Tilastollinen vastepintamallinnus:
kokeiden suunnittelu, regressiomallin analyysi ja vasteen optimointi

Robert Piché ja Keijo Ruohonen
Tampereen teknillinen yliopisto

2010

Sisältö

Luku 1

REGRESSIO

Mallinnustilanteessa suure y riippuu suureista $x_1, \ldots, x_k$ tunnetun tai tuntemattoman funktion ϕ kautta, ts.

$$y = \phi(x_1, \ldots, x_k).$$

Suure y on tällöin ns. *vaste* eli *selitettävä muuttuja* ja muuttujat $x_1, \ldots, x_k$ ovat ns. *faktoreita* eli *selittäviä muuttujia*. Faktoreiden arvoja kutsutaan *tasoiksi*. Funktio ϕ on ns. *todellinen vastefunktio* ja se sisältää systeemin kaikki fysiikan ilmiöt, mittausprosessin mukaan lukien.

response

levels

Funktio ϕ on yleensä tuntematon tai sitten niin mutkikas, ettei sitä voida sellaisenaan käyttää. Otamme käyttöön muotoa

$$y = f(x_1, \ldots, x_k) + \epsilon,$$

olevan *empirisen mallin*, missä funktio f on olettamamme funktion ϕ approksimaatio, ja *virhetermi* ϵ on satunnaismuuttuja.

Malli saadaan käyttöön, kun ensin on saatu kokeiden tuloksena tietty määrä faktorien arvoyhdelmiä ja niitä vastaavat vasteen arvot. *Datana* on kerätty N kappaletta faktorien arvoyhdelmiä sekä niitä vastaavat vasteen arvot:

koe	faktorien tasot	vaste
1	$x_{11}, \ldots, x_{1k}$	y_1
2	$x_{21}, \ldots, x_{2k}$	y_2
$\vdots$	$\vdots$	$\vdots$
N	$x_{N1}, \ldots, x_{Nk}$	y_N

Regressioanalyysi on mallin f *parametrien* estimointi.

Mallia voidaan käyttää

- arvioimaan erilaisten faktoreiden vaikutusta vasteeseen tai

- ennustamaan vasteen arvon sellaisille faktorien arvoyhdelmille, joille vastaavia kokeita ei ole tehty, mm. vasteen optimoinnin yhteydessä.

Jos datan keruu on kallista, vaarallista tai muuten hankalaa, kannattaa valita faktorien arvoyhdelmät etukäteen niin, että saatu malli olisi mahdollisimman käyttökelpoinen mahdollisimman pienellä arvoyhdelmien lukumäärällä. Tällainen valinta on *tilastollinen kokeiden suunnittelu*. Jos data on jo kerätty tai jos faktoreiden tasot eivät ole tiedossa tai niihin ei voida vaikuttaa, ei kokeiden suunnittelua tarvita.

design of experiments

Tässä luvussa kerrataan tilastomatematiikan peruskurssien opettamia tietoja regressiomallin rakentamisesta, mallia koskevien hypoteesien testaamisesta ja mallin sopivuuden arvioinnista. Keskitymme lineaarisen regressiomenetelmän osiin, joita tarvitaan vastepintamallinnuksessa.

1.1 Matriisilaskentaa ja multinormaalijakauma

Tässä kerrataan ja käsitellään lyhyesti eräitä tilastollisten monimuuttujamenetelmien tarvitsemia matriisilaskennan ja normaalijakauman käsitteitä.

Matriisilaskentaa

Aluksi eräitä määritelmiä. Neliömatriisi $\mathbf{A}$ on *symmetrinen*, jos $\mathbf{A}^T = \mathbf{A}$, ja *idempotentti*, jos $\mathbf{A}^2 = \mathbf{A}$. Idempotentin matriisin ainoat mahdolliset ominaisarvot ovat 0 ja 1, sillä jos $\mathbf{Ax} = \lambda\mathbf{x}$, niin myös $\mathbf{A}^2\mathbf{x} = \lambda\mathbf{x}$ ja toisaalta $\mathbf{A}^2\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, joten $\lambda^2 = \lambda$. Jos symmetrinen matriisi on ei-singulaarinen, niin sen käänteismatriisi on myös symmetrinen.

eigenvalues

Matriisin *rivirangi* (vast. *sarakerangi*) on sen suurin lineaarisesti riippumattomien rivien (vast. sarakkeiden) lukumäärä. Tunnetusti matriisin $\mathbf{A}$ rivi- ja sarakerangit ovat samat, tätä yhteistä arvoa kutsutaan matriisin *asteeksi* eli *rangiksi*, merkitään $\mathrm{rank}(\mathbf{A})$. Edelleen neliömatriisin rangi on sen nollasta eroavien ominaisarvojen lukumäärä (moninkertaiset ominaisarvot otetaan mukaan kertalukunsa osoittama määrä). Näin ollen idempotentin matriisin rangi on sen 1-ominaisarvojen lukumäärä.

diagonal elements

Neliömatriisin $\mathbf{A}$ *jälki*, merkitään $\mathrm{trace}(\mathbf{A})$, on sen lävistäjäalkioiden summa. Jäljellä on seuraavat ominaisuudet:

1. $\mathrm{trace}(\mathbf{A} + \mathbf{B}) = \mathrm{trace}(\mathbf{A}) + \mathrm{trace}(\mathbf{B})$;
2. $\mathrm{trace}(c\mathbf{A}) = c\,\mathrm{trace}(\mathbf{A})$ (c on skalaari);
3. $\mathrm{trace}(\mathbf{A}^T) = \mathrm{trace}(\mathbf{A})$;

4. $\mathrm{trace}(\mathbf{B}^T\mathbf{A}) = \mathrm{trace}(\mathbf{A}\mathbf{B}^T) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_{ij}b_{ij}$, kun $\mathbf{A}$ ja $\mathbf{B}$ ovat $n \times m$-matriiseja;

5. $\mathrm{trace}(\mathbf{A})$ on $\mathbf{A}$:n ominaisarvojen summa neliömatriisille $\mathbf{A}$.

Ominaisuudesta 5. johtuen idempotentin matriisin rangi on sen jälki.

Merkitään $\mathbf{0}_n$:llä n-vektoria, jonka kaikki alkiot ovat nollia (*nollavektori*), $\mathbf{1}_n$:llä n-vektoria, jonka kaikki alkiot ovat ykkösiä (*ykkösvektori*), $\mathbf{O}_n$:llä $n \times n$-matriisia, jonka kaikki alkiot ovat nollia (*nollamatriisi*), ja vielä $\mathbf{I}_n$:llä $n \times n$-identiteettimatriisia. Seuraavia erikoismatriiseja tarvitaan usein:

$$\mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^T \quad , \quad \mathbf{K}_n = \mathbf{J}_n - \mathbf{I}_n \quad , \quad \mathbf{M}_n = \mathbf{I}_n - \frac{1}{n}\mathbf{J}_n$$

$\mathbf{J}_n$ on $n \times n$-matriisi, jonka kaikki alkiot ovat ykkösiä. Matriisit $\mathbf{J}_n$, $\mathbf{K}_n$ ja $\mathbf{M}_n$ ovat symmetrisiä. Seuraavat ominaisuudet ovat todettavissa suoralla laskulla:

(i) $\mathbf{1}_n^T\mathbf{1}_n = n$

(ii) $\mathbf{J}_n\mathbf{1}_n = n\mathbf{1}_n$

(iii) $\mathbf{K}_n\mathbf{1}_n = (n-1)\mathbf{1}_n$

(iv) $\mathbf{M}_n\mathbf{1}_n = \mathbf{0}_n$

(v) $\mathbf{J}_n^2 = n\mathbf{J}_n$

(vi) $\mathbf{K}_n^2 = (n-1)\mathbf{J}_n - \mathbf{K}_n$

(vii) $\mathbf{M}_n^2 = \mathbf{M}_n$ (eli $\mathbf{M}_n$ on idempotentti)

(viii) $\mathbf{J}_n\mathbf{K}_n = (n-1)\mathbf{J}_n$

(ix) $\mathbf{J}_n\mathbf{M}_n = \mathbf{O}_n$

(x) $\mathbf{K}_n\mathbf{M}_n = -\mathbf{M}_n$

(xi) $n(\mathbf{K}_n + \mathbf{M}_n) = (n-1)\mathbf{J}_n$

Matriiseja on usein edullista käsitellä jaettuina lohkoihin:

$$\mathbf{A} = \left(\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1k} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2k} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{A}_{\ell 1} & \mathbf{A}_{\ell 2} & \cdots & \mathbf{A}_{\ell k} \end{array}\right).$$

Lohkomuodossa olevien matriisien transpoosi ja tulo saadaan suoraan lohkojen avulla:

$$\left(\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1k} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2k} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{A}_{\ell 1} & \mathbf{A}_{\ell 2} & \cdots & \mathbf{A}_{\ell k} \end{array}\right)^T = \left(\begin{array}{c|c|c|c} \mathbf{A}_{11}^T & \mathbf{A}_{21}^T & \cdots & \mathbf{A}_{\ell 1}^T \\ \hline \mathbf{A}_{12}^T & \mathbf{A}_{22}^T & \cdots & \mathbf{A}_{\ell 2}^T \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{A}_{1k}^T & \mathbf{A}_{2k}^T & \cdots & \mathbf{A}_{\ell k}^T \end{array}\right)$$

ja

$$\left(\begin{array}{c|c|c|c} \mathbf{A}_{11} & \mathbf{A}_{12} & \cdots & \mathbf{A}_{1k} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} & \cdots & \mathbf{A}_{2k} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{A}_{\ell 1} & \mathbf{A}_{\ell 2} & \cdots & \mathbf{A}_{\ell k} \end{array}\right) \left(\begin{array}{c|c|c|c} \mathbf{B}_{11} & \mathbf{B}_{12} & \cdots & \mathbf{B}_{1m} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} & \cdots & \mathbf{B}_{2m} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{B}_{k1} & \mathbf{B}_{k2} & \cdots & \mathbf{B}_{km} \end{array}\right) = \left(\begin{array}{c|c|c|c} \mathbf{C}_{11} & \mathbf{C}_{12} & \cdots & \mathbf{C}_{1m} \\ \hline \mathbf{C}_{21} & \mathbf{C}_{22} & \cdots & \mathbf{C}_{2m} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \mathbf{C}_{\ell 1} & \mathbf{C}_{\ell 2} & \cdots & \mathbf{C}_{\ell m} \end{array}\right),$$

missä

$$\mathbf{C}_{ij} = \sum_{t=1}^{k} \mathbf{A}_{it}\mathbf{B}_{tj}$$

(huomaa kertojärjestys), olettaen, että kaikki esiintyvät matriisikertolaskut ovat määriteltyjä. Lohkokertosääntö muistuttaa ”tavallista” matriisien kertosääntöä $c_{ij} = \sum_{t=1}^{k} a_{it}b_{tj}$, ja voidaan sitä käyttäen todistaa helposti. Eräs erikoistapaus on ns. *toinen matriisikertosääntö*

$$\left(\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_k \end{array}\right) \left(\begin{array}{c} \mathbf{b}_1^T \\ \hline \mathbf{b}_2^T \\ \hline \vdots \\ \hline \mathbf{b}_k^T \end{array}\right) = \sum_{t=1}^{k} \mathbf{a}_t\mathbf{b}_t^T.$$

Summalausekkeet ja matriisit liittyvät toisiinsa seuraavilla kaavoilla, jotka ovat helposti todettavissa. Merkitään

$$\mathbf{A} = \left(\begin{array}{c|c|c|c} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_k \end{array}\right) \quad \text{ja} \quad \mathbf{B} = \left(\begin{array}{c} \mathbf{b}_1^T \\ \hline \mathbf{b}_2^T \\ \hline \vdots \\ \hline \mathbf{b}_k^T \end{array}\right).$$

Silloin

column mean

1. $\frac{1}{k}\mathbf{A}\mathbf{1}_k = \frac{1}{k}\sum_{t=1}^{k} \mathbf{a}_t$ (ns. $\mathbf{A}$:n *sarakekeskiarvo*, merkitään $\bar{\mathbf{a}}$);

row mean

2. $\frac{1}{k}\mathbf{1}_k^T\mathbf{B} = \frac{1}{k}\sum_{t=1}^{k} \mathbf{b}_t^T$ (ns. $\mathbf{B}$:n *rivikeskiarvo*, merkitään $\bar{\mathbf{b}}^T$);

3. $\mathbf{A}\mathbf{J}_k\mathbf{B} = \sum_{t=1}^{k}\sum_{s=1}^{k} \mathbf{a}_t\mathbf{b}_s^T$;

4. $\mathbf{A}\mathbf{K}_k\mathbf{B} = \sum_{t=1}^{k}\sum_{\substack{s=1\\s\neq t}}^{k}\mathbf{a}_t\mathbf{b}_s^T$;

5. $\mathbf{A}\mathbf{M}_k = \mathbf{A} - \bar{\mathbf{a}}\mathbf{1}_k^T$ (vähennetään $\mathbf{A}$:n sarakkeista sen sarakekeskiarvo eli *keskitetään sarakkeet*);

6. $\mathbf{M}_k\mathbf{B} = \mathbf{B} - \mathbf{1}_k\bar{\mathbf{b}}^T$ (vähennetään $\mathbf{B}$:n riveistä sen rivikeskiarvo eli *keskitetään rivit*);

7. $\mathbf{A}\mathbf{M}_k\mathbf{B} = \mathbf{A}\mathbf{M}_k^2\mathbf{B} = (\mathbf{A} - \bar{\mathbf{a}}\mathbf{1}_k^T)(\mathbf{B} - \mathbf{1}_k\bar{\mathbf{b}}^T)$.

Kohdan 7. seurauksena erikoisesti

$$\mathbf{A}\mathbf{M}_k\mathbf{A}^T = (\mathbf{A} - \bar{\mathbf{a}}\mathbf{1}_k^T)(\mathbf{A} - \bar{\mathbf{a}}\mathbf{1}_k^T)^T.$$

Multinormaalijakauma

Satunnaisvektorilla $\mathbf{x}$ (n-vektori) on ns. *multinormaalijakauma* $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, jos sen tiheysfunktio on *multinormal distribution*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Sigma})}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right).$$

Tässä $\boldsymbol{\mu} = E(\mathbf{x})$ (odotusarvo(vektori)) ja $\boldsymbol{\Sigma} = V(\mathbf{x}) = E((\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T)$ (varianssi(matriisi)). Mikäli $\boldsymbol{\mu} = \mathbf{0}_n$ ja $\boldsymbol{\Sigma} = \mathbf{I}_n$, on kyseessä ns. *standardimultinormaalijakauma*. *expectation*

Todetaan seuraavat multinormaalijakauman ominaisuudet:

1. Jos $\mathbf{x}$:llä on n-ulotteinen $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-jakauma, $\mathbf{C}$ on $m \times n$-matriisi, jonka rivirangi on täysi (eli m), ja $\mathbf{b}$ on m-vektori, niin satunnaisvektorilla $\mathbf{C}\mathbf{x}+\mathbf{b}$ on m-ulotteinen $\mathrm{N}(\mathbf{C}\boldsymbol{\mu}+\mathbf{b}, \mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$-jakauma.

2. Jos $\mathbf{x}$:llä on n-ulotteinen $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$-jakauma, $\mathbf{C}_1$ on $m_1 \times n$-matriisi, $\mathbf{C}_2$ on $m_2 \times n$-matriisi, $\mathbf{b}_1$ on m_1-vektori ja $\mathbf{b}_2$ on m_2-vektori, niin satunnaisvektorit $\mathbf{C}_1\mathbf{x}+\mathbf{b}_1$ ja $\mathbf{C}_2\mathbf{x}+\mathbf{b}_2$ ovat riippumattomat tarkalleen silloin, kun $\mathbf{C}_1\boldsymbol{\Sigma}\mathbf{C}_2^T = \mathbf{O}$. *independent*

3. Jos $\mathbf{x}$:llä on n-ulotteinen $\mathrm{N}(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n)$-jakauma ja $s^2 = \frac{1}{n-1}\mathbf{x}^T\mathbf{M}_n\mathbf{x}$ (otosvarianssi) niin satunnaismuuttujalla *sample variance*

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

on χ^2-jakauma $n-1$ vapausasteella. Yleisesti, jos $\mathbf{A}$ on symmetrinen idem- *degrees of freedom*

potentti $n \times n$ matriisi, niin satunnaismuuttujalla

$$\frac{1}{\sigma^2}(\mathbf{x} - \mu\mathbf{1}_n)^T\mathbf{A}(\mathbf{x} - \mu\mathbf{1}_n)$$

on χ^2-jakauma $\mathrm{rank}(\mathbf{A})$ vapausasteella.

sample mean

4. Jos $\mathbf{x}$:llä on n-ulotteinen $\mathrm{N}(\mu\mathbf{1}_n, \sigma^2\mathbf{I}_n)$-jakauma, $\bar{x} = \frac{1}{n}\mathbf{1}_n^T\mathbf{x}$ (otoskeskiarvo) ja $s^2 = \frac{1}{n-1}\mathbf{x}^T\mathbf{M}_n\mathbf{x}$ (otosvarianssi) niin satunnaismuuttujalla

$$\frac{(\bar{x} - \mu)\sqrt{n}}{s}$$

on t-jakauma $n-1$ vapausasteella. (Huomaa, että $(\bar{x} - \mu)\sqrt{n}/\sigma$ on standardinormaalisti jakautunut ja $s^2(n-1)/\sigma^2$ on χ^2-jakautunut $n-1$ vapausasteella ja että nämä satunnaismuuttujat ovat riippumattomat. Yleisesti, jos u on standardinormaalisti jakautunut, v on χ^2-jakautunut m vapausasteella ja u ja v ovat riippumattomat, niin $u\sqrt{m}/\sqrt{v}$ on t-jakautunut m vapausasteella.)

5. Jos $\mathbf{x}_1$:llä on n_1-ulotteinen $\mathrm{N}(\boldsymbol{\mu}_1, \sigma^2\mathbf{I}_{n_1})$-jakauma, $\mathbf{x}_2$:llä on n_2-ulotteinen $\mathrm{N}(\boldsymbol{\mu}_2, \sigma^2\mathbf{I}_{n_2})$-jakauma sekä $\mathbf{x}_1$ ja $\mathbf{x}_2$ ovat riippumattomat, niin satunnaismuuttujalla

$$\frac{\mathbf{x}_1^T\mathbf{M}_{n_1}\mathbf{x}_1/(n_1 - 1)}{\mathbf{x}_2^T\mathbf{M}_{n_2}\mathbf{x}_2(n_2 - 1)}$$

on F-jakauma vapausastein $n_1 - 1$ ja $n_2 - 1$. (Huomaa, että $\mathbf{x}_1^T\mathbf{M}_{n_1}\mathbf{x}_1/\sigma^2$ ja $\mathbf{x}_2^T\mathbf{M}_{n_2}\mathbf{x}_2/\sigma^2$ ovat riippumattomat χ^2-jakautuneet satunnaismuuttujat vapausastein $n_1 - 1$ ja $n_2 - 1$, vastaavasti. Yleisesti riippumattomien, vapausastein m_1 ja m_2 χ^2-jakautuneiden vapausasteillaan jaettujen satunnaismuuttujien osamäärä on F-jakautunut vapausastein m_1 ja m_2.)

1.2 Lineaarinen regressiomalli

Ensimmäisen kertaluvun regressiomalli on muotoa

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon.$$

missä kertoimet $\beta_0, \beta_1, \ldots, \beta_k$ ovat mallin *parametrit*. Jos merkitään

$$\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_k \end{pmatrix} \quad \text{ja} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix},$$

voidaan tällainen 1. kertaluvun regressiomalli kirjoittaa muotoon

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon.$$

Yleisesti *d:nnen kertaluvun regressiomalli* on muotoa $y = p(x_1, \ldots, x_k) + \epsilon$ oleva malli, missä p on muuttujien $x_1, \ldots, x_k$ d:nnen asteen polynomi, jonka kertoimet ovat parametrejä. Polynomin p ei tarvitse sisältää kaikkia mahdollisia termejä. Itse asiassa polynomiaalinen regressio voidaan palauttaa 1. kertaluvun regressioksi ottamalla uusia faktoreita käyttöön. Esimerkiksi toisen kertaluvun regressiomallissa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

ainoa korkeamman kuin ensimmäisen asteen termi on $\beta_{12}x_1x_2$. Otetaan uusi muuttuja x_3, kirjoitetaan sen arvoksi suureen x_1x_2 arvo, ja valitaan parametri β_{12} sen kertoimeksi. Näin saadaan 1. kertaluvun regressiomalli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_3 + \epsilon.$$

Polynomimallit ovat *lineaarisia* regressiomalleja, ts. ne ovat parametriensä lineaariyhdelmä + virhetermi.

linear combination

Regressiomallien parametrit voidaan estimoida *pienimmän neliösumman keinolla*. Muodostetaan koedatasta ns. *datamatriisi* $\mathbf{X}$ sekä *vastevektori* $\mathbf{y}$:

method of least squares

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & x_{N2} & \cdots & x_{Nk} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}.$$

Valitaan parametrien $\boldsymbol{\beta}$ estimaatti $\mathbf{b}$ siten, että neliösumma

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})$$

minimoituu. Neliösumman gradientti $\mathbf{b}$:n suhteen on $-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{b})$ ja merkitsemällä se nollavektoriksi saadaan lineaarinen yhtälöryhmä

$$\mathbf{X}^T\mathbf{X}\mathbf{b} = \mathbf{X}^T\mathbf{y},$$

josta ratkaistaan $\mathbf{b}$:

$$\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Tällöin tietysti oletetaan, että $\mathbf{X}^T\mathbf{X}$ on ei-singulaarinen ja erityisesti, että $N \geq k+1$. Matriisit $\mathbf{X}^T\mathbf{X}$ ja $(\mathbf{X}^T\mathbf{X})^{-1}$ ovat symmetrisiä.

Koska 1. kertaluvun malli on muotoa $y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$, liittyvät vastevektori ja datamatriisi toisiinsa yhtälöllä

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \quad , \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix},$$

missä $\boldsymbol{\epsilon}$ on satunnaisvektori. Satunnaismuuttujat $\epsilon_1, \epsilon_2, \ldots, \epsilon_N$ ovat riippumattomia (sillä kokeet suoritetaan toisistaan riippumattomasti) ja niillä on kullakin $\mathrm{N}(0, \sigma^2)$-jakauma. (Odotusarvo on 0, sillä systemaattinen virhe voidaan sisällyttää parametriin β_0.) Satunnaisvektorilla $\boldsymbol{\epsilon}$ on siis $\mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_N)$-multinormaalijakauma. Koska $\boldsymbol{\epsilon}$ on satunnaisvektori, niin samoin on $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ sekä edelleen

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}.$$

Vaikka $\boldsymbol{\epsilon}$:n komponentit ovat riippumattomia satunnaismuuttujia, eivät $\mathbf{b}$:n komponentit sitä yleisesti ole. Välittömästi todetaan nimittäin, että

$$E(\mathbf{b}) = E(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}) = \boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T E(\boldsymbol{\epsilon}) = \boldsymbol{\beta}$$

ja

$$V(\mathbf{b}) = V(\boldsymbol{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\epsilon}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T V(\boldsymbol{\epsilon})\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.$$

Siispä $\mathbf{b}$:llä on $\mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$-multinormaalijakauma ja sen komponentit ovat riippumattomat tarkalleen silloin, kun $\mathbf{X}^T\mathbf{X}$ on lävistäjämatriisi (jolloin myös $(\mathbf{X}^T\mathbf{X})^{-1}$ on lävistäjämatriisi).

Kun mallin parametrien estimaatti $\mathbf{b}$ on saatu, voidaan muita faktorien tasoja $\boldsymbol{\xi}$ vastaava vasteen arvo ennustaa:

$$\hat{y} = \boldsymbol{\xi}^T\mathbf{b}.$$

Tässä $\mathbf{b}$ on satunnaisvektori, joten ennustus $\hat{y}$ on satunnaismuuttuja. Edelleen

$$E(\hat{y}) = \boldsymbol{\xi}^T E(\mathbf{b}) = \boldsymbol{\xi}^T\boldsymbol{\beta}$$

ja

$$V(\hat{y}) = \boldsymbol{\xi}^T V(\mathbf{b})\boldsymbol{\xi} = \sigma^2\boldsymbol{\xi}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi}.$$

Erityisesti voidaan ”ennustaa” datamatriisissa esiintyviä faktorien arvoyhdelmiä vastaavat vasteet:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

Erotus $\mathbf{y} - \hat{\mathbf{y}} =: \mathbf{r}$ on ns. *residuaalivektori*, datan avulla lausuttuna

$$\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}.$$

Ideaalisesti residuaalissa $\mathbf{r}$ on vain ”kohinaa” eli virhetermin $\boldsymbol{\epsilon}$ vaikutus. Residuaalivektorin pituuden neliö

$$\|\mathbf{r}\|^2 = \mathbf{r}^T\mathbf{r} = (\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b}) =: \mathrm{SSE}$$

on ns. *residuaalin neliösumma*. Koska $E(\mathrm{SSE}) = (N - k - 1)\sigma^2$, niin (olettaen, että $N > k + 1$) *sum of squares of residuals (errors)*

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

on virhetermin varianssin σ^2 estimaatti. Jos merkitään

$$(\mathbf{X}^T\mathbf{X})^{-1} =: \mathbf{C} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{0k} \\ c_{10} & c_{11} & \cdots & c_{1k} \\ \vdots & \vdots & \ddots & \vdots \\ c_{k0} & c_{k1} & \cdots & c_{kk} \end{pmatrix},$$

niin $V(b_i) = \sigma^2 c_{ii}$. Näin ollen $V(b_i)$:n estimaatiksi käy $s^2 c_{ii}$. *Standardivirhe* $\mathrm{se}(b_i) := \sqrt{s^2 c_{ii}}$ on parametrin estimoitu keskihajonta. Vastaavasti *standard error* *standard deviation*

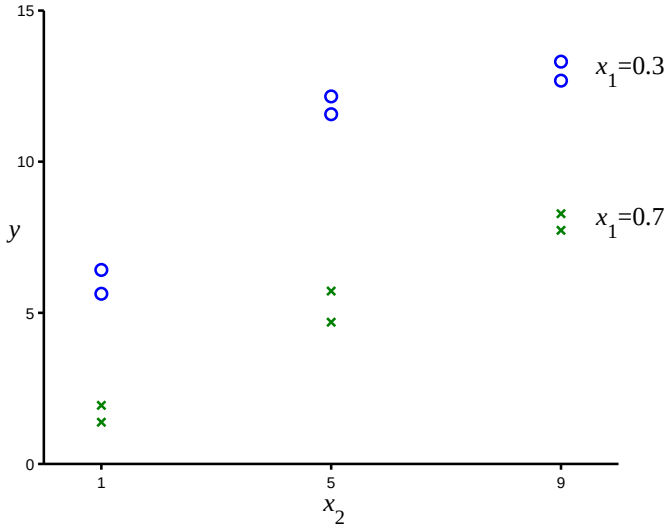
$$\mathrm{se}(\hat{y}) := \sqrt{s^2 \boldsymbol{\xi}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi}}$$

on vasteen ennustuksen estimoitu keskihajonta.

Esimerkki

Määritä ensimmäisen kertaluvun regressiomalli datalle:

x_1	x_2	y
0.3	1	5.63
0.3	1	6.42
0.7	1	1.38
0.7	1	1.94
0.3	5	11.57
0.3	5	12.16
0.7	5	5.72
0.7	5	4.69
0.3	9	12.68
0.3	9	13.31
0.7	9	8.28
0.7	9	7.73



ja estimoi parametrien standardivirheet. Ennusta mallin avulla vasteen y arvo tasoilla $x_1 = 0.5$, $x_2 = 4$.

Ratkaisu

Malli on siis muotoa

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

Datamatriisi ja vastevektori ovat:

$$\mathbf{X} = \begin{pmatrix} 1 & 0.3 & 1 \\ 1 & 0.3 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.3 & 5 \\ 1 & 0.3 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.3 & 9 \\ 1 & 0.3 & 9 \\ 1 & 0.7 & 9 \\ 1 & 0.7 & 9 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 5.63 \\ 6.42 \\ 1.38 \\ 1.94 \\ 11.57 \\ 12.16 \\ 5.72 \\ 4.69 \\ 12.68 \\ 13.31 \\ 8.28 \\ 7.73 \end{pmatrix}.$$

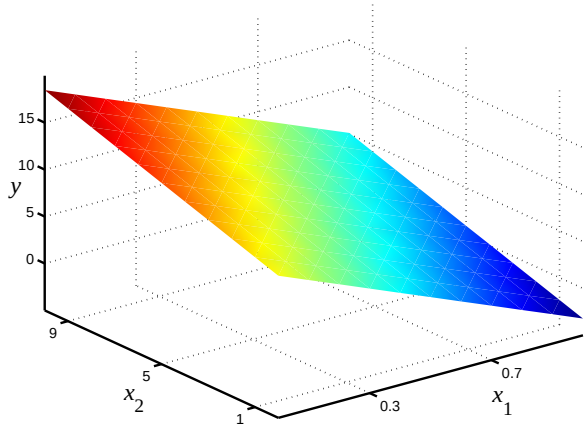
Pienimmän neliösumman menetelmällä saadaan tarkkojen parametrien $\boldsymbol{\beta}$ estimaatti $\mathbf{b}$:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 10.14 \\ -13.35 \\ 0.83 \end{pmatrix},$$

josta varianssianalyysissa käytetty matriisi $\mathbf{C}$:

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.80 & -1.04 & -0.04 \\ -1.04 & 2.08 & 0.00 \\ -0.04 & 0.00 & 0.01 \end{pmatrix}.$$

Regressiomalli on siis $\hat{y} = 10.14 - 13.35x_1 + 0.83x_2$. Vastepinta on taso:



”Ennustamme” datan vasteet ja laskemme residuaalivektorin:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \begin{pmatrix} 6.97 \\ 6.97 \\ 1.63 \\ 1.63 \\ 10.30 \\ 10.30 \\ 4.96 \\ 4.96 \\ 13.62 \\ 13.62 \\ 8.29 \\ 8.29 \end{pmatrix}, \quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -1.34 \\ -0.55 \\ -0.25 \\ 0.31 \\ 1.28 \\ 1.87 \\ 0.76 \\ -0.27 \\ -0.94 \\ -0.31 \\ -0.01 \\ -0.56 \end{pmatrix}.$$

Residuaalin neliösumma:

$$\mathrm{SSE} = \|\mathbf{r}\|^2 = \mathbf{r}^T\mathbf{r} = 9.30.$$

Virhetermin varianssin σ^2 estimaatti s^2:

$$s^2 = \frac{\mathrm{SSE}}{N-k-1} = \frac{\mathrm{SSE}}{12-2-1} = 1.03.$$

Estimoidut parametrien keskihajonnat eli standardivirheet:

$$\begin{aligned} \mathrm{se}(b_0) &= \sqrt{s^2 c_{00}} = \sqrt{1.03 \cdot 0.80} = 0.91, \\ \mathrm{se}(b_1) &= \sqrt{s^2 c_{11}} = \sqrt{1.03 \cdot 2.08} = 1.47, \\ \mathrm{se}(b_2) &= \sqrt{s^2 c_{22}} = \sqrt{1.03 \cdot 0.01} = 0.09. \end{aligned}$$

Nyt voidaan ennustaa mallia käyttäen vasteen arvo, kun $x_1 = 0.5$ ja $x_2 = 4$:

$$\boldsymbol{\xi} = \begin{pmatrix} 1 \\ 0.5 \\ 4 \end{pmatrix}, \quad \hat{y} = \boldsymbol{\xi}^T\mathbf{b} = (1\ 0.5\ 4) \begin{pmatrix} 10.14 \\ -13.35 \\ 0.83 \end{pmatrix} = 10.14 - 13.35 \cdot 0.5 + 0.83 \cdot 4 = 6.79.$$

Ennusteen keskihajonta estimoidaan:

$$\mathrm{se}(\hat{y}) = \sqrt{s^2 \boldsymbol{\xi}^T \mathbf{C} \boldsymbol{\xi}} = 0.31.$$

Harjoitustehtävät

1. Muodosta satunnainen datamatriisi $\mathbf{X}$, jossa on $N = 10$ koetta ja $k = 5$ faktoria. Valitse jokin parametrivektori $\boldsymbol{\beta}$ ja generoi simuloitu vastevektori $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, missä $\boldsymbol{\epsilon}$ on $\mathrm{N}(\mathbf{0}, 0.01\mathbf{I})$. Estimoi nyt $\boldsymbol{\beta}$ eli etsi $\mathbf{b}$ pienimmän neliösumman keinolla. Laske myös parametrien estimoidut standardivirheet $\mathrm{se}(b_i)$.

   Kokeile miten käy, kun vaihdat matriisin $\mathbf{X}$ 3. sarakkeeksi sen 2. sarakkeen lisättynä pienellä luvulla δ. Käy läpi arvot $\delta = 0.1, 0.01, 0.0001, 0.000001$ ja seuraa mitä tapahtuu estimaatille $\mathbf{b}$.

2. Mitkä ovat faktorit, parametrit ja vasteet ja miten saadaan datamatriisit, kun mallit muunnetaan 1. kertaluvun lineaarisiksi malleiksi, ja kun ϵ on $\mathrm{N}(0, \sigma^2)$-normaalijakautunut virhetermi?

   (a) $y = \beta_0 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{23} x_2 x_3 + \epsilon$
   (b) $y = \beta_0 + \beta_1 \sin(\omega_1 t + \phi_1) + \beta_2 \sin(\omega_2 t + \phi_2) + \epsilon$ (t on aikamuuttuja)
   (c) $y = C\lambda t^k$ (t on aikamuuttuja ja $\lambda = e^\epsilon$ on lognormaali virhetermi)

3. Johda kaavat

   (a) $\mathbf{r} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\boldsymbol{\epsilon}$
   (b) $\mathbf{X}^T\mathbf{r} = \mathbf{0}$
   (c) $E(\mathbf{b}\mathbf{r}^T) = \mathbf{O}$
   (d) $E(\mathrm{SSE}) = (N - k - 1)\sigma^2$

1.3 Hypoteesien testaaminen

Varianssianalyysiä käyttäen voidaan testata ns. *lineaarisia hypoteeseja,* ts. muotoa

$$\mathrm{H}_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$$

olevia hypoteeseja, missä $\mathbf{A}$ on $q \times (k + 1)$-matriisi, jonka rivirangi on täysi, ts. sen rivit ovat lineaarisesti riippumattomat, ja $\mathbf{d}$ on q-vektori. Vielä oletetaan, että $q < k + 1$. Valitsemalla $\mathbf{A}$ ja $\mathbf{d}$ sopivasti saadaan hyvinkin monenlaisia testejä. Vastahypoteesi on $\mathrm{H}_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{d}$.

Regressiomallin hypoteesintestauksen perustulos on

Lause 1.1. *Jos* $\mathrm{H}_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}$ *on tosi, niin (aiemmin mainituin normaalisuusoletuksin) suureella*

$$\frac{(\mathbf{A}\mathbf{b} - \mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{d})/q}{\mathrm{SSE}/(N - k - 1)}$$

on F-jakauma vapausastein q ja $N-k-1$ (taas kerran olettaen, että $N > k+1$).

Todistus. Ensinnäkin (kts. kohdan 1.2 harjoitustehtävä 3c) satunnaisvektorit $\mathbf{b}$ ja $\mathbf{r}$ ovat riippumattomia. Näin ollen ovat myös suureet $(\mathbf{Ab}-\mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{Ab}-\mathbf{d})$ ja $\mathrm{SSE} = \mathbf{r}^T\mathbf{r}$ riippumattomat. Edelleen suureella $\frac{1}{\sigma^2}\mathrm{SSE}$ on χ^2-jakauma $N-k-1$ vapausasteella. Vielä pitää näyttää, että suureella

$$\frac{1}{\sigma^2}(\mathbf{Ab}-\mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{Ab}-\mathbf{d})$$

on χ^2-jakauma q vapausasteella, kun H_0 on tosi.

Koska vektorilla $\mathbf{b}$ on $\mathrm{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$-jakauma, on vektorilla $\mathbf{Ab}-\mathbf{d}$ jakauma

$$\mathrm{N}(\mathbf{A}\boldsymbol{\beta}-\mathbf{d}, \sigma^2\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T) = \mathrm{N}(\mathbf{0}_q, \sigma^2\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T).$$

Selvästi $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$ on symmetrinen ja positiivisemidefiniitti. Koska $\mathbf{A}$:lla on täysi rivirangi ja $\mathbf{X}^T\mathbf{X}$ on ei-singulaarinen, on myös $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$ ei-singulaarinen ja siis positiividefiniitti. Schurin lauseen mukaan se voidaan kirjoittaa muotoon $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, missä $\mathbf{Q}$ on ortogonaalimatriisi ja $\boldsymbol{\Lambda}$ on lävistäjämatriisi, jonka lävistäjällä ovat $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$:n (positiiviset) ominaisarvot. Näin ollen on matriisilla $(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}$ neliöjuuri $\mathbf{Q}\sqrt{\boldsymbol{\Lambda}^{-1}}\mathbf{Q}^T =: \mathbf{B}$, missä lävistäjämatriisi $\sqrt{\boldsymbol{\Lambda}^{-1}}$ saadaan matriisista $\boldsymbol{\Lambda}^{-1}$ ottamalla sen lävistäjäalkioista neliöjuuret. Ilmeisesti $\mathbf{B}$ on symmetrinen ei-singulaarinen matriisi. Nyt vektorin $\mathbf{B}(\mathbf{Ab}-\mathbf{d})$ jakauma on

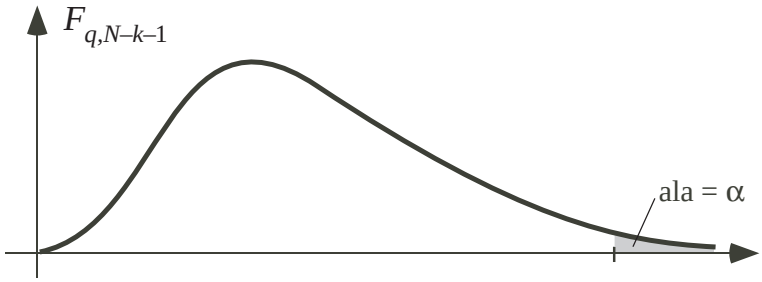
$$\mathrm{N}(\mathbf{0}_q, \sigma^2\mathbf{B}\mathbf{B}^{-2}\mathbf{B}^T) = \mathrm{N}(\mathbf{0}_q, \sigma^2\mathbf{I}_q).$$

Suureella

$$\frac{1}{\sigma^2}(\mathbf{Ab}-\mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{Ab}-\mathbf{d}) = \frac{1}{\sigma^2}(\mathbf{B}(\mathbf{Ab}-\mathbf{d}))^T\mathbf{B}(\mathbf{Ab}-\mathbf{d})$$

on näin ollen $\chi^2(q)$-jakauma. □

Hypoteesin testaaminen sujuu tavalliseen tapaan. Merkitsevyystaso α kiinnitetään. Jos testisuure osuu varjostetulle häntäalueelle kuvassa, hylätään H_0. Mitä ”huonommin” H_0 pitää paikkansa, sitä suurempi pyrkii $\|\mathbf{Ab}-\mathbf{d}\|$ ja F-testisuure olemaan. *level of significance*



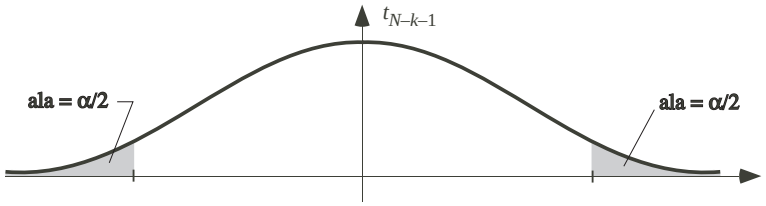
Jos $q = 1$, on hypoteesi H_0 muotoa $\mathbf{a}^T\boldsymbol{\beta} = d$, ja F-testin suure on

$$\frac{N-k-1}{\mathrm{SSE}} \frac{(\mathbf{a}^T\mathbf{b} - d)^2}{\mathbf{a}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{a}} = \frac{1}{s^2} \frac{(\mathbf{a}^T\mathbf{b} - d)^2}{\mathbf{a}^T\mathbf{C}\mathbf{a}}$$

Hypoteesin testaamiseen voidaan yhtä hyvin ottaa $N - k - 1$ vapausasteella t-jakautunut suure

$$t_i = \frac{\mathbf{a}^T\mathbf{b} - d}{\sqrt{s^2\mathbf{a}^T\mathbf{C}\mathbf{a}}}$$

eli F-testisuureen neliöjuuri. Hypoteesi $\mathrm{H}_0 : \mathbf{a}^T\boldsymbol{\beta} = d$ hylätään, jos suure t_i osuu varjostetulle alueelle kuvassa. Tätä t-suuretta käyttäen voidaan myös tehdä yksipuolisia testejä.



Valinta $\mathbf{a}^T = (0, \ldots, 0, 1, 0, \ldots, 0)$, missä 1 on i:s alkio, antaa hypoteesin $\mathrm{H}_0 : \beta_i = 0$, joka testaa faktorin x_i tarpeellisuutta mallissa. Tällöin F-testisuure on

$$\frac{b_i^2}{s^2 c_{ii}}$$

ja t-testisuure on sen neliöjuuri.

Koko mallin käyttökelpoisuutta puolestaan testaa hypoteesi

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

Jos tätä H_0:aa ei hylätä, ovat käytetyt faktorit huonoja selittäjiä, ts. koko malli voitaisiin yhtä hyvin korvata vakiolla + kohinalla (eli mallilla $y = \beta_0 + \epsilon$). Vastaava $\mathbf{A}$-matriisi on $\left(\begin{array}{c|c} \mathbf{0}_k & \mathbf{I}_k \end{array}\right)$ ja $\mathbf{d} = \mathbf{0}_k$. Tehdään datamatriisissa ja $\mathbf{b}$-vektorissa samanlainen ositus:

$$\mathbf{X} = \left(\begin{array}{c|c} \mathbf{1}_N & \mathbf{D} \end{array}\right) \quad \text{ja} \quad \mathbf{b} = \begin{pmatrix} b_0 \\ \hline \mathbf{b}_1 \end{pmatrix}.$$

design matrix

(Matriisi $\mathbf{D}$ on muuten ns. *suunnittelumatriisi*, jota tarvitaan vielä jatkossa.) Tässä $\mathbf{1}_N$ on N-vektori, jonka kaikki alkiot ovat ykkösiä. Silloin $\mathbf{A}\mathbf{b} = \mathbf{b}_1$ ja

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} \mathbf{1}_N^T \\ \hline \mathbf{D}^T \end{pmatrix} \left(\begin{array}{c|c} \mathbf{1}_N & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{D} \\ \hline \mathbf{D}^T\mathbf{1}_N & \mathbf{D}^T\mathbf{D} \end{array}\right).$$

Edelleen tällöin

$$(\mathbf{A}\mathbf{b} - \mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{d}) = \mathbf{b}_1^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}\mathbf{b}_1 =: \mathrm{SSR},$$

ns. *regression neliösumma.*

Tunnetun lohkomatriisien kääntökaavan[1] mukaan $(\mathbf{X}^T\mathbf{X})^{-1}$:n oikea alalohko eli siis $\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T$ on

$$\left(\mathbf{D}^T\mathbf{D} - \mathbf{D}^T\mathbf{1}_N\frac{1}{N}\mathbf{1}_N^T\mathbf{D}\right)^{-1} = (\mathbf{D}^T\mathbf{M}_N\mathbf{D})^{-1}.$$

Matriisi $\mathbf{M}_N = \mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T$ on ns. *keskitysmatriisi.* Sillä kertominen vähentää vektorista sen keskiarvon. Koska $\mathbf{M}_N\mathbf{1}_N = \mathbf{0}_N$, niin

$$\begin{aligned}\mathrm{SSR} &= \mathbf{b}_1^T\mathbf{D}^T\mathbf{M}_N\mathbf{D}\mathbf{b}_1 = (b_0\mathbf{1}_N + \mathbf{D}\mathbf{b}_1)^T\mathbf{M}_N(b_0\mathbf{1}_N + \mathbf{D}\mathbf{b}_1) = (\mathbf{X}\mathbf{b})^T\mathbf{M}_N\mathbf{X}\mathbf{b} \\ &= \hat{\mathbf{y}}^T\mathbf{M}_N\hat{\mathbf{y}}.\end{aligned}$$

Koska edelleen $\mathbf{X}^T\mathbf{r} = \mathbf{0}_{k+1}$ (kohdan 1.2 harjoitustehtävä 3b), niin $\mathbf{1}_N^T\mathbf{r} = 0$ (tarkastellaan vain $\mathbf{X}$:n ensimmäistä saraketta) ja $\hat{\mathbf{y}}^T\mathbf{r} = \mathbf{b}^T\mathbf{X}^T\mathbf{r} = 0$. Näin ollen

$$\mathbf{r}^T\mathbf{M}_N\hat{\mathbf{y}} = \mathbf{r}^T\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)\hat{\mathbf{y}} = \mathbf{r}^T\hat{\mathbf{y}} - \frac{1}{N}\mathbf{r}^T\mathbf{1}_N\mathbf{1}_N^T\hat{\mathbf{y}} = 0$$

ja

$$\mathbf{r}^T\mathbf{M}_N\mathbf{r} = \mathbf{r}^T\left(\mathbf{I}_N - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^T\right)\mathbf{r} = \mathbf{r}^T\mathbf{r} - \frac{1}{N}\mathbf{r}^T\mathbf{1}_N\mathbf{1}_N^T\mathbf{r} = \mathbf{r}^T\mathbf{r} = \mathrm{SSE}.$$

Ns. *kokonaisneliösumma* *total sum of squares*

$$\mathrm{SST} := (\mathbf{M}_N\mathbf{y})^T(\mathbf{M}_N\mathbf{y})$$

on näin hajotettavissa residuaalin neliösumman ja regression neliösumman summaksi:

$$\begin{aligned}\mathrm{SST} &= (\mathbf{M}_N\mathbf{y})^T(\mathbf{M}_N\mathbf{y}) = \mathbf{y}^T\mathbf{M}_N\mathbf{y} = (\mathbf{r} + \hat{\mathbf{y}})^T\mathbf{M}_N(\mathbf{r} + \hat{\mathbf{y}}) \\ &= \mathbf{r}^T\mathbf{M}_N\mathbf{r} + \hat{\mathbf{y}}^T\mathbf{M}_N\hat{\mathbf{y}} = \mathrm{SSE} + \mathrm{SSR}.\end{aligned}$$

Jakamalla neliösumma vapausasteellaan saadaan aina vastaava *keskineliö:* *mean square*

[1] Jos neliömatriisin $\left(\begin{array}{c|c}\mathbf{U} & \mathbf{V} \\ \hline \mathbf{W} & \mathbf{Z}\end{array}\right)$ neliölohko $\mathbf{U}$ on ei-singulaarinen, niin neliömatriisi on ei-singulaarinen jos ja vain jos $\mathbf{Z} - \mathbf{W}\mathbf{U}^{-1}\mathbf{V}$ on ei-singulaarinen. Sen käänteismatriisi on

$$\left(\begin{array}{c|c}\mathbf{U}^{-1} + \mathbf{U}^{-1}\mathbf{V}(\mathbf{Z} - \mathbf{W}\mathbf{U}^{-1}\mathbf{V})^{-1}\mathbf{W}\mathbf{U}^{-1} & -\mathbf{U}^{-1}\mathbf{V}(\mathbf{Z} - \mathbf{W}\mathbf{U}^{-1}\mathbf{V})^{-1} \\ \hline -(\mathbf{Z} - \mathbf{W}\mathbf{U}^{-1}\mathbf{V})^{-1}\mathbf{W}\mathbf{U}^{-1} & (\mathbf{Z} - \mathbf{W}\mathbf{U}^{-1}\mathbf{V})^{-1}\end{array}\right).$$

$$\mathrm{MSE} := \frac{\mathrm{SSE}}{N-k-1} = s^2\ , \quad \mathrm{MSR} := \frac{\mathrm{SSR}}{k}\ , \quad \mathrm{MST} := \frac{\mathrm{SST}}{N-1}$$

(*residuaalin keskineliö, regression keskineliö* ja *kokonaiskeskineliö*).

Hypoteesin $\mathrm{H}_0 : \beta_1 = \cdots = \beta_k = 0$ testisuure on näin ollen $\mathrm{MSR}/\mathrm{MSE}$ ja sillä on Lauseen 1.1 mukaan F-jakauma vapausastein k ja $N-k-1$. Vastahypoteesi on

$$\mathrm{H}_1 : \text{"ainakin yksi parametreistä } \beta_1, \ldots, \beta_k \text{ on } \neq 0\text{"}.$$

ANalysis Of VAriance

H_0:n hylkääminen merkitsee, että ainakin yhdellä faktorilla on merkittävää vaikutusta vasteeseen. *Varianssianalyysitaulu* eli ANOVA-taulu sisältää kaiken tämän:

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
regressio	k	SSR	MSR	$\frac{\mathrm{MSR}}{\mathrm{MSE}}$	pienin α:n arvo, jolla H_0 hylätään
residuaali	$N-k-1$	SSE	MSE		
kokonaisvariaatio	$N-1$	SST	MST		

coefficient of determination

Neliösummista saadaan myös ns. *determinaatiokerroin* eli *selitysaste*

$$\frac{\mathrm{SSR}}{\mathrm{SST}} =: R^2.$$

R^2 ilmoittaa kuinka suuren suhteellisen osan vastevektorin otosvarianssista regressio selittää. R^2:n neliöjuuri R on ns. *yhteiskorrelaatiokerroin*. Jotkut käyttävät mieluummin ns. *korjattua determinaatiokerrointa*

adjusted R^2

$$1 - \frac{\mathrm{MSE}}{\mathrm{MST}} =: R^2_{\mathrm{A}} = 1 - (1-R^2)\frac{N-1}{N-k-1}.$$

R^2_{A} ilmoittaa kuinka paljon suhteellisesti $V(\epsilon)$:n estimoidusta arvosta voidaan poistaa sovittamalla jokin muu kuin H_0:n mukainen malli $y = \beta_0 + \epsilon$ verrattuna siihen $V(\epsilon)$:n estimoituun arvoon ($= \mathrm{MST}$), joka ko. mallin avulla saadaan.

Esimerkki

Kohdan 1.2. esimerkin vasteen kuvaaja (sivulla 9) näyttää neliölliseltä x_2:n suhteen, varsinkin kun $x_1 = 0.3$, joten tuntuu järkevältä lisätä termin $\beta_{22}x_2^2$ malliin. Koska koejärjestelyissä ei ole testattu kuin kahdella x_1:n arvolla, ei voida ottaa mukaan x_1:n toisen asteen termejä. Yhteisvaikutukset saadaan kertomalla lineaariset ja neliölliset x_2 faktorit ja lineaariset x_1 faktorit keskenään. Määritä siis kolmannen kertaluvun malli

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \beta_{122} x_1 x_2^2 + \epsilon$$

kohdan 1.2 esimerkin datalle. Testaa faktorin $x_1 x_2^2$ tarpeellisuus ja koko mallin käyttökelpoisuus. Laske mallin determinaatiokertoimet R^2 ja R^2_{A}.

Ratkaisu

Nyt siis

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \\ \beta_{22} \\ \beta_{122} \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_{12} \\ b_{22} \\ b_{122} \end{pmatrix}.$$

Datamatriisi ja vastevektori ovat:

$$\mathbf{X} = \begin{pmatrix} 1 & 0.3 & 1 & 0.3 & 1 & 0.3 \\ 1 & 0.3 & 1 & 0.3 & 1 & 0.3 \\ 1 & 0.7 & 1 & 0.7 & 1 & 0.7 \\ 1 & 0.7 & 1 & 0.7 & 1 & 0.7 \\ 1 & 0.3 & 5 & 1.5 & 25 & 7.5 \\ 1 & 0.3 & 5 & 1.5 & 25 & 7.5 \\ 1 & 0.7 & 5 & 3.5 & 25 & 17.5 \\ 1 & 0.7 & 5 & 3.5 & 25 & 17.5 \\ 1 & 0.3 & 9 & 2.7 & 81 & 24.3 \\ 1 & 0.3 & 9 & 2.7 & 81 & 24.3 \\ 1 & 0.7 & 9 & 6.3 & 81 & 56.7 \\ 1 & 0.7 & 9 & 6.3 & 81 & 56.7 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 5.63 \\ 6.42 \\ 1.38 \\ 1.94 \\ 11.57 \\ 12.16 \\ 5.72 \\ 4.69 \\ 12.68 \\ 13.31 \\ 8.28 \\ 7.73 \end{pmatrix}.$$

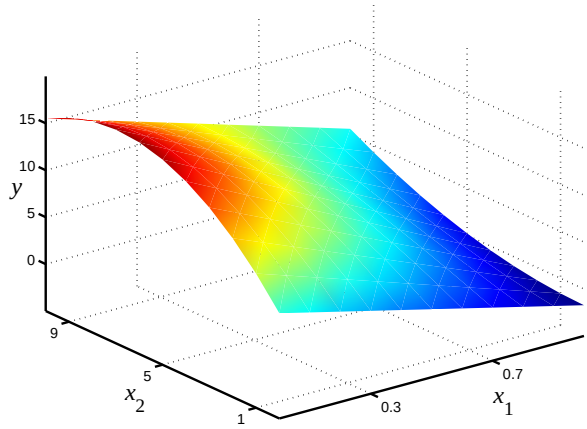
Mallin kertoimet ovat:

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 6.21 \\ -7.93 \\ 3.33 \\ -3.29 \\ -0.24 \\ 0.31 \end{pmatrix}.$$

Varianssianalyysin matriisi $\mathbf{C}$:

$$\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 4.20 & -7.24 & -1.81 & 3.11 & 0.15 & -0.26 \\ -7.24 & 14.49 & 3.11 & -6.23 & -0.26 & 0.52 \\ -1.81 & 3.11 & 1.12 & -1.93 & -0.11 & 0.18 \\ 3.11 & -6.23 & -1.93 & 3.86 & 0.18 & -0.37 \\ 0.15 & -0.26 & -0.11 & 0.18 & 0.01 & -0.02 \\ -0.26 & 0.52 & 0.18 & -0.37 & -0.02 & 0.04 \end{pmatrix}.$$

Malli on siis $\hat{y} = 6.21 - 7.93 \cdot x_1 + 3.33 \cdot x_2 - 3.29 \cdot x_1x_2 - 0.24 \cdot x_2^2 + 0.31 \cdot x_1x_2^2$.
Vastepinta on lievästi kaareva:



”Ennustetut” vasteet ja residuaali:

$$\hat{\mathbf{y}} = \mathbf{Xb} = \begin{pmatrix} 6.03 \\ 6.03 \\ 1.66 \\ 1.66 \\ 11.87 \\ 11.87 \\ 5.21 \\ 5.21 \\ 13.00 \\ 13.00 \\ 8.01 \\ 8.01 \end{pmatrix}, \quad \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}} = \begin{pmatrix} -0.40 \\ 0.40 \\ -0.28 \\ 0.28 \\ -0.30 \\ 0.30 \\ 0.52 \\ -0.52 \\ -0.32 \\ 0.32 \\ 0.28 \\ -0.28 \end{pmatrix}.$$

Residuaalin neliösumma:

$$\text{SSE} = \mathbf{r}^T\mathbf{r} = 1.52.$$

Varianssin estimaatti:

$$s^2 = \frac{\text{SSE}}{N-k-1} = \frac{\text{SSE}}{12-5-1} = 0.25.$$

Estimoidut parametrien keskihajonnat:

$$\begin{aligned} \text{se}(b_0) &= \sqrt{s^2 c_{00}} = \sqrt{0.25 \cdot 4.20} = 1.03, \\ \text{se}(b_1) &= \sqrt{s^2 c_{11}} = \sqrt{0.25 \cdot 14.49} = 1.92, \\ \text{se}(b_2) &= \sqrt{s^2 c_{22}} = \sqrt{0.25 \cdot 1.12} = 0.53, \\ \text{se}(b_{12}) &= \sqrt{s^2 c_{33}} = \sqrt{0.25 \cdot 3.86} = 0.99, \\ \text{se}(b_{22}) &= \sqrt{s^2 c_{44}} = \sqrt{0.25 \cdot 0.01} = 0.05, \\ \text{se}(b_{122}) &= \sqrt{s^2 c_{55}} = \sqrt{0.25 \cdot 0.04} = 0.10. \end{aligned}$$

Testataan kolmannen asteen faktorin $x_1x_2^2$ tarpeellisuus. Hypoteesit siis ovat:

$$\mathrm{H}_0 : \beta_{122} = 0, \quad \mathrm{H}_1 : \beta_{122} \neq 0,$$

eli matriiseina:

$$\mathrm{H}_0 : \mathbf{a}^T\boldsymbol{\beta} = d, \quad \mathrm{H}_1 : \mathbf{a}^T\boldsymbol{\beta} \neq d,$$

kun

$$\mathbf{a}^T = \mathbf{A} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad d = 0.$$

F-testisuure on:

$$\mathrm{F} = \frac{1}{s^2}\frac{(\mathbf{a}^T\mathbf{b} - d)^2}{\mathbf{a}^T\mathbf{C}\mathbf{a}} = 10.32,$$

jolla siis on F-jakauma vapausastein $q = 1$ ja $N - k - 1 = 12 - 5 - 1 = 6$. Testin voi myös tehdä t-testinä jolloin suureeksi saadaan:

$$t = \frac{\mathbf{a}^T\mathbf{b} - d}{\sqrt{s^2\mathbf{a}^T\mathbf{C}\mathbf{a}}} = 3.21,$$

vapausastein $N - k - 1 = 6$. Taulokoista tai ohjelmistoista nähdään, että molemmilla testeillä pienin merkitsevyystaso α, jolla H_0 hylätään on 0.018.

Koko mallin käyttökelpoisuus testataan hypoteesein

$$\mathrm{H}_0 : \beta_1 = \cdots = \beta_k = 0, \quad \mathrm{H}_1 : \beta_i \neq 0 \text{ jollakin termillä } i \in \{1, 2, 12, 22, 122\},$$

eli matriiseina:

$$\mathrm{H}_0 : \mathbf{A}\boldsymbol{\beta} = \mathbf{d}, \quad \mathrm{H}_1 : \mathbf{A}\boldsymbol{\beta} \neq \mathbf{d},$$

kun

$$\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Testisuure on

$$F = \frac{(N - k - 1)}{q \cdot \mathrm{SSE}}(\mathbf{A}\mathbf{b} - \mathbf{d})^T(\mathbf{A}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{d})$$

$$= \frac{(12 - 5 - 1)}{5 \cdot \mathrm{SSE}}(\mathbf{A}\mathbf{b} - \mathbf{d})^T(\mathbf{A}\mathbf{C}\mathbf{A}^T)^{-1}(\mathbf{A}\mathbf{b} - \mathbf{d}) = 143.33.$$

Pienin merkitsevyys α, jolla H_0 hylätään, on $3.72 \cdot 10^{-6}$. Malli siis toimii.

Lasketaan F-suure uudestaan, nyt ANOVA-taulun kautta. Otetaan käyttöön matriisi

$$M_{12} = \mathbf{I}_{12} - \frac{1}{12}\mathbf{1}_{12}\mathbf{1}_{12}^T.$$

Nyt voimme laskea regression neliösumman:

$$\mathrm{SSR} = \hat{\mathbf{y}}^T \mathrm{M}_{12} \hat{\mathbf{y}} = 181.91,$$

ja edelleen kokonaisneliösumman:

$$\mathrm{SST} = \mathrm{SSE} + \mathrm{SSR} = 1.52 + 181.91 = 183.44.$$

Neliösummia vastaavat keskineliöt:

$$\mathrm{MSE} = \frac{\mathrm{SSE}}{N-k-1} = \frac{1.52}{12-5-1} = 0.25,$$

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{k} = \frac{181.91}{5} = 36.38,$$

$$\mathrm{MST} = \frac{\mathrm{SST}}{N-1} = \frac{183.44}{12-1} = 16.68$$

ja F-testisuure:

$$\mathrm{F} = \frac{\mathrm{MSR}}{\mathrm{MSE}} = \frac{36.38}{0.25} = 143.33.$$

Tulokset kootaan ANOVA-tauluksi.

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
regressio	5	181.91	36.38		
residuaali	6	1.52	0.25	143.33	$3.72 \cdot 10^{-6}$
kokonaisvariaatio	11	183.44	16.68		

Lopuksi lasketaan vielä determinaatiokerroin ja korjattu determinaatiokerroin:

$$R^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = 0.9917, \quad R^2_{\mathrm{A}} = 1 - \frac{\mathrm{MSE}}{\mathrm{MST}} = 0.9848.$$

Harjoitustehtävät

1. Taulussa on annettu mittaustulokset eräälle transistorien valmistukseen liittyvälle koesarjalle. Faktori x_1 on käsittelyaika ja faktori x_2 ionikonsentraatio, kumpikin koodattuna välille $[-1, 1]$. (Datan koodauksesta lisää luvun 2 kohdassa 1. Koodaus on tässä tarpeen kertalukueroista johtuen, sillä tyypillinen käsittelyaika on luokkaa 200 min ja ionikonsentraatio $4 \cdot 10^{-14}$.) Vaste

y on vahvistuskerroin.

x_1	x_2	y
-1	-1	1004
1	-1	1636
-1	0.6667	852
1	0.6667	1506
0	-0.4444	1272
0	-0.7222	1270
0	0.6667	1269
-1	-0.1667	903
1	-0.1667	1555
0	-1	1260
0	0.9444	1146
0	-0.1667	1276
0	1	1225
0.1667	-0.1667	1321

(a) Estimoi mallin $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ parametrit sekä ϵ:n varianssi.

(b) Testaa parametrien merkitsevyys $\mathrm{H}_0 : \beta_i = 0$ ($i = 0, 1, 2$) kaksipuolisella t-testillä. Ilmoita kunkin parametrin kohdalla pienin riski, jolla se otetaan mukaan malliin.

(c) Ennusta vasteen arvo suurimmalla kokeissa olleella käsittelyajalla ja ionikonsentraatiolla.

(d) Testaa vielä hypoteesi $\mathrm{H}_0 : \beta_1 = 350$ F-testillä. Ilmoita pienin riski, jolla H_0 voidaan hylätä.

2. Minkälainen matriisi $\mathbf{A}$ ja vektori $\mathbf{d}$ tarvitaan seuraavissa lineaarisissa hypoteesin testauksissa? Malli on $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \epsilon$.

(a) Testataan, onko kaikkien faktorien vaikutus sama, ts. onko $\mathrm{H}_0 : \beta_1 = \cdots = \beta_k$.

(b) Malli on aikanaan estimoitu suurella koesarjalla ja sitä käytetään myöhemmin harvoin. Kun sitä käytetään, testataan ensin varmuuden vuoksi mallin toimivuus suorittamalla uudet kokeet kahdella tyypillisellä faktoriyhdelmällä:

x_1	x_2	$\cdots$	x_k	y
x_1'	x_2'	$\cdots$	x_k'	y'
x_1''	x_2''	$\cdots$	x_k''	y''

Saatujen tulosten perusteella testataan hypoteesi H_0 : ”malli toimii ko. kokeiden perusteella”.

3. Taulussa on annettu mittaustulokset erääseen kemialliseen prosessiin liittyvälle koesarjalle. x_1 on lämpötila ja x_2 reaktantin väkevyys, koodattuna ns. CCD-kokeen muotoon. (CCD-kokeista lisää myöhemmin.) Vaste y on tuotosprosentti.

x_1	x_2	y
-1	-1	43
1	-1	78
-1	1	69
1	1	73
$-\sqrt{2}$	0	48
$\sqrt{2}$	0	76
0	$-\sqrt{2}$	65
0	$\sqrt{2}$	74
0	0	76
0	0	79
0	0	83
0	0	81

Malli on neliöllinen eli 2. kertalukua:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \epsilon.$$

   (a) Testaa mallin merkitsevyys, ts. laadi ANOVA-taulu.

   (b) Laske myös kertoimet R^2, R, R^2_{A} ja R_{A}.

4. Mallin merkitsevyystestin F-testisuure esitetään usein muodossa

$$F = \frac{R^2(N-k-1)}{(1-R^2)k}.$$

   (a) Totea, että kaava on oikein.

   (b) N-vektorin $\mathbf{y}$ ns. *otosvarianssi* eli alkioiden varianssi on $V_{\mathbf{y}} = \frac{1}{N}\mathbf{y}^T\mathbf{M}_N\mathbf{y}$. Toisaalta kahden N-vektorin $\mathbf{y}_1$ ja $\mathbf{y}_2$ ns. *otoskovarianssi* eli alkioittainen kovarianssi on $\mathrm{cov}_{\mathbf{y}_1\mathbf{y}_2} = \frac{1}{N}\mathbf{y}_1^T\mathbf{M}_N\mathbf{y}_2$ ja *otoskorrelaatiokerroin* on

$$\rho_{\mathbf{y}_1\mathbf{y}_2} = \frac{\mathrm{cov}_{\mathbf{y}_1\mathbf{y}_2}}{\sqrt{V_{\mathbf{y}_1}}\sqrt{V_{\mathbf{y}_2}}}.$$

   Totea, että $R = \rho_{\mathbf{y}\hat{\mathbf{y}}}$, ts. yhteiskorrelaatiokerroin on vastevektorin ja ennustetun vastevektorin otoskorrelaatiokerroin (tästä sen nimi).

1.4 Mallin epäsopivuuden testaus toistokokein

lack of fit

Regressiomallin *epäsopivuus* tarkoittaa sitä, että lisäämällä uusia faktoreita tai entisistä faktoreista muodostettuja uusia (korkeampiasteisia) faktoreita, residuaalia voidaan ”pienentää”.

Jos siis malli

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon$$

on epäsopiva, tarkoittaa se sitä, että *jokin* laajennettu malli

$$y = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{z}^T\boldsymbol{\gamma} + \epsilon',$$

missä $\mathbf{z} = (z_1, \ldots, z_\ell)^T$ on uusien tai entisistä kertomalla tai muuten saatujen faktorien muodostama vektori ja $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_\ell)^T$ on uusi parametrivektori, on ”parempi”.

Huomaa, että sovitettaessa jälkimmäinen malli pienimmän neliösumman keinolla vastevektoriin $\mathbf{y}$ ja datamatriisiin

$$\left(\begin{array}{c|c} \mathbf{X} & \mathbf{Z} \end{array} \right),$$

missä $\mathbf{X}$ on aikaisempi datamatriisi ja $\mathbf{Z}$ uusia faktoreita vastaavista sarakkeista muodostettu ”jatke”, eivät parametrit $\boldsymbol{\beta}$ saa (välttämättä) samoja arvoja kuin sovitettaessa alkuperäistä mallia. Tämä johtuu siitä, että uudet selittävät faktorit voivat osittain selittää samoja tekijöitä, kuin vanhat faktorit. Uusien faktorien osuutta voidaan erotella kirjoittamalla

$$\mathbf{Z} = \hat{\mathbf{Z}} + (\mathbf{Z} - \hat{\mathbf{Z}}),$$

missä matriisi $\hat{\mathbf{Z}}$ saadaan ennustamalla matriisin $\mathbf{Z}$ sarakkeet vanhaa mallia käyttäen ja matriisissa $(\mathbf{Z} - \hat{\mathbf{Z}})$ on se, mitä uudet faktorit selittävät ja vanhat eivät. Matriisin $\hat{\mathbf{Z}}$ sarakkeet ovat siis vanhan mallin datamatriisin sarakkeiden lineaariyhdelmiä,

$$\hat{\mathbf{Z}} = \mathbf{X}\mathbf{E},$$

ja matriisin $(\mathbf{Z} - \hat{\mathbf{Z}})$ sarakkeet ja vanhan mallin datamatriisin sarakkeet ovat ortogonaalisia,

$$\mathbf{X}^T(\mathbf{Z} - \hat{\mathbf{Z}}) = \mathbf{O},$$

joten $\mathbf{E} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ (ns. *aliasmatriisi*).

Hypoteesi, jonka mukaan malli *ei* ole tarkasteltujen uusien faktorien kannalta epäsopiva, on näin ollen

$$\mathrm{H}_0 : (\mathbf{Z} - \hat{\mathbf{Z}})\boldsymbol{\gamma} = \mathbf{0}_N.$$

Vastahypoteesi on tietysti $\mathrm{H}_1 : (\mathbf{Z} - \hat{\mathbf{Z}})\boldsymbol{\gamma} \neq \mathbf{0}_N$. Jos halutaan testata, onko mallia yleensä ottaen mahdollista parantaa, pitää verrata virhetermin aiheuttamaa varianssia vasteen selittämättä jääneen osan aiheuttamaan varianssiin. Jos

jälkimmäinen on ”huomattavasti” suurempi, on mallin sopivuutta mahdollista parantaa uusia faktoreita käyttäen.

replicate observations

Testisuure tällaiselle testaukselle saadaan, jos mukana on *toistokokeita*, ts. datamatriisissa on samoja rivejä. Oletetaan, että datamatriisissa $\mathbf{X}$ on *erilaisia* rivejä m kappaletta. Huomaa, että $m \geq k+1$, muuten $\mathbf{X}^T\mathbf{X}$ on singulaarinen. Kootaan mainitut erilaiset rivit $m \times (k+1)$-matriisiksi $\mathbf{X}_1$. Silloin voidaan kirjoittaa

$$\mathbf{X} = \mathbf{T}\mathbf{X}_1$$

sopivasti valitulle $N \times m$-matriisille $\mathbf{T}$. Huomaa, että $\mathbf{T}$:llä on täysi sarakerangi, ts. sen sarakkeet ovat lineaarisesti riippumattomat, ja että $\mathbf{T}\mathbf{1}_m = \mathbf{1}_N$. Itse asiassa $\mathbf{T}$ saadaan identiteettimatriisista $\mathbf{I}_m$ toistamalla sen rivejä sopivasti.

Laajin mahdollinen malli, joksi alkuperäinen malli voidaan täydentää, saadaan, kun lisätään datamatriisiin $\mathbf{X}$ suurin mahdollinen määrä sarakkeita, jotka ovat aikaisemmista sarakkeista lineaarisesti riippumattomia, samalla säilyttäen toistetut rivit. Tällaiseen malliin ei nimittäin voida lisätä yhtäkään uutta selittäjää, joka ei, toistokokeiden puitteissa, riippuisi lineaarisesti aikaisemmista. Täydennetään $\mathbf{X}_1$ ensin $m \times m$-matriisiksi lisäämällä siihen $m-k-1$ aikaisemmista lineaarisesti riippumatonta saraketta:

$$\left(\begin{array}{c|c} \mathbf{X}_1 & \mathbf{Z}_1 \end{array} \right) =: \mathbf{X}_2.$$

Matriisin $\mathbf{X}$ täydennys on sen jälkeen $N \times m$-matriisi

$$\mathbf{T}\mathbf{X}_2 = \left(\begin{array}{c|c} \mathbf{T}\mathbf{X}_1 & \mathbf{T}\mathbf{Z}_1 \end{array} \right) = \left(\begin{array}{c|c} \mathbf{X} & \mathbf{Z} \end{array} \right) =: \mathbf{X}_3,$$

missä $\mathbf{Z} = \mathbf{T}\mathbf{Z}_1$.

Alkuperäisestä datamallista (*Malli I*)

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

saadaan näin laajennettu datamalli (*Malli II*)

$$\mathbf{y} = \mathbf{X}_3 \left(\frac{\boldsymbol{\beta}}{\boldsymbol{\gamma}} \right) + \boldsymbol{\epsilon}' = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}'$$

Mallista II saatu ennustevektori on

$$\hat{\mathbf{y}}_{\mathrm{II}} = \mathbf{X}_3(\mathbf{X}_3^T\mathbf{X}_3)^{-1}\mathbf{X}_3^T\mathbf{y} = \mathbf{T}\mathbf{X}_2(\mathbf{X}_2^T\mathbf{T}^T\mathbf{T}\mathbf{X}_2)^{-1}\mathbf{X}_2^T\mathbf{T}^T\mathbf{y} = \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T\mathbf{y},$$

joka *ei riipu* $\mathbf{Z}_1$*:stä, ts. siitä, miten* $\mathbf{X}_1$ *täydennetään!* Näin ollen saatava testi ei myöskään riipu mallin laajennustavasta, kunhan toistojen rakenne (eli $\mathbf{T}$) säilytetään. Mallista II saatava residuaali on

$$\mathbf{r}_{\mathrm{II}} = (\mathbf{I}_N - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T)\mathbf{y}$$

ja tämän residuaalin neliösumma on

$$\mathbf{r}_{\mathrm{II}}^T\mathbf{r}_{\mathrm{II}} =: \mathrm{SSPE},$$

ns. *puhtaan virheen neliösumma.* *sum of squares for pure error*

Yritetään selittää Mallin I residuaalivektori

$$\mathbf{r} = (\mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{y}$$

Mallin II avulla. Jos tämä onnistuu tarpeeksi hyvin, ei Malli I ole sopiva, vaan se voidaan täydentää sopivammaksi. Merkitään lyhyyden vuoksi

$$\mathbf{P} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T \quad \text{ja} \quad \mathbf{R} = \mathbf{I}_N - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T.$$

Silloin todetaan helpolla laskulla, että $\mathbf{P}$ ja $\mathbf{R}$ ovat symmetrisiä idempotentteja matriiseja ja että

$$\mathbf{RX} = \mathbf{O} \quad , \quad \mathbf{RZ} = \mathbf{O} \quad , \quad \mathbf{RP} = \mathbf{PR} = \mathbf{R} \quad , \quad \mathbf{PX} = \mathbf{O},$$

$$\mathrm{rank}(\mathbf{P}) = \mathrm{trace}(\mathbf{P}) = N - k - 1,$$

$$\mathrm{rank}(\mathbf{R}) = \mathrm{trace}(\mathbf{R}) = N - m.$$

Selitettäessä Mallin II avulla Mallin I residuaalia $\mathbf{r}$ on selittämättä jäävä residuaali $\mathbf{r}_{\mathrm{II}} = \mathbf{R}\mathbf{y} = \mathbf{RP}\mathbf{y} = \mathbf{R}\mathbf{r}$, jonka neliösumma on nimenomaan SSPE. Kokonaisneliösumma on puolestaan $\mathbf{r}^T\mathbf{r}$ eli Mallin I residuaalin neliösumma SSE. Edelleen regression neliösumma tässä selitysyrityksessä on

$$\mathrm{SSE} - \mathrm{SSPE} =: \mathrm{SSLOF},$$

ns. *epäsopivuuden neliösumma.* Matriisimuodossa *sum of squares due to lack of fit*

$$\mathrm{SSLOF} = \mathbf{y}^T(\mathbf{P} - \mathbf{R})\mathbf{y}.$$

Matriisi $\mathbf{P} - \mathbf{R}$ on symmetrinen idempotentti matriisi, jonka rangi on

$$\mathrm{trace}(\mathbf{P} - \mathbf{R}) = \mathrm{trace}(\mathbf{P}) - \mathrm{trace}(\mathbf{R}) = m - k - 1.$$

Neliösumma SSPE vastaa sitä osaa residuaalivarianssista, joka johtuu virhetermistä. Siihen ei voida vaikuttaa mallilla, olipa tämä kuinka hyvä tahansa. SSLOF vastaa taas sitä osaa residuaalivarianssista, joka johtuu mallin huonosta selittävyydestä eli epäsopivuudesta.

Mutta: *Residuaali* $\mathbf{r}$ *ei ole oikeaa vasteen tyyppiä,* sillä sillä on singulaarinen normaalijakauma (ts. $\mathbf{P}$ on singulaarinen). Näin ollen saatujen neliösummien jakaumat ja vapausasteet sekä niihin perustuva ANOVA katsotaan erikseen. Huomaa kuitenkin, että SSPE on Mallin II residuaalin neliösumma, joten sillä on χ^2-jakauma $N - m$ vapausasteella.

Lause 1.2. *Jos hypoteesi* $\mathrm{H}_0 : \mathbf{PZ}\boldsymbol{\gamma} = \mathbf{0}_N$ *on tosi Mallille II, niin suureella*

$$\frac{\mathrm{SSLOF}/(m-k-1)}{\mathrm{SSPE}/(N-m)}$$

on F-jakauma vapausastein $m-k-1$ *ja* $N-m$ *(olettaen tietysti, että* $m > k+1$*).*

Todistus. Pitää näyttää, että SSLOF ja SSPE ovat riippumattomasti χ^2-jakautuneet vapausastein $m-k-1$ ja $N-m$, vastaavasti. Hypoteesin H_0 voimassaollessa

$$(\mathbf{P}-\mathbf{R})\mathbf{y} = (\mathbf{P}-\mathbf{R})(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}') = (\mathbf{P}-\mathbf{R})\boldsymbol{\epsilon}'$$

ja

$$\mathbf{R}\mathbf{y} = \mathbf{R}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}') = \mathbf{R}\boldsymbol{\epsilon}'.$$

Koska $\mathbf{P}-\mathbf{R}$ ja $\mathbf{R}$ ovat symmetrisiä idempotentteja matriiseja, $\mathbf{R}(\mathbf{P}-\mathbf{R}) = \mathbf{O}_N$ ja satunnaisvektorilla $\boldsymbol{\epsilon}'$ on $\mathrm{N}(\mathbf{0}_N, \sigma^2\mathbf{I}_N)$-multinormaalijakauma, on lause oikea. □

Lauseessa esiintyvä $\mathbf{Z}$ on tietysti se laajin mahdollinen, jolla alkuperäistä datamatriisia $\mathbf{X}$ täydennetään. Vastahypoteesi on $\mathrm{H}_1 : \mathbf{PZ}\boldsymbol{\gamma} \neq \mathbf{0}_N$.

SSPE:llä on siis vapausasteita $N-m$ ja SSLOF:llä $m-k-1$. Vastaavat keskineliöt ovat näin ollen

$$\frac{\mathrm{SSPE}}{N-m} =: \mathrm{MSPE} \quad \text{ja} \quad \frac{\mathrm{SSLOF}}{m-k-1} =: \mathrm{MSLOF}$$

(*puhtaan virheen keskineliö* ja *epäsopivuuden keskineliö*). Varianssianalyysitaulu on siten

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
epäsopivuus	$m-k-1$	SSLOF	MSLOF	$\frac{\mathrm{MSLOF}}{\mathrm{MSPE}}$	pienin α:n arvo, jolla H_0 hylätään
puhdas virhe	$N-m$	SSPE	MSPE		
residuaali	$N-k-1$	SSE	MSE		

Jos hypoteesia H_0 ei hyväksytä, voidaan mallia parantaa täydentämällä sitä sopivilla faktoreilla. Huomaa, että jos erityisesti täydentävät faktorit ovat entisistä laskien saatuja korkean asteen faktoreita, niin edellä esitetty toistettujen rivien säilyminen täydennettäessä on automaattista. Näin ollen esitetty testi on erityisen sopiva juuri tällaista täydennystä ajatellen. Jos mallia päätetään täydentää, ei tietystikään mukaan välttämättä kannata ottaa ”kaikkia mahdollisia” lisäselittäjiä, vaan vain sopivasti valitut lisäfaktorit. Tilastolliset ohjelmistot tarjoavatkin korkeampiasteisten faktorien osalta monia (puoli)automaattisia lisäys- ja valintamenetelmiä (ns. *askeltava regressio*).

Epäsopivuustesti voidaan tehdä muutenkin kuin toistokokeita käyttäen. Matriisista $\mathbf{T}$:kin käytettiin nimittäin vain sen ominaisuuksia

(i) $\mathbf{T}$:llä on täysi sarakerangi (jotta $\mathbf{T}^T\mathbf{T}$ olisi ei-singulaarinen) ja

(ii) hajotelmassa $\mathbf{X} = \mathbf{T}\mathbf{X}_1$ on $\mathbf{X}_1$:llä täysi sarakerangi $k+1$ (jotta se voidaan täydentää ei-singulaariseksi $m \times m$-matriisiksi $\mathbf{X}_2$).

Mikä tahansa matriisi, joka toteuttaa nämä ehdot, kelpaisi periaatteessa $\mathbf{T}$:n tilalle. Tällöin ei kyseessä olisi välttämättä enää koetoistoihin perustuva testi. Valitsemalla $\mathbf{T}$ eri tavoin saadaan erilaisia epäsopivuustestejä, tosin näin saadut testit ovat yleensä heikompia kuin toistoihin perustuvat. Ks. myös CHRISTENSEN ja artikkeli JOGLEKAR, G. & SCHUENMEYER, J.H. & LARICCIA, V.: Lack-of-Fit Testing When Replicates Are Not Available, *The American Statistician* **43** (1989), 135–143.

Esimerkki

Testaame kohdan 1.2. mallin sopivuuden toistokokein. Datamatriisi $\mathbf{X}$ on

$$\mathbf{X} = \begin{pmatrix} 1 & 0.3 & 1 \\ 1 & 0.3 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.3 & 5 \\ 1 & 0.3 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.3 & 9 \\ 1 & 0.3 & 9 \\ 1 & 0.7 & 9 \\ 1 & 0.7 & 9 \end{pmatrix},$$

jossa on 6 erilaista riviä, eli $m = 6$. Erilaisten rivien matriisi $\mathbf{X}_1$:

$$\mathbf{X}_1 = \begin{pmatrix} 1 & 0.3 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.3 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.3 & 9 \\ 1 & 0.7 & 9 \end{pmatrix}.$$

Matriisi $\mathbf{T}$, joka toteuttaa yhtälön $\mathbf{X} = \mathbf{T}\mathbf{X}_1$ on selvästi:

$$\mathbf{T} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Käyttöön otetaan matriisit $\mathbf{P}$ ja $\mathbf{R}$, jotka siis määriteltiin:

$$\mathbf{P} = \mathbf{I}_N - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \quad \mathbf{R} = \mathbf{I}_N - \mathbf{T}(\mathbf{T}^T\mathbf{T})^{-1}\mathbf{T}^T.$$

Nyt saadaan mallin II residuaali:

$$\mathbf{r}_{\mathrm{II}} = \mathbf{R}\mathbf{y} = \begin{pmatrix} -0.40 \\ 0.40 \\ -0.28 \\ 0.28 \\ -0.30 \\ 0.30 \\ 0.52 \\ -0.52 \\ -0.32 \\ 0.32 \\ 0.28 \\ -0.28 \end{pmatrix},$$

puhtaan virheen neliösumma:

$$\mathrm{SSPE} = \mathbf{r}_{\mathrm{II}}^T \mathrm{r}_{\mathrm{II}} = 1.52,$$

ja edelleen epäsopivuuden neliösumma:

$$\mathrm{SSLOF} = \mathrm{SSE} - \mathrm{SSPE} = 9.30 - 1.52 = 7.78.$$

Vastaavat keskineliöt:

$$\mathrm{MSPE} = \frac{\mathrm{SSPE}}{N - m} = \frac{1.52}{12 - 6} = 0.25,$$

$$\text{MSLOF} = \frac{\text{SSLOF}}{m-k-1} = \frac{7.78}{6-2-1} = 2.59.$$

Hypoteesit epäsopivuuden testaukseen ovat

$$\text{H}_0 : \text{'Malli on sopiva'}, \quad \text{H}_1 : \text{'Malli on epäsopiva'},$$

eli matriiseina:

$$\text{H}_0 : \mathbf{PZ}\boldsymbol{\gamma} = \mathbf{0}_N, \quad \text{H}_1 : \mathbf{PZ}\boldsymbol{\gamma} \neq \mathbf{0}_N.$$

ja niitä vastaava F-testisuure:

$$F = \frac{\text{MSLOF}}{\text{MSPE}} = \frac{2.59}{0.25} = 10.21.$$

Pienin merkitsevyys, jolla H_0 hylätään, on 0.009. Mallia voi siis parantaa, joka onkin todettu luvun 1.3. esimerkissä. Kootaan tulokset varianssianalyysitauluksi.

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
epäsopivuus	3	7.78	2.59		
puhdas virhe	6	1.52	0.25	10.21	0.009
residuaali	9	9.30	1.03		

Harjoitustehtävät

1. Tehtävän 1.3-3 mittaustuloksissa on toistoja, ns *keskustoistoja*. Testaa näitä käyttäen mallin sopivuus.

2. Minkälainen on epäsopivuustestin matriisi $\mathbf{T}$ seuraavissa tilanteissa, liittyen siihen, miten kokeet on järjestetty datamatriisissa? Matriisi $\mathbf{T}^T\mathbf{T}$ on aina $m \times m$-lävistäjämatriisi, millainen?

   (a) Kokeet tehdään toistoineen järjestyksessä, ts. ensin ensimmäisen yhdelmän kokeet toistoineen (t_1 kpl), sitten toisen yhdelmän kokeet toistoineen (t_2 kpl), jne.

   (b) Toistot tehdään sarjassa, ts. tehdään ensin kokeet kaikilla eri faktoriyhdelmillä ja sitten toistetaan samaa koesarjaa t kertaa.

   (c) Ensin tehdään kokeet kaikilla eri faktoriyhdelmillä ja sitten vasta toistot yhdelmä kerrallaan järjestyksessä, ts. ensin tehdään ensimmäisen yhdelmän (mahdolliset) toistot (s_1 toistettua kertaa), sitten toisen yhdelmän (mahdolliset) toistot (s_2 toistettua kertaa), jne.

1.5 Mallin riittävyys

model adequacy

Mallin *riittävyys* tarkoittaa sitä, että mallin yhteydessä sovitut oletukset (riippumattomuudet, normaalisuus, varianssien samuus, jne.) pitävät paikkansa. Koska ANOVA saattaa olla hyvinkin herkkä poikkeamille näistä oletuksista, testataan usein oletusten voimassaoloa. Testauksessa käytetään residuaalia $\mathbf{r}$. Jos malli on riittävä, ei residuaalissa ole juurikaan muuta virhettä kuin $\mathrm{N}(\mathbf{0}_N, \sigma^2\mathbf{I}_N)$-jakautuneen satunnaismuuttujan $\boldsymbol{\epsilon}$ aiheuttamaa "kohinaa". Ellei näin ole, on mahdollisia syitä useita.

Epänormaalisuus

Jos virhetermin $\boldsymbol{\epsilon}$ jakauma ei olekaan multinormaali, vaan jotakin muuta, ei ANOVAn tuloksiin ole paljoakaan luottamista. Epänormaalisuuden toteamiseksi voidaan residuaalivektorin $\mathbf{r}$ komponenttien olettaa olevan otos ja tutkia, voiko tämän otoksen katsoa olevan peräisin normaalijakautuneesta satunnaismuuttujasta, esimerkiksi piirtämällä vastaava pylväsdiagrammi. Parempi menettely on järjestää residuaalit suuruusjärjestykseen

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(N)}$$

normal probability plot, Q-Q plot

ja piirtää ns. *normaalitodennäköisyyskuvion* pisteet

$$\left(r_{(j)}, \Phi^{-1}\left(\frac{j}{N+1}\right)\right), \quad j = 1, \ldots, N,$$

missä Φ^{-1} on käänteinen standardinormaalikertymä. Kuvion pisteiden pitäisi olla kutakuinkin samalla suoralla. Usein kertymän argumentti $j/(N+1)$ korvataan jollain muulla kaavalla, esimerkiksi

$$\frac{j-\frac{1}{2}}{N}, \quad \frac{j-\frac{1}{3}}{N+\frac{1}{3}}, \quad \text{tai} \quad \frac{j-\frac{3}{8}}{N+\frac{1}{4}}.$$

Jakaumatestausta varten on olemassa omia tilastollisiakin testejä, mm. *Kolmogorovin ja Smirnovin testi* ja *Cramerin ja von Misesin testi.*

outliers

Toisinaan sattuu, että yksi tai useampikin $\mathbf{r}$:n komponenteista on itseisarvoltaan muita huomattavasti suurempi. Tällaisia komponentteja kutsutaan *ulkolaisiksi.* Ne ovat merkkejä joko siitä, että vastaava koe on virheellinen tai sitten siitä, että muut kokeet onkin tehty tilanteen kannalta huonolla alueella. Ulkolaisten esiintyessä on aina selvitettävä, mistä ne johtuvat, sillä ANOVA on osoittautunut herkäksi ulkolaisten esiintymiselle. Useinkaan ei ole selvää, onko poikkeava komponentti ulkolainen vai sattuman oikusta syntynyt poikkeava arvo. Ulkolaisten tunnistamiseksi on erityisiä testejäkin. Koska residuaalin i:nnen komponentin

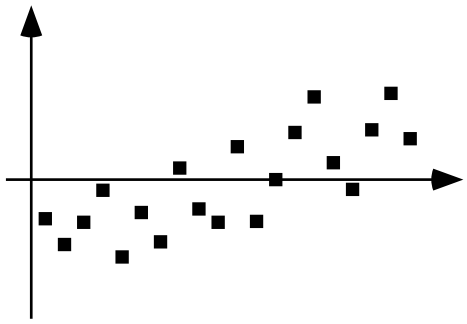
varianssi on $V(r_i) = \sigma^2(1 - h_{ii})$, missä h_{ii} on matriisin $\mathbf{H} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ i:s lävistäjäalkio, eräs tällainen testi on laskea ns. *studentoitu residuaali*

$$\frac{1}{\sqrt{s^2(1 - h_{ii})}} r_i =: e_i.$$

Satunnaismuuttuja e_i on likimain standardinormaalisesti jakautunut, joten nyrkkisääntö on: jos e_i on itseisarvoltaan ≥ 3, on syytä epäillä sitä ulkolaiseksi.

Korrelointi

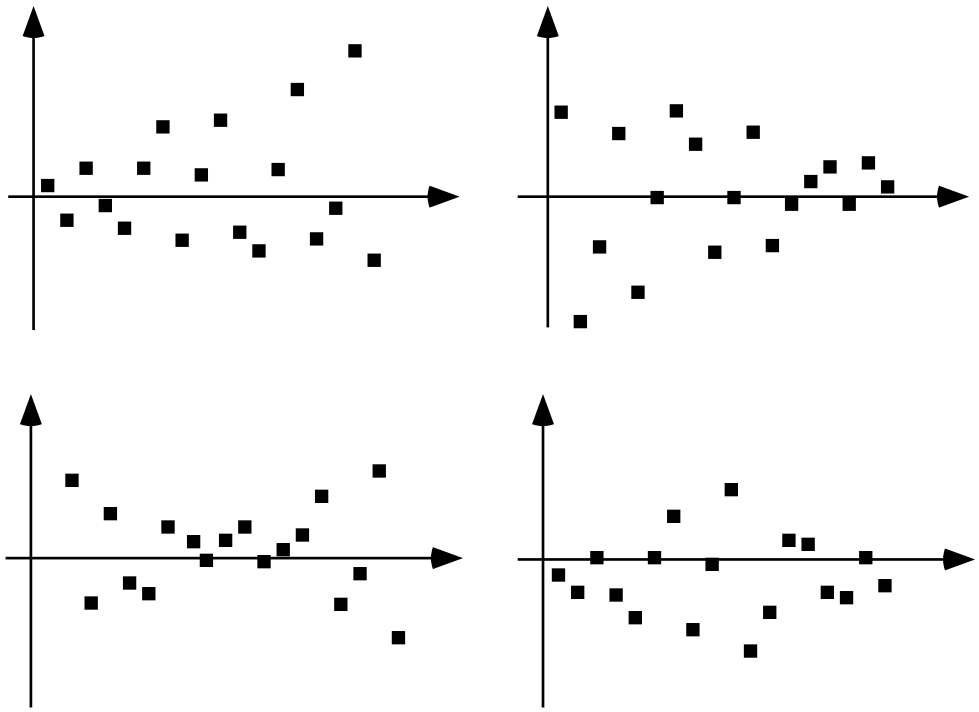
Vaikka $\boldsymbol{\epsilon}$:n komponentit olisivatkin normaalijakautuneita, voi niiden välillä olla korrelaatiota, ts. ne eivät ole riippumattomia. Asia paljastuu usein piirrettäessä $\mathbf{r}$:n komponentit kokeiden fysikaalisen suoritusjärjestyksen funktiona. Siksi järjestys on syytä olla satunnaistettu ja tallennettu. Korrelointi näkyy tällaisesta kuvaajasta usein selvästi seuraavan kuvion tapaan, sillä se johtuu tällöin ajallisesta yhteydestä.



Residuaali ei saa korreloida muidenkaan muuttujien kanssa eikä erityisesti vasteen kanssa (tehtävä 1.4-3). Piirtämällä residuaali vs. ennustettu vaste paljastuu usein mallin tämänkaltainen riittämättömyys. Eo. kuvio olisi nytkin hälyttävä.

Heterogeeninen varianssi

Vaikkei epänormaalisuutta tai korrelaatiota esiinnykään, voi malli osoittautua riittämättömäksi vielä sen vuoksi, että $\boldsymbol{\epsilon}$:n komponenttien varianssit eivät ole samat. Usein tämä näkyy piirrettäessä $\mathbf{r}$:n komponentit suoritusjärjestyksen funktiona kuten edellä: hajonta on jossakin suurempaa kuin muualla. Alla on neljä hälyttävää kuviota.

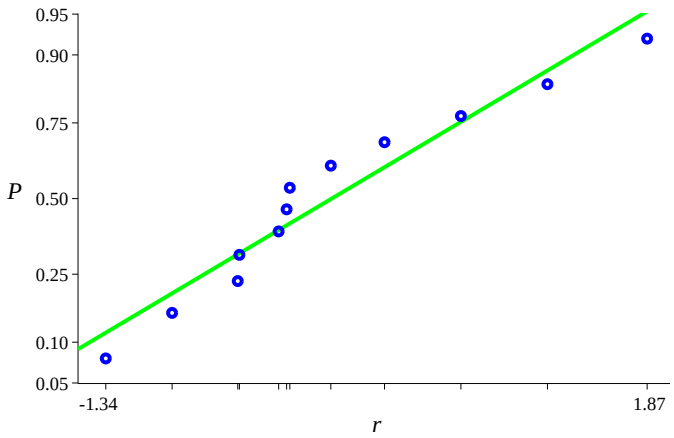


Esimerkki

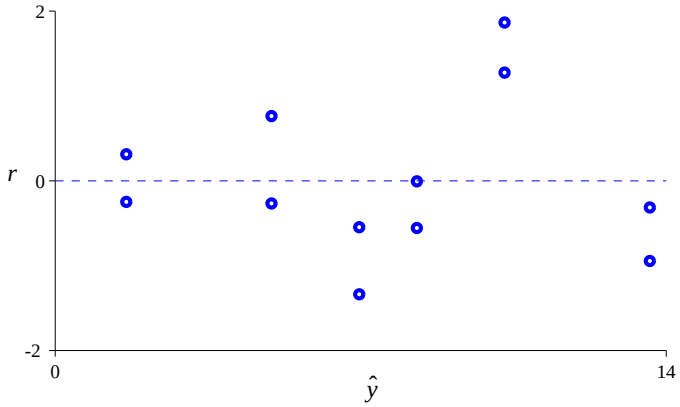
Tarkistamme kohdan 1.2 esimerkin lineaarisen mallin riittävyyden. Kokeet tehtiin järjestyksessä $(12, 9, 3, 10, 4, 7, 2, 11, 8, 6, 1, 5)$. Mallinnuksen tulokset ovat

x_1	x_2	y	$\hat{y}$	r	h_{ii}	e
0.7	9	7.73	8.29	−0.56	0.29	−0.65
0.3	9	12.68	13.62	−0.94	0.29	−1.10
0.7	1	1.38	1.63	−0.25	0.29	−0.29
0.3	9	13.31	13.62	−0.31	0.29	−0.37
0.7	1	1.94	1.63	0.31	0.29	0.36
0.7	5	5.72	4.96	0.76	0.17	0.82
0.3	1	6.42	6.97	−0.55	0.29	−0.64
0.7	9	8.28	8.29	−0.01	0.29	−0.01
0.7	5	4.69	4.96	−0.27	0.17	−0.29
0.3	5	12.16	10.30	1.87	0.17	2.01
0.3	1	5.63	6.97	−1.34	0.29	−1.56
0.3	5	11.57	10.30	1.28	0.17	1.37

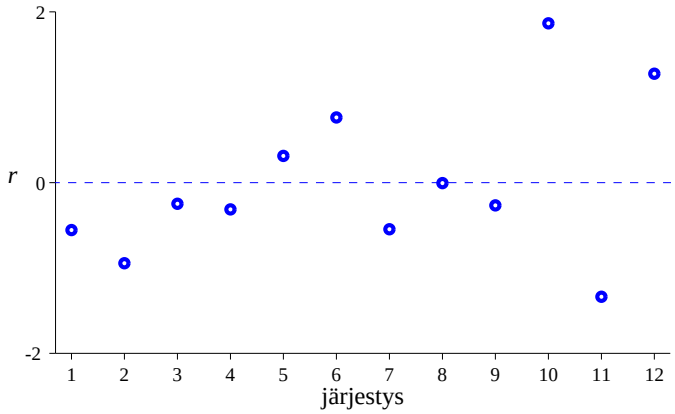
Taulukon viimeisessä sarakkeessa olevat arvot ovat kaikki itseisarvoltaan reilusti alle 3, joten ulkolaisia ei näytä olevan. Normaalitodennäköisyyskuvio on



Residuaalin kuvaaja ennustetun vasteen funktiona on



Mitään sen kummempaa merkkiä epänormaalisuudesta tai korrelaatiosta näistä kuvista ei paljastu. Piirretään vielä residuaalin kuvaaja koejärjestyksen funktiona:



Tiettyä huolta virheen varianssin kasvamisesta ajan kuluessa voisi tämän kuvan perusteella tuntea.

Harjoitustehtävät

1. Tarkista tehtävän 1.3-1 mallin riittävyys.
2. Tarkista tehtävän 1.3-3 mallin riittävyys.
3. Todista residuaalin ja ennustetun vasteen riippumattomuus.
4. Johda kaava $V(r_i) = \sigma^2(1 - h_{ii})$ ja todista, että $0 \leq h_{ii} \leq 1$.

Luku 2

KOKEIDEN SUUNNITTELU

2.1 Datan muunnokset

Jos $\mathbf{X} = (\,\mathbf{1}_N \,|\, \mathbf{D}\,)$ on $N \times (k+1)$-datamatriisi (jossa matriisi $\mathbf{D}$ on *suunnittelumatriisi*) ja $\mathbf{L}$ on ei-singulaarinen $(k+1) \times (k+1)$-matriisi, jonka ensimmäinen sarake on $(1, 0, \ldots, 0)^T$, niin $\mathbf{XL}$ on myös $N \times (k+1)$-datamatriisi, joka sisältää saman informaation kuin $\mathbf{X}$. Tällainen muunnos on datan *affiinimuunnos*. Matriisi $\mathbf{L}$ on siis muotoa

affine transformation

$$\mathbf{L} = \left(\begin{array}{c|c} 1 & \boldsymbol{\ell}^T \\ \hline \mathbf{0}_k & \mathbf{K} \end{array}\right),$$

missä $\boldsymbol{\ell}$ on k-vektori ja $\mathbf{K}$ on ei-singulaarinen $k \times k$-matriisi, ja uusi datamatriisi

$$\mathbf{XL} = (\,\mathbf{1}_N \,|\, \mathbf{1}_N\boldsymbol{\ell}^T + \mathbf{DK}\,)$$

on oikean muotoinen.

Koska

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{XLL}^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

on uutta datamatriisia $\mathbf{XL}$ vastaava parametrivektori $\mathbf{L}^{-1}\boldsymbol{\beta} =: \boldsymbol{\gamma}$. Edelleen pienimmän neliösumman keinon antama parametrivektorin $\boldsymbol{\gamma}$ estimaatti on

$$\mathbf{g} = ((\mathbf{XL})^T\mathbf{XL})^{-1}(\mathbf{XL})^T\mathbf{y} = \mathbf{L}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{L}^T)^{-1}\mathbf{L}^T\mathbf{X}^T\mathbf{y} = \mathbf{L}^{-1}\mathbf{b}$$

ja ”uusi” ennustevektori on $\mathbf{XLg} = \mathbf{Xb} = \hat{\mathbf{y}}$ eli sama kuin ”vanha”. Näin ollen myöskin residuaali pysyy datan affiinimuunnoksessa samana ja itse asiassa kaikki neliösummat SSE, SST ja SSR sekä vastaavat keskineliöt pysyvät samana. Mallin merkitsevyys ei siis muutu. Toisaalta

$$V(\mathbf{g}) = \sigma^2((\mathbf{XL})^T\mathbf{XL})^{-1} = \sigma^2\mathbf{L}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{L}^T)^{-1}$$

voi hyvinkin olla ”edullisempaa” muotoa kuin $V(\mathbf{b})$, ts. $\mathbf{g}$:n komponenttien välillä voi olla vähemmän korrelaatiota kuin $\mathbf{b}$:n komponenttien välillä ja niiden variansit voivat olla pienempiä kuin $\mathbf{b}$:n komponenttien variansit.

Tavallinen ensimmäisen kertaluvun mallin datan affiinimuunnos on *skaalaus*, jota vastaava matriisi $\mathbf{K}$ on lävistäjämatriisi,

$$\mathbf{K} = \begin{pmatrix} 1/p_1 & & \\ & \ddots & \\ & & 1/p_k \end{pmatrix},$$

jonka lävistäjäalkiot ovat nollasta eroavia. Selittäjä x_i korvautuu skaalauksessa selittäjällä $\frac{x_i}{p_i} + \ell_i$, missä ℓ_i on vektorin $\boldsymbol{\ell}$ i:s alkio. Skaalaus ei vaikuta mallin ensimmäisen kertaluvun termien merkitsevyyteen, sillä yhtälöstä $\gamma_i = p_i\beta_i$ $(1 \leq i \leq k)$ seuraa, että hypoteesi $\mathrm{H}_0 : \beta_i = 0$ on yhtä kuin hypoteesi $\mathrm{H}_0 : \gamma_i = 0$.

Skaalauksen tarkoituksena on, paitsi vaihtaa selittävien muuttujien asteikot ”sopivammiksi”, muuntaa keinotekoisesti selittävät muuttujat tyypillisten arvojensa suhteen samaan asemaan. Tyypillisten arvojen kokoero saattaa nimittäin alunperin olla monia dekadeja, mikä aiheuttaa mm. numeerista epätarkkuutta laskuissa. Tällöin suoritetaan ensin skaalaus ja vasta sitten mallin sovitus.

Erityinen skaalauksen muoto on datan *standardointi*, jossa p_i on x_i:n otoshajonta ja ℓ_i on x_i:n otosvariaatiokertoimen vastaluku, ts. valitaan

$$p_i = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{ji} - \overline{x}_i)^2} \quad \text{ja} \quad \ell_i = -\overline{x}_i/p_i,$$

missä $\overline{x}_i$ on x_i:n otoskeskiarvo. Matriisimuodossa:

$$p_i = (\mathbf{D}^T\mathbf{M}_N\mathbf{D})_{ii}\,, \quad \boldsymbol{\ell}^T = -\frac{1}{N}\mathbf{1}_N^T\mathbf{D}\mathbf{K}.$$

Jos data on kunkin faktorin osalta tasavälistä, käytetään usein *koodausta*, joka myös on eräs skaalauksen muoto. Tällöin

$$p_i = \frac{\max(x_{1i}, \ldots, x_{Ni}) - \min(x_{1i}, \ldots, x_{Ni})}{2}$$

$$\ell_i = -\frac{1}{p_i}\frac{\min(x_{1i}, \ldots, x_{Ni}) + \max(x_{1i}, \ldots, x_{Ni})}{2}.$$

Lähinnä koodausta käytetään tilanteessa, missä kullakin faktorilla on kaksi tasoa tai kolme tasavälistä tasoa, sillä tällöin koodatut arvot ovat $\{-1, 1\}$ tai $\{0, 1, -1\}$.

Esimerkki

Suorita standardointi ja koodaus kohdan 1.2 esimerkin datalle ja sovita malli ensin standardoituun ja sitten koodattuun dataan.

Ratkaisu

Standardointia varten lasketaan aluksi ensimmäisen asteen faktorien keskiarvot:

$$\begin{aligned}\bar{x}_1 &= \frac{1}{N}\sum_{i=1}^{N} x_{i1} = \frac{1}{12}(0.3 + 0.3 + 0.7 + \cdots + 0.7) = 0.50,\\ \bar{x}_2 &= \frac{1}{N}\sum_{i=1}^{N} x_{i2} = \frac{1}{12}(1 + 1 + \cdots + 9) = 5.00\end{aligned}$$

ja otoshajonnat:

$$\begin{aligned}p_{s1} &= \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{j1} - \bar{x}_1)^2} = \sqrt{\frac{(0.3-0.5)^2 + \cdots + (0.7-0.5)^2}{12-1}} = 0.21,\\ p_{s2} &= \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_{j2} - \bar{x}_2)^2} = \sqrt{\frac{(1-5)^2 + \cdots + (9-5)^2}{12-1}} = 3.41.\end{aligned}$$

Vektorin $\boldsymbol{\ell}_s$ komponentit:

$$\ell_{s1} = -\bar{x}_2/p_{s2} = -5.00/3.41 = -1.47.$$

$\mathbf{K}_s$-matriisi ja $\boldsymbol{\ell}_s$-vektori:

$$\mathbf{K}_s = \begin{pmatrix} 1/p_{s1} & 0 \\ 0 & 1/p_{s2} \end{pmatrix} = \begin{pmatrix} 1/0.21 & 0 \\ 0 & 1/3.41 \end{pmatrix} = \begin{pmatrix} 4.79 & 0 \\ 0 & 0.29 \end{pmatrix}, \quad \boldsymbol{\ell}_s = \begin{pmatrix} -2.40 \\ -1.47 \end{pmatrix}.$$

Saadaan matriisi $\mathbf{L}_s$:

$$\mathbf{L}_s = \left(\begin{array}{c|c} 1 & \boldsymbol{\ell}_s^T \\ \hline \mathbf{0}_2 & \mathbf{K}_s \end{array}\right) = \left(\begin{array}{c|cc} 1 & -2.40 & -1.47 \\ \hline 0 & 4.79 & 0 \\ 0 & 0 & 0.29 \end{array}\right).$$

Standardoidaan datamatriisi:

$$\mathbf{X}_s = \mathbf{X}\mathbf{L}_s = \begin{pmatrix} 1 & 0.3 & 1 \\ 1 & 0.3 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.3 & 5 \\ 1 & 0.3 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.3 & 9 \\ 1 & 0.3 & 9 \\ 1 & 0.7 & 9 \\ 1 & 0.7 & 9 \end{pmatrix} \begin{pmatrix} 1 & -2.40 & -1.47 \\ 0 & 4.79 & 0 \\ 0 & 0 & 0.29 \end{pmatrix} = \begin{pmatrix} 1 & -0.96 & -1.17 \\ 1 & -0.96 & -1.17 \\ 1 & 0.96 & -1.17 \\ 1 & 0.96 & -1.17 \\ 1 & -0.96 & 0 \\ 1 & -0.96 & 0 \\ 1 & 0.96 & 0 \\ 1 & 0.96 & 0 \\ 1 & -0.96 & 1.17 \\ 1 & -0.96 & 1.17 \\ 1 & 0.96 & 1.17 \\ 1 & 0.96 & 1.17 \end{pmatrix}.$$

Matriisi $\mathbf{C}_s$ on harvempi kuin alkuperäisellä datalla:

$$\mathbf{C}_s = (\mathbf{X}_s^T\mathbf{X}_s)^{-1} = \begin{pmatrix} 0.08 & 0.00 & 0.00 \\ 0.00 & 0.09 & 0.00 \\ 0.00 & 0.00 & 0.09 \end{pmatrix}.$$

Lasketaan parametrien estimaatit:

$$\mathbf{b}_s = \mathbf{C}_s\mathbf{X}_s^T\mathbf{y} = \begin{pmatrix} 7.63 \\ -2.79 \\ 2.84 \end{pmatrix}.$$

Virhetermin varianssin σ^2 estimaatti:

$$\mathrm{SSE}_s = (\mathbf{y} - \mathbf{X}_s\mathbf{b}_s)^T(\mathbf{y} - \mathbf{X}_s\mathbf{b}_s) = 9.30, \quad s_s^2 = \frac{\mathrm{SSE}_s}{N-k-1} = 1.03.$$

Estimoidut parametrien keskihajonnat:

$$\mathrm{se}(b_{s0}) = \sqrt{s_s^2 c_{s00}} = \sqrt{1.03 \cdot 0.08} = 0.29,$$
$$\mathrm{se}(b_{s1}) = \sqrt{s_s^2 c_{s11}} = \sqrt{1.03 \cdot 0.09} = 0.30,$$
$$\mathrm{se}(b_{s2}) = \sqrt{s_s^2 c_{s22}} = \sqrt{1.03 \cdot 0.09} = 0.30.$$

Koodaus:

$$p_{c1} = \frac{\max(x_{11}, \ldots, x_{N1}) - \min(x_{11}, \ldots, x_{N1})}{2} = \frac{0.7 - 0.3}{2} = 0.2,$$

$$p_{c2} = \frac{\max(x_{12},\ldots,x_{N2}) - \min(x_{12},\ldots,x_{N2})}{2} = \frac{9-1}{2} = 4,$$

$$\ell_{c1} = -\frac{1}{p_1}\frac{\min(x_{11},\ldots,x_{N1}) + \max(x_{11},\ldots,x_{N1})}{2} = -\frac{1}{0.2}\frac{0.3+0.7}{2} = -2.5,$$

$$\ell_{c2} = -\frac{1}{p_2}\frac{\min(x_{12},\ldots,x_{N2}) + \max(x_{12},\ldots,x_{N2})}{2} = -\frac{1}{4}\frac{1+9}{2} = -1.25.$$

$\mathbf{K}_c$-matriisi ja $\boldsymbol{\ell}_c$-vektori:

$$\mathbf{K}_c = \begin{pmatrix} 1/p_c & 0 \\ 0 & 1/p_c \end{pmatrix} = \begin{pmatrix} 1/0.2 & 0 \\ 0 & 1/4 \end{pmatrix} = \begin{pmatrix} 5 & 0 \\ 0 & 0.25 \end{pmatrix}, \quad \boldsymbol{\ell}_c = \begin{pmatrix} -2.5 \\ -1.25 \end{pmatrix}.$$

Saadaan matriisi $\mathbf{L}_c$:

$$\mathbf{L}_c = \left(\begin{array}{c|c} 1 & \boldsymbol{\ell}_c^T \\ \hline \mathbf{0}_2 & \mathbf{K}_c \end{array}\right) = \left(\begin{array}{c|cc} 1 & -2.5 & -1.25 \\ \hline 0 & 5 & 0 \\ 0 & 0 & 0.25 \end{array}\right).$$

Koodataan datamatriisi:

$$\mathbf{X}_c = \mathbf{X}\mathbf{L}_c = \begin{pmatrix} 1 & 0.3 & 1 \\ 1 & 0.3 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.7 & 1 \\ 1 & 0.3 & 5 \\ 1 & 0.3 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.7 & 5 \\ 1 & 0.3 & 9 \\ 1 & 0.3 & 9 \\ 1 & 0.7 & 9 \\ 1 & 0.7 & 9 \end{pmatrix} \begin{pmatrix} 1 & -2.5 & -1.25 \\ 0 & 5 & 0 \\ 0 & 0 & 0.25 \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 \\ 1 & -1 & -1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}.$$

Matriisi $\mathbf{C}_c$ on harvempi kuin alkuperäisellä datalla:

$$\mathbf{C}_c = (\mathbf{X}_c^T\mathbf{X}_c)^{-1} = \begin{pmatrix} 0.08 & 0.00 & 0.00 \\ 0.00 & 0.08 & 0.00 \\ 0.00 & 0.00 & 0.13 \end{pmatrix}.$$

Lasketaan parametrien estimaatit:

$$\mathbf{b}_c = \mathbf{C}_c\mathbf{X}_c^T\mathbf{y} = \begin{pmatrix} 7.63 \\ -2.67 \\ 3.33 \end{pmatrix}.$$

Virhetermin varianssin σ^2 estimaatti:

$$\mathrm{SSE}_c = (\mathbf{y} - \mathbf{X}_c\mathbf{b}_c)^T(\mathbf{y} - \mathbf{X}_c\mathbf{b}_c) = 9.30, \quad s_c^2 = \frac{\mathrm{SSE}_c}{N-k-1} = 1.03.$$

Estimoidut parametrien keskihajonnat:

$$\mathrm{se}(b_{c0}) = \sqrt{s_c^2 c_{c00}} = \sqrt{1.03 \cdot 0.08} = 0.29,$$

$$\mathrm{se}(b_{c1}) = \sqrt{s_c^2 c_{c11}} = \sqrt{1.03 \cdot 0.08} = 0.29,$$

$$\mathrm{se}(b_{c2}) = \sqrt{s_c^2 c_{c22}} = \sqrt{1.03 \cdot 0.13} = 0.36.$$

Esitetään vielä alkuperäinen, standardoitu ja koodattu data vierekkäin:

alkuperäinen		standardoitu		koodattu	
x_1	x_2	x_1	x_2	x_1	x_2
0.3	1	−0.96	−1.17	−1	−1
0.3	1	−0.96	−1.17	−1	−1
0.7	1	0.96	−1.17	1	−1
0.7	1	0.96	−1.17	1	−1
0.3	5	−0.96	0	−1	0
0.3	5	−0.96	0	−1	0
0.7	5	0.96	0	1	0
0.7	5	0.96	0	1	0
0.3	9	−0.96	1.17	−1	1
0.3	9	−0.96	1.17	−1	1
0.7	9	0.96	1.17	1	1
0.7	9	0.96	1.17	1	1

Harjoitustehtävät

1. Kohdan 1.3. esimerkissä JMP ohjelma vaihtaa automaattisesti mallin $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{22} x_2^2 + \beta_{122} x_1 x_2^2 + \epsilon$ muotoon $y = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \gamma_{12}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2) + \gamma_{22}(x_2 - \bar{x}_2)^2 + \gamma_{122}(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)^2 + \epsilon$. Muodosta käytetty affiinimuunnoksen matriisi $\mathbf{L}$ ja tarkista, että hypoteesien $\mathrm{H}_0 : \gamma_{122} = 0$ ja $\mathrm{H}_0 : \beta_{122} = 0$ testisuureet ovat samat.

2. Tehtävässä 1.3.1 oli annettu mittaustulokset eräälle transistorien valmistukseen liittyvälle koesarjalle. Data oli koodattu. Ensimmäisen alkuperäisen faktorin vaihtelukeskipiste on 225 ja toisen $4.36 \cdot 10^{-14}$, vastaavat vaihteluvälit ovat 60 sekä $0.72 \cdot 10^{-14}$.

(a) Muodosta käytetty skaalausmatriisi $\mathbf{L}$ ja etsi sitä käyttäen alkuperäinen datamatriisi $\mathbf{X}$.

(b) Laske $(\mathbf{X}^T\mathbf{X})^{-1}$ ja vertaa sitä koodatun datan vastaavaan matriisiin.

(c) Laske alkuperäisen mallin parametrit.

2.2 Ortogonaalisuus ja kiertosymmetrisyys

Suunnittelun sanotaan olevan *ortogonaalinen*, jos $\mathbf{X}^T\mathbf{X}$ on lävistäjämatriisi.

Lause 2.1. *Suunnittelu on ortogonaalinen täsmälleen silloin, kun*

i $\mathbf{D}^T\mathbf{D}$ *on lävistäjämatriisi ja*

ii $\mathbf{1}_N^T\mathbf{D} = \mathbf{0}_k^T$, *eli* $\mathbf{D}$*:n sarakesummat ovat nollia.*

Toisin sanoen, suunnittelu on ortogonaalinen täsmälleen silloin, kun faktoreita vastaavat sarakkeet ovat kohtisuorassa toisiaan vastaan ja myös vektoria $\mathbf{1}_N$ vastaan.

Todistus. Koska

$$\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c} \mathbf{1}_N^T \\ \hline \mathbf{D}^T \end{array}\right) \left(\begin{array}{c|c} \mathbf{1}_N^T & \mathbf{D} \end{array}\right) = \left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{D} \\ \hline \mathbf{D}^T\mathbf{1}_N & \mathbf{D}^T\mathbf{D} \end{array}\right),$$

niin ilmeisesti $\mathbf{X}^T\mathbf{X}$ on lävistäjämatriisi tarkalleen silloin, kun ehdot (i) ja (ii) toteutuvat. □

Ortogonaalista suunnittelua käytettäessä $V(\mathbf{b}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ on lävistäjämatriisi, ts. parametriestimaatit $b_0, \ldots, b_k$ ovat riippumattomat. Edelleen tällöin käänteismatriisin $\mathbf{C} = (\mathbf{X}^T\mathbf{X})^{-1}$ laskeminen on helppoa ja tarkkaa.

Ortogonaalinen suunnittelu on optimaalinen seuraavassa mielessä. Koodatun datamatriisin sarakkeet $\mathbf{x}_i$ toteuttavat epäyhtälön

$$\|\mathbf{x}_i\|^2 \leq N \quad (0 \leq i \leq k).$$

Voidaan näyttää (kts. harjoitustehtävä 1), että

$$c_{ii} \geq \frac{1}{\|\mathbf{x}_i\|^2} \quad (0 \leq i \leq k),$$

ja tätä epäyhtälöä käyttäen saadaan alaraja regressiomallin parametrien varianssien summalle:

$$\sum_i V(b_i) = \sigma^2 \mathrm{trace}(\mathbf{C}) = \sum_i c_{ii} \geq \sum_i \|\mathbf{x}_i\|^{-2} \geq \frac{k+1}{N}.$$

Tämä alaraja saavutetaan, kun suunnittelu on ortogonaalinen ja kaikki data-arvot ovat 1 tai -1.

rotatable

Ensimmäisen kertaluvun mallin suunnittelun sanotaan olevan *kiertosymmetrinen*, jos matriisi $\mathbf{X}^T\mathbf{X}$ säilyy samana, kun dataan tehdään mielivaltainen ortogonaalinen muunnos, ts. $\mathbf{X}^T\mathbf{X}$ on ”koordinaatistosta riippumaton”. Ortogonaalinen muunnos on sama, kuin muotoa

$$\mathbf{Q} = \left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K} \end{array}\right)$$

oleva affiinimuunnos, missä $\mathbf{K}$ on $k \times k$-ortogonaalimatriisi.

Lause 2.2. *Suunnittelu on kiertosymmetrinen täsmälleen silloin, kun*

(i) $\mathbf{D}$*:n sarakesummat ovat nollia, ts.* $\mathbf{1}_N^T\mathbf{D} = \mathbf{0}_k^T$ *ja*

(ii) $\mathbf{D}^T\mathbf{D}$ *on muotoa* $\lambda\mathbf{I}_k$*, missä* λ *on vakio.*

Todistus. Oletetaan, että suunnittelu on kiertosymmetrinen. Sovelletaan mielivaltaista ortogonaalimuunnosta:

$$\begin{aligned}
(\mathbf{XQ})^T\mathbf{XQ} = \mathbf{Q}^T\mathbf{X}^T\mathbf{XQ} &= \left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K}^T \end{array}\right)\left(\begin{array}{c} \mathbf{1}_N^T \\ \hline \mathbf{D}^T \end{array}\right)\left(\begin{array}{c|c} \mathbf{1}_N & \mathbf{D} \end{array}\right)\left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K} \end{array}\right) \\
&= \left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K}^T \end{array}\right)\left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{D} \\ \hline \mathbf{D}^T\mathbf{1}_N & \mathbf{D}^T\mathbf{D} \end{array}\right)\left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K} \end{array}\right) \\
&= \left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{D} \\ \hline \mathbf{K}^T\mathbf{D}^T\mathbf{1}_N & \mathbf{K}^T\mathbf{D}^T\mathbf{D} \end{array}\right)\left(\begin{array}{c|c} 1 & \mathbf{0}_k^T \\ \hline \mathbf{0}_k & \mathbf{K} \end{array}\right) \\
&= \left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{DK} \\ \hline \mathbf{K}^T\mathbf{D}^T\mathbf{1}_N & \mathbf{K}^T\mathbf{D}^T\mathbf{DK} \end{array}\right).
\end{aligned}$$

Jotta tämä olisi

$$\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c|c} N & \mathbf{1}_N^T\mathbf{D} \\ \hline \mathbf{D}^T\mathbf{1}_N & \mathbf{D}^T\mathbf{D} \end{array}\right),$$

on oltava

$$\mathbf{K}^T\mathbf{D}^T\mathbf{1}_N = \mathbf{D}^T\mathbf{1}_N \quad \text{ja} \quad \mathbf{K}^T\mathbf{D}^T\mathbf{DK} = \mathbf{D}^T\mathbf{D},$$

olipa $\mathbf{K}$ mikä tahansa ortogonaalimatriisi. Mutta, jotta kaikki ortogonaalimuunnokset pitäisivät $\mathbf{D}^T\mathbf{1}_N$:n samana, pitää sen olla $= \mathbf{0}_k$, ts. (i) pätee.

Toisaalta $\mathbf{D}^T\mathbf{D}$ on symmetrinen matriisi, joten se on diagonalisoitavissa ortogonaalimuunnoksella. Näin ollen $\mathbf{D}^T\mathbf{D}$:n on oltava valmiiksi lävistäjämatriisi.

Silloin taas $\mathbf{D}^T\mathbf{D}$:n lävistäjäalkiot voidaan permutoida mielivaltaiseen järjestykseen ortogonaalimuunnoksella. Näin ollen lävistäjäalkioiden on oltava samoja. Siispä myös (ii) pätee.

Selvästi suunnittelu on kiertosymmetrinen, jos (i) ja (ii) pätevät. □

Kiertosymmetrisessä suunnittelussa ei ole mahdollista ”parantaa” mallia siirtymällä ”uusiin koordinaatteihin”, ts. esimerkiksi $V(\mathbf{b})$ pysyy samana. Malli ei voi tällöin myöskään ”huonontua”. Erityisesti ennusteen varianssi

$$V(\hat{y}) = \sigma^2 \boldsymbol{\xi}^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi} = \sigma^2 \left(\frac{1}{N} + \frac{\xi_1^2 + \cdots + \xi_k^2}{\lambda} \right)$$

riippuu vain datavektorin $\boldsymbol{\xi}$ pituudesta. Tästä itse asiassa tulee nimi ”kiertosymmetrinen”: ennusteen varianssi ei riipu suunnasta.

Lauseista 2.1 ja 2.2 seuraa, että jokainen kiertosymmetrinen suunnittelu on myös ortogonaalinen, mutta ei kääntäen. Tärkeä ortogonaalisten/kiertosymmetristen suunnittelujen ominaisuus on se, että niistä faktoreita poistamalla eli *typistämällä* saadut suunnittelut ovat myös ortogonaalisia/kiertosymmetrisiä. (Tämä seuraa varsin suoraan yo. lauseista.)

Esimerkki

Selvitä, ovatko edellisessä esimerkissä lasketut datat ortogonaalisia tai kiertosymmetrisiä.

Ratkaisu

Suunnittelumatriisit ovat

$$\mathbf{D} = \begin{pmatrix} 0.3 & 1 \\ 0.3 & 1 \\ 0.7 & 1 \\ 0.7 & 1 \\ 0.3 & 5 \\ 0.3 & 5 \\ 0.7 & 5 \\ 0.7 & 5 \\ 0.3 & 9 \\ 0.3 & 9 \\ 0.7 & 9 \\ 0.7 & 9 \end{pmatrix}, \quad \mathbf{D}_s = \begin{pmatrix} -0.96 & -1.17 \\ -0.96 & -1.17 \\ 0.96 & -1.17 \\ 0.96 & -1.17 \\ -0.96 & 0 \\ -0.96 & 0 \\ 0.96 & 0 \\ 0.96 & 0 \\ -0.96 & 1.17 \\ -0.96 & 1.17 \\ 0.96 & 1.17 \\ 0.96 & 1.17 \end{pmatrix}, \quad \mathbf{D}_c = \begin{pmatrix} -1 & -1 \\ -1 & -1 \\ 1 & -1 \\ 1 & -1 \\ -1 & 0 \\ -1 & 0 \\ 1 & 0 \\ 1 & 0 \\ -1 & 1 \\ -1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Data on ortogonaalista mikäli $\mathbf{X}^T\mathbf{X}$ on lävistäjämatriisi. Lasketaan nämä matriisit:

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 12.00 & 6.00 & 60.00 \\ 6.00 & 3.48 & 30.00 \\ 60.00 & 30.00 & 428.00 \end{pmatrix},$$

$$\mathbf{X}_s^T\mathbf{X}_s = \begin{pmatrix} 12.00 & 0.00 & 0.00 \\ 0.00 & 11.00 & 0.00 \\ 0.00 & 0.00 & 11.00 \end{pmatrix},$$

$$\mathbf{X}_c^T\mathbf{X}_c = \begin{pmatrix} 12.00 & 0.00 & 0.00 \\ 0.00 & 12.00 & 0.00 \\ 0.00 & 0.00 & 8.00 \end{pmatrix}.$$

Nähdään, että standardoitu ja koodattu data ovat ortogonaalisia, mutta alkuperäinen ei.

Kiertosymmetrisyyden ensimmäinen ehto on $\mathbf{1}_N^T\mathbf{D} = \mathbf{0}_k^T$:

$$\mathbf{1}_N^T\mathbf{D} = \begin{pmatrix} 6.00 & 60.00 \end{pmatrix},$$

$$\mathbf{1}_N^T\mathbf{D}_s = \begin{pmatrix} 0.00 & 0.00 \end{pmatrix},$$

$$\mathbf{1}_N^T\mathbf{D}_c = \begin{pmatrix} 0.00 & 0.00 \end{pmatrix}.$$

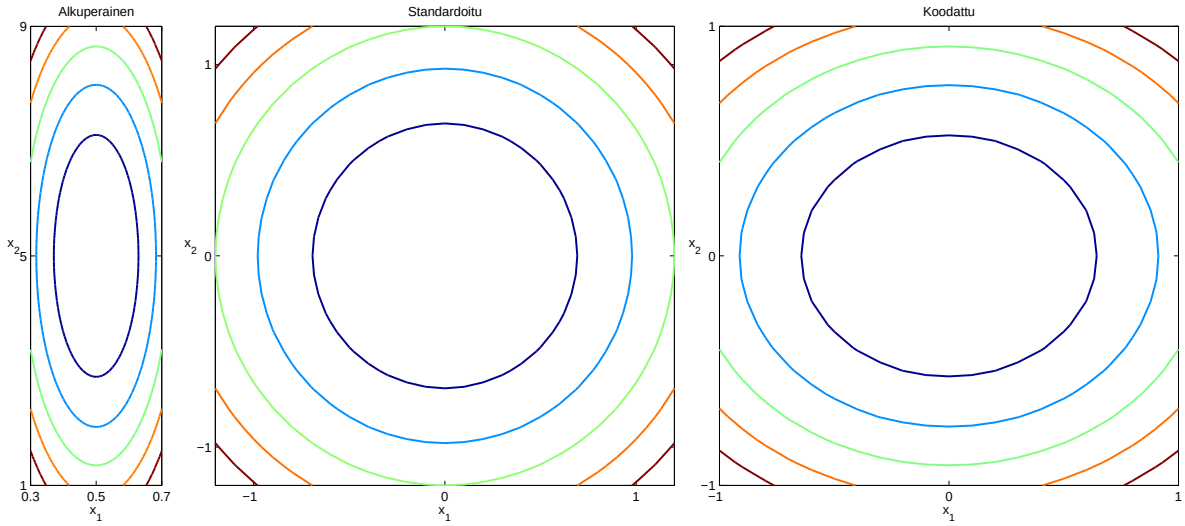
Alkuperäinen data ei siis ole kiertosymmetristä, muiden osalta on tarkistettava toinen ehto $\mathbf{D}^T\mathbf{D} = \lambda\mathbf{I}_k$:

$$\mathbf{D}_s^T\mathbf{D}_s = \begin{pmatrix} 11.00 & 0.00 \\ 0.00 & 11.00 \end{pmatrix},$$

$$\mathbf{D}_c^T\mathbf{D}_c = \begin{pmatrix} 12.00 & 0.00 \\ 0.00 & 8.00 \end{pmatrix}.$$

Vain standardoitu data on tässä tapauksessa kiertosymmetristä.

Tarkastellaan asiaa vielä kuvista, joissa on piirretty $V(\hat{y})$:n tasa-arvo käyrät:



Kiertosymmetrisessä standardoidussa datassa käyrät ovat origokeskeisiä ympyröitä, eli varianssi riippuu vain datavektorin $\boldsymbol{\xi}$ pituudesta. Alkuperäisellä ja koodatulla datalla suunnallakin on merkitystä.

Harjoitustehtävät

1. (a) Olkoot $\mathbf{x}$ ja $\mathbf{y}$ $k+1$-vektoreita ja $\mathbf{C}$ symmetrinen positiividefiniitti matriisi. Johda epäyhtälö $(\mathbf{x}^T\mathbf{C}\mathbf{x})(\mathbf{y}^T\mathbf{C}^{-1}\mathbf{y}) \geq |\mathbf{x}^T\mathbf{y}|^2$. (Vihje: matriisilla on symmetrinen positiividefiniitti neliöjuuri $\mathbf{C}^{1/2}$, ja Cauchyn epäyhtälö on $|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\| \cdot \|\mathbf{y}\|$.)

   (b) Todista tekstin epäyhtälö $c_{ii} \geq \|\mathbf{x}_i\|^{-2}$. (Vihje: edellisen kohdan erikoistapaus on $\mathbf{e}_i^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{e}_i \geq \|\mathbf{X}\mathbf{e}_i\|^{-2}$, missä $\mathbf{e}_i$ on identiteettimatriisin $\mathbf{I}_{k+1}$ i:s sarake.)

   (c) Todista $Nh_{ii} \geq 1$, missä h_{ii} on matriisin $\mathbf{H} := \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ i:s lävistäjäalkio.

2. Paljon käytetty kahden faktorin koetyyppi on ns. *monikulmiokoe*. Vastaavia ns. *monitahokaskokeita* on myös suuremmille faktorimäärille. Valitettavasti vain säännöllisiä monitahokkaita on äärellinen määrä — $\mathbb{R}^3$:ssa viisi, $\mathbb{R}^4$:ssä kuusi ja muissa vain kolme — joten idea on yleiskäyttöinen vain dimensiossa 2. *equiradial design*

   Monikulmiokokeen datamatriisi on $N \times 3$-matriisi, missä 2. ja 3. sarake muodostuvat sellaisen origokeskisen säännöllisen R-säteisen N-kulmion ($N \geq 3$) kärkien koordinaateista $\mathbb{R}^2$:ssa, jonka yksi kärki on positiivisel-

la x_1-akselilla:

$$x_{\ell 1} = R\cos\frac{2\pi\ell}{N} \quad \text{ja} \quad x_{\ell 2} = R\sin\frac{2\pi\ell}{N} \quad (\ell = 0, \ldots, N-1),$$

kompleksitasossa

$$x_{\ell 1} + ix_{\ell 2} = Re^{\frac{\ell}{N}2\pi i} \quad (\ell = 0, \ldots, N-1).$$

(i on imaginääriyksikkö.) Tietysti data viedään jälleen käytännön mittauksia varten sopivalla käänteisellä skaalauksella $\mathbf{L}^{-1}$ ”reaalimaailmaan”. Näytä, että monikulmiokoe on kiertosymmetrinen. (*Vihje:* Laske summat

$$\sum_{\ell=0}^{N-1} Re^{\frac{\ell}{N}2\pi i} \quad \text{sekä} \quad \sum_{\ell=0}^{N-1} (Re^{\frac{\ell}{N}2\pi i})^2$$

ja tarkastele reaali- ja imaginääriosia.)

Usein monikulmiokokeeseen lisätään keskustoistoja origossa epäsopivuustestiä varten. Ne parantavat tilannetta myös muuten, jos mukana on toisen kertaluvun faktoreita, kuten usein on — toistoja pitää olla silloin nimenomaan N kappaletta. Keskustoistot eivät hävitä kiertosymmetrisyyttä.

2.3 Simplex-koe

saturated design

Suunnittelua, missä kokeiden lukumäärä N on sama kuin lineaarisen mallin parametrien lukumäärä $k+1$, kutsutaan *kyllästetyksi*. Silloin kyseessä on interpolaatio, ja mallin sopivuus on täydellinen.

Ensimmäsen kertaluvun mallin kiertosymmetrinen kyllästetty suunnittelu, joka toteuttaa ehdon

$$\mathbf{X}^T\mathbf{X} = (k+1)\mathbf{I}_{k+1},$$

kutsutaan *simplex-kokeeksi*. Se voidaan muodostaa seuraavasti:

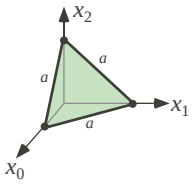
1. Valitse ei-singulaarinen $(k+1) \times (k+1)$ matriisi $\mathbf{W}$, jonka ensimmäinen sarake on $\mathbf{1}_{k+1}$.

2. Laske sen QR-hajotelma $\mathbf{W} = \mathbf{QR}$, missä $\mathbf{Q}$ on ortogonaalimatriisi ja $\mathbf{R}$ on yläkolmiomatriisi.

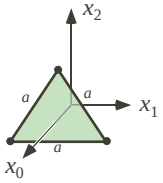
3. Valitse $\mathbf{X} = \pm\sqrt{k+1}\mathbf{Q}$, missä merkki valitaan niin, että datamatriisin ensimmäinen sarake on $\mathbf{1}_{k+1}$.

Simplex-kokeen suunnittelumatriisi muodostuu origokeskisen $k+1$-kärkisen monitahokkaan eli simpleksin koordinaateista $\mathbb{R}^k$:ssa. Esimerkiksi $\mathbb{R}^2$:ssa tällainen simpleksi on tasasivuinen origokeskinen kolmio. Sama tasasivuinen kolmio syntyy $\mathbb{R}^3$:een leikattaessa ensimmäistä oktanttia tasolla

$$x + y + z = \frac{a}{\sqrt{2}}.$$



Rotaatiolla saadaan kolmio x_1x_2-tason suuntaiseksi, jolloin sen kärkien ensimmäiset koordinaatit ovat samat:



Kolmion kärjet origoon yhdistävät janat

$$\left(\frac{a}{\sqrt{6}}, 0, \frac{a}{\sqrt{3}}\right), \left(\frac{a}{\sqrt{6}}, \frac{a}{2}, -\frac{a}{\sqrt{12}}\right), \left(\frac{a}{\sqrt{6}}, -\frac{a}{2}, -\frac{a}{\sqrt{12}}\right)$$

ovat edelleen kohtisuorassa toisiaan vastaan (ortogonaalisuus). Asettamalla $a = \sqrt{6}$ saadaan datamatriisi

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 2/\sqrt{2} \\ 1 & \sqrt{3/2} & -1/\sqrt{2} \\ 1 & -\sqrt{3/2} & -1/\sqrt{2} \end{bmatrix},$$

joka toteuttaa $\mathbf{X}^T\mathbf{X} = 3\mathbf{I}_3$.

Käytännössä simplex-data muunnetaan sopivalle asteikolle skaalauksella. Koe suoritetaan skaalatulla datalla, mutta mallina käytetään simplex-datan mallia, josta haluttaessa voidaan päästä skaalauksella ”reaalimaailmaan”. Suunnittelua voidaan haluttaessa typistää, ts. ottaa mukaan pienempi määrä faktoreita. Kuten edellä todettiin, tämä ei poista ortogonaalisuutta eikä kiertosymmetrisyyttä. Typistetty simplex-koe ei ole kyllästetty, joten voidaan tehdä t-testejä ja ANOVAa.

Erikoistapaus simplex-kokeesta on ns. *Plackettin ja Burmanin koe.* Datamatriisi on tällöin (mahdollisen koodauksen jälkeen) alkioista ± 1 koostuva $(k+1) \times (k+1)$-matriisi $\mathbf{X}$, joka toteuttaa ehdon

$$\mathbf{X}^T\mathbf{X} = (k+1)\mathbf{I}_{k+1}.$$

Tällaista ± 1-matriisia $\mathbf{X}$ kutsutaan *Hadamardin matriisiksi*. Hadamardin matriisi on *standardimuodossa,* jos sen ensimmäinen sarake on $\mathbf{1}$ ja ensimmäinen rivi $\mathbf{1}^T$. (Jokainen Hadamardin matriisi voidaan saattaa tällaiseksi kertomalla sen rivejä ja sarakkeita sopivasti -1:llä. Tämä säilyttää Hadamardin ominaisuuden, kuten voi todeta.) Hadamardin $m \times m$-matriisilla $\mathbf{H}$ on seuraavat ominaisuudet:

(i) Standardimuotoisen Hadamardin matriisin sarakesummat ensimmäistä saraketta lukuunottamatta ovat $= 0$, ts. sarakkeissa on yhtä monta $+1$:tä ja -1:tä.

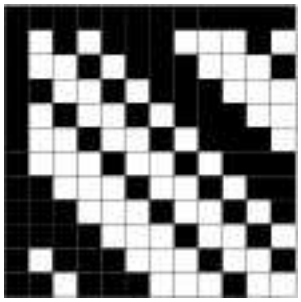
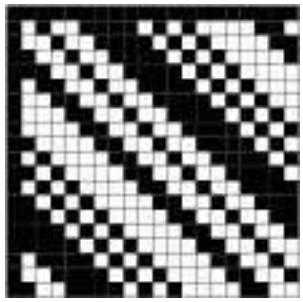
(ii) Joko $m = 2$ tai sitten m on neljällä jaollinen luku.

(iii) $\mathbf{H}$:n kahden rivin välinen etäisyys on aina $\sqrt{2m}$. Tästä ja kohdasta (i) seuraa, että Plackettin ja Burmanin koe on simplex-koe, koska rivin ensimmäinen alkio on 1.

Nämä ominaisuudet ovat kutakuinkin helposti todettavissa (harjoitustehtävä 1). 2×2 Hadamardin matriisi on

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

12×12 ja 20×20 Hadamardin matriisit ovat


ja

,

missä musta ruutu tarkoittaa 1 ja valkoinen ruutu -1.

Jo saaduista Hadamardin matriiseista saa uusia isompia Hadamardin matriiseja ns. *Kroneckerin tuloa* käyttämällä. Yleisesti $n_1 \times m_1$-matriisin

$$\mathbf{A} = \begin{pmatrix} a_{11} & \cdots & a_{1m_1} \\ \vdots & \ddots & \vdots \\ a_{n_1} & \cdots & a_{n_1 m_1} \end{pmatrix}$$

ja $n_2 \times m_2$-matriisin $\mathbf{B}$ Kroneckerin tulo on $n_1 n_2 \times m_1 m_2$-matriisi

$$\mathbf{A} = \left(\begin{array}{c|c|c} a_{11}\mathbf{B} & \cdots & a_{1m_1}\mathbf{B} \\ \hline \vdots & \ddots & \vdots \\ \hline a_{n_1}\mathbf{B} & \cdots & a_{n_1 m_1}\mathbf{B} \end{array}\right) =: \mathbf{A} \otimes \mathbf{B}.$$

Lohkomatriisien kertolaskukaavasta seuraa melko välittömästi, että mikäli matriisitulot $\mathbf{AC}$ ja $\mathbf{BD}$ ovat määritellyt, niin

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}),$$

ja lohkomatriisin transponointikaavasta puolestaan, että $(\mathbf{A} \otimes \mathbf{B})^T = \mathbf{A}^T \otimes \mathbf{B}^T$. Jos nyt $m_1 \times m_1$-matriisi $\mathbf{H}_1$ ja $m_2 \times m_2$-matriisi $\mathbf{H}_2$ ovat Hadamardin matriiseja, niin samoin on niiden Kroneckerin tulo $\mathbf{H}_1 \otimes \mathbf{H}_2$, sillä

$$\begin{aligned}(\mathbf{H}_1 \otimes \mathbf{H}_2)^T(\mathbf{H}_1 \otimes \mathbf{H}_2) &= (\mathbf{H}_1^T \otimes \mathbf{H}_2^T)(\mathbf{H}_1 \otimes \mathbf{H}_2) = (\mathbf{H}_1^T\mathbf{H}_1) \otimes (\mathbf{H}_2^T\mathbf{H}_2) \\ &= (m_1\mathbf{I}_{m_1}) \otimes (m_2\mathbf{I}_{m_2}) = m_1m_2\mathbf{I}_{m_1m_2}\end{aligned}$$

ja $\mathbf{H}_1 \otimes \mathbf{H}_2$:n ensimmäinen sarake on $\mathbf{1}_{m_1m_2}$.

Esimerkki

Etsi simplex-koe, kun $k = 3$, ja typistetty Plackettin ja Burmanin koe, kun $k = 9$ ja $N = 24$.

Ratkaisu

Muodostetaan simplex-koe kohdan 2.2. kolmivaiheisella menettelyllä.

1. Valitaan 4×4 matriisiksi $\mathbf{W}$:

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix},$$

   jossa kolme viimeistä saraketta on valittu lineaarisesti riippumattomiksi, jotta matriisi olisi ei-singulaarinen.

2. Muodostetaan QR-hajotelma:

$$\mathbf{W} = \mathbf{QR},$$

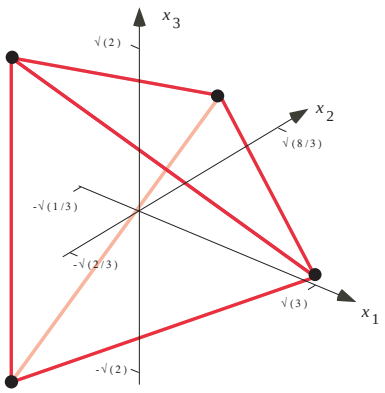
   kun

$$\begin{aligned}\mathbf{Q} &= \begin{pmatrix} -1/2 & \sqrt{3}/6 & \sqrt{6}/6 & -\sqrt{2}/2 \\ -1/2 & -\sqrt{3}/2 & 0 & 0 \\ -1/2 & \sqrt{3}/6 & -\sqrt{6}/3 & 0 \\ -1/2 & \sqrt{3}/6 & \sqrt{6}/6 & \sqrt{2}/2 \end{pmatrix}, \\ \mathbf{R} &= \begin{pmatrix} -2 & -1/2 & -1/2 & -1/2 \\ 0 & -\sqrt{3}/2 & \sqrt{3}/6 & \sqrt{3}/6 \\ 0 & 0 & -\sqrt{6}/3 & \sqrt{6}/6 \\ 0 & 0 & 0 & \sqrt{2}/2 \end{pmatrix}.\end{aligned}$$

3. Valitaan:

$$\mathbf{X} = -\sqrt{k+1}\mathbf{Q} = -2\mathbf{Q} = \begin{pmatrix} 1 & -1/\sqrt{3} & -2/\sqrt{3} & \sqrt{2} \\ 1 & \sqrt{3} & 0 & 0 \\ 1 & -1/\sqrt{3} & \sqrt{8/3} & 0 \\ 1 & -1/\sqrt{3} & -2/\sqrt{3} & -\sqrt{2} \end{pmatrix}.$$

Datapisteet ovat tetraedrin kärjet:



Plackettin ja Burmanin koetta varten lasketaan 24×24 Hadamardin matriisi lähtemällä tekstissä annetuista 2×2 ja 12×12 Hadamardin matriiseista.

$$\mathbf{H}_{24} = \mathbf{H}_2 \otimes \mathbf{H}_{12} = \left(\begin{array}{c|c} \mathbf{H}_{12} & \mathbf{H}_{12} \\ \hline \mathbf{H}_{12} & -\mathbf{H}_{12} \end{array}\right) = \begin{pmatrix} \mathbf{h}_1 & \mathbf{h}_2 & \cdots & \mathbf{h}_{24} \end{pmatrix}.$$

Typistetään tämä suunnittelu nyt 9:n faktorin suunnitteluksi. Voidaan valita siis matriisin $\mathbf{H}_{24}$ sarakkeista kymmenen ensimmäistä. Tällöin jokainen koe olisi toistettu. Valitaankin sen sijaan sarakeita lähtien $13.$ sarakkeesta, jolloin toistokokeita ei tule. Nyt jälkimmäinen koetusina on ensimmäinen merkki vaihdettuna, lukuun-

ottamatta ensimmäistä $\mathbf{1}_N$-saraketta:

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{24} & \mathbf{h}_{13} & \mathbf{h}_{14} & \cdots & \mathbf{h}_{21} \end{pmatrix}$$

$$= \left(\begin{array}{cccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 & -1 \\
1 & 1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 & 1 \\
1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & 1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & -1 \\
1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 \\
1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \\
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
\hline
1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 \\
1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 & -1 \\
1 & -1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 & -1 \\
1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 & 1 \\
1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 & 1 \\
1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\
1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 & 1 & 1
\end{array}\right).$$

Harjoitustehtävät

1. Todista Hadamardin matriisin ominaisuudet (i–iii).

2. Suunnittele typistetty Plackettin ja Burmanin koe, jossa on 10 faktoria ja 20 koetta.

3. Alla on 2×2-lohkottujen matriisien kertokaava. Tässä tietysti oletetaan, että esiintyvät matriisikertolaskut ovat sallittuja.

$$\left(\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right) \left(\begin{array}{c|c} \mathbf{B}_{11} & \mathbf{B}_{12} \\ \hline \mathbf{B}_{21} & \mathbf{B}_{22} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{A}_{11}\mathbf{B}_{11} + \mathbf{A}_{12}\mathbf{B}_{21} & \mathbf{A}_{11}\mathbf{B}_{12} + \mathbf{A}_{12}\mathbf{B}_{22} \\ \hline \mathbf{A}_{21}\mathbf{B}_{11} + \mathbf{A}_{22}\mathbf{B}_{21} & \mathbf{A}_{21}\mathbf{B}_{12} + \mathbf{A}_{22}\mathbf{B}_{22} \end{array}\right)$$

   Miten tästä saadaan Kroneckerin tulon $\otimes$ osittelukaava

$$(\mathbf{A} \otimes \mathbf{B})(\mathbf{C} \otimes \mathbf{D}) = (\mathbf{AC}) \otimes (\mathbf{BD}),$$

kun $\mathbf{A}$ ja $\mathbf{C}$ ovat 2×2-matriiseja?

2.4 Kahden tason kokeet

two-level factorial design

Kahden tason kokeella tarkoitetaan koetta, jossa $(k+1) \times N$-datamatriisin $\mathbf{X}$ sarakkeissa (ensimmäistä saraketta lukuunottamatta) esiintyy vain kahta eri tason arvoa. Koodauksen jälkeen ne ovat 1 ja -1. Jatkossa oletetaankin koodaus valmiiksi suoritetuksi. Plackettin ja Burmanin kokeet ovat siis kahden tason kokeita.

Kahden tason kokeen malli on

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{1 \le i < j \le k} \beta_{ij} x_i x_j + \cdots + \beta_{1\ldots k} x_1 \cdots x_k$$

tai tästä joitakin termejä pois jättämällä saatu malli. Jotta datamatriisin $\mathbf{X}$ sarakerangi säilyy täydellisenä, faktorien kvadraattisia ja korkeampia potensseja ei oteta mukaan malliin, sillä

$$x_i \in \{1, -1\} \quad \Rightarrow \quad x_i^{2n+1} = x_i \quad \text{ja} \quad x_i^{2n} = 1.$$

main effects
interaction effects

Ensimmäisen kertaluvun faktoreita kutsutaan *päävaikutuksiksi* ja niiden tuloja *yhdysvaikutuksiksi*.

2^k factorial design

Jos mallissa on kaikki mahdolliset termit mukana, kyseessä on *täydellinen* 2^k*-koe.* Täydellisessä 2^k-kokeessa on mukana

$$1 + \binom{k}{1} + \binom{k}{2} + \cdots + \binom{k}{k} = (1+1)^k = 2^k$$

termiä. Mahdollisia erilaisia datamatriisin rivejä on toisaalta myös 2^k kappaletta. Jos toistettuja rivejä ei ole mukana, voidaan rivit järjestää siten, että 2. sarakkeessa on ensin 2^{k-1} kappaletta -1:stä ja sitten 2^{k-1} kappaletta 1:stä, 3. sarakkeessa on ensin 2^{k-2} kappaletta -1:stä, sitten 2^{k-2} kappaletta 1:stä, sitten 2^{k-2} kappaletta -1:stä ja lopuksi 2^{k-2} kappaletta 1:stä, jne., $k+1$:nnessä sarakkeessa -1:t ja 1:t vuorottelevat. Esimerkiksi ensimmäisen kertaluvun 2^3-kokeen tällä tavoin esitetty datamatriisi on

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Ensimmäisen kertaluvun täydellinen 2^k-koe on näin ollen aina kiertosymmetrinen, sillä ilmeisesti $\mathbf{X}^T\mathbf{X} = 2^k\mathbf{I}_{k+1}$ (Lause 1.4).

Täydellisessä 2^k-kokeessa on useinkin käytännön kannalta liian monta koetta. Ns. *osittaisissa 2^k-kokeissa* faktorien määrää karsitaan (ja datamatriisin rivilukua pienennetään) aivan omalla tavallaan kieltämällä tietyt yhdysvaikutukset. Yhdysvaikutuksen *kielto* tarkoittaa sitä, että sen arvo kiinnitetään ± 1:ksi. Kiellettäessä tiettyjä yhdysvaikutuksia päätetään samalla, etteivät ne ole tarkastelun kannalta tärkeitä. Kiellettyjen vaikutuksien sanotaan *sekoittuvan* vakiotermiin.

2^k fractional factorial design

sacrifice

confounding

Jos kielletään vaikutukset $z_1, \ldots, z_m$, on kiellettävä myös kaikki näistä keskenään kertomalla saadut vaikutukset, sillä näiden arvot tulevat myös kiinnitetyksi. Jos siis tapauksessa $k = 5$ päätetään kieltää

$$x_1x_2x_4 \quad \text{ja} \quad x_1x_3x_5$$

on myös kiellettävä

$$x_1x_2x_4 \cdot x_1x_3x_5 = x_2x_3x_4x_5.$$

Alinta kertalukua olevan kielletyn termin aste on ns. *kokeen resoluutio*. Resoluutio-III kokeissa päävaikutukset sekoittuvat kahden faktorin yhdysvaikutuksiin mutta eivät sekoitu keskenään. Resoluutio-IV kokeissa päävaikutukset eivät sekoitu keskenään eikä kahden faktorin yhdysvaikutuksiin, mutta kahden faktorin yhdysvaikutukset sekoittuvat keskenään. Resoluutio-V kokeissa päävaikutukset ja kahden faktorin yhdysvaikutukset eivät sekoitu keskenään.

Kun vaikutukset $z_1, \ldots, z_m$ on kielletty, ts. niiden arvot kiinnitetty, jätetään datamatriisiin vain ne rivit, jotka toteuttavat nämä kiinnitykset. Itse malliin ei oteta mukaan kiellettyjen vaikutuksien termejä. Toisaalta kiinnitykset samaistavat tiettyjä vaikutuksia merkkiä vaille ja näistä otetaan mukaan malliin vain yksi, jottei datamatriisiin tule lineaarisesti riippuvia sarakkeita. Tällaisia vaikutuksia kutsutaan toistensa *aliaksiksi*. Esimerkiksi yo. kiinnitysten puitteissa malliin ei saa ottaa mukaan molempia termejä $\beta_{34}x_3x_4$ ja $\beta_{25}x_2x_5$, sillä

$$x_3x_4 = (\pm x_2x_3x_4x_5)x_3x_4 = \pm x_2x_5,$$

missä merkki $\pm$ valitaan siten, että $\pm x_2x_3x_4x_5 = 1$. Kaikki ko. kiinnityksen aliakset ovat

$$\begin{array}{ccccccc}
1 & = & |x_1x_2x_4| & = & |x_1x_3x_5| & = & |x_2x_3x_4x_5| \\
|x_1| & = & |x_2x_4| & = & |x_3x_5| & = & |x_1x_2x_3x_4x_5| \\
|x_2| & = & |x_1x_4| & = & |x_1x_2x_3x_5| & = & |x_3x_4x_5| \\
|x_3| & = & |x_1x_2x_3x_4| & = & |x_1x_5| & = & |x_2x_4x_5| \\
|x_4| & = & |x_1x_2| & = & |x_1x_3x_4x_5| & = & |x_2x_3x_5| \\
|x_5| & = & |x_1x_2x_4x_5| & = & |x_1x_3| & = & |x_2x_3x_4| \\
|x_2x_3| & = & |x_1x_3x_4| & = & |x_1x_2x_5| & = & |x_4x_5| \\
|x_3x_4| & = & |x_1x_2x_3| & = & |x_1x_4x_5| & = & |x_2x_5|
\end{array}$$

Vaikutuksien $x_1x_2x_4$ ja $x_1x_3x_5$ kielto antaa ensimmäisen kertaluvun mallin osittaisen 2^5-koesuunnittelun, jossa on $N = 2^{5-2} = 8$ koetta. Suunnittelu on kiertosymmetrinen, sillä se on täydellisen 2^k-kokeen typistys.

Seuraavasta taulukosta löytyy hyödyllisiä osittaisia 2^k-kokeita.

nimi	k	N	kiellot
2^{3-1}_{III}	3	4	$x_1x_2x_3$
2^{4-1}_{IV}	4	8	$x_1x_2x_3x_4$
2^{5-1}_{V}	5	16	$x_1x_2x_3x_4x_5$
2^{5-2}_{III}	5	8	$x_1x_2x_4$, $x_1x_3x_5$
2^{6-1}_{VI}	6	32	$x_1x_2x_3x_4x_5x_6$
2^{6-2}_{IV}	6	16	$x_1x_2x_3x_5$, $x_2x_3x_4x_6$
2^{6-3}_{III}	6	8	$x_1x_2x_4$, $x_1x_3x_5$, $x_2x_3x_6$
2^{7-2}_{IV}	7	32	$x_1x_2x_3x_6$, $x_1x_2x_4x_5x_7$
2^{7-3}_{IV}	7	16	$x_1x_2x_3x_4$, $x_2x_3x_4x_6$, $x_1x_3x_4x_7$
2^{7-4}_{III}	7	8	$x_1x_2x_4$, $x_1x_3x_5$, $x_2x_3x_6$, $x_1x_2x_3x_7$
2^{8-3}_{IV}	8	32	$x_1x_2x_3x_6$, $x_1x_2x_4x_7$, $x_2x_3x_4x_5x_8$
2^{8-4}_{IV}	8	16	$x_2x_3x_4x_5$, $x_1x_3x_4x_6$, $x_1x_2x_3x_7$, $x_1x_2x_4x_8$

Kyllästetyt osittaiset 2^k-kokeet, kuten 2^{3-1}_{III} ja 2^{7-4}_{III}, ovat samalla Plackettin ja Burmanin kokeita.

Esimerkki

principal fraction

Sitä kiellettyjen faktoreiden arvojen kiinnitystä, joka antaa kullekin niistä arvon 1, kutsutaan *pääositukseksi*. Etsi tekstissä olevan 2^{5-2}_{III}-kokeen pääosituksen ensimmäisen kertaluvun mallin datamatriisi.

Ratkaisu

Taulukosta nähdään että 2^{5-2}_{III} mallissa tulisi kieltää esimerkiksi faktorit $x_1x_2x_4$ ja $x_1x_3x_5$. Huomataan, että $x_1x_2x_4 = 1 \Leftrightarrow x_1x_2 = x_4$ ja $x_1x_3x_5 = 1 \Leftrightarrow x_1x_3 = x_5$. Toisaalta aliastaulustakin nähdään, että $x_4 = x_1x_2$ ja $x_5 = x_1x_3$ (merkit valittu pääosituksen mukaiseksi).

Datamatriisi saadaan, kun annetaan faktoreille x_1, x_2 ja x_3 kaikki mahdolliset

arvoyhdelmät, ja asetetaan $x_4 = x_1x_2$ ja $x_5 = x_1x_3$:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Huomataan, että jos faktorit x_4 ja x_5 osoittautuvat tarpeettomiksi, niin samoja koetuloksia voidaan käyttää sovittamaan ensimmäisen kertaluvun malli, missä on vain faktorit x_1, x_2 ja x_3. Kannattaa siis numeroida faktorit niin, että x_4 ja x_5 ovat ne faktorit, joiden uskotaan olevan vähemmän tärkeitä.

Harjoitustehtävät

1. Etsi 2^{3-1}_{III}-kokeen datamatriisi ja aliastaulu. Näytä, että koe on ekvivalentti erään Plackettin ja Burmanin kokeen kanssa.

2. Etsi resoluutio-III osittainen 2^k-koesuunnittelu, jolla on mahdollisimman pieni kokeiden lukumäärä N, kun $k = 9$. Löytyykö kiertosymmetrinen kahden tason koe, jolla on pienempi N?

3. 2^{4-1}_{IV}-koetta käytetään sovittamaan ensimmäisen kertaluvun malli. Jos todellinen vastefunktio on

$$y = \beta_0 + \sum_{i=1}^{4} \beta_i x_i + \beta_{12} x_1 x_2,$$

niin ovatko parametrien β_0, β_1, β_2, β_3 ja β_4 estimaatit harhattommia, eli *unbiased* toteutuuko $E(\beta_i) = \beta_i$ $(0 \leq i \leq 4)$?

2.5 Toisen kertaluvun regressiomalli

Toisen kertaluvun malli tarvitaan erikoisesti silloin, kun faktorien tarkastelualueella on maksimia tai minimiä. Täydellinen toisen kertaluvun malli on muotoa

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{1 \leq i < j \leq k} \beta_{ij} x_i x_j + \epsilon.$$

Termejä on $1 + 2k + \binom{k}{2}$ kappaletta. Tarkastelemme tässä kohdassa toisen kertaluvun mallin koesuunnittelun ortogonaalisuutta ja kiertosymmetrisyyttä.

Matriisia $\frac{1}{N}\mathbf{X}^T\mathbf{X}$ kutsutaan *momenttimatriisiksi*. Momentit määritellään

$$\begin{aligned}
[i] &:= \frac{1}{N}\sum_{t=1}^{N} x_{ti} \\
[ij] &:= \frac{1}{N}\sum_{t=1}^{N} x_{ti}x_{tj} = [ji] \\
[ijm] &:= \frac{1}{N}\sum_{t=1}^{N} x_{ti}x_{tj}x_{tm} = [jim] = \cdots \\
[ijmn] &:= \frac{1}{N}\sum_{t=1}^{N} x_{ti}x_{tj}x_{tm}x_{tn} = [imjn] = \cdots
\end{aligned}$$

Esimerkiksi tapauksessa $k = 2$ momenttimatriisi on

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c|cc|cc|c}
1 & [1] & [2] & [11] & [22] & [12] \\
\hline
[1] & [11] & [12] & [111] & [122] & [112] \\
[2] & [12] & [22] & [112] & [222] & [122] \\
\hline
[11] & [111] & [122] & [1111] & [1122] & [1112] \\
[22] & [122] & [112] & [1122] & [2222] & [1222] \\
\hline
[12] & [112] & [122] & [1112] & [1222] & [1122]
\end{array}\right)$$

Momenttimatriisin yleinen lohkorakenne on siis

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c|c|c|c}
1 & [m] & [mm] & [mn]\ (m < n) \\
\hline
[i] & [im] & [imm] & [imn]\,(m < n) \\
\hline
[ii] & [iim] & [iimm] & [iimn]\ (m < n) \\
\hline
[ij]\ (i < j) & [ijm]\,(i < j) & [ijmm]\ (i < j) & [ijmn]\ (i < j, m < n)
\end{array}\right)$$

Koesuunnittelu on ortogonaalinen jos momenttimatriisi on lävistäjämatriisi. Oletetaan ensin, että faktorit $x_1, \ldots, x_k$ ovat skaalattuja niin, että $\bar{x}_i = [i] = 0$ ja $[ii] = 1$. Tämä skaalaus on samanlainen kuin standardointi, paitsi että käytetty hajonta on

$$\sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_{ji} - \overline{x}_i)^2}.$$

Seuraavaksi korvataan affiinimuunnoksella $\mathbf{L}$ mallin neliöfaktorit uusilla muotoa

$$x_i^2 + p_i x_i + q_i$$

olevilla faktoreilla. Uuden mallin momenttimatriisin rakenne on

$$\frac{1}{N}(\mathbf{LX})^T\mathbf{LX} =$$

$$\left(\begin{array}{c|c|c|c} 1 & 0 & 1+q_m & [mn]\ (m<n) \\ \hline 0 & [im] & [imm]+p_m[im] & [imn]\ (m<n) \\ \hline 1+q_i & [iim]+p_i[im] & \begin{array}{c}[iimm]+p_i[imm]+\\ p_m[iim]+p_ip_m[im]\\ +q_i+q_m+q_iq_m\end{array} & \begin{array}{c}[iimn]+p_i[imn]\\ +q_i[mn]\ (m<n)\end{array} \\ \hline [ij]\ (i<j) & [ijm]\ (i<j) & \begin{array}{c}[ijmm]+p_m[ijm]\\ +q_m[ij]\ (i<j)\end{array} & [ijmn]\ (i<j, m<n) \end{array}\right)$$

Jotta uutta mallia käyttävä koe olisi ortogonaalinen, on affiinimuunnoksen kertoimien oltava

$$p_i = -[iii] \quad \text{ja} \quad q_i = -1$$

ja momenttien oltava

1. $[ij] = 0$, kun $i \neq j$,
2. $[ijm] = 0$, kun ei ole kyseessä tapaus $i = j = m$,
3. $[iijj] = 1$, kun $i \neq j$ ja
4. $[ijmn] = 0$, kun ei ole kyseessä tapaus 3. eikä tapaus $i = j = m = n$.

Ts. ainoat nollasta eroavat momentit ovat muotoa $[ii]$, $[iii]$, $[iijj]$ tai $[iiii]$ (ja näistäkin $[iii]$ voi olla nolla). Ortogonalisoituvan kokeen momenttimatriisi on siis muotoa

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c|c|c|c} 1 & \mathbf{0}_k^T & \mathbf{1}_k^T & \mathbf{0}^T \\ \hline \mathbf{0}_k & \mathbf{I}_k & \mathbf{S}_3 & \mathbf{O}^T \\ \hline \mathbf{1}_k & \mathbf{S}_3 & \mathbf{S}_4 - \mathbf{I}_k + \mathbf{1}_k\mathbf{1}_k^T & \mathbf{O}^T \\ \hline \mathbf{0} & \mathbf{O} & \mathbf{O} & \mathbf{I} \end{array}\right),$$

missä $\mathbf{S}_3$ ja $\mathbf{S}_4$ ovat lävistäjämatriiseja, joiden i:s lävistäjäalkio on $[iii]$ ja $[iiii]$. Usein riittää pelkkä tieto siitä, että koe on periaatteessa ortogonalisoitavissa eo. tapaan. Ortogonaalisuuden hyvät puolet kun näkyvät ilman varsinaista ortogonalisointiakin!

Kiertosymmetriaa ei voida määritellä toisen kertaluvun malleja käyttäville kokeille samalla tavalla kuin edellä tehtiin ensimmäisen kertaluvun mallin tapauksessa. Toisaalta määritelmäksi voidaan ottaa sivulla 43 mainittu kiertosymmetrisen kokeen ennusteen varianssia koskeva ominaisuus: Toisen kertaluvun malliin perustuva koe on *kiertosymmetrinen*, jos sitä käyttäen saadun ennusteen varianssi riippuu vain datavektorin ensimmäisen kertaluvun osan pituudesta.

Lause 2.3. *Toisen kertaluvun kokeeseen perustuva malli on kiertosymmetrinen, jos momenttimatriisi on muotoa*

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \left(\begin{array}{c|c|c|c} 1 & \mathbf{0}_k^T & \lambda_1\mathbf{1}_k^T & \mathbf{0}^T \\ \hline \mathbf{0}_k & \lambda_1\mathbf{I}_k & \mathbf{O}_k & \mathbf{O}^T \\ \hline \lambda_1\mathbf{1}_k & \mathbf{O}_k & \lambda_2(2\mathbf{I}_k + \mathbf{1}_k\mathbf{1}_k^T) & \mathbf{O}^T \\ \hline \mathbf{0} & \mathbf{O} & \mathbf{O} & \lambda_2\mathbf{I} \end{array}\right).$$

Jotta $\mathbf{X}^T\mathbf{X}$ olisi ei-singulaarinen, on ilmeisesti oltava $\lambda_1 \neq 0$ ja $\lambda_2 \neq 0$, mutta myös $k\lambda_1^2 \neq (k+2)\lambda_2$, kuten todistuksesta huomataan. Jotta lauseessa annettu kiertosymmetrinen koe olisi vielä ortogonalisoitavissa, on edellä olevan mukaan ilmeisesti oltava $\lambda_2 = 1$; standardoinnista johtuen on tällöin myös $\lambda_1 = 1$.

Todistus. Käyttämällä lohkomatriisin kääntökaavaa (ks. sivun 15 alaviite) lohkoille $(1,1)$, $(1,3)$, $(3,1)$ ja $(3,3)$ todetaan, että käänteismatriisi on

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{N}\left(\begin{array}{c|c|c|c} 1 + \lambda_1^2\mathbf{1}_k^T\mathbf{Y}\mathbf{1}_k & \mathbf{0}_k^T & -\lambda_1\mathbf{1}_k^T\mathbf{Y} & \mathbf{0}^T \\ \hline \mathbf{0}_k & \lambda_1^{-1}\mathbf{I}_k & \mathbf{O}_k & \mathbf{O}^T \\ \hline -\lambda_1\mathbf{Y}\mathbf{1}_k & \mathbf{O}_k & \mathbf{Y} & \mathbf{O}^T \\ \hline \mathbf{0} & \mathbf{O} & \mathbf{O} & \lambda_2^{-1}\mathbf{I} \end{array}\right),$$

missä

$$\mathbf{Y} := \left(\lambda_2(2\mathbf{I}_k + \mathbf{1}_k\mathbf{1}_k^T) - \lambda_1^2\mathbf{1}_k\mathbf{1}_k^T\right)^{-1} = \frac{1}{2\lambda_2}\left(\mathbf{I}_k + \frac{\lambda_2 - \lambda_1^2}{2\lambda_2}\mathbf{1}_k\mathbf{1}_k^T\right)^{-1},$$

edellyttäen tietysti, että ko. käänteismatriisi on olemassa. Shermanin ja Morrisonin kääntökaavaa[1] käyttäen todetaankin, että

$$\mathbf{Y} = \frac{1}{2\lambda_2}\left(\mathbf{I}_k + \frac{\lambda_1^2 - \lambda_2}{(k+2)\lambda_2 - k\lambda_1^2}\mathbf{1}_k\mathbf{1}_k^T\right).$$

Lasketaan sitten ennusteen $\hat{y}$ varianssi. Datavektori $\boldsymbol{\xi}$ ositetaan seuraavasti:

$$\boldsymbol{\xi} = \left(\begin{array}{c} 1 \\ \hline \mathbf{d} \\ \hline \mathbf{z}_1 \\ \hline \mathbf{z}_2 \end{array}\right),$$

[1] Shermanin ja Morrisonin kääntökaava on

$$(\mathbf{A} + \mathbf{u}\mathbf{v}^T)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}(\mathbf{A}^{-1}\mathbf{v})^T}{1 + \mathbf{v}^T\mathbf{A}^{-1}\mathbf{u}}$$

ja se on voimassa, mikäli $\mathbf{v}^T\mathbf{A}^{-1}\mathbf{u} \neq -1$.

missä vektorissa $\mathbf{d}$ on ensimmäisen kertaluvun faktorit, vektorissa $\mathbf{z}_1$ ovat $\mathbf{d}$:n komponenttien neliöt ja vektorissa $\mathbf{z}_2$ niiden sekatulot. Tutkittava varianssi on $V(\hat{y}) = \sigma^2 \boldsymbol{\xi}^T (\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi}$. Lohkokertolasku osoittaa, että

$$\frac{N}{\sigma^2}V(\hat{y}) = 1 + \lambda_1^2 \mathbf{1}_k^T \mathbf{Y}\mathbf{1}_k - 2\lambda_1 \mathbf{z}_1^T \mathbf{Y}\mathbf{1}_k + \frac{1}{\lambda_1}\mathbf{d}^T\mathbf{d} + \mathbf{z}_1^T\mathbf{Y}\mathbf{z}_1 + \frac{1}{\lambda_2}\mathbf{z}_2^T\mathbf{z}_2.$$

Nyt $\mathbf{Y}\mathbf{1}_k = c\mathbf{1}_k$, missä

$$c = \frac{1}{2\lambda_2}\left(1 + \frac{\lambda_1^2 - \lambda_2}{(k+2)\lambda_2 - k\lambda_1^2}k\right),$$

joten $\mathbf{1}_k^T\mathbf{Y}\mathbf{1}_k = ck$ ja $\mathbf{z}_1^T\mathbf{Y}\mathbf{1}_k = c\|\mathbf{d}\|^2$. Toisaalta

$$\mathbf{z}_2^T\mathbf{z}_2 = \sum_{1\le i<j\le k}\xi_i^2\xi_j^2 = \frac{1}{2}\mathbf{z}_1^T(\mathbf{1}_k\mathbf{1}_k^T - \mathbf{I}_k)\mathbf{z}_1,$$

joten

$$\mathbf{z}_1^T\mathbf{Y}\mathbf{z}_1 + \frac{1}{\lambda_2}\mathbf{z}_2^T\mathbf{z}_2 = \mathbf{z}_1^T\left(\mathbf{Y} + \frac{1}{2\lambda_2}(\mathbf{1}_k\mathbf{1}_k^T - \mathbf{I}_k)\right)\mathbf{z}_1 = d\mathbf{z}_1^T\mathbf{1}_k\mathbf{1}_k^T\mathbf{z}_1 = d\|\mathbf{d}\|^4.$$

missä

$$d = \frac{1}{2\lambda_2}\left(1 + \frac{\lambda_1^2 - \lambda_2}{(k+2)\lambda_2 - k\lambda_1^2}\right).$$

Kaiken kaikkiaan siis $V(\hat{y})$ riippuu vain $\|\mathbf{d}\|$:sta. □

Lauseessa annettu karakterisaatio on itse asiassa täydellinen, ts. muunlaisia kiertosymmetrisiä toisen kertaluvun malliin perustuvia kokeita ei ole. (Asian todistus ei ole aivan helppo.) Samantapainen karakterisaatio voidaan antaa minkä tahansa kertaluvun mallia käyttävän kokeen kiertosymmetrisyydelle. Ks. KHURI & CORNELL ja JOHN. Alkuperäisviite on BOX, G.E.P. & HUNTER, J.S.: Multifactor Experimental Designs for Exploring Response Surfaces, *Annals of Mathematical Statistics* **28** (1957), 195–242.

Esimerkki

Kuusikulmion koepisteet ovat

$$x_{j1} = R\cos(2\pi j/6),\ x_{j2} = R\sin(2\pi j/6), \quad (1 \le j \le 6)$$

johon lisätään n_0 koetoistoa origossa. Näytä, että $2.$ kertaluvun malli on kiertosymmetrinen ja ortogonalisoituva sopivalla n_0:n arvolla.

Ratkaisu

Kokeiden lukumäärä on $N = 6 + n_0$. Koska

$$\frac{1}{N}\sum_{j=1}^{N} x_{ji} = 0 \quad \text{ja} \quad \frac{1}{N}\sum_{j=1}^{N} x_{ji}^2 = \frac{3R^2}{N}$$

niin skaalaukseen riittää, että asetetaan $R = \sqrt{N/3}$. Momenttimatriisi on silloin

$$\frac{1}{N}\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & N/4 & N/12 & 0 \\ 1 & 0 & 0 & N/12 & N/4 & 0 \\ 0 & 0 & 0 & 0 & 0 & N/12 \end{pmatrix},$$

joka on kiertosymmetrisen kokeen momenttimatriisi, jossa $\lambda_1 = 1$ ja $\lambda_2 = N/12$. Koe on ortogonalisoituva, kun $N = 12$, eli kun toistokokeiden lukumäärä on $n_0 = 6$.

Harjoitustehtävät

1. Alla on erään muovikemian kokeen data. Datamatriisi on ns. *Box–Behnken-matriisi*. Faktorit (koodattuina) ovat lämpötila, seoksen sekoitusnopeus ja erään aineksen lisäysnopeus. Vaste on hartsin viskositeetti.

x_1	x_2	x_3	y
-1	-1	0	53
1	-1	0	58
-1	1	0	59
1	1	0	56
-1	0	-1	64
1	0	-1	45
-1	0	1	35
1	0	1	60
0	-1	-1	59
0	1	-1	64
0	-1	1	53
0	1	1	65
0	0	0	65
0	0	0	59
0	0	0	62

Onko $2.$ kertaluvun malli ortogonalisoituva? kiertosymmetrinen? Sovita $2.$ kertaluvun malli.

2. Kolmen tason kokeessa faktoreilla on kullakin kolme tasoyhdelmää datamatriisissa, koodauksen jälkeen -1, 0 ja 1. *Täydellisessä 3^k-kokeessa* ovat mukana kaikki 3^k eri tasoyhdelmää, kukin kerran. Todista, että nämä kokeet ovat toisen kertaluvun mallille ortogonalisoituvia.

2.6 CCD-kokeet

Central Composite Design

Suosittu toisen kertaluvun mallin koesuunnittelu on ns. *CCD-koe*. Sen datamatriisi muodostetaan kolmesta osasta:

1. *Faktoriaaliosa* koostuu kahden tason kokeesta, jonka f faktoritasoa koodataan ± 1:ksi. Ensimmäisen kertaluvun osuus datamatriisin faktoriaaliosasta on muotoa $\left(\begin{array}{c|c} \mathbf{1}_f & \mathbf{F} \end{array}\right)$, missä $\mathbf{F}$ on $f \times k$-matriisi. Faktoriaaliosa voi olla osittainen 2^k-koe, mutta resoluution on oltava vähintään V, jotta ensimmäisen ja toisen kertaluvun vaikutukset eivät sekoitu keskenään.

2. *Aksiaaliosa* saadaan pisteistä, jotka ovat $\mathbb{R}^k$:n akseleilla etäisyydellä α origosta. Ensimmäisen kertaluvun osuus datamatriisin aksiaaliosasta on muotoa
$$\left(\begin{array}{c|c} \mathbf{1}_k & \alpha\mathbf{I}_k \\ \hline \mathbf{1}_k & -\alpha\mathbf{I}_k \end{array}\right).$$

3. *Keskusosa* koostuu n_0 koetoistosta origossa. Ensimmäisen kertaluvun osuus datamatriisin keskusosasta on muotoa $\left(\begin{array}{c|c} \mathbf{1}_{n_0} & \mathbf{O} \end{array}\right)$. Keskusosan koetoistoja käyttäen voidaan testata mallin epäsopivuutta.

Ilmeisesti $N = f + 2k + n_0$ ja ensimmäisen kertaluvun osuus datamatriisista on
$$\left(\begin{array}{c|c} \mathbf{1}_f & \mathbf{F} \\ \hline \mathbf{1}_k & \alpha\mathbf{I}_k \\ \hline \mathbf{1}_k & -\alpha\mathbf{I}_k \\ \hline \mathbf{1}_{n_0} & \mathbf{O} \end{array}\right).$$

Faktoriaaliosa valitaan siten, että se on keskitetty, ts.
$$\mathbf{F}^T\mathbf{1}_f = \mathbf{0}_k.$$

Täydellisen tai osittaisen 2^k-kokeen muodostavalle faktoriaaliosalle tämä toteutuu automaattisesti. Tällöin koko ensimmäisen kertaluvun data on keskitetty. Skaalattaessa sarakkeiden yhteinen toisen asteen momentti on
$$S^2 := [ii] = \frac{1}{N}\sum_{t=1}^{N} x_{ti}^2 = \frac{1}{N}(f + 2\alpha^2).$$

Skaalattu datamatriisi on näin

$$\mathbf{X} = \left(\begin{array}{c|c|c|c} \mathbf{1}_f & \frac{1}{S}\mathbf{F} & \frac{1}{S^2}\mathbf{1}_f\mathbf{1}_k^T & \frac{1}{S^2}\mathbf{G} \\ \hline \mathbf{1}_k & \frac{\alpha}{S}\mathbf{I}_k & \frac{\alpha^2}{S^2}\mathbf{I}_k & \mathbf{O} \\ \hline \mathbf{1}_k & -\frac{\alpha}{S}\mathbf{I}_k & \frac{\alpha^2}{S^2}\mathbf{I}_k & \mathbf{O} \\ \hline \mathbf{1}_{n_0} & \mathbf{O} & \mathbf{O} & \mathbf{O} \end{array}\right),$$

missä $\mathbf{G}$ on faktorien sekatuloista muodostuva osuus ($f \times g$-matriisi). Jotta voitaisiin yleensä ottaen päästä ortogonalisoituviin ja/tai kiertosymmetrisiin CCD-kokeisiin, pitää faktoriaaliosa valita siten, että (aikaisemmin mainitun ehdon $\mathbf{F}^T\mathbf{1}_f = \mathbf{0}_k$ lisäksi)

$$\mathbf{F}^T\mathbf{F} = f\mathbf{I}_k \quad , \quad \mathbf{G}^T\mathbf{1}_f = \mathbf{0}_g \quad , \quad \mathbf{G}^T\mathbf{G} = f\mathbf{I}_g \quad \text{ja} \quad \mathbf{G}^T\mathbf{F} = \mathbf{O}.$$

Jälleen täydellisen tai osittaisen 2^k-kokeen muodostavalle faktoriaaliosalle tämä on automaattista, sillä

$$\begin{aligned}
(\mathbf{F}^T\mathbf{F})_{i,j} &= \sum_{t=1}^{N} F_{ti}F_{tj} = \begin{cases} f & \text{jos } i = j \\ 0 & \text{jos } i \neq j \end{cases} \\
(\mathbf{G}^T\mathbf{1}_f)_{(ij)} &= \sum_{t=1}^{N} F_{ti}F_{tj} = 0 \\
(\mathbf{G}^T\mathbf{G})_{(ij),(mn)} &= \sum_{t=1}^{N} F_{ti}F_{tj}F_{tm}F_{tn} = \begin{cases} f & \text{jos } (ij) = (mn) \\ 0 & \text{jos } (ij) \neq (mn) \end{cases} \\
(\mathbf{G}^T\mathbf{F})_{(ij),m} &= \sum_{t=1}^{N} F_{ti}F_{tj}F_{tm} = 0
\end{aligned}$$

Tällöin

$$\begin{aligned}
\mathbf{X}^T\mathbf{X} &= \left(\begin{array}{c|c|c|c} N & \mathbf{0}_k^T & \frac{f}{S^2}\mathbf{1}_k^T + 2\frac{\alpha^2}{S^2}\mathbf{1}_k^T & \mathbf{0}_g^T \\ \hline \mathbf{0}_k & \frac{f}{S^2}\mathbf{I}_k + 2\frac{\alpha^2}{S^2}\mathbf{I}_k & \mathbf{O} & \mathbf{O}^T \\ \hline \frac{f}{S^2}\mathbf{1}_k + 2\frac{\alpha^2}{S^2}\mathbf{1}_k & \mathbf{O} & \frac{f}{S^4}\mathbf{J}_k + 2\frac{\alpha^4}{S^4}\mathbf{I}_k & \mathbf{O}^T \\ \hline \mathbf{0}_g & \mathbf{O} & \mathbf{O} & \frac{f}{S^4}\mathbf{I}_g \end{array}\right) \\
&= N\left(\begin{array}{c|c|c|c} 1 & \mathbf{0}_k^T & \mathbf{1}_k^T & \mathbf{0}_g^T \\ \hline \mathbf{0}_k & \mathbf{I}_k & \mathbf{O} & \mathbf{O}^T \\ \hline \mathbf{1}_k & \mathbf{O} & \frac{f\mathbf{J}_k + 2\alpha^4\mathbf{I}_k}{S^2(f+2\alpha^2)} & \mathbf{O}^T \\ \hline \mathbf{0}_g & \mathbf{O} & \mathbf{O} & \frac{f\mathbf{I}_g}{S^2(f+2\alpha^2)} \end{array}\right).
\end{aligned}$$

Koe on nyt ortogonalisoitavissa, jos

$$\frac{f}{S^2(f+2\alpha^2)} = \frac{fN}{(f+2\alpha^2)^2} = 1 \quad \text{eli} \quad \alpha = \sqrt{\frac{1}{2}\left(\sqrt{fN} - f\right)}$$

ja kiertosymmetrinen, jos $\alpha = \sqrt[4]{f}$. Sekä kiertosymmetrisen että ortogonalisoituvan kokeen aikaansaamiseksi on näin ollen valittava

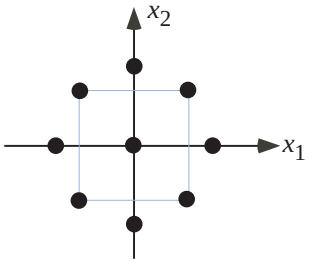
$$\alpha = \sqrt[4]{f} \quad \text{ja} \quad n_0 = 4 - 2k + 4\sqrt{f}$$

(olettaen, että f on neliö).

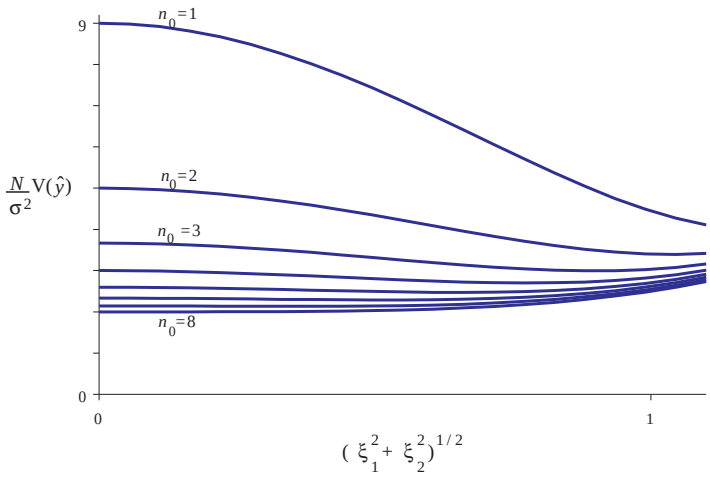
Esimerkki

Kahden faktorin ($k = 2$) täydelliseen 2^k-kokeeseen ($f = 2^2 = 4$) perustuva CCD-koe toisen kertaluvun mallille on ortogonalisoituva ja kiertosymmetrinen, kun valitaan $\alpha = \sqrt[4]{f} = \sqrt{2}$ ja $n_0 = 4 - 2k + 4\sqrt{f} = 8$. Silloin datamatriisin ensimmäisen kertaluvun faktorit ovat

x_1	x_2
-1	-1
-1	1
1	-1
1	1
$\sqrt{2}$	0
0	$\sqrt{2}$
$-\sqrt{2}$	0
0	$-\sqrt{2}$
0	0
$\vdots$	$\vdots$
0	0



Jos valitaan pienempi määrä origon toistoja, niin koe ei enää ole ortogonalisoituva mutta pysyy kiertosymmetrisenä. Kannattaa kuitenkin ottaa $n_0 \geq 3$ toistoja mallin epäsopivuustestin suorittamiseksi ja ennusteen varianssin pienentämiseksi tarkastelualueella:

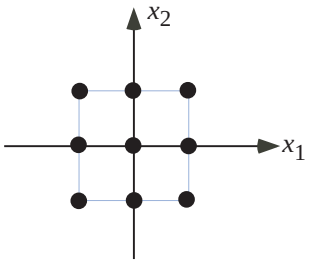


Harjoitustehtävät

1. Etsi toisen kertaluvun mallin ortogonalisoituva kiertosymmetrinen CCD-koe, kun $k = 5$.

2. Jos tarkastelualueen rajoitukset ovat muotoa $c_i \leq x_i \leq d_i$, niin voidaan asettaa CCD-kokeen parametrin arvoksi $\alpha = 1$, jotta koodatut aksiaalikoepisteet olisivat tarkastelualueen rajapinnalla. Silloin CCD-koe on kolmen tason koe. Tapauksessa $k = 2$ datamatriisin ensimmäisen kertaluvun faktorit ovat

$$\begin{array}{rr} \underline{x_1} & \underline{x_2} \\ -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \\ & \\ 1 & 0 \\ 0 & 1 \\ -1 & 0 \\ 0 & -1 \\ & \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{array}$$



Piirrä ennusteen varianssin $NV(\hat{y})/\sigma^2$ tasa-arvokäyriä (Matlabin komento `contour`) alueella $-1 \leq \xi_1, \xi_2 \leq 1$ kun $n_0 = 0, 1, 2$. Mitä voi päätellä?

2.7 Optimaaliset kokeet

On paljon tehtäviä, joissa edellisissä kohdissa esitettyjä ”klassisia” kokeita ei voi käyttää. Mm. koealueen muoto, kokeiden lukumäärä tai regressiomalli saattavat olla sen verran erikoisia, että tarvitaan räätälöity koesuunnittelu. Viime vuosina kehitetyt optimointialgoritmit tarjoavat tehokkaita ja monipuolisia työkaluja tällaisiin koesuunnittelutehtäviin.

Optimaalisen kokeen suunnittelussa etsitään koepisteiden sijainnit koealueella niin, että tavoitefunktio minimoituu. Tavoitefunktioita on useita. *A-optimaalisen* kokeen tavoitefunktio on regressiomallin parametrien skaalattujen varianssien summa, eli

$$J_{\mathrm{A}}(\mathbf{X}) = \frac{1}{\sigma^2}\sum_i V(b_i) = \sum_i c_{ii} = \mathrm{trace}((\mathbf{X}^T\mathbf{X})^{-1}),$$

Esimerkiksi, jos koealue on kuutio $\mathcal{X} = \{1\} \times [-1, 1]^k$, niin kohdan 2.2 tuloksien perusteella tiedetään, että ortogonaalinen kahden tason koe on A-optimaalinen.

Eniten käytetty optimaalisuuskriteeri on *D-optimaalisuus*, jonka tavoitefunktio on

$$J_{\mathrm{D}}(\mathbf{X}) = \frac{1}{\det(\mathbf{X}^T\mathbf{X})} = \det((\mathbf{X}^T\mathbf{X})^{-1}).$$

Determinantti on positiivinen, koska se on positiividefiniitin matriisin $(\mathbf{X}^T\mathbf{X})^{-1}$ ominaisarvojen tulo. D-optimaalinen suunnittelu pysyy D-optimaalisena datan affiinimuunnoksella, koska

$$J_{\mathrm{D}}(\mathbf{XL}) = \frac{1}{\det(\mathbf{L}^T\mathbf{X}^T\mathbf{XL})} = \frac{1}{\det(\mathbf{L}^T)\det(\mathbf{X}^T\mathbf{X})\det(\mathbf{L})} \propto J_{\mathrm{D}}(\mathbf{X}).$$

Koesuunnittelu voidaan siis tehdä koodatuilla faktoreilla. Tätä edullista piirettä ei ole muilla regressiomallin parametrien laatua kuvaavilla optimaalisuuskriteereillä, kuten A-optimaalisuudella.

D-optimaalinen koesuunnittelu minimoi regressiomallin parametrien $\mathbf{b}$ luottamusellipsoidien tilavuuden. Sillä lauseen 1.1 mukaan, suure *confidence ellipsoid*

$$\frac{(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\mathbf{b}-\boldsymbol{\beta})}{(k+1)s^2}$$

on F-jakautunut vapausastein $k+1$ ja $N-k-1$, joten epäyhtälö

$$(\mathbf{b}-\boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\mathbf{b}-\boldsymbol{\beta}) \le (k+1)s^2F_{k+1,N-k-1}(\alpha) =: f^2$$

määrittelee $\mathbf{b}$-keskisen luottamusellipsoidin parametriavaruudessa. Lineaarikuvaus $\mathbf{b} \mapsto \frac{1}{f}(\mathbf{X}^T\mathbf{X})^{1/2}\mathbf{b}$ muuttaa ellipsoidin $k+1$-ulotteiseksi yksikkösäteiseksi palloksi. Ellipsoidin tilavuus on $\left(\det\left(\frac{1}{f}(\mathbf{X}^T\mathbf{X})^{1/2}\right)\right)^{-1} = J_{\mathrm{D}}^{1/2}f^{k+1}$ kertaa pallon tilavuus.

Optimaalisuuskriteeri voi myös perustua vastepinnan varianssiin. Kuten mainittiin kohdassa 2.1, ennusteen varianssin arvo pysyy muuttumattomana affiinimuunnoksen suhteen. Ennusteen normalisoitu varianssi faktoripisteessä $\boldsymbol{\xi}$ on

$$g(\boldsymbol{\xi}, \mathbf{X}) := \frac{N}{\sigma^2} V(\hat{y}) = N\boldsymbol{\xi}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi}.$$

G-optimaalisen kokeen tavoitefunktio on

$$J_{\mathrm{G}}(\mathbf{X}) = \max\{g(\boldsymbol{\xi}, \mathbf{X}) \,|\, \boldsymbol{\xi} \in \mathcal{X}\},$$

eli koesuunnittelussa pyritään minimoimaan vastepinnan suurin varianssi koealueella. Kohdan 2.6 esimerkissä arvioitiin CCD-kokeiden origotoistojen lukumäärän n_0 valinta nimenomaan G-optimaalisuuskriteerin näkökulmasta.

Datamatriisin i:nnen rivin faktoriyhdelmän vastaava normalisoitu varianssi on

$$g_i := \frac{N}{\sigma^2} V(\hat{y}_i) = Nh_{ii},$$

missä h_{ii} on matriisin $\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ i:s lävistäjäalkio. Arvojen g_i keskiarvo ei voi olla maksimiarvoa isompi, joten

$$\begin{aligned} J_{\mathrm{G}}(\mathbf{X}) \;&\geq\; \max_i g_i \geq \frac{1}{N}\sum_i g_i = \sum_i h_{ii} = \mathrm{trace}(\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T) \\ &= \mathrm{trace}(\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}) = \mathrm{trace}(\mathbf{I}_{k+1}) = k+1, \end{aligned}$$

eli $k+1$ on G-optimaalisen tavoitefunktion alaraja.

Kun ensimmäisen kertaluvun mallin datamatriisi ja datavektori jaetaan lohkoihin tutun tapaan

$$\mathbf{X} = (\mathbf{1} \,|\, \mathbf{D}) \quad \text{ja} \quad \boldsymbol{\xi} = \begin{pmatrix} 1 \\ \mathbf{d} \end{pmatrix},$$

niin ennusteen normalisoitu varianssi on

$$g = 1 + N(\mathbf{d} - \bar{\mathbf{d}})^T(\mathbf{D}^T\mathbf{D} - N\bar{\mathbf{d}}\bar{\mathbf{d}}^T)^{-1}(\mathbf{d} - \bar{\mathbf{d}}),$$

missä $\bar{\mathbf{d}} := \frac{1}{N}\mathbf{D}^T\mathbf{1}_N$ on datan keskipiste. Koska matriisi $(\mathbf{D}^T\mathbf{D} - N\bar{\mathbf{d}}\bar{\mathbf{d}}^T)^{-1}$ on positiividefiniitti, niin $g \geq 1$ ja $g = 1$ vain datan keskipisteessä.

Tilastolliset ohjelmistot käyttävät algoritmejä, jotka etsivät optimaaliset suunnittelut diskretoidulla koealueella, eli jatkumo $\mathcal{X}$ korvataan äärellisellä pistejoukolla. Tällaisen kombinatorisen optimointitehtävän globaalisen minimin löytäminen on yleensä raskas laskentatehtävä, sillä jos jokainen faktori diskretoidaan käyttäen m tasoa, niin mahdollisia koesuunnitteluja on m^{kN}. Siksi käytetään heuristisia algoritmejä, joiden antama ratkaisu ei välttämättä ole globaali minimi, mutta joka on yleensä aivan käyttökelpoinen.

Esimerkki

Tutkitaan yhden muuttujan ensimmäisen kertaluvun mallin $\hat{y} = b_0 + b_1 x$ koesuunnitteluja, kun koealue on $-1 \leq x \leq 1$.

Ensimmäisen kertaluvun mallin suunnittelu $\mathbf{D} = [-1, 1]^T$ antaa

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix},$$

joten $J_\mathrm{D}(\mathbf{X}) = 1/\det(\mathbf{X}^T\mathbf{X}) = 1/4$. Mallin parametrien luottamusellipsoidi on

$$(\mathbf{b} - \boldsymbol{\beta})^T\mathbf{X}^T\mathbf{X}(\mathbf{b} - \boldsymbol{\beta}) = 2(b_0 - \beta_0)^2 + 2(b_1 - \beta_1)^2 \leq f^2$$

eli $f/\sqrt{2}$-säteinen kiekko. Ennusteen normalisoitu varianssi on

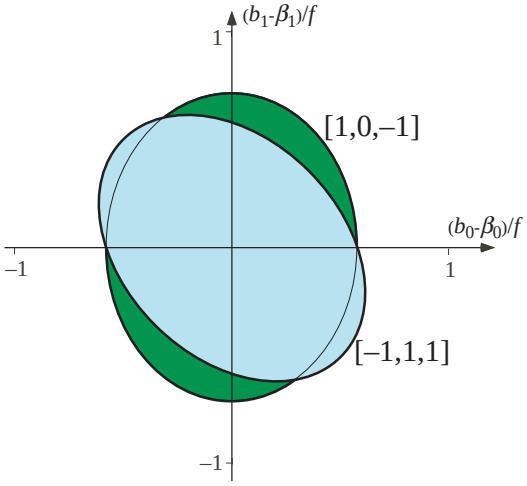
$$g(\boldsymbol{\xi}, \mathbf{X}) = N\boldsymbol{\xi}^T(\mathbf{X}^T\mathbf{X})^{-1}\boldsymbol{\xi} = 2(\,1\;\;\xi\,)\begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}\begin{pmatrix} 1 \\ \xi \end{pmatrix} = 1 + \xi^2,$$

joten $J_\mathrm{G}(\mathbf{X}) = \max\limits_{-1\leq\xi\leq 1} 1 + \xi^2 = 2 = k + 1$, eli koe on G-optimaalinen.

Ensimmäisen kertaluvun mallin suunnittelut $\mathbf{D}_1 = [-1, 0, 1]^T$ ja $\mathbf{D}_2 = [-1, 1, 1]$ antavat

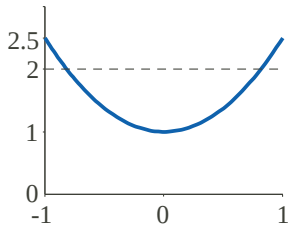
$$\mathbf{X}_1^T\mathbf{X}_1 = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix} \quad \text{ja} \quad \mathbf{X}_2^T\mathbf{X}_2 = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix},$$

joten $J_\mathrm{D}(\mathbf{X}_1) = 1/6$ ja $J_\mathrm{D}(\mathbf{X}_2) = 1/8$, eli suunnittelu $[-1, 1, 1]$ on suunnittelua $[-1, 0, 1]$ parempi D-kriteerin suhteen. Mallin parametrien luottamusellipsoidit ovat



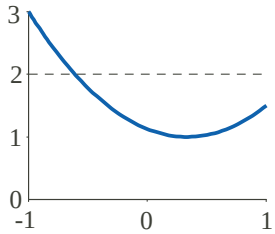
Ennusteen normalisoidut varianssit ovat

$$g(\boldsymbol{\xi}, \mathbf{X}_1) = 1 + \frac{3}{2}\xi^2$$



ja

$$g(\boldsymbol{\xi}, \mathbf{X}_2) = \frac{9}{8} - \frac{3}{4}\xi + \frac{9}{8}\xi^2.$$



joten $J_{\mathrm{G}}(\mathbf{X}_1) = 2.5$ ja $J_{\mathrm{G}}(\mathbf{X}_2) = 3$, eli suunnittelu $[-1, 0, 1]$ on suunnittelua $[-1, 1, 1]$ parempi G-kriteerin suhteen.

Harjoitukset

1. Näytä, että jos datamatriisi on muotoa

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \hline \mathbf{X}_1 \end{pmatrix},$$

   (eli kaikki kokeet toistetaan), niin $g(\boldsymbol{\xi}, \mathbf{X}) = g(\boldsymbol{\xi}, \mathbf{X}_1)$.

2. Etsi sellainen $\zeta \in [-1, 1]$ arvo, että koepisteet $[-1, 0, \zeta, 1]$ antavat D-optimaalisen kokeen toisen kertaluvun mallille $\hat{y} = b_0 + b_1 x + b_2 x^2$.

3. Etsi kahden muuttujan toisen kertaluvun mallin D-optimaalinen koe tilastoohjelmistoa käyttäen, kun $N = 10$ ja koealue on $-1 \leq x_1, x_2 \leq 1$. Vertaile tämän suunnittelun ja kohdan 2.6 tehtävän 2 suunnittelun J_{D} ja J_{G} arvoja, kun $n_0 = 2$.

Luku 3

VASTEEN OPTIMOINTI

Vasteen optimoinnilla tarkoitetaan sellaisen faktorien tasoyhdelmän löytämistä, jolla vaste saa maksimi- tai minimiarvon. Tähän voidaan käyttää tavallista numeerista optimointia, mutta silloin koedatan kohinaa ei oteta huomioon. Tässä luvussa esitetään tilastollisen vastepintamallinnuksen menetelmään perustuva vasteen optimointi. Menettely on kolmivaiheinen.

Seulonta. Ensin luetteloidaan kaikki mahdollisesti vasteeseen vaikuttavat kvantitatiiviset muuttujat ja määrätään niiden käyttöalue. Seulonnan koesarjan tarkoitus on poistaa muuttujia, joilla varianssianalyysin mukaan ei ole merkittävää vaikutusta vasteeseen. Tähän voidaan käyttää kahden tason kokeita. Regressiomallin riittävyys voidaan myös tarkistaa tässä vaiheessa.

screening

region of operability

Gradienttimenetelmä. Seuraavaksi iteroidaan gradienttimenetelmää kunnes vaste ei enää olennaisesti kasva, varianssianalyysi tai epäsopivuuden testaus ilmoittaa toisen kertaluvun mallin tarpeellisuuden tai tullaan käyttöalueen reunalle.

method of steepest descent / ascent

Ääriarvotarkastelu tai harjuanalyysi. Sovitetaan täydellinen toisen kertaluvun malli ja lasketaan sen kriittinen piste. Jos ääriarvo on käyttöalueen sisällä, tarkastellaan sen laatu (maksimi, minimi, tai satulapiste); muussa tapauksessa suoritetaan ns. *harjuanalyysi* .

stationary point

ridge analysis

Seulonnan menetelmät ovat esitettynä edellisissä luvuissa, joten tässä luvussa keskitytään muiden vaiheiden menettelyihin.

3.1 Gradienttimenetelmä

Tämä vaihe on iteratiivinen. Valitaan lähtöpiste $\mathbf{d}_0$ ja toistetaan seuraavat toimenpiteet.

1. Valitaan jokin ensimmäisen kertaluvun mallin datamatriisi $\mathbf{X} = \left(\begin{array}{c|c} \mathbf{1}_N & \mathbf{D} \end{array} \right)$ käyttöalueen sisällä siten, että $\mathbf{d}_0$ on datan keskipiste:

$$\frac{1}{N}\mathbf{D}^T\mathbf{1}_N = \mathbf{d}_0.$$

region of experimentation

Datan kattamaa aluetta kutsutaan *koealueeksi*. Ensimmäisissä iteraatioissa, joissa ollaan kaukana ääriarvosta, voidaan käyttää kahden tason kokeita. Myöhemmissä iteraatioissa kannattaa lisätä toistokokeita keskipisteessä vastepinnan kaarevuuden testamiseen. Erityisen edullista on käyttää kiertosymmetrisiä kokeita, joille ennusteen varianssi on suunnasta riippumaton.

2. Suoritetaan vastaavat kokeet ja sovitetaan ensimmäisen kertaluvun malli

$$y = \mathbf{x}^T\boldsymbol{\beta} + \epsilon.$$

3. Testataan mahdollisten koetoistojen avulla mallin epäsopivuus. Jos malli osoittautuu epäsopivaksi, siirrytään suoraan seuraavaan vaiheeseen (s. 74), jossa sovitetaan toisen kertaluvun malli.

4. Suoritetaan varianssianalyysi. Jos malli ei osoittaudu merkitseväksi, siirrytään suoraan seuraavaan vaiheeseen.

5. Käytetään estimoituja parametrejä $\mathbf{b}$ seuraavalla tavalla. Etsitään yksikkövektori $\mathbf{n}$, jonka suuntaan vaste mallin mukaan kasvaa nopeimmin (maksimointi) tai vähenee nopeimmin (minimointi), ns. *viettosuunta.* Tämä suunta on gradientin

$$\mathrm{grad}(\mathbf{x}^T\mathbf{b}) = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{pmatrix} =: \mathbf{b}_1$$

suunta tai sille vastakkainen suunta.

6. Valitaan jokin askelpituus Δ ja kokeita suorittamalla etsitään vasteet pisteissä

$$\mathbf{d}_0 + i\Delta\mathbf{n} \quad (i = 1, 2, \ldots),$$

kunnes vaste ei enää merkittävästi kasva (maksimointi) tai vähene (minimointi), tai tullaan käyttöalueen reunalle. Olkoon tällainen piste uusi lähtöpiste $\mathbf{d}_0$.

Jos tullaan käyttöalueen reunalle, vastepinnan optimin etsintä jatkuu reunalla. Käyttöalueen reuna oletetaan koostuvan hypertasoista muotoa $\{\mathbf{d} \mid \mathbf{q}^T(\mathbf{d} - \mathbf{d}_{\mathrm{reuna}}) = 0\}$. Etsintäpolku leikkaa käyttöalueen reunan, kun

$$\mathbf{q}^T(\mathbf{d}_0 + i\Delta\mathbf{n} - \mathbf{d}_{\mathrm{reuna}}) = 0,$$

eli kun $i = \mathbf{q}^T(\mathbf{d}_{\text{reuna}} - \mathbf{d}_0)/(\Delta\mathbf{q}^T\mathbf{n})$. Otetaan silloin käyttöön uudet vapaat faktorit $\mathbf{d}' := (x'_1, \ldots, x'_{k-1})^T$, jotka liittyvät alkuperäisiin rajoitettuihin faktoreihin yhtälön

$$\mathbf{d} = \mathbf{d}_{\text{reuna}} + \mathbf{P}\mathbf{d}'$$

kautta, missä $k \times (k-1)$ matriisin $\mathbf{P}$ sarakkeet ovat lineaarisesti riippumattomia ja ortogonaalisia vektorin $\mathbf{q}$ kanssa.

7. Toistetaan kohtien 1–6 menettely kunnes joko

   (a) tullaan ”ulos” kohdista 3. tai 4. tai

   (b) vaste ei enää olennaisesti kasva (maksimointi) tai vähene (minimointi).

Esimerkki

Halutaan maksimoida erään kemiallisen prosessin tuotteen suhteellinen puhtaus (prosenttina). Vasteeseen vaikuttaa kaksi muuttujaa, joiden käyttöalue on $[60, 150] \times [0, \infty]$. Alkupisteeksi valitaan $(80, 60)^T$, ja koealueeksi $[70, 90] \times [30, 90]$. Suoritetaan täydellinen 2^2-koe, ja jokainen koe toistetaan. Koodatut muuttujat ovat

$$X_1 := \frac{x_1 - 80}{(90-70)/2}, \quad X_2 := \frac{x_2 - 60}{(90-30)/2}.$$

Koetulokset ovat

x_1	x_2	X_1	X_2	y
70	30	-1	-1	49.8, 48.1
70	90	-1	$+1$	65.7, 69.4
90	30	$+1$	-1	57.3, 52.3
90	90	$+1$	$+1$	73.1, 77.8

Ensimmäisen kertaluvun mallin

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

parametrien estimaatit $\mathbf{b}$ ovat

$$\begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{8} & 0 \\ 0 & 0 & \frac{1}{8} \end{pmatrix} \begin{pmatrix} 493.5 \\ 27.7 \\ 78.5 \end{pmatrix} = \begin{pmatrix} 61.69 \\ 3.44 \\ 9.81 \end{pmatrix}.$$

Regressiomalli on siis $\hat{y} = 61.69 + 3.44X_1 + 9.81X_2$. Varianssianalyysitaulu mallin epäsopivuuden testaamiseen on

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
epäsopivuus	1	2.10	2.10		
puhdas virhe	4	31.84	7.96	0.264	0.63
residuaali	5	33.94	6.79		

ja siinä on näytettä mallin sopivuudesta.

Varianssianalyysitaulu mallin käyttökelpoisuuden testamiseen on

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
regressio	2	864.81	432.41		
residuaali	5	33.94	6.79	63.71	$3 \cdot 10^{-4}$
kokonaisvariaatio	7	898.75	128.39		

ja siinä ei ole syytä hyväksyä hypoteesia $\beta_1 = \beta_2 = 0$. Testataan vielä jokaisen faktorin tarpeellisuutta mallissa. Hypoteesin $\beta_1 = 0$ testisuure on

$$t = \frac{b_1 - 0}{\sqrt{\text{MSE} \cdot c_{11}}} = \frac{3.44}{\sqrt{6.79/8}} = 3.73,$$

jonka merkitsevyys on 0.0135. Hypoteesin $\beta_2 = 0$ testisuure on

$$t = \frac{b_2 - 0}{\sqrt{\text{MSE} \cdot c_{22}}} = \frac{9.81}{\sqrt{6.79/8}} = 10.65,$$

jonka merkitsevyys on $1.3 \cdot 10^{-4}$. Voidaan siis päätellä, että molemmat faktorit vaikuttavat vasteeseen.

Viettosuunta vasteen maksimointiin on

$$\mathbf{n} = \frac{1}{\sqrt{3.44^2 + 9.81^2}} \begin{pmatrix} 3.44 \\ 9.81 \end{pmatrix} = \begin{pmatrix} 0.331 \\ 0.944 \end{pmatrix},$$

joten maksimin etsinnän pisteet ovat

$$\begin{aligned} X_1 &= 0.331\Delta,\ 2 \cdot 0.331\Delta,\ \cdots \\ X_2 &= 0.944\Delta,\ 2 \cdot 0.944\Delta,\ \cdots \end{aligned}$$

Vastaavat alkuperäisten muuttujien arvot ovat

$$\begin{aligned} x_1 &= 80 + 10 \cdot 0.331\Delta,\ 80 + 10 \cdot 2 \cdot 0.331\Delta,\ \cdots \\ x_2 &= 60 + 30 \cdot 0.944\Delta,\ 60 + 30 \cdot 2 \cdot 0.944\Delta,\ \cdots \end{aligned}$$

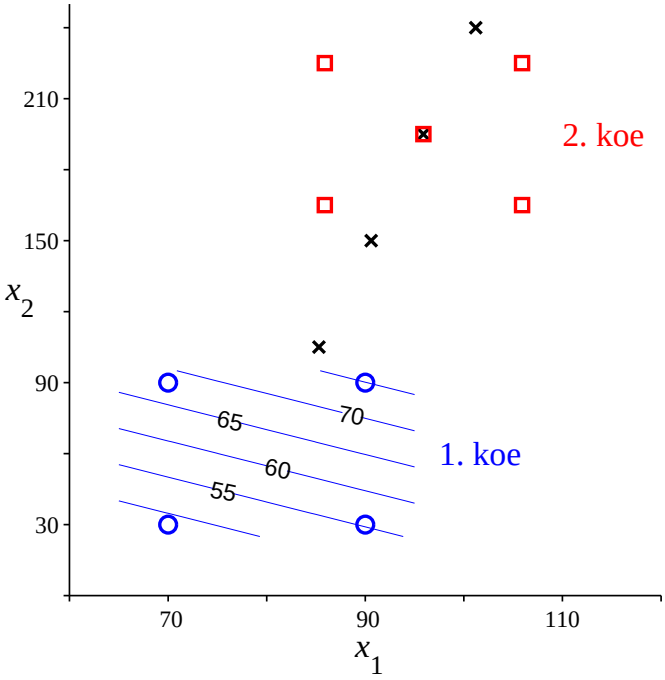
Valitsemme $0.944\Delta = 1.5$ ja suoritetaan kokeita.

x_1	x_2	y
85.3	105	74.3
90.6	150	83.2
95.9	195	84.7
101.2	240	80.1

Suurin vaste tulee pisteessä $(95.9, 195)^T$, joten seuraava iteraatio lähtee sieltä. Valitaan koealueeksi $[85.9, 105.9] \times [165, 225]$ ja suoritetaan täydellinen 2^2-koe, johon on lisätty keskustoistoja. Edellisen koesarjan mittaus keskipisteessä otetaan myös mukaan. Koetulokset ovat

x_1	x_2	X_1	X_2	y
85.9	165	-1	-1	82.2
85.9	225	-1	$+1$	75.8
105.9	165	$+1$	-1	88.5
105.9	225	$+1$	$+1$	83.8
95.9	195	0	0	84.7
95.9	195	0	0	81.9

Mallin sovitus, epäsopivuuden tarkastelu, merkitsevyyden tarkastelu ja seuraavan etsintäsuunnan määrittely jätetään lukijalle. Alla olevassa kuvassa näkyy ensimmäiseen kokeseen (o) perustuvat ennustetut tasa-arvokäyrät, ensimmäisen etsintäpolun kokeet ($\times$) ja toisen kokeen pisteet ($\square$).



Harjoitustehtävät

1. Suorita esimerkin toinen iteraatio loppuun.

2. Alla on erään kokeen data. Onko ensimmäisen kertaluvun malli merkitsevä tai sopiva? Mihin suuntaan lähdet ja mistä pisteestä minimoitaessa kahden faktorin vastetta?

x_1	x_2	y
30	150	39.3
30	160	40.0
40	150	40.9
40	160	41.5
35	155	40.3
35	155	40.5
35	155	40.7
35	155	40.2
35	155	40.6

3. Esitä rajoituksen reuna muodossa $\{(x_1, x_2, x_3)^T \mid \mathbf{q}^T(\mathbf{d} - \mathbf{d}_{\text{reuna}}) = 0\}$ ja esitä reunapisteet uusien faktoreiden (x_1', x_2') funktiona.

   (a) $x_1 \geq 0$ (b) $x_1 + x_2 \leq 10$

3.2 Ääriarvotarkastelu

Kun ensimmäisen kertaluvun malli ei enää sovi, sovitetaan täydellinen toisen kertaluvun malli

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{1 \leq i \leq j \leq k} \beta_{ij} x_i x_j + \epsilon.$$

Merkitään

$$\mathbf{d} := \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$$

ja

$$\mathbf{B} := \frac{1}{2}(\mathbf{B}' + (\mathbf{B}')^T),$$

missä $\mathbf{B}'$ on yläkolmiomatriisi

$$\mathbf{B}' := \begin{pmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1k} \\ 0 & \beta_{22} & \cdots & \beta_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta_{kk} \end{pmatrix}.$$

Tällöin

$$\mathbf{d}^T\mathbf{B}'\mathbf{d} = \operatorname{trace}(\mathbf{d}^T\mathbf{B}'\mathbf{d}) = \operatorname{trace}(\mathbf{B}'\mathbf{d}\mathbf{d}^T) = \sum_{1\le i\le j\le k} \beta_{ij}x_ix_j$$

ja vastaavasti

$$\mathbf{d}^T(\mathbf{B}')^T\mathbf{d} = \sum_{1\le i\le j\le k} \beta_{ij}x_ix_j.$$

Siispä myös

$$\mathbf{d}^T\mathbf{B}\mathbf{d} = \sum_{1\le i\le j\le k} \beta_{ij}x_ix_j$$

ja malli voidaan kirjoittaa matriisimuotoon

$$y = \mathbf{x}^T\boldsymbol{\beta} + \mathbf{d}^T\mathbf{B}\mathbf{d} + \epsilon.$$

Myöskin ennuste voidaan kirjoittaa matriisimuotoon:

$$\hat{y}(\mathbf{d}) = b_0 + \mathbf{d}^T\mathbf{b}_1 + \mathbf{d}^T\mathbf{E}\mathbf{d},$$

missä matriisi $\mathbf{E}$ saadaan ottamalla $\mathbf{B}$:ssä β_{ij}:n paikalle b_{ij}.

Valitaan sellainen toisen kertaluvun mallin datamatriisi $\mathbf{X} = \left(\begin{array}{c|c} \mathbf{1}_N & \mathbf{D} \end{array} \right)$, että $\frac{1}{N}\mathbf{D}^T\mathbf{1}_N = \mathbf{d}_0$, missä $\mathbf{d}_0$ on edellisessä vaiheessa viimeksi saatu datan keskipiste. Usein voidaan käyttää edellisen vaiheen dataa joko sellaisenaan tai täydentäen. Tehdään varianssianalyysi ja testataan (mahdollisten) koetoistojen avulla mallin epäsopivuus. Jos malli ei osoittaudu merkitseväksi tai osoittautuu epäsopivaksi, tehdään koealue isommaksi tai pienemmäksi.

Seuraavaksi käytetään estimoitua mallia ääriarvon etsintään. Muodostetaan

$$\operatorname{grad}(b_0 + \mathbf{d}^T\mathbf{b}_1 + \mathbf{d}^T\mathbf{E}\mathbf{d}) = \mathbf{b}_1 + 2\mathbf{E}\mathbf{d}$$

ja merkitään se $= \mathbf{0}_k$:ksi. Jos $\mathbf{E}$ on singulaarinen, siirrytään harjuanalyysiin (kohta 3.3). Muussa tapauksessa etsitään *kriittinen piste*

$$-\frac{1}{2}\mathbf{E}^{-1}\mathbf{b}_1 =: \mathbf{z}.$$

Jos kriittinen piste sijaitsee koealueen sisällä, tarkastellaan sen laatu matriisin $\mathbf{E}$ ominaisarvojen $\lambda_1, \ldots, \lambda_k$ avulla. Koska $\mathbf{E}$ on symmetrinen matriisi, sen ominaisarvot ovat reaaliset.

- Jos kaikki ominaisarvot ovat positiiviset, $\mathbf{E}$ on positiividefiniitti ja $\mathbf{z}$ on minimipiste, jossa saavutetaan minimaalinen vaste.
- Jos kaikki ominaisarvot ovat negatiiviset, $\mathbf{E}$ on negatiividefiniitti ja $\mathbf{z}$ on maksimipiste, jossa saavutetaan maksimaalinen vaste.
- Muussa tapauksessa $\mathbf{z}$ on satulapiste.

Jos kriittinen piste on satulapiste tai sijaitsee koealueen ulkopuolella, siirrytään harjuanalyysiin.

Koealueen sisällä sijaitsevan kriittisen pisteen laatu voidaan selvittää tarkemmin ominaisvektoreiden avulla. Kirjoitetaan estimoitu malli muotoon

$$\hat{y}(\mathbf{d}) = \hat{y}(\mathbf{z}) + (\mathbf{d} - \mathbf{z})^T \mathbf{E} (\mathbf{d} - \mathbf{z}).$$

Etsitään matriisin $\mathbf{E}$ Schurin hajotelma:

$$\mathbf{E} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^T,$$

missä $\mathbf{Q}$ on ortogonaalimatriisi ja $\mathbf{\Lambda}$ on lävistäjämatriisi, jonka lävistäjäalkiot ovat $\lambda_1, \ldots, \lambda_k$. Huomaa, että $\mathbf{Q}$:n sarakkeet ovat (järjestyksessä) ominaisarvoihin $\lambda_1, \ldots, \lambda_k$ liittyviä ominaisvektoreita. Merkitään $\mathbf{Q}^T(\mathbf{d} - \mathbf{z}) =: \mathbf{e}$, jolloin malli on ns. *kanonista muotoa*

$$\hat{y}(\mathbf{d}) = \hat{y}(\mathbf{z}) + \mathbf{e}^T \mathbf{\Lambda} \mathbf{e} \quad \text{eli} \quad \hat{y}(\mathbf{d}) = \hat{y}(\mathbf{z}) + \sum_{i=1}^{k} \lambda_i e_i^2.$$

Ominaisvektorit määrävät vastepinnan pääakselin suunnat. Ominaisarvojen itseisarvot määrävät pinnan kaarevuuden eli kuinka nopeasti vaste muuttuu siirryttäessä pois kriittisestä pisteestä $\mathbf{z}$ pääakseleita pitkin.

Esimerkki

Edellisen kohdan kemiallisen prosessin puhtauden maksimoinnissa on kokeiden edetessä päästy seuraavaan dataan.

x_1	x_2	X_1	X_2	y
125.9	171.9	-1	-1	93.6
125.9	218.1	-1	$+1$	91.7
145.9	171.9	$+1$	-1	92.5
145.9	218.1	$+1$	$+1$	92.9
135.9	195.0	0	0	96.2
135.9	195.0	0	0	97.0

Sovitetaan ensimmäisen kertaluvun malli

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

koodattuun dataan. Saadaan parametrien estimaatit

$$\mathbf{b} = \begin{pmatrix} 93.98 \\ 0.03 \\ -0.38 \end{pmatrix}.$$

Estimoitu malli on siis

$$\hat{y}(X_1, X_2) = 93.98 + 0.03X_1 - 0.38X_2.$$

Testataan mallin epäsopivuus toistokokein.

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
epäsopivuus	2	21.86	10.93		
puhdas virhe	1	0.32	0.32	34.16	0.12
residuaali	3	22.18	7.39		

Mallin epäsopivuutta on syytä epäillä. Mallin merkitsevyys

variaation lähde	vapausasteet	neliösummat	keskineliöt	F-suure	merkitsevyys
regressio	2	0.57	0.28		
residuaali	3	22.18	7.39	0.04	0.96
kokonaisvariaatio	5	22.75	4.55		

Malli ei ole merkitsevä. Sovitetaan täydellinen toisen kertaluvun malli

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \epsilon.$$

Jotta vastepinnan kaarevuus saadaan paremmin mukaan, on tehty lisäkokeita. Huomaa, että kaikki tehdyt kokeet muodostavat CCD-koesuunnittelun.

x_1	x_2	X_1	X_2	y
121.75	195.0	$-\sqrt{2}$	0	92.7
150.04	195.0	$\sqrt{2}$	0	92.8
135.9	162.3	0	$-\sqrt{2}$	93.4
135.9	227.7	0	$\sqrt{2}$	92.7

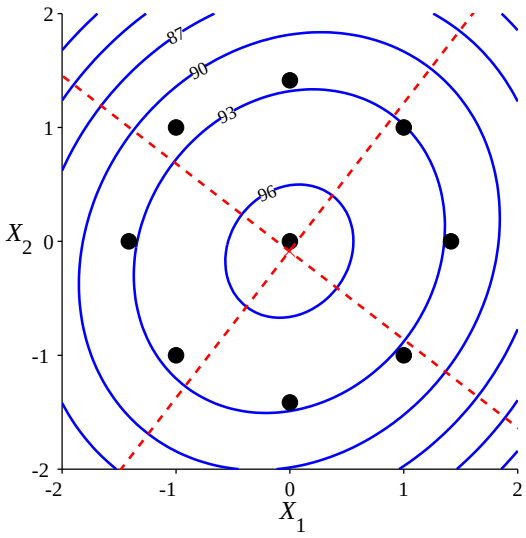
Toisen asteen sovituksesta saadaan parametrien estimaatit

$$\mathbf{b} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_{11} \\ b_{22} \\ b_{12} \end{pmatrix} = \begin{pmatrix} 96.60 \\ 0.03 \\ -0.31 \\ -1.98 \\ -1.83 \\ 0.58 \end{pmatrix},$$

eli estimoitu malli on

$$\hat{y}(X_1, X_2) = 96.60 + 0.03X_1 - 0.31X_2 - 1.98X_1^2 - 1.83X_2^2 + 0.58X_1X_2.$$

Vastepinnan tasa-arvokäyrät näyttävät, että vasteen maksimi löytyy läheltä koe-alueen keskipistettä.



Epäsopivuuden testauksen varianssitaulusta nähdään, että malli on sopiva:

variaation lähde	vapausasteet	neliösummat	keskineliöt	F	merkitsevyys
epäsopivuus	3	0.13	0.04		
puhdas virhe	1	0.32	0.32	0.14	0.92
residuaali	4	0.45	5.09		

Malli on myös merkitsevä:

variaation lähde	vapausasteet	neliösummat	keskineliöt	F-suure	merkitsevyys
regressio	5	25.45	5.09		
residuaali	4	0.45	0.11	44.85	$1.3 \cdot 10^{-3}$
kokonaisvariaatio	9	25.91	2.87		

Kirjoitetaan estimoitu malli sivun 75 muodossa.

$$\begin{aligned}\hat{y}(\mathbf{d}) &= b_0 + \mathbf{d}^T\mathbf{b}_1 + \mathbf{d}^T\mathbf{E}\mathbf{d} \\ &= 96.6 + (X_1\, X_2)\begin{pmatrix} 0.03 \\ -0.31 \end{pmatrix} + (X_1\, X_2)\begin{pmatrix} -1.98 & 0.29 \\ 0.29 & -1.83 \end{pmatrix}\begin{pmatrix} X_1 \\ X_2 \end{pmatrix}\end{aligned}$$

Vastepinnan kriittinen piste löytyy gradientin nollakohdasta:

$$\mathbf{z} = -\frac{1}{2}\mathbf{E}^{-1}\mathbf{b}_1 = \begin{pmatrix} 0.00 \\ -0.09 \end{pmatrix}.$$

Siis koodattuna $Z_1 = 0.00$, $Z_2 = -0.09$ eli $z_1 = 135.9$, $z_2 = 193.0$. Vasteen arvo kriittisessä pisteessä on (tehtävän 1 kaavaa käyttäen)

$$\begin{aligned}\hat{y}(Z_1, Z_2) &= b_0 + \frac{1}{2}(b_1 Z_1 + b_2 Z_2) = 96.60 + \frac{1}{2}(0.03 \cdot 0.00 - 0.31 \cdot (-0.09)) \\ &= 96.61.\end{aligned}$$

Katsotaan vielä kanoninen muoto. Schurin hajotelman $\mathbf{E} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ matriisit ovat

$$\mathbf{Q} = \begin{pmatrix} -0.79 & -0.61 \\ 0.61 & -0.79 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} -2.20 & 0 \\ 0 & -1.61 \end{pmatrix}.$$

Ominaisarvot löytyvät $\mathbf{\Lambda}$:n diagonaalilta ja matriisin $\mathbf{Q}$ pystyrivit määräävät pinnan pääakselien suunnat. Pisteen laatu nähdään matriisin $\mathbf{E}$ ominaisarvoista $\lambda_1 = -2.20$, $\lambda_2 = -1.61$. Matriisi on negatiividefiniitti, eli piste on maksimi, kuten pitääkin. Kanoninen muoto on

$$\hat{y} = \hat{y}(Z_1, Z_2) + \lambda_1 e_1^2 + \lambda_2 e_2^2 = 96.61 - 2.20e_1^2 - 1.61e_2^2,$$

missä $\mathbf{d} = \mathbf{z} + \mathbf{Q}\mathbf{e}$, eli

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 0.00 \\ -0.09 \end{pmatrix} + \begin{pmatrix} -0.79 \\ 0.61 \end{pmatrix} e_1 + \begin{pmatrix} -0.61 \\ -0.79 \end{pmatrix} e_2.$$

Koska $|\lambda_1| > |\lambda_2|$, vastepinta on vähän kaarevampi (eli vaste muuttuu nopeammin) ensimmäisen pääakselin suuntaan kuin toisen akselin suuntaan.

Harjoitustehtävät

1. Näytä, että vaste kriittisessä pisteessä on $\hat{y}(\mathbf{z}) = b_0 + \frac{1}{2}\mathbf{b}_1^T\mathbf{z}$.
2. Vastetta minimoitaessa päädyttiin estimoituun 2. kertaluvun malliin, jossa

   $$b_0 = 0.67, \quad \mathbf{b}_1 = \begin{pmatrix} 1.22 \\ 3.96 \\ -14.52 \end{pmatrix} \quad \text{ja} \quad \mathbf{E} = \begin{pmatrix} 13.37 & 13.50 & -2.49 \\ 13.50 & 23.98 & -10.81 \\ -2.49 & -10.81 & 10.08 \end{pmatrix}.$$

   Missä minimi saavutetaan (vai saavutetaanko ollenkaan) ja mikä on minimivaste?

3.3 Harjuanalyysi

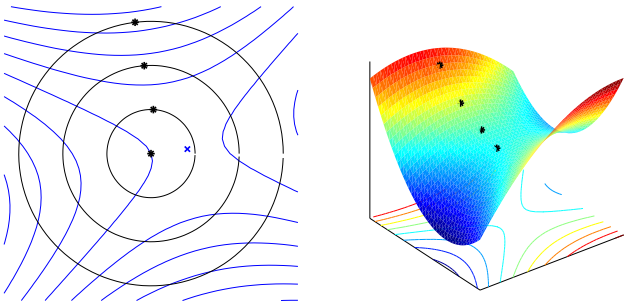
Edellisessä kohdassa vasteen optimointi voi jäädä kesken kahdesta syystä:

1. Toisen kertaluvun vastepinta ei saavuta optimiarvoa koska $\mathbf{E}$ ei ole definiitti, tai

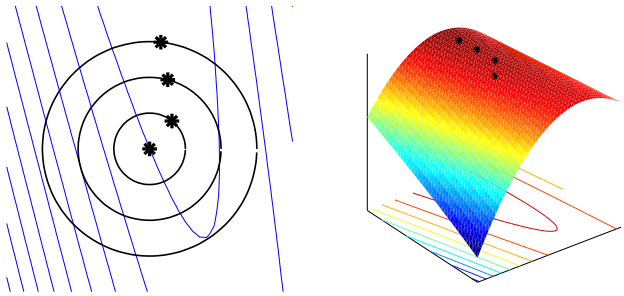
2. optimipiste voi olla niin kaukana koealueen keskipisteestä $\mathbf{d}_0$, ettei siihen kannata luottaa.

ridge analysis
trust-region method

Tällaisissa tapauksissa vasteen parantamiseen voidaan käyttää *harjuanalyysiä*, joka tunnetaan myös numeerisen optimoinnin alalla *luottamisalueen menetelmänä*. Idea on seuraava: valitaan jokin askelpituus Δ ja etsitään pisteet $\mathbf{d}_i$ $(i = 1, 2, \ldots)$ pallojen pinnassa $\{\mathbf{d} \,|\, \|\mathbf{d} - \mathbf{d}_0\| = i\Delta\}$, jotka optimoivat ennustetun vasteen. Kokeita suoritetaan näissä pisteissä kunnes saadut vasteet eivät enää parane. Sitten suoritetaan taas ensimmäisen tai toisen kertaluvun mallin sovittamisen koesarja ja palataan kohdan 3.1 tai kohdan 3.2 menettelyyn. Kahden muuttujan maksimointitehtävässä harjuanalyysi sujuu esimerkiksi näin



tai näin



Seuraava lause auttaa pallon pisteen etsimisessä.

Lause 3.1. *Jos* $\mu > \max\{\lambda_1, \ldots, \lambda_k\}$ *(vast.* $\mu < \min\{\lambda_1, \ldots, \lambda_k\}$*) niin*

$$\mathbf{d}_i = \mathbf{d}_0 + (\mathbf{E} - \mu\mathbf{I})^{-1}(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0)$$

on ainoa globaali vasteen $\hat{y}(\mathbf{d})$ maksimoija (vast. minimoija) joukossa $\{\mathbf{d} \mid \|\mathbf{d} - \mathbf{d}_0\| = \|\mathbf{d}_i - \mathbf{d}_0\|\}$.

Todistus: Matriisin $\mathbf{E}$ Schurin hajotelmaa käyttäen saadaan

$$\mathbf{d}^T(\mathbf{E} - \mu\mathbf{I})\mathbf{d} = \mathbf{d}^T\mathbf{Q}(\mathbf{\Lambda} - \mu\mathbf{I})\mathbf{Q}^T\mathbf{d} = \sum_{j=1}^{k}(\lambda_j - \mu)q_j^2$$

missä $\mathbf{d}$ on mielivaltainen k-vektori ja q_j on vektorin $\mathbf{Q}^T\mathbf{d}$ j:s komponentti. Jos $\mu > \max\{\lambda_1, \ldots, \lambda_k\}$ niin $\mathbf{d}^T(\mathbf{E}-\mu\mathbf{I})\mathbf{d} < 0$ jokaisella nollasta poikkeavalla vektorille $\mathbf{d}$, eli $\mathbf{E} - \mu\mathbf{I}$ on negatiividefiniitti. Vastaavasti, jos $\mu < \min\{\lambda_1, \ldots, \lambda_k\}$, niin $\mathbf{E}-\mu\mathbf{I}$ on positiividefiniitti. Molemmissa tapauksissa $\mathbf{E}-\mu\mathbf{I}$ on ei-singulaarinen ja $\mathbf{d}_i$ on hyvin määritelty. Jos $\|\mathbf{d} - \mathbf{d}_0\| = \|\mathbf{d}_i - \mathbf{d}_0\|$, niin

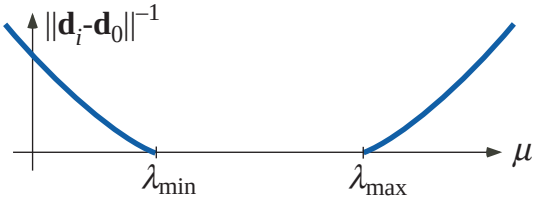
$$\begin{aligned}
\hat{y}(\mathbf{d}) - \hat{y}(\mathbf{d}_i) &= (\mathbf{b}_1 + 2\mathbf{E}\mathbf{d}_i)^T(\mathbf{d} - \mathbf{d}_i) + (\mathbf{d} - \mathbf{d}_i)^T\mathbf{E}(\mathbf{d} - \mathbf{d}_i) \\
&= (\mathbf{d} - \mathbf{d}_i)^T(\mathbf{E} - \mu\mathbf{I})(\mathbf{d} - \mathbf{d}_i) \\
&\quad + (\mu\mathbf{d} - \mu\mathbf{d}_i + \mathbf{b}_1 + 2\mathbf{E}\mathbf{d}_i)^T(\mathbf{d} - \mathbf{d}_i) \\
&= (\mathbf{d} - \mathbf{d}_i)^T(\mathbf{E} - \mu\mathbf{I})(\mathbf{d} - \mathbf{d}_i) + \mu(\|\mathbf{d} - \mathbf{d}_0\|^2 - \|\mathbf{d}_i - \mathbf{d}_0\|^2) \\
&= (\mathbf{d} - \mathbf{d}_i)^T(\mathbf{E} - \mu\mathbf{I})(\mathbf{d} - \mathbf{d}_i),
\end{aligned}$$

josta väite seuraa. □

Pallon säde riippuu suureesta μ seuraavasti:

$$\begin{aligned}
\|\mathbf{d}_i - \mathbf{d}_0\| &= \|(\mathbf{E} - \mu\mathbf{I})^{-1}(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0)\| \\
&= \|\mathbf{Q}(\mathbf{\Lambda} - \mu\mathbf{I})^{-1}\mathbf{Q}^T(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0)\| \\
&= \|(\mathbf{\Lambda} - \mu\mathbf{I})^{-1}\mathbf{Q}^T(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0)\| \\
&= \sqrt{\sum_{j=1}^{k}\frac{p_j^2}{(\lambda_j - \mu)^2}}
\end{aligned}$$

missä p_j on vektorin $\mathbf{Q}^T(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0)$ j:s alkio. Näin $\|\mathbf{d}_i - \mathbf{d}_0\|^{-1}$ on vähenevä μ:n funktio, kun $\mu < \lambda_{\min}$, ja kasvava funktio, kun $\mu > \lambda_{\max}$. Piirtämällä $\|\mathbf{d}_i - \mathbf{d}_0\|^{-1}$ voidaan määritellä arvo μ halutulle säteen arvolle $i\Delta$, ja sitten laskea harjuanalyysin pisteet $\mathbf{d}_i$ lauseen 3.1 kaavaa käyttäen.



Esimerkki

Olkoon kahden muuttujan maksimointitehtävä, missä

$$\mathbf{b}_1 = \begin{pmatrix} 0.93 \\ 0.38 \end{pmatrix}, \quad \mathbf{E} = -\begin{pmatrix} 0.96 & 0.21 \\ 0.21 & 0.04 \end{pmatrix}, \quad \mathbf{d}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Etsi kriittinen piste ja harjuanalyysin pisteet $\mathbf{d}_1$ ja $\mathbf{d}_2$, kun askelpituus on $\Delta = 0.5$.

Ratkaisu

Etsitään kriittinen piste

$$\mathbf{z} = -\frac{1}{2}\mathbf{E}^{-1}\mathbf{b}_1 = \begin{pmatrix} 3.74 \\ -14.87 \end{pmatrix}.$$

Piste on selvästi liian kaukana pisteestä $\mathbf{d}_0$ ollaakseen tarkka arvio todellisen vastepinnan kriittisestä pisteestä.

Tehdään harjuanalyysi. Vasteen maksimipiste ympyrän $\|\mathbf{d}-\mathbf{d}_0\| = i\Delta$ kehällä saadaan lauseesta 3.1:

$$\mathbf{d}_i = \mathbf{d}_0 + (\mathbf{E} - \mu_i\mathbf{I})^{-1}(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0).$$

Ensin on etsittävä $\mu_1, \mu_2 > \max\{\lambda_1, \lambda_2\}$, joilla $\|\mathbf{d}_1 - \mathbf{d}_0\| = \Delta$ ja $\|\mathbf{d}_2 - \mathbf{d}_0\| = 2\Delta$.

$\mathbf{E}$:n Schurin hajotelman $\mathbf{E} = \mathbf{Q}^T\mathbf{\Lambda}\mathbf{Q}$ matriisit ovat

$$\mathbf{Q} = \begin{pmatrix} -0.98 & 0.21 \\ -0.21 & -0.98 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} -1.01 & 0 \\ 0 & 0.01 \end{pmatrix}.$$

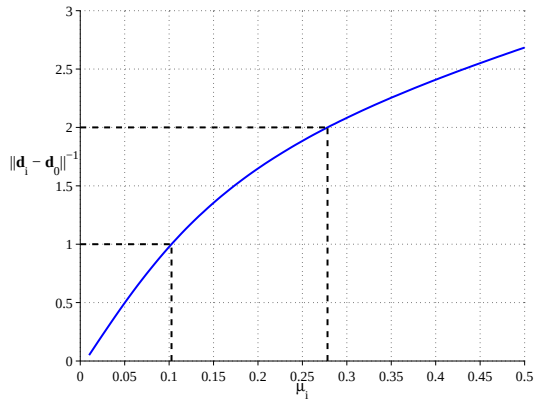
$\mathbf{E}$:n ominaisarvoista nähdään, että $\mu_1, \mu_2 > 0.01$. Vektori $\mathbf{p}$:

$$\mathbf{p} = \mathbf{Q}^T(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0) = \begin{pmatrix} 0.49 \\ 0.09 \end{pmatrix}.$$

Nyt

$$\begin{aligned} \|\mathbf{d}_i - \mathbf{d}_0\|^{-1} &= \frac{1}{\sqrt{(\lambda_1 - \mu_i)^{-2}p_1^2 + (\lambda_2 - \mu_i)^{-2}p_2^2}} \\ &= \frac{1}{\sqrt{(-1.01 - \mu_i)^{-2} \cdot 0.49^2 + (0.01 - \mu_i)^{-2} \cdot 0.09^2}}. \end{aligned}$$

Piirretään tämä, kun $\mu_i \in \{0.01, \ldots, 0.5\}$:



Kuvasta nähdään, että $\|\mathbf{d}_1 - \mathbf{d}_0\|^{-1} = 2$ eli $\|\mathbf{d}_1 - \mathbf{d}_0\| = 0.5$, kun μ_1 on noin 0.28. Numeerisella menetelmällä tai laskemalla yhtälö auki päästään tarkempaan $\mu_1 = 0.2783$. Lasketaan sitten $\mathbf{d}_1$:

$$\mathbf{d}_1 \;=\; \mathbf{d}_0 + (\mathbf{E} - \mu_1\mathbf{I})^{-1}(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0) = \begin{pmatrix} 0.31 \\ 0.39 \end{pmatrix}.$$

Toinen piste $\mathbf{d}_2$ löytyy vastaavasti. Edellisestä kuvasta nähdään, että $\|\mathbf{d}_2 - \mathbf{d}_0\| = 1$, kun μ_2 on noin 0.10, tarkemmin $\mu_2 = 0.1027$. Saadaan toinen piste

$$\mathbf{d}_2 \;=\; \mathbf{d}_0 + (\mathbf{E} - \mu_2\mathbf{I})^{-1}(-\frac{1}{2}\mathbf{b}_1 - \mathbf{E}\mathbf{d}_0) = \begin{pmatrix} 0.25 \\ 0.97 \end{pmatrix}.$$

Menettelyä voi tarkastella sivun 80 toisista kuvista, jossa on piirretty lisäksi kolmas harjuanalyysin piste.

Harjoitustehtävät

1. Todista: jos $\mathbf{E}$ ei ole negatiividefiniitti niin vasteen $\hat{y}(\mathbf{d})$ maksimoija pallossa $\{\mathbf{d} \mid \|\mathbf{d} - \mathbf{d}_0\| \leq \Delta\}$ sijaitsee pallon reunalla.

2. Jos koealueen keskipiste on satulapinnan kriittinen piste niin vastepinnan kasvusuuntia on useampi. Onko tämä ristiriidassa lauseen 3.1 kanssa?

3. Olkoon kahden muuttujan maksimointitehtävä, missä

$$\mathbf{b}_1 = \begin{pmatrix} 0.93 \\ 0.38 \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} -0.8676 & -0.6161 \\ -0.6161 & 1.8676 \end{pmatrix}, \quad \mathbf{d}_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

   Etsi harjuanalyysin pisteet $\mathbf{d}_1$, $\mathbf{d}_2$ ja $\mathbf{d}_3$, kun askelpituus on $\Delta = 0.6$.

Kirjallisuus

ATKINSON, A.C. & DONEV, A.N.: *Optimum Experimental Designs.* Oxford (1992).

BOX, G.E.P. & DRAPER, N.R.: *Empirical Model-Building and Response Surfaces.* Wiley (1987)

BOX, G.E.P. & HUNTER, W.G. & HUNTER, J.S.: *Statistics for Experimenters.* Wiley (1978)

CHRISTENSEN, R.: *Plane Answers to Complex Questions. The Theory of Linear Models.* Springer–Verlag (1996)

DAVIES, O.L. (toim.): *The Design and Analysis of Industrial Experiments.* Oliver and Boyd (1967)

DRAPER, N.R. & SMITH, H.: *Applied Regression Analysis.* Wiley (1998)

EVERITT, B.S & DUNN, G.: *Applied Multivariate Data Analysis.* Arnold (2001)

GUENTHER, W.C.: *Analysis of Variance.* Prentice–Hall (1964)

JOHN, P.W.M.: *Statistical Design and Analysis of Experiments.* SIAM (1998)

JOHNSON, R.A. & WICHERN, D.W.: *Applied Multivariate Statistical Analysis.* Prentice–Hall (1998)

JOHNSTON, J.: *Econometric Methods.* McGraw–Hill (1996)

KHURI, A.I. & CORNELL, J.A.: *Response Surfaces. Designs and Analyses.* Marcel Dekker (1996)

MYERS, R.H. & MONTGOMERY, D.C.: *Response Surface Methodology. Process and Product Optimization Using Designed Experiments.* Wiley (1995)

MONTGOMERY, D.C.: *Design and Analysis of Experiments.* Wiley (1996)