Robustit menetelmät logistisessa regressiossa

Teemu Palviainen
Helsingin yliopisto
Valtiotieteellinen tiedekunta
Tilastotiede
Pro gradu -tutkielma
Joulukuu 2013

Sisältö

Luku 1

Johdanto

Tämän tutkielman tarkoituksena on tarkastella robustisuutta logistisessa regressiomallissa. Mallin ollessa robusti, voidaan todeta, että se ei ole herkkä huomattavasti muista havainnoista poikkeaville havainnoille. Tarkastelun kohteena on muutama valikoitu lähestymistapa, joista yhdessä (Carroll & Pederson) lähtökohtana on se, että pienellä otoskoolla luokitteluvirheet tekevät mallista helposti harhaisen. Tilanteeseen tarjotaan ratkaisuksi muun muassa luokitteluvirhe-estimaattia, jonka avulla harhaisuutta on saatu minimoitua. Luokitteluvirhe-estimaatit ovat myös yhtenä esimerkkitapauksena, kun tutkielman loppuosassa vertaillaan logistisen regressiomallin antamia estimaatteja eri robustisuusmenetelmillä.

Eräs toinen käsiteltävä lähestymistapa (Croux & Haesbroeck) käsittelee Biancon ja Yohain estimaattorin soveltamista logistiseen regressiomalliin. Tämän tapauksen yhteydessä tuodaan esille myös tappio-funktio (loss function), jota voidaan soveltaa robustisuustarkasteluiden yhteydessä ja jos ja vain jos suurimman uskottavuuden estimaattori on johdettavissa. Tämä lähestymistapa on sovellettavissa myös painotettujen muuttujien tapauksessa.

Molempien lähestymistapojen antamia tuloksia on pyritty havainnollistamaan erilaisten outliereiden (voimakkaasti muista havainnoista poikkeava havainto) avulla sekä vertailemaan niitä tavanomaisen logistisen regressiomallin antamiin tuloksiin. Tutkielma rakentuu siten, että aluksi on esitelty eri regressiomenetelmät keskeisine tuloksineen jonka jälkeen menetelmien antamia tuloksia on vertailtu keskenään robustisuustarkastelun näkökulmasta käyttäen esimerkkiaineistoa sekä simuloitua tilannetta.

Robustia logistista regressiota on myös aikaisemmin tutkittu muista eri näkökulmista, joihin ei tässä tutkielmassa syvemmin perehdytä. Yhtenä esimerkkinä mainittakoon ns. luokittelu-kohina robusti logistinen regressio (Label-noise robust logistic regression, [3]), joka perustuu osittain myös tässäkin tutkielmassa esiteltyyn luokitteluvirheen todennä-

köisyyden estimointiin. Menetelmän käyttöä perustellaan sillä, että usein aineistossa annettujen luokkien määritelmiin luotetaan sokeasti, vaikka takeita näiden oikeellisuudesta ei läheskään aina ole. Tämä saattaa johtaa usein jo olemassa olevien mallin parametrien luotettavuuden heikkenemiseen. Havaintojen väärinluokittelu sovitettaessa logistista regressiomallia on ongelmallista etenkin silloin, kun kyseessä ovat pienet havaintoaineistot (esimerkiksi jotkin lääketieteeseen liittyvät aineistot, joissa tutkittavien tapausten määrä on esimerkiksi 10). Ongelmaa onkin tähän asti pyritty korjaamaan siten, että ongelmalliseksi epäiltyjä havaintoyksikköjä on poistettu tai luokiteltu uudelleen. Tällöin kuitenkin riski menettää hyödyllistä informaatiota kasvaa huomattavasti, varsinkin jos tutkittavia tapauksia aineistossa on vähän. [3]

Toisena esimerkkinä mainittakoon robusti logistinen regressio, joka hyödyntää katkaistua logistista tappio-funktiota (Robust penalized logistic regression, RPLR, [13]). Katkaisulla tarkoitetaan tässä yhteydessä sitä, että tavallisen logistisen mallin tappio-funktio on konveksi, mutta katkaisu tekee siitä ei-konveksin. Kyseisen lähestymistavan yhteyttä muihin, jo aikaisemmin kehiteltyihin robusteihin logistisiin malleihin on myös pyritty selvittämään. Mallin teoreettinen tarkastelu osoittaa, että kyseinen malli on yhtäpitävä Fisherin kehittämien tulosten (liittyen luokittelun kalibrointiin) kanssa sekä robustimpi outliereita kohtaan. Saman tarkastelun yhteydessä on säätelyparametrin valintaan kehitetty erityinen menetelmä, jota kutsutaan nimellä estimoitu yleistetty ristiinvalidointi (Estimated generalized approximate cross validation, EGACV). Samaisen menetelmän esittelyn yhteydessä on päädytty tuloksiin, joissa numeerisilla esimerkeillä on osoitettu, että tappio-funktion katkaiseminen tuottaa tarkempia estimaatteja, kun kiinnitetään huomiota luokittelun laatuun sekä luokittelutodennäköisyyksien estimointiin. [13]

Luku 2

Logistinen regressio

2.1 Logit-malli

Ennen robustien menetelmien esittelyä, on hyvä palauttaa mieleen tavallinen logistinen regressiomalli, joka on selitettävän 0/1 -muuttujan Y ja selittävien muuttujien $x_1, ..., x_p$ mukana ollessa muotoa

$$g(\pi) = \eta = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}^t\boldsymbol{\beta},$$

jossa $g(\pi)$ on logit-linkki, $\boldsymbol{x} = (1, x_1, ..., x_p)^t$, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ ja $\pi = P(Y = 1|\boldsymbol{x})$. Nyt nähdään, että

$$\frac{\pi}{1-\pi} = \exp(\boldsymbol{x}^t\boldsymbol{\beta}),$$

jolloin

$$\pi = \frac{\exp(\boldsymbol{x}^t\boldsymbol{\beta})}{1 + \exp(\boldsymbol{x}^t\boldsymbol{\beta})}.$$

Mallin log-uskottavuusfunktio on

$$l(\boldsymbol{\pi}; \boldsymbol{y}) = \sum_{i=1}^{n}\left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i)\right], \tag{2.1}$$

jossa n on erilaisten profiilien $\boldsymbol{x_i}$ lukumäärä ja $\boldsymbol{y} = (y_1, ..., y_n)^t$, jossa havainnot $y_1, ..., y_n$ ovat riippumattomia $Bin(m_i, \pi_i)$ jakautuneita satunnaismuuttujia, kun $i = 1, ..., n$ ja m_i on toistojen lukumäärä sekä $\boldsymbol{\pi} = (\pi_1, ..., \pi_n)$. Nyt siis tiedetään, että logistisen mallin tapauksessa

$$g(\pi_i) = \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \sum_{j=0}^{p} x_{ij}\beta_j = \boldsymbol{x}_i^t\boldsymbol{\beta}.$$

Sijoittamalla tämä yhtälöön (2.1), saadaan log-uskottavuusfunktioksi

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{n}\sum_{j=0}^{p} y_i x_{ij}\beta_j - \sum_{i=1}^{n} m_i \log\left(1 + \exp\left(\sum_{j=0}^{p} x_{ij}\beta_j\right)\right),$$

jossa $\sum_{i=1}^{n}\sum_{j=0}^{p} y_i x_{ij}\beta_j = \sum_{i=1}^{n} y_i \boldsymbol{x}_i^t\boldsymbol{\beta} = \boldsymbol{\beta}^t(\boldsymbol{X}^t\boldsymbol{y})$ eli log-uskottavuus riippuu $\boldsymbol{y}$:stä ainoastaan lineaarikombinaation $\boldsymbol{X}^t\boldsymbol{y}$ kautta, joten $\boldsymbol{X}^t\boldsymbol{y}$ on tyhjentävä tunnusluku $\boldsymbol{\beta}$:lle. [12]

2.2 Suurimman uskottavuuden estimointi

Derivoimalla aiemmin esitetty log-uskottavuusfunktio (2.1) π_i:n suhteen, saadaan

$$\frac{\partial l}{\partial \pi_i} = \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}.$$

Käyttämällä ketjusääntöä, nähdään että derivaatta β_r:n suhteen on

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)}\frac{\partial \pi_i}{\partial \beta_r}.$$

Yleistettyjen lineaaristen mallien tapauksessa on hyvä tuoda esiin vielä seuraava tulos

$$\frac{\partial \pi_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i}\frac{\partial \eta_i}{\partial \beta_r} = \frac{d\pi_i}{d\eta_i}x_{ir}.$$

Nyt derivaataksi β_r:n suhteen saadaan siis

$$\frac{\partial l}{\partial \beta_r} = \sum_{i=1}^{n} \frac{y_i - m_i\pi_i}{\pi_i(1-\pi_i)} \frac{d\pi_i}{d\eta_i} x_{ir}. \tag{2.2}$$

Fisherin informaation alkio rs on

$$-E\left(\frac{\partial^2 l}{\partial \beta_r \partial \beta_s}\right) = \sum_{i=1}^{n} \frac{m_i}{\pi_i(1-\pi_i)} \frac{\partial \pi_i}{\partial \beta_r} \frac{\partial \pi_i}{\partial \beta_s} = \sum_{i=1}^{n} m_i \frac{(d\pi_i/d\eta_i)^2}{\pi_i(1-\pi_i)} x_{ir}x_{is} = \{\boldsymbol{X}^t \boldsymbol{W} \boldsymbol{X}\}_{rs},$$

jossa

$$\boldsymbol{W} = \text{diag}\left\{ m_i \left(\frac{d\pi_i}{d\eta_i}\right)^2 \Big/ \pi_i(1-\pi_i) \right\} \quad \text{ja} \quad \mathbf{X} = \begin{pmatrix} \boldsymbol{x}_1^t \\ \vdots \\ \boldsymbol{x}_n^t \end{pmatrix}.$$

Logistisen mallin tapauksessa $\frac{\partial \pi_i}{\partial \eta_i} = \pi_i(1-\pi_i)$, joten yhtälö (2.2) saadaan matriisimuotoon

$$\frac{\partial l}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^t(\boldsymbol{y} - \boldsymbol{\mu}),$$

jossa $\boldsymbol{\mu} = (m_1\pi_1, ..., m_m\pi_m)^t$. Lisäksi diagonaalimatriisi $\boldsymbol{W}$ saadaan muotoon

$$\boldsymbol{W} = \text{diag}\{m_i\pi_i(1-\pi_i)\}.$$

Parametriestimaatit saadaan laskettua numeerisesti iteratiivisesti uudelleenpainotetun pienimmän neliösumman menetelmän (IRLS, iteratively re-weighted least-squares) avulla. Alkuarvon $\hat{\boldsymbol{\beta}}_0$ avulla saadaan laskettua vektorit $\hat{\boldsymbol{\pi}}_0$ ja $\hat{\boldsymbol{\eta}}_0$. Käyttämällä kyseisiä vektoreita, saadaan määritellyksi vektori $\boldsymbol{z}$, jonka komponentit ovat

$$z_i = \hat{\eta}_i + \frac{y_i - m_i\hat{\pi}_i}{m_i} \frac{1}{\hat{\pi}_i(1-\hat{\pi}_i)}.$$

Parametrivektorin $\boldsymbol{\beta}$ uudeksi estimaatin arvoksi saadaan

$$\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}^t \hat{\boldsymbol{W}} \boldsymbol{X})^{-1} \boldsymbol{X}^t \hat{\boldsymbol{W}} \boldsymbol{z}, \tag{2.3}$$

jossa $\hat{\boldsymbol{W}}$ ja $\boldsymbol{z}$ on laskettu käyttämällä alkuarvoa $\hat{\boldsymbol{\beta}}_0$. Seuraavassa vaiheessa $\hat{\boldsymbol{\beta}}_1$ toimii uutena alkuarvona laskettaessa yhtälön (2.3) oikealla puolella olevia suureita $\hat{\boldsymbol{W}}$ ja $\boldsymbol{z}$. Iteraatioita jatketaan, kunnes päästään riittävän lähelle lopullista ratkaisua. [12]

Luku 3

Robustit menetelmät

3.1 Mallows-tyyppiset estimaattorit

Ohessa yleiskatsaus menetelmistä, joissa estimaattori logistiselle regressiomallille saadaan ratkaisuna yhtälöstä

$$\mathbf{0} = \sum_{i=1}^{n} w_i \boldsymbol{x}_i \{y_i - F(\boldsymbol{x}_i^t \boldsymbol{\beta}) - c(\boldsymbol{x}_i, \boldsymbol{\beta})\}, \tag{3.1}$$

jossa $\{w_i\}$ sisältää mallissa tarvittavat painot ja $F(\boldsymbol{x}_i^t \boldsymbol{\beta}) = P(y = 1 | \boldsymbol{x}_i)$. Jos $w_i \equiv 1$ ja $c(\boldsymbol{x}_i, \boldsymbol{\beta}) \equiv \mathbf{0}$, niin yhtälö (3.1) antaa tavallisen logistisen regressioestimaatin tiivistämättömälle aineistolle. Jos taas $w_i = w(\boldsymbol{x}_i, \boldsymbol{x}_i^t \boldsymbol{\beta})$ ja $c(\boldsymbol{x}_i, \boldsymbol{\beta}) \equiv \mathbf{0}$, niin painot riippuvat ainoastaan asetelmastaan ja puhutaan niin sanotusta Mallows-luokasta, joka on seuraavaksi tarkasteltavien menetelmien oleellinen osa. [9,11] Mallows-luokan tuloksissa painot eivät riipu vasteesta suoraan, vaan asetelman $\{\boldsymbol{x}_i\}$ ja parametrin $\boldsymbol{\beta}$ kautta. Tämä perustuu siihen, että havainnot, jotka poikkeavat voimakkaasti, vaikuttavat liikaa aineistossa ja näihin on kohdistettava eri toimenpiteitä harhan korjaamiseksi, kuten alaspäin painottamista (downweighting). Estimaattori $\hat{\boldsymbol{\beta}}$ on välttämättä tarkentuva, koska estimointiyhtälö (3.1) on harhaton. Vaikka Mallows-luokan estimaatit ovat vähemmän tehokkaita kuin perinteisen logistisen mallin SU-estimaatit, voidaan niille laskea helposti harha, joka voi olla pienempi kuin SU-estimaatilla varsinkin, jos aineistossa on epätavallisia $\boldsymbol{x}_i$ arvoja. [5]

Olkoon

$$\boldsymbol{V}_{nj}(\boldsymbol{\beta}) = n^{-1} \sum_{i=1}^{n} w_i^j \boldsymbol{x}_i \boldsymbol{x}_i^t F^{(1)}(\boldsymbol{x}_i^t \boldsymbol{\beta}),$$

$$\boldsymbol{T}_{nj}(\boldsymbol{\beta}) = \boldsymbol{V}_{n1}^{-1}(\boldsymbol{\beta}) n^{-1} \sum_{i=1}^{n} w_i^{(2-j)} \boldsymbol{x}_i \boldsymbol{x}_i^t \boldsymbol{V}_{n1}^{-1}(\boldsymbol{\beta}) \boldsymbol{V}_{n2}(\boldsymbol{\beta}) \boldsymbol{V}_{n1}^{-1}(\boldsymbol{\beta}) \boldsymbol{x}_i F^{(j)}(\boldsymbol{x}_i^t \boldsymbol{\beta}),$$

jossa yläindeksi (j) tarkoittaa j:nnettä derivaattaa ja edellä $w_i^{(2-j)}$ kuvaa $(2-j)$:ttä $w(\boldsymbol{u}, v)$:n derivaattaa v:n suhteen kohdassa $\boldsymbol{u} = \boldsymbol{x}_i$ ja $v = \boldsymbol{x}_i^t \boldsymbol{\beta}$. Yhtälön (3.1) ratkaisu $\hat{\boldsymbol{\beta}}$ on asymptoottisesti normaalijakautunut:

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \underset{as}{\sim} N(\boldsymbol{0}, \boldsymbol{V}_{n1}^{-1}(\boldsymbol{\beta}) \boldsymbol{V}_{n2}(\boldsymbol{\beta}) \boldsymbol{V}_{n1}^{-1}(\boldsymbol{\beta})). \tag{3.2}$$

Estimaattorin $\hat{\boldsymbol{\beta}}$ kovarianssimatriisille saadaan tarkentuva estimaatti ([5])

$$n^{-1} \boldsymbol{V}_{n1}^{-1}(\hat{\boldsymbol{\beta}}) \boldsymbol{V}_{n2}(\hat{\boldsymbol{\beta}}) \boldsymbol{V}_{n1}^{-1}(\hat{\boldsymbol{\beta}}).$$

3.1.1 Vipuarvojen alaspäin painotus

Oletetaan että $w_i = w(\boldsymbol{x}_i, \boldsymbol{x}_i^t \boldsymbol{\beta}) = w(\boldsymbol{x}_i)$, jolloin alaspäin painotus kohdistuu suoraan vipuarvoihin. Nyt $\frac{\partial}{\partial v} w(u, v) = 0$, joten $w^{(1)} = T_{n1} = 0$. Estimaattorin $\hat{\boldsymbol{\beta}}$ harhalle voidaan antaa arvio ([5])

$$E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \approx -(2n)^{-1} \{\boldsymbol{T}_{n2}(\boldsymbol{\beta}) + 2\boldsymbol{T}_{n1}(\boldsymbol{\beta})\}.$$

Esimerkiksi, kun $p = 1$, $\beta_0 = 0$ ja ottamalla huomioon, että $T_{n1} = 0$, niin edellinen tulos saadaan muotoon

$$E(\hat{\beta} - \beta) \approx -(2n)^{-1} \frac{n^{-1} \sum_{i=1}^{n} w_i x_i^3 F^{(2)}(x_i \beta) n^{-1} \sum_{i=1}^{n} w_i^2 x_i^2 F^{(1)}(x_i \beta)}{\{n^{-1} \sum_{i=1}^{n} w_i x_i^2 F^{(1)}(x_i \beta)\}^3}.$$

Poikkeavien havaintojen kohdalla voidaan todeta, että eri menetelmien välillä on eroja siinä, kuinka suuri harha mallissa lopulta on. Useimmiten esimerkiksi edellä esitettyä menetelmää käyttäen päästään SU-estimaattia vähemmän harhaiseen tulokseen. Varsinaisten lukujen tuottaminen jätetään tilasto-ohjelmiston tehtäväksi, joten menetelmiä onkin pyritty havainnollistamaan tutkielman seuraavissa osissa. Yleisesti voidaan todeta, että mallin saaminen robustimmaksi vähentää sen tehokkuutta. Toisin sanoen robustisuustarkastelu tapahtuu, ainakin osittain, mallin tehokkuuden kustannuksella. [5]

3.1.2 Ennusteiden alaspäin painotus

Seuraavaksi esitellään robusti estimointimenetelmä, joka perustuu tavallisesta logistisesta mallista johdettavaan luokitteluvirhemalliin (missclassification model). Luokitteluvirhemallissa kukin havaintoyksikkö luokitellaan väärin todennäköisyydellä ξ, jonka lisäksi $\boldsymbol{\beta}_{Mc}$ kuvaa estimoitavaa parametria. Logistisessa mallissa vastaavaa parametriestimaattia merkitään $\boldsymbol{\beta}$:lla. Tavallinen logistinen malli (ks. kpl 2.1) voidaan kirjoittaa muodossa

$$P(y = 1|\boldsymbol{x}) = F(\boldsymbol{x}^t\boldsymbol{\beta}),$$

jossa

$$F(v) = \{1 + \exp(-v)\}^{-1} = \frac{\exp(v)}{1 + \exp(v)}$$

on logistisen jakauman kertymäfunktio. Luokitteluvirhemalli on muotoa

$$P(y = 1|\boldsymbol{x}) = F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc}) + \xi\{1 - 2F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc})\} = G(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc}, \xi). \tag{3.3}$$

Kun $\xi = 0$, niin $P(y = 1|\boldsymbol{x}) = F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc})$ ja kun $\xi = 1$, niin $P(y = 1|\boldsymbol{x}) = 1 - F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc})$. Menetelmän yhtenä tarkoituksena on tarjota keinoja korjata otoskoon tuottamaa harhaa, jonka lisäksi sillä on käyttöä tilanteessa, jossa tarkastelun kohteena on luokitteluvirheen mahdollisuus. Mallia on aikaisemmin tutkittu pienillä ξ:n arvoilla ja keskeisimpiä havaintoja on ollut, että korjatun luokitteluvirhemallin suurimman uskottavuuden estimaatti on ainoastaan likimain tarkentuva, eikä tuloksia suurella otoskoolla ole saatavilla. Menetelmän suurimpana hyötynä onkin nimenomaan helposti tulkittava säätelyparametri [5]. Käyttämällä painofunktiota, joka on muotoa $w_i = w(\boldsymbol{x}_i, \boldsymbol{x}_i^t\boldsymbol{\beta}) = w(\boldsymbol{x}_i^t\boldsymbol{\beta})$, äärimmäiset sovitetut todennäköisyydet saadaan alaspäin painotettua. Esimerkki tällaisesta painofunktiosta on

$$w(\boldsymbol{x}^t\boldsymbol{\beta}) = [F(\boldsymbol{x}^t\boldsymbol{\beta})\{1 - F(\boldsymbol{x}^t\boldsymbol{\beta})\}]^c[F(\boldsymbol{x}^t\boldsymbol{\beta})^\lambda + \{1 - F(\boldsymbol{x}^t\boldsymbol{\beta})\}^\lambda],$$

jossa c ja λ ovat säätövakioita. Kun sijoitetaan (c, λ) arvoiksi $(0, 0)$ tai $(1, -1)$, saadaan tavallinen logistinen malli. Arvoilla $(1, 0)$ painot ovat suoraan verrannollisia y:n variansseihin annetuilla x:n arvoilla ja tällöin malli reagoi outliereihin, joilla on erittäin alhainen tai erittäin korkea ennustettu todennäköisyys. Tämän osoittaa seuraavaksi esiteltävä luokitteluvirhe-estimaatti. [5]

Palataan jo aikaisemmin esiteltyyn yhtälöön

$$P(y=1|\boldsymbol{x}) = F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc}) + \xi\{1 - 2F(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc})\} = G(\boldsymbol{x}^t\boldsymbol{\beta}_{Mc}, \xi), \tag{3.4}$$

jolloin SU-estimaatti $\hat{\boldsymbol{\beta}}_{Mc}$ on nk. M-estimaatti, joka saadaan ratkaistua yhtälöstä

$$\mathbf{0} = \sum_{i=1}^{n} w(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\boldsymbol{x}_i\{y_i - G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\},$$

jossa

$$w(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi) = (1-2\xi)F(\boldsymbol{x}_i^t\boldsymbol{\beta})\{1 - F(\boldsymbol{x}_i^t\boldsymbol{\beta})\}[G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\{1 - G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\}]^{-1}.$$

Logistisessa mallissa $\hat{\boldsymbol{\beta}}_{Mc}$ ei konvergoi kohti $\boldsymbol{\beta}_{Mc}$:ta, joka on määritelty mallissa (3.4), vaan kohti $\boldsymbol{\beta}_{Mc*}$:ta, joka on ratkaisu yhtälöön

$$\mathbf{0} = \lim_{n\to\infty}\left[n^{-1}\sum_{i=1}^{n} w_i(\boldsymbol{x}_i^t\boldsymbol{\beta}_{Mc*}, \xi)\boldsymbol{x}_i\{F(\boldsymbol{x}_i^t\boldsymbol{\beta}) - G(\boldsymbol{x}_i^t\boldsymbol{\beta}_{Mc*}, \xi)\}\right].$$

Logistisessa mallissa $\hat{\boldsymbol{\beta}}_{Mc}$ on ei-tarkentuva $\boldsymbol{\beta}$:n estimaattori, joten Copas (1988, [6]) ehdotti tilalle harhakorjattua versiota, joka on sopiva pienille ξ:n arvoille. Tämä harhakorjattu estimaatti on ei-tarkentuva logistisessa mallissa, mutta käytännössä siitä on huolta ainoastaan suurilla ξ:n arvoilla. [5]

Olkoon siis $\hat{\boldsymbol{\beta}}_{Mc}$ ratkaisu yhtälölle

$$\mathbf{0} = \sum_{i=1}^{n} w_i(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\boldsymbol{x}_i\{y_i - G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\}.$$

Kuten aikaisemmin on mainittu, estimaatti $\hat{\boldsymbol{\beta}}_{Mc}$ on tarkentuva $\boldsymbol{\beta}_{Mc}$:n estimaatti mallissa (3.4), mutta ei-tarkentuva $\boldsymbol{\beta}$:n estimaatti logistisessa mallissa. Tätä varten voidaan käyttää korjausta $\hat{\boldsymbol{\beta}}_{Mc}$:lle, joka on ainakin approksimatiivisesti tarkentuva logistisen mallin kohdalla, kun nimellinen luokitteluvirhetaso ξ on pieni [5]. Tämä estimaatti ei ole kuitenkaan tarkentuva yleisesti. Seuraaksi esitellään $\boldsymbol{\beta}$:n tarkentuva estimaatti, joka perustuu luokitteluvirhe-estimaattiin $\hat{\boldsymbol{\beta}}_{Mc}$ ja jolla on helposti tulkittava säätövakio. Yksinkertaisinta on käyttää luokitteluvirhe-estimaatista saatuja painoja $w(\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{Mc}, \xi)$. Määritellään $\hat{\boldsymbol{\beta}}$ ratkaisuksi yhtälölle

$$\mathbf{0} = \sum_{i=1}^{n} w(\boldsymbol{x}_i^t \hat{\boldsymbol{\beta}}_{Mc}, \xi)\boldsymbol{x}_i\{y_i - F(\boldsymbol{x}_i^t\boldsymbol{\beta})\}. \tag{3.5}$$

Yhtälöllä on yleensä yksikäsitteinen ratkaisu, koska se on estimoiva yhtälö painotetussa logistisessa regressiossa painoilla $w(\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{Mc}, \xi)$. Voidaan osoittaa, että $\hat{\boldsymbol{\beta}}$:n asymptoottinen jakauma on aiemmin esitetyn tuloksen (3.3) mukainen, kun painoina on $w_i = w(\boldsymbol{x}_i^t\hat{\boldsymbol{\beta}}_{Mc*}, \xi)$. Säätövakion voidaan nyt tulkita olevan luokitteluvirheen oletettu nimellisarvo. Yhtälön (3.5) voidaan tulkita myös korjaavan estimointiyhtälöä (kts. s.10)

$$\mathbf{0} = \sum_{i=1}^{n} w_i(\boldsymbol{w}_i^t\boldsymbol{\beta}, \xi)\boldsymbol{x}_i\{y_i - G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\},$$

jotta tämä saataisiin harhattomaksi logistisessa mallissa. Yhtälön (3.5) avulla voidaan myös selvittää mallin lopullisen estimaatin käyttäytymistä.

Edelleen sivulla 12 esitetyn yhtälön

$$w_i(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi) = (1-2\xi)F(\boldsymbol{x}_i^t\boldsymbol{\beta})\{1 - F(\boldsymbol{x}_i^t\boldsymbol{\beta})\}[G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\{1 - G(\boldsymbol{x}_i^t\boldsymbol{\beta}, \xi)\}]^{-1}$$

avulla voidaan todeta, että painot ovat pieniä silloin, kun $|\boldsymbol{x}_i^t\boldsymbol{\beta}|$ on suuri ja ξ vaikuttaa siihen, kuinka suuria sovitettujen todennäköisyyksien on oltava ennen kuin merkittävää alaspäin painotusta tapahtuu. Suurille ξ:n arvoille painot ovat lähes suoraan verrannollisia $F(\boldsymbol{x}^t\boldsymbol{\beta})\{1 - F(\boldsymbol{x}^t\boldsymbol{\beta})\}$:n suhteen. Lopuksi voidaan kuitenkin todeta, että ratkaisu $\hat{\boldsymbol{\beta}}$ yhtälölle (3.5) sekä luokitteluvirhe-estimaatti $\hat{\boldsymbol{\beta}}_{Mc}$ eivät ole tarkasti ottaen robusteja, koska niiden influenssifunktiot ovat rajoittamattomia. Ongelmia saattaa tulla silloin, kun $||\boldsymbol{x}_i||$ on suuri, mutta $|\boldsymbol{x}_i^t\boldsymbol{\beta}|$ pieni. [5]

3.1.3 Kvasi-uskottavuus estimaattori

Mallows-tyypin estimaattoreihin kuuluu myös tässä kappaleessa esiteltävä kvasi-uskottavuus estimaattori, josta tämän tutkielman yhteydessä käytetään lyhennettä Mqle (Mallows quasi-likelihood estimator, [4]). Kyseinen estimaattori kuuluu nk. M-estimaattoreihin, ja on ratkaisu seuraavalle estimointiyhtälölle

$$\sum_{i=1}^{n} \boldsymbol{\psi}(y_i, \mu_i) = \mathbf{0}, \tag{3.6}$$

jossa $\boldsymbol{\psi}(y,\mu) = \boldsymbol{\nu}(y,\mu)w(\boldsymbol{x})\frac{\partial}{\partial\boldsymbol{\beta}}\mu - \boldsymbol{a}(\boldsymbol{\beta})$ ja $\boldsymbol{\nu}(y_i,\mu_i) = \psi_c(r_i)\frac{1}{V^{1/2}(\mu_i)} = V(\mu_i) = \mathrm{Var}(y_i)$, jossa $r_i = \frac{y_i-\mu_i}{V^{1/2}(\mu_i)}$ on Pearsonin residuaali. Lisäksi on syytä tuoda muistutuksena esille tässäkin yhteydessä keskeinen tulos: $\eta_i = g(\mu_i) = \boldsymbol{x}_i^t\boldsymbol{\beta}$, jolloin $\mu_i = \mu_i(\boldsymbol{\beta}) = g^{-1}(\boldsymbol{x}_i^t\boldsymbol{\beta})$. Huberin funktio ψ_c on määritelty seuraavasti:

$$\psi_c(r) = \max(-c, \min(r,c)) = \begin{cases} r & , |r| \leq c, \\ c\,\mathrm{sign}(r) & , |r| > c. \end{cases}$$

Varsinainen Mallows kvasi-uskottavuus estimaattori saadaan ratkaisuna estimointiyhtälölle

$$\sum_{i=1}^{n}\left[\psi_c(r_i)w(\boldsymbol{x}_i)\frac{1}{V^{1/2}(\mu_i)}\frac{\partial}{\partial\boldsymbol{\beta}}\mu_i - \boldsymbol{a}(\boldsymbol{\beta})\right] = \mathbf{0}, \tag{3.7}$$

jossa $\boldsymbol{a}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} E[\psi_c(r_i)]w(\boldsymbol{x}_i)\frac{1}{V^{1/2}(\mu_i)}\frac{\partial\mu_i}{\partial\boldsymbol{\beta}}$. Korjaustermi $\boldsymbol{a}(\boldsymbol{\beta})$ takaa Fisher-tarkentuvuuden.

Lisäksi, kun tiedetään, että $p_i = \frac{\mu_i}{n_i}$ ja $\eta_i = \log\left(\frac{p_i}{1-p_i}\right)$, niin $\mu_i = \frac{n_i\exp(\eta_i)}{1+\exp(\eta_i)}$. Näin ollen lopullinen estimointiyhtälö logistiselle regressiolle on ([4])

$$\sum_{i=1}^{n}\left[\psi_c(r_i)w(\boldsymbol{x}_i)\sqrt{\frac{\mu_i(n_i-\mu_i)}{n_i}}\boldsymbol{x}_i - \boldsymbol{a}(\boldsymbol{\beta})\right] = \mathbf{0}. \tag{3.8}$$

3.2 Schweppe-tyyppiset estimaattorit

Scweppe-tyyppisten estimaattoreiden ([10]) yleinen määritelmä on

$$v(\boldsymbol{x}) = 1/w(\boldsymbol{x}) \quad \text{ja} \quad \eta(\boldsymbol{x},r) = w(\boldsymbol{x})\psi(r/w(\boldsymbol{x})),$$

joista saadaan

$$\sum_{i=1}^{n} w(\boldsymbol{x}_i)\psi((\boldsymbol{y}_i - \boldsymbol{x}_i^t T_n)/w(\boldsymbol{x}_i))\boldsymbol{x}_i = \mathbf{0}. \tag{3.9}$$

Edellisestä saadaan ratkaisuksi

$$d \cdot \psi_b(|d|(\boldsymbol{x}^t\boldsymbol{\beta}^{-1}\boldsymbol{x})^{1/2})/|d|(\boldsymbol{x}^t\boldsymbol{\beta}^{-1}\boldsymbol{x})^{1/2} = \frac{d}{|d|}(\boldsymbol{x}^t\boldsymbol{\beta}^{-1}\boldsymbol{x})^{-1/2}\psi_b(|d|(\boldsymbol{x}^t\boldsymbol{\beta}^{-1}\boldsymbol{x})^{1/2}). \quad (3.10)$$

Erityisesti tässä tutkielmassa tarkasteltava Schweppe-tyyppisen estimaattorin erikoistapaus, jota kutsutaan harhattomaksi estimaatiksi ja jolla on rajoitettu influenssifunktio (Conditionally unbiased influence function, CUBIF), voidaan esittää muodossa

$$\sum_{i=1}^{n} \boldsymbol{\psi}_{cond}(\boldsymbol{y_i}, \boldsymbol{x_i}, \boldsymbol{\beta}, \boldsymbol{B}) = \boldsymbol{0}, \quad (3.11)$$

jossa

$$\boldsymbol{\psi}_{cond}(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B}) = d(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B})w_b(|d(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B})|(\boldsymbol{x}^t\boldsymbol{B}^{-1}\boldsymbol{x})^{1/2})\boldsymbol{x}, \quad (3.12)$$

kun $w_b(a) = \psi_b(a)/a$ ja ψ_b on Huberin funktio

$$d(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B}) = \boldsymbol{y} - F(\boldsymbol{x}^t\boldsymbol{\beta}) - c(\boldsymbol{x}^t\boldsymbol{\beta}, b/(\boldsymbol{x}^t\boldsymbol{B}^{-1}\boldsymbol{x})^{1/2}), \quad (3.13)$$

jossa $F(v) = \frac{\exp(v)}{1+\exp(v)}$ ja on muotoa $w_2(|d(\boldsymbol{y},\boldsymbol{x}, \boldsymbol{\beta})|)$. Schweppe-tyypin estimaattorit ovat yhteydessä myös Mallows-tyypin estimaattoreiden tuloksiin, mutta suurin ero perustuu pistemääräfunktioon (3.12), jossa w_b faktoroituu kahteen osaan. Ensimmäinen osa riippuu ainoastaan $\boldsymbol{x}$:stä ja on muotoa $w_1((\boldsymbol{x}^t\boldsymbol{B}^{-1}\boldsymbol{x})^{1/2})$. Toinen osa riippuu funktiosta $d(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B})$ ja on muotoa $w_2(|d(\boldsymbol{y}, \boldsymbol{x}, \boldsymbol{\beta}, \boldsymbol{B})|)$.

Yhtälössä (3.13) $c(\cdot, \cdot)$ on korjaustermi, jonka avulla $\hat{\boldsymbol{\beta}}$ saadaan ehdollisesti Fisher-tarkentuvaksi ja matriisi $\boldsymbol{B}$ valitaan siten, että itsestandardoitu sensitiivisyys (self-standardized sensitivity) saa arvon b. On kuitenkin huomattava, että estimaattori $\hat{\boldsymbol{\beta}}$ on ehdollisesti Fisher-tarkentuva ja sillä on rajoitettu influenssi kaikilla matriisin $\boldsymbol{B}$ arvoilla. Korjaustermi voidaan esittää myös muodossa

$$c(r, s) = \begin{cases} sF(r)/(1 - F(r)) - F(r), & \text{kun } r < 0, s < 1 - F(r), \\ 1 - F(r) - s(1 - F(r))/F(r), & \text{kun } r > 0, s < F(r), \\ 0, & \text{muulloin.} \end{cases} \quad (3.14)$$

Itsestandardoitu sensitiivisyys on b^2 (eli $s(\boldsymbol{\psi}_{cond}) = b$, jossa $\boldsymbol{\psi}_{cond}$ seuraa yhtälöstä 3.12) ja $s(\boldsymbol{\psi})^2 = \sup_{\boldsymbol{y},\boldsymbol{x}} \sup_{\boldsymbol{\lambda}\neq\boldsymbol{0}} \frac{(\boldsymbol{\lambda}^t IF_\psi)^2}{\boldsymbol{\lambda}^t V(\psi)\boldsymbol{\lambda}}$, jossa $V(\boldsymbol{\psi}) = V(\boldsymbol{\psi},\boldsymbol{\beta}) = E_{\boldsymbol{\beta}}[IF_\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta})IF_\psi(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta})^t]$. Influenssifunktio IF noudattaa sivun 21 määritelmää estimaattorille T eli $IF(w,T,H_0) = \lim_{\varepsilon\to 0} \frac{T((1-\varepsilon)H_0+\varepsilon\Delta_w)-T(H_0)}{\varepsilon}$, jossa H_0 on mallin jakauma ja $\Delta_{\boldsymbol{w}}$ on Diracin jakauma, jonka koko todennäköisyysmassa on pisteessä $\boldsymbol{w} = (\boldsymbol{x}^t, y)^t$, jossa $\boldsymbol{x} \in \mathbb{R}^{p+1}$ ja $y \in \{0,1\}$. Nyt nähdään, että

$$E_{\boldsymbol{\beta}}[\boldsymbol{\psi}_{cond}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{B})\boldsymbol{\psi}_{cond}(\boldsymbol{y},\boldsymbol{x},\boldsymbol{\beta},\boldsymbol{B})^t] = \boldsymbol{B}. \tag{3.15}$$

Koska aina pätee, että $s(\boldsymbol{\psi})^2 \geq p+1$, niin välttämättä täytyy olla $b^2 \geq p+1$, jotta yhtälöllä (3.15) olisi ratkaisu. [10]

3.3 Biancon & Yohain estimaattori

3.3.1 Painottamaton Biancon & Yohain estimaattori

Seuraavaksi esitellään menetelmä, joka osittain perustuu jo johdannossakin mainittuun tappiofunktioon ρ (loss-function). Myös tämän estimaattorin painotettu versio ([7]) esitellään seuraavissa kappaleissa.

Olkoot Y_i, $i = 1,...,n$, riippumattomia Bernoulli-jakautuneita muuttujia, joiden onnistumistodennäköisyydet ovat

$$P(Y_i = 1|\boldsymbol{x_i}) = F(\boldsymbol{x}_i^t\boldsymbol{\beta}),$$

jossa F on aidosti kasvava kertymäfunktio. Tässä yhteydessä ja esiteltäessä menetelmiä seuraavissa kappaleissa, käytetään jo sivulla 8 esitettyä logistisen jakauman kertymäfunktiota ([7])

$$F(v) = 1/(1+\exp(-v)) = \frac{\exp(v)}{1+\exp(v)}.$$

Olkoon $\hat{\boldsymbol{\beta}}$ otoksesta $X_n = \{(\boldsymbol{x}_1, y_1),...,(\boldsymbol{x}_n, y_n)\}$ laskettu $\boldsymbol{\beta}$:n estimaattori. Suurimman uskottavuuden estimaattori $\hat{\boldsymbol{\beta}}^{ML}$ on määritelty seuraavalla tavalla:

$$\hat{\boldsymbol{\beta}}_n^{ML} = \arg\max_{\boldsymbol{\beta}} \log L(\boldsymbol{\beta}; X_n) = \arg\min_{\beta} \sum_{i=1}^{n} d(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i), \tag{3.16}$$

jossa $\log L(\boldsymbol{\beta}, X_n)$ on ehdollisen log-uskottavuusfunktion arvo kohdassa $\boldsymbol{\beta}$ ja $d(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i)$ on mallin devianssikomponentti, joka on muotoa

$$d(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i) = -y_i \log F(\boldsymbol{x}_i^t \boldsymbol{\beta}) - (1 - y_i) \log \{1 - F(\boldsymbol{x}_i^t \boldsymbol{\beta})\}.$$

Kuten jo aikaisemminkin on todettu, SU-estimaattori on asymptoottisesti kaikista tehokkain estimaattori, mutta aineistossa mahdollisesti mukana olevilla outliereilla voi olla suuri vaikutus estimaattorin arvoihin. Nyt esiteltävä lähestymistapa tarjoaa ratkaisuksi yhtälössä (3.16) olevan funktion d korvaamista jollakin muulla funktiolla. Estimaattori $\hat{\boldsymbol{\beta}}$ voidaan nyt määritellä seuraavasti:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \varphi(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i), \tag{3.17}$$

jossa φ on positiivinen ja lähes kaikkialla differentioituva funktio. Funktion φ pitää täyttää ehto $\varphi(s; 0) = \varphi(-s; 1)$ kaikilla pistemäärillä s, jossa pistemäärä $s = \boldsymbol{x}_i^t \boldsymbol{\beta}$ on parametrivektorin $\boldsymbol{\beta}$ lineaarikombinaatio. Tulevissa esityksissä funktion φ sijasta käytetään funktiota $\phi(s) = \varphi(s; 0)$. Havaintoarvoa $y = 0$ vastaava funktion $\phi(s)$ arvo antaa pistemäärän s vaikutuksen yhtälössä (3.17) minimoitavaan objektifunktioon. Funktion oletetaan olevan ei-laskeva. Näinollen, niiden tulisi saada suurempi paino funktiota minimoitaessa. Vaatimuksena on edelleen, että $\lim_{s \to -\infty} \phi(s) = 0$, josta seuraa, että suurella negatiivisella pistemäärällä ei ole suurta vaikutusta minimoitavaan objektifunktioon.

Kuten yhtälöstä (3.17) voidaan päätellä, estimaatti $\hat{\boldsymbol{\beta}}$ kuuluu nk. M-estimaatteihin, joilla on ensimmäisen kertaluvun ehtona minimikohdalle

$$\frac{1}{n} \sum_{i=1}^{n} \Psi(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i) x_i = 0, \tag{3.18}$$

jossa $\Psi(s; 0) = \partial \varphi(s; 0) / \partial s$ ja $\Psi(s; 1) = -\Psi(-s; 0)$. Täytyy kuitenkin huomata, että ensimmäisen kertaluvun ehdot saattavat saada useita eri ratkaisuja, jolloin joudutaan

käyttämään apuna yhtälöä (3.17), jotta saataisiin valittua lopullinen estimaatti. Tulevissa luvuissa käytetään merkintää $\psi(s) = \Psi(s;0) = \phi^{'}(s)$.

Yhtälön (3.17) erikoistapauksia on käsitelty kirjallisuudessa aiemmin, joista jatkon kannalta tärkeimmät esitellään seuraavaksi. SU-estimaattori kuuluu M-estimaattoreiden luokkaan, jossa $\phi_{ML}(s) = -\log(1 - F(s))$. Pregibon esitteli estimaattorista robustimman version, joka on muotoa

$$\hat{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \lambda(d(\boldsymbol{x}_i^t\boldsymbol{\beta}; y_i)),$$

jossa λ on aidosti kasvava Huberin funktio. Tämän estimaattorin tarkoituksena oli antaa vähemmän painoa havainnoille, joilla on mallin estimaatteihin vain lievä vaikutus. Se ei kuitenkaan tarpeeksi painottanut alaspäin malliin voimakkaasti vaikuttaneita, muista poikkeavia havaintoja. Tämä luonnollisesti heikensi mallin käytettävyyttä. [14]

Myöhemmin Bianco & Yohai ([7]) kehittivät tarkentuvan ja robustimman version Pregibonin estimaattorista käyttämällä rajoitettua funktiota ρ, jolloin estimaattoriksi saadaan

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \sum_{i=1}^{n} \{\rho(d(\boldsymbol{x}_i^t\boldsymbol{\beta})) + C(\boldsymbol{x}_i^t\boldsymbol{\beta})\}, \tag{3.19}$$

jossa $C(\boldsymbol{x}_i^t\boldsymbol{\beta})$ on harhaa korjaava termi, joka saadaan kun

$$C(s) = G(F(s)) + G(1 - F(s)) - G(1),$$

jossa

$$G(t) = \int_0^t \rho^{'}(-\log v)dv.$$

Termin $G(1)$ vähentäminen Biancon & Yohain alkuperäisestä harhan korjauksesta on tarpeen, mikäli halutaan pysyä yhtälön (3.17) mukaisessa määrittelyssä. Merkitään Biancon & Yohain estimaattoria tästä eteenpäin notaatiolla BY. Estimaattorin yhteydessä voidaan esittää funktio ϕ, joka on muotoa

$$\phi_{BY}(s) = \rho(-\log(1 - F(s))) + G(F(s)) + G(1 - F(s)) - G(1), \tag{3.20}$$

ja joka menee nollaan, kun $s \to -\infty$. BY-estimaattorin muodosta nähdään, että funktio ϕ_{BY} riippuu ainoastaan ρ-funktion valinnasta. Biancon & Yohain alkuperäisessä artikkelissa ([2]) BY-estimaattorissa käytetään seuraavaa ρ-funktiota:

$$\rho(t) = \begin{cases} t - \frac{t^2}{2c} & \text{jos } t \leq c, \\ \frac{c}{2} & \text{muulloin,} \end{cases} \tag{3.21}$$

jossa c on säätöparametri. BY-estimaattorin voidaan yleisesti todeta olevan tiettyjen ehtojen vallitessa tarkentuva ja asymptoottisesti normaalijakautunut, joka on lisäksi tehokas [2]. Tehtyjen numeeristen kokeiden yhteydessä voitiin pääosin vahvistaa tulosten oikeellisuus, mutta lisäksi havaittiin, että tappio-funktiota (3.21) käytettäessä BY-estimaattori ei ollut kaikissa tapauksissa olemassa. Vaikka kyseessä olisi otos ilman outliereita, saattoi käydä niin, että BY-estimaattorin arvot kohosivat lähes äärettömään, johtuen siitä että yhtälön (3.17) minimi saavutettiin parametriavaruuden reunalla. Ratkaisuksi onkin esitetty ρ-funktiota, joka takaa BY-estimaattorin olemassaolon aina silloin, kun SU-estimaattori on olemassa. [7]

Tarkastellaan nyt M-estimaattoreiden olemassaoloon liittyvää ongelmaa. On osoitettu, että SU-estimaattori on olemassa vain silloin, kun havaintojen välistä päällekäisyyttä on havaittavissa, kun $y_i = 0$ ja $y_i = 1$. Päällekäisyydellä tarkoitetaan tässä tapauksessa sitä, että hypertason avulla ei voida erottaa selittäviä muuttujia ryhmiin, joissa toisessa on havainnot, kun $y_i = 1$ ja toisessa havainnot, kun $y_i = 0$. Tilanne voidaan muotoilla siten, että $I^1 = \{i \in \{1, ..., n\} | y_i = 1\}$ on havainnot, joille $y_i = 1$ ja $I^0 = \{i \in \{1, ..., n\} | y_i = 0\}$ tämän komplementti. Tällöin ei päde tilanne, jossa $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ ja

$$\boldsymbol{x}_i^t \boldsymbol{\beta} \geq 0 \qquad \forall i \in I^1 \qquad \& \qquad \boldsymbol{x}_i^t \boldsymbol{\beta} \leq 0 \qquad \forall i \in I^0. \tag{3.22}$$

Erityisesti tämä muoto poissulkee tilanteen, jossa kaikki y_i ovat samoja.

Muidenkin tarkasteltavien estimaattoreiden kohdalla pätee lause 1, jonka mukaan sama olemassaoloehto, kattaen joitakin rajoituksia funktiolle $\psi = \phi'$, pätee ensimmäisen kertaluvun muodolle. On kuitenkin huomioitava, että ϕ:n ollessa ei-kasvava on ψ:n oltava positiivinen funktio. [7]

Lause 1. *Olkoon* φ*:*$\mathbb{R}^2 \to \mathbb{R}$ *positiivinen funktio ja* $\phi(s) = \phi(s;0) = \phi(-s;1)$*. Oletetaan että* ϕ *on ei-kasvava ja jatkuva funktio, jolla on jatkuva derivaatta* ψ*, jolloin* $\lim_{s\to-\infty}\phi(s) = 0$*. Olkoon logistisen regressiomallin estimaattori* $\hat{\boldsymbol{\beta}}$*, joka on määritelty yhtälössä (3.17). Oletetaan, että kolme seuraavaa ehtoa pätevät.*

(1) *Otoksessa esiintyy päällekäisyyttä.*
(2) *On olemassa sellainen* $L_0 > 0$*, että* ψ *kasvaa välillä* $]-\infty, -L_0]$ *ja joko laskee tai kasvaa välillä* $[L_0, +\infty[$*.*
(3) $\lim_{s\to\infty}\psi(st)/\psi(-s) = \infty \quad \forall t > 0.$

Tällöin estimaattori $\hat{\boldsymbol{\beta}}$ *on olemassa ja äärellinen.*

Ehdon (1) täyttyminen tarvitaan myös SU-estimaattorin olemassaoloon. Ehto (2) kattaa kaksi ψ-funktion tapaa käyttäytyä. Se joko kasvaa kuten SU-estimaattorin tapauksessa tai se voi olla redescending-funktio kuten BY-estimaattorin tapauksessa. ψ-funktion muoto määrittelee myöhemmin esiteltävän influenssifunktion muodon. [1]

Ehto (3) toteutuu kasvavilla ψ-funktioilla, kunhan ϕ toteuttaa ehdon $\lim_{s\to-\infty}\psi(s) = 0$. Redescending-funktion tapauksessa on toteuduttava se, että ψ-funktion on uudelleen laskettava kohti nollaa nopeammin oikein luokiteltujen havaintojen ($s < 0$) puolella kuin virheellisesti luokiteltujen havaintojen ($s > 0$) puolella. Nyt kyseessä oleva BY-estimaattori, joka käyttää funktiota (3.21) tappio-funktionaan, ei täytä ehtoa (3). ρ-funktion derivaatta, joka on muotoa

$$\rho'(t) = \begin{cases} 1 - \frac{t}{c} & \text{jos } t \le c, \\ 0 & \text{muulloin,} \end{cases}$$

katoaa kokonaan argumenttiensa arvojen ollessa tarpeeksi suuria. Väärin luokitellut havainnot ovat taipuvaisia tulemaan herkästi alaspäin painotetuiksi, vaikka nämä havainnot saattavat sisältää suurimman osan informaatiosta käytettäessä logistista regressiota. Tätä tilannetta varten voidaan ottaa käyttöön toisenlainen ρ-funktio. Ajatuksena on valita funktio, jonka derivaatta on muotoa

$$\rho'(t) = \begin{cases} e^{-\sqrt{d}} & \text{jos } t \le d, \\ e^{-\sqrt{t}} & \text{muulloin,} \end{cases} \tag{3.23}$$

annetulle vakiolle d. Vakio d tulisi valita siten, että saavutetaan tasapaino robustisuuden ja tehokkuuden välillä: kun d kasvaa, myös tehokkuus kasvaa mutta robustisuus heikkenee ja päinvastoin. Croux & Haesbroeck ([7]) ehdottavat käytettäväksi arvoa $d = 0,5$. Funktioilla ρ ja G, jotka määriteltiin funktiossa (3.21) on analyyttiset muodot

$$\rho(t) = \begin{cases} te^{-\sqrt{d}} & \text{jos } t \leq d, \\ -2e^{-\sqrt{t}}(1+\sqrt{t}) + e^{-\sqrt{d}}(2(1+\sqrt{d})+d) & \text{muulloin,} \end{cases} \tag{3.24}$$

ja

$$G(t) = \begin{cases} te^{-\sqrt{-\log t}} + e^{1/4}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2}+\sqrt{-\log t})) - e^{-1/4}\sqrt{\pi} & \text{jos } t \leq e^{-d}, \\ e^{-\sqrt{d}}t - e^{-1/4}\sqrt{\pi} + e^{1/4}\sqrt{\pi}\Phi(\sqrt{2}(\frac{1}{2}+\sqrt{d})) & \text{muulloin,} \end{cases}$$

jossa Φ on standardoidun normaalijakauman kertymäfunktio. Jatkossa tämänkin tutkielman osalta käytetään edellä esityttyä muotoa BY-estimaattorista, jonka etuihin kuuluu se, että se on aina olemassa otoksilla, joissa esiintyy havaintojen välistä päällekäisyyttä. [7]

Edellä näytettiin kuinka ψ-funktion käyttäytyminen on tärkeää M-estimaattorin olemassaolon kannalta. Kyseinen funktio myös määrittelee muodon influenssifunktiolle. Tehdään selväksi, että influenssifunktio estimaattorille T kohdassa H_0, jossa H_0 on mallin jakauma, on määritelty siten, että

$$IF(w, T, H_0) = \lim_{\varepsilon \to 0} \frac{T((1-\varepsilon)H_0 + \varepsilon\Delta_w) - T(H_0)}{\varepsilon}, \tag{3.25}$$

jossa $\Delta_{\boldsymbol{w}}$ on Diracin jakauma, jonka koko todennäköisyysmassa on pisteessä $\boldsymbol{w} = (\boldsymbol{x}^t, y)^t$, jossa $\boldsymbol{x} \in \mathbb{R}^{p+1}$ ja $y \in \{0, 1\}$. M-tyypin estimaattoreille (määritelty yhtälössä (3.17) ja (3.18)) influenssifunktioksi saadaan

$$IF(w, T, H_0) = -E_{H_0}\left[\left.\frac{\partial^2\varphi(t;y)}{\partial t^2}\right|_{t=\boldsymbol{x}^t\boldsymbol{\beta}_0} \boldsymbol{x}\boldsymbol{x}^t\right]^{-1} \Psi(\boldsymbol{x}^t\boldsymbol{\beta}_0; y)\boldsymbol{x}. \tag{3.26}$$

Yhtälössä (3.26) on mainittavaa se, että vakiomatriisin oikeanpuoleinen osa jakautuu edelleen kahdeksi osaksi, joista toinen $\Psi(\boldsymbol{x}^t\boldsymbol{\beta}_0; y)$ riippuu ainoastaan pistemäärän $s = \boldsymbol{x}^t\boldsymbol{\beta}_0$ saamasta arvosta sekä selitettävästä muuttujasta. Toinen osa on kovariaatin arvo $\boldsymbol{x}$.

Logistisessa regressiomallissa esiintyvät outlierit voivat esiintyä erilaisissa yhteyksissä. Analogia lineaariseen regressioon saadaan, kun havaintoa $\boldsymbol{w} = (\boldsymbol{x}, y)$ kutsutaan vipupisteeksi silloin, kun $\boldsymbol{x}$ on outlier kovariaattiavaruudessa. Y-suuntainen outlier on havainto, joka ei ole vipupiste, mutta jonka residuaali $y - F(\boldsymbol{x}^t\boldsymbol{\beta}_0)$ on itseisarvoltaan suuri. Lisäksi

vipupisteet voivat olla sekä "hyviä"että "huonoja", riippuen siitä, onko niillä suuri vai pieni residuaali suhteessa käytettyyn sovitukseen.

SU-estimaattorissa Ψ-funktio on muotoa $\Psi(s, y) = F(s) - y$, jolloin y-suuntaisten outliereiden vaikutus SU-estimaattoriin on näin ollen rajoitettu, koska aina pätee että $|F(\boldsymbol{x}^t\boldsymbol{\beta}_0) - y| \leq 1$. Hyvä vipupiste reagoi poikkeavaan $\boldsymbol{x}$:n arvoon pienellä residuaalin $F(\boldsymbol{x}^t\boldsymbol{\beta}_0) - y$ arvolla. Kun $\boldsymbol{x}$ kasvaa, niin residuaalin arvo $F(\boldsymbol{x}^t\boldsymbol{\beta}_0) - y$ tyypillisesti lähestyy nollaa nopeammin kuin $\boldsymbol{x}$:n arvo ääretöntä, jolloin vaikutus jää pieneksi. Voidaan näin ollen sanoa, että hyvät vipuarvot ovat SU-estimaattorille harmittomia, mutta huonot vipuarvot saavat influenssifunktion arvot kasvamaan suureksi. Esimerkiksi, kun tarkastellaan virheellisesti luokiteltua havaintoa, jolloin $\boldsymbol{x}^t\boldsymbol{\beta}_0(\boldsymbol{y}_0 - 0.5) < 0$, ja annetaan $\boldsymbol{x}$:n pituuden $\|\boldsymbol{x}\|$ kasvaa kohti ääretöntä, niin influenssifunktion arvo lähenee ääretöntä. [7]

3.3.2 Painotettu Biancon & Yohain estimaattori

Jotta saataisiin kaikkialla rajoitettu influenssifunktio, voidaan alaspäin painotettaviin vipuarvoihin lisätä painotuskerroin. Aluksi on pyrittävä tunnistamaan x-suuntaiset outlierit. Ratkaisuksi tähän on esitetty Mahalanobis-etäisyyden laskemista kullekin havainnolle. Etäisyys perustuu keskiarvovektoriin $\overline{\boldsymbol{x}}$ sekä kovarianssimatriisiin $\boldsymbol{C}$. Mahalanobis-etäisyydeksi saadaan tällöin $MD_i = \{(\boldsymbol{x}_{i(-1)} - \overline{\boldsymbol{x}}_{\boldsymbol{(-1)}})^t \boldsymbol{C}^{-1} (\boldsymbol{x}_{i(-1)} - \overline{\boldsymbol{x}}_{\boldsymbol{(-1)}})\}^{1/2}$, jossa $\boldsymbol{x}_{i(-1)} = (x_{i1}, ..., x_{ip})^t$ ja $\overline{\boldsymbol{x}}_{(-1)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_{i(-1)}$. Tämä menetelmä ei välttämättä kuitenkaan ole robusti, koska $\overline{\boldsymbol{x}}_{(-1)}$ ja $\boldsymbol{C}$ ovat erittäin herkkiä poikkeaville havainnolle. Tämän vuoksi näiden tilalle onkin ehdotettu sijaintiin ja sirontaan perustuvia robusteja estimaattoreita. MCD-estimaattori (Minimum Covariance Determinant) etsii h havainnon osajoukon, jolla osajoukkoa vastaava kovarianssimatriisin determinantti minimoituu. Näin ollen kyseisestä optimaalisesta osajoukosta lasketut klassiset estimaatit antavat MCD-estimaatit sijainnille ja sironnalle. Tässä tapauksessa $h = \lfloor 3n/4 \rfloor$ tuottaa 25% murtumispiste-estimaattorin. Lisäksi MCD-estimaattorilla on normaali konvergointinopeus ($n^{1/2}$) ja kohtuullinen tehokkuus. Merkitään 'robustia Mahalanobis' -etäisyyttä RD_i:llä, jossa $i = 1, ..., n$. [7,8,15]

BY-estimaattorin painotettu versio saa siis muodon

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \omega_i \varphi_{BY}(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i), \tag{3.27}$$

jossa funktio φ_{BY} on sama kuin BY-estimaattorissa ja painot on laskettu laskevan funktion W avulla. Nyt $\omega_i = W(RD_i)$ on robustin etäisyyden funktio, jossa

$$W(t) = \begin{cases} 1 & \text{jos } t^2 \leq \chi^2_{p,0,975}, \\ 0 & \text{muulloin,} \end{cases} \tag{3.28}$$

jossa $\chi^2_{p,0,975}$ on χ^2_p -jakauman $97,5\%$ kvantiilipiste. Yhtälön (3.27) ratkaisua kutsutaan nyt WBY-estimaattoriksi. Se säilyttää tarkentuvuutensa ilman jakaumiin liittyviä lisäoletuksia, koska painotus perustuu yksinomaan $\boldsymbol{x}$-arvoihin. Seuraavan apulauseen avulla saadaan influenssifunktio painotetuille versioille M-estimaattoreista. Olkoon

$$D_G(\boldsymbol{x}_{(-1)}) = \sqrt{(\boldsymbol{x}_{(-1)} - \boldsymbol{\mu}(G))^t \boldsymbol{\Sigma}(G)^{-1} (\boldsymbol{x}_{(-1)} - \boldsymbol{\mu}(G))}, \tag{3.29}$$

jossa $\boldsymbol{x}_{(-1)} = (x_1, ..., x_p)^t$ ja $(\boldsymbol{\mu}(G), \boldsymbol{\Sigma}(G))$ ovat sijainti- ja skaalafunktionaalit jakaumalle G. [7]

Apulause 1. *Olkoon T_W painotettu M-tyypin funktionaali, joka toteuttaa ehdon*

$$E_H[h(x, y, T_W(H))x \quad W(D_G(x))] = 0 \tag{3.30}$$

jossa $(x_{(-1)}, y) \sim H, x_{(-1)} \sim G, x = (1, x_{(-1)}^t)^t, h : \mathbb{R}^p \times \mathbb{R} \times \Gamma \to \mathbb{R}$, jossa Γ on avoin Euklidisen parametriavaruuden osa. Oletetaan, että T_W on määritelty riittävän suuressa jakaumaluokassa. Olkoon H_0 mallin jakauma ja $x_{(-1)} \sim G_0$ ja $(x_{(-1)}, y) \sim H_0$. Asetetaan $\beta_0 := T_W(H_0)$ ja määritellään

$$\Omega(a, \beta) = E_{H_0}[h(\boldsymbol{x}, y, \boldsymbol{\beta})|\boldsymbol{x}_{(-1)} = \boldsymbol{a}]$$

kaikilla $\boldsymbol{a} \in \mathbb{R}^p \quad ja \quad \boldsymbol{\beta} \in \Gamma$. Jos seuraavat ehdot täyttyvät:

(1) *h on melkein kaikkialla differentioituva,*
(2) *Yhtälössä (3.29) käytettyjen sijainti- ja sirontafunktionaalien influenssifunktiot ovat olemassa G_0:ssa,*
(3) *Ehdollisen Fisher-tarkentuvuuden ominaisuudet pätevät H_0:lla t.s. $\Omega(\boldsymbol{a}, \boldsymbol{\beta}_0) = 0 \quad \forall \boldsymbol{a}$,*
(4) *Jakaumalla G_0 on differentioituva tiheysfunktio,*
(5) *Painofunktio W on rajoitettu ja melkein kaikkialla jatkuva,*

niin tällöin

$$IF(\boldsymbol{w}, T_W, H_0) = -E_{H_0}\left[X\frac{\partial h(\boldsymbol{x}, y, \boldsymbol{\beta})}{\partial \beta}\bigg|_{\beta_0} W(D_{G_0}(\boldsymbol{x}_{(-1)}))\right]^{-1} h(\boldsymbol{x}, y, \boldsymbol{\beta}_0) z W(D_{G_0}(\boldsymbol{x}_{(-1)})),$$

jossa $\boldsymbol{w} = (\boldsymbol{x}^t, y)^t$.

Ehto (4) vaatii selittäjien olevan jatkuvia, mikä on melko luonnollista, sillä painot ovat laskettu käyttämällä mallimatriisin $\boldsymbol{X}$ rivejä. On huomattava myös, että G_0:n ei tarvitse olla symmetrinen. Itse asiassa ehtoa (4) voidaan lieventää siten, oletetaan ainoastaan, että jakaumalla on olemassa jatkuva tiheysfunktio ympäristössä $\{\boldsymbol{x}_{(-1)} \in \mathbb{R}^p : \mathrm{W}$ on ei-differentioituva kohdassa $D_{G_0}(\boldsymbol{x}_{(-1)})\}$. Tärkein ehto on (3), jota kutsutaan ehdolliseksi Fisher-tarkentuvuudeksi.

Apulause 1 on sovellettevissa myös lineaariselle regressiolle, mutta tässä yhteydessä keskitytään vain BY:n logistisen regression estimaattoriin. Näin ollen, oletetaan että $h(\boldsymbol{x}, y, \boldsymbol{\beta}) = \Psi_{BY}(\boldsymbol{x}^t\boldsymbol{\beta}, y)$. Nähdään nyt, että

$$E_{H_0}[\Psi_{BY}(\boldsymbol{x}^t\boldsymbol{\beta}, y)|\boldsymbol{x}_{(-1)}] = 0,$$

jolloin ehdollinen Fisher-tarkentuvuus pätee BY-estimaattorille. Täten apulauseen 1:n seurauksena saadaan

Lause 2. *Olkoon T_W WBY-funktionaali. Oletetaan, että φ_{BY} on lähes kaikkialle kahdesti differentioituva ja, että apulauseen 1 ehdot 3, 4 ja 5 pätevät. Näinollen mallijakauman H_0 kohdalla saadaan tulokseksi*

$$IF(w, T_W, H_0) = -E_{H_0}\left[\left.\frac{\partial^2\varphi_{BY}(t;y)}{\partial t^2}\right|_{t=x^t\beta_0} W(D_{G_0}(\boldsymbol{x}_{(-1)}))\boldsymbol{x}\boldsymbol{x}^t\right]^{-1} W(D_{G_0}(\boldsymbol{x}_{(-1)}))\Psi(\boldsymbol{x}^t\boldsymbol{\beta}_0; y)\boldsymbol{x},$$

jossa $\boldsymbol{w} = (\boldsymbol{x}_{(-1)}^t, \boldsymbol{y})^t$, $\boldsymbol{x} = (1, \boldsymbol{x}_{(-1)}^t)^t$ ja $x \sim G_0$ kun $((1, \boldsymbol{x}_{(-1)}^t), y) \sim H_0$.

Voidaan huomata, että WBY:n influenssifunktio ei riipu valituista sijainti- tai sirontafunktionaaleista, joita käytettiin etäisyyksien laskemisessa yhtälön (3.29) yhteydessä. Kuten voidaan odottaa, WBY:n influessifunktion tärkein ominaisuus on se, että se on rajattu, koska yhtälöä (3.28) vastaava painofunktio menee nollaan suurten vipupisteiden kohdalla. Tietenkin oheinen painotustilanne myös vähentää hyvien vipuarvojen painoa, mikä taas ei ole välttämättä tarpeellista, ja mikä johtaa tehokkuuden vähenemiseen. [7]

3.4 Painotettu suurimman uskottavuuden estimaattori

Lause 2 ei ole validi ainoastaan WBY-estimaattorille, vaan myös jokaiselle painotetulle versiolle estimaattorista, joka on saatu yhtälöstä (3.17), kunhan ne ovat ehdollisesti Fisher-konsistentteja. Esimerkiksi painotus voidaan tehdä myös tavalliselle SU-estimaattorille käyttämällä yhtälöä (3.28), jolloin tulokseksi saadaan painotettu suurimman uskottavuuden estimaattori (WML). Tämä estimaattori on helppo laskea ja sillä on rajoitettu influenssifunktio. Suurin ero on siinä, että WML-estimaattori on vähemmän resistantti y-suuntaisille outliereille kuin BY tai WBY-estimaattorit ([7]). Tässä tapauksessa estimaattori $\hat{\boldsymbol{\beta}}$ voidaan määritellä seuraavasti:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} w_i d(\boldsymbol{x}_i^t \boldsymbol{\beta}; y_i). \tag{3.31}$$

Luku 4

Tulosten havainnollistaminen esimerkkiaineiston avulla

4.1 Aineiston esittely

Edellä esitettyjä tuloksia pyritään seuraavaksi havainnollistamaan logistiselle regressiolle tarkoitetun esimerkkiaineiston avulla. Aineistosta on tarkoitus saada laskettua oleellisia lukuja, kuten regressiomalli ja keskivirhe. Tarkoitus on vertailla tavallisen logistisen mallin antamia lukuja sellaisiin lukuihin, jotka on saatu robustisuusestimaattorilla varustetusta logistisesta mallista. Tuloksia pyritään havainnollistamaan outliereiden avulla. Käytetty aineisto perustuu Yhdysvalloissa toteutettuun laajaan kliiniseen tutkimukseen, jossa tutkittiin suunielun syövän esiintyvyyttä. Tutkittavat potilaat jaoteltiin satunnaisesti kahteen eri ryhmään, joista toisessa hoitona käytettiin pelkkää sädehoitoa ja toisessa sädehoitoa yhdessä kemoterapeuttisen aineen kanssa. Tutkimuksen yksi päätarkoituksista oli verrata näitä kahta hoitomuotoa ja selvittää niiden yhteyttä potilaan selviytymiseen. Aineistoon on otettu mukaan vain kuuden suurimman tutkimukseen osallistuneen laitoksen keräämä tieto.

Lista muuttujista on seuraava:

\# (Tapauksen numero)

INST (Laitoksen numero)

SEX (Sukupuoli, 1=mies, 2=nainen)

TX (Hoito, 1=standardi, 2=koe)

GRADE (Laatu, 1=hyvin eriytynyt, 2=kohtalaisesti eriytynyt, 3=huonosti eriytynyt, NA=puuttuva tieto)

AGE (Ikä diagnoosihetkellä)

COND (Kunto, 1=ei työkyvyttömyyttä, 2=rajoittunut työkyky, 3=tarvitsee apua tullakseen toimeen, 4=vuodelevossa, NA=puuttuva tieto)

SITE (Sijainti, 1=kitakaari, 2=nielurisat, 3=takimmainen pylväs, 4=nielu, 5=takaseinämä)

T_STAGE (Kasvaimen tila, 1=enintään 2cm halkaisijaltaan, 2=halkaisijaltaan 2-4cm minimaalisesti tunkeutuneena, 3=suurempi kuin 4cm, 4=massiivinen leviävä kasvain)

N_STAGE (Etäispesäkkeiden tila, 1=ei kliinisiä todisteita etäispesäkkeistä, 2=yksittäinen etäispesäke halkaisijaltaan 3cm tai vähemmän, 3=useampia etäispesäkkeitä)

ENTRY_DT (Tutkimuksen aloitusajankohta, päivä ja vuosi, dddyy)

STATUS (Tila, 0=selviytynyt, 1=kuollut)

TIME (Elinaika, aika jonka potilas elänyt diagnoosin saamisesta lähtien)

Aineisto on nähtävillä lähteen [16] osoittamassa www-osoitteessa.

Tutkimus sisältää useita muuttujia, joilla oletetaan olevan vaikutusta selviytymisaikaan, joka on myös tärkein outliereiden vaikutusten tarkastelun kannalta. Tärkeimmät muuttujat selviytymisajan suhteen ovat sukupuoli, kasvaimen tila, etäispesäkkeiden tila, ikä, kunto ja laatu. Kasvaimen sijainti ja erot eri tutkimuslaitosten välillä vaativat myös tarkastelua [16]. Tarkasteltaessa tilannetta, jossa mukana outliereita, luotiin nk. saastutettu aineisto, jossa havaintoyksikön 32 arvoja muutettiin muuttujien age ja time kohdalta siten, että age $64 \Rightarrow 18$ ja time $1565 \Rightarrow 4000$. Aineisto pysyi muilta osin täysin ennallaan.

#	inst	sex	tx	grade	age	cond	site	tstage	nstage	entrydt	status	time
32	2	2	1	2	18	1	1	2	3	32468	1	4000

Taulukko 1. Outlier muuttujilla age ja time.

4.2 Tulosten tulkintaa

Pyrittäessä osoittamaan robustisuusestimaattoreiden toimivuus, laskettiin regressiomallit alkuperäisellä sekä saastutetulla aineistolla ilman erillistä robustisuusestimaattoria sekä tämän tutkielman kannalta keskeisimpien robustisuusestimaattoreiden kanssa. Tulokset on laskettu pelkän logistisen regressiomallin (glm) avulla, jonka lisäksi myös luokitteluvirheestimaateilla (misclass), Mallows-estimaateilla (Mqle) sekä Biancon & Yohain (BY) estimaateilla varustetulla mallilla.

Estimaattoreiden vertailu on toteutettu siten, että kunkin mallin antamat estimaatit lasketaan ensin alkuperäisellä aineistolla, jonka jälkeen saastutetulla aineistolla. Saaduista tuloksista lasketaan keskivirheiden erotuksen itseisarvot, joita vertailemalla voidaan osoittaa, että robusteilla estimaattoreilla on vaikutusta. Mallin estimaatit eivät siis muutu niin paljoa robustisuusestimaattorin kanssa kuin ilman sitä, kun tarkasteltavana on havaintoaineisto, jossa on mukana outliereita. Tulosteet, josta erot regressiokertoimien ja keskivirheiden välillä näkyvät, on esitetty liitteessä A.

Kuten jo aikaisemmin on mainittu, menetelmien mahdollista robustisuutta voidaan tarkastella vertaamalla valittujen muuttujien estimaatteja alkuperäisen ja saastutetun aineiston välillä, jossa kunkin muuttujan kohdalla esitetään alkuperäisestä sekä saastutetusta aineistosta saadut estimaatit. Mitä pienempiä erot ovat, sitä enemmän mallia voidaan pitää robustina.

Seuraavan sivun taulukoissa on esitetty estimaatit sekä keskivirheet alkuperäisellä ja saastutetulla aineistolla.

	Sex	TX	Grade	Age	Cond	Time
Estimaatti (glm)	0,27350	-0,69450	0,41912	0,04390	-0,21482	-0,00540
(Keskivirhe)	(0,56534)	(0,49408)	(0,56666)	(0,02328)	(0,74265)	(0,00086)
Estimaatti (misclass)	0,09155	-0,63477	0,49402	0,04405	-0,25607	-0,00570
(Keskivirhe)	(0,56523)	(0,50512)	(0,57275)	(0,02388)	(0,75324)	(0,00093)
Estimaatti (Mqle)	-0,35260	-0,61511	0,57107	0,03349	-0,22237	-0,00621
(Keskivirhe)	(0,61570)	(0,56513)	(0,62605)	(0,02666)	(0,85422)	(0,00113)
Estimaatti (BY)	0,23695	-0,61603	-0,15963	0,04445	-0,48996	-0,00535
(Keskivirhe)	(0,67635)	(0,50523)	(0,35969)	(0,03779)	(0,60110)	(0,00103)

Taulukko 2. Estimaatit ja keskivirheet alkuperäisellä aineistolla.

	Sex	TX	Grade	Age	Cond	Time
Estimaatti (glm)	0,11106	-0,46310	0,36189	0,01474	0,21337	-0,00388
(Keskivirhe)	(0,51378)	(0,43724)	(0,51325)	(0,01998)	(0,65855)	(0,00065)
Estimaatti (misclass)	-0,13482	-0,66186	0,44670	0,04672	-0,27867	-0,00644
(Keskivirhe)	(0,57897)	(0,53147)	(0,59593)	(0,02528)	(0,78775)	(0,00106)
Estimaatti (Mqle)	-0,45274	-0,64193	0,57612	0,03397	-0,20256	-0,00654
(Keskivirhe)	(0,63027)	(0,58096)	(0,64066)	(0,02740)	(0,87910)	(0,00121)
Estimaatti (BY)	0,08590	-0,53587	-0,25056	0,01891	-0,02934	-0,00469
(Keskivirhe)	(0,70744)	(0,50621)	(0,42691)	(0,03514)	(0,88999)	(0,00073)

Taulukko 3. Estimaatit ja keskivirheet saastutetulla (age $64 \Rightarrow 18$ ja time $1565 \Rightarrow 4000$) aineistolla.

Taulukoista 2 ja 3 voidaan vertailla eri menetelmien antamia estimaatteja alkuperäisen ja saastutetun aineiston välillä. Voidaan havaita, että erot ovat pääsääntöisesti suurimmat glm-estimaattien kohdalla. Esimerkiksi muuttujan 'Cond' kohdalla glm-estimaattien tapauksessa ero on havaittavissa muun muassa estimaatin etumerkin muuttumisena. Toisaalta taas BY-estimaattorin kohdalla muuttujalla 'Cond' estimaattien erotus on lähes yhtä suuri kuin glm-tapauksessa. Mitä pienemmät erot eri menetelmien antamilla estimaateilla (ja myös keskivirheillä) on alkuperäisen ja saastutetun aineiston välillä, sitä robustimpana mallia voidaan pitää. Huomataan, että menetelmät 'misclass' ja 'Mqle' ovat suhteellisen robusteja kyseisellä aineistolla, mutta BY-estimaattien kohdalla eroja on selvästi enemmän. Yhtenä mahdollisena syynä tähän voidaan pitää aineiston pientä kokoa ($N = 195$). Robustisuustarkasteluja tehdessä olisikin hyvä tehdä estimointi useammalla kuin yhdellä menetelmällä, jotta saatuja tuloksia voisi vertailla ja tehdyille havainnoille saataisiin tukea.

	Sex	TX	Grade	Age	Cond	Time
Estimaatti (glm)	0,24733	-0,20140	0,19065	0,00729	1,01098	-0,00076
(Keskivirhe)	(0,41762)	(0,35380)	(0,42458)	(0,01614)	(0,54786)	(0,00040)
Estimaatti (misclass)	-0,12099	-0,70689	0,72407	0,03708	-0,23378	-0,00689
(Keskivirhe)	(0,59652)	(0,55177)	(0,62118)	(0,02598)	(0,82740)	(0,00115)
Estimaatti (Mqle)	-0,51820	-0,71138	0,87684	0,02148	-0,01460	-0,00714
(Keskivirhe)	(0,65929)	(0,61291)	(0,68689)	(0,02872)	(0,95305)	(0,00136)
Estimaatti (BY)	-0,52214	-0,33372	-0,61436	0,01400	1,51532	-0,00439
(Keskivirhe)	(0,55434)	(0,42006)	(0,35103)	(0,02548)	(0,78340)	(0,00120)

Taulukko 4. Estimaatit ja keskivirheet saastutetulla (age 64 $\Rightarrow$ 18 ja time 1565 $\Rightarrow$ 9000) aineistolla, jossa mukana suurempi outlier.

Taulukossa 4 on tulokset, jossa saastutetun aineiston havaintoyksikön 32 muuttujan time arvoa nostettiin 4000:sta 9000:een. Muuttujan age arvo pysyi edelleen 18:ssa. Huomataan, että glm-estimaatit muuttuvat yhä voimakkaammin verrattuna tilanteeseen, jossa alkuperäinen aineisto. Robusteimpia tässäkin tilanteessa ovat 'misclass' eli luokitteluvirhe-estimoitu malli ja 'Mqle' eli Mallows-estimaatein varustettu malli. Näiden mallien regressiokertoimet eivät muutu läheskään niin paljoa, vaikka aineistossa on mukana nyt voimakkaasti poikkeava havainto. On kuitenkin huomattava, että BY-estimaattorin antama regressiokerroin muuttujalle 'Cond' muuttui suuremman outlierin myötä vielä enemmän. Tässä tapauksessa estimaattorin antamaa tulosta voidaan pitää kyseenalaisena. Tilastoohjelman antamat tulosteet löytyvät liitteestä A.

Luku 5

Tulosten havainnollistaminen simuloinnin avulla

Tässä kappaleessa estimaattoreiden robustisuutta on pyritty tarkastelemaan kokeellisen tilanteen avulla. Tilanne mukailee Crouxin & Haesbroeckin ([7]) artikkelissa ollutta simulointikoetta. Kokeella pyrittiin osoittamaan käytettyjen estimaattoreiden käyttökelpoisuus simuloidun aineiston avulla. Estimaattoreina käytettiin tavallisen glm-estimaattorin (glm) lisäksi tutkielmassa aiemmin luvuissa 3.1.2 ja 3.1.3 esiteltyjä luokitteluvirhe-estimaattoria (misclass) ja kvasi-uskottavuus -estimaattoria (Mqle) sekä luvussa 3.3 esiteltyä Biancon ja Yohain estimaattoria (BY). Lisäksi simulointi on tehty käyttämällä luvussa 3.3.2 esiteltyä painotettua Biancon ja Yohain estimaattoria (WBY) ja luvussa 3.4 esiteltyä painotettua suurimman uskottavuuden estimaattoria (WML).

Simuloinnit tehtiin kahdessa eri tilanteessa, joissa toisessa käytettiin dimensiota $p = 2$ ja toisessa $p = 10$ otoskoon ollessa $n = 100$. Tilanteessa I (merkitty taulukkoon 5 ja 6) selittävät muuttujat $(x_{i1}, x_{i2})^t$ noudattavat standardoitua normaalijakaumaa parametrein $N_2(\mathbf{0}, \boldsymbol{I}_2)$. Virhetermit ε_i noudattavat logistista jakaumaa yhtälön $P(\varepsilon_i \leq u) = F(u) = \frac{1}{1+\exp(-u)} = \frac{\exp(u)}{1+\exp(u)}$ nojalla. Selitettävä muuttuja luotiin simulointitilanteeseen noudattaen seuraavaa yhtälöä:

$$y_i = \begin{cases} 0 & \text{jos } \boldsymbol{x}_i^t\boldsymbol{\beta} + \varepsilon_i \leq 0, \\ \\ 1 & \text{jos } \boldsymbol{x}_i^t\boldsymbol{\beta} + \varepsilon_i > 0, \end{cases} \tag{5.1}$$

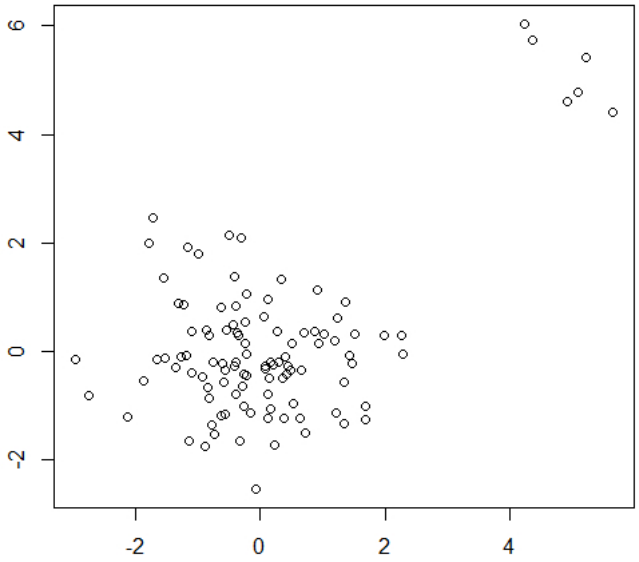
jossa todelliseksi parametrin arvoksi on asetettu $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^t = (0, 2, 2)^t$, kun $p = 2$ ja $\boldsymbol{\beta} = (0, 0, ..., 0)^t$, kun $p = 10$. Selitettävän muutujan y_i käyttäytyminen logistisen mallin tilanteessa saadaan muotoon $P(y_i = 1|\boldsymbol{x}_i) = P(\boldsymbol{x}_i^t\boldsymbol{\beta} + \varepsilon_i > \mathbf{0}|\boldsymbol{x}_i) = P(\varepsilon_i >$

$-\boldsymbol{x}_i^t\boldsymbol{\beta}|\boldsymbol{x}_i) = 1 - P(\varepsilon_i \leq -\boldsymbol{x}_i^t\boldsymbol{\beta}|\boldsymbol{x}_i) = 1 - F(-\boldsymbol{x}_i^t\boldsymbol{\beta}) = 1 - \frac{1}{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})} = \frac{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})-1}{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})} = \frac{\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}_i^t\boldsymbol{\beta})}$. Kyseessä on siis binäärinen muuttuja y, kovariaatit $\boldsymbol{x}_i$ ja todennäköisyys P. Nähdään, että F on sekä jatkuva että kasvava jakaumafunktio. Lisäksi F muodostaa logistisen funktion. Huomioon on otettu myös residuaalin ε_i yhteys logistiseen funktioon [2]. Toisessa simulointitilanteessa II aineistoon on lisätty kuusi outlieria, alkuperäisten 100 havainnon lisäksi. Tilanteessa II outliereina toimivat vipuarvot on saatu jakaumasta $N_2(5 \cdot \mathbf{1}_2; 0, 5^2 \cdot \boldsymbol{I}_2)$. Tilanne on havainnollistettu myös kuvassa 1, jossa käytetty otos on kontaminoitu suurilla outliereilla dimensiolla $p = 2$.

Estimaattoreille tehtiin kullekin $m = 1000$ ajoa ja taulukosta 5 ja 6 nähdään saadut tulokset tilanteissa I ja II, jossa vertailtavina ovat kulmakerroin-estimaattien eri komponentit. Näitä on havainnollistettu harhalla (Bias) ja keskineliövirheellä (MSE), jotka on saatu seuraavista yhtälöistä:

$$\text{Bias} = \left\| \frac{1}{m} \sum_{i=1}^{m} \hat{\boldsymbol{\beta}}_{i(-1)} - \boldsymbol{\beta}_{(-1)} \right\| \qquad \& \qquad \text{MSE} = \frac{1}{m} \sum_{i=1}^{m} \|\hat{\boldsymbol{\beta}}_{i(-1)} - \boldsymbol{\beta}_{(-1)}\|^2,$$

joissa $\|\cdot\|$ on Euklidinen normi, $\boldsymbol{\beta}_{(-1)} = (\beta_1, ..., \beta_p)^t$ ja $\hat{\boldsymbol{\beta}}_{i(-1)} = (\hat{\beta}_1, ..., \hat{\beta}_p)^t$. Yleisesti voitiin havaita, että kun aineistossa ei ollut outliereita, käyttäytyivät estimaattorit hyvin samankaltaisesti. Tilanteessa II, jossa $p = 10$, havaittiin, että myös tällöin estimaattorit käyttäytyivät hyvin samankaltaisesti. Ainoa tilanne, jossa eroja estimaattoreiden välille saatiin oli tilanne II, jossa $p = 2$. Tällöin WBY- ja WML-estimaattorit olivat ainoat, joiden harhat ja keskivirheet eivät muuttuneet voimakkaasti outliereiden takia. Myöskään BY-estimaattorin harha ja keskivirhe ei muuttunut kovin voimakkaasti, mutta kuitenkin enemmän kuin WBY- ja WML-estimaattoreiden kohdalla



Kuva 1. Simuloitu tilanne otoskoolla $n = 100$, jossa $p = 2$ ja kuusi äärimmäistä vipuarvoa lisättynä. Selittävät muuttujat (x_{i1}, x_{i2}) ovat yhteydessä arvoihin y_i.

$p = 2$	I		II	
	Bias	MSE	Bias	MSE
glm	0,222	0,653	2,689	7,273
WML	0,183	0,580	0,223	0,593
misclass	0,276	0,721	2,669	7,170
Mqle	0,289	0,845	2,669	7,165
BY	0,292	0,919	1,039	2,501
WBY	0,245	0,821	0,517	0,998

Taulukko 5. Harhat ja keskineliövirheet glm-, WML-, misclass-, Mqle-, BY- ja WBY-estimaattoreille tilanteessa $p = 2$, kun simulointeja on tehty 1000 kertaa ilman outliereita (I) ja outliereiden kanssa (II).

$p = 10$	I		II	
	Bias	MSE	Bias	MSE
glm	0,029	0,600	0,228	0,605
WML	0,028	0,718	0,024	0,669
misclass	0,018	0,612	0,217	0,588
Mqle	0,023	0,621	0,218	0,597
BY	0,030	0,637	0,205	0,622
WBY	0,029	0,642	0,186	0,577

Taulukko 6. Sama tilanne kuin taulukossa 5, mutta nyt $p = 10$.

Taulukoista 5 ja 6 nähdään simuloinnin antama tulos kyseisille menetelmille. Keskeisimpänä havaintona voidaan pitää sitä, miten vaihtelu on kohtalaisen vähäistä eri menetelmien välillä $p = 10$ tilanteessa: vain WML-estimaattorin kohdalla harha pysyy samalla tasolla keskineliövirheen kuitenkin kasvaessa korkeimmaksi. Sen sijaan $p = 2$ tilanteessa outliereiden mukana ollessa BY-, WBY- ja WML-estimaattorit on ainoat, jotka osoittautuvat robusteiksi kyseisessä tilanteessa. Näistä kolmesta WBY- ja WML-estimaattoreilla keskineliövirheet pysyvät samalla tasolla outliereiden mukana ollessa, vaikka myös BY-estimaattori osoittautuu kohtalaisen robustiksi. Kun $p = 10$, niin nähdään, että kaikkien menetelmien paitsi WML:n kohdalla harha kasvaa tilanteiden I ja II välillä. Muutos on samaa suuruusluokkaa kaikilla muilla paitsi WML-estimaattorilla, jonka harha jopa laskee.

Tilanteessa $p = 2$ yhtenä syynä sille, että vain BY-, WBY- ja WML-estimaattorit vaikuttavat robusteilta, voitaneen pitää sitä, että menetelmät glm, misclass ja Mqle eivät juuri kyseisessä simulointitilanteessa ole erityisen robusteja, mikä voi myös johtua siitä, että harha ja keskineliövirhe ovat ainoita tarkasteltavia tunnuslukuja. Simuloinnissa käytetyt R-ohjelman koodit löytyvät tutkielman liitteestä C. [7]

Luku 6

Pohdintaa

Tutkielmassa on kokeellisesti selvitetty logistisen regressiomallin tilanteessa eri robustisuusestimaattoreiden käyttökelpoisuutta. Yhtenä isona havaintona voidaan pitää estimaattoreiden epästabiiliutta, joka ilmenee esimerkiksi siten, että menetelmien toimivuus vaihtelee sen mukaan hyvin voimakkaasti, minkälainen aineisto on kulloinkin käytettävissä. Myös erot tässäkin tutkielmassa käytetyn empiirisen aineiston ja simulointitilanteen välillä on suurehkoja tarkasteltujen menetelmien välillä. Riippuu myös hyvin paljon tarkasteltavasta parametrista, mikä menetelmä on vähiten herkkä voimakkasti poikkeaville havainnoille eli on robusti. Esimerkiksi menetelmien robustisuusaste vaihtelee senkin mukaan, tarkastellaanko keskivirhettä ja harhaa yhdessä vai pelkkiä regressiokertoimia. Kaikissa tapauksissa ei myöskään ole niin, että varsinainen robustisuusestimaattori olisi robustein kaikista menetelmistä, vaan tavallisen glm-estimaattorinkin tulokset saattavat olla kaikista käyttökelpoisimpia, varsinkin jos outlierit eivät ole kovin voimakkasti poikkeavia.

Toisena merkittävänä havaintona tutkielmaa tehtäessä huomattiin, että robusteja menetelmiä logistiselle regressiolle on tutkittu suhteellisen vähän, mikä käy ilmi myös pikaisella kirjallisuushaulla. Myös ohjelmien pienoinen keskeneräisyys oli havaittavissa, sillä esimerkiksi monia R-funktioita ei oltu selvästi kokeiltu lukuisilla erilaisilla aineistoilla, koska vastaan tuli muun muassa paljon odottamattomia virheilmoituksia, joita ei välttämättä olisi jollain toisella aineistolla tullut. Johtopäätöksenä voidaankin sanoa, että ei ole yhtä oikeaa robustisuusestimaattoria, jota käyttämällä saataisiin luotettavat tulokset tilanteessa kuin tilanteessa.

Kirjallisuutta

[1] Albert, A. & Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrika* 71, 1-10.

[2] Bianco, A.M. & Yohai, V.J. (1996). Robust estimation in the logistic regression model. Teoksessa: Rieder, H. (Toim.), Robust Statistics, Data Analysis, and Computer Intensive Methods, *Lecture Notes in Statistics*, Vol. 109. Springer, New York, 17-34.

[3] Bootkrajang, J. & Kabán, A. (2012). Label-noise Robust Logistic Regression and Its Applications, *ECML PKDD'12 Proceedings of the 2012 European conference on Machine Learning and Knowledge Discovery in Databases*, Vol. 1, 143-158.

[4] Cantoni, E. & Ronchetti, E. (2001). Robust inference for generalized linear models, *Journal of the American Statistical Association*, Vol. 96, No. 455, 1022-1030.

[5] Carroll, R.J. & Pederson, S. (1993). On Robustness in the Logistic Regression Model, *Journal of the Royal Statistical Society.* Series B (Methodological), Vol. 55, No. 3, 693-706.

[6] Copas, J.B. (1988). Binary regression models for contaminated data (with discussion), *Journal of the Royal Statistical Society.* Series B, Vol. 50, 225-265.

[7] Croux, C. & Haesbroeck, G. (2003). Implementing the Bianco and Yohai estimator for logistic regression, *Computational Statistics & Data Analysis* 44, 273-295.

[8] Croux, C. & Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator, *Journal of Multivariate Analysis.* 71, 161-190.

[9] Ekholm, A. & Palmgren, J. (1987). Correction for misclassification using doubly sampled data, *Journal of Official Statistics.* 3, 419-429.

[10] Künsch, H.R. & Stefanski L.A. & Carroll R.J. (1989). Conditionally Bounded-Influence Estimation in General Regression Models, with Applications to Generalized Linear Models, *Journal of the American Statistical Association*. Vol. 84, No. 406, 460-466.

[11] Mallows, C.L. (1975). On some topics in robustness, *Technical Memorandum, Bell Telephone Laboratories*, Murray Hill.

[12] McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models*, 2nd edition, Chapman & Hall/CRC, London.

[13] Park, S.Y. & Liu, Y. (2011). Robust penalized logistic regression with truncated loss functions, *The Canadian Journal of Statistics*, Vol. 39, No. 2, 300-323.

[14] Pregibon, D. (1982). Resistant fits for some commonly used logistic models with medical applications, *Biometrics* 38, 485-498.

[15] Rousseeuw, P.J. & Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*. 85, 633-639.

[16] University of Massachusetts Amherst. (2004). Statistical Software Information, Index of Logistic Regression Datasets.

http://www.umass.edu/statdata/statdata/data/pharynx.dat (Haettu: Joulukuu 2012)

Liite A

Tilasto-ohjelman antamien tulosteiden vertailu

Tavallisella glm-estimaattorilla varustetun mallin antamat tulokset:

	Estimate	Std. Error	z value	P(>|z|)
(Intercept)	1.9995306	1.4405337	1.388	0.1651
sex2	0.2734950	0.5653415	0.484	0.6285
tx2	-0.6495008	0.4940757	-1.315	0.1887
grade2	0.4191254	0.5666558	0.740	0.4595
grade3	-0.2737268	0.6812493	-0.402	0.6878
age	0.0439020	0.0232848	1.885	0.0594
cond2	-0.2148207	0.7426512	-0.289	0.7724
cond3	-1.7642162	1.2799430	-1.378	0.1681
time	-0.0053963	0.0008611	-6.267	3.68e-10

Taulukko 8. Malli (glm) alkuperäisellä (#32, age=64, time=1565) aineistolla.

	Estimate	Std. Error	z value	P(>|z|)
(Intercept)	2.6481412	1.3195115	2.007	0.0448
sex2	0.1110626	0.5137809	0.216	0.8289
tx2	-0.4630964	0.4372441	-1.059	0.2895
grade2	0.3618875	0.5132540	0.705	0.4808
grade3	-0.3869741	0.6130156	-0.631	0.5279
age	0.0147378	0.0199765	0.738	0.4607
cond2	0.2133711	0.6585548	0.324	0.7459
cond3	-0.8625264	1.2222555	-0.706	0.4804
time	-0.0038757	0.0006464	-5.996	2.02e-09

Taulukko 9. Malli (glm) saastutetulla (#32, age=18, time=4000) aineistolla.

	Estimate	Std. Error	z value	P(>|z|)
(Intercept)	1.0218336	1.0498387	0.973	0.3304
sex2	-0.2473343	0.4176175	-0.592	0.5537
tx2	-0.2013970	0.3537951	-0.569	0.5692
grade2	0.1906481	0.4245802	0.449	0.6534
grade3	-0.6673523	0.5009283	-1.332	0.1828
age	0.0072852	0.0161380	0.451	0.6517
cond2	1.0109780	0.5478579	1.845	0.0650
cond3	0.6802631	1.1612850	0.586	0.5580
time	-0.0007625	0.0004014	-1.900	0.0575

Taulukko 10. Malli (glm) saastutetulla (#32, age=18, time=9000) aineistolla.

Luokitteluvirhe-estimaatilla (misclass) varustetun mallin antamat tulokset:

	Value	Std. Error	t value
(Intercept)	2.11210474	1.4713017163	1.4355347
sex2	0.09155351	0.5652276016	0.1619764
tx2	-0.63477176	0.5051195728	-1.2566762
grade2	0.49401652	0.5727469459	0.8625389
grade3	-0.16825302	0.6956067380	-0.2418795
age	0.04405010	0.0238836302	1.8443635
cond2	-0.25606644	0.7532407532	-0.3399530
cond3	-1.82957915	1.3083054784	-1.3984342
time	-0.00570406	0.0009252004	-6.1652155

Taulukko 11. Malli (misclass) alkuperäisellä (#32, age=64, time=1565) aineistolla.

	Value	Std. Error	t value
(Intercept)	2.478583464	1.555791627	1.5931333
sex2	-0.134816701	0.578965425	-0.2328579
tx2	-0.661857409	0.531467588	-1.2453392
grade2	0.446697283	0.595928078	0.7495825
grade3	-0.133137564	0.731132495	-0.1820977
age	0.046717789	0.025282120	1.8478588
cond2	-0.278669366	0.787753582	-0.3537519
cond3	-2.051111851	1.356951007	-1.5115593
time	-0.006435092	0.001062345	-6.0574392

Taulukko 12. Malli (misclass) saastutetulla (#32, age=18, time=4000) aineistolla.

	Value	Std. Error	t value
(Intercept)	3.123948112	1.652239313	1.8907359
sex2	-0.120986786	0.596517806	-0.2028218
tx2	-0.706888824	0.551765745	-1.2811394
grade2	0.724072384	0.621178749	1.1656426
grade3	0.106622933	0.756949517	0.1408587
age	0.037075811	0.025983513	1.4268976
cond2	-0.233780731	0.827409891	-0.2825452
cond3	-2.169627925	1.384447666	-1.5671433
time	-0.006885198	0.001151277	-5.9804861

Taulukko 13. Malli (misclass) saastutetulla (#32, age=18, time=9000) aineistolla.

Mallows-luokan estimaattorilla (Mqle) varustetun mallin antamat tulokset:

	Estimate	Std. Error	z-value	P(>|z|)
(Intercept)	3.063034	1.702650	1.799	0.072
sex2	-0.352596	0.615700	-0.573	0.567
tx2	-0.615109	0.565128	-1.088	0.276
grade2	0.571068	0.626049	0.912	0.362
grade3	-0.234884	0.763393	-0.308	0.758
age	0.033488	0.026655	1.256	0.209
cond2	-0.222370	0.854221	-0.260	0.795
cond3	-0.711461	2.050413	-0.347	0.729
time	-0.006207	0.001131	-5.487	4.08e-08

Taulukko 14. Malli (Mqle) alkuperäisellä (#32, age=64, time=1565) aineistolla.

	Estimate	Std. Error	z-value	P(>|z|)
(Intercept)	3.323031	1.768756	1.879	0.0603
sex2	-0.452740	0.630270	-0.718	0.4726
tx2	-0.641934	0.580956	-1.105	0.2692
grade2	0.576122	0.640664	0.899	0.3685
grade3	-0.247701	0.783203	-0.316	0.7518
age	0.033397	0.027399	1.219	0.2229
cond2	-0.202557	0.879096	-0.230	0.8178
cond3	-0.608047	2.182290	-0.279	0.7805
time	-0.006535	0.001206	-5.417	6.07e-08

Taulukko 15. Malli (Mqle) saastutetulla (#32, age=18, time=4000) aineistolla.

	Estimate	Std. Error	z-value	P(>|z|)
(Intercept)	4.243185	1.964476	2.160	0.0308
sex2	-0.518200	0.659285	-0.786	0.4319
tx2	-0.711384	0.612907	-1.161	0.2458
grade2	0.876836	0.686888	1.277	0.2018
grade3	-0.029017	0.823825	-0.035	0.9719
age	0.021484	0.028719	0.748	0.4544
cond2	-0.014604	0.953047	-0.015	0.9878
cond3	-0.445225	2.368105	-0.188	0.8509
time	-0.007135	0.001364	-5.232	1.67e-07

Taulukko 16. Malli (Mqle) saastutetulla (#32, age=18, time=9000) aineistolla.

Biancon & Yohain estimaattorilla (BY) varustetun mallin antamat tulokset:

	Estimate	Std. Error
(Intercept)	3.280644	2.917757278
sex	0.2369514	0.676346633
tx	-0.6160304	0.505230643
grade	-0.1596262	0.359688475
age	0.04444824	0.037788554
cond	-0.4899563	0.601104606
time	-0.005354641	0.001029293

Taulukko 17. Malli (BY) alkuperäisellä (#32, age=64, time=1565) aineistolla.

	Estimate	Std. Error
(Intercept)	4.251761	3.2698618285
sex	0.08589748	0.7074406047
tx	-0.5358721	0.5062055963
grade	-0.2505576	0.4269077936
age	0.01890681	0.0351452967
cond	-0.02936098	0.8899924751
time	-0.004688359	0.0007267457

Taulukko 18. Malli (BY) saastutetulla (#32, age=18, time=4000) aineistolla.

	Estimate	Std. Error
(Intercept)	2.385204	1.832535866
sex	-0.5221411	0.554335178
tx	-0.3337273	0.420063284
grade	-0.6143554	0.351030655
age	0.01400322	0.025477408
cond	1.515317	0.783401331
time	-0.004390417	0.001196555

Taulukko 19. Malli (BY) saastutetulla (#32, age=18, time=9000) aineistolla.

Liite B

Esimerkkiaineiston analyysissä käytetyt R-koodit

```
############################
## alkuperainen aineisto ##
############################
pharynx = read.table("pharynx.txt",header=TRUE)
pharynx$grade[pharynx$grade==9]<-NA  #Puuttuva tieto
pharynx$cond[pharynx$cond==9]<-NA  #Puuttuva tieto
pharynx$cond[pharynx$cond==0]<-NA  #Puuttuva tieto
pharynx$cond[pharynx$cond==4]<-3  #Yhdistetaan luokat 3 ja 4
#pharynx$cond[pharynx$cond==3]<-2  #Yhdistetaan luokat 2, 3 ja 4
#pharynx$cond[pharynx$cond==4]<-2  #Yhdistetaan luokat 2, 3 ja 4
pharynx$sex<-factor(pharynx$sex)
pharynx$tx<-factor(pharynx$tx)
pharynx$grade<-factor(pharynx$grade)
pharynx$cond<-factor(pharynx$cond)
pharynx$site<-factor(pharynx$site)
pharynx$tstage<-factor(pharynx$tstage)
pharynx$nstage<-factor(pharynx$nstage)
```

```
#############################################################
## muutettu havaintoyksikon 32 arvoja muuttujista age ja time #
#############################################################
pharynxout = read.table("pharynxout.txt",header=TRUE)
pharynxout$grade[pharynxout$grade==9]<-NA  #Puuttuva tieto
pharynxout$cond[pharynxout$cond==9]<-NA  #Puuttuva tieto
pharynxout$cond[pharynxout$cond==0]<-NA  #Puuttuva tieto
pharynxout$cond[pharynxout$cond==4]<-3 #Yhdistetaan luokat 3 ja 4
#harynxout$cond[pharynxout$cond==3]<-2 #Yhdistetaan luokat 2, 3 ja 4
#pharynxout$cond[pharynxout$cond==4]<-2 #Yhdistetaan luokat 2, 3 ja 4
pharynxout$sex<-factor(pharynxout$sex)
pharynxout$tx<-factor(pharynxout$tx)
pharynxout$grade<-factor(pharynxout$grade)
pharynxout$cond<-factor(pharynxout$cond)
pharynxout$site<-factor(pharynxout$site)
pharynxout$tstage<-factor(pharynxout$tstage)
pharynxout$nstage<-factor(pharynxout$nstage)
```

```
##########################################
## Mallinnus alkuperaiselle aineistolle  ##
##########################################
c1 <- glm(status~sex+tx+grade+age+cond+time,data=pharynx,family=binomial)
summary(c1)
library(robust)
c2 <- glmRob(status~sex+tx+grade+age+cond+time,data=pharynx,family=binomial,method="misclass")
summary(c2)
library(robustbase)
c3 <- glmrob(status~sex+tx+grade+age+cond+time,data=pharynx,family=binomial,method="Mqle")
summary(c3)
x0 <- pharynx[,c(2:6,12)]
y <- pharynx[,11]
ok <- complete.cases(x0,y)  #Palauttaa loogisen vektorin, joka ilmoittaa milla henkiloilla on taydelliset tiedot
x0<-x0[ok,]   #Kaytetaan vain tapauksia, joilla taydelliset tiedot
y<-y[ok]   #Kaytetaan vain tapauksia, joilla taydelliset tiedot
c4 <- BYlogreg(x0,y,initwml=FALSE)
print(c4)


##########################################
## Mallinnus saastutetulle aineistolle   ##
##########################################
c1b = glm(status~sex+tx+grade+age+cond+time,data=pharynxout,family=binomial)
summary(c1b)
c2b = glmRob(status~sex+tx+grade+age+cond+time,data=pharynxout,family=binomial,method="misclass")
summary(c2b)
c3b = glmrob(status~sex+tx+grade+age+cond+time,data=pharynxout,family=binomial,method="Mqle")
summary(c3b)
x0 <- pharynxout[,c(2:6,12)]
y <- pharynxout[,11]
ok <- complete.cases(x0,y)  #Palauttaa loogisen vektorin, joka ilmoittaa milla henkiloilla on taydelliset tiedot
x0<-x0[ok,]  #Kaytetaan vain tapauksia, joilla taydelliset tiedot
y<-y[ok]  #Kaytetaan vain tapauksia, joilla taydelliset tiedot
c4b <- BYlogreg(x0,y,initwml=FALSE)
print(c4b)
```

Liite C

Simuloinnissa käytetty R-ohjelma

```
##########################################
# p=2, valmiina WBY-estimaattoria varten #
##########################################
sim4<-function(n=100,beta0=0,beta1=c(2,2),m=100,yout=NULL,Xout=NULL)
{
  require(robust)
  require(robustbase)
  p<-length(beta1)
  beta<-c(beta0,beta1)
  s1<-rep(0,p)
  s2<-0
  i<-0
  while(i<m)
  {
    X<-matrix(rnorm(n*p),n,p)
    Z<-cbind(1,X)
    g<-Z%*%beta+rlogis(n)
    y<-as.numeric(g>0)
    if(!is.null(yout)){
      y<-c(y,yout)
      X<-rbind(X,Xout)
    }
    #mod<-glmRob(y~X, family=binomial, method="misclass")
    mod <- glmrob(y~X, family=binomial, method= "WBY")
    #mod<-WMLlogreg(X, y)
    i<-i+1
    if(i%%10==0)print(i)
    beta1hat<-mod$coefficients[-1]
    delta<-beta1hat-beta1
    s1<-s1+delta
    s2<-s2+sum(delta^2)
  }
  s1<-s1/m
  bias<-sqrt(t(s1)%*%s1)[1]
  mse<-s2/m
  list(bias=bias,mse=mse,y=y,X=X)
}
```

```
############################
# Outliereiden määrittely #
############################
Xout<-matrix(rnorm(12,5,0.5),6,2)
yout<-rep(0,6)

###############################################
#Simuloinnin käynnistäminen ilman outliereita #
###############################################
res<-sim4(m=1000,beta1=c(2,2))

#############################################################################
# Simuloinnin käynnistäminen outliereiden kanssa sekä kuvaajan tulostaminen #
#############################################################################
res<-sim4(m=1000,beta1=c(2,2),Xout=Xout,yout=yout)
plot(res$X)

########################################
# Harhan ja keskineliövirheen tulostus #
########################################
res$bias
res$mse
```

```
########################################
# p=10, valmiina WBY-estimaattoria varten #
########################################
sim4<-function(n=100,beta0=0,beta1=rep(0,10),m=100,yout=NULL,Xout=NULL)
{
  require(robust)
  require(robustbase)
  p<-length(beta1)
  beta<-c(beta0,beta1)
  s1<-rep(0,p)
  s2<-0
  i<-0
  while(i<m)
  {
    X<-matrix(rnorm(n*p),n,p)
    Z<-cbind(1,X)
    g<-Z%*%beta+rlogis(n)
    y<-as.numeric(g>0)
    if(!is.null(yout)){
      y<-c(y,yout)
      X<-rbind(X,Xout)
    }
    #mod<-glmRob(y~X, family=binomial, method="misclass")
    mod <- glmrob(y~X, family=binomial, method= "WBY")
    #mod<-WMLlogreg(X, y)
    i<-i+1
    if(i%%10==0)print(i)
    beta1hat<-mod$coefficients[-1]
    delta<-beta1hat-beta1
    s1<-s1+delta
    s2<-s2+sum(delta^2)
  }
  s1<-s1/m
  bias<-sqrt(t(s1)%*%s1)[1]
  mse<-s2/m
  list(bias=bias,mse=mse,y=y,X=X)
}

############################
# Outliereiden määrittely #
############################
Xout<-matrix(rnorm(60,5,0.5),6,10)
yout<-rep(0,6)

################################################
#Simuloinnin käynnistäminen ilman outliereita #
################################################
res<-sim4(m=1000,beta1=rep(0,10))

#############################################################################
# Simuloinnin käynnistäminen outliereiden kanssa sekä kuvaajan tulostaminen #
#############################################################################
res<-sim4(m=1000,beta1=rep(0,10),Xout=Xout,yout=yout)
plot(res$X)
```

```
########################################
# Harhan ja keskineliövirheen tulostus #
########################################
res$bias
res$mse
```