Sakotettujen logististen regressiomenetelmien vertailu

Jani Reinikainen

Pro gradu -tutkielma
Huhtikuu 2026

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

Tarkastajat:
Apul.prof. Joni Virta
Dos. Pekka Nieminen

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Jani Reinikainen: Sakotettujen logististen regressiomenetelmien vertailu
Pro gradu -tutkielma, 40 s., 2 liites.
Tilastotiede
Huhtikuu 2026

Tässä tutkielmassa vertaillaan sakotettujen logististen regressiomenetelmien mukaisia odotettuja ennustevirheitä poikkileikkausaineistoja, joissa vastemuuttujat noudattavat logistista regressiomallia, havaittaessa. Varsinaisille sakotetuille menetelmille vertailukohtana toimivan suurimman uskottavuuden menetelmän lisäksi tarkasteltuja menetelmiä ovat Akaiken informaatiokriteeriin perustuva paras osajoukko ja askeltavat menetelmät, logistinen harju- ja LASSO-regressio sekä sen höllennetty versio. Koska nämä menetelmät olettavat aineiston noudattavan logistista regressiomallia, on ennustevirheeksi valittu Kullback-Leibler-informaatio.

Menetelmien puhtaan empiirisen vertailun sijaan niiden mukaisten odotettujen ennustevirheiden vertailu perustetaan KL-informaation odotusarvon asymptoottiseen approksimaatioon. Sen ja informaatioepäyhtälön perusteella suurimman uskottavuuden estimaattorin osoitetaan tuottavan asymptoottisesti pienimmän mahdollisen odotetun ennustevirheen asymptoottisesti normaalien ja harhattomien estimaattorien joukossa parametriavaruuden nollamittaista osaa lukuun ottamatta. Tästä nähdään, että sakotettu estimaattori ei voi kuulua tähän joukkoon ollakseen asymptoottisesti perusteltavissa kaikkialla parametriavaruudessa.

Logistisen harjuregression käyttämän sakon todetaan puolestaan olevan luonteeltaan sellaista, että se tuottaa asymptoottisin perustein aina jollain menetelmäparametrin arvolla pienemmän odotetun ennustevirheen kuin suurimman uskottavuuden menetelmä. Koska logistisen LASSO-regression mukainen sakko ei vastaavin perustein samaan kykene, jos kaikille regressiokertoimille estimoidaan aina sama nollasta poikkeava merkki, perustellaan logistisen harjuregression tuottavan muita menetelmiä pienemmän ennustevirheen odotusarvon tällaisia aineistoja havaittaessa.

Osana vertailtujen menetelmien mukaisten odotettujen ennustevirheiden asymptoottisten approksimaatioiden muodostamista tässä työssä johdetaan myös logistisen LASSO-regression ja sen höllennetyn version asymptoottiset jakaumat niiden ehdollisia jakaumia ja valintatodennäköisyyksiä hyödyntämällä. Yhdessä niistä simulointiin esitetyn asymptoottisen LARS-algoritmin kanssa nämä tulokset tarjoavat myös mielenkiintoisen ja uuden näkökulman logistisen LASSO-regression mukaiseen odotettuun ennustevirheeseen.

Asiasanat: sakotettu logistinen regressio, informaatioepäyhtälö, KL-informaatio, paras osajoukko, logistinen harju- ja LASSO-regressio, höllennetty LASSO.

Sisällys

1 Johdanto

Satunnaismuuttujan Y_i ehdollisen odotusarvon $\mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i)$ ennustaminen havaittujen selittävien muuttujien arvojen muodostaman vektorin $\mathbf{x}_i \in \mathbb{R}^p$ perusteella erillisestä opetusaineistosta muodostettua tilastollista mallia käyttäen on eräs tärkeimmistä tilastotieteen ja koneoppimisen sovelluskohteista. Kun vastemuuttujan Y_i arvojoukoksi voidaan valita $\{0, 1\}$, on saatu ennuste käytännössä arvio ehdollisesta todennäköisyydestä $\mathrm{P}(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)$. Tällaisia ennusteita tarvitaan muun muassa luoton takaisinmaksuriskiä ja potilaan sairastumisriskiä arvioitaessa.

Vaikka selittävien muuttujien arvot ovat ainakin osin ennalta määrättyjä kokeellisessa tutkimuksessa, ovat ne havainnoivassa tutkimuksessa monesti vastemuuttujan tapaan satunnaisia. Tällöin tilastoyksiköihin liittyvät havaitut arvot $(y_i, \mathbf{x}_i)$ ovat peräisin satunnaisvektoreista $(Y_i, \mathbf{X}_i)$, joille usein oletetaan sama jakauma. Kun lisäksi oletetaan, että tilastoyksiköt ovat toisistaan riippumattomia, voidaan ennustamiseen käytetyn mallin perusjoukkoa edustavaksi opetusaineistoksi valita yksinkertaisesti joku tällaisten vektoreiden satunnaisotos, jonka koko on $n \gg p$. Nämä ovat tyypillisiä oletuksia edellä mainituissa havainnoivan tutkimuksen sovelluskohteissa.

1.1 Tutkimuksen tavoite ja uudet tulokset

Tämän tutkimuksen tarkoituksena on vertailla teoreettisesti menetelmiä, jotka olettavat edellä kuvatun kaltaisen aineiston vastemuuttujan noudattavan luvussa 2 tarkemmin kuvattua logistista regressiomallia tämän oletuksen toteutuessa. Suurimman uskottavuuden menetelmän lisäksi tässä työssä käsiteltyjä menetelmiä ovat Akaiken informaatiokriteeriin perustuva paras osajoukko ja askeltavat menetelmät, logistinen harju- ja LASSO-regressio sekä sen höllennetty versio. Menetelmien vertailu perustetaan odotettuun ennustevirheeseen, sillä se on käytännössä myös se tunnusluku, jonka minimi ristiinvalidoimalla pyritään löytämään. Ennustevirheeksi tähän on valittu Kullback-Leibler-informaatio, sillä logistisen regression yhteydessä sen minimointi vastaa ennusteen log-uskottavuuden odotusarvon maksimointia.

Vaikka tämä ei ole ensimmäinen tutkimus [12, 28, 10], joka näitä menetelmiä vertailee, on se kuitenkin tiettävästi ainoa, jossa KL-informaatioon perustuvia odotettuja ennustevirheitä tarkastellaan teoreettisesti luvussa 2.4 johdetun asymptoottisen approksimaation pohjalta. Lisäksi osana menetelmien odotetun ennustevirheen teoreettista tarkastelua tässä työssä perustellaan oletettavasti ensimmäistä kertaa, että menetelmäparametrilla $\lambda = \lambda_0\sqrt{n}$, jossa λ_0 on sopivasti valittu vakio, logistinen harjuregressio tuottaa asymptoottisesti muita tarkasteltuja menetelmiä pienemmän odotetun ennustevirheen silloin, kun parametrien merkit poikkeavat nollasta ja opetusaineiston jakauma on sellainen, että ne estimoidaan käytännössä aina oikein.

Ehkä työn merkittävin teoreettinen tulos on kuitenkin logistisen LASSO-regression asymptoottisen jakauman johtaminen sen ehdollisten jakaumien ja niiden valintatodennäköisyyksien kautta luvussa 3.3.1 tavalla, joka laajentaa ehdolliseen päättelyyn liittyviä tuloksia [17, 26, 24] mahdollistaen kaikkien mallien asymptoottisten valintatodennäköisyyksien laskemisen luvussa 3.3.2 kuvatulla algoritmilla, kun oikea malli tunnetaan. Tässä työssä näitä tuloksia käytetään osana teoreettista tarkastelua logistisen LASSO-regression KL-informaation odotusarvon asymptoottisessa ap-

proksimaatiossa, ja ne johdetaan luvussa 3.4 myös höllennetylle logistiselle LASSO-regressiolle. Suurimman uskottavuuden estimaattorin asymptoottisen optimaalisuuden todistus luvussa 3 odotetun ennustevirheen suhteen asymptoottisesti normaalien ja harhattomien estimaattorien joukossa on myös mainitsemisen arvoinen tulos.

1.2 Tutkimuksen rakenne ja käytetyt merkinnät

Koska tieto siitä, miten tarkasteltu menetelmä muodostaa estimaattinsa aineiston tuottaneen logistisen regressiomallin tuntemattomasta parametrivektorista $\boldsymbol{\beta}_0 \in \mathbb{R}^p$, auttaa ymmärtämään sen tilastollisia ominaisuuksia, käydään tässä työssä jokaisen käsitellyn menetelmän osalta läpi, miten ja millä oletuksilla se tämän estimaatin ratkaisee, kun aineisto on havaittu. Tämän jälkeen jokaisen menetelmän osalta tarkastellaan sen mukaisen estimaattorin asymptoottista jakaumaa. Asymptoottisesti se määrää suoraan estimaattorin odotetun ennustevirheen, ja saadaan johdettua, kun estimaatin muodostamiseen käytettyä aineistoa pidetään satunnaisena.

Luvun 2 tarkoituksena on esitellä sekä logistinen regressiomalli että teoria, jota myöhemmin tarvitaan valittujen menetelmien mukaisten odotettujen ennustevirheiden arvioinnissa. Kun luvuissa 2.1 ja 2.2 on kuvattu myöhemmin käsiteltyjen sakotettujen menetelmien perustana toimiva suurimman uskottavuuden menetelmä, johdetaan luvussa 2.3 estimaattorien pienintä asymptoottisesti saavutettavissa olevaa odotettua ennustevirhettä rajoittava informaatioepäyhtälö. Luvussa 2.4 käydään vielä läpi, miten KL-informaation odotusarvon asymptoottinen approksimaatio saadaan muodostettua suurimman uskottavuuden estimaattorille.

Sakotettua logistista regressiota käsittelevän luvun 3 tarkoituksena on esitellä vastaavasti kuin edellä tähän työhön valitut luvuissa 3.1, 3.2, 3.3 ja 3.4 tarkemmin kuvatut sakotetut menetelmät. Aluksi kuitenkin käydään läpi teoreettisesti, milloin estimoitavaan parametrivektoriin kohdistuvasta sakosta voi ylipäätään olla hyötyä. Sekä logistisessa harju- että LASSO-regressiossa ja sen höllennetyssä versiossa menetelmäparametrin $\lambda = \lambda_0\sqrt{n}$ oletetaan riippuvan positiivisesta vakiosta λ_0. Tämä oletus eroaa hieman muun muassa Knightin ja Fun [15] käyttämästä $\lambda/\sqrt{n} \to \lambda_0 \geq 0$. Logistisen LASSO-regression osalta luvussa 3.3.2 käydään myös vielä läpi, miten sen ehdottomasta asymptoottisesta jakaumasta voidaan simuloida.

Matemaattisten merkintöjen seuraamisen helpottamiseksi tässä työssä käytetään seuraavia yleisiä käytäntöjä. Satunnaismuuttujaa merkitään tyypillisesti isolla kirjaimella X, ja sen havaittua arvoa vastaavalla pienellä kirjaimella x. Lisäksi vektorit $\mathbf{x}$ ja matriisit $\mathbf{X}$, joiden riveihin viitataan $\mathbf{X}_{j\cdot}$, lihavoidaan. Koska matriiseja merkitään aina isolla kirjaimella ja vektoreita silloin, kun ne ovat satunnaismuuttujia, riippuu merkinnän lopullinen tulkinta kuitenkin kontekstista. Muut merkinnät, siinä määrin kuin tarpeellista, määritellään sitä mukaa, kun niitä käytetään. Normilla $\|\cdot\|$ viitataan tässä Frobenius-normiin

$$\|\mathbf{X}\| = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbf{X}_{ij}^2}, \quad \mathbf{X} \in \mathbb{R}^{n\times p},$$

joka vektoreille ja skalaareille on sama kuin euklidinen normi.

2 Logistinen regressiomalli

Alan Agrestin [1, luku 1] mukaan logistisessa regressiomallissa aineiston oletetaan koostuvan Bernoulli-jakautuneiden riippumattomien satunnaismuuttujien Y_1,...,Y_n havaituista arvoista. Lisäksi näiden vastemuuttujien ehdollisten odotusarvojen π_i oletetaan riippuvan niitä selittävistä muuttujista $\mathbf{X}_i = (X_{i1}, ..., X_{ip})$ sekä tuntemattomasta parametrivektorista $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ logistisen funktion $\pi = \text{logit}^{-1}$ kautta

$$\pi_i = \mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \text{P}(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \pi(\mathbf{x}_i'\boldsymbol{\beta}_0) = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_0)}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta}_0)}.$$

Joseph Berkson [3] perusteli logistisen funktion käytön tällaisessa mallissa jo vuonna 1944 toteamalla sen seuraavan hyvin tietyllä annoksen logaritmilla x_{i2} kuolleiden yksilöiden määrän m_i osuutta kokeessa, jossa sitä annettiin yksilöille J_i. Tämän käytännönläheisen suurten lukujen lakiin nojaavan perustelun

$$\frac{M_i}{|J_i|} = \frac{1}{|J_i|}\sum_{j \in J_i} Y_j \xrightarrow{P} \mathbb{E}\left(Y_i|\mathbf{X}_i = (1, x_{i2})\right) = \pi_i, \quad \text{kun } |J_i| \to \infty$$

lisäksi Berkson totesi logistisen funktion olevan laskennallisesti yksinkertaisemman ja muodoltaan varsin samanlaisen kuin siihen aikaan yleisesti käytetyn, yksilöiden annosherkkyyden oletetulla normaalisuudella perustellun normaalijakauman kertymäfunktion $\pi_i = \Phi(\mathbf{x}_i'\boldsymbol{\beta}_0)$. Logistisen funktion suosioon on vaikuttanut osaltaan myös esimerkiksi David Coxin [7] mainitsema regressiokerrointen β_j tulkinta

$$\frac{\pi(\mathbf{x}_i'\boldsymbol{\beta} + \beta_j)}{1 - \pi(\mathbf{x}_i'\boldsymbol{\beta} + \beta_j)} : \frac{\pi(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \pi(\mathbf{x}_i'\boldsymbol{\beta})} = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta} + \beta_j)}{\exp(\mathbf{x}_i'\boldsymbol{\beta})} = \exp(\beta_j)$$

vetosuhteiden yhteydessä. Ensimmäinen yhtäsuuruus edellä on varsin ilmeinen, sillä logistisen funktion jatkossa tärkeäksi käänteisfunktioksi ja samalla myös mallin vaihtoehtoiseksi muotoiluksi saadaan ratkaistua

$$\pi_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta}_0)}{1 + \exp(\mathbf{x}_i'\boldsymbol{\beta}_0)} = (1 - \pi_i)\exp(\mathbf{x}_i'\boldsymbol{\beta}_0) \Leftrightarrow \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i'\boldsymbol{\beta}_0 = \text{logit}(\pi_i).$$

2.1 Suurimman uskottavuuden estimointi

Ronald Fisherin [9] vuonna 1922 esittämässä suurimman uskottavuuden menetelmässä on tarkoituksena löytää havaitun aineiston $\mathbf{y} = (y_1, ..., y_n)$ oletetusti tuottaneen satunnaismuuttujan $\mathbf{Y} = (Y_1, ..., Y_n)$ jakauman määränneen tilastollisen mallin $f(\mathbf{y}; \boldsymbol{\beta})$ maksimikohta, kun sitä pidetään vain parametrivektorin $\boldsymbol{\beta}$ funktiona. Aluksi on siis tarpeen määritellä logistisen regressiomallin mukainen f. Koska $Y_1, ..., Y_n$ ovat riippumattomia ja koska $\text{P}(Y_i = y_i|\mathbf{X}_i = \mathbf{x}_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, y_i \in \{0, 1\}$, saadaan satunnaisvektorin $\mathbf{Y}$ ehdolliseksi yhteispistetodennäköisyysfunktioksi eli malliksi

$$f(\mathbf{y}; \boldsymbol{\beta}) \overset{\amalg}{=} \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i} = \prod_{i=1}^{n} \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} \left(\frac{1}{1 - \pi_i}\right)^{-1}, \quad \mathbf{y} \in \{0, 1\}^n.$$

Käytännössä $f(\mathbf{y};\boldsymbol{\beta})$:n sijaan voidaan selvittää myös funktion $l(\boldsymbol{\beta}) = \log f(\mathbf{y};\boldsymbol{\beta})$ maksimikohta, sillä se on kummallakin funktiolla sama luonnollisen logaritmin aidon kasvavuuden ansiosta. Kun logit-funktion määritelmä muistetaan, nähdään, että maksimoitavaksi funktioksi saadaan

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) - \log\left(\frac{1}{1-\pi_i}\right)\right) = \sum_{i=1}^{n} \left(y_i \mathbf{x}_i'\boldsymbol{\beta} - \log\big(1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\big)\right).$$

Koska parametriavaruus eli parametrivektorin sallittujen arvojen muodostama joukko on tässä koko $\mathbb{R}^p$, täytyy funktion l paikallisten ääriarvokohtien löytyä gradienttivektorin

$$\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right)\mathbf{x}_i = \sum_{i=1}^{n} \big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta})\big)\mathbf{x}_i = \mathbf{X}'\big(\mathbf{y} - \boldsymbol{\pi}(\mathbf{X}\boldsymbol{\beta})\big)$$

jossa $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]'$ ja jossa $\boldsymbol{\pi}(\mathbf{X}\boldsymbol{\beta})_i = \pi(\mathbf{x}_i'\boldsymbol{\beta}), i = 1, ..., n$, nollakohdista, sillä se on olemassa kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p$. Käytännössä näitä nollakohtia on kuitenkin vain yksi olettaen, että se saavutetaan ja että matriisin $\mathbf{X}$ aste on p, sillä selvästikin funktion l Hessen matriisi

$$\begin{aligned}
\mathbf{H}(\boldsymbol{\beta}) = \nabla^2 l(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})\left(1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\right) - \exp(\mathbf{x}_i'\boldsymbol{\beta})^2}{\left(1+\exp(\mathbf{x}_i'\boldsymbol{\beta})\right)^2}\mathbf{x}_i\mathbf{x}_i' \\
&= -\sum_{i=1}^{n} \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\left(1 - \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})}\right)\mathbf{x}_i\mathbf{x}_i' \\
&= -\sum_{i=1}^{n} \pi(\mathbf{x}_i'\boldsymbol{\beta})\left(1-\pi(\mathbf{x}_i'\boldsymbol{\beta})\right)\mathbf{x}_i\mathbf{x}_i' \\
&= -\mathbf{X}'\mathbf{V}(\mathbf{X}\boldsymbol{\beta})\mathbf{X} = -\left(\mathbf{V}(\mathbf{X}\boldsymbol{\beta})^{1/2}\mathbf{X}\right)'\left(\mathbf{V}(\mathbf{X}\boldsymbol{\beta})^{1/2}\mathbf{X}\right),
\end{aligned}$$

jossa $\mathbf{V}(\mathbf{X}\boldsymbol{\beta}) = \mathrm{diag}[\pi(\mathbf{x}_1'\boldsymbol{\beta})\left(1-\pi(\mathbf{x}_1'\boldsymbol{\beta})\right) \ \cdots \ \pi(\mathbf{x}_n'\boldsymbol{\beta})\left(1-\pi(\mathbf{x}_n'\boldsymbol{\beta})\right)]$, on negatiivisesti semidefiniitti: $\mathbf{a}'\mathbf{H}(\boldsymbol{\beta})\mathbf{a} = -\sum_{i=1}^{n}\mathbf{V}(\mathbf{X}\boldsymbol{\beta})_{ii}(\mathbf{a}'\mathbf{x}_i)^2 \leq 0, \forall\, \mathbf{a} \in \mathbb{R}^p$. Kun lisäksi huomioidaan, että $\mathrm{r}\left(\mathbf{H}(\boldsymbol{\beta})\right) = \mathrm{r}\left(\mathbf{V}(\mathbf{X}\boldsymbol{\beta})^{1/2}\mathbf{X}\right) = \mathrm{r}(\mathbf{X})$ asteisen matriisin $\mathbf{H}(\boldsymbol{\beta})$ kaikki ominaisarvot poikkeavat nollasta, kun $\mathrm{r}(\mathbf{X}) = p$, nähdään, että se on jopa negatiivisesti definiitti taaten sen, että $l(\boldsymbol{\beta})$ on aidosti konkaavi. Vaikka funktion l maksimikohta $\hat{\boldsymbol{\beta}}$ on siten aikaisempien oletusten perusteella yhtälön $\nabla l(\boldsymbol{\beta}) = \mathbf{0}$ ainoa ratkaisu, ei sitä yleisesti voida esittää suljetussa muodossa.

Koska edeltä nähdään, että Hessen matriisi parametrivektorin $\boldsymbol{\beta}$ funktiona on jatkuva, on yhtälön $\nabla l(\boldsymbol{\beta}) = \mathbf{0}$ numeerisen ratkaisun kannalta hyödyllistä huomata, että gradienttivektorille on Taylorin lauseen perusteella voimassa yhtälö

$$\nabla l(\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \nabla l(\boldsymbol{\beta}) + \mathbf{H}(\boldsymbol{\beta})\boldsymbol{\epsilon} + \mathbf{r} = \nabla l(\boldsymbol{\beta}) + \int_0^1 \mathbf{H}(\boldsymbol{\beta} + t\boldsymbol{\epsilon})\boldsymbol{\epsilon}\, dt,$$

jossa $\boldsymbol{\epsilon} \in \mathbb{R}^p$ ja $\mathbf{r} = \nabla l(\boldsymbol{\beta} + \boldsymbol{\epsilon}) - \nabla l(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta})\boldsymbol{\epsilon}$. Kun aikaisempien oletusten mukaan $\mathbf{H}(\boldsymbol{\beta})$ on kääntyvä kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p$, saadaan tästä ratkaistua valitsemalla askelpituudeksi $\alpha = 1$ Newtonin menetelmän päivityskaava

$$\begin{aligned}
\nabla l\big(\boldsymbol{\beta}^{(q)} + (\boldsymbol{\beta}^{(q+1)} - \boldsymbol{\beta}^{(q)})\big) = \nabla l(\boldsymbol{\beta}^{(q)}) + \mathbf{H}(\boldsymbol{\beta}^{(q)})(\boldsymbol{\beta}^{(q+1)} - \boldsymbol{\beta}^{(q)}) + \mathbf{r} &= \mathbf{0} \\
\Leftrightarrow \boldsymbol{\beta}^{(q+1)} = \boldsymbol{\beta}^{(q)} - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\boldsymbol{\beta}^{(q)}) - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\mathbf{r} & \\
\approx \boldsymbol{\beta}^{(q)} - \alpha\mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\boldsymbol{\beta}^{(q)}) = T(\boldsymbol{\beta}^{(q)}), &
\end{aligned}$$

merkitsemällä, että $\boldsymbol{\beta}^{(q+1)} = T(\boldsymbol{\beta}^{(q)})$, jolloin $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(q+1)} - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\mathbf{r}$. Käytännössä päivityskaava siis tuottaa valitun alkuarvon $\boldsymbol{\beta}^{(0)}$ perusteella jonon $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}, ...$, jonka toivotaan suppenevan kohti gradienttivektorin $\nabla l(\boldsymbol{\beta})$ nollakohtaa.

Koska $\mathbf{V}(\mathbf{X}\boldsymbol{\beta})_{ii}$ ja $D_{\boldsymbol{\beta}}\,\mathbf{V}(\mathbf{X}\boldsymbol{\beta})_{ii} = \big(1-2\pi(\mathbf{x}_i'\boldsymbol{\beta})\big)D_{\boldsymbol{\beta}}\,\pi(\mathbf{x}_i'\boldsymbol{\beta})$ ovat jatkuvia, saadaan Cauchy-Schwarzin epäyhtälön ja differentiaalilaskennan väliarvolauseen mukaan

$$\begin{aligned}|\mathbf{V}(\mathbf{X}\boldsymbol{\beta})_{ii} - \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^*)_{ii}| &= \big|\big(1-2\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big)\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big(1-\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big)\mathbf{x}_i'(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\big| \\ &\leq C\|\mathbf{x}_i\|\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|,\end{aligned}$$

jollain $\boldsymbol{\beta}^i \in \mathbb{R}^p$ ja kaikilla $\boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{R}^p$, sillä selvästikin vakio C voidaan tässä valita siten, että $|\big(1-2\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big)\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big(1-\pi(\mathbf{x}_i'\boldsymbol{\beta}^i)\big)| \leq C < 1$. Kun huomioidaan, että tällöin myös $\mathbf{H}(\boldsymbol{\beta})$ on Lipschitz-jatkuva, sillä

$$\begin{aligned}\|\mathbf{H}(\boldsymbol{\beta}) - \mathbf{H}(\boldsymbol{\beta}^*)\| &\leq \|-\mathbf{X}'\|\|(\mathbf{V}(\mathbf{X}\boldsymbol{\beta}) - \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^*))\|\|\mathbf{X}\| \\ &\leq \|\mathbf{X}\|^2\sqrt{\textstyle\sum_{i=1}^n C^2\|\mathbf{x}_i\|^2\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2} \\ &= C\|\mathbf{X}\|^3\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\| = L\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|,\end{aligned}$$

voidaan Newtonin menetelmän tuottaman jonon Nocedalin ja Wrightin [22, luku 3] mukaisesti osoittaa logistisen regression yhteydessä, tehtyjen oletusten perusteella, suppenevan neliöllisesti kohti arvoa $\hat{\boldsymbol{\beta}}$, kunhan alkuarvo ei ole liian kaukana siitä.

Jotta menetelmän tuottaman jonon suppeneminen olisi neliöllistä, tarvitaan, että $\limsup_{q\to\infty} \|\boldsymbol{\beta}^{(q+1)} - \hat{\boldsymbol{\beta}}\|/\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\|^2 \leq M \in \mathbb{R}_{>0}$. Tätä varten on hyödyllistä huomata, että funktion $\mathbf{H}(\boldsymbol{\beta})$ jatkuvuuden perusteella $\|\mathbf{H}(\boldsymbol{\beta}) - \mathbf{H}(\hat{\boldsymbol{\beta}})\| \leq \frac{1}{2}\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\|^{-1}$ kaikilla $\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\| \leq r\}$, kun $r > 0$ on riittävän pieni, sillä tästä saadaan kyseisessä joukossa voimassa olevan epäyhtälön

$$\begin{aligned}\|\mathbf{H}(\boldsymbol{\beta})^{-1}\| &= \|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1} - \mathbf{H}(\boldsymbol{\beta})^{-1}\big(\mathbf{H}(\boldsymbol{\beta}) - \mathbf{H}(\hat{\boldsymbol{\beta}})\big)\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\| \\ &\leq \|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\| + \|\mathbf{H}(\boldsymbol{\beta})^{-1}\|\|\mathbf{H}(\boldsymbol{\beta}) - \mathbf{H}(\hat{\boldsymbol{\beta}})\|\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\| \\ \Leftrightarrow \|\mathbf{H}(\boldsymbol{\beta})^{-1}\| &\leq \frac{\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\|}{1 - \|\mathbf{H}(\boldsymbol{\beta}) - \mathbf{H}(\hat{\boldsymbol{\beta}})\|\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\|} \leq 2\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\|\end{aligned}$$

sekä Taylorin lauseen ja huomion $\nabla l(\hat{\boldsymbol{\beta}}) = \mathbf{0}$ perusteella Newtonin menetelmän päivityskaavaan perustuva epäyhtälö

$$\begin{aligned}\left\|\boldsymbol{\beta}^{(q+1)} - \hat{\boldsymbol{\beta}}\right\| &= \left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}} - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\left(\nabla l(\boldsymbol{\beta}^{(q)}) - \nabla l(\hat{\boldsymbol{\beta}})\right)\right\| \\ &= \left\|\mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\left(\mathbf{H}(\boldsymbol{\beta}^{(q)}) - \int_0^1 \mathbf{H}(\boldsymbol{\beta}^{(q)} + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(q)}))\,dt\right)(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}})\right\| \\ &\leq \left\|\mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\right\| \int_0^1 \left\|\mathbf{H}(\boldsymbol{\beta}^{(q)}) - \mathbf{H}(\boldsymbol{\beta}^{(q)} + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(q)}))\right\| \left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\right\| dt \\ &\leq \left\|\mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\right\| \left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\right\| \int_0^1 L\left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\right\| t\,dt \\ &\leq L\left\|\mathbf{H}(\hat{\boldsymbol{\beta}})^{-1}\right\| \left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\right\|^2 = \tilde{L}\left\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}\right\|^2,\end{aligned}$$

josta neliöllinen suppeneminen sopivalla alkuarvolla seuraa.

Newtonin menetelmän voitaisiin vielä Nocedalin ja Wrightin [22, luku 3] mukaisesti osoittaa tässä suppenevan kaikilla alkuarvoilla $\boldsymbol{\beta}^{(0)}$, kun askelpituudeksi $\alpha = \rho^k$ valitaan viivahaulla kiinteiden vakioiden $\rho \in (0,1)$ ja $c \in (0, \frac{1}{2})$ perusteella riittävän parannuksen $l(\boldsymbol{\beta}^{(q+1)}) \geq l(\boldsymbol{\beta}^{(q)}) - c\alpha \nabla l(\boldsymbol{\beta}^{(q)})' \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1} \nabla l(\boldsymbol{\beta}^{(q)})$ jollain $k \in \mathbb{N}_0$ antavista suurin, olettamalla, että $\|\mathbf{H}(\boldsymbol{\beta}^{(q)})\| \|\mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\| \leq M \in \mathbb{R}_{>0}$ kaikilla q. Tyydytään kuitenkin toteamaan, että jatkossa logististen regressiomallien estimointiin käytetty R-ohjelmiston glm-funktio ei aina suppene [20], vaikka se joskus askelpituuksia α puolittaakin. Koska Newtonin menetelmän päivityskaava voidaan esittää Hastien et al. [13, luku 4] mukaisesti aikaisemmin käytetyillä merkinnöillä myös muodossa

$$\begin{aligned}
\boldsymbol{\beta}^{(q+1)} &= \boldsymbol{\beta}^{(q)} - \alpha \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1} \nabla l(\boldsymbol{\beta}^{(q)}) \\
&= \boldsymbol{\beta}^{(q)} + \alpha \big(\mathbf{X}' \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)}) \mathbf{X}\big)^{-1} \mathbf{X}' \big(\mathbf{y} - \boldsymbol{\pi}(\mathbf{X}\boldsymbol{\beta}^{(q)})\big) \\
&= \big(\mathbf{X}' \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)}) \mathbf{X}\big)^{-1} \mathbf{X}' \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)}) \big(\mathbf{X}\boldsymbol{\beta}^{(q)} + \alpha \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)})^{-1} \big(\mathbf{y} - \boldsymbol{\pi}(\mathbf{X}\boldsymbol{\beta}^{(q)})\big)\big) \\
&= \big(\mathbf{X}' \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)}) \mathbf{X}\big)^{-1} \mathbf{X}' \mathbf{V}(\mathbf{X}\boldsymbol{\beta}^{(q)}) \mathbf{z}(\alpha, \mathbf{y}, \mathbf{X}\boldsymbol{\beta}^{(q)}),
\end{aligned}$$

nähdään, että se periaatteessa vastaa glm-funktion käyttämän iteratiivisesti uudelleen painotetun pienimmän neliösumman menetelmän yhtä iteraatiota.

2.2 Tarkentuvuus ja asymptoottinen normaalisuus

Vaikka suurimman uskottavuuden menetelmällä periaatteessa etsitäänkin tässä vain ehdollisen uskottavuusfunktion $f(\mathbf{y}; \boldsymbol{\beta}) = f(\mathbf{y}|\mathbf{X}; \boldsymbol{\beta}) = f(\mathbf{y}, \mathbf{X}; \boldsymbol{\beta}) / f(\mathbf{X})$ maksimikohta $\hat{\boldsymbol{\beta}}$ tietyllä aineistolla $(\mathbf{y}, \mathbf{X})$, voidaan sen tilastollisia ominaisuuksia tutkia, kun aineisto ja siten myös suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X})$ ovat satunnaismuuttujia. Tilastollisen päättelyn kannalta mielenkiinto kohdistuu erityisesti tämän estimaattorin asymptoottiseen jakaumaan, joka seuraavassa johdetaan Neweyn ja McFaddenin [21] esittämiä yleisiä tuloksia hyödyntäen.

Oletetaan jatkossa, että myös matriisi $\mathbf{X} = [\mathbf{x}_1 \ \cdots \ \mathbf{x}_n]'$ on otos jostain jakaumasta, jonka toisten momenttien matriisi $\mathbb{E}(\mathbf{X}_1 \mathbf{X}_1')$ on kääntyvä ja siten positiivisesti definiitti. Koska tällöin $\mathbb{E}\big((\mathbf{X}_1'(\boldsymbol{\beta} - \boldsymbol{\beta}_0))^2\big) = (\boldsymbol{\beta} - \boldsymbol{\beta}_0)' \mathbb{E}(\mathbf{X}_1 \mathbf{X}_1')(\boldsymbol{\beta} - \boldsymbol{\beta}_0) > 0$ kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p \setminus \{\boldsymbol{\beta}_0\}$, niin $\mathrm{P}\big(\mathbf{X}_1'(\boldsymbol{\beta} - \boldsymbol{\beta}_0) \neq 0\big) = \mathrm{P}(\mathbf{X}_1'\boldsymbol{\beta} \neq \mathbf{X}_1'\boldsymbol{\beta}_0) > 0$, kun $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$. Kun vielä huomioidaan, että sekä $\pi(\cdot)$ että $1 - \pi(\cdot)$ ovat aidosti monotonisia, nähdään, että malli on identifioituva: $\mathrm{P}\big(f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}) \neq f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}_0)\big) > 0$, kun $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$.

Koska $f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}) / f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}_0)$ on tällöin vakiomuuttuja vain, jos $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, voidaan todellinen parametrivektori $\boldsymbol{\beta}_0$ osoittaa funktion $\mathbb{E}\big(\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})\big)$ ainoaksi maksimikohdaksi $\boldsymbol{\beta}_0^*$. Tämä tunnistettavuus seuraa luonnollisen logaritmin aidon konkaavisuuden vuoksi kaikilla $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$ voimassa olevasta Jensenin epäyhtälöstä

$$\begin{aligned}
&\mathbb{E}\big(\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})\big) - \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}_0)\big) \\
&= \mathbb{E}\left(\log\left(\frac{f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})}{f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}_0)}\right)\right) < \log \mathbb{E}\left(\frac{f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}) f(\mathbf{X}_1')}{f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta}_0) f(\mathbf{X}_1')}\right) = 0,
\end{aligned}$$

joka on määritelty, sillä epäyhtälön $\log(1 + \exp(z)) \leq |z| + \log 2$ perusteella

$$\begin{aligned}
\mathbb{E}\big(\big|\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})\big|\big) &= \mathbb{E}\big(\big|Y_1 \mathbf{X}_1'\boldsymbol{\beta} - \log\big(1 + \exp(\mathbf{X}_1'\boldsymbol{\beta})\big)\big|\big) \\
&\leq \mathbb{E}(|Y_1 \mathbf{X}_1'\boldsymbol{\beta}| + |\mathbf{X}_1'\boldsymbol{\beta}| + \log 2) \leq 2\mathbb{E}(\|\mathbf{X}_1\|)\|\boldsymbol{\beta}\| + \log 2 < \infty
\end{aligned}$$

kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p$ odotusarvon $\mathrm{tr}\big(\mathbb{E}(\mathbf{X}_1 \mathbf{X}_1')\big) = \mathbb{E}(\mathbf{X}_1' \mathbf{X}_1)$ olemassaolon ansiosta.

Käytännössä havaintojen riippumattomuudesta ja samoin jakautuneisuudesta vahvan suurten lukujen lain perusteella seuraava satunnaisten funktioiden $\frac{1}{n}l(\boldsymbol{\beta})$ pisteittäinen suppeneminen melkein varmasti

$$\forall \boldsymbol{\beta} \in \mathbb{R}^p, \tfrac{1}{n}l(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\mathbf{X}_i';\boldsymbol{\beta}) \xrightarrow{a.s.} \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\boldsymbol{\beta})\big)$$

yhdistettynä $\boldsymbol{\beta}_0$:n tunnistettavuuteen ja log-uskottavuusfunktion konkaavisuuteen riittää tässä Neweyn ja McFaddenin [21] mukaan takaamaan, että $\hat{\boldsymbol{\beta}}$ on olemassa yhtä lähestyvällä todennäköisyydellä ja että $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$. Tätä estimaattorin ominaisuutta, joka seuraavaksi vielä osoitetaan, kutsutaan yleisesti tarkentuvuudeksi.

Ensinnäkin oleellista on huomata, että myös $Q_0(\boldsymbol{\beta}) = \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\boldsymbol{\beta})\big)$ on konkaavi, sillä pisteittäinen suppeneminen säilyttää konkaavisuuden. Koska konkaavi funktio on jatkuva määrittelyjoukkonsa sisäpisteissä, on selvää, että Q_0 on jatkuva myös joukossa $C = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| \leq r\}, r > 0$. Koska satunnaisten konkaavien funktioiden pisteittäinen stokastinen suppeneminen joukossa $\mathbb{R}^p$ johtaa Andersenin ja Gillin [2] mukaan myös niiden tasaiseen stokastiseen suppenemiseen kaikissa $\mathbb{R}^p$:n kompakteissa osajoukoissa, nähdään erityisesti joukon C osalta, että

$$\lim_{n\to\infty} \mathrm{P}\left(\sup_{\boldsymbol{\beta}\in C} \left|\tfrac{1}{n}l(\boldsymbol{\beta}) - \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\boldsymbol{\beta})\big)\right| > \epsilon\right) = 0, \quad \forall \epsilon > 0.$$

Olkoon $\epsilon > 0$ ja $\boldsymbol{\beta}_m$ funktion $Q_n(\boldsymbol{\beta}) = \frac{1}{n}l(\boldsymbol{\beta})$ maksimikohta joukossa C. Koska tällöin $Q_n(\boldsymbol{\beta}_m) - \epsilon/3 > Q_n(\boldsymbol{\beta}_0) - 2\epsilon/3$, saadaan edellä todetusta tasaisesta suppenemisesta $\sup_{\boldsymbol{\beta}\in C} |Q_n(\boldsymbol{\beta}) - Q_0(\boldsymbol{\beta})| \xrightarrow{P} 0$ yhtä lähestyvällä todennäköisyydellä

$$Q_0(\boldsymbol{\beta}_m) > Q_n(\boldsymbol{\beta}_m) - \epsilon/3 > Q_n(\boldsymbol{\beta}_0) - 2\epsilon/3 > Q_0(\boldsymbol{\beta}_0) - \epsilon.$$

Kun nyt $S = \{\boldsymbol{\beta} \in C : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| < \min\{r, \delta\}\}$ ja $\delta > 0$, nähdään, että C:ssä jatkuvana funktiona Q_0 saavuttaa arvon $\sup_{\boldsymbol{\beta}\in C\setminus S} Q_0(\boldsymbol{\beta})$ jollain $\tilde{\boldsymbol{\beta}}_0$ kompaktissa joukossa $C \setminus S$. Koska $\boldsymbol{\beta}_0$ on tunnistettavissa, voidaan valita $\epsilon = Q_0(\boldsymbol{\beta}_0) - Q_0(\tilde{\boldsymbol{\beta}}_0)$, jolloin edellisen epäyhtälön perusteella $Q_0(\boldsymbol{\beta}_m) > Q_0(\tilde{\boldsymbol{\beta}}_0)$ yhtä lähestyvällä todennäköisyydellä. Tästä nähdään, että $\boldsymbol{\beta}_m \xrightarrow{P} \boldsymbol{\beta}_0$, sillä edellinen on voimassa kaikilla $\delta > 0$.

Koska kaikille $\boldsymbol{\beta} \in \mathbb{R}^p \setminus C$ on lisäksi jollain $\lambda \in (0,1)$ olemassa konveksi kombinaatio $\lambda\boldsymbol{\beta}_m + (1-\lambda)\boldsymbol{\beta} \in C \setminus S$, kun $\boldsymbol{\beta}_m \in S$, nähdään funktion Q_n konkaavisuuden perusteella tällöin, että $Q_n(\boldsymbol{\beta}_m) \geq Q_n(\lambda\boldsymbol{\beta}_m + (1-\lambda)\boldsymbol{\beta}) \geq \lambda Q_n(\boldsymbol{\beta}_m) + (1-\lambda)Q_n(\boldsymbol{\beta})$ ja siten edelleen että $(1-\lambda)Q_n(\boldsymbol{\beta}_m) \geq (1-\lambda)Q_n(\boldsymbol{\beta})$. Tästä seuraa, että estimaattori $\hat{\boldsymbol{\beta}}$ on tarkentuva $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$, sillä selvästikin $\boldsymbol{\beta}_m$ on funktion Q_n maksimikohta myös joukossa $\mathbb{R}^p$ todennäköisyydellä, joka lähestyy yhtä.

Siirrytään seuraavaksi tarkastelemaan estimaattorin $\hat{\boldsymbol{\beta}}$ asymptoottista jakaumaa. Koska Taylorin lauseen mukaan $\mathbf{0} = \nabla l(\hat{\boldsymbol{\beta}}) = \nabla l(\boldsymbol{\beta}_0) + \mathbf{H}(\tilde{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, jollain vektorilla $\tilde{\boldsymbol{\beta}} \in \{\boldsymbol{\beta}_0 + t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) : 0 < t < 1\}$, nähdään, että $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = -\mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\nabla l(\boldsymbol{\beta}_0)$. Tässä matriisien $h(\boldsymbol{\beta};\mathbf{X}_i) = -\pi(\mathbf{X}_i'\boldsymbol{\beta})(1-\pi(\mathbf{X}_i'\boldsymbol{\beta}))\mathbf{X}_i\mathbf{X}_i'$, jotka ovat riippumattomia ja samoin jakautuneita, otoskeskiarvo $\frac{1}{n}\mathbf{H}(\boldsymbol{\beta})$ on aina jatkuva. Kun lisäksi huomioidaan, että $\|h(\boldsymbol{\beta};\mathbf{X}_i)\| \leq \|\mathbf{X}_i\mathbf{X}_i'\| = \|\mathbf{X}_i\|^2 = \mathrm{tr}(\mathbf{X}_i'\mathbf{X}_i) = \mathrm{tr}(\mathbf{X}_i\mathbf{X}_i'), \forall \boldsymbol{\beta} \in \mathbb{R}^p$ ja että $\mathrm{tr}\big(\mathbb{E}(\mathbf{X}_i\mathbf{X}_i')\big) < \infty$ tehtyjen oletusten perusteella, nähdään, että siihen voidaan soveltaa kompaktissa joukossa C tasaista suurten lukujen lakia

$$\sup_{\boldsymbol{\beta}\in C} \left\| -\frac{1}{n}\sum_{i=1}^{n} \pi(\mathbf{X}_i'\boldsymbol{\beta})\left(1-\pi(\mathbf{X}_i'\boldsymbol{\beta})\right)\mathbf{X}_i\mathbf{X}_i' + \mathbb{E}\left(\pi(\mathbf{X}_1'\boldsymbol{\beta})\left(1-\pi(\mathbf{X}_1'\boldsymbol{\beta})\right)\mathbf{X}_1\mathbf{X}_1'\right) \right\| \xrightarrow{P} 0.$$

Vektorin $\tilde{\boldsymbol{\beta}}$ määritelmän perustella on myös selvää, että $\tilde{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$, sillä $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$, josta $n^{-1}\mathbf{H}(\tilde{\boldsymbol{\beta}}) \xrightarrow{P} \mathrm{E}\big(h(\boldsymbol{\beta}_0; \mathbf{X}_1)\big)$ edellisen perusteella seuraa. Koska matriisin kääntäminen on jatkuva operaatio, saadaan vielä, että $n\mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1} \xrightarrow{P} \mathrm{E}\big(h(\boldsymbol{\beta}_0; \mathbf{X}_1)\big)^{-1}$.

Gradienttivektoriin $\nabla l(\boldsymbol{\beta}) = \sum_{i=1}^{n}\big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta})\big)\mathbf{x}_i$ voidaan toisaalta soveltaa keskeistä raja-arvolausetta, sillä $\frac{1}{n}\nabla l(\boldsymbol{\beta}_0)$ on riippumattomien ja samoin jakautuneiden muuttujien otoskeskiarvo. Koska $\mathrm{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \pi(\mathbf{x}_i'\boldsymbol{\beta}_0)$, nähdään esimerkiksi iteroidun odotusarvon perusteella summattavista termeistä, että

$$\mathrm{E}\big(\big(Y_i - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)\mathbf{X}_i\big) = \mathrm{E}\big(\big(\pi(\mathbf{X}_i'\boldsymbol{\beta}_0) - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)\mathbf{X}_i\big) = \mathbf{0}$$

ja edelleen niiden ehdollisten varianssien avulla niistä että

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}_0) = \mathrm{E}\Big(\big(Y_i - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)^2\mathbf{X}_i\mathbf{X}_i'\Big) &= \mathrm{E}\big(\mathrm{Var}(Y_i|\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i'\big) \\ &= \mathrm{E}\big(\pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big(1 - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)\mathbf{X}_i\mathbf{X}_i'\big)\,.\end{aligned}$$

Kun nämä tulokset yhdistetään, saadaan keskeisen raja-arvolauseen ja Slutskyn lauseen perusteella estimaattorin $\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X})$ asymptoottiseksi jakaumaksi

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) &= -\sqrt{n}\mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\nabla l(\boldsymbol{\beta}_0) \\ &= -n\mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)\mathbf{X}_i\right) \\ &\xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \mathrm{E}\big(h(\boldsymbol{\beta}_0; \mathbf{X}_1)\big)^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathrm{E}\big(h(\boldsymbol{\beta}_0; \mathbf{X}_1)\big)^{-1}\Big) = \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big),\end{aligned}$$

sillä $\mathcal{I}(\boldsymbol{\beta}_0) = -\mathrm{E}\big(h(\boldsymbol{\beta}_0; \mathbf{X}_1)\big)$ on oletuksen $\mathrm{E}(\mathbf{X}_1\mathbf{X}_1') > \mathbf{0}$ perusteella positiivisesti definiitti: $\mathbf{a}'\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{a} = \mathrm{E}\big(\pi(\mathbf{X}_1'\boldsymbol{\beta}_0)\big(1 - \pi(\mathbf{X}_1'\boldsymbol{\beta}_0)\big)(\mathbf{a}'\mathbf{X}_1)^2\big) > 0, \forall \mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$. Käytännön kannalta on myös hyvä huomata, että Hessen matriisista $-\frac{1}{n}\mathbf{H}(\hat{\boldsymbol{\beta}})$ saadaan tällöin Fisherin informaatiomatriisin $\mathcal{I}(\boldsymbol{\beta}_0)$ tarkentuva estimaattori.

Todetaan vielä, että suurimman uskottavuuden estimaattori on tarkentuva ja asymptoottisesti normaali aikaisempien oletusten perusteella, vaikka $f(\mathbf{y}; \boldsymbol{\beta})$ ei olisi aineiston tuottanut malli millään $\boldsymbol{\beta} \in \mathbb{R}^p$. Todellisen parametrivektorin $\boldsymbol{\beta}_0$ sijaan tällöin tosin estimoidaan vain funktion $\mathrm{E}\big(\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})\big)$ oletettua maksimikohtaa $\boldsymbol{\beta}_0^* \in \mathbb{R}^p$. Koska $\|h(\boldsymbol{\beta}; \mathbf{X}_i)\| + \|\nabla \log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta})\| \leq \mathrm{tr}(\mathbf{X}_i\mathbf{X}_i') + \|\mathbf{X}_i\|, \forall \boldsymbol{\beta} \in \mathbb{R}^p$ ja $\mathrm{tr}\big(\mathrm{E}(\mathbf{X}_i\mathbf{X}_i')\big) + \mathrm{E}\big(\|\mathbf{X}_i\|\big) < \infty$, niin $\nabla^2\mathrm{E}\big(\log f(Y_1|\mathbf{X}_1'; \boldsymbol{\beta})\big) = \mathrm{E}\big(h(\boldsymbol{\beta}; \mathbf{X}_1)\big) < \mathbf{0}$. Tästä nähtävä Q_0:n aito konkaavisuus riittää $\boldsymbol{\beta}_0^*$:n tunnistettavuuteen, ja johtaa vastaavasti kuin edellä tarkentuvuuden $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0^*$ lisäksi yhtälöön $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^* = -\mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\nabla l(\boldsymbol{\beta}_0^*)$. Vaikka tässäkin saadaan vastaavin perustein kuin edellä, että

$$\tfrac{1}{n}\mathbf{H}(\tilde{\boldsymbol{\beta}}) \xrightarrow{P} \mathrm{E}\big(h(\boldsymbol{\beta}_0^*; \mathbf{X}_1)\big) = -\mathrm{E}\big(\pi(\mathbf{X}_1'\boldsymbol{\beta}_0^*)\big(1 - \pi(\mathbf{X}_1'\boldsymbol{\beta}_0^*)\big)\mathbf{X}_1\mathbf{X}_1'\big) = -\mathcal{I}(\boldsymbol{\beta}_0^*),$$

on syytä muistaa, että odotusarvo otetaan aina aineiston määränneen jakauman suhteen. Kun lisäksi gradienttivektorin $\nabla l(\boldsymbol{\beta})$ summattavista termeistä huomataan $\boldsymbol{\beta}_0^*$:n määritelmän perusteella, että $\mathrm{E}\big(\nabla \log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta}_0^*)\big) = \nabla\mathrm{E}\big(\log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta}_0^*)\big) = \mathbf{0}$ ja että $\mathrm{E}\big(\nabla \log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta}_0^*)\nabla \log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta}_0^*)'\big) = \mathrm{E}\big(\big(Y_i - \pi(\mathbf{X}_i'\boldsymbol{\beta}_0^*)\big)^2\mathbf{X}_i\mathbf{X}_i'\big) = \mathcal{J}(\boldsymbol{\beta}_0^*)$, saadaan vastaavasti kuin edellä keskeisen raja-arvolauseen ja Slutskyn lauseen perusteella vektoria $\boldsymbol{\beta}_0^*$ estimoitaessa $\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X})$:n asymptoottiseksi jakaumaksi

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big).$$

Jos nyt $\mathrm{P}\big(\mathbf{X}_1 \in \{\mathbf{x}_1 \in \mathbb{R}^p : \mathrm{E}(Y_1|\mathbf{X}_1 = \mathbf{x}_1) \neq \pi(\mathbf{x}_1'\boldsymbol{\beta}_0^*)\}\big) = 0$, voidaan mallia pitää olennaisesti oikeana. Tällöin $\mathcal{J}(\boldsymbol{\beta}_0^*) = \mathcal{I}(\boldsymbol{\beta}_0^*)$ vastaavasti kuin edellä.

2.3 Informaatioepäyhtälö ja asymptoottinen tehokkuus

Tarkastellaan seuraavaksi, mitä edellisessä luvussa kuvatun logistisen regressiomallin todellista parametrivektoria $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ estimoivan aineiston funktion $\boldsymbol{\delta} = \boldsymbol{\delta}(\mathbf{Y}, \mathbf{X})$ lineaarikombinaatioiden varianssien alarajoista voidaan sanoa, ja aloitetaan johtamalla ne harhattomille [1] estimaattoreille $\boldsymbol{\delta}_u$ asettava informaatioepäyhtälö Lehmannin ja Casellan [19, luku 2] esittämiä tuloksia hyödyntäen. Olkoon $\delta_a = \mathbf{a}'\boldsymbol{\delta}, \mathbf{a} \in \mathbb{R}^p \setminus \{\mathbf{0}\}$ minkä tahansa estimaattorin $\boldsymbol{\delta}$ lineaarikombinaatio, jolle $\nabla_{\boldsymbol{\beta}}\mathbb{E}(\delta_a) \in \mathbb{R}^p$ on olemassa ja saadaan osittaisderivoimalla odotusarvon sisällä. Tällöin Cauchy-Schwarzin epäyhtälöä olemassa oleviksi oletettuihin variansseihin sovellettaessa saadaan

$$\mathrm{Var}(\delta_a)\mathrm{Var}\big(\mathbf{b}'\nabla l(\boldsymbol{\beta}_0)\big) \geq \mathrm{Cov}\big(\delta_a, \mathbf{b}'\nabla l(\boldsymbol{\beta}_0)\big)^2 \Leftrightarrow \mathrm{Var}(\delta_a) \geq \frac{\mathbf{b}'\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{b}}{\mathbf{b}'\mathbf{C}\mathbf{b}}, \forall \mathbf{b} \in \mathbb{R}^p \setminus \{\mathbf{0}\},$$

kun $\boldsymbol{\gamma} = \big(\mathrm{Cov}\big(\delta_a, \frac{\partial}{\partial\beta_1} l(\boldsymbol{\beta}_0)\big), ..., \mathrm{Cov}\big(\delta_a, \frac{\partial}{\partial\beta_p} l(\boldsymbol{\beta}_0)\big)\big)$ ja $\mathbf{C} = \mathrm{Cov}\big(\nabla l(\boldsymbol{\beta}_0)\big)$. Koska tämä epäyhtälö on voimassa kaikilla $\mathbf{b} \neq \mathbf{0}$, nähdään edelleen, että

$$\begin{aligned}\mathrm{Var}(\delta_a) \geq \max_{\mathbf{b}} \frac{\mathbf{b}'\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{b}}{\mathbf{b}'\mathbf{C}\mathbf{b}} &= \max_{\mathbf{b}} \frac{\mathbf{b}'\mathbf{Q}_C'\mathbf{D}_C^{1/2}\mathbf{D}_C^{-1/2}\mathbf{Q}_C\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_C'\mathbf{D}_C^{-1/2}\mathbf{D}_C^{1/2}\mathbf{Q}_C\mathbf{b}}{\mathbf{b}'\mathbf{Q}_C'\mathbf{D}_C\mathbf{Q}_C\mathbf{b}} \\ &= \max_{\mathbf{v}} \frac{\mathbf{v}'\mathbf{D}_C^{-1/2}\mathbf{Q}_C\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_C'\mathbf{D}_C^{-1/2}\mathbf{v}}{\mathbf{v}'\mathbf{v}},\end{aligned}$$

jossa $\mathbf{Q}_C'\mathbf{D}_C\mathbf{Q}_C$ on kääntyväksi oletetun matriisin $\mathbf{C}$ ominaisarvohajotelma [2], sillä muuttujanvaihdossa tarvittavan lineaarikuvauksen $\mathbf{v} = \mathbf{D}_C^{1/2}\mathbf{Q}_C\mathbf{b}$ määräävä matriisi on kääntyvä. Merkitään edelleen, että $\mathbf{A} = \mathbf{D}_C^{-1/2}\mathbf{Q}_C\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_C'\mathbf{D}_C^{-1/2}$, jolloin sen ominaisarvohajotelmaa hyödynnettäessä epäyhtälö saa muodon

$$\mathrm{Var}(\delta_a) \geq \max_{\mathbf{v}} \frac{\mathbf{v}'\mathbf{Q}_A'\mathbf{D}_A\mathbf{Q}_A\mathbf{v}}{\mathbf{v}'\mathbf{v}} = \max_{\mathbf{v}} \frac{\mathbf{v}'\mathbf{Q}_A'\mathbf{D}_A\mathbf{Q}_A\mathbf{v}}{\mathbf{v}'\mathbf{Q}_A'\mathbf{Q}_A\mathbf{v}} = \max_{\|\mathbf{w}\|=1} \mathbf{w}'\mathbf{D}_A\mathbf{w} = \max_{\|\mathbf{w}\|=1} \sum_{i=1}^{p} \lambda_i w_i^2,$$

jossa $\mathbf{w} = \tilde{\mathbf{w}}\|\tilde{\mathbf{w}}\|^{-1}$, lineaarikuvauksen $\tilde{\mathbf{w}} = \mathbf{Q}_A\mathbf{v}$ bijektiivisyyden perusteella.

Koska $\mathbf{w}$ on yksikkövektori, saadaan oikean puolen summan maksimi valitsemalla vektori $\mathbf{w}$ siten, että sen suurinta ominaisarvoa λ_{max} vastaava komponentti w_{max} on yksi muiden ollessa nolla. Todetaan vielä, että kun $\mathbf{E}$ ja $\mathbf{F}$ ovat neliömatriiseja, niin matriisien $\mathbf{EF}$ ja $\mathbf{FE}$ ominaisarvot ovat samat, sillä aina, kun $\mathbf{x}$ on matriisin $\mathbf{EF}$ ominaisvektori ja λ sitä vastaava ominaisarvo, niin $\mathbf{FEFx} = \lambda\mathbf{Fx}$. Näin ollen matriisilla $\mathbf{A}$ on samat ominaisarvot kuin matriisilla $\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{Q}_C'\mathbf{D}_C^{-1/2}\mathbf{D}_C^{-1/2}\mathbf{Q}_C = \boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{C}^{-1}$. Kun vielä huomataan, että $\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{C}^{-1}\boldsymbol{\gamma} = \boldsymbol{\gamma}'\mathbf{C}^{-1}\boldsymbol{\gamma}\boldsymbol{\gamma}$ ja että $\boldsymbol{\gamma}\boldsymbol{\gamma}'$:n asteen perusteella matriisilla $\boldsymbol{\gamma}\boldsymbol{\gamma}'\mathbf{C}^{-1}$ on vain yksi nollasta poikkeava ominaisarvo, saadaan

$$\mathrm{Var}(\delta_a) \geq \max_{\|\mathbf{w}\|=1} \sum_{i=1}^{p} \lambda_i w_i^2 = \lambda_{\max} = \boldsymbol{\gamma}'\mathbf{C}^{-1}\boldsymbol{\gamma}.$$

Palautetaan edellisestä luvusta mieleen, että $\mathbb{E}\big(\nabla \log f(Y_i|\mathbf{X}_i'; \boldsymbol{\beta}_0)\big) = \mathbf{0}$ kaikilla i, jolloin gradienttivektorin komponenteista nähdään, että $\mathbb{E}(\delta_a)\mathbb{E}\left(\frac{\partial}{\partial\beta_j} l(\boldsymbol{\beta}_0)\right) = 0$,

[1] Lehmannin ja Casellan määritelmässä harhattomuutta vaaditaan kaikilla $\boldsymbol{\beta}_0 \in \mathbb{R}^p$.

[2] Symmetrisen matriisin $\mathbf{A} \in \mathbb{R}^{p\times p}$ ominaisarvohajotelma on $\mathbf{Q}_A'\mathbf{D}_A\mathbf{Q}_A$, ja siinä ominaisarvot löytyvät matriisista $\mathbf{D}_A = \mathrm{diag}[\lambda_1 \cdots \lambda_p]$. Ominaisvektorit sisältäville matriiseille pätee $\mathbf{Q}_A'\mathbf{Q}_A = \mathbf{I}$.

ja siten aluksi tehtyjen oletusten perusteella vektorin $\boldsymbol{\gamma}$ komponenteista iteroitua odotusarvoa hyödyntäen, että

$$\begin{aligned}\mathrm{Cov}\big(\delta_a, \tfrac{\partial}{\partial\beta_j} l(\boldsymbol{\beta}_0)\big) &= \mathbb{E}\left(\delta_a(\mathbf{Y},\mathbf{X})\tfrac{\partial}{\partial\beta_j}\log f(\mathbf{Y}|\mathbf{X};\boldsymbol{\beta}_0)\right)\\ &= \mathbb{E}\left(\textstyle\sum_{\mathbf{z}\in\{0,1\}^n}\delta_a(\mathbf{z},\mathbf{X})\left(\tfrac{\partial}{\partial\beta_j}\log f(\mathbf{z}|\mathbf{X};\boldsymbol{\beta}_0)\right) f(\mathbf{z}|\mathbf{X};\boldsymbol{\beta}_0)\right)\\ &= \mathbb{E}\left(\textstyle\sum_{\mathbf{z}\in\{0,1\}^n}\delta_a(\mathbf{z},\mathbf{X})\tfrac{\partial}{\partial\beta_j} f(\mathbf{z}|\mathbf{X};\boldsymbol{\beta}_0)\right)\\ &= \tfrac{\partial}{\partial\beta_j}\mathbb{E}\big(\delta_a(\mathbf{Y},\mathbf{X})\big)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.\end{aligned}$$

Kun havaintojen riippumattomuuden sekä edellisen luvun tulosten perusteella todetaan vielä gradienttivektorin kovarianssimatriisista, että

$$\begin{aligned}\mathbf{C} = \mathbb{E}\big(\nabla l(\boldsymbol{\beta}_0)\nabla l(\boldsymbol{\beta}_0)'\big) &= \sum_{i=1}^n\sum_{j=1}^n \mathbb{E}\big(\nabla\log f(Y_i|\mathbf{X}_i';\boldsymbol{\beta}_0)\nabla\log f(Y_j|\mathbf{X}_j';\boldsymbol{\beta}_0)'\big)\\ &\overset{\perp\!\!\!\perp}{=} \sum_{i=1}^n \mathbb{E}\Big(\big(Y_i-\pi(\mathbf{X}_i'\boldsymbol{\beta}_0)\big)^2\mathbf{X}_i\mathbf{X}_i'\Big) = n\mathcal{I}(\boldsymbol{\beta}_0),\end{aligned}$$

saadaan informaatioepäyhtälö, kun $b(\mathbf{a}'\boldsymbol{\delta}) = \mathbb{E}(\mathbf{a}'\boldsymbol{\delta}) - \mathbf{a}'\boldsymbol{\beta}$, lopulta muotoon [3]

$$\begin{aligned}\mathrm{Var}(\delta_a) = \mathbf{a}'\mathrm{Cov}(\boldsymbol{\delta})\mathbf{a} &\geq \tfrac{1}{n}\nabla_{\boldsymbol{\beta}}\mathbb{E}(\delta_a)|'_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\nabla_{\boldsymbol{\beta}}\mathbb{E}(\delta_a)|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}\\ &= \tfrac{1}{n}\mathbf{a}'(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})\mathcal{I}(\boldsymbol{\beta}_0)^{-1}(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})'\mathbf{a}\\ \Leftrightarrow \mathbf{a}'\big(\mathrm{Cov}(\boldsymbol{\delta}) - \tfrac{1}{n}(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})\mathcal{I}(\boldsymbol{\beta}_0)^{-1}(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})'\big)\mathbf{a} &\geq 0,\end{aligned}$$

josta $\mathrm{Cov}(\boldsymbol{\delta}) - \frac{1}{n}(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})\mathcal{I}(\boldsymbol{\beta}_0)^{-1}(\mathbf{I}+\mathbf{J}_{b(\boldsymbol{\delta})|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}})' \geq \mathbf{0}$ voidaan todeta.

Käytännössä kaikille edellä mainitut ehdot täyttäville harhattomille estimaattoreille $\boldsymbol{\delta}_u$ tämä tarkoittaa sitä, että $\mathrm{Cov}(\sqrt{n}\boldsymbol{\delta}_u) \geq \mathcal{I}(\boldsymbol{\beta}_0)^{-1}$. Jotta olisi selvää, että tämä tulos ei suoraan yleisty asymptoottiseen tarkasteluun, todetaan seuraavaksi, että kaikille asymptoottisesti harhattomille estimaattoreille $\sqrt{n}(\boldsymbol{\delta}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathbf{Z}$, joiden rajajakaumalla on äärellinen kovarianssimatriisi $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\Sigma}_{\boldsymbol{\delta}}$, saadaan Portmanteaulauseesta kaikilla $\mathbf{a}\in\mathbb{R}^p\setminus\{\mathbf{0}\}$ voimassa oleva epäyhtälö

$$\begin{aligned}&\liminf_{n\to\infty}\mathbf{a}'\mathbb{E}\big(\sqrt{n}(\boldsymbol{\delta}_n-\boldsymbol{\beta}_0)\sqrt{n}(\boldsymbol{\delta}_n-\boldsymbol{\beta}_0)'\big)\mathbf{a}\\ &= \liminf_{n\to\infty} n\mathbb{E}\big((\mathbf{a}'\boldsymbol{\delta}_n - \mathbf{a}'\boldsymbol{\beta}_0)^2\big) \geq \mathrm{Var}(\mathbf{a}'\mathbf{Z}) = \mathbf{a}'\boldsymbol{\Sigma}_{\boldsymbol{\delta}}\mathbf{a},\end{aligned}$$

kun matriisi $\mathbf{W}_{\boldsymbol{\delta}} = \liminf_{n\to\infty}\mathbb{E}\big(\sqrt{n}(\boldsymbol{\delta}_n-\boldsymbol{\beta}_0)\sqrt{n}(\boldsymbol{\delta}_n-\boldsymbol{\beta}_0)'\big)$ on olemassa. Tästä nähtävä tulos $\mathbf{W}_{\boldsymbol{\delta}} \geq \boldsymbol{\Sigma}_{\boldsymbol{\delta}}$ voidaan tulkita siten, että tämän estimaattorin asymptoottisen jakauman kovarianssimatriisi on optimistinen arvio sen varsinaisesta keskineliövirhematriisista otoskoon ollessa suuri. Päättely toki vaatii, että $n\to\infty$.

Edellisen tuloksen perusteella on selvää, että informaatioepäyhtälö ei vielä takaa, että $\boldsymbol{\Sigma}_{\boldsymbol{\delta}} \geq \mathcal{I}(\boldsymbol{\beta}_0)^{-1}$, vaikka estimaattori olisi harhaton. Tässä määritellyssä säännöllisessä logistisessa regressiomallissa tämä epäyhtälö on kuitenkin voimassa kaikille asymptoottisesti normaaleille ja harhattomille estimaattoreille parametriavaruuden nollamittaista osaa lukuun ottamatta [19, luku 6]. Koska $\hat{\boldsymbol{\beta}}$ on tällainen estimaattori ja $\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$, sanotaan, että se on asymptoottisesti tehokas. On kuitenkin hyvä huomata, että se ei ole harhaton [4].

[3] Tässä on hyvä huomata, että $\mathbf{J}_{\boldsymbol{\delta}} = \left[\tfrac{\partial}{\partial\beta_j}\delta_i\right] \implies \mathbf{J}'_{\boldsymbol{\delta}}\mathbf{a} = \left(\tfrac{\partial}{\partial\beta_1}\boldsymbol{\delta}'\mathbf{a}, ..., \tfrac{\partial}{\partial\beta_p}\boldsymbol{\delta}'\mathbf{a}\right) = \nabla_{\boldsymbol{\beta}}(\mathbf{a}'\boldsymbol{\delta})$.

[4] Cordeiron ja McCullaghin [6] mukaan harhan karkea arvio on $\boldsymbol{\beta}_0 p/n$.

2.4 Ennustevirhe ja odotettu ennustevirhe

Vaikka suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\beta}}$ onkin log-uskottavuusfunktion maksimikohta estimointiin käytetyssä aineistossa, on hyvä muistaa, että varsinainen mielenkiinto kohdistuu todelliseen parametrivektoriin $\boldsymbol{\beta}_0 \in \mathbb{R}^p$. Koska luvun 2.2 perusteella $\boldsymbol{\beta}_0$:n tunnistettavuus logistisessa regressiomallissa tarkoittaa sitä, että kaikille kiinteille estimaateille $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}_0$ on voimassa epäyhtälö

$$\mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\boldsymbol{\beta}_0)\big) - \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\hat{\boldsymbol{\beta}})\big) = -\mathbb{E}\left(\log \frac{f(Y_1|\mathbf{X}_1';\hat{\boldsymbol{\beta}})}{f(Y_1|\mathbf{X}_1';\boldsymbol{\beta}_0)}\right) > 0,$$

voidaan siinä esiintyvää Kullback-Leibler-informaatioksi $D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}})$ kutsuttua odotusarvojen erotusta käyttää estimaatin ennustevirheenä. Käytännössä eri tavoilla saatujen estimaattien vertailuun tarvitaan kuitenkin vain näistä jälkimmäistä odotusarvoa, jonka tarkentuva estimaattori vahvan suurten lukujen lain mukaan estimointiin käytetystä aineistosta riippumattomassa testiaineistossa laskettu

$$\frac{1}{n}l(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} l(\hat{\boldsymbol{\beta}}; Y_i|\mathbf{X}_i') = \frac{1}{n}\sum_{i=1}^{n} \log f(Y_i|\mathbf{X}_i';\hat{\boldsymbol{\beta}}) \xrightarrow{a.s.} \mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\hat{\boldsymbol{\beta}})\big)$$

on, sillä selvästikin $\mathbb{E}\big(\log f(Y_1|\mathbf{X}_1';\boldsymbol{\beta}_0)\big)$ on estimaatista riippumaton vakio.

Kun estimaattoria pidetään jälleen estimointiin käytetystä opetusaineistosta riippuvana satunnaismuuttujana $\hat{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X})$, saadaan KL-informaatioon perustuvasta ennustevirheestä siitä riippuva satunnaismuuttuja. Tarkastellaan seuraavaksi tämän ennustevirheen odotusarvoa testiaineistoon kuuluvaa yksittäistä havaintoa $(y_0, \mathbf{x}_0)$ hyödyntäen Konishin ja Kitagawan [16, luku 3] mukaisesti Taylorin polynomiin, jossa $h(\boldsymbol{\beta};\mathbf{x}_0) = -\pi(\mathbf{x}_0'\boldsymbol{\beta})\big(1-\pi(\mathbf{x}_0'\boldsymbol{\beta})\big)\mathbf{x}_0\mathbf{x}_0'$, perustuvan approksimaation

$$\begin{aligned}\log f(y_0|\mathbf{x}_0';\hat{\boldsymbol{\beta}}) = l(\hat{\boldsymbol{\beta}};y_0|\mathbf{x}_0') \approx{}& l(\boldsymbol{\beta}_0;y_0|\mathbf{x}_0') + \nabla l(\boldsymbol{\beta}_0;y_0|\mathbf{x}_0')'(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\\ &+ \tfrac{1}{2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'h(\boldsymbol{\beta}_0;\mathbf{x}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\end{aligned}$$

avulla. Koska opetus- ja testiaineisto ovat riippumattomia, $\nabla l(\boldsymbol{\beta}_0;Y_0|\mathbf{X}_0') \perp\!\!\!\perp \hat{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X})$ ja $\hat{\boldsymbol{\beta}}(\mathbf{Y},\mathbf{X}) \perp\!\!\!\perp h(\boldsymbol{\beta}_0;\mathbf{X}_0)$. Kun lisäksi muistetaan, että $\mathbb{E}\big(\nabla l(\boldsymbol{\beta}_0;Y_0|\mathbf{X}_0')\big) = \mathbf{0}$ ja että $\mathbb{E}\big(h(\boldsymbol{\beta}_0;\mathbf{X}_0)\big) = -\mathcal{I}(\boldsymbol{\beta}_0)$, saadaan edelliseen approksimaatioon perustuen, että [5]

$$\begin{aligned}\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}})\big) &= \mathbb{E}\big(l(\boldsymbol{\beta}_0;Y_0|\mathbf{X}_0')\big) - \mathbb{E}\big(l(\hat{\boldsymbol{\beta}};Y_0|\mathbf{X}_0')\big)\\ &\approx -\tfrac{1}{2}\mathrm{tr}\big(\mathbb{E}\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'h(\boldsymbol{\beta}_0;\mathbf{X}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)\big)\big)\\ &= \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\mathbb{E}\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\big)\big)\\ &= \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}) + b(\hat{\boldsymbol{\beta}})b(\hat{\boldsymbol{\beta}})'\big)\big),\end{aligned}$$

jossa $b(\hat{\boldsymbol{\beta}}) = \mathbb{E}(\hat{\boldsymbol{\beta}}) - \boldsymbol{\beta}_0$ on estimaattorin harha, sillä selvästikin

$$\begin{aligned}\mathbb{E}\big((\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'\big) ={}& \mathbb{E}\big(\big(\hat{\boldsymbol{\beta}}-\mathbb{E}(\hat{\boldsymbol{\beta}})+b(\hat{\boldsymbol{\beta}})\big)\big(\hat{\boldsymbol{\beta}}-\mathbb{E}(\hat{\boldsymbol{\beta}})+b(\hat{\boldsymbol{\beta}})\big)'\big)\\ ={}& \mathrm{Cov}(\hat{\boldsymbol{\beta}}) + \big(\mathbb{E}(\hat{\boldsymbol{\beta}})-\mathbb{E}(\hat{\boldsymbol{\beta}})\big)b(\hat{\boldsymbol{\beta}})'\\ &+ b(\hat{\boldsymbol{\beta}})\big(\mathbb{E}(\hat{\boldsymbol{\beta}})-\mathbb{E}(\hat{\boldsymbol{\beta}})\big)' + b(\hat{\boldsymbol{\beta}})b(\hat{\boldsymbol{\beta}})'.\end{aligned}$$

[5] Huomaa, että tulos $\mathbb{E}\big(l(\boldsymbol{\beta}_0^*;Y_0|\mathbf{X}_0')\big) - \mathbb{E}\big(l(\hat{\boldsymbol{\beta}};Y_0|\mathbf{X}_0')\big) \approx \frac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0^*)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}) + b(\hat{\boldsymbol{\beta}})b(\hat{\boldsymbol{\beta}})'\big)\big)$ saadaan, vaikka $\boldsymbol{\beta}_0^*$ ei olisi todellinen parametrivektori.

Suurimman uskottavuuden estimaattorin asymptoottisen harhattomuuden ja jakauman $\mathrm{N}\big(\boldsymbol{\beta}_0, n^{-1}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ perusteella sen KL-informaatioon perustuvan ennustevirheen odotusarvon approksimaatioksi saadaan siten $0.5\,p/n$, kun opetusaineiston koko n on suuri ja $n(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)'$ on tasaisesti integroituva. Tulos ei kuitenkaan ole tarkka, sillä Taylorin polynomiin perustuvan approksimaation virhe, joka tosin katoaa, kun $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$, on $O(\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|^3)$ funktion h osittaisderivaattojen jatkuvuuden perusteella. Lisäksi asymptoottiset tulokset periaatteessa vaativat, että $n \to \infty$.

Jotta arvion $0.5\,p/n$ tarkkuudesta saisi paremman kuvan, on taulukossa 1 esitetty Monte Carlo -estimaatit KL-informaatioon perustuvan ennustevirheen odotusarvosta, kun aineiston määränneen logistisen regressiomallin parametrivektorista $\boldsymbol{\beta}_0 = (1,1,1,1,1,0,0,0,0,0)$ estimoidaan suurimman uskottavuuden menetelmällä ensimmäiset $\beta_{0_1}, ..., \beta_{0_p}$. Selittävistä muuttujista X_{11} on vakio loppujen noudattaessa multinormaalijakaumaa, jossa $\mathrm{Cov}(X_{1i}, X_{1j}) = 1$, kun $i = j$ ja muuten ρ.

	p	ρ	n=200	n=500	n=1000	n=2000	n=5000
$\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 \| \hat{\boldsymbol{\beta}})\big)$	5	0	0.014	0.005	0.0026	0.0013	0.0005
	5	0.7	0.014	0.005	0.0026	0.0013	0.0005
	5	−0.1	0.014	0.005	0.0025	0.0013	0.0005
	10	0	0.031	0.011	0.0052	0.0026	0.0010
	10	0.7	0.034	0.011	0.0052	0.0026	0.0010
	10	−0.1	0.030	0.010	0.0052	0.0026	0.0010
$0.5\ p/n$	5	-	0.013	0.005	0.0025	0.0013	0.0005
	10	-	0.025	0.010	0.0050	0.0025	0.0010

Taulukko 1: Estimoidut KL-informaation odotusarvot.

3 Sakotettu logistinen regressio

Sakotetulla logistisella regressiolla tarkoitetaan tässä menetelmää, jossa todellisen parametrivektorin $\boldsymbol{\beta}_0$ estimaatti saadaan suurimman uskottavuuden menetelmään liittyvän log-uskottavuusfunktion $l(\boldsymbol{\beta})$ sijaan funktion $l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - p(\boldsymbol{\beta})$ maksimikohtana $\hat{\boldsymbol{\beta}}^*$. Käytännössä parametrivektoriin $\boldsymbol{\beta}$ kohdistuvan sakon $p(\boldsymbol{\beta})$ tarkoituksena on vähentää suurimman uskottavuuden estimaattorin $\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X})$ satunnaisuuteen liittyvää vaihtelua ilman siihen kohdistuvaa merkittävää harhaa.

Pohditaan aluksi, mitä odotetusta ennustevirheestä voidaan sanoa informaatioepäyhtälön avulla harhattomien ja asymptoottisesti harhattomien estimaattorien osalta. Olkoon $\boldsymbol{\delta}_u$ hypoteettinen harhaton estimaattori, jonka $D_{\boldsymbol{\beta}}\mathbb{E}(\boldsymbol{\delta}_u) \in \mathbb{R}^{p\times p}$ on olemassa ja saadaan osittaisderivoimalla komponenteittain odotusarvon sisällä. Kun $\boldsymbol{\delta}_u^*$ on mikä tahansa muu tällainen harhaton estimaattori, jonka kovarianssimatriisi on olemassa, ja $\mathrm{Cov}(\sqrt{n}\boldsymbol{\delta}_u)^{-1} = \mathcal{I}(\boldsymbol{\beta}_0)$ Cholesky-hajotelmanaan $\mathbf{LL}'$, saadaan

$$\begin{aligned}\mathbb{E}(D_{KL}\big(\boldsymbol{\beta}_0 \| \boldsymbol{\delta}_u^*\big)) - \mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 \| \boldsymbol{\delta}_u)\big) &\approx \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\boldsymbol{\delta}_u^*) - \mathrm{Cov}(\boldsymbol{\delta}_u)\big)\big) \\ &= \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{A}\big) = \tfrac{1}{2}\mathrm{tr}(\mathbf{L}'\mathbf{A}\mathbf{L}),\end{aligned}$$

jossa $\mathbf{A} = \mathrm{Cov}(\boldsymbol{\delta}_u^*) - \mathrm{Cov}(\boldsymbol{\delta}_u)$. Koska matriisin jälki on summa sen ominaisarvoista ja koska $\mathbf{a}'\mathbf{L}'\mathbf{A}\mathbf{L}\mathbf{a} \geq 0$ kaikilla $\mathbf{a} \in \mathbb{R}^p$ informaatioepäyhtälöstä saatavan $\mathbf{A} \geq \mathbf{0}$

perusteella, nähdään, että tämä KL-informaatioon perustuva ennustevirheiden odotusarvojen approksimaatioiden erotus ei voi olla nollaa pienempi.

Suurimman uskottavuuden estimaattorin $\hat{\boldsymbol{\beta}}$ tarkentuvuuden ja asymptoottisen tehokkuuden perusteella vastaavaa tulosta estimaattorien asymptoottisiin kovarianssimatriiseihin sovellettaessa nähdään lisäksi, että mikään muu asymptoottisesti normaali ja harhaton estimaattori ei voi alittaa sen approksimatiivista odotettua ennustevirhettä $0.5p/n$ otoksissa, joissa asymptotiikan voidaan olettaa toimivan [6]. Käytännössä sakosta $p(\boldsymbol{\beta})$ voi siis olla merkittävää hyötyä vain, jos se aiheuttaa harhaa tai tuottaa estimaattorin, johon informaatioepäyhtälöä ei voida soveltaa.

Tarkastellaan ensin harhan merkitystä hyödyntämällä kääntyvän matriisin $\mathcal{I}(\boldsymbol{\beta}_0)$ ominaisarvohajotelmaa $\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}$. Kun $\hat{\boldsymbol{\gamma}}^* = \mathbf{Q}\hat{\boldsymbol{\beta}}^*$ on sakotetun estimaattorin esitys matriisin $\mathbf{Q}$ sisältämien ominaisvektorien määräämässä kannassa, saadaan odotetun ennustevirheen approksimaatioksi edellisen luvun perusteella

$$\begin{aligned}
\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}}^*)\big) &\approx \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) + b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)\big) \\
&= \tfrac{1}{2}\mathrm{tr}\big(\boldsymbol{\Lambda}\big(\mathrm{Cov}(\mathbf{Q}\hat{\boldsymbol{\beta}}^*) + b(\mathbf{Q}\hat{\boldsymbol{\beta}}^*)b(\mathbf{Q}\hat{\boldsymbol{\beta}}^*)'\big)\big) \\
&= \frac{1}{2}\sum_{j=1}^{p}\Lambda_{jj}\big(\mathrm{Var}(\hat{\gamma}_j^*) + b(\hat{\gamma}_j^*)^2\big),
\end{aligned}$$

josta Hastien et al. [13, luku 7] mainitsemaan harhan ja varianssin väliseen kompromissiin liittyvä yhteys muunnetun estimaattorin osalta voidaan todeta. Koska $\mathcal{I}(\boldsymbol{\beta}_0)$:n positiivisen definiittisyyden perusteella $\Lambda_{jj} > 0$ kaikilla j, voi jonkin ominaisvektorin suuntaisesta harhasta $b(\hat{\gamma}_k^*)$ olla yksinään hyötyä edellä vain, jos se vähentää ominaisvektorien suuntaisten varianssien ominaisarvoilla painotettua summaa enemmän kuin $\Lambda_{kk}b(\hat{\gamma}_k^*)^2$. Harhan myötä approksimaation virhe $O(\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\|^3)$ tosin kasvaa.

Suurimman uskottavuuden estimaattorin asymptoottisen jakauman osalta nähdään lisäksi, että se hyötyisi edellä harhasta aina, kun jokin kaikkia muita parametreja kohtaan ortogonaalinen β_k on itseisarvoltaan pieni. Tällöin vektori $\mathbf{e}_k$, jonka k:s komponentti on yksi muiden ollessa nollia, on selvästikin matriisin $\mathcal{I}(\boldsymbol{\beta}_0)$ ominaisvektori $\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{e}_k = \mathcal{I}(\boldsymbol{\beta}_0)_{kk}\mathbf{e}_k$, ja siten $\hat{\gamma}_j^* = \mathbf{e}_k'\hat{\boldsymbol{\beta}} = \hat{\beta}_k$ jollain j. Koska lohkodiagonaalisen matriisin käänteismatriisi on myös lohkodiagonaalinen, vaikuttaa parametrin β_k estimoimatta jättäminen vain matriisin $\mathbf{Q}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{Q}'$ alkioon $\mathbf{e}_k'\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{e}_k \approx n\mathrm{Var}(\hat{\beta}_k)$. Periaatteessa odotetun ennustevirheen approksimaatiosta saataisiin siten pienempi vaikka olettamalla, että $\beta_k = 0$, kun $\mathrm{Var}(\hat{\gamma}_j^*) = \mathrm{Var}(\hat{\beta}_k) > \beta_{0_k}^2$.

Itse asiassa, jos keskitytään parametriavaruuden osajoukkoon, jonka mitta on nolla, ja oletetaan, että parametrivektori on harva ja järjestetty yhdessä selittävien muuttujien kanssa siten, että sen viimeiset $p - k \in \{1, ..., p-1\}$ komponenttia ovat nollia, on selvää, että suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\beta}}_{(k)}$, joka estimoi vain ensimmäiset k parametria, voi suurissa otoksissa saavuttaa likimain $0.5(p-k)/n$ parannuksen odotettuun ennustevirheeseen [7]. Tämä on helppoa todeta huomaamalla, että tällöin mallin, jossa $\boldsymbol{\beta}_0 \in \mathbb{R}^k$, Fisherin informaatiomatriisi saadaan ottamalla lohko $\mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{11}} \in \mathbb{R}^{k\times k}$ matriisin $\mathcal{I}(\boldsymbol{\beta}_0) = \mathbb{E}\big(\pi(\mathbf{X}_1'\boldsymbol{\beta}_0)\big(1 - \pi(\mathbf{X}_1'\boldsymbol{\beta}_0)\big)\mathbf{X}_1\mathbf{X}_1'\big) \in \mathbb{R}^{p\times p}$ vasemmasta yläkulmasta siinä esiintyvää vektoria $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ koskevan havainnon

[6] Luvun 2.3 mukaisesti tämä pätee parametriavaruuden nollamittaista osaa lukuun ottamatta.

[7] Tässä on syytä huomioida luvussa 2.4 odotusarvon approksimaatiota koskevat huomiot.

$\mathbf{x}_i'\boldsymbol{\beta}_0 = \sum_{j=1}^{p} x_{ij}\beta_{0_j} = \sum_{j=1}^{k} x_{ij}\beta_{0_j}$ perusteella, johtaen approksimaatioon

$$\begin{aligned}
\mathbb{E}(D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}}_{(k)})) &\approx \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)}) + b(\hat{\boldsymbol{\beta}}_{(k)})b(\hat{\boldsymbol{\beta}}_{(k)})'\big)\big) \\
&\approx \tfrac{1}{2}\mathrm{tr}\left(\begin{bmatrix} \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{11}} & \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{12}} \\ \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{21}} & \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{22}} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{11}} & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{12}} \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{21}} & \mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{22}} \end{bmatrix}\right) \\
&= \tfrac{1}{2}\mathrm{tr}\left(\begin{bmatrix} \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{11}}\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{11}} & \mathbf{0} \\ \mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{21}}\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{11}} & \mathbf{0} \end{bmatrix}\right) \\
&= \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)_{\mathbf{11}}\mathrm{Cov}(\hat{\boldsymbol{\beta}}_{(k)})_{\mathbf{11}}\big),
\end{aligned}$$

josta haluttu tulos, jota estimoituihin parametreihin kohdistuva harha mahdollisesti vielä parantaisi, voidaan todeta vastaavasti kuin luvussa 2.4. Sakotettujen estimaattorien osalta tämä tulos vaatisi sitä, että niiden täytyisi pystyä valitsemaan aidosti selittävät muuttujat valitun kokoisissa opetusaineistoissa todennäköisyydellä yksi.

3.1 Paras osajoukko ja askeltavat menetelmät

Ehkä yksinkertaisin tapa vähentää estimointiin liittyvää varianssia, on sisällyttää estimoitavaan malliin vain osa selittävistä muuttujista. Esimerkiksi parhaan osajoukon menetelmässä on Hastien et al. [13, luku 3] mukaan tarkoitus löytää funktion $l(\boldsymbol{\beta}_{(k)})$ maksimoiva vakion ja $k-1$ selittävän muuttujan osajoukko jokaisella $k = 1, ..., p$. Koska opetusaineistossa $l(\hat{\boldsymbol{\beta}}_{(k)}) \leq l(\hat{\boldsymbol{\beta}}_{(k+1)})$, ei näin löydettyjä malleja voida kuitenkaan suoraan verrata keskenään, etenkin, kun muistetaan, että mallin, joka sisältää ainakin aidosti selittävät muuttujat $\{X_{ij} : \beta_j \neq 0\}$, odotetun ennustevirheen approksimaatio on suoraan verrannollinen estimoitujen parametrien määrään.

Ennustevirheen kannalta eri määrän selittäviä muuttujia sisältävistä malleista paras on luonnollisestikin se, jonka mukainen $D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}}_{(k)})$ [8] on pienin. Vaikka tähän tarvittava $\mathbb{E}\big(l(\hat{\boldsymbol{\beta}}_{(k)}; Y_0|\mathbf{X}_0')\big)$ periaatteessa kuuluisikin estimoida riippumattomassa testiaineistossa, voidaan sitä arvioida Konishin ja Kitagawan [16, luku 3] mukaan myös pelkän opetusaineiston avulla Takeuchin informaatiokriteeriin (TIC) perustuen. Käytännössä olennaista tässä on huomata, että vahvaa suurten lukujen lakia opetusaineistoon sovellettaessa syntyvä harha, kun $\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}) \in \mathbb{R}^k$ estimoi parametrivektoria $\boldsymbol{\beta}_0^*$, voidaan jakaa kolmeen eri osaan

$$\begin{aligned}
b_k &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} l\big(\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}); Y_i|\mathbf{X}_i'\big)\right) - \mathbb{E}\big(l\big(\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}); Y_0|\mathbf{X}_0'\big)\big) \\
&= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} l\big(\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}); Y_i|\mathbf{X}_i'\big)\right) - \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\beta}_0^*; Y_i|\mathbf{X}_i')\right) \\
&\quad + \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{\beta}_0^*; Y_i|\mathbf{X}_i')\right) - \mathbb{E}\big(l(\boldsymbol{\beta}_0^*; Y_0|\mathbf{X}_0')\big) \\
&\quad + \mathbb{E}\big(l(\boldsymbol{\beta}_0^*; Y_0|\mathbf{X}_0')\big) - \mathbb{E}\big(l\big(\hat{\boldsymbol{\beta}}(\mathbf{Y}, \mathbf{X}); Y_0|\mathbf{X}_0'\big)\big) \\
&= D_1 + D_2 + D_3,
\end{aligned}$$

[8] Tässä ja jatkossa voidaan olettaa, että vektoriin $\boldsymbol{\beta}_{(k)} \in \mathbb{R}^k$ on lisätty $p-k$ nolla-arvoista komponenttia siitä puuttuvien komponenttien tilalle, kun sitä verrataan vektoriin $\boldsymbol{\beta}_0 \in \mathbb{R}^p$.

joista viimeisimmän approksimaatioksi johdettiin jo aikaisemmin $D_3 \approx 0.5k/n$, kun $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}_0$, ja $D_3 \approx 0.5\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0^*)\mathrm{Cov}(\hat{\boldsymbol{\beta}})\big) \approx \frac{1}{2n}\mathrm{tr}\big(\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big)$ muuten. Koska D_2 on nolla havaintojen samoinjakautuneisuuden perusteella, voi harhaa tulla lisää vain termistä D_1. Sen osalta on hyvä huomata, että log-uskottavuusfunktio voidaan nyt esittää lähellä pistettä $\hat{\boldsymbol{\beta}}$ Taylorin polynomina

$$\sum_{i=1}^{n} l(\boldsymbol{\beta}_0^*; y_i|\mathbf{x}_i') = l(\boldsymbol{\beta}_0^*) \approx l(\hat{\boldsymbol{\beta}}) + \nabla l(\hat{\boldsymbol{\beta}})'(\boldsymbol{\beta}_0^* - \hat{\boldsymbol{\beta}}) + \tfrac{1}{2}(\boldsymbol{\beta}_0^* - \hat{\boldsymbol{\beta}})'\mathbf{H}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}_0^* - \hat{\boldsymbol{\beta}}),$$

jossa $\nabla l(\hat{\boldsymbol{\beta}}) = \mathbf{0}$, ja josta siten saadaan ratkaistua approksimaatio

$$D_1 = \tfrac{1}{n}\mathbb{E}\big(l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}_0^*)\big) \approx -\tfrac{1}{2n}\mathbb{E}\big((\boldsymbol{\beta}_0^* - \hat{\boldsymbol{\beta}})'\mathbf{H}(\hat{\boldsymbol{\beta}})(\boldsymbol{\beta}_0^* - \hat{\boldsymbol{\beta}})\big).$$

Koska odotusarvon sisällä $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*)'\big(-\frac{1}{n}\mathbf{H}(\hat{\boldsymbol{\beta}})\big)\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^*) \xrightarrow{d} \mathbf{Z}'\mathcal{I}(\boldsymbol{\beta}_0^*)\mathbf{Z}$, jossa $\mathbf{Z} \sim \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big)$, Slutskyn ja jatkuvan kuvauksen lauseen mukaan, saadaan tasainen integroitavuus oletettaessa D_1:n approksimaatioksi edelleen

$$D_1 \approx \tfrac{1}{2n}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0^*)\mathbb{E}(\mathbf{Z}\mathbf{Z}')\big) = \tfrac{1}{2n}\mathrm{tr}\big(\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big) \approx D_3.$$

Kun opetusaineiston log-uskottavuusfunktion maksimista $l(\hat{\boldsymbol{\beta}})$ vähennetään sen kokoa vastaavan harhan $nb_k = nD_1 + nD_2 + nD_3$ arvio $\mathrm{tr}\big(\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big)$, saadaan siitä aikaisemmin tehtyjen oletusten perusteella KL-informaatioon perustuvan ennustevirheen kannalta oleellisen odotusarvon $n\mathbb{E}\big(l(\hat{\boldsymbol{\beta}}_{(k)}; Y_0|\mathbf{X}_0')\big)$ asymptoottisesti harhaton estimaattori. Jos lisäksi oletetaan, että $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}_0$ kaikilla k, on arvio aina Akaiken informaatiokriteerin (AIC) mukainen $b_k \approx k/n$. Vaikka tämä on selvästikin väärin, kun oikeassa mallissa on vakion lisäksi muitakin selittäviä muuttujia, esitän seuraavassa perustelun, miksi AIC:n mukainen b_k on silti käyttökelpoinen.

Merkitään aluksi, että estimoitavassa mallissa vektorit $\mathbf{X}_{(k)_i} \in \mathbb{R}^k$ on saatu poistamalla osa parametrivektorin $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ mukaisten satunnaisvektorien $\mathbf{X}_i \in \mathbb{R}^p$ komponenteista. Koska luvussa 2.2 todettiin, että malli on olennaisesti oikea, kun

$$\mathrm{P}(\mathbf{X}_1 \in \{\mathbf{x}_1 \in \mathbb{R}^p : \mathbb{E}(Y_1|\mathbf{X}_1 = \mathbf{x}_1) \neq \pi(\mathbf{x}_{(k)_1}'\boldsymbol{\beta}_0^*)\}) = 0,$$

myös siinä mielessä, että $nb_k \approx \mathrm{tr}\big(\mathcal{J}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big) = \mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0^*)\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big) = k$, on tässä hyödyllistä huomata, että ehdollisessa odotusarvossa

$$\begin{aligned}
\mathcal{J}(\boldsymbol{\beta}_0^*)_{\mathbf{X}_1=\mathbf{x}_1} &= \mathbb{E}\big(\big(Y_1 - \pi(\mathbf{X}_{(k)_1}'\boldsymbol{\beta}_0^*)\big)^2\mathbf{X}_{(k)_1}\mathbf{X}_{(k)_1}'|\mathbf{X}_1 = \mathbf{x}_1\big) \\
&= \mathbb{E}\big(\big(Y_1\big(1 - 2\pi(\mathbf{X}_{(k)_1}'\boldsymbol{\beta}_0^*)\big) + \pi(\mathbf{X}_{(k)_1}'\boldsymbol{\beta}_0^*)^2\big)\mathbf{X}_{(k)_1}\mathbf{X}_{(k)_1}'|\mathbf{X}_1 = \mathbf{x}_1\big) \\
&= \big(\mathbb{E}(Y_1|\mathbf{X}_1 = \mathbf{x}_1)\big(1 - 2\pi(\mathbf{x}_{(k)_1}'\boldsymbol{\beta}_0^*)\big) + \pi(\mathbf{x}_{(k)_1}'\boldsymbol{\beta}_0^*)^2\big)\mathbf{x}_{(k)_1}\mathbf{x}_{(k)_1}'
\end{aligned}$$

esiintyvä $\mathcal{J}(\boldsymbol{\beta}_0^*)_{\mathbf{X}_1=\mathbf{x}_1}$ on ehdollisesta odotusarvosta $\mathbb{E}(Y_1|\mathbf{X}_1 = \mathbf{x}_1) = \pi(\mathbf{x}_1'\boldsymbol{\beta}_0)$ löytyvän lineaarikombinaation $\mathbf{x}_1'\boldsymbol{\beta}_0$ jatkuva funktio. Näin ollen $\mathcal{J}(\boldsymbol{\beta}_0^*)_{\mathbf{X}_1=\mathbf{x}_1}$ saadaan mielivaltaisen lähelle matriisia $\mathcal{I}(\boldsymbol{\beta}_0^*)_{\mathbf{X}_1=\mathbf{x}_1}$ valitsemalla $\mathbf{x}_1'\boldsymbol{\beta}_0$ tarpeeksi läheltä pistettä $\mathbf{x}_{(k)_1}'\boldsymbol{\beta}_0^*$. Jos näin tehdään kaikille paitsi mahdollisesti nollamittaiseen joukkoon kuuluville $\mathbf{x}_i$, saadaan harhan nb_k approksimaatio mielivaltaisen lähelle arvoa k kuvauksen $\mathrm{tr}\big(\mathcal{J}(\boldsymbol{\beta}_0^*)_{\mathbf{X}_1=\mathbf{x}_1}\mathcal{I}(\boldsymbol{\beta}_0^*)^{-1}\big)$ jatkuvuuden perusteella.

Tarkastellaan log-uskottavuusfunktiota seuraavaksi vektorin $\mathbf{X}\boldsymbol{\beta} \in [-\eta, \eta]^n$ funktiona [9], ja pohditaan, miten epäyhtälö $l(\mathbf{X}\hat{\boldsymbol{\beta}}) - l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}) \leq c \in \mathbb{R}_{\geq 0}$ vaikuttaa etäisyyteen $d_n = \|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}\|$. Koska nyt $-\nabla^2 l(\mathbf{X}\boldsymbol{\beta})_{ii} \geq \pi(\eta)\big(1 - \pi(\eta)\big) = m > 0$ kaikilla i, kun $\eta > 0$, nähdään saadusta vahvasta konkaavisuudesta, että

$$\begin{aligned} l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}) &\leq l(\mathbf{X}\hat{\boldsymbol{\beta}}) + \nabla l(\mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)} - \mathbf{X}\hat{\boldsymbol{\beta}}) - \tfrac{md_n^2}{2} \\ \Leftrightarrow d_n^2 &\leq \tfrac{2}{m}\big(l(\mathbf{X}\hat{\boldsymbol{\beta}}) - l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)})\big), \end{aligned}$$

sillä $\nabla l(\mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{X}\boldsymbol{\beta} = \nabla l(\hat{\boldsymbol{\beta}})'\boldsymbol{\beta} = 0$ kaikilla $\boldsymbol{\beta} \in \mathbb{R}^p$. Käytännössä alkuperäinen epäyhtälö $l(\mathbf{X}\hat{\boldsymbol{\beta}}) - l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}) \leq c$ ei siis voi olla voimassa kiinteällä c otoskoon kasvaessa rajatta, ellei $d_n^2 \leq \frac{2c}{m} < \infty$. Toisin sanoen, aineiston kokoa kasvattamalla joko lähes kaikki $(\mathbf{x}'_{(k)_i}\boldsymbol{\beta}_0^* - \mathbf{x}_i'\boldsymbol{\beta}_0 + o_p(1))^2$ saadaan mielivaltaisen lähelle nollaa etäisyyden d_n rajoittuneisuuden ansiosta, kuten harhan nb_k arvion k perustelun yhteydessä tarvitaan, tai erotus $l(\mathbf{X}\hat{\boldsymbol{\beta}}) - l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)})$ kasvaa suuremmaksi kuin c.

Vaikka esitetty perustelu onkin puutteellinen, sillä se ei ota kantaa siihen, mitä lähellä tarkalleen ottaen eri yhteyksissä tarkoittaa, on kuitenkin selvää, että jos $l(\mathbf{X}\hat{\boldsymbol{\beta}}) - l(\mathbf{X}_{(k)}\hat{\boldsymbol{\beta}}_{(k)}) > p - k$ ja $b_k > k/n$, ei todellinen b_k ole mallin valinnan kannalta merkityksellinen. Toisaalta jos d_n on edellä pieni, koska vain $\boldsymbol{\beta}_0$:n mukaisesta mallista löytyvien selittävien muuttujien indeksien J mukaiset $\sum_{j\in J} x_{ij}\hat{\beta}_j$ ovat selitettävissä $\boldsymbol{\beta}_0^*$:n mukaisesta mallista löytyvien selittävien muuttujien lineaarikombinaatioiden avulla riittävän hyvin, voi $\boldsymbol{\beta}_0^*$:n mukaisen mallin valinta olla joka tapauksessa perusteltua myös harhan ja varianssin välisen kompromissin perusteella.

Koska edellä todettiin, että eri määrän selittäviä muuttujia sisältävien mallien vertailu voidaan perustaa harhakorjatulla otoskeskiarvolla $\frac{1}{n}\big(l(\hat{\boldsymbol{\beta}}_{(k)}) - k\big)$ estimoituun odotusarvoon $\mathbb{E}\big(l(\hat{\boldsymbol{\beta}}_{(k)}; Y_0|\mathbf{X}_0')\big)$, voidaan AIC:iin perustuva parhaan osajoukon menetelmä esittää myös Hastien et al. [13, luku 3] mukaisesti optimointitehtävänä

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \big(y_i\mathbf{x}_i'\boldsymbol{\beta} - \log\big(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\big)\big) - \lambda\|\boldsymbol{\beta}\|_0 \right\}, \lambda = 1,$$

jossa $\boldsymbol{\beta}$-vektorin L0-normi $nb_{\|\boldsymbol{\beta}\|_0} \approx \|\boldsymbol{\beta}\|_0 = p(\boldsymbol{\beta})$ korjaa edellä johdettua approksimatiivista asymptoottista harhaa toimien samalla yleisen kehikon mukaisena sakkona. Kun $i(\boldsymbol{\beta})$ palauttaa vektorin $\boldsymbol{\beta}$ nollasta poikkeavien komponenttien indeksit ja $M = \{i(\boldsymbol{\beta}) : \beta_1 \neq 0\}$ vertailtavien mallien mukaan, voidaan sakotetun estimaattorin $\hat{\boldsymbol{\beta}}^*$ kertymäfunktio esittää eri estimaattorien ehdollisten kertymäfunktioiden valintatodennäköisyyksillä painotettuna summana

$$F(\boldsymbol{\beta}) = \sum_{m\in M} F\big(\boldsymbol{\beta}|i(\hat{\boldsymbol{\beta}}^*) = m\big)\mathrm{P}\big(i(\hat{\boldsymbol{\beta}}^*) = m\big)$$

kokonaistodennäköisyyden kaavaa hyödyntämällä. Huomionarvoista tässä on, että yleisesti $F\big(\boldsymbol{\beta}|i(\hat{\boldsymbol{\beta}}^*) = m\big)$ ei vastaa minkään tietyn indeksijoukon m mukaisen estimaattorin ehdotonta kertymäfunktiota, sillä tapahtuma $\{i(\hat{\boldsymbol{\beta}}^*) = m\}$ valita tietyn m mukainen malli AIC:n perusteella riippuu tyypillisesti myös tarkastellun estimaattorin jakaumasta. Kun lisäksi huomioidaan, että todennäköisyys valita AIC:llä

[9] Vastaavasti kuin luvussa 2.1 nähdään, että $\nabla l(\boldsymbol{\eta}) = \big(y_1 - \pi(\eta_1), ..., y_n - \pi(\eta_n)\big) = \big(\mathbf{y} - \pi(\boldsymbol{\eta})\big)$ ja että $\nabla^2 l(\boldsymbol{\eta}) = -\mathrm{diag}\big[\pi(\eta_1)\big(1 - \pi(\eta_1)\big) \cdots \pi(\eta_n)\big(1 - \pi(\eta_n)\big)\big]$.

$\hat{\boldsymbol{\beta}}_{(k)} \in \{\boldsymbol{\beta} \in \mathbb{R}^k : i(\boldsymbol{\beta}) = i(\boldsymbol{\beta}_0), \forall \beta_j \neq 0\}$ sijaan $\hat{\boldsymbol{\beta}}_{(l)} \in \{\boldsymbol{\beta} \in \mathbb{R}^{k+d} : i(\boldsymbol{\beta}) = i(\boldsymbol{\beta}_0)\}$ on säännöllisissä malleissa, kun $d > 0$, uskottavuusosamäärän testin perusteella asymptoottisesti $\mathrm{P}(\chi^2_d > 2d)$ [10], nähdään myös, että mitään tiettyä mallia ei valita todennäköisyydellä yksi edes aineiston kasvaessa rajatta.

Jotta edellä kuvatun estimaattorin $\hat{\boldsymbol{\beta}}^*$ odotetusta ennustevirheestä saisi paremman kuvan, on taulukossa 2 esitetty Monte Carlo -estimaatit sekä mallien valintatodennäköisyyksistä että niiden mukaisten KL-informaatioiden odotusarvoista, kun aineiston määränneessä logistisessa regressiomallissa parametrivektoria $\boldsymbol{\beta}_0 = (1, 0.5, 0)$ vastaavista selittävistä muuttujista X_{11} on vakio muiden noudattaessa multinormaalijakaumaa, jossa $\mathrm{Cov}(X_{1i}, X_{1j}) = 1$, kun $i = j$ ja muuten 0.7. Vaikka $\hat{\boldsymbol{\beta}}^*$ vaikuttaakin tässä selvästi huonommalta valinnalta kuin $\hat{\boldsymbol{\beta}}_{(3)}$, kun $m = \{1, 2, 3\}$, on syytä muistaa, että niiden mukaiset odotetut ennustevirheet ovat identtiset rajoituttaessa opetusaineistoihin, joissa $i(\hat{\boldsymbol{\beta}}^*) = \{1, 2, 3\}$.

	m	n=200	n=500	n=1000	n=2000	n=5000
$\mathrm{P}\big(i(\hat{\boldsymbol{\beta}}^*)=m\big)$	1	0.04	0.00	0.00	0.00	0.00
	1,2	0.76	0.85	0.84	0.84	0.85
	1,3	0.10	0.01	0.00	0.00	0.00
	1,2,3	0.10	0.14	0.16	0.16	0.15
$\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 \| \hat{\boldsymbol{\beta}}^*) \| i(\hat{\boldsymbol{\beta}}^*)=m\big)$	1	0.03	0	-	-	-
	1,2	0.005	0.002	0.0010	0.0005	0.0002
	1,3	0.018	0.01	0.0123	-	-
	1,2,3	0.015	0.006	0.0028	0.0014	0.0006
$\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 \| \hat{\boldsymbol{\beta}}^*)\big)$	-	0.008	0.003	0.0013	0.0006	0.0003
$\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 \| \hat{\boldsymbol{\beta}}_{(3)})\big)$	-	0.008	0.003	0.0015	0.0007	0.0003

Taulukko 2: Estimoidut KL-informaation odotusarvot.

Mahdollisten selittävien muuttujien määrän p kasvaessa parhaan osajoukon menetelmä kärsii muun muassa Bertsimasin ja Kingin [4] esittämistä algoritmisista parannuksista huolimatta siitä, että periaatteessa mahdollisia vakion sisältäviä malleja on kaikkiaan 2^{p-1}. Laskennallisesti kevyempänä vaihtoehtona sille voidaan käyttää esimerkiksi Hastien et al. [13, luku 3] mainitsemia, tosin vain hakupolulle osuvan lokaalin maksimin löytäviä askeltavia menetelmiä:

- Eteenpäin askellettaessa jokaisella $k = 2, ..., \min\{\|\hat{\boldsymbol{\beta}}^*\|_0 + 1, p\}$ askeleella valitaan paras malli, jossa edellisellä askeleella saatuun $k-2$ selittävää muuttujaa ja vakion sisältäneeseen malliin on lisätty yksi uusi selittävä muuttuja.

- Taaksepäin askellettaessa jokaisella $k = p-1, ..., \max\{\|\hat{\boldsymbol{\beta}}^*\|_0 - 1, 1\}$ askeleella valitaan paras vakion sisältävä malli, joka saadaan poistamalla jokin edellisellä askeleella käytetyistä selittävistä muuttujista.

Pahimmassakin tapauksessa askeltavissa menetelmissä vertailtavia malleja on siten vain $1+(p-1)p/2$. Vaikka näitä menetelmiä voidaan vielä laajentaa sallimalla askelia molempiin suuntiin, ei näitä laajennoksia käsitellä tässä.

[10] Uskottavuusosamäärän testin mukaan $2\big(l(\hat{\boldsymbol{\beta}}_{(l)}) - l(\hat{\boldsymbol{\beta}}_{(k)})\big) \underset{as}{\sim} \chi^2_d$ jo näillä kahdella mallilla.

3.2 Logistinen harjuregressio

Toinen yksinkertainen tapa vähentää estimointiin liittyvää varianssia on perustaa sakko menetelmäparametrilla $\lambda > 0$ säädeltävään parametrivektorin $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2)$ osavektorin neliöityyn euklidiseen normiin $p(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}_2\|^2 = \lambda\sum_{j=2}^{p}\beta_j^2$. Koska tämä Hastien et al. [13, luku 3] mukainen sakko logistisessa harjuregressiossa on sitä suurempi, mitä suurempia termit β_j^2 ovat, vaikuttavat selittävien muuttujien mittayksiköt siihen, miten voimakkaasti mitäkin parametria sakotetaan. Tämä on varsin ilmeistä, sillä malli $\text{logit}(\pi_i) = \beta_1 + (c_1\beta_2)(x_{i2}/c_1) + ... + (c_{p-1}\beta_p)(x_{ip}/c_{p-1})$, josta estimoidaan parametrit $\beta_1, c_1\beta_2, ..., c_{p-1}\beta_p$, on olennaisesti sama kaikilla $c_1, ..., c_{p-1} > 0$.

Käytännössä eri mittayksiköiden aiheuttamalta ongelmalta voidaan välttyä kohdistamalla sakko vektorin $\boldsymbol{\beta}_2$ sijaan vektoriin $\mathbf{C}_2\boldsymbol{\beta}_2 = \text{diag}[s_{x_2} \cdots s_{x_p}]\boldsymbol{\beta}_2$, joka on saatu valitsemalla $c_j = s_{x_{j+1}}$ kaikilla $j = 1, .., p-1$ otoskeskihajontoja hyödyntäen. Tällöin logistinen harjuregressio, jossa luvun 3 mukaisesti $l^*(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - p(\boldsymbol{\beta})$, voidaan esittää menetelmäparametrista $\lambda > 0$ riippuvana optimointitehtävänä

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}}{\text{argmax}} \left\{ \sum_{i=1}^{n} \big(y_i \mathbf{x}_i'\boldsymbol{\beta} - \log\big(1 + \exp(\mathbf{x}_i'\boldsymbol{\beta})\big)\big) - \lambda\|\mathbf{C}_2\boldsymbol{\beta}_2\|^2 \right\},$$

jossa vakioon β_1 ei kohdistu sakkoa[11]. Koska maksimoitavan funktion $l^*(\boldsymbol{\beta})$ gradienttivektoriksi saadaan tässä $\nabla l(\boldsymbol{\beta}) - (0, 2\lambda\mathbf{C}_2'\mathbf{C}_2\boldsymbol{\beta}_2)$, nähdään, että se on aidosti konkaavi, sillä $\mathbf{a}'\nabla^2 l(\boldsymbol{\beta})\mathbf{a} - \mathbf{a}'\nabla^2 p(\boldsymbol{\beta})\mathbf{a} < 0$ kaikilla $\mathbf{a} \in \mathbb{R}^p \setminus \mathbf{0}$ sakkoon liittyvän matriisin $-\nabla^2 p(\boldsymbol{\beta}) = -2\lambda\text{diag}[0\ s_{x_2}^2 \cdots s_{x_p}^2]$ negatiivisen semidefiniittisyyden perusteella.

Logistinen harjuregressio voitaisiin siis tässä ratkaista Newtonin menetelmällä, ja sen tuottaman jonon voitaisiin vastaavasti kuin luvussa 2.1 osoittaa suppenevan neliöllisesti kohti arvoa $\hat{\boldsymbol{\beta}}^*$ funktion $\nabla^2 l^*(\boldsymbol{\beta})$ Lipschitz-jatkuvuuden, joka seuraa siitä, että $\forall\boldsymbol{\beta}, \boldsymbol{\beta}^* \in \mathbb{R}^p, \|\nabla^2 l^*(\boldsymbol{\beta}) - \nabla^2 l^*(\boldsymbol{\beta}^*)\| = \|\nabla^2 l(\boldsymbol{\beta}) - \nabla^2 l(\boldsymbol{\beta}^*)\|$, perusteella. Tässä työssä logistisen harjuregression mukaisen mallin sovittamiseen käytetään kuitenkin seuraavassa luvussa tarkemmin kuvattavaa coordinate descent -menetelmää hyödyntävää R-ohjelmiston glmnet-lisäpakettia [11] [25] sen helppokäyttöisyyden vuoksi.

3.2.1 Asymptoottinen jakauma ja odotettu ennustevirhe

Pidetään jälleen edellä kuvattua estimaattoria sovittamiseen käytetystä aineistosta riippuvana satunnaismuuttujana $\hat{\boldsymbol{\beta}}^*(\mathbf{Y}, \mathbf{X})$, ja johdetaan sen asymptoottinen jakauma Cessien ja Houwelingenin [5] mukaisesti. Todetaan aluksi, että gradienttivektori $\nabla l^*(\hat{\boldsymbol{\beta}}^*)$ voidaan esittää lähellä pistettä $\boldsymbol{\beta}_0$ Taylorin polynomina

$$\mathbf{0} = \nabla l^*(\hat{\boldsymbol{\beta}}^*) \approx \nabla l(\boldsymbol{\beta}_0) - \nabla p(\boldsymbol{\beta}_0) + \big(\nabla^2 l(\boldsymbol{\beta}_0) - \nabla^2 p(\boldsymbol{\beta}_0)\big)(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0),$$

jonka virhe on $O(\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\|^2)$ ja josta saadaan ratkaistua approksimaatio

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &\approx \boldsymbol{\beta}_0 + \big(\nabla^2 l(\boldsymbol{\beta}_0) - \nabla^2 p(\boldsymbol{\beta}_0)\big)^{-1}\big(\nabla p(\boldsymbol{\beta}_0) - \nabla l(\boldsymbol{\beta}_0)\big) \\ &= \big(\nabla^2 l(\boldsymbol{\beta}_0) - \nabla^2 p(\boldsymbol{\beta}_0)\big)^{-1}\big(\nabla^2 l(\boldsymbol{\beta}_0)\boldsymbol{\beta}_0 - \nabla l(\boldsymbol{\beta}_0)\big)\end{aligned}$$

[11] Jos sakon halutaan kohdistuvan myös vakioon, voidaan se kohdistaa koko vektoriin $\boldsymbol{\beta}$. On kuitenkin syytä huomata, että tällöin vakiota vastaavaa selittävää muuttujaa ei saada samalle asteikolle muiden kanssa, sillä $s_{\mathbf{x}_1} = 0$. Lisäksi luvun 2.1 mukainen yhtälö $\frac{\partial}{\partial\beta_1} l(\boldsymbol{\beta}) = 0$ ei sakon vuoksi enää takaa, että $\bar{y} = \frac{1}{n}\sum_{i=1}^{n}\pi(\mathbf{x}_i'\hat{\boldsymbol{\beta}}^*)$ kuten aikaisemmin.

havainnon $\nabla p(\boldsymbol{\beta}) = (0, 2\lambda \mathbf{C}_2' \mathbf{C}_2 \boldsymbol{\beta}_2) = 2\lambda \text{diag}[0 \; s_{x_2}^2 \cdots s_{x_p}^2]\boldsymbol{\beta} = \nabla^2 p(\boldsymbol{\beta})\boldsymbol{\beta}$ perusteella. Koska vastaavasta Taylorin polynomista suurimman uskottavuuden estimaattorille voidaan ratkaista $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}_0 - \nabla^2 l(\boldsymbol{\beta}_0)^{-1} \nabla l(\boldsymbol{\beta}_0)$, saadaan edellisestä vielä logistisen harjuregression mukaisen estimaattorin asymptoottisen jakauman kannalta varsin hyödyllinen approksimaatio

$$\begin{aligned}\hat{\boldsymbol{\beta}}^* &\approx \big(\nabla^2 l(\boldsymbol{\beta}_0) - \nabla^2 p(\boldsymbol{\beta}_0)\big)^{-1} \nabla^2 l(\boldsymbol{\beta}_0)\hat{\boldsymbol{\beta}} \\ &= \hat{\boldsymbol{\beta}} + \big(\nabla^2 l(\boldsymbol{\beta}_0) - \nabla^2 p(\boldsymbol{\beta}_0)\big)^{-1} \nabla^2 p(\boldsymbol{\beta}_0)\hat{\boldsymbol{\beta}}.\end{aligned}$$

Jos nyt opetusaineiston oletetaan edelleen olevan sen verran suuri, että estimaattorin $\hat{\boldsymbol{\beta}}$ asymptoottista jakaumaa $\text{N}\big(\boldsymbol{\beta}_0, \frac{1}{n}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ ja tuloksia $\frac{1}{n}\nabla^2 l(\boldsymbol{\beta}_0) \xrightarrow{P} -\mathcal{I}(\boldsymbol{\beta}_0)$ ja $S_{x_j} \xrightarrow{P} \sigma_{x_j}$ voidaan hyödyntää, saadaan sakotetun estimaattorin $\hat{\boldsymbol{\beta}}^*$ approksimatiiviseksi asymptoottiseksi jakaumaksi [12] edellisen perusteella

$$\text{N}\big(\boldsymbol{\beta}_0 - \tfrac{2\lambda}{n}\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathbf{C}^2\boldsymbol{\beta}_0, \tfrac{1}{n}\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\big),$$

kun $\mathbf{C}^2 = \text{diag}[0 \; \sigma_{x_2}^2 \cdots \sigma_{x_p}^2]$. Koska $\hat{\boldsymbol{\beta}}^*$ ei selvästikään ole harhaton, kun $\lambda_0 = \frac{\lambda}{\sqrt{n}} > 0$, saadaan tasainen integroitavuus oletettaessa sen KL-informaatioon perustuvan ennustevirheen odotusarvoksi luvussa 2.4 esitetyn approksimaation avulla tässä

$$\begin{aligned}\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0 || \hat{\boldsymbol{\beta}}^*)\big) &\approx \tfrac{1}{2}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\text{Cov}(\hat{\boldsymbol{\beta}}^*) + b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)\big) \\ &\approx \tfrac{1}{2n}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\big) \\ &\quad + \tfrac{2\lambda^2}{n^2}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathbf{C}^2\boldsymbol{\beta}_0\boldsymbol{\beta}_0'\mathbf{C}^2\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\big).\end{aligned}$$

Koska matriisi $\mathbf{C}$ ei ole kääntyvä, oletetaan seuraavaksi yksinkertaisuuden vuoksi, että mallissa ei ole vakiota, jolloin $\mathbf{C} = \text{diag}[\sigma_{x_1} \cdots \sigma_{x_p}]$, ja merkitään, että $\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}$ on matriisin $\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1}$ ominaisarvohajotelma, jolloin termin $\frac{1}{2}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\text{Cov}(\hat{\boldsymbol{\beta}}^*)\big)$ osalta nähdään matriisin jäljen ominaisuuksia hyödyntämällä, että

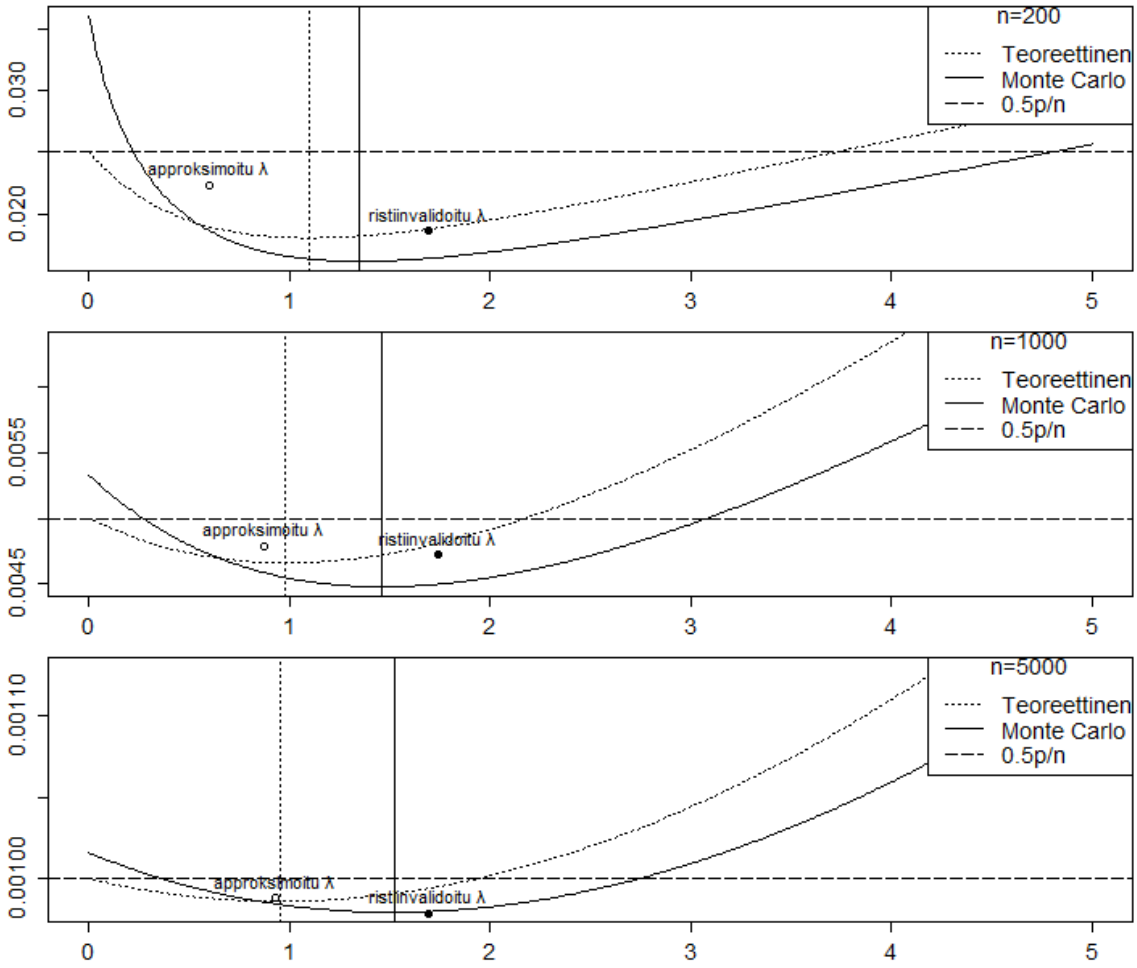
$$\begin{aligned}\tfrac{1}{2}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\text{Cov}(\hat{\boldsymbol{\beta}}^*)\big) &\approx \tfrac{1}{2n}\text{tr}\Big(\big(\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\big)^2\Big) \\ &= \tfrac{1}{2n}\text{tr}\Big(\big(\big(\mathbf{C}\big(\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1} + \tfrac{2\lambda}{n}\mathbf{I}\big)\mathbf{C}\big)^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\big)^2\Big) \\ &= \tfrac{1}{2n}\text{tr}\Big(\big((\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q} + \tfrac{2\lambda}{n}\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}\big)^2\Big) \\ &= \tfrac{1}{2n}\text{tr}\Big(\big(\mathbf{Q}'(\boldsymbol{\Lambda} + \tfrac{2\lambda}{n}\mathbf{I})^{-1}\boldsymbol{\Lambda}\mathbf{Q}\big)^2\Big) \\ &= \frac{1}{2n}\text{tr}\left(\left(\mathbf{Q}'\text{diag}\left[\frac{\lambda_1}{\lambda_1 + \frac{2\lambda}{n}} \; \cdots \; \frac{\lambda_p}{\lambda_p + \frac{2\lambda}{n}}\right]\mathbf{Q}\right)^2\right) \\ &= \frac{1}{2n}\sum_{j=1}^{p}\left(\frac{\lambda_j}{\lambda_j + \frac{2\lambda}{n}}\right)^2.\end{aligned}$$

Koska tämän arvion ja sitä vastaavan suurimman uskottavuuden menetelmän mukaisen arvion $\frac{1}{2}\text{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\text{Cov}(\hat{\boldsymbol{\beta}})\big) \approx 0.5p/n$ erotus on aina negatiivinen ja $O(\lambda_0)$, kun $\lambda_0 > 0$, edellä summattavista termeistä binomisarjaa hyödyntämällä saadun tuloksen $\big(1 + 2(\lambda_j\sqrt{n})^{-1}\lambda_0\big)^{-2} = 1 + O(\lambda_0)$ perusteella, voidaan opetusaineistoissa, joissa

[12] Tarkalleen ottaen $\frac{1}{n}\nabla^2 p(\boldsymbol{\beta}_0) = \frac{2\lambda_0}{\sqrt{n}}\text{diag}[0 \; s_{x_2}^2 \cdots s_{x_p}^2] \xrightarrow{P} \mathbf{0}$.

n on riittävän suuri ja kiinteä, aina löytää λ_0, jolla saadaan edellisen approksimaation perusteella pienempi odotettu ennustevirhe kuin estimaattorilla $\hat{\boldsymbol{\beta}}$. Tähän perusteluun tarvitaan tosin vielä sitä, että termin $\frac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)$ approksimaatio on edellä $\frac{2\lambda_0^2}{n}O(1) = O(\lambda_0^2)$, ja lähestyy siten nollaa nopeammin kuin $O(\lambda_0)$.

Jotta edellä johdetun odotetun ennustevirheen approksimaation tarkkuudesta saisi paremman käsityksen, on kuvassa 1 esitetty vielä, miten hyvin sen avulla saadut teoreettiset odotusarvot vastaavat Monte Carlo -menetelmällä estimoituja todellisia KL-informaation odotusarvoja menetelmäparametrin λ funktiona [13] logistisessa regressiomallissa, jossa $\boldsymbol{\beta}_0 = (1, 1, 1, 1, 1, 0, 0, 0, 0, 0)$ ja $\mathbf{X}_1 \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Kovarianssimatriisissa $\Sigma_{ij} = 0.7$, kun $i \neq j$ ja muuten yksi. Koska $\hat{\boldsymbol{\beta}}^*$ on suurimman uskottavuuden estimaattori, kun $\lambda = 0$, nähdään kuvasta 1 myös, miten kaikilla tähän valituilla $n \in \{200, 1000, 5000\}$ on olemassa $\lambda > 0$, jolla estimaattori $\hat{\boldsymbol{\beta}}^*$ tuottaa pienemmän odotetun ennustevirheen kuin $\hat{\boldsymbol{\beta}}$.



Kuva 1: Odotettu ennustevirhe menetelmäparametrin λ funktiona.

Käytännössä sopiva menetelmäparametri λ valitaan Hastien et al. [13, luku 7] mukaan tyypillisesti k-kertaisella ristiinvalidoinnilla, jossa opetusaineisto $(\mathbf{y}, \mathbf{X})$ jaetaan k:hon likimain yhtä suureen osaan, joista jokaista $(\mathbf{y}_i, \mathbf{X}_i)$ käytetään vuorollaan

[13] Jotta funktioiden minimikohdat erottuisivat selvemmin, on ne merkitty pystysuorilla.

testiaineistona muiden osien $(\mathbf{y}_{-i}, \mathbf{X}_{-i})$ toimiessa opetusaineistona. Näin saaduista osista lasketusta keskiarvosta $\frac{1}{n}\sum_{i=1}^{k} \log f\big(\mathbf{y}_i; \hat{\boldsymbol{\beta}}^*(\mathbf{y}_{-i}, \mathbf{X}_{-i})\big)$, jossa summattavat termit riippuvat toisistaan päällekkäisten opetusaineistojen kautta, saadaan yleisesti käytetty estimaatti odotetusta ennustevirheestä $\mathbb{E}\big(\log f\big(Y_0|\mathbf{X}_0'; \hat{\boldsymbol{\beta}}^*(\mathbf{Y}, \mathbf{X})\big)\big)$.

Koska menetelmäparametrin λ valinta etukäteen valituista kandidaateista S_λ ristiinvalidoinnilla estimoitujen odotusarvojen $\mathbb{E}\big(\log f\big(Y_0|\mathbf{X}_0'; \hat{\boldsymbol{\beta}}^*(\mathbf{Y}, \mathbf{X})\big)\big)$ perusteella tai muuten tuottaa tyypillisesti korkeamman odotetun ennustevirheen kuin optimaalinen menetelmäparametri, on kuvaan 1 lisätty myös pisteet, joista näkee sekä ristiinvalidointiin $\hat{\lambda}_{cv}$ että estimaattoriin $\hat{\lambda}_{kl}$ perustuvien menetelmien mukaiset odotetut menetelmäparametrit ja ennustevirheet, kun $S_\lambda = \{0, 0.01, 0.02, ..., 5\}$. Estimaattorilla $\hat{\lambda}_{kl}$ tarkoitetaan tässä yksinkertaisesti aiemmin johdetun odotetun ennustevirheen approksimaation, jossa tuntemattomat parametrit on korvattu niiden tarkentuvilla estimaattoreilla

$$\tfrac{1}{n}\mathbf{X}'\mathbf{V}(\mathbf{X}\hat{\boldsymbol{\beta}})\mathbf{X} \xrightarrow{P} \mathcal{I}(\boldsymbol{\beta}_0), \quad \operatorname{diag}[S_{x_1}^2 \cdots S_{x_p}^2] \xrightarrow{P} \mathbf{C}^2, \quad \hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0,$$

minimikohtaa joukossa S_λ. Vaikka estimaattori $\hat{\lambda}_{kl}$ onkin varsin mielenkiintoinen, eikä siitä näyttäisi löytyvän viitteitä kirjallisuudessa, käytetään tässä työssä jatkossa vain ristiinvalidointiin perustuvaa menetelmää, joka toimi keskimäärin paremmin.

3.2.2 Logistisen harjuregression optimaalisuus

Tarkastellaan vielä, miten KL-informaation odotusarvon approksimaatiossa esiintyvä todellinen parametrivektori $\boldsymbol{\beta}_0$ vaikuttaa odotettuun ennustevirheeseen. Oletetaan taas, että mallissa ei ole vakiota, jolloin standardoitu parametrivektori saadaan lineaarimuunnoksena $\boldsymbol{\gamma}_0 = \mathbf{C}\boldsymbol{\beta}_0$, jossa $\mathbf{C} = \operatorname{diag}[\sigma_{x_1} \cdots \sigma_{x_p}]$. Kun lisäksi merkitään, että $\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}$ on jälleen matriisin $\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1}$ ominaisarvohajotelma, saadaan odotetun ennustevirheen approksimaatiossa esiintyvän termin $\frac{1}{2}\operatorname{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)$ asymptoottiseksi approksimaatioksi matriisin jäljen ominaisuuksien avulla

$$\begin{aligned}
&\tfrac{2\lambda^2}{n^2}\operatorname{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\mathbf{C}^2\boldsymbol{\beta}_0\boldsymbol{\beta}_0'\mathbf{C}^2\big(\mathcal{I}(\boldsymbol{\beta}_0) + \tfrac{2\lambda}{n}\mathbf{C}^2\big)^{-1}\big) \\
&= \tfrac{2\lambda^2}{n^2}\operatorname{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1}\big(\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1} + \tfrac{2\lambda}{n}\mathbf{I}\big)^{-1}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0'\big(\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1} + \tfrac{2\lambda}{n}\mathbf{I}\big)^{-1}\mathbf{C}^{-1}\big) \\
&= \tfrac{2\lambda^2}{n^2}\operatorname{tr}\big(\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}(\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q} + \tfrac{2\lambda}{n}\mathbf{Q}'\mathbf{Q})^{-1}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0'(\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q} + \tfrac{2\lambda}{n}\mathbf{Q}'\mathbf{Q})^{-1}\big) \\
&= \tfrac{2\lambda^2}{n^2}\operatorname{tr}\big(\mathbf{Q}'\boldsymbol{\Lambda}(\boldsymbol{\Lambda} + \tfrac{2\lambda}{n}\mathbf{I})^{-1}\mathbf{Q}\boldsymbol{\gamma}_0\boldsymbol{\gamma}_0'\mathbf{Q}'(\boldsymbol{\Lambda} + \tfrac{2\lambda}{n}\mathbf{I})^{-1}\mathbf{Q}\big) \\
&= \tfrac{2\lambda^2}{n^2}\left\|\operatorname{diag}\left[\frac{\sqrt{\lambda_1}}{\lambda_1 + \frac{2\lambda}{n}} \cdots \frac{\sqrt{\lambda_p}}{\lambda_p + \frac{2\lambda}{n}}\right]\mathbf{Q}\boldsymbol{\gamma}_0\right\|^2.
\end{aligned}$$

Jos tässä esiintyvää neliöityä euklidista normia merkitään $\|\mathbf{w}\|^2 = \|\mathbf{D}\mathbf{Q}\boldsymbol{\gamma}_0\|^2$, nähdään, että se saavuttaa miniminsä kiinteällä rajoitteella $\tau = \|\mathbf{w}\|_1$ Jensenin epäyhtälön mukaan, kun $|w_1| = \cdots = |w_p|$, neliöfunktion aidon konveksisuuden ansiosta. Toisin sanoen tietyllä menetelmäparametrin arvolla $\lambda > 0$ ja rajoitteella τ logistisen harjuregression mukainen sakko on optimaalisin harhan suhteen silloin, kun ominaisvektoreilla $\mathbf{Q}$ kierretyn [14] ja matriisilla $\mathbf{D}$ skaalatun vektorin $\boldsymbol{\gamma}_0$ komponenttien itseisarvot ovat yhtä suuria. Luonnollisestikin optimaalisuuteen vaikuttaa toki eniten se, miten pieneksi itseisarvojen summaa koskeva rajoite τ voidaan valita.

[14] Ominaisvektorit voidaan valita siten, että $\det(\mathbf{Q}) = 1$.

Koska rajoitteen τ suhteen optimaalisin $\mathbf{w}$ on vektorin $\boldsymbol{\gamma}_0$ L1-normia koskevan rajoitteen suhteen optimaalisin, kun $\mathbf{DQ} \propto \mathbf{I}$, oletetaan yksinkertaisuuden vuoksi, että $\mathbf{C}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)\mathbf{C}^{-1} = c\mathbf{I}$ jollain vakiolla $c > 0$. Tällöin KL-informaation odotusarvon approksimaatioksi tähänastiset tulokset yhdistettäessä saadaan

$$\begin{aligned}\mathbb{E}\big(D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}}^*)\big) &\approx \frac{p}{2n}\left(\frac{c}{c+\frac{2\lambda}{n}}\right)^2 + \frac{2\lambda^2}{n^2}\left(\frac{\sqrt{c}}{c+\frac{2\lambda}{n}}\right)^2 \|\boldsymbol{\gamma}_0\|^2 \\ &= \frac{npc^2 + 4\lambda^2 c\,\|\boldsymbol{\gamma}_0\|^2}{2(nc+2\lambda)^2},\end{aligned}$$

jossa optimaalisuus kiinteillä vakioilla $\lambda > 0, \tau > 0$ tarkoittaa nyt yksinkertaisesti standardoitujen parametrien itseisarvojen yhtäsuuruutta. Koska tämän lausekkeen minimikohdaksi menetelmäparametrin $\lambda > 0$ funktiona löydetään $\lambda^* = 0.5p/\|\boldsymbol{\gamma}_0\|^2$ sen derivaatan nollakohdasta, saadaan sen minimiksi tässä erikoistapauksessa

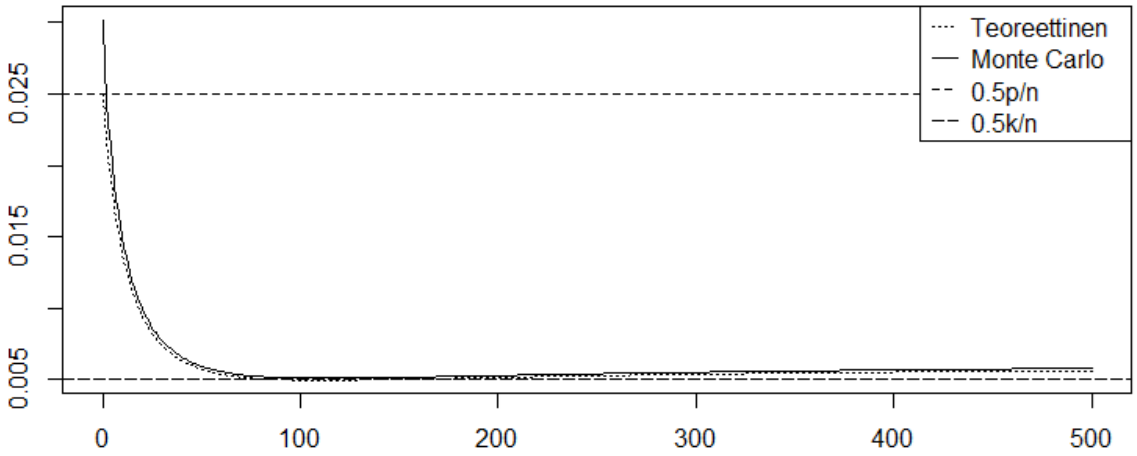
$$\frac{npc^2 + 4(\lambda^*)^2 c\,\|\boldsymbol{\gamma}_0\|^2}{2(nc+2\lambda^*)^2} = \frac{p}{2n}\left(\frac{c}{c+\frac{p}{n\|\boldsymbol{\gamma}_0\|^2}}\right).$$

Vaikka tämä approksimaatio ei pädekään yleisesti, on se silti mielenkiintoinen, sillä sen avulla voidaan perustella, että suotuisissa olosuhteissa logistinen harjuregressio voi tuottaa jopa pienemmän odotetun ennustevirheen kuin oraakkeli, joka etukäteen tietää, mitkä parametrit poikkeavat nollasta. Eli kun todellisessa parametrivektorissa $\boldsymbol{\beta}_0$ on vain $k < p$ nollasta poikkeavaa komponenttia, voi logistinen harjuregressio tuottaa asymptoottisin perustein silti pienemmän odotetun ennustevirheen kuin oraakkelina toimiva suurimman uskottavuuden menetelmä, joka estimoi vain nämä nollasta poikkeavat k parametria, kun

$$\begin{aligned}&\frac{p}{2n}\left(\frac{c}{c+\frac{p}{n\|\boldsymbol{\gamma}_0\|^2}}\right) < \frac{k}{2n} \Leftrightarrow pc < k\left(c + \frac{p}{n\|\boldsymbol{\gamma}_0\|^2}\right) \\ &\Leftrightarrow c(p-k) < \frac{kp}{n\|\boldsymbol{\gamma}_0\|^2} \Leftrightarrow \frac{1}{k}\sum_{j=1}^{p}\gamma_{0_j}^2 < \frac{p}{nc(p-k)}.\end{aligned}$$

Helpoiten tämä jollain $\tau > 0$ käy, jos kaikki parametrit $|\gamma_{0_j}| > 0$ ovat yhtä suuria, opetusaineisto on pieni, ja nollasta poikkeavia parametreja on paljon.

Jotta tämän asymptoottisesti perustellun epäyhtälön tarkkuudesta saisi paremman käsityksen, on kuvassa 2 esitetty vielä, miten tarkasti logistinen harjuregressio saavuttaa edellä kuvatun oraakkelin mukaisen odotetun ennustevirheen $0.5k/n$, kun $c \approx 0.28$, tehtyjen oletusten mukaisessa logistisessa regressiomallissa, jossa $n = 200$ ja standardoidun parametrivektorin $\boldsymbol{\gamma} \in \mathbb{R}^{10}$ kaikki $k = 2$ nollasta poikkeavaa komponenttia ovat $\gamma_{0_j} = \sqrt{p/(ncp - nck)} \approx 0.15$. Koska tämä asetelma on logistiselle harjuregressiolle varsin epäedullinen, on mielenkiintoista havaita, miten hyvin tässä johdetut teoreettiset tulokset vastaavat Monte Carlo -menetelmällä saatuja. Selvästikään logistista harjuregressiota ei siis voida sivuuttaa vain sen perusteella, että mallin oletetaan olevan harva.



Kuva 2: Odotettu ennustevirhe menetelmäparametrin λ funktiona.

3.3 Logistinen LASSO-regressio

Robert Tibshiranin [27] vuonna 1996 esittämään pienimmän itseiskutistamisen ja valinnan operaattoriin (LASSO) perustuva logistinen LASSO-regressio eroaa logistisesta harjuregressiosta optimointitehtävänä vain menetelmäparametrilla $\lambda > 0$ säädeltävän parametrivektorin $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_2)$ osavektoriin kohdistuvan sakon $p(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}_2\|_1$ osalta. Koska L1-normi riippuu L2-normin tapaan parametrien β_j itseisarvoista, voidaan selittävien muuttujien eri mittayksiköiden aiheuttamalta ongelmalta jälleen välttyä esittämällä logistinen LASSO-regressio optimointitehtävänä

$$\hat{\boldsymbol{\beta}}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \big(y_i \mathbf{x}_i' \boldsymbol{\beta} - \log\big(1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})\big) \big) - \lambda \|\mathbf{C}_2 \boldsymbol{\beta}_2\|_1 \right\},$$

jossa sakko kohdistuu vektorin $\boldsymbol{\beta}_2$ sijaan vektoriin $\mathbf{C}_2\boldsymbol{\beta}_2 = \operatorname{diag}[s_{x_2} \cdots s_{x_p}]\boldsymbol{\beta}_2$ [15].

Koska tällöin $p(\boldsymbol{\beta}) = \lambda \sum_{j=2}^{p} |s_{x_j}\beta_j|$, nähdään, että maksimoitavan funktion $l^*(\boldsymbol{\beta})$ gradienttivektori $\nabla l(\boldsymbol{\beta}) - \lambda\big(0, \operatorname{sgn}(\beta_2)s_{x_2}, ..., \operatorname{sgn}(\beta_p)s_{x_p}\big)$ ei ole määritelty, kun $\beta_j = 0$ jollain j. On kuitenkin syytä huomioida, että $l^*(\boldsymbol{\beta})$ on tässä joka tapauksessa aidosti konkaavin $l(\boldsymbol{\beta})$:n ja kaikilla $\gamma \in [0, 1]$ voimassa olevan epäyhtälön

$$\begin{aligned}
-\gamma p(\boldsymbol{\beta}) - (1-\gamma)p(\boldsymbol{\beta}^*) &= -\gamma\lambda \sum_{j=2}^{p} |s_{x_j}\beta_j| - (1-\gamma)\lambda \sum_{j=2}^{p} |s_{x_j}\beta_j^*| \\
&= -\lambda \sum_{j=2}^{p} \big(\gamma |s_{x_j}\beta_j| + (1-\gamma)|s_{x_j}\beta_j^*|\big) \\
&\leq -\lambda \sum_{j=2}^{p} |s_{x_j}\big(\gamma\beta_j + (1-\gamma)\beta_j^*\big)| = -p\big(\gamma\boldsymbol{\beta} + (1-\gamma)\boldsymbol{\beta}^*\big)
\end{aligned}$$

perusteella konkaavin $-p(\boldsymbol{\beta})$:n summana aidosti konkaavi. Newtonin menetelmää ei nyt kuitenkaan voida käyttää, sillä edes $\nabla l^*(\hat{\boldsymbol{\beta}}^*)$ ei ole välttämättä määritelty.

[15] Jos sakon halutaan kohdistuvan vakioon, on syytä huomioida logistisen harjuregression yhteydessä esitetyt huomiot.

Merkitään nyt, että $l_Q(\boldsymbol{\beta};\tilde{\boldsymbol{\beta}}) = l(\tilde{\boldsymbol{\beta}}) + \nabla l(\tilde{\boldsymbol{\beta}})'(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) + \frac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})'\mathbf{H}(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})$ on funktion l Taylorin polynomi kehitettynä pisteessä $\tilde{\boldsymbol{\beta}}$, ja todetaan, että se on aidosti konkaavi, sillä tässä $\mathbf{H}(\tilde{\boldsymbol{\beta}}) < \mathbf{0}$ kaikilla $\tilde{\boldsymbol{\beta}}$. Koska funktiota $-l^*(\boldsymbol{\beta})$ approksimoiva $g:\mathbb{R}^p \to \mathbb{R}, g(\boldsymbol{\beta}) = g(\boldsymbol{\beta};\tilde{\boldsymbol{\beta}}) = -l_Q(\boldsymbol{\beta};\tilde{\boldsymbol{\beta}}) + p(\boldsymbol{\beta})$ on tällöin aidosti konveksi, ei sen alidifferentiaali $\boldsymbol{\partial} g(\boldsymbol{\beta})$ [16] ole Rockafellarin [23, lause 23.4] mukaan tyhjä joukko millään $\boldsymbol{\beta}$. Funktion l^* maksimikohtaa voidaan siten etsiä proksimaalisella Newtonin menetelmällä, joka valitun alkuarvon $\boldsymbol{\beta}^{(0)}$ ja päivityskaavan $\mathbf{0} \in \boldsymbol{\partial} g(\boldsymbol{\beta}^{(q+1)};\boldsymbol{\beta}^{(q)})$ perusteella tuottaa luvussa 2.1 kuvatun Newtonin menetelmän tapaan jonon $\boldsymbol{\beta}^{(1)}, \boldsymbol{\beta}^{(2)}, \boldsymbol{\beta}^{(3)}, ...,$ jonka toivotaan suppenevan kohti funktion l^* maksimikohtaa $\hat{\boldsymbol{\beta}}^*$.

Koska $\mathbf{H}(\boldsymbol{\beta})$ on tässä täyttä astetta kaikilla $\boldsymbol{\beta}$, saadaan edellä mainitusta ja tässä työssä käytetyn R-ohjelmiston glmnet-lisäpaketin käyttämästä päivityskaavasta konveksin analyysin merkintöjä [17] hyödyntämällä, kun $\boldsymbol{\beta} = \boldsymbol{\beta}^{(q+1)}$ ja $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(q)}$, ratkaistua tavallista Newtonin menetelmää vastaava päivityskaava

$$\nabla l(\tilde{\boldsymbol{\beta}}) + \mathbf{H}(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \in \boldsymbol{\partial} p(\boldsymbol{\beta}) \Leftrightarrow \boldsymbol{\beta} \in \tilde{\boldsymbol{\beta}} - \mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\nabla l(\tilde{\boldsymbol{\beta}}) + \mathbf{H}(\tilde{\boldsymbol{\beta}})^{-1}\boldsymbol{\partial} p(\boldsymbol{\beta}),$$

jonka askelpituuden voidaan todeta olevan yksi. Koska glmnet käyttää Friedmanin et al. [11] mukaan takautuvan viivahakualgoritmin sijaan aina tätä askelpituutta, ei proksimaalisen Newtonin menetelmän suppenemista voida tässä osoittaa.

Osoitetaan kuitenkin seuraavassa Leen et al. [18] mukaisesti, että jos $\boldsymbol{\beta}^{(q)}$ on tarpeeksi lähellä pistettä $\hat{\boldsymbol{\beta}}^*$ ja $\boldsymbol{\beta}^{(q+1)}$ ratkaistaan tarkasti, niin proksimaalinen Newtonin menetelmä askelpituudella yksi suppenee tavallisen Newtonin menetelmän tapaan neliöllisesti, olettaen, että Hessen matriisin $\mathbf{H}(\boldsymbol{\beta})$ suurin ominaisarvo on $-m < 0$ kaikilla $\boldsymbol{\beta}^{(q)}$. Määritellään aluksi, että $\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{\mathbf{x}'\mathbf{A}\mathbf{x}}$ ja että kaikilla $\mathcal{H} > \mathbf{0}$

$$\begin{aligned}\operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}}) =& \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p}\left\{\tfrac{1}{2}\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_{\mathcal{H}}^2 + p(\boldsymbol{\beta})\right\}\\ \Leftrightarrow& \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}}) \in \tilde{\boldsymbol{\beta}} - \mathcal{H}^{-1}\boldsymbol{\partial} p\big(\operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\big),\end{aligned}$$

jolloin $\mathbf{0} \in \boldsymbol{\partial} g(\boldsymbol{\beta}^{(q+1)};\boldsymbol{\beta}^{(q)}) \Leftrightarrow \boldsymbol{\beta}^{(q+1)} = \operatorname{prox}_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})}\big(\boldsymbol{\beta}^{(q)} - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\boldsymbol{\beta}^{(q)})\big)$. Kun nyt vielä huomioidaan, että kaikilla aligradienteilla $\mathbf{v}_{\boldsymbol{\beta}} \in \boldsymbol{\partial} p(\boldsymbol{\beta})$ ja $\mathbf{v}_{\tilde{\boldsymbol{\beta}}} \in \boldsymbol{\partial} p(\tilde{\boldsymbol{\beta}})$ on suoraan alidifferentiaalin määritelmän perusteella voimassa

$$\begin{aligned}p(\boldsymbol{\beta}) - p(\tilde{\boldsymbol{\beta}}) + p(\tilde{\boldsymbol{\beta}}) - p(\boldsymbol{\beta}) = 0 &\geq \mathbf{v}'_{\tilde{\boldsymbol{\beta}}}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) + \mathbf{v}'_{\boldsymbol{\beta}}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta})\\ &= -(\mathbf{v}_{\boldsymbol{\beta}} - \mathbf{v}_{\tilde{\boldsymbol{\beta}}})'(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}),\end{aligned}$$

ja nähdään vektoreiden $\mathbf{u} = \mathcal{H}^{1/2}\mathbf{x}, \mathbf{x} \in \mathbb{R}^p$ ja $\mathbf{v} = \mathcal{H}^{1/2}\mathbf{y}, \mathbf{y} \in \mathbb{R}^p$ kohdalla epäyhtälön $\mathbf{x}'\mathcal{H}\mathbf{y}\|\mathbf{y}\|_{\mathcal{H}}^2 \leq (\mathbf{x}'\mathcal{H}\mathbf{y})^2 = (\mathbf{u}'\mathbf{v})^2 \leq (\mathbf{u}'\mathbf{u})(\mathbf{v}'\mathbf{v}) = \|\mathbf{x}\|_{\mathcal{H}}^2\|\mathbf{y}\|_{\mathcal{H}}^2 \leq \mathbf{x}'\mathcal{H}\mathbf{y}\|\mathbf{x}\|_{\mathcal{H}}^2$ seuraavan Cauchy-Schwarzin epäyhtälöstä, kun $\mathbf{x}'\mathcal{H}\mathbf{y} \geq \|\mathbf{y}\|_{\mathcal{H}}^2$, saadaan neliöllisen suppenemisen kannalta tärkeä tulos

$$\begin{aligned}&\big(\mathcal{H}\big(\boldsymbol{\beta} - \operatorname{prox}_{\mathcal{H}}(\boldsymbol{\beta})\big) - \mathcal{H}\big(\tilde{\boldsymbol{\beta}} - \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\big)\big)'\big(\operatorname{prox}_{\mathcal{H}}(\boldsymbol{\beta}) - \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\big) \geq 0\\ \Leftrightarrow&(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})'\mathcal{H}\big(\operatorname{prox}_{\mathcal{H}}(\boldsymbol{\beta}) - \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\big) \geq \|\operatorname{prox}_{\mathcal{H}}(\boldsymbol{\beta}) - \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\|_{\mathcal{H}}^2\\ \Longrightarrow&\|\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\|_{\mathcal{H}} \geq \|\operatorname{prox}_{\mathcal{H}}(\boldsymbol{\beta}) - \operatorname{prox}_{\mathcal{H}}(\tilde{\boldsymbol{\beta}})\|_{\mathcal{H}}.\end{aligned}$$

[16] Alidifferentiaalin määritelmän $\boldsymbol{\partial} f(\mathbf{x}) = \{\mathbf{v} \in \mathbb{R}^p : f(\mathbf{y}) - f(\mathbf{x}) \geq \mathbf{v}'(\mathbf{y}-\mathbf{x}), \forall \mathbf{y} \in \mathbb{R}^p\}$ perusteella on selvää, että $\mathbf{0} \in \boldsymbol{\partial} f(\mathbf{x})$, jos ja vain jos $\mathbf{x} \in \operatorname{argmin} f(\mathbf{x})$. Lisäksi jos f on konveksi ja sillä on gradientti pisteessä $\mathbf{x}$, niin $\boldsymbol{\partial} f(\mathbf{x}) = \{\nabla f(\mathbf{x})\}$ Rockafellarin [23, lause 25.1] mukaan.

[17] Jatkossa $\mathbf{A}\boldsymbol{\partial} f(\mathbf{x}) + \mathbf{B}\boldsymbol{\partial} f(\mathbf{y}) + \mathbf{c}$ on joukko $\{\mathbf{A}\mathbf{v}_{\mathbf{x}} + \mathbf{B}\mathbf{v}_{\mathbf{y}} + \mathbf{c} : \mathbf{v}_{\mathbf{x}} \in \boldsymbol{\partial} f(\mathbf{x}), \mathbf{v}_{\mathbf{y}} \in \boldsymbol{\partial} f(\mathbf{y})\}$.

Kun nyt vielä muistetaan, että luvun 2.1 mukaan Hessen matriisi on Lipschitz-jatkuva, ja huomioidaan sen suurinta ominaisarvoa koskevasta oletuksesta seuraavat varsin ilmeiset tulokset [18], saadaan epäyhtälö [19]

$$\begin{aligned}
\sqrt{m}\|\boldsymbol{\beta}^{(q+1)} - \hat{\boldsymbol{\beta}}^*\| \leq & \|\boldsymbol{\beta}^{(q+1)} - \hat{\boldsymbol{\beta}}^*\|_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})} \\
= & \big\|\mathrm{prox}_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})}(\boldsymbol{\beta}^{(q)} - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\boldsymbol{\beta}^{(q)})) \\
& - \mathrm{prox}_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})}(\hat{\boldsymbol{\beta}}^* - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\hat{\boldsymbol{\beta}}^*))\big\|_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})} \\
\leq & \big\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^* - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\big(\nabla l(\boldsymbol{\beta}^{(q)}) - \nabla l(\hat{\boldsymbol{\beta}}^*)\big)\big\|_{-\mathbf{H}(\boldsymbol{\beta}^{(q)})} \\
\leq & \tfrac{1}{\sqrt{m}}\big\|\mathbf{H}(\boldsymbol{\beta}^{(q)})(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^*) - \nabla l(\boldsymbol{\beta}^{(q)}) + \nabla l(\hat{\boldsymbol{\beta}}^*)\big\| \\
= & \tfrac{1}{\sqrt{m}}\left\|\left(\mathbf{H}(\boldsymbol{\beta}^{(q)}) - \int_0^1 \mathbf{H}\big(\boldsymbol{\beta}^{(q)} + t(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}^{(q)})\big)dt\right)(\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^*)\right\| \\
\leq & \tfrac{1}{\sqrt{m}}\big\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^*\big\| \int_0^1 L\big\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^*\big\| t\, dt = \tfrac{L}{2\sqrt{m}}\big\|\boldsymbol{\beta}^{(q)} - \hat{\boldsymbol{\beta}}^*\big\|^2,
\end{aligned}$$

josta neliöllinen suppeneminen seuraa. Koska seuraavaksi kuvattu glmnetin coordinate descent -menetelmän toteutus ei ratkaise päivityskaavasta $\mathbf{0} \in \partial g(\boldsymbol{\beta}^{(q+1)}; \boldsymbol{\beta}^{(q)})$ arvoa $\boldsymbol{\beta}^{(q+1)}$ tarkasti, ei neliöllistä suppenemista tässä työssä kuitenkaan tarkalleen ottaen edelliseen vedoten voida osoittaa.

Todetaan aluksi Friedmanin et al. [11] mukaisesti, että edellä esitetty, proksimaalisen Newtonin menetelmän päivityskaavassa esiintyvä funktio $l_Q(\boldsymbol{\beta}) = l_Q(\boldsymbol{\beta}; \boldsymbol{\beta}^{(q)})$ voidaan esittää merkittäessä, että $w_i = \pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big(1-\pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big)$, myös muodossa

$$\begin{aligned}
l_Q(\boldsymbol{\beta}) &= l(\boldsymbol{\beta}^{(q)}) + \nabla l(\boldsymbol{\beta}^{(q)})'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}) + \tfrac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)})'\mathbf{H}(\boldsymbol{\beta}^{(q)})(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}) \\
&= l(\boldsymbol{\beta}^{(q)}) + \sum_{i=1}^n \big(\big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big)\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}) - \tfrac{1}{2}w_i\big(\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)})\big)^2\big) \\
&= l(\boldsymbol{\beta}^{(q)}) - \frac{1}{2}\sum_{i=1}^n w_i\big(-2w_i^{-1}\big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big)\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)}) + \big(\mathbf{x}_i'(\boldsymbol{\beta} - \boldsymbol{\beta}^{(q)})\big)^2\big) \\
&= C(\boldsymbol{\beta}^{(q)}) - \frac{1}{2}\sum_{i=1}^n w_i\big(w_i^{-1}\big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big) - \mathbf{x}_i'\boldsymbol{\beta} + \mathbf{x}_i'\boldsymbol{\beta}^{(q)}\big)^2.
\end{aligned}$$

Koska $\mathrm{logit}(\pi_i) = \mathbf{z}_i'\boldsymbol{\gamma} = (\beta_1 + \sum_{j=2}^p \beta_j\bar{x_j}) + \sum_{j=2}^p \beta_j(x_{ij} - \bar{x_j}) = \mathbf{x}_i'\boldsymbol{\beta}$, kun selittävät muuttujat ovat vakiota lukuun ottamatta standardoitu, saadaan tästä merkittäessä, että $u_i = w_i^{-1}\big(y_i - \pi(\mathbf{x}_i'\boldsymbol{\beta}^{(q)})\big) + \mathbf{x}_i'\boldsymbol{\beta}^{(q)} = w_i^{-1}\big(y_i - \pi(\mathbf{z}_i'\boldsymbol{\gamma}^{(q)})\big) + \mathbf{z}_i'\boldsymbol{\gamma}^{(q)}$ kaikilla i, proksimaalisen Newtonin menetelmän päivityskaavaksi

$$\begin{aligned}
\boldsymbol{\beta}^{(q+1)} = \mathrm{argmin}_{\boldsymbol{\beta}}\, g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(q)}) &= \mathrm{argmin}_{\boldsymbol{\beta}} \left\{ \frac{1}{2n}\sum_{i=1}^n w_i(u_i - \mathbf{z}_i'\boldsymbol{\gamma})^2 + \frac{\lambda}{n}\|\boldsymbol{\gamma}_2\|_1 \right\} \\
&= \mathrm{argmin}_{\boldsymbol{\beta}}\, \tilde{g}(\boldsymbol{\gamma}; \boldsymbol{\gamma}^{(q)}),
\end{aligned}$$

joka merkittäessä, että $\lambda_{glmnet} = \lambda/n$, on sama kuin glmnetin käyttämä.

[18] Jos $\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}$ on matriisin $\mathcal{H} > \mathbf{0}$ ominaisarvohajotelma ja m sen pienin ominaisarvo, nähdään, että $\|\mathbf{x}\|_{\mathcal{H}}^2 = \mathrm{tr}(\mathbf{x}'\mathbf{Q}'\boldsymbol{\Lambda}\mathbf{Q}\mathbf{x}) = \mathrm{tr}(\boldsymbol{\Lambda}\mathbf{Q}\mathbf{x}(\mathbf{Q}\mathbf{x})') \geq m\,\mathrm{tr}(\mathbf{Q}\mathbf{x}\mathbf{x}'\mathbf{Q}') = m\|\mathbf{x}\|^2$. Lisäksi hyödyllistä on huomata, että tällöin $\|\mathcal{H}^{-1}\mathbf{x}\|_{\mathcal{H}}^2 = \mathbf{x}'\mathcal{H}^{-1}\mathbf{x} \leq m^{-1}\|\mathbf{x}\|^2$.

[19] Huomaa, että $\mathbf{0} \in \nabla l(\hat{\boldsymbol{\beta}}^*) - \partial p(\hat{\boldsymbol{\beta}}^*) \Leftrightarrow \hat{\boldsymbol{\beta}}^* \in \hat{\boldsymbol{\beta}}^* - \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\nabla l(\hat{\boldsymbol{\beta}}^*) + \mathbf{H}(\boldsymbol{\beta}^{(q)})^{-1}\partial p(\hat{\boldsymbol{\beta}}^*)$.

Koska funktion $\tilde{g}$ minimikohdassa $\mathbf{0} \in \boldsymbol{\partial}\tilde{g}(\boldsymbol{\gamma}^{(q+1)}; \boldsymbol{\gamma}^{(q)})$, nähdään, kun merkitään $\tilde{u}_{ij} = u_i - \sum_{k \neq j} z_{ik}\gamma_k$, aligradienttien komponenttien osalta, että [20]

$$\sum_{i=1}^{n} w_i z_{ij}(\tilde{u}_{ij} - \mathrm{z}_{ij}\gamma_j) \in \begin{cases} \{0\} & \text{kun } j = 1 \\ \lambda\partial|\gamma_j| & \text{kun } j \neq 1 \end{cases}$$

$$\implies \gamma_1 = \sum_{i=1}^{n} w_i z_{i1}\tilde{u}_{i1} / \sum_{i=1}^{n} w_i z_{i1}^2, \;\; \gamma_j \in \left(\sum_{i=1}^{n} w_i z_{ij}\tilde{u}_{ij} - \lambda\partial|\gamma_j| \right) / \sum_{i=1}^{n} w_i \mathrm{z}_{ij}^2, \quad j \neq 1,$$

ja edelleen komponenteista $\gamma_2, ..., \gamma_p$, että

$$\gamma_j = \operatorname{sgn}\left(\sum_{i=1}^{n} w_i z_{ij}\tilde{u}_{ij} \right) \max\left\{ 0, \left| \sum_{i=1}^{n} w_i z_{ij}\tilde{u}_{ij} \right| - \lambda \right\} / \sum_{i=1}^{n} w_i \mathrm{z}_{ij}^2, \quad j \neq 1,$$

kun huomioidaan, että $\lambda\partial|\gamma_j| = \{\lambda \operatorname{sgn}(\gamma_j)\}$, kun $\gamma_j \neq 0$, ja muuten $[-\lambda, \lambda]$ alidifferentiaalin määritelmän perusteella. Käytännössä glmnet ratkaisee funktion $\tilde{g}$ minimikohdan iteratiivisesti coordinate descent -menetelmällä, jossa minimikohdan arviota $\tilde{\boldsymbol{\gamma}}$ päivitetään asettamalla jokainen komponentti $\tilde{\gamma}_j$ vuorollaan arvoon γ_j^*, joka saadaan ratkaisemalla $0 \in \partial\tilde{g}_j(\gamma_j^*; \boldsymbol{\gamma}^{(q)})$ päivitetyn arvion $\tilde{\boldsymbol{\gamma}}$ mukaisilla termeillä $\tilde{u}_{ij}$.

Todetaan vielä lopuksi, että ainoa arvo, johon edellä kuvattu menetelmä voi tässä konvergoitua on $\boldsymbol{\gamma}^{(q+1)}$, sillä jokaisen päivityssyklin aikana funktion $\tilde{g}$ saama arvo $\tilde{g}(\tilde{\boldsymbol{\gamma}}; \boldsymbol{\gamma}^{(q)})$ on yksittäisen koordinaatin muutoksen jälkeen aina pienempi kuin aikaisemmin. Lisäksi jos mikään $\tilde{\gamma}_j$ ei muutu yksittäisen päivityssyklin aikana, nähdään, että $\mathbf{0} \in \boldsymbol{\partial}\tilde{g}(\tilde{\boldsymbol{\gamma}}; \boldsymbol{\gamma}^{(q)})$. Käytännössä glmnet tosin palauttaa arvon $\tilde{\boldsymbol{\gamma}}$ jo siinä vaiheessa, kun $\tilde{g}(\tilde{\boldsymbol{\gamma}}; \boldsymbol{\gamma}^{(q)})$ ei enää juurikaan muutu edellisestä päivityssyklistä.

3.3.1 Asymptoottinen jakauma ja odotettu ennustevirhe

Tarkastellaan edellä johdetun estimaattorin jakaumaa, kun sitä pidetään jälleen sovittamiseen käytetystä aineistosta riippuvana satunnaismuuttujana $\hat{\boldsymbol{\beta}}^*(\mathbf{Y}, \mathbf{X})$, ja keskitytään aluksi tilanteeseen, jossa $\beta_{0_j} \neq 0$ ja $\mathrm{P}\big(\operatorname{sgn}(\hat{\beta}_j^*) \neq \operatorname{sgn}(\beta_{0_j})\big) \approx 0$ kaikilla j mahdollista vakiota lukuun ottamatta. Koska tällöin gradienttivektori $\nabla l^*(\hat{\boldsymbol{\beta}}^*)$ voidaan esittää vastaavasti kuin logistisen harjuregression yhteydessä lähellä pistettä $\boldsymbol{\beta}_0$ Taylorin polynomina

$$\mathbf{0} = \nabla l^*(\hat{\boldsymbol{\beta}}^*) \approx \nabla l(\boldsymbol{\beta}_0) - \nabla p(\boldsymbol{\beta}_0) + \nabla^2 l(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0),$$

jonka virhe on $O(\|\hat{\boldsymbol{\beta}}^* - \boldsymbol{\beta}_0\|^2)$ ja josta saadaan ratkaistua

$$\hat{\boldsymbol{\beta}}^* \approx \boldsymbol{\beta}_0 - \nabla^2 l(\boldsymbol{\beta}_0)^{-1}\nabla l(\boldsymbol{\beta}_0) + \nabla^2 l(\boldsymbol{\beta}_0)^{-1}\nabla p(\boldsymbol{\beta}_0),$$

nähdään, että tuloksia $S_{x_j} \xrightarrow{P} \sigma_{x_j}$, $\frac{\sqrt{n}}{n}\nabla l(\boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)\big)$ ja $\frac{1}{n}\nabla^2 l(\boldsymbol{\beta}_0) \xrightarrow{P} -\mathcal{I}(\boldsymbol{\beta}_0)$ hyödyntämällä estimaattorin $\hat{\boldsymbol{\beta}}^*$ asymptoottiseksi jakaumaksi saadaan

$$\mathrm{N}\big(\boldsymbol{\beta}_0 - \tfrac{\lambda}{n}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big(0, \sigma_{x_2}\operatorname{sgn}(\beta_{0_2}), ..., \sigma_{x_p}\operatorname{sgn}(\beta_{0_p})\big), \tfrac{1}{n}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big).$$

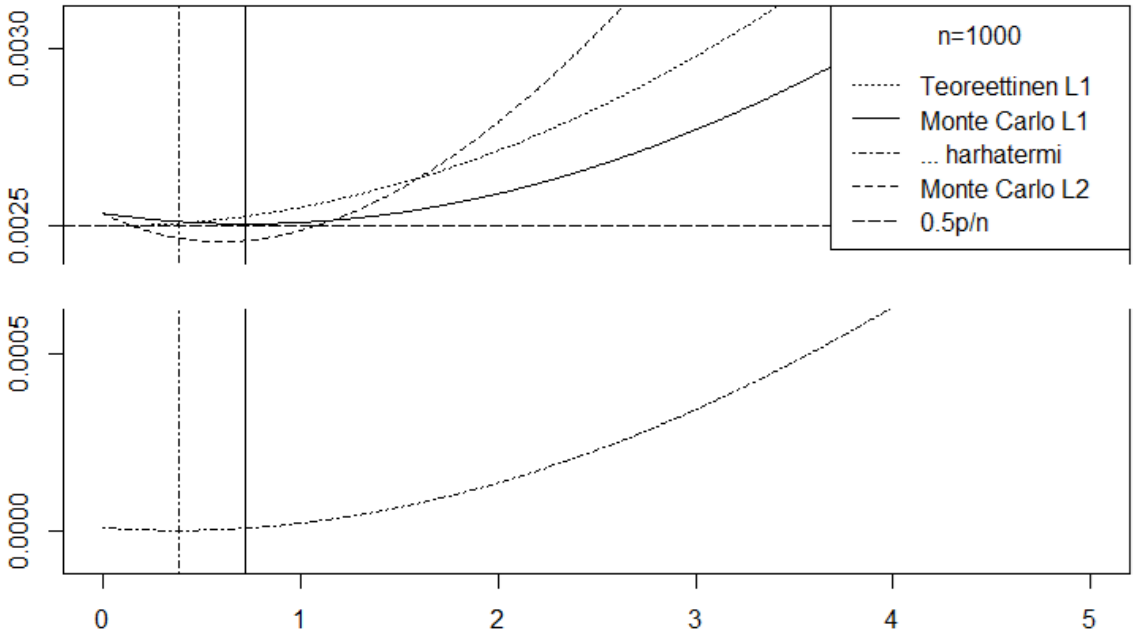
[20] Logistinen harjuregressio eroaa tästä siinä, että $\partial p_j(\boldsymbol{\gamma}) = \{2\lambda\gamma_j\}$, kun $j \neq 1$.

Tasainen integroitavuus oletettaessa KL-informaatioon perustuvan ennustevirheen odotusarvoksi saadaan luvussa 2.4 esitetyn approksimaation avulla siten

$$\begin{aligned}\mathbb{E}(D_{KL}(\boldsymbol{\beta}_0||\hat{\boldsymbol{\beta}}^*)) &\approx \tfrac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) + b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)\big)\\ &\approx \tfrac{p}{2n} + \tfrac{\lambda^2}{2n^2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{B}\big),\end{aligned}$$

kun $\mathbf{B} = \big(0, \sigma_{x_2}\mathrm{sgn}(\beta_{0_2}), ..., \sigma_{x_p}\mathrm{sgn}(\beta_{0_p})\big)\big(0, \sigma_{x_2}\mathrm{sgn}(\beta_{0_2}), ..., \sigma_{x_p}\mathrm{sgn}(\beta_{0_p})\big)'$.

Vaikka tämän perusteella L1-normin mukainen sakko näyttääkin aiheuttavan vain harhaa vähentämättä estimointiin liittyvää varianssia, kun todennäköisyys estimoida parametrin $\beta_{0_j} \neq 0$ merkki väärin on kaikilla j lähes nolla, ei tilanne kuitenkaan ole käytännössä näin paha, kuten kuvasta 3, jossa eri tavalla estimoidut KL-informaation odotusarvot sekä harhatermi $\frac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)$ menetelmäparametrin λ funktiona on esitetty, nähdään [21]. Itse asiassa tässä logistisessa regressiomallissa, jossa $\boldsymbol{\beta}_0 = (1, 1, -1, 1, -1)$ ja $\mathbf{X}_i \sim \mathrm{N}\big(\mathbf{0}, 0.7\mathbf{I} + 0.3(1, ..., 1)(1, ..., 1)'\big)$, L1-normin mukainen sakko näyttäisi aluksi jopa vaimentavan Cordeiron ja McCullaghin [6] mainitsemaa logistiseen regressioon liittyvän harhan, jonka karkea arvio on $\boldsymbol{\beta}_0 p/n$, vaikutusta odotettuun ennustevirheeseen.



Kuva 3: Odotettu ennustevirhe ja harhatermi menetelmäparametrin λ funktiona.

L2-normin mukaisen sakon ylivertaisuus edellisen esimerkin kaltaisissa tilanteissa todettaessa on syytä huomioida, että Robert Tibshirani [27] motivoi LASSO-menetelmän sillä, että se yhdistää L0- ja L2-normiin perustuvien sakkojen hyvät puolet. Tarkastellaan siis seuraavaksi tilannetta, jossa matriisin $n^{-1}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ diagonaalialkiot ovat sen verran suuria ja parametrivektorin $\boldsymbol{\beta}_0$ komponentit sen verran pieniä, että niiden suurimman uskottavuuden estimaatit vaihtavat usein jopa merkkiä. Tällöin sakotetun estimaattorin $\hat{\boldsymbol{\beta}}^*$ kertymäfunktio voidaan luvun 3.1 tapaan

[21] Jotta valittujen funktioiden minimikohdat erottuisivat selvemmin, on ne merkitty pystysuorilla.

esittää muuttujien m valintaan ehdollistettujen kertymäfunktioiden valintatodennäköisyyksillä painotettuna summana

$$F(\boldsymbol{\beta}) = \sum_{m \in M} F\big(\boldsymbol{\beta} | i(\hat{\boldsymbol{\beta}}^*) = m\big) \mathrm{P}\big(i(\hat{\boldsymbol{\beta}}^*) = m\big)$$

kokonaistodennäköisyyden kaavaa hyödyntämällä.

Vaikka tyypillisesti [17, 26, 24] näitä ehdollisia jakaumia $\hat{\boldsymbol{\beta}}^* | i(\hat{\boldsymbol{\beta}}^*) = m$ tarkastellaan vain sovitettuun malliin m ehdollistuen, on tähän työhön valittu suurimman uskottavuuden estimaattoriin $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ perustuva menetelmä, joka seuraavassa johdetaan, sillä sen avulla voidaan hahmotella kaikkien mallien asymptoottiset valintatodennäköisyydet ja ehdolliset jakaumat, kun ne ovat olemassa. Käytännössä tämä sakotetun estimaattorin $\hat{\boldsymbol{\beta}}^*$ jatkuvuutta ja paloittaista lineaarisuutta menetelmäparametrin λ funktiona hyödyntävä lähestymistapa vaatii tosin valitettavasti sen, että suurimman uskottavuuden estimaattorin $\hat{\boldsymbol{\beta}}$ asymptoottinen jakauma tunnetaan.

Palautetaan nyt mieleen, että alidifferentiaalin määritelmän mukaan pisteessä $\hat{\boldsymbol{\beta}}^*$ on voimassa yhtälö $\nabla l(\hat{\boldsymbol{\beta}}^*) = \lambda \hat{\mathbf{C}} \mathbf{v}$, jossa $\hat{\mathbf{C}} = \mathrm{diag}[0 \ s_{x_2} \cdots \ s_{x_p}]$, jollain $\mathbf{v} \in \boldsymbol{\partial} \|\hat{\boldsymbol{\beta}}^*\|_1$. Kun lisäksi huomioidaan, että $\nabla l(\hat{\boldsymbol{\beta}}^*)$ voidaan esittää lähellä pistettä $\hat{\boldsymbol{\beta}}$ Taylorin polynomina

$$\nabla l(\hat{\boldsymbol{\beta}}^*) \approx \nabla l(\hat{\boldsymbol{\beta}}) + \nabla^2 l(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = \nabla^2 l(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$$

jonka virhe on $O(\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|^2)$, saadaan pisteessä $\hat{\boldsymbol{\beta}}^*$ voimassa olevasta yhtälöstä ratkaistua vastaavan virheen $O_p(1/n)$ sisältävä approksimaatio

$$\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} \approx \lambda \nabla^2 l(\hat{\boldsymbol{\beta}})^{-1} \hat{\mathbf{C}} \mathbf{v} \Leftrightarrow \hat{\boldsymbol{\beta}}^* \approx \hat{\boldsymbol{\beta}} + \lambda \nabla^2 l(\hat{\boldsymbol{\beta}})^{-1} \hat{\mathbf{C}} \mathbf{v},$$

joka on likimain $\hat{\boldsymbol{\beta}}^* \approx \hat{\boldsymbol{\beta}} - \frac{\lambda}{n} \mathcal{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{C} \mathbf{v}$, kun $\mathbf{C} = \mathrm{diag}[0 \ \sigma_{x_2} \cdots \sigma_{x_p}]$, otoskoon ollessa niin suuri, että tuloksia $n \nabla^2 l(\hat{\boldsymbol{\beta}})^{-1} \xrightarrow{P} -\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ ja $S_{x_j} \xrightarrow{P} \sigma_{x_j}$ voidaan soveltaa siihen. Oletetaan toistaiseksi, että tässä approksimaatiossa ei ole virhettä.

Todetaan seuraavaksi, että $v_1 = 0$, koska tässä mallin oletetaan sisältävän vakion, ja mainitaan Leen et al. [17] tapaan, että $v_j \in \{-1, 1\}$, jos $\hat{\beta}_j^* \neq 0$, kun $j > 1$, jolloin edellä esitetyksi approksimaatioksi, kun $\boldsymbol{\delta} = \sqrt{n}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)$ ja $\lambda_0 = \frac{\lambda}{\sqrt{n}}$, saadaan

$$\begin{aligned}
\begin{bmatrix} \hat{\boldsymbol{\delta}}_A^* \\ \hat{\boldsymbol{\delta}}_B^* \end{bmatrix} &\approx \begin{bmatrix} \hat{\boldsymbol{\delta}}_A \\ \hat{\boldsymbol{\delta}}_B \end{bmatrix} - \begin{bmatrix} \lambda_0 \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_A \mathbf{C} \mathbf{v} \\ \lambda_0 \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_B \mathbf{C} \mathbf{v} \end{bmatrix} \\
&= \begin{bmatrix} \hat{\boldsymbol{\delta}}_A \\ \hat{\boldsymbol{\delta}}_B \end{bmatrix} - \begin{bmatrix} \lambda_0 \big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{AA} \mathbf{C}_{AA} \mathbf{v}_A + \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{AB} \mathbf{C}_{BB} \mathbf{v}_B \big) \\ \lambda_0 \big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BA} \mathbf{C}_{AA} \mathbf{v}_A + \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BB} \mathbf{C}_{BB} \mathbf{v}_B \big) \end{bmatrix},
\end{aligned}$$

jossa $\mathbf{v}_B \in [-1, 1]^{p - |m|}$, kun $A = A(m)$ on operaattori, joka muodostaa vektorin $\hat{\boldsymbol{\beta}}^*$ nollasta poikkeavien indeksien $m = i(\hat{\boldsymbol{\beta}}^*)$ mukaisia osavektoreita tai -matriiseja operaattorin $B = B(m)$ toimiessa vastaavasti joukkoon m kuulumattomille indekseille.

Edellä esitetty jako on varsin hyödyllinen, sillä sen avulla voidaan ratkaista ne menetelmäparametrin $S_\lambda \subset \mathbb{R}_{>0}$ arvot, joilla yhtälö $\hat{\boldsymbol{\delta}}^* = \hat{\boldsymbol{\delta}} - \lambda_0 \mathcal{I}(\boldsymbol{\beta}_0)^{-1} \mathbf{C} \mathbf{v}$ on voimassa tehtyjen oletusten perusteella kiinteillä $\hat{\boldsymbol{\delta}}$ ja $\mathbf{v}_A$. Muodostetaan siis ensin termin $\sqrt{n} \hat{\boldsymbol{\beta}}_B^* = \hat{\boldsymbol{\delta}}_B^* + \sqrt{n} \boldsymbol{\beta}_{0B}$ nollavektoriksi asettamisesta seuraava yhtälö

$$\begin{aligned}
&\hat{\boldsymbol{\delta}}_B + \sqrt{n} \boldsymbol{\beta}_{0B} = \lambda_0 \big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BA} \mathbf{C}_{AA} \mathbf{v}_A + \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BB} \mathbf{C}_{BB} \mathbf{v}_B \big) \\
&\quad \Leftrightarrow \mathbf{C}_{BB} \mathbf{v}_B = \big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BB} \big)^{-1} \big(\tfrac{1}{\lambda_0} (\hat{\boldsymbol{\delta}}_B + \sqrt{n} \boldsymbol{\beta}_{0B}) - \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BA} \mathbf{C}_{AA} \mathbf{v}_A \big),
\end{aligned}$$

ja merkitään, että $\mathbf{E} = \big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BB}\big)^{-1}$ ja $\mathbf{F} = \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BA}$, jolloin edellisestä ja oletuksesta $\mathbf{v}_A \in \{0\} \times \{-1,1\}^{|m|-1}, \mathbf{v}_B \in (-1,1)^{p-|m|}$ saadaan ehto

$$
\begin{aligned}
&-\sigma_{x_{B_j}} < \mathbf{E}'_{j.}(\tfrac{1}{\lambda_0}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}) - \mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A) < \sigma_{x_{B_j}} \\
\Leftrightarrow\ &\mathbf{E}'_{j.}\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A - \sigma_{x_{B_j}} < \tfrac{1}{\lambda_0}\mathbf{E}'_{j.}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}) < \mathbf{E}'_{j.}\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A + \sigma_{x_{B_j}} \\
\Longrightarrow\ &\begin{cases}
\lambda_0 \in \big(b_j(1), b_j(-1)\big) & \text{kun } c_j > 1, d_j > 0 \\
\lambda_0 \in \big(b_j(1), \infty\big) & \text{kun } c_j \in (-1,1], d_j > 0 \\
\lambda_0 \in \big(b_j(-1), \infty\big) & \text{kun } c_j \in [-1,1), d_j < 0 \\
\lambda_0 \in \big(b_j(-1), b_j(1)\big) & \text{kun } c_j < -1, d_j < 0 \\
\lambda_0 \in \emptyset & \text{kun } c_j \geq 1, d_j < 0 \text{ tai } c_j \leq -1, d_j > 0
\end{cases} \\
=\ &\begin{cases}
\lambda_0 \in \big(b_j\big(\mathrm{sgn}(d_j)\big), b_j\big(-\mathrm{sgn}(d_j)\big)\big) & \text{kun } \mathrm{sgn}(d_j)c_j > 1 \\
\lambda_0 \in \big(b_j\big(\mathrm{sgn}(d_j)\big), \infty\big) & \text{kun } c_j \in (-1,1) \cup \{\mathrm{sgn}(d_j)\} \\
\lambda_0 \in \emptyset & \text{kun } \mathrm{sgn}(d_j)c_j \leq -1
\end{cases}
\end{aligned}
$$

jossa on merkintöjen yksinkertaistamiseksi käytetty määritelmiä

$$
b_j(a) = \frac{\mathbf{E}'_{j.}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B})}{\mathbf{E}'_{j.}\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A + a\sigma_{x_{B_j}}}, a \in \{-1,1\},\ c_j = \frac{\mathbf{E}'_{j.}\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A}{\sigma_{x_{B_j}}},\ d_j = \mathbf{E}'_{j.}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}),
$$

sekä oletettu, että $\lambda_0 > 0, d_j \neq 0$. Näistä ensimmäinen oletus on ilmeinen, ja jälkimmäinen voidaan perustella sillä, että tapahtuman $E_j = \{\boldsymbol{\delta} \in \mathbb{R} : \mathbf{E}'_{j_d}\boldsymbol{\delta}_B = \epsilon\}$ todennäköisyys $\mathrm{P}(\hat{\boldsymbol{\delta}} \in E_j)$ on nolla kaikilla j ja $\epsilon \in \mathbb{R}$ estimaattorin $\hat{\boldsymbol{\delta}} \xrightarrow{d} \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ lineaarikombinaatioiden mukaisten rajajakaumien jatkuvuuden perusteella.

Näitä vektoria $\hat{\boldsymbol{\delta}}^*_B$ koskevia ehtoja tarkasteltaessa on mielenkiintoista havaita, että ne säilyttävät satunnaisen luonteensa aineiston koon kasvaessa rajatta vain siinä tapauksessa, että $\boldsymbol{\beta}_{0B}$ on nollavektori. Jos näin ei ole, karkaavat näissä ehdoissa esiintyvät menetelmäparametrin λ_0 alarajaan liittyvät termit $b_j\big(\mathrm{sgn}(d_j)\big)$ matriisin $\mathbf{E}$ kääntyvyyden perusteella lineaarikombinaatioiden $\sqrt{n}\mathbf{E}'_{j.}\boldsymbol{\beta}_{0B}$ mukana äärettömään, ellei ehdoksi tule $\lambda_0 \in \emptyset$ ainakin jollain j. Koska yleisesti jako on mahdollinen vain, jos kaikki vektoria $\hat{\boldsymbol{\delta}}^*_B$ koskevat ehdot toteutuvat, tarkoittaa tämä käytännössä sitä, että tarkastellun jaon mukainen malli on asymptoottisesti mahdollinen vain, jos $\boldsymbol{\beta}_{0B} = \mathbf{0}$, kuten edellä todettiin ja estimaattorin $\hat{\boldsymbol{\beta}}^*$ tarkentuvuuteen tarvitaan.

Tarkastellaan seuraavaksi vektoria $\sqrt{n}\hat{\boldsymbol{\beta}}^*_A$, ja todetaan, että sen komponenttien merkkien on vakiota lukuun ottamatta vastattava vektorin $\mathbf{v}_A$ komponentteja. Merkitään tätä varten, että $\mathbf{K} = \big(\mathcal{I}(\boldsymbol{\beta}_0\ )^{-1}\big)_{AA}$, $\mathbf{L} = \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{AB}$ ja $\mathbf{M} = \big(\mathcal{I}(\boldsymbol{\beta}_0)_{AA}\big)^{-1}$, ja muodostetaan kohtaa, jossa vektorin $\sqrt{n}\hat{\boldsymbol{\beta}}^*_A = \hat{\boldsymbol{\delta}}^*_A + \sqrt{n}\boldsymbol{\beta}_{0A}$ komponentin merkki voi vaihtua, koskeva yhtälö [22] [23]

$$
\begin{aligned}
\hat{\boldsymbol{\delta}}_{A_j} + \sqrt{n}\boldsymbol{\beta}_{0A_j} &= \lambda_0\big(\mathbf{K}\mathbf{C}_{AA}\mathbf{v}_A + \mathbf{L}\mathbf{E}(\tfrac{1}{\lambda_0}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}) - \mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A)\big)_j \\
\Leftrightarrow \hat{\boldsymbol{\delta}}_{A_j} + \sqrt{n}\boldsymbol{\beta}_{0A_j} &= \lambda_0\mathbf{K}'_{j.}\mathbf{C}_{AA}\mathbf{v}_A + \mathbf{L}'_{j.}\mathbf{E}(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}) - \lambda_0(\mathbf{L}\mathbf{E}\mathbf{F})'_{j.}\mathbf{C}_{AA}\mathbf{v}_A \\
\Leftrightarrow \lambda_0 &= \frac{\hat{\boldsymbol{\delta}}_{A_j} - \mathbf{L}'_{j.}\mathbf{E}\hat{\boldsymbol{\delta}}_B + \sqrt{n}(\boldsymbol{\beta}_{0A_j} - \mathbf{L}'_{j.}\mathbf{E}\boldsymbol{\beta}_{0B})}{(\mathbf{K} - \mathbf{L}\mathbf{E}\mathbf{F})'_{j.}\mathbf{C}_{AA}\mathbf{v}_A}, \quad \mathbf{M}'_{j.}\mathbf{C}_{AA}\mathbf{v}_A \neq 0,
\end{aligned}
$$

[22] Jos $\hat{\beta}^*_j \neq 0$ kaikilla j, käytetään tämän sijaan yhtälöä $\lambda_0 = (\hat{\delta}_j + \sqrt{n}\beta_{0_j})/\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)'_{j.}\mathbf{C}\mathbf{v}$

[23] Huomaa, että $\mathbf{K} - \mathbf{L}\mathbf{E}\mathbf{F} = \mathbf{M} = \big(\mathcal{I}(\boldsymbol{\beta}_0)_{AA}\big)^{-1}$ on matriisin $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ Schurin komplementti.

jonka toteutuminen ei riipu menetelmäparametrista λ_0, kun $\mathbf{M}'_{j\cdot}\mathbf{C}_{AA}\mathbf{v}_A = 0$. Koska tämän yhtälön eri indekseillä $j > 1$ saatujen nollaa suurempien ratkaisujen muodostaman joukon S^A_λ järjestetyt alkiot $\boldsymbol{\lambda} = (\lambda_1, .., \lambda_{|S^A_\lambda|})$ määräävät avoimet välit $A_\lambda = \{(0, \lambda_1), (\lambda_1, \lambda_2), ..., (\lambda_{|S^A_\lambda|}, \infty)\}$, joista korkeintaan yksi on mahdollinen, saadaan tästä tarkastellun jaon mukaista vektoria $\hat{\boldsymbol{\delta}}^*_A$ koskevat ehdot.

Tässäkin on mielenkiintoista havaita, että joukko A_λ voi säilyttää satunnaisen luonteensa ja välttyä olemasta joukko $\{(0, \infty)\}$ aineiston koon n kasvaessa rajatta vain, jos $\boldsymbol{\beta}_{0A_j} - \mathbf{L}'_{j\cdot}\mathbf{E}\boldsymbol{\beta}_{0B} = 0$ jollain vektorin $\hat{\boldsymbol{\delta}}^*_A$ komponentin indeksillä $j > 1$. Koska asymptoottisesti $\boldsymbol{\beta}_{0B}$ ei voi olla muuta kuin nollavektori, voi näin käydä vain, jos $\boldsymbol{\beta}_{0A_j} = 0$. Vaikka edellä kuvattu satunnaisuuden väheneminen periaatteessa seuraakin estimaattorin $\hat{\boldsymbol{\beta}}^*$ tarkentuvuudesta, tarjoavat edellä johdetut ehdot tietoa myös siitä, miten se parametrivektorin $\boldsymbol{\beta}_0$ lineaarikombinaatioiden muodostamien ja opetusaineiston koon vahvistamien signaalien voimistuessa tapahtuu.

Muodostetaan nyt mallien ja merkkivektoreiden asymptoottiset valintatapahtumat edellä johdetun perusteella niin, että niitä voidaan käyttää myös riittävän suurissa otoksissa. Määritellään kuitenkin ensin merkintöjen yksinkertaistamiseksi, että funktio $\psi_A(\hat{\boldsymbol{\delta}}, \mathbf{v}_A, n)$ muodostaa avoimet välit A_λ, ja palauttaa tyhjän joukon sijaan niistä sen, jossa $\operatorname{sgn}\big(\hat{\beta}^*_{A_j}(\lambda_0)\big) = \mathbf{v}_{A_j}$ kaikilla $j > 1$, kun sellainen on. Määritellään lisäksi, että funktio $\psi_B(\hat{\boldsymbol{\delta}}, \mathbf{v}_A, n)$ palauttaa vektoria $\hat{\boldsymbol{\delta}}^*_B$ koskeviin ehtoihin liittyvien mahdollisesti tyhjien avointen välien leikkauksen, jolloin vektoria $\hat{\boldsymbol{\delta}}^*$ koskeva yhtälö toteutuu kiinteillä $\hat{\boldsymbol{\delta}}$ ja $\mathbf{v}_A$, kun $\lambda_0 \in \mathbb{R}_{>0}$ kuuluu joukkoon $\psi_A(\hat{\boldsymbol{\delta}}, \mathbf{v}_A, n) \cap \psi_B(\hat{\boldsymbol{\delta}}, \mathbf{v}_A, n)$ olettaen, että tarkasteltuun jakoon liittyvä osa $\hat{\boldsymbol{\delta}}^*_B$ on olemassa.

Koska malli $m = i(\hat{\boldsymbol{\beta}}^*)$ on sama kaikilla $\mathbf{v}_A \in \{0\} \times \{-1, 1\}^{|m|-1}$, todetaan Leen et al. [17] tapaan, että käytännössä valintatapahtumat on helpointa muodostaa jokaiselle indeksijoukolle $i(\hat{\boldsymbol{\beta}}^*)$ ja merkkivektorille $\mathbf{v}_A$ erikseen. Koska vain vakion sisältävässä mallissa vektorissa $\mathbf{v}_A$ on vain yksi komponentti, voidaan sen valinta määrittää tapahtumalla

$$T_1(\lambda_0, m, \mathbf{v}_A, n) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \lambda_0 \in \psi_B(\boldsymbol{\delta}, \mathbf{v}_A, n)\}.$$

Kun mallissa on vakion lisäksi muitakin selittäviä muuttujia, joidenkin komponenttien $\hat{\beta}^*_j$ ollessa kuitenkin estimoituna nollaksi, sen ja siihen liittyvän vektorin $\mathbf{v}_A$ valinnan määrittää tapahtuma

$$T_2(\lambda_0, m, \mathbf{v}_A, n) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \lambda_0 \in \psi_A(\boldsymbol{\delta}, \mathbf{v}_A, n) \cap \psi_B(\boldsymbol{\delta}, \mathbf{v}_A, n)\}.$$

Sellaisen mallin, jossa mikään vektorin $\hat{\boldsymbol{\beta}}^*$ komponenteista ei ole nolla, ja siihen liittyvän vektorin $\mathbf{v} = \mathbf{v}_A$, valinnan määrää puolestaan tapahtuma

$$T_3(\lambda_0, m, \mathbf{v}_A, n) = \{\boldsymbol{\delta} \in \mathbb{R}^p : \lambda_0 \in \psi_A(\boldsymbol{\delta}, \mathbf{v}_A, n)\}.$$

Osoitetaan seuraavaksi, että edellä määritellyistä tapahtumista aineiston koon kasvaessa rajatta muodostuvat tapahtumat $\lim_{n\to\infty} T_j(\lambda_0, m, \mathbf{v}_A, n) = T_j(\lambda_0, m, \mathbf{v}_A)$ ovat jatkuvuusjoukkoja rajajakauman suhteen kaikilla j, sillä Slutskyn ja Portmanteau-lauseen mukaan approksimaatiossa $\hat{\boldsymbol{\delta}}^* = \hat{\boldsymbol{\delta}} - \lambda_0 \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{C}\mathbf{v} + o_p(1)$ esiintyvä satunnaisvektori $\hat{\boldsymbol{\delta}} + o_p(1) \xrightarrow{d} \mathbf{Z}$ voidaan tällöin todennäköisyyttä $\mathrm{P}\big(\hat{\boldsymbol{\delta}} \in T_j(\lambda_0, m, \mathbf{v}_A)\big)$ arvioitaessa korvata sen rajajakaumaa $\mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ noudattavalla satunnaisvektorilla $\mathbf{Z}$, ja silti saada approksimaatio, jonka virhe on $o(1)$. Koska kaikki Borel-joukot,

joiden reunan todennäköisyys on nolla rajajakauman suhteen, ovat jatkuvuusjoukkoja, keskitytään tähän ehtoon nimenomaan jakauman $\mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ suhteen.

Todetaan ensin funktion $\psi_B(\hat{\boldsymbol{\delta}},\mathbf{v}_A,n)$ palauttaman joukon reunan määräytyvän termien $b_j(a)$ ja d_j perusteella, sillä c_j on vakio. Kun $b_j(a)$ tunnistetaan vektorin $\hat{\boldsymbol{\delta}}$ jatkuvaksi muunnokseksi, nähdään vastaavasti kuin termin d_j osalta, että se ei voi saada mitään tiettyä arvoa nollasta poikkeavalla todennäköisyydellä. Koska näitä reunan määrittäviä termejä on vain rajallinen määrä, on selvää, että myös sen todennäköisyys on nolla. Myös funktion $\psi_A(\hat{\boldsymbol{\delta}},\mathbf{v}_A,n)$ palauttama joukko voidaan todeta jatkuvuusjoukoksi vastaavin perustein. Sen reunan määrittävien yhtälöiden $\hat{\boldsymbol{\beta}}^*_{A_j} = 0$ osalta riittää havaita, että niiden ratkaisut ovat vektorin $\hat{\boldsymbol{\delta}}$ jatkuvia muunnoksia termin $\mathbf{MC}_{AA}\mathbf{v}_A$ ollessa vakio. Joukon $\psi_A(\hat{\boldsymbol{\delta}},\mathbf{v}_A,n) \cap \psi_B(\hat{\boldsymbol{\delta}},\mathbf{v}_A,n)$ reunan osalta taas riittää tunnistaa se leikkaavien joukkojen reunojen yhdisteen osajoukoksi.

Koska lopullisena tavoitteena tässä on estimoida eri mallien ja niihin liittyvien merkkivektoreiden $\mathbf{v}_A$ valintatodennäköisyyksiä myös silloin, kun n on vielä äärellinen, on hyvä huomata, että tapahtumien $T_j(\lambda_0, m, \mathbf{v}_A, n)$ ja $T_j(\lambda_0, m, \mathbf{v}_A)$ symmetrisen erotuksen todennäköisyys lähestyy nollaa joka tapauksessa kaikilla j signaalina toimivien termien $\sqrt{n}\mathbf{E}'_{k\cdot}\boldsymbol{\beta}_{0B}$, $\sqrt{n}(\boldsymbol{\beta}_{0A_k} - \mathbf{L}'_{k\cdot}\mathbf{E}\boldsymbol{\beta}_{0B})$ vahvistuessa sekä estimaattoria $\hat{\boldsymbol{\delta}}$ että sen rajajakaumaa käytettäessä. Tämä tulos, jonka tarkempi perustelu on jätetty liitteeseen A, tarkoittaa käytännössä sitä, että todennäköisyyttä $\mathrm{P}\big(i(\hat{\boldsymbol{\beta}}^*) = m\big)$ voidaan arvioida summalla

$$\sum_{\mathbf{v}_A \in \{0\}\times\{-1,1\}^{|m|-1}} \mathrm{P}\big(\mathbf{Z} \in T_j(\lambda_0, m, \mathbf{v}_A, n)\big) + o(1),$$

joka voidaan laskea esimerkiksi hylkäysotannalla tapahtumien erillisyyden perusteella. Luonnollisestikin samalla voidaan tuottaa myös otos indeksijoukon m mukaisesta mallista siirtämällä simuloituja vektoreita $\hat{\boldsymbol{\beta}}^i$ tiettyyn merkkiyhdistelmään liittyvän asymptoottisin perustein saadun poikkeaman verran

$$\begin{aligned}
\begin{bmatrix} \hat{\boldsymbol{\beta}}^*_A \\ \mathbf{0} \end{bmatrix} &= \begin{bmatrix} \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\beta}}_B \end{bmatrix} - \begin{bmatrix} \frac{\lambda}{n}\big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{AA}\mathbf{C}_{AA}\mathbf{v}_A + \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{AB}\mathbf{C}_{BB}\mathbf{v}_B\big) \\ \frac{\lambda}{n}\big(\big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BA}\mathbf{C}_{AA}\mathbf{v}_A + \big(\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)_{BB}\mathbf{C}_{BB}\mathbf{v}_B\big) \end{bmatrix} \\
&= \begin{bmatrix} \hat{\boldsymbol{\beta}}_A \\ \hat{\boldsymbol{\beta}}_B \end{bmatrix} - \begin{bmatrix} \mathbf{LE}\hat{\boldsymbol{\beta}}_B + \frac{\lambda}{n}\mathbf{MC}_{AA}\mathbf{v}_A \\ \hat{\boldsymbol{\beta}}_B \end{bmatrix}.
\end{aligned}$$

Estimaattorin $\hat{\boldsymbol{\beta}}^*$ tietyn mallin m ja merkkivektorin $\mathbf{v}_A$ valintaan ehdollistetusta jakaumasta simulointi edellä kuvatulla tavalla voidaan perustella sillä, että tapahtumien $T_j(\lambda_0, m, \mathbf{v}_A, n) \cap T_*$ ja $T_j(\lambda_0, m, \mathbf{v}_A) \cap T_*$ symmetrinen erotus, kun T_* on mikä tahansa jatkuvuusjoukko jakauman $\mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ suhteen, on aina alkuperäisten valintatapahtumien symmetrisen erotuksen osajoukko kaikilla j. Koska simulointi tehdään joka tapauksessa jollain äärellisellä n, on valintatapahtuman todennäköisyyden arvioon sisältyvä virhe $o(1)$ syytä huomioida etenkin niiden mallien ja merkkivektoreiden osalta, joiden asymptoottinen valintatodennäköisyys on nolla, ehdollisen todennäköisyyden $\mathrm{P}\big(\hat{\boldsymbol{\delta}} \in T_* | i(\hat{\boldsymbol{\beta}}^*) = m\big)$ approksimaatiota

$$\frac{\sum_{\mathbf{v}_A \in \{0\}\times\{-1,1\}^{|m|-1}} \mathrm{P}\big(\mathbf{Z} \in T_* \cap T_j(\lambda_0, m, \mathbf{v}_A, n)\big)}{\sum_{\mathbf{v}_A \in \{0\}\times\{-1,1\}^{|m|-1}} \mathrm{P}\big(\mathbf{Z} \in T_j(\lambda_0, m, \mathbf{v}_A, n)\big)}.$$

muodostettaessa. Kun parametrivektori on sellainen, että $\boldsymbol{\beta}_{0B} = \mathbf{0}$ ja $\mathbf{v}_{A_k}\boldsymbol{\beta}_{0A_k} \geq 0$ kaikilla $k > 1$, estimaattorin $\hat{\boldsymbol{\beta}}^*$ ehdollinen jakauma $\hat{\boldsymbol{\beta}}^*|T_j(\lambda_0, m, \mathbf{v}_A)$ on olemassa myös asymptoottisesti. Koska $\hat{\boldsymbol{\beta}}^*_A$ on kiinteillä $\mathbf{A} = \begin{bmatrix}\mathbf{I} & -\mathbf{LE}\end{bmatrix}$, m ja $\mathbf{v}_A$ estimaattorin $\hat{\boldsymbol{\beta}}$ jatkuva muunnos $\hat{\boldsymbol{\beta}}^*_A = \mathbf{A}\hat{\boldsymbol{\beta}} - \frac{\lambda}{n}\mathbf{MC}_{AA}\mathbf{v}_A + o_p(1)$, saadaan sen ehdolliseksi jakaumaksi tällöin jollain j Slutskyn ja jatkuvan kuvauksen lauseen perusteella

$$\hat{\boldsymbol{\beta}}^*_A|T_j(\lambda_0, m, \mathbf{v}_A) \underset{as}{\sim} \mathrm{N}\big(\boldsymbol{\beta}_{0A} - \tfrac{\lambda}{n}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\mathbf{C}_{AA}\mathbf{v}_A, \tfrac{1}{n}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\big)\big|T_j(\lambda_0, m, \mathbf{v}_A),$$

joka vastaa Taylorin ja Robert Tibshiranin [26] esittämää. Tässä on hyödynnetty sitä, että $\mathbf{A}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{A}' = \begin{bmatrix}\mathbf{K} - \mathbf{LEF} & \mathbf{0}\end{bmatrix}\mathbf{A}' = \mathbf{M}$.

3.3.2 Asymptoottinen LARS-algoritmi

Käytännössä simulointi ilman täydellistä hakua parempaa käsitystä siitä, mihin tapahtumiin $T_j(\lambda_0, m, \mathbf{v}_A, n)$ simuloidut vektorit $\hat{\boldsymbol{\beta}}^i$ liittyvät on kuitenkin laskennallisesti varsin tuhlaileva menetelmä, kun tavoitteena on approksimoida kaikkien mallien valintatodennäköisyyksiä ja jakaumia, varsinkin kun huomioidaan, että nyt jokaista vertailtavaa mallia m kohden on myös $2^{|m|-1}$ mahdollista merkkiyhdistelmää. Kuvataan siis seuraavaksi lyhyesti Efron et al. [8] esittämää, LASSO-regressioon tarkoitettua LARS-algoritmia muistuttava tällaiseen simulointiin tässä työssä käytetty menetelmä, joka hyödyntää valintatapahtumat huomioivan funktion

$$\hat{\boldsymbol{\beta}}^{*i}(\lambda) = \hat{\boldsymbol{\beta}}^i - \tfrac{\lambda}{n}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathbf{Cv}(\hat{\boldsymbol{\beta}}^i, \lambda),$$

jatkuvuutta ja paloittaista lineaarisuutta.

Vaikka jatkuvuus kohdissa, joissa malli m ja merkkivektori $\mathbf{v}_A$ vaihtuvat, vaatiikin tarkemman perustelun, saadaan siitä tässä varsin yksinkertainen olettamalla simulointiin käytetyn jakauman jatkuvuuteen vedoten, että vain yksi vektorin $\hat{\boldsymbol{\beta}}^{*i}(\lambda)$ komponenttien merkeistä pyrkii vaihtumaan kerrallaan. Kun tätä kohtaa merkitään λ_*, nähdään matriisin $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ kääntyvyyden perusteella, että

$$\lim_{\lambda \to \lambda_* -} \hat{\boldsymbol{\beta}}^{*i}(\lambda) = \lim_{\lambda \to \lambda_* +} \hat{\boldsymbol{\beta}}^{*i}(\lambda),$$

jos ainoa ero toispuoleisissa raja-arvoissa liittyy siihen, tulkitaanko merkkiään vaihtavan komponentin indeksin liittyvän joukkoon A vai ei. Koska sille puolelle rajaarvoa, jossa vektorin $\hat{\boldsymbol{\beta}}^{*i}(\lambda)$ komponentin arvo on nolla, voidaan edellisen tulkintaeron perusteella aina löytää yksikäsitteinen malli ja merkkivektori, nähdään että $\hat{\boldsymbol{\beta}}^{*i}(\lambda)$ on jatkuva, ja vektorit $\mathbf{v}(\hat{\boldsymbol{\beta}}^i, \lambda)$ on yksinkertaista muodostaa.

Näiden tulosten pohjalta nähdään, että täydellistä hakua suorituskykyisempi menetelmä saadaan, kun ensin selvitetään kuuluuko menetelmäparametri λ_0 joukkoon $\psi_B(\hat{\boldsymbol{\delta}}^i, \mathbf{v}_A, n)$, jossa $\mathbf{v}_A = \mathbf{v}_{\{1\}} = 0$, vai ei. Jos se kuuluu, on mallissa vain vakio, ja jos se ei kuulu, lisätään joukkoon A funktion ψ_B palauttamaan alarajaan $b_j(\mathrm{sgn}(d_j))$ liittynyt indeksi B_j, ja jatketaan oikean mallin ja merkkivektorin etsimistä lisäämällä vektorin $\mathbf{v}_A$ uuteen indeksiin merkki $\mathrm{sgn}(d_j)$, ja toistamalla vastaavaa menettelyä kunnes oikea malli on löytynyt. Koska termin $d_j = \mathbf{E}'_{j\cdot}(\hat{\boldsymbol{\delta}}^i_B + \sqrt{n}\boldsymbol{\beta}_{0B})$ merkki ei aina ole sama kuin termin $\hat{\boldsymbol{\beta}}^i_{B_j}$, on kuitenkin varauduttava myös siihen, että jokin vektorin $\hat{\boldsymbol{\beta}}^{*i}_A$ komponenteista vaihtaa merkkiä. Jos välin ψ_A alaraja on suurempi kuin λ_0

ja välin ψ_B mahdollinen alaraja, on joukosta A ja vektorista $\mathbf{v}_A$ syytä poistaa tähän alarajaan liittyneen komponentin indeksin mukaiset alkiot, ja jatkaa etsimistä.

Jotta edellä esitetyn menetelmän tarkkuudesta saisi paremman kuvan, on taulukkoon 3 koottu sekä Monte Carlo -menetelmällä että simuloimalla suurimman uskottavuuden estimaattorin asymptoottisesta jakaumasta saadut arviot mallien m valintatodennäköisyyksistä ja odotetuista ennustevirheistä. Tässä aineiston määränneessä logistisessa regressiomallissa, jossa $\boldsymbol{\beta}_0 = (1, 0.5, 0.25, 0, 0)$, selittävät muuttujat noudattivat vakiota lukuun ottamatta multinormaalijakaumaa, jossa kovarianssimatriisin diagonaalialkiot olivat yksi muiden alkioiden ollessa 0.7 opetusaineistossa, jonka koko oli $n = 200$. Estimoitaessa menetelmäparametri oli $\lambda = 2.7$.

m	Monte-Carlo		Simulointi $\boldsymbol{\theta_0}$		Simulointi $\hat{\boldsymbol{\theta}}_1$		Simulointi $\hat{\boldsymbol{\theta}}_2$	
	P_m	D_{KL}	P_m	D_{KL}	P_m	D_{KL}	P_m	D_{KL}
1,2,3,4,5	0.17	0.009	0.17	0.009	0.16	0.008	0.36	0.011
1,3,4,5	0.02	0.02	0.02	0.02	0.12	0.008	0.03	0.02
1,2,4,5	0.07	0.01	0.07	0.01	0.09	0.01	0.05	0.02
1,4,5	0.00	0	0.00	0	0.04	0.011	0.00	0
1,2,3,5	0.18	0.008	0.18	0.008	0.07	0.01	0.28	0.011
1,3,5	0.01	0.02	0.01	0.02	0.03	0.01	0.03	0.02
1,2,5	0.06	0.01	0.06	0.01	0.02	0.01	0.04	0.02
1,5	0.00	0	0.00	0	0.00	0	0.00	0
1,2,3,4	0.18	0.008	0.18	0.008	0.13	0.008	0.11	0.014
1,3,4	0.01	0.02	0.01	0.02	0.11	0.008	0.01	0.02
1,2,4	0.06	0.01	0.06	0.01	0.08	0.01	0.02	0.02
1,4	0.00	0	0.00	0	0.03	0.01	0.00	0
1,2,3	0.17	0.007	0.17	0.007	0.06	0.01	0.05	0.02
1,3	0.01	0.02	0.01	0.02	0.03	0.01	0.01	0.02
1,2	0.04	0.01	0.04	0.01	0.02	0.01	0.01	0.02
1	0.00	0	0.00	0	0.00	0	0.00	0
mikä tahansa	1.00	0.009	1.00	0.009	1.00	0.009	1.00	0.013

Taulukko 3: Estimoidut mallien valintatodennäköisyydet ja odotetut ennustevirheet.

Monte Carlo -menetelmällä tässä tarkoitetaan sitä, että logistinen LASSO-regressio on suoritettu yhteensä sataantuhanteen edellisen mallin mukaiseen toisistaan riippumattomaan opetusaineistoon. Sen yhteydessä KL-informaation odotusarvo on estimoitu erillisten testiaineistojen avulla, jotta se edustaisi mahdollisimman tarkasti todellista odotettua ennustevirhettä. Simuloinnilla tässä taas tarkoitetaan vektoreiden $\hat{\boldsymbol{\beta}}^i$ tuottamista suoraan suurimman uskottavuuden estimaattorin asymptoottisesta jakaumasta, joko tunnettua monikkoa $\boldsymbol{\theta_0} = \big(\boldsymbol{\beta_0}, \mathcal{I}(\boldsymbol{\beta}_0), \mathbf{C}\big)$ tai sen yksittäisestä aineistosta k estimoitua tarkentuvaa vastinetta $\hat{\boldsymbol{\theta}}_k = \big(\hat{\boldsymbol{\beta}}_k, -\frac{1}{n}\nabla^2 l(\hat{\boldsymbol{\beta}})_k, \hat{\mathbf{C}}_k\big)$ käyttäen. Taulukkoon 3 on valittu kymmenestä tällaisella monikolla $\hat{\boldsymbol{\theta}}_k$ suoritetusta simuloinnista sekä paras että huonoin ehdottoman odotetun ennustevirheen arvion tarkkuuden perusteella. Koska kaikissa simuloinneissa asymptoottisiin jakaumiin perustuvan approksimaation $\frac{1}{2}\mathrm{tr}\big(\mathcal{I}(\boldsymbol{\beta}_0)\big(\mathrm{Cov}(\hat{\boldsymbol{\beta}}^*) + b(\hat{\boldsymbol{\beta}}^*)b(\hat{\boldsymbol{\beta}}^*)'\big)\big)$ on oletettu olevan riittävän tarkka, voidaan odotettu ennustevirhe niiden osalta estimoida suoraan niissä tuotetuista otoksista ilman erillistä testiaineistoa.

Vaikka monikkoon $\hat{\boldsymbol{\theta}}_1$ liittynyt simulointi tuottikin tässä lähes saman ehdottoman ennustevirheen odotusarvon arvion kuin Monte Carlo -menetelmä, on syytä huomata, että se on selvästikin osittain sattumaa, sillä iteroidun odotusarvon perusteella ehdottoman ennustevirheen odotusarvo on ehdollisten odotettujen ennustevirheiden niihin liittyvien mallien valintatodennäköisyyksillä, jotka monikon $\hat{\boldsymbol{\theta}}_1$ osalta selvästikin erosivat oikeista, painotettu keskiarvo. Valitettavasti näiden arvioiden virheet eivät tässä mene nollaan edes aineiston koon n kasvaessa rajatta, sillä satunnaisvektorissa $\hat{\boldsymbol{\delta}} = \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ esiintyvän parametrivektorin $\boldsymbol{\beta}_0$ korvaaminen sen estimaatilla $\hat{\boldsymbol{\beta}}^e$ aiheuttaa aina virheen $\sqrt{n}(\hat{\boldsymbol{\beta}}^e - \boldsymbol{\beta}_0) = O_p(1)$.

Koska tässä työssä ei ole tarkoituksena muodostaa luottamusvälejä, tyydytään tässä kohtaa toteamaan, että suurimman uskottavuuden estimaattorin asymptoottisesta jakaumasta sen todellisia monikon $\boldsymbol{\theta}_0$ mukaisia parametreja käyttäen saadut estimaatit mallien valintatodennäköisyyksistä vastasivat varsin hyvin empiirisiä jo näinkin pienessä opetusaineistossa. Kun lisäksi huomioidaan, että jopa niihin ehdollisiin jakaumiin $\hat{\boldsymbol{\beta}}^*|T_j(\lambda_0, m, \mathbf{v}_A)$, joita ei asymptoottisesti ole olemassa, perustuneet odotetun ennustevirheen approksimaatiot vastasivat varsin hyvin empiirisiä, tarjoavat nämä tulokset yhdessä varsin mielenkiintoisen ja ainakin jossain määrin uuden näkökulman estimaattorin $\hat{\boldsymbol{\beta}}^*$ odotettuun ennustevirheeseen.

3.4 Höllennetty logistinen LASSO-regressio

Höllennetystä logistisesta LASSO-regressiosta käytetään tässä Hastien et al. [14] mukaista määritelmää, johon myös R-ohjelmiston glmnet-lisäpaketin toteutus siitä perustuu. Käytännössä höllentämisellä tarkoitetaan yksinkertaisesti sitä, että tietyllä menetelmäparametrin λ arvolla saatua logistisen LASSO-regression mukaista estimaattia $\hat{\boldsymbol{\beta}}^*_\lambda$ siirretään kohti sen mukaiseen indeksijoukkoon $m = i(\hat{\boldsymbol{\beta}}^*_\lambda)$ rajoitettua suurimman uskottavuuden estimaattia $\hat{\boldsymbol{\beta}}_\lambda$. Koska kiinteillä menetelmäparametreilla $\lambda > 0$, $\gamma \in [0, 1)$ höllennetty estimaattori

$$\hat{\boldsymbol{\beta}}^*_{\lambda,\gamma} = \gamma\hat{\boldsymbol{\beta}}^*_\lambda + (1-\gamma)\hat{\boldsymbol{\beta}}_\lambda$$

on estimaattorien $\hat{\boldsymbol{\beta}}^*_\lambda$ ja $\hat{\boldsymbol{\beta}}_\lambda$ painotettu keskiarvo, ei se tarkalleen ottaen ole luvussa 3 esitetyn sakotetun logistisen regressiomenetelmän määritelmän mukainen.

Johdetaan seuraavaksi höllennetyn estimaattorin $\hat{\boldsymbol{\beta}}^*_{\lambda,\gamma}$ luvussa 3.3.1 määriteltyihin tapahtumiin $T_j(\lambda_0, m, \mathbf{v}_A)$ ehdollistetut jakaumat. Todetaan ensin tätä varten, että tiettyyn indeksijoukkoon $m = i(\hat{\boldsymbol{\beta}}^*_\lambda)$ rajoitettu suurimman uskottavuuden estimaattori voidaan, kun $\boldsymbol{\beta}_{0B} = \mathbf{0}$, luvun 2.2 perusteella esittää rajoittamattoman suurimman uskottavuuden estimaattorin $\hat{\boldsymbol{\beta}}$ lineaarimuunnoksena

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0) &= \begin{bmatrix}\mathbf{I}\\ \mathbf{0}\end{bmatrix} \mathcal{I}(\boldsymbol{\beta}_{0A})^{-1} \begin{bmatrix}\mathbf{I} & \mathbf{0}\end{bmatrix} \mathcal{I}(\boldsymbol{\beta}_0)\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1)\\
&= \begin{bmatrix}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\\ \mathbf{0}\end{bmatrix} \begin{bmatrix}\mathcal{I}(\boldsymbol{\beta}_0)_{AA} & \mathcal{I}(\boldsymbol{\beta}_0)_{AB}\end{bmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1)\\
&= \begin{bmatrix}\mathbf{I} & -\mathbf{LE}\\ \mathbf{0} & \mathbf{0}\end{bmatrix} \sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o_p(1),
\end{aligned}$$

jossa on hyödynnetty luvussa 3.3.1 käytettyjen merkintöjen lisäksi käänteismatriisin määritelmästä $\mathcal{I}(\boldsymbol{\beta}_0)^{-1}\mathcal{I}(\boldsymbol{\beta}_0) = \mathbf{I}$ seuraavaa yhtälöä

$$\begin{aligned}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\mathcal{I}(\boldsymbol{\beta}_0)_{AB} &= (\mathbf{K} - \mathbf{LEF})\mathcal{I}(\boldsymbol{\beta}_0)_{AB} \\ &= -\mathbf{L}\mathcal{I}(\boldsymbol{\beta}_0)_{BB} - \mathbf{LE}(\mathbf{I} - \mathbf{E}^{-1}\mathcal{I}(\boldsymbol{\beta}_0)_{BB}) = -\mathbf{LE}.\end{aligned}$$

Vaikka tämä tulos ei olekaan voimassa yleisesti, on hyvä muistaa, että tapahtumien $T_j(\lambda_0, m, \mathbf{v}_A)$ asymptoottinen todennäköisyys on nolla, kun $\boldsymbol{\beta}_{0B} \neq \mathbf{0}$ tai $\mathbf{v}_{A_k}\boldsymbol{\beta}_{0A_k} < 0$ jollain $k > 1$. Näin ollen edellinen tulos on riittävä ja varsin hyödyllinen, sillä kun $\boldsymbol{\beta}_{0B} = \mathbf{0}$, höllennetty estimaattori voidaan tiettyä m kiinteällä merkkivektorilla $\mathbf{v}_A$ tarkasteltaessa nähdä luvun 3.3.1 perusteella jatkuvana muunnoksena

$$\begin{aligned}\hat{\boldsymbol{\beta}}^*_{\lambda,\gamma A} &= \gamma\hat{\boldsymbol{\beta}}_A - \gamma\mathbf{LE}\hat{\boldsymbol{\beta}}_B - \tfrac{\gamma\lambda}{n}\mathbf{MC}_{AA}\mathbf{v}_A + (1-\gamma)\hat{\boldsymbol{\beta}}_{\lambda A} + o_p(1) \\ &= \begin{bmatrix}\gamma\mathbf{I} + (1-\gamma)\mathbf{I} & -\gamma\mathbf{LE} - (1-\gamma)\mathbf{LE}\end{bmatrix}\hat{\boldsymbol{\beta}} - \tfrac{\gamma\lambda}{n}\mathbf{MC}_{AA}\mathbf{v}_A + o_p(1)\end{aligned}$$

josta vastaavin perustein kuin luvussa 3.3.1 saadaan höllennetyn estimaattorin tietyn mallin m ja merkkivektorin $\mathbf{v}_A$ valintaan ehdollistetuksi asymptoottiseksi jakaumaksi, kun se on olemassa, aikaisempia merkintöjä käyttäen

$$\hat{\boldsymbol{\beta}}^*_{\lambda,\gamma A}|T_j(\lambda_0, m, \mathbf{v}_A) \underset{as}{\sim} \mathrm{N}\big(\boldsymbol{\beta}_{0A} - \tfrac{\gamma\lambda}{n}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\mathbf{C}_{AA}\mathbf{v}_A, \tfrac{1}{n}\mathcal{I}(\boldsymbol{\beta}_{0A})^{-1}\big)\big|T_j(\lambda_0, m, \mathbf{v}_A).$$

Kun menetelmäparametriksi valitaan $\gamma = 0$, on mielenkiintoista havaita, että estimaattorin $\hat{\boldsymbol{\beta}}^*_{\lambda,\gamma}$ tietyn m mukaiset ehdottomat jakaumat ovat samat kuin luvussa 3.1 esitetyt. Toisin sanoen valinnalla $\gamma = 0$ höllennetty logistinen LASSO-regressio ei voi tuottaa pienempää odotettua ennustevirhettä kuin AIC:iin perustuva paras osajoukko tai askeltava menetelmä, ellei se valitse malleja optimaalisemmin.

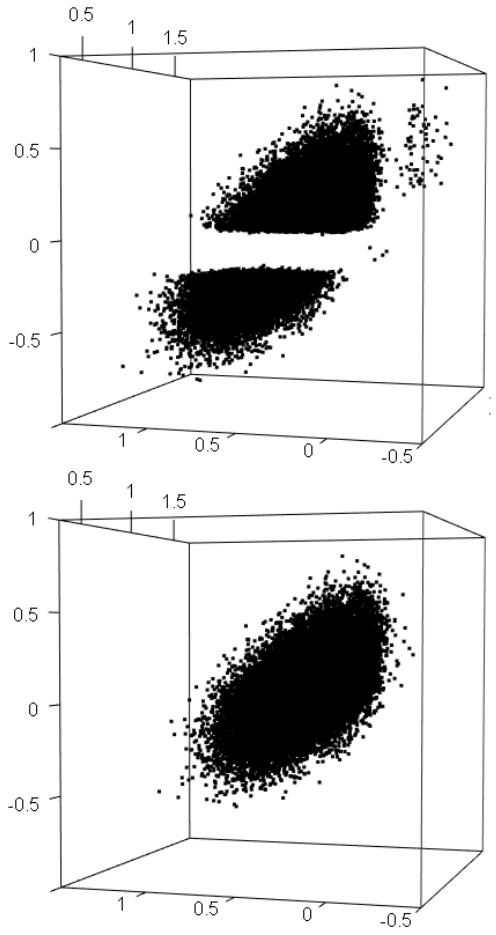
m	AIC		LASSO (λ=2.7)		Höllennetty (λ=6.74)	
	P_m	D_{KL}	P_m	D_{KL}	P_m	D_{KL}
1	0.04	0.026	0.00	0	0.02	0.026
1,2	0.76	0.005	0.42	0.005	0.66	0.005
1,3	0.10	0.018	0.02	0.018	0.04	0.017
1,2,3	0.10	0.015	0.55	0.007	0.29	0.009
mikä tahansa	1.00	0.0079	1.00	0.0065	1.00	0.0068

Taulukko 4: Estimoidut mallien valintatodennäköisyydet ja odotetut ennustevirheet.

Verrataan seuraavaksi tähän liittyen logistista LASSO-regressiota sekä sen höllennettyä versiota valinnalla $\gamma = 0$ Akaiken informaatiokriteeriin perustuvaan parhaan osajoukon menetelmään vektoriin $\boldsymbol{\beta}_0 = (1, 0.5, 0)$ perustuvan logistisen regressiomallin, jossa selittävät muuttujat vakiota lukuun ottamatta noudattavat jakaumaa $\mathrm{N}\big(\mathbf{0}, 0.3\mathbf{I} + 0.7(1, ..., 1)(1, ..., 1)'\big)$, määräämässä opetusaineistossa, jonka koko on $n = 200$. Vaikka taulukosta 4 nähdään, että indeksijoukon $m = \{1, 2\}$ mukainen malli valitaan todennäköisimmin AIC:n perusteella, on sen mukainen ehdoton odotettu ennustevirhe silti huonoin, sillä se valitsee indeksijoukot $m \not\ni 2$ muita menetelmiä useammin. Vertailu ei tosin ole täysin reilu, sillä sekä tavallisen että höllennetyn logistisen LASSO-regression osalta menetelmäparametrin λ arvoksi on valittu Monte Carlo -menetelmän perusteella optimaalisin.

Höllennettyä ja tavallista logistista LASSO-regressiota verrattaessa on puolestaan mielenkiintoista havaita, että valinnalla $\gamma = 0$ höllentäminen vaatii tavallista suuremman menetelmäparametrin λ arvon ollakseen odotetun ennustevirheen kannalta optimaalista. Tämä on tosin odotettua, kun huomioidaan, että tällainen höllentäminen käytännössä poistaa asymptoottisen harhan kokonaan valitun mallin sisältäessä kaikki aidosti selittävät muuttujat. Ehkä se, että logistisen LASSO-regression mukainen harha on tässä hyödyllistä, on kuitenkin jossain määrin yllättävää.

Käytännössä logistisen LASSO-regression odotetun ennustevirheen pienuutta selittää tässä osaltaan se, että se saavuttaa opetusaineistoissa, joissa $i(\hat{\boldsymbol{\beta}}^*_\lambda) = \{1, 2, 3\}$, tähän malliin liittyvän suurimman uskottavuuden estimaattorin approksimatiivista asymptoottista ehdotonta odotettua ennustevirhettä $0.5p/n = 0.0075$ pienemmän odotetun ennustevirheen luvussa 3 käsitellyn harhan ja varianssin välisen kompromissin ansiosta. Jotta tästä ilmiöstä saisi paremman käsityksen, on kuvassa 4 vielä esitetty edellisessä luvussa esitetyllä tavalla simuloidut menetelmäparametrilla $\lambda = 2.7$ saatuun indeksijoukkoon $m = \{1, 2, 3\}$ liittyneet suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\beta}}$ sekä niistä asymptoottisin perustein saadun harhan lisäämisen jälkeen saadut estimaatit $\hat{\boldsymbol{\beta}}^*_\lambda$. Selvästikin harhan lisääminen simuloituihin estimaatteihin siirtää tässä ne lähemmäksi toisiaan vähentäen samalla satunnaisuuteen liittyvää vaihtelua.



Kuva 4: Estimaatit $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{\beta}}^*_\lambda$ aineistoissa, joissa $i(\hat{\boldsymbol{\beta}}^*_\lambda) = \{1, 2, 3\}$.

4 Yhteenveto

Tämän tutkimuksen tarkoituksena oli vertailla teoreettisesti menetelmiä, joilla ennustetaan ehdollista todennäköisyyttä $\mathrm{P}(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i)$, tilanteessa, jossa ne saavat opetusaineistokseen perusjoukkoa edustavan riippumattoman otoksen satunnaisvektorista $(Y_1, \mathbf{X}_1)$. Suurimman uskottavuuden menetelmän lisäksi tässä työssä käsiteltyjä menetelmiä olivat Akaiken informaatiokriteeriin perustuva paras osajoukko ja askeltavat menetelmät, logistinen harju- ja LASSO-regressio sekä sen höllennetty versio. Koska ne kaikki olettavat havaitun aineiston noudattavan logistista regressiomallia, perustettiin niiden teoreettinen vertailu KL-informaation asymptoottisella approksimaatiolla arvioituun odotettuun ennustevirheeseen.

Vaikka suurimman uskottavuuden menetelmä ei ole sakotettu menetelmä, valittiin sen mukainen, KL-informaatioon perustuvan odotetun ennustevirheen suhteen asymptoottisesti optimaalinen harhaton estimaattori $\hat{\boldsymbol{\beta}}$ teoreettiseksi vertailukohdaksi muiden menetelmien mukaisille estimaattoreille. Tarkalleen ottaen tämä asymptoottinen optimaalisuus osoitettiin tässä KL-informaation odotusarvon approksimaation suhteen asymptoottisesti normaalien ja harhattomien estimaattorien joukossa parametriavaruuden nollamittaista osaa lukuun ottamatta. Tästä saatiin myös välttämätön ehto sakotettujen menetelmien sakolle: sen täytyy aiheuttaa harhaa tai tuottaa estimaattori, johon informaatioepäyhtälöä ei voida soveltaa, ollakseen asymptoottisesti perusteltavissa kaikkialla parametriavaruudessa.

Logistisen harjuregression käyttämän L2-normin mukaisen sakon tuottaman harhan perusteltiin tässä johtavan asymptoottisesti aina suurimman uskottavuuden menetelmää pienempään odotettuun ennustevirheeseen sopivasti valitulla menetelmäparametrilla λ. Vaikka logistinen LASSO-regressio ja siten myös sen höllennetty versio tuottavatkin asymptoottisesti muita menetelmiä suuremman odotetun ennustevirheen, kun parametrien merkit poikkeavat nollasta ja opetusaineiston jakauma on sellainen, että ne estimoidaan käytännössä aina oikein, nähtiin näiden menetelmien kuitenkin empiirisesti alittavan suurimman uskottavuuden menetelmän mukaisen odotetun ennustevirheen. Asymptoottisesti L1-normin mukaisen sakon ongelmana on, että se aiheuttaa edellä kuvatussa tilanteessa vain harhaa.

Kun matriisin $\frac{1}{n}\mathcal{I}(\boldsymbol{\beta}_0)^{-1}$ diagonaalialkioiden arvot kasvavat ja osa parametrivektorin $\boldsymbol{\beta}_0$ komponenttien itseisarvoista lähestyy nollaa, ajaudutaan kuitenkin jossain vaiheessa tilanteeseen, jossa menetelmät, jotka olettavat estimoitavan vektorin $\boldsymbol{\beta}_0$ kuuluvan parametriavaruuden nollamittaiseen joukkoon $\{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\|_0 < p\}$, tuottavat usein muita menetelmiä pienemmän odotetun ennustevirheen. Varmaa tämä ei kuitenkaan ole, sillä logistisen harjuregression osoitettiin voivan tuottaa jopa pienemmän odotetun ennustevirheen kuin suurimman uskottavuuden menetelmä, joka etukäteen tietää, mitkä parametrit poikkeavat nollasta. Menetelmäparametrilla $\gamma = 0$ höllennetyn logistisen LASSO-regression nähtiin tässä taas tuottavan AIC:iin perustuvaa parasta osajoukkoa ja askeltavia menetelmiä pienemmän odotetun ennustevirheen, jos ja vain jos se valitsee mallit niitä optimaalisemmin.

Lisäksi logistisen LASSO-regression ja sen höllennetyn version asymptoottiset jakaumat johdettiin osana tätä tutkimusta niiden ehdollisten jakaumien ja valintatodennäköisyyksien kautta tavalla, joka mahdollistaa niistä simuloinnin tässä kehitetyllä asymptoottisella LARS-algoritmilla. Nämä tulokset täydentävät ehdolliseen

päättelyyn ja LARS-algoritmiin liittyvää aikaisempaa tutkimusta ja mahdollistavat odotetun ennustevirheen teoreettisen tarkastelun puhtaasti suurimman uskottavuuden estimaattorin asymptoottisen jakauman pohjalta. Uutta tässä lähestymistavassa on se, että ennustevirheenä toimivan KL-informaation odotusarvo esitetään ehdollisia jakaumia ja niiden tarkasti määriteltyjä valintatapahtumia hyödyntäen. Koska logistisen harjuregression yhteydessä tarkasteltiin, milloin sen mukainen sakko on optimaalisinta, tarjoaa tämä tutkimus myös uusia työkaluja logistisen LASSO- ja harjuregression ennustekyvyn teoreettiseen vertailuun.

Todetaan vielä lopuksi, että tätä tutkimusta voitaisiin laajentaa ainakin kahteen eri suuntaan. Ensinnäkin vertailuun voitaisiin ottaa mukaan lisää sakotettuja menetelmiä, jotka olettavat havaitun aineiston noudattavan logistista regressiomallia. Toisekseen odotetun ennustevirheen lisäksi teoreettiseen tarkasteluun voitaisiin ottaa mukaan myös ennustevirheenä toimivan KL-informaation varianssi. On nimittäin mahdollista, että tutkija saattaisi haluta suosia pienempään ennustevirheen varianssiin johtavaa menetelmää, vaikka se tuottaisikin hieman suuremman odotetun ennustevirheen.

Lähteet

[1] Alan Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2015.

[2] Per Kragh Andersen ja Richard Gill. "Cox's Regression Model for Counting Processes: A Large Sample Study". *The Annals of Statistics* 10.4 (1982), s. 1100–1120.

[3] Joseph Berkson. "Application of the Logistic Function to Bio-Assay". *Journal of the American Statistical Association* 39.227 (1944), s. 357–365.

[4] Dimitris Bertsimas ja Angela King. "Logistic Regression: From Art to Science". *Statistical Science* 32.3 (2017), s. 367–384.

[5] Saskia le Cessie ja Johannes van Houwelingen. "Ridge Estimators in Logistic Regression". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 41.1 (1992), s. 191–201.

[6] Gauss Cordeiro ja Peter McCullagh. "Bias Correction in Generalized Linear Models". *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3 (1991), s. 629–643.

[7] David Cox. "The Regression Analysis of Binary Sequences". *Journal of the Royal Statistical Society: Series B (Methodological)* 20.2 (1958), s. 215–242.

[8] Bradley Efron, Trevor Hastie, Iain Johnstone ja Robert Tibshirani. "Least Angle Regression". *The Annals of Statistics* 32.2 (2004), s. 407–451.

[9] Ronald Fisher. "On the Mathematical Foundations of Theoretical Statistics". *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 222.594-604 (1922), s. 309–368.

[10] Egill Fridgeirsson, Ross Williams, Peter Rijnbeek, Marc Suchard ja Jenna Reps. "Comparing Penalization Methods for Linear Models on Large Observational Health Data". *Journal of the American Medical Informatics Association* 31.7 (2024), s. 1514–1521.

[11] Jerome Friedman, Trevor Hastie ja Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software* 33.1 (2010), s. 1–22.

[12] Christopher Greenwood, George Youssef, Primrose Letcher, Jacqui Macdonald, Lauryn Hagg, Ann Sanson, Jenn Mcintosh, Delyse Hutchinson, John Toumbourou, Matthew Fuller-Tyszkiewicz ja Craig Olsson. "A Comparison of Penalised Regression Methods for Informing the Selection of Predictive Markers". *PLOS ONE* 15.11 (2020), s. 1–14.

[13] Trevor Hastie, Robert Tibshirani ja Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. painos. Springer Series in Statistics. Springer, 2009.

[14] Trevor Hastie, Robert Tibshirani ja Ryan Tibshirani. "Best Subset, Forward Stepwise or LASSO? Analysis and Recommendations Based on Extensive Comparisons". *Statistical Science* 35.4 (2020), s. 579–592.

[15] Keith Knight ja Wenjiang Fu. “Asymptotics for LASSO-Type Estimators”. *The Annals of Statistics* 28.5 (2000), s. 1356–1378.

[16] Sadanori Konishi ja Genshiro Kitagawa. *Information Criteria and Statistical Modeling*. Springer Series in Statistics. Springer Science & Business Media, 2008.

[17] Jason Lee, Dennis Sun, Yuekai Sun ja Jonathan Taylor. “Exact Post-Selection Inference, with Application to the LASSO”. *The Annals of Statistics* 44.3 (2016), s. 907–927.

[18] Jason Lee, Yuekai Sun ja Michael Saunders. “Proximal Newton-Type Methods for Minimizing Composite Functions”. *SIAM Journal on Optimization* 24.3 (2014), s. 1420–1443.

[19] Erich Lehmann ja George Casella. *Theory of Point Estimation*. 2. painos. Springer Texts in Statistics. Springer-Verlag New York, 1998.

[20] Ian Marschner. “glm2: Fitting Generalized Linear Models with Convergence Problems”. *The R Journal* 3.2 (2011), s. 12–15.

[21] Whitney Newey ja Daniel McFadden. “Large Sample Estimation and Hypothesis Testing”. Teoksessa: *Handbook of Econometrics*. Toim. Robert Engle ja Daniel McFadden. Vol. 4. Elsevier, 1994. Luku 36, s. 2111–2245.

[22] Jorge Nocedal ja Stephen Wright. *Numerical Optimization*. 2. painos. Springer Series in Operations Research and Financial Engineering. Springer Science & Business Media, 2006.

[23] Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics and Physics. Princeton University Press, 1997.

[24] Xiang-yu Shi, Bo Liang ja Qi Zhang. “Post-Selection Inference of Generalized Linear Models Based on the LASSO and the Elastic Net”. *Communications in Statistics - Theory and Methods* 51.14 (2022), s. 4739–4756.

[25] Kenneth Tay, Balasubramanian Narasimhan ja Trevor Hastie. “Elastic Net Regularization Paths for All Generalized Linear Models”. *Journal of Statistical Software* 106.1 (2023), s. 1–31.

[26] Jonathan Taylor ja Robert Tibshirani. “Post-Selection Inference for ℓ_1-Penalized Likelihood Models”. *Canadian Journal of Statistics* 46.1 (2018), s. 41–61.

[27] Robert Tibshirani. “Regression Shrinkage and Selection via the LASSO”. *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), s. 267–288.

[28] Yan Yan, Zhizhou Yang, Tara Semenkovich, Benjamin Kozower, Bryan Meyers, Ruben Nava, Daniel Kreisel ja Varun Puri. “Comparison of Standard and Penalized Logistic Regression in Risk Model Development”. *JTCVS Open* 9 (2022), s. 303–316.

A Liite: Valintatapahtumien symmetrinen erotus

Osoitetaan, että tulos $\mathrm{P}\big(\hat{\boldsymbol{\delta}} \in T_j(\lambda_0, m, \mathbf{v}_A, n)\big) = \mathrm{P}\big(\hat{\boldsymbol{\delta}} \in T_j(\lambda_0, m, \mathbf{v}_A)\big) + o(1)$ on voimassa kaikilla $j = 1, 2, 3$ ja että vastaava tulos on voimassa myös satunnaisvektorin $\hat{\boldsymbol{\delta}} \xrightarrow{d} \mathbf{Z} \sim \mathrm{N}\big(\mathbf{0}, \mathcal{I}(\boldsymbol{\beta}_0)^{-1}\big)$ rajajakaumalle. Määritellään tätä varten luvussa 3.3.1 käytettyjä merkintöjä hyödyntäen, että

$$\mathbf{r}_k'\hat{\boldsymbol{\delta}} = \mathbf{E}_{k\cdot}'\hat{\boldsymbol{\delta}}_B, \quad s_k(a) = \mathbf{E}_{k\cdot}'\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A + a\sigma_{x_{B_k}}, \quad t_k = \mathbf{E}_{k\cdot}'\boldsymbol{\beta}_{0B},$$
$$\mathbf{u}_k'\hat{\boldsymbol{\delta}} = \mathbf{v}_{A_k}\hat{\boldsymbol{\delta}}_{A_k} - \mathbf{v}_{A_k}\mathbf{L}_{k\cdot}'\mathbf{E}\hat{\boldsymbol{\delta}}_B, \quad q_k = \mathbf{v}_{A_k}\mathbf{M}_{k\cdot}'\mathbf{C}_{AA}\mathbf{v}_A, \quad w_k = \mathbf{v}_{A_k}(\boldsymbol{\beta}_{0A_k} - \mathbf{L}_{k\cdot}'\mathbf{E}\boldsymbol{\beta}_{0B}),$$

jolloin luvussa 3.3.1 esitetystä epäyhtälöstä ja yhtälöstä

$$\mathbf{E}_{k\cdot}'\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A - \sigma_{x_{B_k}} < \tfrac{1}{\lambda_0}\mathbf{E}_{k\cdot}'(\hat{\boldsymbol{\delta}}_B + \sqrt{n}\boldsymbol{\beta}_{0B}) < \mathbf{E}_{k\cdot}'\mathbf{F}\mathbf{C}_{AA}\mathbf{v}_A + \sigma_{x_{B_k}}$$
$$\operatorname{sgn}\big(\hat{\boldsymbol{\delta}}_{A_k} - \mathbf{L}_{k\cdot}'\mathbf{E}\hat{\boldsymbol{\delta}}_B + \sqrt{n}(\boldsymbol{\beta}_{0A_k} - \mathbf{L}_{k\cdot}'\mathbf{E}\boldsymbol{\beta}_{0B}) - \lambda_0\mathbf{M}_{k\cdot}'\mathbf{C}_{AA}\mathbf{v}_A\big) = \mathbf{v}_{A_k}$$

saadaan tapahtumaa $T_j(\lambda_0, m, \mathbf{v}_A, n)$ kaikilla j vastaava tapahtuma

$$\mathcal{T}_n = \left(\bigcap_{k=1}^{p-|m|}\big\{\boldsymbol{\delta} \in \mathbb{R}^p : \lambda_0 s_k(-1) - \sqrt{n}t_k < \mathbf{r}_k'\boldsymbol{\delta} < \lambda_0 s_k(1) - \sqrt{n}t_k\big\}\right)$$
$$\bigcap\left(\bigcap_{k=2}^{|m|}\big\{\boldsymbol{\delta} \in \mathbb{R}^p : \mathbf{u}_k'\boldsymbol{\delta} > \lambda_0 q_k - \sqrt{n}w_k\big\}\right).$$

Todetaan myös, että aineiston koon kasvaessa rajatta $n \to \infty$, tästä saadaan kaikilla $j = 1, 2, 3$ tapahtumaa $T_j(\lambda_0, m, \mathbf{v}_A)$ vastaava joukko $\mathcal{T}_0 = \lim_{n\to\infty}\mathcal{T}_n$, joka edellisen määritelmän perusteella on tyhjä aina, kun $t_k \neq 0$ tai $w_k < 0$ leikkaavissa joukoissa. Luvussa 3.3.1 näin nähtiin käyvän aina, kun $\boldsymbol{\beta}_{0B} \neq \mathbf{0}$ tai $\mathbf{v}_{A_k}\boldsymbol{\beta}_{0A_k} < 0$ jollain $k > 1$. Jos tapahtuma $\mathcal{T}_0$ ei ole tyhjä joukko, se voidaan esittää muodossa

$$\mathcal{T}_0 = \left(\bigcap_{k:t_k=0}\big\{\boldsymbol{\delta} \in \mathbb{R}^p : \lambda_0 s_k(-1) < \mathbf{r}_k'\boldsymbol{\delta} < \lambda_0 s_k(1)\big\}\right)$$
$$\bigcap\left(\bigcap_{k>1:w_k=0}\{\boldsymbol{\delta} \in \mathbb{R}^p : \mathbf{u}_k'\boldsymbol{\delta} > \lambda_0 q_k\}\right).$$

Todetaan seuraavaksi, että varsinaisen mielenkiinnon kohteena oleva todennäköisyys $\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n)$ voidaan esittää summana

$$\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n) = \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0) + \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n \setminus \mathcal{T}_0) - \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0 \setminus \mathcal{T}_n)$$
$$\implies |\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n) - \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0)| \leq \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n \setminus \mathcal{T}_0) + \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0 \setminus \mathcal{T}_n).$$

Jotta haluttu tulos $\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n) = \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0) + o(1)$ saataisiin, on summasta seuraavassa epäyhtälössä oikealla puolella olevien termien mentävä nollaan aineiston koon kasvaessa rajatta. Tätä varten on hyödyllistä huomata, että

$$\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0 \setminus \mathcal{T}_n) \leq \sum_{k>1:w_k>0}\mathrm{P}(\mathbf{u}_k'\hat{\boldsymbol{\delta}} \leq \lambda_0 q_k - \sqrt{n}w_k)$$

ja että

$$\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n \setminus \mathcal{T}_0) \leq \sum_{k:t_k \neq 0} \mathrm{P}(\lambda_0 s_k(-1) - \sqrt{n} t_k < \mathbf{r}_k' \hat{\boldsymbol{\delta}} < \lambda_0 s_k(1) - \sqrt{n} t_k) + \sum_{k>1:w_k<0} \mathrm{P}(\mathbf{u}_k' \hat{\boldsymbol{\delta}} > \lambda_0 q_k - \sqrt{n} w_k).$$

Koska satunnaisvektorin $\hat{\boldsymbol{\delta}}$ lineaarikombinaatiot, vaikka ne sisältäisivät virheen $o_p(1)$, suppenevat jakaumiltaan Slutskyn ja jatkuvan kuvauksen lauseen perusteella normaalijakaumiksi, ovat ne Prohorovin lauseen perusteella tasaisesti tiukkoja. Käytännössä jokaista $\epsilon > 0$ kohden on siis olemassa vakiot M_{r_k} ja M_{u_k}, joilla epäyhtälöt

$$\sup_{n \in \mathbb{N}} \mathrm{P}(|\mathbf{r}_k' \hat{\boldsymbol{\delta}}| > M_{r_k}) < \epsilon, \quad \sup_{n \in \mathbb{N}} \mathrm{P}(|\mathbf{u}_k' \hat{\boldsymbol{\delta}}| > M_{u_k}) < \epsilon$$

ovat voimassa kaikilla k. Kun vielä huomioidaan, että nämä epäyhtälöt rajoittavat häntätodennäköisyyksien summaa, on selvää, että

$$\mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_n \setminus \mathcal{T}_0) \to 0, \quad \mathrm{P}(\hat{\boldsymbol{\delta}} \in \mathcal{T}_0 \setminus \mathcal{T}_n) \to 0$$

kuten haluttiin. Edellisen perusteella on myös ilmeistä, että

$$\mathrm{P}\big(\mathbf{Z} \in T_j(\lambda_0, m, \mathbf{v}_A, n)\big) = \mathrm{P}\big(\mathbf{Z} \in T_j(\lambda_0, m, \mathbf{v}_A)\big) + o(1).$$