_{A(4.20)}$ saatuun ennusteen seltyskertoimen $Q^{(25)}$ arvoon.

6.6 Osittainen pienimmän neliösumman regressio

Kun teräksen karkenevuutta mallitetaan osittaisella pienimmän neliösumman regressiolla eli PLS -regressiollla, niin malli on

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{X}_{A+1} \quad \text{ja} \quad \boldsymbol{Y} = \boldsymbol{T}\boldsymbol{C}' + \boldsymbol{Y}_{A+1}, \tag{6.24}$$

missä matriisit X ja Y ovat kuten regressiomallissa (6.3), matriisit $T = (t_1,...,t_A)$, $C = (c_1,...,c_A)$ ja $P = (p_1,...,p_A)$ saadaan PLS-algoritmin mukaan, matriisit X_{A+1} ja Y_{A+1} ovat jäännösmatriiseja ja valittujen PLS-komponenttiparien lukumäärä on A (ks. kappale 5.1). Mallissa (6.24) kovarianssimatriisin $X'Y$ selityskerroin R^2_{XY} saadaan yhtälöstä

$$R^2_{X'Y} = 1 - \mathrm{tr}(X'_{A+1}Y_{A+1}Y'_{A+1}X_{A+1})/\mathrm{tr}(X'YY'X) \qquad (6.25)$$

(Höskuldsson, 1988). Tämä selityskerroin R^2_{XY} kertoo, kuinka paljon malli (6.24) on selittänyt kovarianssimatriista $X'Y$. Rakennematriisin X selityskerroin R^2_X saadaan yhtälöstä

$$R^2_X = 1 - \mathrm{tr}(X'_{A+1}X_{A+1})/\mathrm{tr}(X'X) \qquad (6.26)$$

(vrt. (6.22)) (Höskuldsson, 1988). Tämä selityskerroin R^2_X kertoo, kuinka paljon rakennematriisin X varianssista on mallissa (6.24) käytetty selittämään havaintomatriisia Y. PLS-komponenttien lukumäärä A kappaleen 5.5 valintamenetelmin sekä vastaavat kovarianssi- ja rakennematriisin selityskertoimet R^2_{XY} ja R^2_X ovat taulukossa 6.8.

Taulukko 6.8: PLS-regressiomallin PLS-komponenttien lukumäärä A sekä vastaavat kovarianssi- ja rakennematriisin selityskertoimien arvot

Valintamenetelmä	PLS-komp. lkm. A	R^2_X	R^2_{XY}
A_{cross}	9	0.773	1.000
$A_{\|Y\|}$	12	0.875	1.000
S_p [a]	16	1.000	1.000

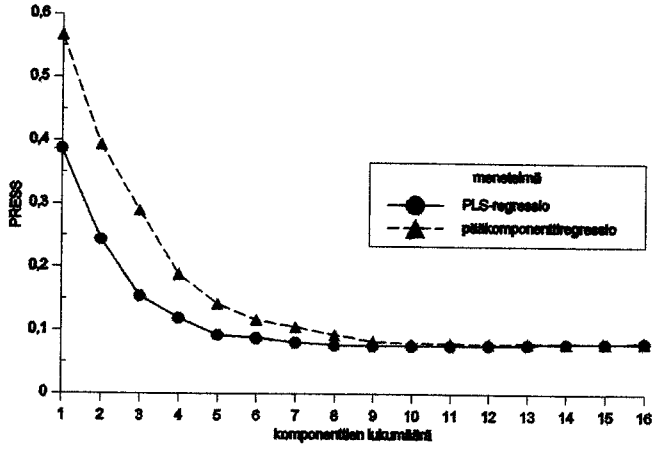
[a] Asetettu PLS-komponenttien lukumääräksi 16

PLS-regressiolla saatujen estimointiaineiston sovitteiden ja testiaineiston ennusteiden selityskertoimien arvot ovat taulukossa 6.9.

Kun regressiomalli (6.3) PLS-regressiolla ja tarvittavien PLS-komponenttien lukumäärä valitaan *cross-validation* menetelmällä, niin käytetään Wangin ja Chowin (1994) lauseketta (6.19) ennusteen virheneliösummalle, jolloin

$$\text{PRESS}(A) = \sum_{j=1}^{3} \sum_{i=1}^{176} \frac{y'_{A+1ij} y_{A+1ij}}{(1 - h_{ii})^2}, \qquad (6.27)$$

missä y_{A+1ij} on havainnon y_{ij} jäännös ja $\boldsymbol{H}$ on projektiomatriisi (5.73) PLS-algoritmin askeleessa A ja diag($\boldsymbol{H}$) = diag(h_{ii}), i = 1,...,176. PLS-regressiolla ja pääkomponenttiregressiolla saatujen ennusteiden virheneliösummat, jotka on jaettu selitettävin muuttujien lukumäärällä kolme, ovat valittujen komponenttien lukumäärän A funktiona kuvassa 6.2.



Kuva 6.2: PLS-regressiolla ja pääkomponenttiregressiolla saatujen ennusteiden virheneliösummat

Taulukko 6.9: PLS-regressiolla saatujen sovitteiden ja ennusteiden selityskertoimien arvot

A	$R^{2(1.5)}$	$R^{2(13)}$	$R^{2(25)}$	R^2	$Q^{2(1.5)}$	$Q^{2(13)}$	$Q^{2(25)}$	Q^2
A_{CROSS}	0.972	0.917	0.911	0.933	0.912	0.917	0.841	0.902
$A_{\|\|F\|\|}$	0.974	0.918	0.917	0.937	0.970	0.924	0.913	0.934
S_p [a]	0.974	0.919	0.918	0.937	0.970	0.921	0.910	0.931

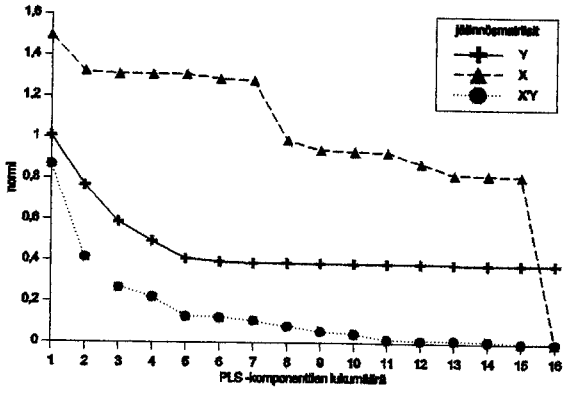
[a] Asetettu PLS-komponenttien lukumääräksi 16

Kun verrataan PLS-regressiolla saatua ennusteen virheneliösummaa (6.27) monimuuttujaisella pääkomponenttiregressiolla saatuun ennusteen virheneliösummaan (6.20), niin edellinen pienenee nopeammin kuin jälkimmäinen komponenttien lukumäärän A funktiona. PLS-regressios-

sa saavutetaan ennusteen virheneliösumman minimi yhdeksällä komponentilla, kun vastaavasti pääkomponenttiregressiossa saavutetaan ennusteen virheneliösumman minimi 11 komponentilla. Täten PLS-regressiomallin ulottuvuus on pienempi kuin pääkomponenttiregressiomallin ulottuvuus, kun komponentit on valittu *cross-validation* -menetelmällä.

Kun PLS-komponenttien lukumäärä valitaan *cross-validation* -menetelmällä, niin $R^{2(25)} = 0.911$, $Q^{2(1.5)} = 0.912$ ja $Q^{2(25)} = 0.841$ ovat pieniä verrattuna muilla menetelmillä saatuihin vastaaviin selityskertoimien arvoihin. Täten PLS-regressiomalli, jossa PLS-komponenttien lukumäärä valitaan *cross-validation* -menetelmällä, ei anna parasta testiaineiston ennustetta.

Jäännösmatriisien $\boldsymbol{Y}_{A+1}$, $\boldsymbol{X}_{A+1}$ ja $\boldsymbol{X}_{A+1}'\boldsymbol{Y}_{A+1}$ normit PLS-komponenttien lukumäärän A funktiona ovat kuvassa 6.3. Koska PLS-algoritmi perustuu epäyhtälöön (5.65), niin $||\boldsymbol{X}_A'\boldsymbol{Y}_A|| \leq ||\boldsymbol{X}_{A+1}'\boldsymbol{Y}_{A+1}||$ (ks. kuva 6.3).



Kuva 6.3: PLS-regressiossa jäännösmatriisien $\boldsymbol{Y}_{A+1}$, $\boldsymbol{X}_{A+1}$ ja $\boldsymbol{X}_{A+1}'\boldsymbol{Y}_{A+1}$ normit PLS-komponenttien lukumäärän A funktiona.

Kuten kuvasta 6.3 nähdään, niin normi $||\boldsymbol{X}_{A+1}'\boldsymbol{Y}_{A+1}||$ vakiintuu PLS-komponenttien lukumäärällä 6. Normi $||\boldsymbol{X}_{A+1}||$ vakiintuu PLS-komponenttien lukumäärällä 9, paitsi $||\boldsymbol{X}_{A+1}|| = 0$, kun $A = 16$. Valitaan kuvan 6.3 perusteella PLS-komponenttien lukumääräksi $A = 12$, jolloin normi $||\boldsymbol{Y}_{A+1}||$ on vakiintunut ja tällöin myös normi $||\boldsymbol{X}_{A+1}'\boldsymbol{Y}_{A+1}||$ on vakiintunut.

Kun PLS-regressiossa valitaan PLS-komponenttien lukumääräksi A selittävien muuttujien lukumäärä p, niin $\boldsymbol{X}_{p+1} = \boldsymbol{0}$. Kun mallissa (6.24) PLS-komponenttien lukumääräksi on valittu $A = 12$, niin saadut sovitteiden selityskertoimien arvot ovat lähes yhtä suuret kuin PLS-komponettien lukumäärällä $A = 16$ saadut sovitteiden selityskertoimien arvot. Koska PLS-komponenttien lukumäärällä $A = 12$ saadaan hiukan suuremmat testiaineiston ennusteiden selityskertoimien arvot

kuin valinnalla $A = 16$, niin PLS-komponenttien lukumäärällä $A = 16$ ylisovitetaan mallia (6.24).

Kun verrataan taulukossa 6.9 kaikkia PLS-regressiolla saatuja sovitteiden ja ennusteiden selityskertomien arvoja, niin valitaan PLS-komponenttien lukumäärän A valintamenetelmäksi jäännösmatriisin $\boldsymbol{Y}_{A+1}$ normin graafinen tarkastelu, koska tällä valintamenetelmällä saadaan parhaimmat testiaineiston ennusteiden selityskertoimien arvot. Täten valitaan mallin (6.24) PLS-komponenttien lukumääräksi $A = 12$.

7 Yhteenveto

Tässä tutkielmassa tehtävänä oli muodostaa monimuuttujainen regressiomalli teräksen karkenevuudelle. Muodostettava regressiomalli oli monimuuttujainen, koska Jominy-kokeessa karkaistaessa terästä sen kovuuteen vaikuttaa etäisyys terässauvan alapäästä ja tavoitteena oli mallittaa teräksen karkenevuutta eri etäisyyksiltä koesauvan alapäästä. Muodostettavassa regressiomallissa selittävinä muuttujina olivat teräksen seosaineiden pitoisuudet, jotka olivat multikollineaarisia, sillä selittävä muuttuja x_{Bjab} oli seosaineiden B, Ti ja N pitoisuuksien linearikombinaatio. Tällöin seosaineiden pitoisuuksien rakennematriisin neliömatriisin käänteismatriisia ei ole olemassa, mutta laskusuorituksissa olevasta pyöristysvirheistä johtuen, tämä käänteismatriisi saatiin. Koska laskettaessa regressiomallin regressiokertoimien pienimmän neliösumman estimaatteja käytettiin tätä teräksen seosaineiden pitoisuuksien neliömatriisin käänteismatriisin approksimaatiota, niin regressiokertoimien pienimmän neliösumman estimaatit eivät ole luotettavia.

Harjaregressiolla, pääkomponenttiregressiolla ja osittaisella pienimmän neliösumman regressiolla approksimoitiin paremmin kääntyviä teräksen seosaineiden pitoisuuksien rakennematriisin neliömatriiseja. Näillä menetelmillä muodostettuja teräksen karkenevuuden regressiomallien riittävyyttä arvioitiin sovitteen ja ennusteen selityskertoimien avulla. Kun estimoitu teräksen karkenevuuden regressiomalli sovitettiin estimointiaineistoon eli havaintoihin, joista regressiokertoimen estimaatit oli laskettu, niin regressiomallin sovitteen selityskertoimen arvoksi saatiin pienimmän neliösumman menetelmällä 0.937. Tämä sovitteen selityskerroin on suurin mahdollinen saaduista sovitteen selityskertoimista, koska pienimmän neliösumman menetelmä minimoi sovitteen ja havaintojen välisen jäännösneliösumman. Kun estimoitu teräksen karkenevuuden regressiomalli oli sovitettu testiaineistoon, jonka havainnot eivät olleet mukana estimoitaessa regressiokertoimien estimaatteja, niin ennusteen suurimmat selityskertoimien Q^2 arvot olivat pienimmän neliösumman menetelmällä $Q^2 = 0.931$, harjaregressiolla $Q^2 = 0.932$, pääkomponenttiregressiolla $Q^2 = 0.932$ ja PLS-regressiolla $Q^2 = 0.934$. Vaikka säännöllistämismenetelmät regressioanalyysissä perustuvat regressiokertoimien estimaattoreiden keskineliövirheen minimointiin ja teräksen karkenevuuden regressiomallissa regressiokertoimien arvot olivat tuntemattomia, niin tästä huolimatta säännöllistämismenetelmillä pystyttiin kutistamaan regressiokertoimien pienimmän neliösumman estimaattoria kohti origoa ja vähentämään regressiomallin ulottuvuutta siten, että säännöllistämismenetelmillä saatiin

keskimäärin hiukan tarkempia selitettävien muuttujien ennusteita havaitsemattomissa otospisteissä kuin pienimmän neliösumman menetelmällä. Koska PLS-regressiolla saatiin suurin ennusteen selityskertoimen Q^2 arvo, niin teräksen karkenevuuden regressiomalliksi valitaan malli (6.24), missä PLS-komponenttien lukumäärä on 12, joka on pienempi kuin selittävien muuttujien lukumäärä 16. Koska tällä mallilla on suurin ennusteen selityskertoimen Q^2 arvo, niin sillä on Copasin (1983) mukaan paras validiteetti muodostetuista teräksen karkenevuuden regressiomalleista.

Lähteet

Brown, P. J. & Payne C. (1975): Election Night Forecasting. *Journal of the Royal Statistical Society* **A 4**, 463 - 482.

Brown, P. J. & Zidek J.V. (1980): Adaptive Multivariate Ridge Regression. *The Annals of Statistics* **8**, 64 - 74.

Brown, P. J. & Zidek J. V. (1982): Multivariate Regression Shrinkage Estimators with Unknown Covariance Matrix. *Scandinavian Journal of Statistical* **9**, 209 - 215.

Brown, P. J. (1993): *Measurement, Regression, and Calibration.* Oxford: Clarendoon Press.

Copas, J. B. (1983): Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society* **B 3**, 311 - 354.

Draper, N.R. & Smith, H. (1981): *Applied Regression Analysis.* New York: Wiley.

Garthwaite, P. H. (1994): An Interpretation of Partial Least Squares: *Journal of the American Statistical Association* **89**, 122 - 127

Gunst, R. F. & Mason, R. L. (1980): Regression Analysis and Its Applications: A Data-Oriented Approach. *Statistics: textbook and monographs* **141**.

Frank, I. E. & Friedman, J. H. (1993): Statistical view of chemometrics regression tools. *Technometrics* **35**, 127 - 131.

Füle, Erika (1995): On egological regression and ridge estimation. *Communications in Statistics.Simula.* **24**, 385-398.

Haaland, D. M. & Thomas, E. V. (1988): Partial Least-Squares Methods for Spectral Analyses. 1. Relation to other Quantitative Calibration Methods and the Extraction of Quantitative Information. *Analytical Chemistry* **60**, 1193 - 1201.

Heikka, R., Minkkinen, P. & Taavatsainen, V.-M. (1994): Comparision of variable selection and regression methods in multivariate calibration of a process analyzer. *Process Control and Quality* **6**, 47 - 54.

Helland, I. S. (1988): On the structure of partial least squares regression. *Communications in Statistics* **17**, 581 - 607.

Hoerl, A. E. & Kennard, R. W. (1970a): Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics* **12**, 55 - 67.

Hoerl, A. E. & Kennard, R. W. (1970b): Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics* **12**, 69 - 82.

Huhtala, K. (1991): Teräksen karkenevuuden mallittamisesta. *Suomen tilastoseuran vuosikirja 1991*. Helsinki: Haakapaino.

Höskulsson, A. (1988): PLS Regression Methods. *Journal of Chemometrics* **2**, 211 - 228.

Jackson, J. E. (1991): *A User's Guide to Principal Components*. New York: John Wiley & Sons, Inc.

Liski, E. & Puntanen, S. (1976): *Regressioanalyysin perusteita*. Tampere: Tampereen Pikakopio Oy.

Leskinen, E. (1977): Multikollineaarisuuden ongelmista regressioanalyysissä. *Jyväskylän yliopiston tilastotieteen laitoksen julkaisuja* **7**.

Leskinen, E. (1981): On Principal Component Analysis and Estimation in Regression. *Department of Statistics University of Jyväskylä*, **4**.

Naes,T., Irgens, C. & Martens, H. (1986): Comparision of Linear Statistical Method for Calibration of NIF Instruments. *Applied Statistics* **35**, 195 - 206.

Otto, M. & Wegscheider, W. (1985): Spectrophotometric Multicomponent Analysis Applied to Trace Metal Determinations. *Analytical Chemistry* **57**, 63 - 69.

Stone, M. & Brooks, R. J. (1990): Continuum Regression: Cross-validated Sequentially Constructed Prediction Embacing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *Journal of the Royal Statistical Society* **B 52**, 237 - 269.

Ranta, E. (1991): *Biometria: Tilastotiedettä ekologeille*. Helsinki: Yliopistopaino.

Wang, S.-G. & Chow, S-C (1994): Advanced linear models - Theory and Applications. *Statistics:textbook and monographs* **141**.

Liite A

Taulukko A.1: Estimointi- ja testiaineiston teräksen seosaineiden pitoisuuksien keskiarvot ja -hajonnat

Seos	Estimointiaineisto		Testiaineisto	
	Keskiarvo	Keskihajonta	Keskiarvo	Keskihajonta
C	0.2547	0.0937	0.2615	0.0989
Si	0.2614	0.0369	0.2593	0.0389
Mn	1.1992	0.1580	1.1977	0.1575
P	0.0145	0.0036	0.0135	0.0029
S	0.0217	0.0157	0.0235	0.0183
Cr	0.4922	0.3225	0.4941	0.3206
Ni	0.1689	0.0795	0.1643	0.0725
Mo	0.0360	0.0270	0.0421	0.0500
V	0.0056	0.0020	0.0054	0.0020
Ti	0.0333	0.0064	0.0327	0.0072
Cu	0.1784	0.0460	0.1723	0.0449
Als	0.0272	0.0074	0.0280	0.0069
B	0.0031	0.0005	0.0031	0.0006
Mb	0.0041	0.0009	0.0039	0.0009
N	0.0100	0.0017	0.0100	0.0015
Bjab	0.0029	0.0018	0.0027	0.0022

Taulukko A.2: Teräksen seosaineiden pitoisuuksien korrelaatiomatriisi

	C	Si	Mn	P	S	Cr
C	1.0000	.0501	.7996	.0971	-.2345	-.7379
Si	.0501	1.0000	.1049	-.0685	.1419	.1634
Mn	.7996	.1049	1.0000	.0558	-.1972	-.6779
P	.0971	-.0685	.0558	1.0000	.1400	-.0700
S	-.2345	.1419	-.1972	.1400	1.0000	.4523
Cr	-.7379	.1634	-.6779	-.0700	.4523	1.0000
Ni	-.2537	-.2359	-.4550	-.1676	-.1083	.1046
Mo	-.4137	.4589	-.2799	-.1671	.4453	.6232
V	-.1465	.2389	-.0160	-.0473	.2255	.3481
Ti	.3383	-.1032	.3214	-.0334	-.3383	-.3689
Cu	-.0969	-.0524	-.1825	.2552	.2028	.1062
Als	.0365	.0616	-.0865	-.1144	.2895	.2562
B	-.3293	.1335	-.2565	.0635	.3615	.3447
Mb	.0622	.1318	.1149	-.1092	.1096	.0739
N	-.4683	.1086	-.3721	-.2238	.0413	.4627
Bjab	.4752	-.1133	.4195	.1587	-.1729	-.4859

	Ni	Mo	V	Ti	Cu	Als
C	-.2537	-.4137	-.1465	.3383	-.0969	.0365
Si	-.2359	.4589	.2389	-.1032	-.0524	.0616
Mn	-.4550	-.2799	-.0160	.3214	-.1825	-.0865
P	-.1676	-.1671	-.0473	-.0334	.2552	-.1144
S	-.1083	.4453	.2255	-.3383	.2028	.2895
Cr	.1046	.6232	.3481	-.3689	.1062	.2562
Ni	1.0000	-.0038	-.1069	.1155	-.0953	.1476
Mo	-.0038	1.0000	.4561	-.4172	.0022	.1443
V	-.1069	.4561	1.0000	-.0564	.0465	.1827
Ti	.1155	-.4172	-.0564	1.0000	-.1301	.1146
Cu	-.0953	.0022	.0465	-.1301	1.0000	-.0231
Als	.1476	.1443	.1827	.1146	-.0231	1.0000
B	.0108	.4566	.1843	-.2837	.2044	.0642
Mb	.0176	.0891	.2098	.1264	.0531	.1815
N	.2095	.4197	.2322	-.0833	-.0957	.0116
Bjab	-.0549	-.4751	-.1450	.7346	.0271	.1139

	B	Mb	N	Bjab
C	-.3293	.0622	-.4683	.4590
Si	.1335	.1318	.1086	-.0965
Mn	-.2565	.1149	-.3721	-.3990
P	.0635	-.1092	-.2238	.1607
S	.3615	.1096	.0413	-.1483
Cr	.3447	.0739	.4627	-.4611
Ni	.0108	.0176	.2095	-.0513
Mo	.4566	.0891	.4197	-.4506
V	.1843	.2098	.2322	-.1435
Ti	-.2837	.1264	-.0833	.7153
Cu	.2044	.0531	-.0957	.0343
Als	.0642	.1815	.0116	.1494
B	1.0000	.0358	.1460	.0074
Mb	.0358	1.0000	-.0109	.1031
N	.1460	-.0109	1.0000	-.6741
Bjab	.0271	.1012	-.6724	1.0000

Taulukko A.3: Teräksen seosaineiden pitoisuuksien ja teräksen kovuuksien $y^{(1.5)}$, $y^{(13)}$ ja $y^{(25)}$ korrelaatiokertoimet

Seos	Teräksen kovuudet		
	$y^{(1.5)}$	$y^{(13)}$	$y^{(25)}$
C	0.985	0.934	0.847
Si	0.38	0.51	180
Mn	0.796	0.781	0.737
P	00.096	1.000	0.057
S	-0.262	-0.305	-0.097
Cr	-0.736	-0.671	-0.370
Ni	-0.250	-0.244	-0.229
Mo	-0.427	-0.360	-0.101
V	-0.178	-0.144	0.054
Ti	0.347	0.378	0.299
Cu	-0.097	-0.041	-0.047
Als	0.003	0.040	0.208
B	-0.340	-0.276	-0.188
Mb	0.051	0.062	0.174
N	-0.444	-0.453	-0.336
Bjab	0.463	0.489	0.356

Taulukko A.4: Estimointiaineiston rakennematriisin X neliömatriisin $X'X$ ominaisarvot sekä vastaavat ominaisarvot laskettuna testiaineistolle

Indeksi	Estimointiaineisto	Testiaineisto
1	4.484	4.798
2	2.051	2.051
3	1.706	1.706
4	1.580	1.580
5	0.942	0.942
6	0.882	0.882
7	0.825	0.825
8	0.689	0.689
9	0.889	0.669
10	0.618	0.618
11	0.563	0.563
12	0.449	0.449
13	0.260	0.260
14	0.152	0.152
15	0.125	0.125
16	0.005	0.005

Taulukko A.5: Regressiomallin (6.3) parametrin $B = (\beta^{(1.5)}, \beta^{(13)}, \beta^{(25)})$ pienimmän neliösumman estimaattori $\boldsymbol{B}_{PNS}$ ja suluissa sen varianssit.

Seosaine	$\boldsymbol{B}_{PNS}$		
	$\boldsymbol{b}^{(1.5)}_{PNS}$	$\boldsymbol{b}^{(13)}_{PNS}$	$\boldsymbol{b}^{(25)}_{PNS}$
C	0.9800 (0.0007)	0.9611 (0.0022)	0.9389 (0.0022)
Si	-0.0047 (0.0002)	-0.0026 (0.0008)	-0.0482 (0.0008)
Mn	0.0426 (0.0008)	0.2810 (0.0025)	0.5832 (0.0025)
P	0.0009 (0.0002)	0.0011 (0.0006)	0.0215 (0.0006)
S	-0.0221 (0.0003)	-0.1836 (0.0009)	-0.1444 (0.0009)
Cr	0.0437 (0.0008)	0.4110 (0.0024)	0.7521 (0.0023)
Ni	0.0052 (0.0003)	0.0962 (0.0009)	0.1900 (0.0009)
Mo	-0.0076 (0.0005)	0.1166 (0.0017)	0.2137 (0.0017)
V	-0.0382 (0.0002)	-0.0664 (0.0007)	-0.0612 (0.0007)
Ti	-0.0537 (0.0096)	-0.1749 (0.0302)	-0.1725 (0.0301)
Cu	0.0106 (0.0002)	0.1170 (0.0006)	0.0982 (0.0006)
Als	-0.0277 (0.0002)	0.0149 (0.0008)	0.0109 (0.0008)
B	-0.0271 (0.0016)	-0.0637 (0.0051)	-0.0739 (0.0050)
Nb	-0.0015 (0.0002)	-0.0078 (0.0006)	0.0041 (0.0006)
N	0.0796 (0.0069)	0.1440 (0.0216)	0.1265 (0.0215)
Bjab	0.0829 (0.0166)	0.3522 (0.0520)	0.3048 (0.0519)

Taulukko A.6: Regressiomallin (6.3) ja (6.4) parametrin A pienimmän neliösumman estimaattori A_{PNS} ja harjaestimaattori $A_H(K_{ad})$, missä K_{ad} on valittu (3.38) mukaan, ja näiden estimaattoreiden varianssit suluissa

Indeksi	A_{PNS}			$A_H(K_{ad})$		
	$a^{(1.5)}{}_{PNS}$	$a^{(13)}{}_{PNS}$	$a^{(25)}{}_{PNS}$	$a^{(1.5)}{}_{H}(k_{ad}^{(1.5)})$	$a^{(13)}{}_{H}(k_{ad}^{(13)})$	$a^{(25)}{}_{H}(k_{ad}^{(25)})$
1	0.3731 (0.4×10^{-4})	0.3230 (1.1×10^{-4})	0.2224 (1.1×10^{-4})	0.3730 (0.3×10^{-4})	0.3226 (1.0×10^{-4})	0.2219 (1.0×10^{-4})
2	-0.2208 (0.8×10^{-4})	-0.2828 (2.5×10^{-4})	-0.3631 (2.5×10^{-4})	-0.2207 (0.7×10^{-4})	-0.2822 (2.3×10^{-4})	-0.3614 (2.2×10^{-4})
3	-0.0690 (1.0×10^{-4})	-0.1108 (3.0×10^{-4})	-0.1644 (3.0×10^{-4})	-0.0690 (0.9×10^{-4})	-0.1105 (2.7×10^{-4})	-0.1635 (2.6×10^{-4})
4	0.1319 (1.0×10^{-4})	0.0520 (3.2×10^{-4})	0.0526 (3.2×10^{-4})	0.1318 (0.9×10^{-4})	0.0519 (2.9×10^{-4})	0.0523 (2.9×10^{-4})
5	-0.1303 (1.7×10^{-4})	-0.0803 (5.4×10^{-4})	-0.1192 (5.4×10^{-4})	-0.1302 (1.6×10^{-4})	-0.0800 (4.9×10^{-4})	-0.1118 (4.9×10^{-4})
6	0.0863 (1.8×10^{-4})	0.0376 (5.7×10^{-4})	0.0679 (5.7×10^{-4})	0.0862 (1.7×10^{-4})	0.0375 (5.2×10^{-4})	0.0672 (5.21×10^{-4})
7	-0.0062 (2.0×10^{-4})	-0.0022 (6.1×10^{-4})	-0.0553 (6.1×10^{-4})	0.0062 (1.8×10^{-4})	-0.0022 (5.6×10^{-4})	-0.0546 (5.5×10^{-4})
8	-0.0259 (2.4×10^{-4})	-0.0727 (7.4×10^{-4})	-0.0144 (7.3×10^{-4})	-0.0259 (2.1×10^{-4})	-0.0722 (6.7×10^{-4})	-0.0142 (6.6×10^{-4})
9	0.0193 (2.4×10^{-4})	0.0844 (7.6×10^{-4})	0.0153 (7.6×10^{-4})	0.0193 (2.2×10^{-4})	0.0839 (6.9×10^{-4})	0.0150 (6.8×10^{-4})
10	0.4164 (2.6×10^{-4})	0.3893 (8.2×10^{-4})	0.3781 (8.2×10^{-4})	0.4157 (2.4×10^{-4})	0.3865 (7.4×10^{-4})	0.3724 (7.4×10^{-4})
11	-0.0158 (2.9×10^{-4})	-0.0231 (9.0×10^{-4})	0.0049 (9.0×10^{-4})	-0.0158 (2.6×10^{-4})	0.0230 (8.2×10^{-4})	0.0048 (8.1×10^{-4})
12	0.0235 (3.6×10^{-4})	-0.0669 (11.2×10^{-4})	-0.0181 (11.2×10^{-4})	0.0235 (3.3×10^{-4})	-0.0662 (10.2×10^{-4})	-0.0177 (10.1×10^{-4})
13	-0.0968 (6.2×10^{-4})	-0.3921 (19.4×10^{-4})	-0.6105 (19.4×10^{-4})	-0.0964 (5.7×10^{-4})	-0.3855 (17.5×10^{-4})	-0.5893 (17.1×10^{-4})
14	-0.06856 (10.7×10^{-4})	-0.8475 (33.5×10^{-4})	-0.9924 (33.4×10^{-4})	-0.6808 (9.7×10^{-4})	-0.8234 (29.7×10^{-4})	-0.9347 (28.7×10^{-4})
15	-0.2784 (13.0×10^{-4})	0.0637 (40.6×10^{-4})	0.4214 (40.1×10^{-4})	-0.2760 (11.8×10^{-4})	0.0615 (35.8×10^{-4})	0.3921 (34.5×10^{-4})
16	-0.1307 (342.5×10^{-4})	-0.4167 (1071.1×10^{-4})	-0.3744 (1068.3×10^{-4})	-0.1067 ($255.6.1 \times 10^{-4}$)	-0.2149 ($505.1.1 \times 10^{-4}$)	-0.1258 (328.13×10^{-4})

Taulukko A.7: Regressiomallin (6.3) ja (6.4) parametrin A harjaestimaattorit $A_H(K_{FF})$, missä K_{FF} on valittu (3.59) mukaan, ja $A_H(K_{CROSS})$, missä K_{CROSS} on valittu (3.61) mukaan, sekä suluissa näiden estimaattoreiden varianssit

Indeksi	$A_H(K_{FF})$			$A_H(K_{CROSS})$		
	$a^{(1.5)}(k_{FF}^{(1.5)})$	$a_H^{(13)}(k_{FF}^{(13)})$	$a_H^{(25)}(k_{FF}^{(25)})$	$a_H^{(1.5)}(0.0072)$	$a_H^{(13)}(0.0072)$	$a_H^{(25)}(0.0072)$
1	0.3730 (0.4×10^{-4})	0.3230 (1.1×10^{-4})	0.2224 (1.1×10^{-4})	0.3725 (0.4×10^{-4})	0.3225 (1.1×10^{-4})	0.2220 (1.1×10^{-4})
2	-0.2207 (0.8×10^{-4})	-0.2828 (2.5×10^{-4})	-0.3631 (2.3×10^{-4})	-0.2200 (0.8×10^{-4})	-0.2818 (2.5×10^{-4})	-0.3618 (2.4×10^{-4})
3	-0.0690 (0.9×10^{-4})	-0.1108 (3.0×10^{-4})	-0.1644 (3.0×10^{-4})	-0.0687 (1.0×10^{-4})	-0.1103 (3.0×10^{-4})	-0.1637 (3.0×10^{-4})
4	0.1319 (1.0×10^{-4})	0.0520 (3.2×10^{-4})	0.0526 (3.2×10^{-4})	0.1313 (1.0×10^{-4})	0.0518 (3.2×10^{-4})	0.0523 (3.2×10^{-4})
5	-0.1302 (1.7×10^{-4})	-0.0803 (5.3×10^{-4})	-0.1129 (5.4×10^{-4})	-0.1293 1.7×10^{-4})	-0.0797 (5.4×10^{-4})	-0.1120 (5.4×10^{-4})
6	0.0862 (1.8×10^{-4})	0.0376 (5.7×10^{-4})	0.06979 (5.7×10^{-4})	0.0856 (1.8×10^{-4})	0.0373 (5.7×10^{-4})	0.06974 (5.7×10^{-4})
7	0.0062 (2.0×10^{-4})	-0.0022 (6.1×10^{-4})	-0.0553 (6.1×10^{-4})	0.0061 (2.0×10^{-4})	-0.0022 (6.1×10^{-4})	-0.0547 (6.1×10^{-4})
8	-0.0259 (2.4×10^{-4})	-0.0726 (7.4×10^{-4})	-0.0144 (7.3×10^{-4})	-0.0257 (2.3×10^{-4})	-0.0719 (7.3×10^{-4})	-0.0143 (7.3×10^{-4})
9	0.0193 (2.4×10^{-4})	0.0844 (7.6×10^{-4})	0.0153 (7.6×10^{-4})	0.0191 (2.4×10^{-4})	0.0835 (7.5×10^{-4})	0.0151 (7.5×10^{-4})
10	0.4158 (2.6×10^{-4})	0.3892 (8.2×10^{-4})	0.3781 (8.2×10^{-4})	0.4116 (2.6×10^{-4})	0.3848 (8.2×10^{-4})	0.3737 (8.2×10^{-4})
11	-0.0158 (2.9×10^{-4})	0.0231 (9.0×10^{-4})	0.0049 (9.0×10^{-4})	-0.0156 (2.9×10^{-4})	0.0229 (9.0×10^{-4})	0.0048 (8.9×10^{-4})
12	0.0235 (3.6×10^{-4})	-0.0669 (11.3×10^{-4})	-0.0181 (11.3×10^{-4})	0.0232 (3.6×10^{-4})	-0.0658 (11.2×10^{-4})	-0.0180 (11.2×10^{-4})
13	-0.0964 (6.2×10^{-4})	-0.3918 (19.5×10^{-4})	-0.6105 (19.4×10^{-4})	-0.0942 (6.1×10^{-4})	-0.3816 (19.1×10^{-4})	-0.5941 (13.1×10^{-4})
14	-0.6814 (10.5×10^{-4})	-0.8456 (33.4×10^{-4})	-0.9923 (33.4×10^{-4})	-0.6545 (10.3×10^{-4})	-0.8091 (32.2×10^{-4})	-0.9473 (32.1×10^{-4})
15	-0.2763 (12.7×10^{-4})	0.0636 (40.1×10^{-4})	0.4214 (40.4×10^{-4})	-0.2632 (12.4×10^{-4})	0.0602 (38.7×10^{-4})	0.3984 (38.6×10^{-4})
16	-0.1091 (285.9×10^{-4})	-0.4009 (1030.0×10^{-4})	-0.3733 (1058.0×10^{-4})	-0.0519 (137.0×10^{-4})	-0.1653 (427.8×10^{-4})	-0.1485 (427.4×10^{-4})