Tilastotieteen pro gradu -tutkielma

Menetelmiä regressiomallin estimointiin kompleksisessa otanta-asetelmassa

Sovellus PISA 2009 -aineistoon

Jaakko Reinikainen

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
10. syyskuuta 2012

JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Reinikainen, Jaakko: Menetelmiä regressiomallin estimointiin kompleksisessa otanta-asetelmassa: sovellus PISA 2009 -aineistoon

Tilastotieteen pro gradu -tutkielma, 47 sivua, kaksi liitettä
10. syyskuuta 2012

Tiivistelmä

Aineiston keruussa käytetään erilaisia otanta-asetelmia kustannussyistä ja tulosten tarkkuutta tavoiteltaessa. Monimutkaisen otanta-asetelman tapauksessa yksinkertaisen regressiomallin oletukset eivät välttämättä ole voimassa, joten mallinnuksessa on käytettävä sellaista menetelmää, joka ottaa otanta-asetelman huomioon. Tässä työssä esitellään otanta-aineiston regressioanalyysissä käytettyjä menetelmiä ja tutkitaan empiirisesti, kuinka yhdenmukaisia tuloksia niillä saadaan, kun sovellusaineistona käytetään Suomen ja Saksan PISA 2009 -aineistoja.

Vertailuissa käytetään lineaarista kiinteiden vaikutusten mallia ja sekamallia. Näissä havaintojen eri suuret poimintatodennäköisyydet otetaan huomioon käyttäen otantapainoja tai lisäämällä selittäjiksi asetelmamuuttujat. Regressiokertoimien keskivirheet lasketaan malliperusteisesti sekamallilla tai asetelmaperusteisesti Fay-modifioidulla tasapainotettujen puoliotosten menetelmällä tai Taylor-linearisoinnilla.

Työn keskeisenä tuloksena nähdään, että eri menetelmien antamat regressiokertoimien estimaatit ja niiden keskivirheet ovat hyvin yhdenmukaisia. Huomataan, että myös yksinkertainen menetelmä, joka ei ota lainkaan otanta-asetelmaa huomioon, voi antaa käytännössä samat estimaatit, koska sopivilla kovariaateilla on mahdollista eliminoida aineiston kompleksisuutta. Pienilläkin menetelmien välisillä eroilla voi kuitenkin olla vaikutusta mallinvalinnan lopputulokseen. Saksan aineistoon sovitetuista malleista nähdään, että on tärkeää huomata kiinteiden vaikutusten mallin ja sekamallin tulkinnallinen ero. Osoittautuu, että analyyseissä voi ja jopa kannattaa käyttää erilaisia menetelmiä, jolloin aineistoa koskeva ymmärrys lisääntyy.

Menetelmien vertailun lisäksi Suomen aineistolla selvitetään, miten eri tekijät selittävät oppilaan tietoisuutta tekstin ymmärtämisen ja muistamisen strategioista. Odotetusti sosioekonomisella asemalla ja kiinnostuksella lukemista kohtaan on positiivinen vaikutus vasteeseen ja tytöt ovat poikia merkitsevästi parempia. Muista selittäjistä muun muassa opiskelustrategioihin kuuluvien mieleenpainamis- ja kontrollistrategioiden hyödyntäminen on yhteydessä vasteeseen. Myös sillä nähdään olevan vaikutusta, minkä tyyppisiä lukemistehtäviä koulussa käytetään.

Avainsanat: Kiinteiden vaikutusten malli, sekamalli, otantapainot, tasapainotettujen puoliotosten menetelmä, Taylor-linearisointi, PISA 2009.

Kiitosmaininnat

Tämä työ ei olisi valmistunut ilman monipuolista yhteistyötä Jyväskylän yliopiston Koulutuksen tutkimuslaitoksen kanssa. Laitoksen innostava ja yhteistyöhaluinen ilmapiiri kannusti minua eteenpäin työn jokaisessa vaiheessa.

Esitän suurimmat kiitokseni tutkielman pääohjaajalle FT Kari Nissiselle, joka keksi hyvin mielenkiintoisen tutkimusongelman ja tuki työni edistymistä alusta loppuun. Hänen kanssaan käymäni keskustelut olivat avartavia ja auttoivat näkemään asioissa pintaa syvemmälle. Lisäksi sain häneltä asiantuntevaa neuvontaa SAS-ohjelmiston käytössä.

Olen kiitollinen FT Sari Sulkuselle tärkeän ja ajankohtaisen lukutaitoon liittyvän sisällöllisen tutkimusongelman keksimisestä ja kaikesta hänen tarjoamastaan avusta etenkin tulosten tulkinnoissa. Osoitan suurkiitokset myös professori Antero Malinille työhöni liittyvistä arvokkaista ideoista ja kommenteista.

Haluan kiittää YTM Eija Puhakkaa, jonka apu erityisesti PISA-tutkimuksen otantaan liittyvissä kysymyksissä oli korvaamatonta. Lisäksi FT Salme Kärkkäinen ja KT, FL Pasi Reinikainen ansaitsevat kiitokset työn loppuvaiheessa antamistaan kommenteista, jotka auttoivat selkeyttämään ja terävöittämään tekstiäni.

Jyväskylässä 30.8.2012
Jaakko Reinikainen

Sisältö

1 Johdanto

Kyselytutkimuksissa vastaajat valitaan usein yksinkertaisen satunnaisotannan sijaan käyttäen jotakin monimutkaisempaa otanta-asetelmaa. Tällä tavoitellaan toisaalta suurempaa tarkkuutta tuloksissa ja toisaalta kustannussäästöjä. Valittu otanta-asetelma on otettava huomioon tehtäessä tilastollisia analyysejä, sillä on mahdollista, että alkioilla ei ole samaa poimintatodennäköisyyttä, eikä havaintoja välttämättä voida pitää toisistaan riippumattomina. Näin ollen otanta-asetelman huomiotta jättäminen saattaisi johtaa harhaisiin estimaatteihin ja virheellisiin johtopäätöksiin niiden merkitsevyydestä (Lohr 1999).

Otanta-asetelma voidaan ottaa huomioon monella eri tavalla, jotka myös saattavat johtaa hieman erilaisiin tuloksiin. Tässä tutkielmassa esitellään ja vertaillaan erilaisia regressiomallin estimoinnissa käytettyjä menetelmiä, jotka ottavat huomioon kompleksisen otanta-asetelman. Eri menetelmät tuottavat erilaisia estimaatteja regressiokertoimille ja niiden keskivirheille.

Menetelmien empiirisessä vertailussa alkioiden eri suuret poimintatodennäköisyydet otetaan huomioon käyttäen otantapainoja tai lisäämällä selittäjiksi muuttujat, joita on hyödynnetty otoksen poiminnassa. Havaintojen klusteroituminen otetaan huomioon estimaattien keskivirheitä laskettaessa malliperusteisesti lineaarisilla sekamalleilla tai asetelmaperusteisesti Taylor-linearisoinnilla tai Fay-modifioidulla tasapainotettujen puoliotosten menetelmällä (BRR-menetelmä, *balanced repeated replication method*), jota käytetään muun muassa PISA-tutkimuksissa. Regressiokertoimien estimaattien laskemisessa käytetään kiinteiden vaikutusten mallia tai sekamallia. Mallinvalinta ja saman mallin sovittaminen tehdään eri menetelmillä, ja näin nähdään, kuinka yhdenmukaisia tuloksia ne antavat.

Sovellusaineistoina vertailuissa käytetään Suomen ja Saksan PISA 2009 -aineistoja. PISA (*Programme for International Student Assessment*) on kansainvälinen 15-vuotiaiden koulutaitoja ja -asenteita mittaava tutkimus. Aineistot on kerätty kaksiasteisella otannalla, jossa ensin koulut on poimittu soveltaen systemaattista ositettua PPS-otantaa (*probability proportional to size sampling*) ja tämän jälkeen oppilaat on poimittu otokseen päätyneistä kouluista systemaattisella otannalla. Koska saman koulun oppilailla opiskeluympäristö ja lukujärjestys ovat samankaltaisia, niin on oletettavaa, että he ovat keskenään homogeenisempi joukko kuin täysin satunnaisesti valitut oppilaat. Tämän huomioon ottaminen on sitä tärkeämpää, mitä voimakkaampaa klusteroituminen on, eli mitä suurempaa koulujen välinen vaihtelu on, tai toisin sanoen mitä pienempää koulujen sisäinen vaihtelu on suhteessa kokonaisvaihteluun. Tutkimuksessa tehtävien menetelmien vertailujen kannalta Suomen ja Saksan aineistojen kiinnostava ero onkin se, että Suomessa koulujen väliset erot ovat hyvin pieniä, kun taas Saksassa erot ovat suhteellisen suuria.

Aineistosta käytetään pääasiassa lukutaitoon liittyviä muuttujia, ja Suomen aineistoon sovitettaviin malleihin liittyy menetelmien vertailun lisäksi sisällöllinen tutkimusongelma. Aikaisemmissa tutkimuksissa on huomattu, että

oppilaan lukutaitoa selittää vahvasti tietoisuus tekstin ymmärtämisen ja muistamisen strategioista (Sulkunen ym. 2010). Siihen opettajien on myös mahdollista vaikuttaa, toisin kuin esimerkiksi oppilaan sukupuoleen tai sosioekonomiseen asemaan, jotka ovat myös lukutaidon merkitseviä selittäjiä. Sisällöllisesti kiinnostavana tutkimusongelmana onkin selvittää, miten eri tekijät puolestaan selittävät tietoisuutta tekstin ymmärtämisen ja muistamisen strategioista.

Seuraavaksi luvussa 2 esitetään lineaaristen mallien yleistä teoriaa ja luvussa 3 otanta-asetelman huomioon ottavaa mallien estimointia. Tämän jälkeen esitellään PISA 2009 -aineistoa. Luvussa 5 suoritetaan menetelmien empiirinen vertailu ja tarkastellaan saatuja tuloksia.

2 Lineaarisista malleista

Oletetaan, että $\boldsymbol{y}$ on vastemuuttujan havaittujen arvojen $n \times 1$ -vektori, ja käytetään sen odotusarvovektorille ja kovarianssimatriisille merkintöjä

$$E(\boldsymbol{y}) = \boldsymbol{\mu} \quad \text{ja} \quad \text{Cov}(\boldsymbol{y}) = \boldsymbol{V}.$$

Voidaan kirjoittaa hyvin yleisesti, että vektorille $\boldsymbol{y}$ pätee

$$\boldsymbol{y} \sim (\boldsymbol{\mu}, \boldsymbol{V}),$$

eli $\boldsymbol{y}$ noudattaa jotakin jakaumaa, jonka odotusarvovektori on $\boldsymbol{\mu}$ ja kovarianssimatriisi $\boldsymbol{V}$.

Täysin yleisessä muodossa $\boldsymbol{\mu}$ sisältää n kappaletta alkioita ja symmetrinen $\boldsymbol{V}$ sisältää $n(n+1)/2$ eri alkiota. Koska $\boldsymbol{y}$ sisältää nyt vähemmän alkioita kuin $\boldsymbol{\mu}$ ja $\boldsymbol{V}$ yhteensä, on $\boldsymbol{\mu}$ ja $\boldsymbol{V}$ spesifioitava jollain tavalla, jotta niiden estimointi tulee mahdolliseksi (McCulloch ja Searle 2001, 115). Se, miten tämä spesifiointi tehdään, riippuu mallinnettavan aineiston tyypistä ja rakenteesta. Tässä luvussa esitetään mallintaminen lineaarisella kiinteiden vaikutusten mallilla ja sekamallilla. Mallin lineaarisuus tarkoittaa sitä, että vasteen odotusarvo $\boldsymbol{\mu}$ voidaan esittää mallin parametrien lineaarisena lausekkeena.

2.1 Kiinteiden vaikutusten malli

Kiinteiden vaikutusten malleja käytetään tutkittaessa selittävien muuttujien vaikutusta vastemuuttujaan. Selittävää muuttujaa käsitellään kiinteänä, jos ollaan kiinnostuttu nimenomaan sen vaikutuksesta vasteeseen ja jos kategorisen muuttujan tapauksessa se sisältää kaikki kiinnostavat tasot. Tässä luvussa esitellään kiinteiden vaikutusten lineaarinen regressiomalli teoksen McCulloch ja Searle (2001) mukaan.

Matriisimuodossa kirjoitettuna malli on

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e},$$

missä $\boldsymbol{y}$ on jatkuvan vastemuuttujan arvojen $n \times 1$ -vektori, $\boldsymbol{X}$ on tunnettu kiinteiden selittävien muuttujien arvojen $n \times p$ -matriisi ja $\boldsymbol{\beta}$ on tuntemattomien kiinteiden vaikutusten $p \times 1$ -parametrivektori, joka sisältää vakioparametrin ja $p-1$ regressiokerrointa. Oletetaan, että jäännökset ovat riippumattomia sekä toisistaan että selittäjistä ja että $n \times 1$ -jäännösvektorille pätee $\boldsymbol{e} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{I})$, missä $\boldsymbol{I}$ on $n \times n$ identtinen matriisi. Seurauksena saadaan, että

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}).$$

2.1.1 Parametrien estimointi

Parametrien $\boldsymbol{\beta}$ ja σ^2 estimointi voidaan tehdä suurimman uskottavuuden menetelmällä, jossa maksimoidaan mallin uskottavuusfunktio parametrien $\boldsymbol{\beta}$ ja

σ^2 suhteen. Normaalisuusoletuksen nojalla uskottavuusfunktioksi saadaan

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right). \tag{1}$$

Maksimoinnin helpottamiseksi käytetään uskottavuusfunktion logaritmia, log-uskottavuutta:

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Derivoimalla log-uskottavuus erikseen parametrien $\boldsymbol{\beta}$ ja σ^2 suhteen saadaan

$$\begin{aligned}
\frac{\partial \log L(\boldsymbol{\beta}, \sigma^2)}{\partial \boldsymbol{\beta}} &= \frac{\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}}{\sigma^2} \quad \text{ja} \\
\frac{\partial \log L(\boldsymbol{\beta}, \sigma^2)}{\partial \sigma^2} &= -\frac{n}{2(\sigma^2)^2} + \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}.
\end{aligned}$$

Asettamalla derivaatat nolliksi ja ratkaisemalla niistä $\boldsymbol{\beta}$ ja σ^2, saadaan uskottavuusfunktion maksimoivat ratkaisupisteet eli suurimman uskottavuuden estimaattorit

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \quad \text{ja} \quad \hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n}.$$

Edellä vaaditaan, että käänteismatriisi $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ on olemassa eli että matriisin $\boldsymbol{X}$ aste on p.

Jos matriisi $\boldsymbol{X}$ ei ole täysiasteinen, niin $\boldsymbol{X}^T\boldsymbol{X}$ ei ole kääntyvä. Tällöin voidaan käyttää yleistettyä käänteismatriisia $(\boldsymbol{X}^T\boldsymbol{X})^-$, jolle pätee

$$\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{X}.$$

Yleistetty käänteismatriisi ei ole yksikäsitteinen kuten ei myöskään sitä käyttäen saatu estimaattori

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{y}.$$

Tällöin kuitenkin $\boldsymbol{X}\hat{\boldsymbol{\beta}}$ on yksikäsitteinen ja harhaton estimaattori vasteen odotusarvolle, eli

$$E\left[\widehat{E(\boldsymbol{y})}\right] = E(\boldsymbol{X}\hat{\boldsymbol{\beta}}) = \boldsymbol{X}\boldsymbol{\beta}.$$

Jos $\boldsymbol{X}$ on täysiasteinen, niin yleistetty käänteismatriisi on yksikäsitteinen ja pätee

$$(\boldsymbol{X}^T\boldsymbol{X})^- = (\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Parametrille σ^2 johdettu estimaattori on kuitenkin alaspäin harhainen. Harhaton estimaattori saadaan asettamalla nimittäjäksi $n - \text{rank}(\boldsymbol{X})$, jolloin estimaattori on muotoa

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n - p}. \tag{2}$$

Regressiokertoimien estimaattori on harhaton, eli $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ ja sen kovarianssimatriisi on

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1},$$

jonka diagonaalialkioiden neliöjuuret ovat regressiokertoimien estimaattien keskivirheitä.

Edellä johdettu regressiokertoimien estimaattori saadaan myös yksinkertaisemmin ilman normaalisuusoletusta pienimmän neliösumman (PNS) menetelmällä. PNS-menetelmässä minimoidaan uskottavuusfunktiossa (1) esiintyvä neliösumma $(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})$ parametrien $\boldsymbol{\beta}$ suhteen. Ilman normaalisuusoletusta saadaan myös johdettua harhaton estimaattori varianssille σ^2. Voidaan osoittaa, että

$$E[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})] = (n-p)\sigma^2,$$

joten asettamalla

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})}{n-p},$$

saadaan varianssiestimaattori (2), joka on harhaton:

$$E(\hat{\sigma}^2) = \frac{E[(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})]}{n-p} = \frac{(n-p)\sigma^2}{n-p} = \sigma^2.$$

Jos havaintojen varianssit eivät ole yhtä suuria, voidaan estimointi suorittaa painotetulla PNS-menetelmällä. Havaintoja painotetaan niiden varianssien käänteisluvuilla, eli käytetään painomatriisia $\boldsymbol{W} = \mathrm{diag}(w_1, \ldots, w_n) = \mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})$. Parametrit $\boldsymbol{\beta}$ ratkaistaan nyt minimoimalla painotettu neliösumma $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. Mitä suurempi havaintoon liittyvä paino w_i on, sitä enemmän se vaikuttaa optimoinnin tulokseen. Regressiokertoimien estimaattoriksi saadaan

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{y} \tag{3}$$

ja sen kovarianssimatriisiksi

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}. \tag{4}$$

Kun edelleen oletetaan havaintojen normaalisuus, saadaan uskottavuusfunktioksi

$$L(\boldsymbol{\beta}, \boldsymbol{W}) = (2\pi)^{-\frac{n}{2}}|\boldsymbol{W}|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T\boldsymbol{W}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right).$$

Uskottavuuden maksimointi parametrien $\boldsymbol{\beta}$ suhteen johtaa estimaattoriin (3). Menetelmä voidaan yleistää korreloituneelle aineistolle, jolloin painomatriisi $\boldsymbol{W}$ ei enää ole diagonaalinen. (Gelman ja Hill 2007, 389.)

2.2 Sekamalli

Kun edellisessä luvussa esiteltyyn kiinteiden vaikutusten malliin lisätään satunnaisvaikutuksia, saadaan sekamalli. Muuttujaa voidaan käsitellä satunnaisvaikutuksena, jos sen tasot eli luokat ovat vain otos jostain mahdollisten luokkien perusjoukosta, eikä mielenkiinto välttämättä kohdistu nimenomaan otokseen päätyneisiin luokkiin, vaan pikemminkin satunnaisvaikutusten jakaumaan. Esimerkiksi koulun, sairaalan tai maantieteellisen alueen vaikutusta voitaisiin pitää satunnaisena. Satunnaisvaikutusten avulla voidaan ottaa huomioon aineiston kovarianssirakenne, joten sekamallit sopivat klusteroituneen aineiston analysointiin. Sekamallien teoria esitetään kirjaan McCulloch ja Searle (2001) perustuen.

Lineaarinen sekamalli on matriisimuodossa kirjoitettuna

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e},$$

missä $\boldsymbol{Z}$ on satunnaisvaikutuksiin liittyvä tunnettu $n \times q$ -design-matriisi, $\boldsymbol{u}$ on satunnaisvaikutusten $q \times 1$ -vektori ja muut kuten kiinteiden vaikutusten mallissa. Oletetaan mallista, että $\boldsymbol{u} \sim N(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$ ja $\mathrm{Cov}(\boldsymbol{u}, \boldsymbol{e}) = \boldsymbol{0}$. Näistä seuraa, että

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T + \boldsymbol{R}),$$

mistä nähdään, että satunnaisvaikutukset eivät esiinny vastemuuttujan odotusarvossa, vaan määrittelevät vasteen kovarianssirakenteen.

Jos matriisin $\boldsymbol{Z}$ alkiot ovat ykkösiä ja nollia ja satunnaisvaikutukset $\boldsymbol{u}$ ovat korreloimattomia, on kyseessä sekamallin erikoistapaus, varianssikomponenttimalli. Kirjoitetaan yksinkertainen kahden varianssikomponentin malli skalaarimuodossa

$$y_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + u_i + e_{ij}\ , \tag{5}$$

missä $\boldsymbol{x}_{ij}^T\boldsymbol{\beta} = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_{p-1}x_{ijp-1}$, $u_i \sim N(0, \sigma_u^2)$ ja $e_{ij} \sim N(0, \sigma_e^2)$. Alaindeksi $i = 1, \ldots, q$ viittaa satunnaisvaikutuksen luokkaan eli klusteriin, joka voi olla esimerkiksi koulu, ja $j = 1, \ldots, n_i$ viittaa klusterin sisäiseen alkioon, esimerkiksi oppilaaseen. Tällöin siis kovarianssimatriisit ovat muotoa $\boldsymbol{D} = \sigma_u^2\boldsymbol{I}$ ja $\boldsymbol{R} = \sigma_e^2\boldsymbol{I}$.

Edellä esitettyä varianssikomponenttimallia voidaan kutsua myös satunnaisen tasoparametrin malliksi, koska mallissa vakioparametri $\beta_0 + u_i$ vaihtelee satunnaisvaikutuksen luokkien välillä. Sekamalliin voidaan lisätä kiinteiden ja satunnaisvaikutusten interaktioita, jotka ovat myös satunnaisia. Kun edelliseen malliin lisätään ensimmäisen selittäjän ja satunnaisvaikutuksen välinen interaktio, on malli muotoa

$$y_{ij} = \beta_0 + (\beta_1 + u_{1i})x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_{p-1}x_{ijp-1} + u_{0i} + e_{ij}\ ,$$

missä $u_{1i} \sim N(0, \sigma_{u_1}^2)$. Kyseessä on satunnaiskertoiminen regressiomalli, jossa siis ensimmäisen selittäjän kulmakertoimet $\beta_1 + u_{1i}$ vaihtelevat satunnaisvaikutuksen luokkien välillä.

Kuten aiemmin mainittiin, satunnaisvaikutuksilla voidaan ottaa huomioon aineiston kovarianssirakenne: samaan satunnaisvaikutuksen luokkaan liittyvät havainnot ovat keskenään korreloituneita. Keskeinen käsite aineiston klusteroituneisuuden voimakkuuden mittaamisessa on sisäkorrelaatio. Varianssikomponenttimallin (5) merkinnöin sisäkorrelaation määritelmä on

$$\mathrm{Cor}(y_{ij}, y_{il}) = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_e^2}, \qquad j \neq l.$$

Sisäkorrelaatio on luku nollan ja ykkösen väliltä ja sen voidaan tulkita kuvaavan, kuinka suuren osan klustereiden välinen vaihtelu kattaa aineiston kokonaisvaihtelusta.

Sekamallissa satunnaisosan mallintamisella saadaan siis otettua huomioon havaintojen välistä riippuvuutta. On kuitenkin huomattava, että kun kiinteiden vaikutusten malliin lisätään satunnaisosa, niin kiinteän osan regressiokertoimien tulkinta muuttuu. Kiinteiden vaikutusten malli on ikään kuin malli koko aineistolle, kun taas sekamallissa kussakin satunnaisvaikutuksen luokassa mallit voivat olla erilaisia. Sekamallissa kiinteiden vaikutusten kertoimet ovat näitä satunnaisvaikutuksen luokkia koskevien mallien keskimääräisiä kertoimia. Tämä mallien tulkinnallinen ero tulee selvemmin esille Saksan aineiston anaalyyseissä luvussa 5.3.2.

2.2.1 Parametrien estimointi ja ennustaminen

Mallin parametrien estimointi voidaan tehdä suurimman uskottavuuden menetelmällä. Merkitään jatkossa $\mathrm{Cov}(\boldsymbol{y}) = \boldsymbol{V}$. Uskottavuusfunktio on nyt normaalisuusoletuksen perusteella

$$L_{ML}(\boldsymbol{\beta}, \boldsymbol{V}) = (2\pi)^{-\frac{n}{2}} |\boldsymbol{V}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right) \qquad (6)$$

ja log-uskottavuus on

$$\log L_{ML}(\boldsymbol{\beta}, \boldsymbol{V}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Menetelmä antaa kiinteiden vaikutusten estimaattoriksi

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y},$$

missä edellytetään, että $(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}$ on olemassa. Jos $\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X}$ ei ole kääntyvä, voidaan käyttää sen yleistettyä käänteismatriisia $(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^-$ vastaavasti kuin kiinteiden vaikutusten mallissa. Estimaattorin kovarianssimatriisiksi saadaan

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1}.$$

Parametrien $\boldsymbol{\beta}$ estimaattorissa esiintyy kovarianssimatriisi $\boldsymbol{V}$, joka on yleensä käytännössä tuntematon, ja pitää myös estimoida. Tässäkin voidaan käyttää suurimman uskottavuuden menetelmää. Uskottavuusfunktioon (6) sijoitetaan parametrien $\boldsymbol{\beta}$ paikalle $(\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{y}$, minkä jälkeen uskottavuus maksimoidaan matriisin $\boldsymbol{V}$ alkioiden suhteen. Toisin sanoen matriisi

$\boldsymbol{V}$ ratkaistaan komponenteittain. Voidaan merkitä $\boldsymbol{V} = \boldsymbol{V}(\boldsymbol{\theta})$ ja esimerkiksi yksinkertaisen varianssikomponenttimallin tapauksessa komponentit olisivat $\boldsymbol{\theta} = (\sigma_u^2, \sigma_e^2)^T$. Maksimointi voidaan suorittaa numeerisilla menetelmillä, kuten Newton-Raphsonin algoritmilla tai EM-algoritmilla.

Suurimman uskottavuuden menetelmällä saatu matriisin $\boldsymbol{V}$ estimaatti $\hat{\boldsymbol{V}}_{ML}$ on alaspäin harhainen, koska menetelmä ei ota huomioon p kappaletta parametrien $\boldsymbol{\beta}$ estimoinnissa menetettäviä vapausasteita. Harha voi olla huomattava, jos matriisin $\boldsymbol{X}$ aste p on suuri suhteessa otoskokoon n. Vapausasteet saadaan otettua huomioon käyttämällä estimoinnissa rajoitettua suurimman uskottavuuden menetelmää (REML, *restricted/residual maximum likelihood*).

REML-menetelmässä valitaan sellainen $n \times (n-p)$ -matriisi $\boldsymbol{K}$, että sen sarakevektorit ovat lineaarisesti riippumattomia ja $\boldsymbol{K}^T\boldsymbol{X} = \boldsymbol{0}$. Tällöin normaalisuusoletuksen nojalla

$$\boldsymbol{K}^T\boldsymbol{y} \sim N(\boldsymbol{0}, \boldsymbol{K}^T\boldsymbol{V}\boldsymbol{K}).$$

Nähdään, että vektorin $\boldsymbol{K}^T\boldsymbol{y}$ jakauma ei riipu ollenkaan mallin kiinteistä vaikutuksista. Matriisi $\boldsymbol{K}$ ikään kuin keskistää aineiston siten, että uskottavuudessa ei esiinny parametreja $\boldsymbol{\beta}$. Nyt kovarianssimatriisi $\boldsymbol{V}$ estimoidaan suurimman uskottavuuden menetelmällä käyttäen vektoria $\boldsymbol{K}^T\boldsymbol{y}$ vektorin $\boldsymbol{y}$ sijaan. Log-uskottavuusfunktioksi saadaan

$$\log L_{REML}(\boldsymbol{V}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{K}^T\boldsymbol{V}\boldsymbol{K}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{K}(\boldsymbol{K}^T\boldsymbol{V}\boldsymbol{K})^{-1}\boldsymbol{K}^T\boldsymbol{y}.$$

Kovarianssimatriisin estimoinnin jälkeen $\boldsymbol{\beta}$ estimoidaan erikseen käyttäen REML-menetelmällä saatua kovarianssimatriisia $\hat{\boldsymbol{V}}_{REML}$. Tällöin

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\hat{\boldsymbol{V}}_{REML}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{V}}_{REML}^{-1}\boldsymbol{y}.$$

Voidaan osoittaa, että REML-menetelmään liittyvä log-uskottavuus on kirjoitettavissa muodossa

$$\log L_{REML}(\boldsymbol{V}) = \log L_{ML}(\boldsymbol{\beta}, \boldsymbol{V}) - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}|,$$

eli ML- ja REML-menetelmiin liittyvät log-uskottavuusfunktiot ovat muuten samat, paitsi REML-menetelmällä mukaan tulee parametrien $\boldsymbol{\beta}$ estimoinnista johtuva niin sanottu sakkotermi.

Satunnaisvaikutusten $\boldsymbol{u}$ yhteydessä käytetään estimoinnin sijaan termiä ennustaminen, koska satunnaisvaikutuksia ei ajatella kiinteiksi, vaan realisaatioiksi jostain jakaumasta. Paras ennuste $BP(\boldsymbol{u})$ (*best predictor*) on ehdollinen odotusarvo

$$\tilde{\boldsymbol{u}} = BP(\boldsymbol{u}) = E(\boldsymbol{u}|\boldsymbol{y}).$$

Paras tarkoittaa tässä pienintä keskineliövirhettä. Kun oletetaan edelleen, että $\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{V})$ ja $\boldsymbol{u} \sim N(0, \boldsymbol{D})$ ja että $(\boldsymbol{y}, \boldsymbol{u})^T$ noudattaa moniulotteista normaalijakaumaa, saadaan

$$E(\boldsymbol{u}|\boldsymbol{y}) = \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Sijoittamalla tuntemattomattomien parametrien $\boldsymbol{\beta}$ paikalle estimaattori $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$, saadaan paras lineaarinen harhaton ennuste $BLUP(\boldsymbol{u})$ (*best linear unbiased predictor*)

$$\tilde{\boldsymbol{u}} = BLUP(\boldsymbol{u}) = \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Ennusteisiin liittyy kutistuminen, eli ennusteiden $\tilde{\boldsymbol{u}}$ varianssi on aina pienempi tai yhtäsuuri kuin todellisten satunnaisvaikutusten $\boldsymbol{u}$ varianssi. Pätee siis

$$\mathrm{Var}(\tilde{u}_i) \leq \mathrm{Var}(u_i).$$

Kovarianssimatriisiksi ennusteille saadaan

$$\mathrm{Cov}(\tilde{\boldsymbol{u}}) = \boldsymbol{D}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{Z}\boldsymbol{D},$$

missä $\boldsymbol{P} = \boldsymbol{V}^{-1}[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}]$.

3 Otanta-asetelman huomioon ottaminen

Yksinkertaisissa regressiomalleissa havaintojen oletetaan olevan riippumattomia. Lisäksi niissä on lähtökohtana, että jokainen havainto edustaa yhtä monta perusjoukon alkiota, eikä havaintoja näin ollen tarvitse painottaa. Nämä oletukset eivät kuitenkaan yleensä päde kompleksisen otanta-asetelman tapauksessa, ja seuraavaksi esitellään tapoja, joilla tämä voidaan ottaa huomioon äärellistä populaatiota koskevassa regressioanalyysissä.

3.1 Otantapainot

Jos alkioiden poimintatodennäköisyydet ovat eri suuria, on tämä otettava huomioon estimoinnissa, tai muuten saadaan harhaisia tuloksia. Havaintoja painotetaan siten, että pienemmällä todennäköisyydellä poimittua alkiota pidetään tärkeämpänä kuin suuremmalla todennäköisyydellä poimittua. Havaintoihin liittyvät painokertoimet eli otantapainot ovatkin poimintatodennäköisyyksien käänteislukuja. Lähdeteoksena tässä luvussa on käytetty kirjaa Lohr (1999).

Olkoon π_i alkion i poimintatodennäköisyys ja $i = 1, \ldots, N$, missä N on populaation koko. Tällöin alkioon i liittyvä otantapaino on $w_i = 1/\pi_i$, ja voidaan ajatella, että havainto i edustaa w_i:tä populaation alkiota. Jos otoskoko on n ja suoritetaan yksinkertainen satunnaisotanta palauttamatta, niin alkion i poimintatodennäköisyys on $\pi_i = n/N$ ja edelleen otantapaino on $w_i = N/n$. Tästä seuraa, että otantapainojen summa on sama kuin populaation koko, eli otos edustaa koko populaatiota.

$$\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} \frac{N}{n} = N.$$

Kaksiasteisessa otannassa alkion poimintatodennäköisyys on ensiasteen otantayksikön eli klusterin poimintatodennäköisyyden ja toisen asteen otantayksikön klusterin sisäisen ehdollisen poimintatodennäköisyyden tulo. Olkoon nyt π_i ensiasteen otantayksikön (esimerkiksi koulun) poimintatodennäköisyys ja $\pi_{j|i}$ toisen asteen otantayksikön (esimerkiksi oppilaan) poimintatodennäköisyys, kun tiedetään, että otosyksikkö i on valittu otokseen. Tällöin siis poimintatodennäköisyydet ovat $\pi_{ij} = \pi_i \pi_{j|i}$ ja otantapainot $w_{ij} = 1/(\pi_i \pi_{j|i})$.

Tarkkuuden parantamiseksi ensiasteen otantayksiköiden poiminnassa käytetään usein PPS-otantaa (*probability proportional to size sampling*). Tällöin poimintatodennäköisyys on suhteessa otantayksikön kokoon, eli suuremmilla otantayksiköillä on suurempi todennäköisyys päätyä otokseen kuin pienemmillä. Otantayksikön koko on jonkin tunnetun positiivisen apumuuttujan arvo. Kaksiasteisessa otannassa ensiasteen otantayksikön koko on usein sen sisältämien alkioiden lukumäärä. PPS-otannassa ensiasteen otantayksikön poimintatodennäköisyys on $\pi_i = \min\{mN_i/N, 1\}$, missä m on poimittavien ensiasteen otantayksiköiden määrä ja N_i on otantayksikön i koko. Luonnollisesti vaaditaan, että $\pi_i \leq 1$. Jos N_i on hyvin suuri, niin on kuitenkin mahdollista, että

$mN_i/N > 1$, jolloin asetetaan $\pi_i = 1$. (Särndal ym. 1992, 87-90.)

Liittämällä otantapainot havaintoihin saadaan laskettua populaatiota koskevia estimaatteja. Esimerkiksi harhaton Horvitz-Thompson-estimaattori kaksiasteisessa otannassa muuttujan y populaation keskiarvolle on

$$\hat{\bar{y}}_{HT} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}},$$

missä n_i on otantayksikön i sisäinen otoskoko. Käytännössä otantapainoja ei aina muodosteta pelkästään poimintatodennäköisyyksien avulla, vaan niissä voidaan ottaa huomioon esimerkiksi vastauskato ja jälkiositteet.

3.2 Otantapainot kiinteiden vaikutusten mallissa

Jos alkioiden poimintatodennäköisyydet ovat eri suuret, niin äärellistä populaatiota koskevan kiinteiden vaikutusten mallin parametrien estimaatit saattavat olla harhaisia, mikäli analyysissä ei käytetä otantapainoja. Seuraavaksi kerrotaan painojen sisällyttämisestä malliin lähdeteoksen Lohr (1999) mukaan.

Tässä luvussa alaindeksillä s viitataan otokseen ja esimerkiksi merkinnöillä $\boldsymbol{y}$ ja $\boldsymbol{X}$ tarkoitetaan vasteen ja selittäjien arvoja koko populaatiossa. Regressiokertoimien estimaatit lasketaan otantapainojen kanssa painotetulla pienimmän neliösumman menetelmällä, jolloin

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{y}_s,$$

missä $\boldsymbol{W}_s = \mathrm{diag}(w_1, w_2, \ldots, w_n)$ on otantapainojen diagonaalimatriisi. Esimerkiksi vektorin $\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{y}_s$ ensimmäinen alkio

$$\sum_{i=1}^{n} x_{i1}y_iw_i$$

on vektorin $\boldsymbol{X}^T\boldsymbol{y}$ ensimmäisen alkion

$$\sum_{i=1}^{N} x_{i1}y_i$$

otanta-asetelman suhteen harhaton estimaattori.

Vaikka regressiokertoimien estimaattori saadaankin painotetulla PNS-menetelmällä, niin sen kovarianssimatriisina ei käytetä matriisia $(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}$, mikä olisi painotetun PNS-menetelmän antama kaava (4), jos otantapainojen matriisi $\boldsymbol{W}_s$ samastettaisiin matriisiin $\mathrm{Cov}(\boldsymbol{y}_s)^{-1}$. Otantapainot eivät kuitenkaan kuvaa aineiston kovarianssirakennetta, vaan ne ovat seurausta otanta-asetelmasta. Regressiokertoimien kovarianssimatriisi voidaan laskea esimerkiksi käyttäen Taylor-linearisointia, joka esitetään seuraavaksi teoksen Wolter (2007, 249-252) mukaan.

Regressiokertoimien estimaattori on epälineaarinen, koska selittävien muuttujien arvot $\boldsymbol{X}_s$ ajatellaan satunnaisiksi kuten vasteen $\boldsymbol{y}_s$ arvot. Näin ollen on käytännöllistä approksimoida estimaattoria lineaarisella Taylorin polynomilla, jolloin varianssin laskeminen tulee helpommaksi. Lasketaan seuraavaksi ensimmäisen asteen approksimaatio soveltaen Taylorin sarjakehitelmää matriisiin $(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}$ matriisin $\boldsymbol{X}^T\boldsymbol{X}$ suhteen.

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\boldsymbol{W}_s(\boldsymbol{X}_s\boldsymbol{\beta}+\boldsymbol{e}_s)\\ &= \boldsymbol{\beta}+(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{e}_s\\ &= \boldsymbol{\beta}+\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}-(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s-\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}+\text{Jäännös}\right]\\ &\quad\times\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{e}_s\,.\end{aligned}$$

Tästä saadaan

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{e}_s+\text{Jäännös},$$

jolloin kovarianssimatriisin approksimaatioksi saadaan

$$\text{Cov}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\text{Cov}(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{e}_s)(\boldsymbol{X}^T\boldsymbol{X})^{-1}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{G}(\boldsymbol{X}^T\boldsymbol{X})^{-1},$$

missä $\boldsymbol{G}=\text{Cov}(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{e}_s)$. Koska $\boldsymbol{\beta}$ on tuntematon, niin myös jäännökset $\boldsymbol{e}_s$ ovat tuntemattomia, joten käytetään estimoituja jäännöksiä $\hat{\boldsymbol{e}}_s=(\hat{e}_1,\ldots,\hat{e}_n)^T$, missä $\hat{e}_i=y_i-\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. Korvataan myös tuntematon $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ sen otokseen perustuvalla estimaattorilla $(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}$, jolloin saadaan Taylor-linearisointiin perustuva estimaattori regressiokertoimien estimaattorin kovarianssimatriisille

$$\text{Cov}(\hat{\boldsymbol{\beta}})=(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}\hat{\boldsymbol{G}}(\boldsymbol{X}_s^T\boldsymbol{W}_s\boldsymbol{X}_s)^{-1}. \tag{7}$$

Jos otanta-asetelma on yksinkertainen satunnaisotanta palauttamatta, niin matriisin $\hat{\boldsymbol{G}}$ alkio (k,l) on

$$\hat{g}_{kl}=\frac{n}{n-1}\sum_{i=1}^{n}\left[\left(x_{ik}\hat{e}_iw_i-\frac{1}{n}\sum_{i'=1}^{n}x_{i'k}\hat{e}_{i'}w_{i'}\right)\left(x_{il}\hat{e}_iw_i-\frac{1}{n}\sum_{i'=1}^{n}x_{i'l}\hat{e}_{i'}w_{i'}\right)\right].$$

Taylor-linearisoinnissa voidaan ottaa huomioon eri suurien poimintatodennäköisyyksien lisäksi erilaisia otanta-asetelmia. Oletetaan H kappaletta ositteita, joiden sisällä suoritetaan kaksiasteinen otanta. Ositteesta h valitaan m_h ensiasteen otantayksikköä, ja ensiasteen otantayksiköstä i valitaan n_{hi} toisen asteen otantayksikköä. Nyt kaavassa (7) matriisin $\hat{\boldsymbol{G}}$ alkio (k,l) on

$$\begin{aligned}\hat{g}_{kl}=\sum_{h=1}^{H}\Bigg[\frac{m_h}{m_h-1}\sum_{i=1}^{m_h}\Bigg(\Bigg(\sum_{j=1}^{n_{hi}}w_{hij}z_{hijk}-\frac{1}{m_h}\sum_{i'=1}^{m_h}\sum_{j=1}^{n_{hi'}}w_{hi'j}z_{hi'jk}\Bigg)\\ \times\Bigg(\sum_{j=1}^{n_{hi}}w_{hij}z_{hijl}-\frac{1}{m_h}\sum_{i'=1}^{m_h}\sum_{j=1}^{n_{hi'}}w_{hi'j}z_{hi'jl}\Bigg)\Bigg)\Bigg],\end{aligned}$$

missä w_{hij} on alkioon (h, i, j) liittyvä otantapaino ja $z_{hijk} = x_{hijk}(y_{hij} - \boldsymbol{x}_{hij}\hat{\boldsymbol{\beta}})$, $k = 1, \ldots, p$. Kaavan (7) muotoisia estimaattoreita kutsutaan usein sandwich-estimaattoreiksi (Goldstein 2011, 94).

Vaihtoehtoisesti regressiokertoimien varianssit voidaan laskea jollakin otoksen uudiskäytön menetelmällä, kuten bootstrap-, jackknife- tai BRR-menetelmällä. Näistä on kehitetty variaatioita erilaisia otanta-asetelmia varten ja luvussa 3.5 esitellään yksi versio BRR-menetelmästä. Bootstrap-menetelmä on kompleksisten otanta-aineistojen tapauksessa muita menetelmiä vähemmän käytetty, ja vertailevissa simulointikokeissa ja muissa empiirisissä tutkimuksissa on todettu, että useimmissa sovelluksissa suurilla otoksilla se ei tuo lisähyötyä verrattuna Taylor-linearisointiin, jackknife- tai BRR-menetelmään, jotka ovat robustimpia kuin bootstrap (Heeringa ym. 2010, 82).

3.3 Otantapainot sekamallissa

Kuten kiinteiden vaikutusten mallissa, myös sekamallissa saatetaan saada harhaisia estimaatteja, jos alkioiden eri suuria poimintatodennäköisyyksiä ei oteta huomioon. Harhaa voi esiintyä, vaikka otos olisi itsepainottuva, eli moniasteisessa otannassa hierarkiassa alimman tason alkioilla olisi samat poimintatodennäköisyydet, jos kuitenkin ylemmän tason alkioiden poimintatodennäköisyydet vaihtelevat (Pfeffermann ym. 1998). Kun oletetaan, että otantapainot ovat riippumattomia satunnaisvaikutuksista, voidaan painot sisällyttää sekamalliin käyttäen Goldsteinin kirjassa (2011) esitettyä menetelmää.

Seuraava esitys koskee varianssikomponenttimallia (5). Skaalataan ensin painot siten, että

$$\sum_{j=1}^{n_i} w_{j|i} = n_i \qquad \text{ja} \qquad \sum_{i=1}^{m} w_i = m \,,$$

missä i viittaa ensiasteen ja j toisen asteen alkioon. Tällöin toisen asteen ehdollisten painojen keskiarvo ja ensiasteen painojen keskiarvo ovat ykkösiä. Lopulliset toisen asteen painot muodostetaan kaavalla

$$w_{ij} = \frac{w_{j|i} w_i \sum_{i=1}^{m} n_i}{\sum_{i=1}^{m} \sum_{j=1}^{n_i} w_{j|i} w_i} \,,$$

joiden summa on n eli otoskoko ja keskiarvo 1. Jos ensiasteen otantayksiköiden painot eivät ole tiedossa, voidaan niiden tilalla käyttää painoja, jotka saadaan kaavalla

$$w_i' = \frac{m \left(\sum_{j=1}^{n_i} w_{ij} \right) \Big/ n_i}{\sum_{i=1}^{m} \left(\sum_{j=1}^{n_i} w_{ij} \right) \Big/ n_i} \,.$$

Olkoon $\boldsymbol{Z}_u$ satunnaisvaikutusten $\boldsymbol{u}$ $n \times q$ -design-matriisi ja $\boldsymbol{Z}_e$ jäännösten $n \times n$ -design-matriisi. Usein on $\boldsymbol{Z}_e = \boldsymbol{I}$. Muodostetaan uudet matriisit

$$\boldsymbol{Z}_u^* = \boldsymbol{W}_1 \boldsymbol{Z}_u \quad \text{ja} \quad \boldsymbol{Z}_e^* = \boldsymbol{W}_2 \boldsymbol{Z}_e \,,$$

missä $\boldsymbol{W}_1 = \text{diag}(w_1^{-1/2}, \ldots, w_i^{-1/2}, \ldots, w_m^{-1/2})$ ja $\boldsymbol{W}_2 = \text{diag}(w_{11}^{-1/2}, \ldots, w_{ij}^{-1/2}, \ldots, w_{mn_m}^{-1/2})$. Regressiokertoimet estimoidaan nyt kuten ilman otantapainoja, mutta käyttäen matriiseja $\boldsymbol{Z}_u^*$ ja $\boldsymbol{Z}_e^*$ matriisien $\boldsymbol{Z}_u$ ja $\boldsymbol{Z}_e$ sijaan. Mallin yhtälö on tällöin

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_u^* \boldsymbol{u} + \boldsymbol{Z}_e^* \boldsymbol{e} \,.$$

Merkitään $\boldsymbol{e}^* = \boldsymbol{Z}_e^* \boldsymbol{e}$. Uuden jäännösvektorin jakauma on

$$\boldsymbol{e}^* \sim N(0, \boldsymbol{Z}_e^* \boldsymbol{R} \boldsymbol{Z}_e^{*T})$$

ja vasteen jakauma

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{Z}_u^* \boldsymbol{D} \boldsymbol{Z}_u^{*T} + \boldsymbol{Z}_e^* \boldsymbol{R} \boldsymbol{Z}_e^{*T}).$$

Kiinteiden vaikutusten mallille tämä menetelmä antaisi tavallisen painotetun regressiomallin. Keskivirheet voidaan laskea käyttämällä sandwich-estimaattoreita tai jotakin otoksen uudiskäytön menetelmää.

Jos oletusta otantapainojen ja satunnaisvaikutusten riippumattomuudesta ei tehdä, tilanne tulee huomattavasti monimutkaisemmaksi. Pfeffermann ym. (1998) ehdottavat tässä tilanteessa pseudo-uskottavuuden ja PWIGLS-menetelmän (*probability-weighted iterative generalized least squares*) käyttöä, ja näyttävät, että myös edellä esitetty menetelmä tuottaa monissa tapauksissa kelvollisia tuloksia.

Käytännön ongelmana on, että edellä esitettyä menetelmää ei ole implementoitu esimerkiksi SAS-ohjelmistoon, jota tämän työn empiirisessä osassa käytetään. Näin ollen sekamallissa otantapainoja käytetään survey-tilastotieteen näkökulmasta väärin käsittämällä ne havaintokohtaisten varianssien käänteislukuina. Ensiasteen otantayksiköihin (kouluihin) liittyviä painoja ei saada sisällytettyä malliin.

3.4 Vaihtoehto otantapainojen käytölle

Oletetaan, että vastemuuttujan arvot y_{ij} ovat yhteydessä selittävien muuttujien arvoihin $\boldsymbol{x}_{ij}$ lineaarisen sekamallin kautta, joka on skalaarimuodossa esitettynä

$$y_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + u_i + e_{ij} \,, \tag{8}$$

missä $i = 1, \ldots, m$ viittaa ensiasteen alkioon ja $j = 1, \ldots, N_i$ toisen asteen alkioon. Oletetaan, että u_i ja e_{ij} ovat normaalisia ja toisistaan riippumattomia sekä $E(e_{ij}) = 0$ ja $\text{Var}(e_{ij}) = \sigma_e^2$.

Valitulla otanta-asetelmalla voi olla vaikutusta siihen, noudattavatko myös otoksen havainnot mallia (8). Esimerkiksi ositetussa otannassa ja PPS-otannassa käytetään jotakin apumuuttujaa otoksen valitsemisessa, jolloin kaikilla alkioilla ei välttämättä ole samaa poimintatodennäköisyyttä, vaan se riippuu joistakin muuttujista $\boldsymbol{X}^P$, jotka sisällytetään selittäjiksi malliin (8). Yläindeksi P viittaa koko populaatioon. On mahdollista, että poimintatodennäköisyys riippuu myös vastemuuttujasta $\boldsymbol{y}^P$. Tällaisissa tilanteissa otos ei suoraan noudata mallia (8), ja tämä voidaan ottaa huomioon käyttämällä otantapainoja.

Jos otanta ei riipu vastemuuttujasta, mutta voi riippua muuttujista $\boldsymbol{X}^P$, jotka sisältyvät malliin, niin otos noudattaa mallia (8), eikä otantapainoja tarvita. Tämä tulos esitetään seuraavaksi teosten Rao (2003, 78-79) ja Valliant ym. (2000, 36-40) mukaan. Kirjoitetaan malli (8) matriisimuodossa äärellisen populaation P osapopulaatiolle i:

$$\boldsymbol{y}_i^P = \boldsymbol{X}_i^P\boldsymbol{\beta} + u_i\mathbf{1}_i^P + \boldsymbol{e}_i^P \,, \tag{9}$$

missä $\boldsymbol{X}_i^P$ on $N_i \times p$ -matriisi ja $\mathbf{1}_i^P = (1,\ldots,1)^T$. Ositetaan malli (9) otokseen kuuluvaan ja kuulumattomaan osaan:

$$\boldsymbol{y}_i^P = \begin{bmatrix} \boldsymbol{y}_i \\ \boldsymbol{y}_i^* \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_i \\ \boldsymbol{X}_i^* \end{bmatrix}\boldsymbol{\beta} + u_i \begin{bmatrix} \mathbf{1}_i \\ \mathbf{1}_i^* \end{bmatrix} + \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i^* \end{bmatrix} \,,$$

missä yläindeksi * tarkoittaa otokseen kuulumatonta osaa. Jos tämä malli pätee otokselle, niin parametreihin $\boldsymbol{\psi} = (\boldsymbol{\beta}^T, \sigma_u^2, \sigma_e^2)^T$ liittyvä tilastollinen päättely perustuu otoksen tiheysfunktioon, joka on populaation tiheysfunktio integroituna otokseen kuulumattomien alkioiden suhteen:

$$f(\boldsymbol{y}_i|\boldsymbol{X}_i^P,\boldsymbol{\psi}) = \int f(\boldsymbol{y}_i,\boldsymbol{y}_i^*|\boldsymbol{X}_i^P,\boldsymbol{\psi})d\boldsymbol{y}_i^* \,.$$

Olkoon $\boldsymbol{a}_i = (a_{i1},\ldots,a_{iN_i})^T$ otoksen indikaattorivektori, jossa $a_{ij} = 1$, jos alkio j kuuluu otokseen ja muuten $a_{ij} = 0$. Jos otanta ei riipu vastemuuttujasta $\boldsymbol{y}_i^P$, eli pätee

$$f(\boldsymbol{a}_i|\boldsymbol{y}_i,\boldsymbol{y}_i^*,\boldsymbol{X}_i^P) = f(\boldsymbol{a}_i|\boldsymbol{X}_i^P) \,, \tag{10}$$

niin saadaan

$$\begin{aligned} f(\boldsymbol{y}_i,\boldsymbol{a}_i|\boldsymbol{X}_i^P,\boldsymbol{\psi}) &= f(\boldsymbol{y}_i|\boldsymbol{X}_i^P,\boldsymbol{\psi})f(\boldsymbol{a}_i|\boldsymbol{y}_i,\boldsymbol{X}_i^P) \\ &= \int f(\boldsymbol{y}_i,\boldsymbol{y}_i^*|\boldsymbol{X}_i^P,\boldsymbol{\psi})f(\boldsymbol{a}_i|\boldsymbol{y}_i,\boldsymbol{y}_i^*,\boldsymbol{X}_i^P)d\boldsymbol{y}_i^* \\ &= f(\boldsymbol{a}_i|\boldsymbol{X}_i^P)\int f(\boldsymbol{y}_i,\boldsymbol{y}_i^*|\boldsymbol{X}_i^P,\boldsymbol{\psi})d\boldsymbol{y}_i^* \\ &= f(\boldsymbol{a}_i|\boldsymbol{X}_i^P)f(\boldsymbol{y}_i|\boldsymbol{X}_i^P,\boldsymbol{\psi}) \,. \end{aligned}$$

Koska $f(\boldsymbol{a}_i|\boldsymbol{X}_i^P)$ ei riipu parametreista $\boldsymbol{\psi}$, niin se on vakio, ja siten esimerkiksi uskottavuusosamäärässä se supistuu pois. Näin ollen otos noudattaa mallia (9), ja parametreihin $\boldsymbol{\psi}$ liittyvä tilastollinen päättely voidaan tehdä tiheysfunktioon $f(\boldsymbol{y}_i|\boldsymbol{X}_i^P, \boldsymbol{\psi})$ perustuen, eikä otantamekanismia tarvitse ottaa muulla tavalla huomioon.

Useimmat otanta-asetelmat toteuttavat ehdon (10). Kyselytutkimuksissa on kuitenkin tyypillistä, ettei kaikilta otokseen valituilta yksiköiltä saada vastausta. Jos vastauskato riippuu vastemuuttujasta, niin ehto (10) ei ole enää voimassa.

Vaikka otanta-asetelma olisi mahdollista ottaa huomioon edellä esitetyllä tavalla käyttämällä asetelmamuuttujia kovariaatteina mallissa (8), on huomattava, että otantapainojen käyttö mallissa luvun 3.2 tai 3.3 tapaan tuo robustisuutta mallin väärinspesifioinnille (Pfeffermann ym. 1998). Toisin sanoen, jos malli on spesifioitu väärin, niin asetelman huomioon ottaminen painoilla kompensoi väärinspesifioinnista aiheutuvaa haittaa. Toisaalta jos malli on spesifioitu oikein, niin painojen käyttö vähentää estimoinnin tehokkuutta (Rabe-Hesketh ja Skrondal 2006). Tällöin siis tavallinen PNS-estimaattori on minimivarianssinen, kun taas painotettu PNS-estimaattori ei ole.

3.5 Tasapainotettujen puoliotosten menetelmä

Parametrien estimaatteihin liittyvä epävarmuus on suurempaa kaksiasteisessa otannassa kuin yksinkertaisessa satunnaisotannassa, jos klustereiden sisäiset havainnot eivät ole riippumattomia. Riippuvuus voidaan ottaa huomioon estimaatin otosvarianssia laskettaessa asetelmaperusteisesti käyttäen tasapainotettujen puoliotosten menetelmää, josta jatkossa käytetään lyhennettä BRR-menetelmä (*balanced repeated replication method*). Se on otoksen uudiskäytön menetelmä, ja sillä voidaan ottaa huomioon klusteroitumisen lisäksi myösositteet ja PPS-otanta.

Otoksen uudiskäytön menetelmiä on käytetty kansainvälisissä koulutustutkimuksissa otosvarianssien estimaattien laskemiseen IEA:n (*International Association for the Evaluation of Educational Achievement*) vuoden 1990 lukutaitotutkimuksesta lähtien (OECD 2009, 69). Etuna näissä menetelmissä on se, että niitä on suhteellisen helppo soveltaa erilaisissa kompleksisissa otanta-asetelmissa myös kompleksisille estimaattoreille (Valliant ym. 2000, 330). PISA-tutkimuksissa käytetään BRR-menetelmää, koska sillä on todettu olevan joitakin hyviä ominaisuuksia verrattuna jackknife-menetelmään, joka on yleisesti koulutustutkimuksissa käytetty otoksen uudiskäytön menetelmä. BRR-menetelmällä on vahvempi teoreettinen perusta kuin jackknife-menetelmällä epäsileiden funktioiden, kuten kvantiilien tapauksessa. Tällaisissa tilanteissa jackknife ei ole tarkentuva, kun taas BRR on (Rust ja Krawchuk 2002, 96).

BRR-menetelmän ideana on tuottaa havaintoaineistosta uusio-otoksia eli replikaatteja, joista jokaisesta lasketaan haluttu estimaatti, ja näiden estimaattien vaihtelusta saadaan otosvarianssi. Tässä tutkielmassa BRR-menetelmää

käytetään regressiokertoimien keskivirheiden estimointiin. Menetelmästä on olemassa eri variaatioita, ja seuraavaksi se esitellään teoksen OECD (2009, 74-75) mukaisesti sellaisena kuin sitä käytetään PISA-tutkimuksen otanta-asetelman tapauksessa.

Kutsutaan tässä luvussa ensiasteen otantayksiköitä lyhyemmin kouluiksi. Ensin järjestetään otokseen valitut koulut siten, että ne jaetaan ositteisiin ja ositteiden sisällä asetetaan suuruusjärjestykseen koon mukaan. Sitten muodostetaan pareja edellä luodun järjestyksen mukaan vierekkäisistä kouluista. Tällöin siis parin koulut ovat keskenään hyvin samankaltaisia, koska ne ovat samasta ositteesta ja lähes saman kokoisia. Näitä pareja kutsutaan pseudo-ositteiksi. Mikäli kouluja on pariton määrä ositteessa, viimeiset kolme muodostavat yhden pseudo-ositteen. Jos kouluja olisi otoksessa esimerkiksi sata, niin pseudo-ositteita olisi 50.

BRR-menetelmässä edetään valitsemalla jokaisesta pseudo-ositteesta yksi koulu, johon liittyvä otantapaino asetetaan nollaksi, ja toisen koulun paino kaksinkertaiseksi. Tällöin uuteen otokseen, niin kutsuttuun puoliotokseen, on valittu puolet kouluista, mutta otantapainojen summa pysyy keskimäärin ennallaan. Jos pseudo-ositteiden määrä on H, niin erilaisia puoliotoksia voitaisiin tehdä 2^H kappaletta. Puoliotoksia muodostetaan kuitenkin vain niin monta, että niiden lukumäärä G on pienin sellainen luvun neljä monikerta, joka on suurempi tai yhtä suuri kuin pseudo-ositteiden määrä. Jos esimerkiksi $H = 50$, niin $G = 52$.

Termillä *tasapainotettu* on tilastotieteessä erilaisia merkityksiä ja puoliotosten yhteydessä sillä tarkoitetaan uusio-otosten tasapainottamista Hadamard-matriisien avulla. Hadamard-matriisi on neliömatriisi, jonka alkiot ovat 1 tai -1 ja jonka rivit ovat ortogonaalisia. Matriisin aste on 1, 2 tai luvun neljä monikerta, mistä tulee vaatimus puoliotosten lukumäärälle. (Valliant ym. 2000, 333.)

Jos puoliotoksia muodostetaan esimerkiksi kahdeksan, niin voidaan käyttää 8×8 -Hadamard-matriisia:

$$\boldsymbol{A} = \begin{bmatrix} +1 & +1 & +1 & +1 & +1 & +1 & +1 & +1 \\ +1 & -1 & +1 & -1 & +1 & -1 & +1 & -1 \\ +1 & -1 & -1 & +1 & +1 & -1 & -1 & +1 \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & -1 & +1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & -1 & +1 & -1 & +1 & +1 & -1 \\ +1 & +1 & -1 & -1 & -1 & -1 & +1 & +1 \end{bmatrix}.$$

Hadamard-matriisit eivät ole yksikäsitteisiä. Tässäkin esimerkiksi rivit voitaisiin järjestää uudelleen, jolloin niiden ortogonaalisuus kuitenkin säilyy.

Määritetään puoliotokset matriisin rivien avulla. Jos matriisin alkio $a_{ij} = 1$, niin puoliotoksessa i pseudo-ositteen j ensimmäisen koulun paino kaksinkertaistetaan ja toisen koulun paino asetetaan nollaksi. Jos $a_{ij} = -1$, niin teh-

dään päinvastoin. Siten edellä olevan matriisin $\boldsymbol{A}$ mukaan ensimmäinen puoliotos koostuisi jokaisen pseudo-ositteen ensimmäisistä kouluista. Toinen puoliotos koostuisi pseudo-ositteiden 1, 3, 5 ja 7 ensimmäisistä kouluista ja pseudo-ositteiden 2, 4, 6 ja 8 toisista kouluista. Näin jatkamalla saataisiin kahdeksan tasapainotettua puoliotosta. Mikäli pseudo-ositteita on seitsemän, voidaan käyttää saman matriisin mitä tahansa seitsemää saraketta, ja saadaan jälleen kahdeksan tasapainotettua puoliotosta. Vastaavasti menetellään tilanteessa, jossa pseudo-ositteita on viisi tai kuusi.

Taulukossa 1 on esitetty, miten puoliotokset muodostuvat matriisin $\boldsymbol{A}$ rivien mukaan kahdeksan pseudo-ositteen tapauksessa. Esimerkissä on 16 koulua, joihin liittyvät otantapainot $w_1, \ldots, w_{16}$.

Taulukko 1: Esimerkki puoliotosten muodostamisesta.

		Puoliotos							
Koulu	Pseudo-osite	1	2	3	4	5	6	7	8
1	1	$2w_1$	$2w_1$	$2w_1$	$2w_1$	$2w_1$	$2w_1$	$2w_1$	$2w_1$
2		0	0	0	0	0	0	0	0
3	2	$2w_3$	0	0	$2w_3$	$2w_3$	0	0	$2w_3$
4		0	$2w_4$	$2w_4$	0	0	$2w_4$	$2w_4$	0
5	3	$2w_5$	$2w_5$	0	0	$2w_5$	$2w_5$	0	0
6		0	0	$2w_6$	$2w_6$	0	0	$2w_6$	$2w_6$
7	4	$2w_7$	0	$2w_7$	0	$2w_7$	0	$2w_7$	0
8		0	$2w_8$	0	$2w_8$	0	$2w_8$	0	$2w_8$
9	5	$2w_9$	$2w_9$	$2w_9$	$2w_9$	0	0	0	0
10		0	0	0	0	$2w_{10}$	$2w_{10}$	$2w_{10}$	$2w_{10}$
11	6	$2w_{11}$	0	0	$2w_{11}$	0	$2w_{11}$	$2w_{11}$	0
12		0	$2w_{12}$	$2w_{12}$	0	$2w_{12}$	0	0	$2w_{12}$
13	7	$2w_{13}$	$2w_{13}$	0	0	0	0	$2w_{13}$	$2w_{13}$
14		0	0	$2w_{14}$	$2w_{14}$	$2w_{14}$	$2w_{14}$	0	0
15	8	$2w_{15}$	0	$2w_{15}$	0	0	$2w_{15}$	0	$2w_{15}$
16		0	$2w_{16}$	0	$2w_{16}$	$2w_{16}$	0	$2w_{16}$	0

Saaduilla painojen replikaateilla lasketaan halutun parametrin (esimerkiksi regressiokertoimen) estimaatit $\hat{\beta}_i$ kaikilla $i = 1, \ldots, G$, missä G on replikaattien eli puoliotosten lukumäärä, ja näitä verrataan koko otoksesta laskettuun estimaattiin $\hat{\beta}$. Näin saadaan estimaatille otosvarianssi:

$$\hat{\sigma}^2 = \frac{1}{G} \sum_{i=1}^{G} (\hat{\beta}_i - \hat{\beta})^2,$$

jonka neliöjuuri on estimaatin keskivirhe.

3.5.1 Fayn modifikaatio

Edellä esitetty menetelmä käyttää jokaisessa uusio-otoksessa vain puolta koko otoksen havainnoista. Tästä voi aiheutua ongelmia pienten osajoukkojen

kohdalla. PISA-tutkimuksissa tämä ongelma vältetään käyttämällä Fayn modifikaatiota BRR-menetelmästä. Otantapainoja ei kerrotakaan luvuilla 0 tai 2, vaan luvuilla k ja $2-k$, missä $0 < k < 1$. Tällöin kaikki havainnot säilyvät uusio-otoksissa. Valinta $k = 0$ johtaisi BRR-menetelmän edellä esitettyyn modifioimattomaan muotoon. Nyt otosvarianssin kaavaksi saadaan

$$\hat{\sigma}^2 = \frac{1}{G(1-k)^2} \sum_{i=1}^{G} (\hat{\beta}_i - \hat{\beta})^2.$$

PISA-tutkimuksissa käytetään arvoa $k = 0.5$, jolloin koulujen painoja kerrotaan luvuilla 0.5 ja 1.5. Taulukosta 2 nähdään, miten edellisen esimerkin tilanteessa puoliotokset muodostetaan, kun käytetään Fayn modifikaatiota arvolla $k = 0.5$.

Taulukko 2: Esimerkki Fay-modifioitujen puoliotosten muodostamisesta.

		Puoliotos							
Koulu	Ps.-os.	1	2	3	4	5	6	7	8
1	1	$1.5w_1$	$1.5w_1$	$1.5w_1$	$1.5w_1$	$1.5w_1$	$1.5w_1$	$1.5w_1$	$1.5w_1$
2		$0.5w_2$	$0.5w_2$	$0.5w_2$	$0.5w_2$	$0.5w_2$	$0.5w_2$	$0.5w_2$	$0.5w_2$
3	2	$1.5w_3$	$0.5w_3$	$0.5w_3$	$1.5w_3$	$1.5w_3$	$0.5w_3$	$0.5w_3$	$1.5w_3$
4		$0.5w_4$	$1.5w_4$	$1.5w_4$	$0.5w_4$	$0.5w_4$	$1.5w_4$	$1.5w_4$	$0.5w_4$
5	3	$1.5w_5$	$1.5w_5$	$0.5w_5$	$0.5w_5$	$1.5w_5$	$1.5w_5$	$0.5w_5$	$0.5w_5$
6		$0.5w_6$	$0.5w_6$	$1.5w_6$	$1.5w_6$	$0.5w_6$	$0.5w_6$	$1.5w_6$	$1.5w_6$
7	4	$1.5w_7$	$0.5w_7$	$1.5w_7$	$0.5w_7$	$1.5w_7$	$0.5w_7$	$1.5w_7$	$0.5w_7$
8		$0.5w_8$	$1.5w_8$	$0.5w_8$	$1.5w_8$	$0.5w_8$	$1.5w_8$	$0.5w_8$	$1.5w_8$
9	5	$1.5w_9$	$1.5w_9$	$1.5w_9$	$1.5w_9$	$0.5w_9$	$0.5w_9$	$0.5w_9$	$0.5w_9$
10		$0.5w_{10}$	$0.5w_{10}$	$0.5w_{10}$	$0.5w_{10}$	$1.5w_{10}$	$1.5w_{10}$	$1.5w_{10}$	$1.5w_{10}$
11	6	$1.5w_{11}$	$0.5w_{11}$	$0.5w_{11}$	$1.5w_{11}$	$0.5w_{11}$	$1.5w_{11}$	$1.5w_{11}$	$0.5w_{11}$
12		$0.5w_{12}$	$1.5w_{12}$	$1.5w_{12}$	$0.5w_{12}$	$1.5w_{12}$	$0.5w_{12}$	$0.5w_{12}$	$1.5w_{12}$
13	7	$1.5w_{13}$	$1.5w_{13}$	$0.5w_{13}$	$0.5w_{13}$	$0.5w_{13}$	$0.5w_{13}$	$1.5w_{13}$	$1.5w_{13}$
14		$0.5w_{14}$	$0.5w_{14}$	$1.5w_{14}$	$1.5w_{14}$	$1.5w_{14}$	$1.5w_{14}$	$0.5w_{14}$	$0.5w_{14}$
15	8	$1.5w_{15}$	$0.5w_{15}$	$1.5w_{15}$	$0.5w_{15}$	$0.5w_{15}$	$1.5w_{15}$	$0.5w_{15}$	$1.5w_{15}$
16		$0.5w_{16}$	$1.5w_{16}$	$0.5w_{16}$	$1.5w_{16}$	$1.5w_{16}$	$0.5w_{16}$	$1.5w_{16}$	$0.5w_{16}$

4 Tutkimusaineisto

PISA (*Programme for International Student Assessment*) on kansainvälinen 15-vuotiaiden oppimistuloksia mittaava arviointiohjelma. Tässä tutkielmassa käytetään vuoden 2009 Suomen ja Saksan PISA-aineistoja. Kokonaisuudessaan Suomen aineisto sisältää tiedot 5810 oppilaasta, jotka on valittu 203 koulusta, ja Saksan aineisto 4979 oppilaasta, jotka ovat 226 koulusta.

Mielenkiinnon kohteena Suomen aineistolla tehtävissä analyyseissä on lukemisen strategioiden hallintaan vaikuttavat tekijät, ja sisällöllisistä syistä aineistosta rajataan osa oppilaista pois. Mukaan otetaan vain ne, jotka ovat tehneet kokeen suomen kielellä ja ovat yhdeksännellä luokalla, jotta kaikilla olisi sama äidinkieli ja kaikki olisivat saaneet suunnilleen yhtä paljon koulutusta. Rajauksen jälkeen aineistossa on 3859 oppilasta. Saksan aineistolla tehtäviin analyyseihin ei liity sisällöllistä mielenkiintoa, joten sille ei tehdä tällaisia rajauksia.

Lisäksi molemmista aineistoista poistetaan oppilaat, joilla on puuttuvaa tietoa jonkin käytössä olevan muuttujan kohdalla. Tämän jälkeen Suomen aineistossa on 3516 oppilasta, joista 1863 tyttöjä ja 1653 poikia sekä Saksan aineistossa 4078 oppilasta, joista tyttöjä 2096 ja poikia 1982. Suomesta on mukana 138 koulua ja Saksasta 213. Jatkossa kaikki analyysit ja muut tarkastelut koskevat näitä rajattuja aineistoja.

4.1 Yleisesti PISA-tutkimuksesta

PISA on OECD:n kolmen vuoden välein toteutettava kansainvälinen tutkimus, jonka kohderyhmänä on OECD-maiden ja useiden muiden maiden tai talousalueiden 15-vuotiaat oppilaat. Päätarkoituksena PISA-tutkimuksessa on selvittää, missä määrin lähellä pakollisen koulutuksen loppua olevat oppilaat ovat omaksuneet moderniin yhteiskuntaan täysipainoisesti osallistumisen kannalta oleellisia tietoja ja taitoja erityisesti lukutaidossa, matematiikassa ja luonnontieteissä. Vuoden 2009 tutkimukseen osallistui 34 OECD-maata ja 41 muuta maata tai talousaluetta. Yhdessä osallistujamaat edustavat lähes 90 % koko maailman taloudesta. Pääpaino tutkimuksessa on arvioida, miten oppilaat osaavat yleistää ja soveltaa oppimaansa niin koulussa kuin koulun ulkopuolella. (OECD 2010b.)

Joka kolmas vuosi järjestettävän tutkimuksen painopiste vaihtuu lukutaidon, matematiikan ja luonnontieteiden välillä tässä järjestyksessä. Ensimmäinen PISA-tutkimus toteutettiin vuonna 2000, jolloin painopisteenä oli lukutaito. Vuonna 2009 alkoi siis uusi kierros, mikä mahdollisti erityisesti lukutaidossa yhdeksän vuoden aikana tapahtuneiden muutosten tutkimisen. Tällöin 20 maassa arvioitiin ensimmäisen kerran myös digitaalisten tekstien lukutaitoa. Suomi ei kuitenkaan osallistunut tähän. Tulevaisuudessa PISA-tutkimuksissa tullaan käyttämään yhä enemmän digitaalisessa muodossa olevia ongelmia ja tekstejä, jotta tutkimus vastaisi paremmin informaatioyhteiskunnan haasteita. (OECD 2010b.)

PISA-tutkimuksessa oppilaat tekevät tehtäviä, joissa on sekä monivalinta- että avoimia tehtäviä. Lisäksi he vastaavat oppilaskyselyyn, jolla selvitetään heidän henkilökohtaisia taustojaan sekä opiskelutottumuksia, -asenteita ja -motivaatiota. Koulujen rehtorit vastaavat koulukyselyyn, jolla kartoitetaan koulun demografisia ominaisuuksia ja oppimisympäristön laatua.

PISA 2009 -tutkimus ei tarjoa pelkästään tarkkaa kuvaa 15-vuotiaiden lukutaidon tasosta, vaan myös lukemiseen liittyvistä asenteista, tottumuksista ja strategioista. Myös matematiikan ja luonnontieteiden osaamista arvioitiin PISA 2009 -tutkimuksessa, mutta selvästi pienemmässä mittakaavassa kuin lukutaitoa. (OECD 2010b.)

Tutkimustuloksia voivat hyödyntää useat eri tahot aina poliitikoista yksittäisen oppilaan vanhempiin. PISA tarjoaa tietoa, joka mahdollistaa erilaisten koulutusjärjestelmien tuloksellisuuden vertailun, ja tuloksia käytetäänkin poliittisen päätöksenteon tukena, kun kehitetään kansallisia koulutusjärjestelmiä. Toisaalta PISA-tutkimukset tuottavat sellaista tietoa oppimisesta, jota rehtorit, opettajat ja oppilaiden vanhemmat voivat hyödyntää omassa toiminnassaan.

4.2 PISA 2009 -tutkimuksen otanta

Perusjoukkona PISA-tutkimuksissa ovat lähtökohtaisesti 15-vuotiaat oppilaat. Käytännössä vuoden 2009 tutkimus oli suunniteltu toteutettavaksi suurimmassa osassa maita huhtikuussa, ja tällöin tarkka perusjoukko koostui oppilaista, joiden iät olivat 15 vuoden ja 3 kuukauden sekä 16 vuoden ja 2 kuukauden välillä, jolloin perusjoukko olisi voitu määritellä 1993 syntyneiksi oppilaiksi (OECD 2012). Suomessa perusjoukoksi määriteltiin aikaisintaan helmikuussa 1993 ja viimeistään tammikuussa 1994 syntyneet oppilaat (Sulkunen ym. 2010).

Otannassa oli vaatimuksena, että vähintään 95 % perusjoukosta on tavoitettavissa tutkimusta varten. Suomessa koulut valittiin peruskouluista sekä sellaisista lukioista ja ammatillisista oppilaitoksista, joissa oli kohderyhmään kuuluvia oppilaita. Otannan ulkopuolelle jätettiin kuitenkin erityiskoulut, joissa opiskeli yhteensä 2.3 % perusjoukosta. Suomessa otokseen päätyneistä oppilaista 91 % osallistui tutkimukseen. (Sulkunen ym. 2010.)

Otos poimittiin kaikissa maissa kaksiasteisella ositetulla otannalla (paitsi Venäjällä kolmiasteisella otannalla). Ensimmäinen aste koostui kouluista, joissa on 15-vuotiaita oppilaita. Koulut poimittiin systemaattisella PPS-otannalla, jossa poimintatodennäköisyys oli suhteessa koulun 15-vuotiaiden oppilaiden estimoituun määrään, mikäli se oli yli 35, muuten todennäköisyys oli suhteessa lukuun 35. Tämä tarkoittaa, että pienten koulujen poimintatodennäköisyydet olivat yhtä suuret. 15-vuotiaiden määrän estimaattina käytettiin Suomessa peruskouluissa koulun yhdeksäsluokkalaisten määrää ja toisen asteen oppilaitoksissa estimointi tehtiin oppilaan syntymävuoden ja -kuukauden mukaan.

Ennen koulujen poimintaa ne oli lähes kaikissa maissa jaettu toisensa poissulkeviin osiin, jotka oli muodostettu parantamaan otokseen perustuvien

estimaattien tarkkuutta. Ositetussa otannassa käytettiin suhteellista kiintiöintiä koulujen lukumäärien suhteen. Ositteita oli kahta eri tyyppiä: eksplisiittisiä ja implisiittisiä. Eksplisiittisessä osituksessa koulut jaetaan ositemuuttujan mukaisiin ryhmiin, joita käsitellään myöhemmin otannassa toisistaan riippumattomina. Implisiittisessä osituksessa koulut järjestetään implisiittisen ositemuuttujan mukaan kunkin eksplisiittisen ositteen sisällä.

Koska kaikissa maissa koulut valittiin systemaattisella PPS-otannalla, niin yhtenä implisiittisenä ositemuuttujana oli 15-vuotiaiden estimoitu määrä. Muuten ositemuuttujien käyttö oli eri maissa hyvin vaihtelevaa. Suomessa eksplisiittisinä ositemuuttujina käytettiin aluetta, sitä sijaitseeko koulu kaupunki- vai maaseutualueella, kieltä ja koulutyyppiä. Nämä muodostivat yhteensä 12 positetta, joista ruotsinkieliset olivat yliotostettuja. Suomen ositteet ja niiden otoskoot on esitetty taulukossa 3. Taulukon neljä viimeistä positetta jäävät aineiston rajausten takia kokonaan pois analyyseistä. Saksassa yhteensä 18 eksplisiittistä positetta muodostuivat koulutyypin ja osavaltion mukaan.

Toisen asteen otantayksiköt olivat valittujen koulujen oppilaita. Koulun kaikista 15-vuotiaista oppilaista poimittiin systemaattisesti 35 oppilasta kaikissa maissa. Jos 15-vuotiaita oli vähemmän kuin 35, niin kaikki otettiin otokseen. (OECD 2012.)

Taulukko 3: Suomen aineiston eksplisiittiset ositteet ja oppilaiden määrät otoksessa ositteittain ennen aineiston rajauksia.

Osite	Otoskoko
Suomenkielinen, Etelä-Suomi, kaupunkialue	1870
Suomenkielinen, Etelä-Suomi, maaseutualue	194
Suomenkielinen, Länsi-Suomi, kaupunkialue	844
Suomenkielinen, Länsi-Suomi, maaseutualue	255
Suomenkielinen, Itä-Suomi, kaupunkialue	388
Suomenkielinen, Itä-Suomi, maaseutualue	212
Suomenkielinen, Pohjois-Suomi, kaupunkialue	414
Suomenkielinen, Pohjois-Suomi, maaseutualue	197
Ruotsinkielinen, peruskoulu, kaupunkialue	1048
Ruotsinkielinen, peruskoulu, maaseutualue	381
Ruotsinkielinen, toisen asteen oppilaitos, kaupunkialue	6
Ruotsinkielinen, toisen asteen oppilaitos, maaseutualue	1

4.2.1 Otantapainojen muodostaminen

PISA-tutkimuksen otanta on suunniteltu siten, että jos vastauskatoa ei ole, niin jokaisen positteen sisällä oppilaiden poimintatodennäköisyydet ja siten myös otantapainot ovat yhtä suuria. Positteiden välillä painot ovat eri suuria, jos jollekin perusjoukon osajoukolle tehdään kustannussyistä aliotos tai lisätarkkuutta tuloksiin tavoiteltaessa yliotos. Otantapainoihin tulee myös vaihtelua,

jos 15-vuotiaiden estimoitu määrä osoittautuu epätarkaksi, eli otoksen poiminnassa käytetty arvio poikkeaa testaushetken todellisesta oppilasmäärästä. Sekä oppilas- että koulutason vastauskato otetaan huomioon otantapainojen korjauksissa. Lisäksi käytetään niin sanottua painojen hienosäätöä, jolla varmistetaan, ettei yksittäisiin havaintoihin kohdistu muista painojen korjauksista johtuvia odottamattoman suuria painoja. Painojen hienosäätö tuo estimaatteihin hieman harhaa, mutta pienentää huomattavasti keskivirheitä. (OECD 2012.)

Lopulliset oppilaspainot w_{ij} oppilaalle j koulussa i muodostetaan käyttäen koulupainoa, oppilaan koulun sisäistä painoa ja neljää korjauskerrointa. Oppilaspaino saadaan kaavalla

$$w_{ij} = w_i w_{j|i} f_i f_{ij} t_i t_{ij} \, ,$$

missä

w_i on koulupaino, koulun i poimintatodennäköisyyden käänteisluku,

$w_{j|i}$ on oppilaan koulun sisäinen paino, käänteisluku oppilaan j poimintatodennäköisyydestä, kun koulu i on valittu,

f_i on koulutason vastauskadon korjauskerroin, kompensoi sellaisen koulun poisjäämistä, joka on samankaltainen kuin koulu i,

f_{ij} on on oppilastason vastauskadon korjauskerroin, kompensoi sellaisen oppilaan poisjäämistä, joka on samankaltainen kuin oppilas ij,

t_i on koulupainon hienosäätökerroin, jolla pienennetään odottamattoman suurta koulupainoa ja

t_{ij} on oppilaspainon hienosäätökerroin, jolla pienennetään kaikkien edellisten tulona syntynyttä odottamattoman suurta oppilaspainoa.

4.3 Tutkimuksen kannalta kiinnostavat muuttujat

Selitettävänä muuttujana tämän tutkimuksen analyyseissä on *tietoisuus tekstin ymmärtämisen ja muistamisen strategioista*. Selittävinä muuttujina kiinnostavia ovat taulukon 4 neljätoista muuta muuttujaa. Jatkuvat muuttujat muodostettiin osioanalyysillä useista diskreeteistä muuttujista eli kysymyksistä, joissa on valmiit vastausvaihtoehdot. Lopulliset jatkuvat muuttujat standardoitiin kansainvälisesti siten, että kaikkien OECD-maiden keskiarvo on 0 ja keskihajonta 1 (OECD 2012). Koulutason muuttuja saatiin koulukyselyistä eli koulun rehtorin antamista vastauksista. Käytössä olevassa Saksan aineistossa ei ole koulutason muuttujia, joten *TEACBEHA* puuttuu Saksaa koskevista tarkasteluista.

Vastemuuttujana käytettävää tietoisuutta tekstin ymmärtämisen ja muistamisen strategioista mitattiin pyytämällä oppilaita arvioimaan annettujen lukemisstrategioiden hyödyllisyyttä kuusiportaisella asteikolla, kun tehtävänä on

Taulukko 4: Kiinnostavat muuttujat PISA 2009 -aineistossa.

	Muuttuja	**Nimi aineistossa**	**Tyyppi**
Luetun ymmärtämisen strategiat	Tietoisuus tekstin ymmärtämisen ja muistamisen strategioista	UNDREM (*Understanding and remembering*)	jatkuva
Opiskelustrategiat	Kontrollistrategioiden hyödyntäminen	CSTRAT (*Control strategies*)	jatkuva
	Mieleenpainamisstrategioiden hyödyntäminen	MEMOR (*Memorisation*)	jatkuva
	Elaborointistrategioiden hyödyntäminen	ELAB (*Elaboration*)	jatkuva
Koulua varten lukeminen	Kaunokirjallisuuden tulkinta	RFSINTRP (*Reading for school: interpretation of literary texts*)	jatkuva
	Epälineaarisia elementtejä sisältävien tekstien lukeminen	RFSNCONT (*Reading for school: use of texts containing non-continuous materials*)	jatkuva
	Kaunokirjallisuuden perinteiset oppisisällöt	RFSTRLIT (*Reading for school: reading activities for traditional literature courses*)	jatkuva
	Painetut lehti- ja ohjetekstit	RFSFUMAT (*Reading for school: use of functional texts*)	jatkuva
Koulutason muuttuja	Opettajien vaikutus koulun ilmapiiriin	TEACBEHA (*Teacher-related factors affecting school climate*)	jatkuva
Muut	Sukupuoli	ST04Q01	dikotominen
	Sosioekonominen asema	ESCS (*Economic, social and cultural status*)	jatkuva
	Kiinnostus lukemista kohtaan	JOYREAD (*Enjoyment of reading*)	jatkuva
	Lukemisen monipuolisuus	DIVREAD (*Diversity in reading*)	jatkuva
	Lukemiseen sitouttaminen	STIMREAD (*Teachers stimulation of reading engagement*)	jatkuva
	Kirjaston käyttö	LIBUSE (*Libraries*)	jatkuva

ymmärtää ja muistaa tekstin sisältämät tiedot. Lopullinen muuttuja muodostettiin vertaamalla oppilaiden arvioita strategioiden hyödyllisyydestä lukutaitoasiantuntijoiden arvioihin. Muita muuttujia koskevissa kysymyksissä pyydettiin kertomaan, miten hyvin väite pätee omalla kohdalla tai kuinka usein toimii kuvauksen mukaisesti. Muuttujien taustalla olevat kysymykset on esitetty liitteessä A.

Taulukossa 5 on esitetty muuttujien keskiarvot ja keskihajonnat. Suomessa oppilaiden sosioekonominen asema ja lukemisen monipuolisuus näyttäisi olevan selvästi OECD-maiden keskiarvoa parempi. Kaunokirjallisuuden tulkintaa on kouluissa taas melko vähän. Suomessa oppilaiden välillä on verrattain suur-

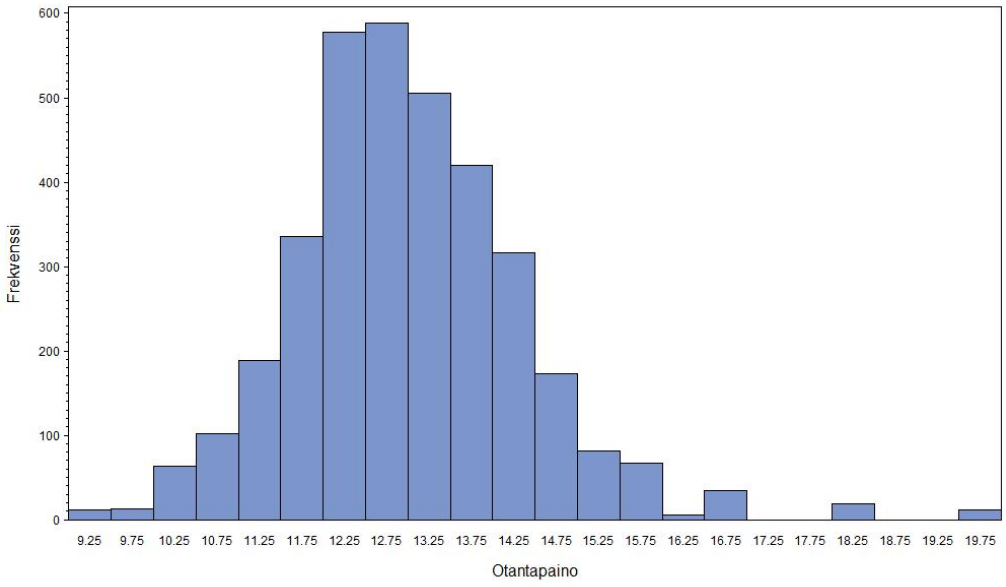
ta vaihtelua tietoisuudessa tekstin ymmärtämisen ja muistamisen strategioista sekä kiinnostuksessa lukemista kohtaan. Sen sijaan opettajien vaikutuksessa koulun ilmapiiriin ei ole kovin suurta vaihtelua.

Saksassa oppilaiden tietoisuus tekstin ymmärtämisen ja muistamisen strategioista näyttää olevan OECD-maiden keskiarvoa selvästi parempaa. Kirjastojen käyttö on huomattavan vähäistä, mutta siinä näyttäisi kuitenkin olevan suurta vaihtelua oppilaiden välillä. Kuten Suomessa myös Saksassa kiinnostus lukemista kohtaan vaihtelee melko paljon.

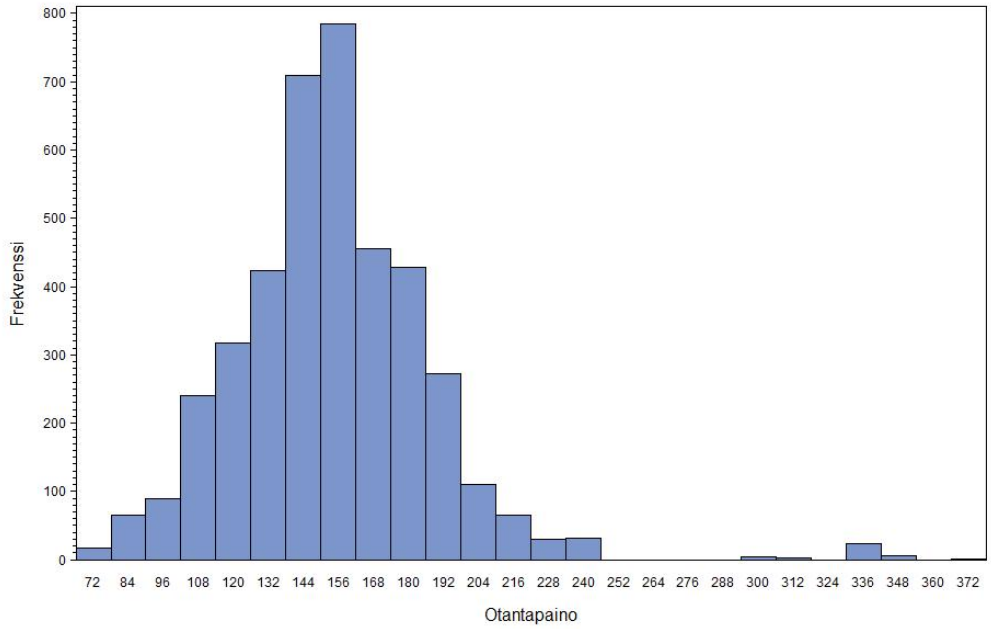
Taulukko 5: Muuttujien keskiarvot ja -hajonnat.

	Suomi		**Saksa**	
Muuttuja	Keskiarvo	Keskihajonta	Keskiarvo	Keskihajonta
UNDREM	0.075	1.011	0.322	0.998
CSTRAT	-0.324	0.962	0.228	0.948
MEMOR	-0.286	0.866	0.224	0.860
ELAB	-0.133	0.949	0.095	0.941
RFSINTRP	-0.406	0.922	0.157	0.889
RFSNCONT	0.069	0.964	0.180	0.957
RFSTRLIT	-0.218	0.930	-0.321	0.909
RFSFUMAT	-0.229	0.981	-0.091	0.866
TEACBEHA	-0.054	0.677		
ESCS	0.306	0.798	0.115	0.938
JOYREAD	0.082	1.041	0.115	1.192
DIVREAD	0.492	0.833	-0.146	0.928
STIMREAD	-0.332	0.790	-0.148	0.899
LIBUSE	-0.011	0.810	-0.500	1.047

Jatkossa tehtävissä analyyseissä keskeisessä osassa ovat otantapainot. Suomen aineistossa painojen keskiarvo on 13.012 ja keskihajonta 1.365 sekä Saksan aineistossa vastaavasti 154.266 ja 34.277. Vaikka PPS-otannalla pyritään saamaan oppilaiden lopulliset poimintatodennäköisyydet suunnilleen yhtä suuriksi, huomataan, että otantapainoissa on kuitenkin muun muassa vastauskadon korjauksista johtuvaa vaihtelua. Kuvissa 1 ja 2 on painojen jakaumia kuvaavat histogrammit.



Kuva 1: Otantapainot Suomen aineistossa.



Kuva 2: Otantapainot Saksan aineistossa.

5 Menetelmien vertailu

Seuraavaksi sovitetaan regressiomalleja kuudella eri tavalla ja vertaillaan niistä saatuja tuloksia. Menetelmien esittelyn jälkeen tutkitaan, johtavatko ne mallinvalinnassa erilaisiin tuloksiin. Sitten sovitetaan malli samoilla muuttujilla kullakin menetelmällä, ja vertaillaan tuloksia. Menetelmien välillä on eroja regressiokertoimien ja niiden keskivirheiden estimoinnissa. Näiden tunnuslukujen lisäksi vertaillaan myös saatuja regressiokertoimien p-arvoja.

5.1 Vertailtavat menetelmät

Vertailuissa käytettävät kuusi menetelmää ovat:

(a) REG-makro: kiinteiden vaikutusten malli painojen kanssa, keskivirheet lasketaan BRR-menetelmällä (luvut 2.1 ja 3.5)

(b) Kiinteiden vaikutusten malli painojen kanssa, keskivirheet lasketaan Taylor-linearisoinnilla (luvut 2.1 ja 3.2)

(c) Sekamalli painojen kanssa (luku 2.2, luvun 3.3 menetelmää ei implementoitu SAS-ohjelmistoon, joten painot käsitetään havaintokohtaisten varianssien käänteislukuina)

(d) Sekamalli ilman painoja, selittäjinä myös asetelmamuuttujat (osite ja yhdeksäsluokkalaisten määrä) (luvut 2.2 ja 3.4)

(e) MIXED-makro: sekamalli painojen kanssa, keskivirheet lasketaan BRR-menetelmällä (luvut 2.2 ja 3.5)

(f) Kiinteiden vaikutusten malli ilman painoja (luku 2.1)

PISA-aineistojen analysointiin on saatavilla valmiita makroja, jotka laskevat keskivirheet asetelmaperusteisesti Fay-modifioidulla BRR-menetelmällä. Makroja on SAS- ja SPSS-ohjelmistoille, ja ne on tuottanut Australian Council for Educational Research, joka vastaa PISAn tilastometodologiasta. Tässä työssä käytettävät SAS-makrot ovat saatavilla teoksesta OECD (2009). Makrot käyttävät kahdeksaakymmentä otantapainojen replikaattia eli puoliotosta, joissa on pseudo-ositteita muodostettaessa otettu huomioon PPS-otanta, ositteet ja oppilaiden klusteroitumisen kouluihin kuten luvussa 3.5 esitettiin.

Vertailuissa käytetään kahta makroa, joista menetelmässä (a) käytettävä REG-makro sovittaa kiinteiden vaikutusten regressiomalleja käyttäen kutakin otantapainojen replikaattia. Näiden mallien regressiokertoimien estimaattien vaihteluja käytetään estimaattien keskivirheinä. BRR-menetelmää käytetään vain keskivirheiden laskemiseen, ja regressiokertoimien estimaatit saadaan kiinteiden vaikutusten mallista käyttämällä varsinaisia otantapainoja. Menetelmässä (e) käytettävä MIXED-makro taas sovittaa lineaarisia sekamalleja, joissa oppilaiden klusteroituminen kouluissa otetaan huomioon asettamalla kouluille satunnainen tasoparametri. Keskivirheiden laskemisessa menetelmä

käyttää samoja otantapainojen replikaatteja kuin edellä mainittu REG-makro ja keskivirheet lasketaan vastaavalla tavalla molemmissa makroissa. Regressiokertoimien estimaatit saadaan sekamallista käyttämällä varsinaisia otantapainoja.

Taylor-linearisointi on paljon käytetty varianssin estimoinnin menetelmä otanta-aineistojen analysoinnissa. Menetelmässä (b) käytetään SAS-ohjelmiston SURVEYREG-proseduuria ja siinä regressiokertoimien keskivirheet lasketaan Taylor-linearisoinnilla, joka ottaa huomioon ositteet ja aineiston klusteroituneisuuden. Regressiokertoimien estimaatit lasketaan käyttämällä painoja kuten menetelmässä (a), ja siten niiden antamat estimaatit ovat täsmälleen samat. Saksan PISA 2009 -aineiston analyyseissä Taylor-linearisointi ei kuitenkaan ota huomioon ositteita, koska aineisto ei sisällä tietoa ositteista.

Menetelmä (c) on samanlainen lineaarinen sekamalli kuin menetelmä (e), mutta keskivirheitä ei lasketa BRR-menetelmällä, vaan ne saadaan suoraan mallista. Tässä käytetään SAS-ohjelmiston MIXED-proseduuria. Menetelmä käyttää estimoinnissa varsinaisia otantapainoja, joten regressiokertoimien estimaatit ovat täsmälleen samat kuin menetelmän (e) estimaatit. SAS-ohjelmistoon ei ole implementoitu sekamallia, joka käsittelisi painoja nimenomaan otantapainoina, kuten luvussa 3.3 esitettiin. Näin ollen painot käsitetään tässä havaintokohtaisten varianssien käänteislukuina, mikä on ainakin perinteisen survey-tilastotieteen näkökulmasta väärin.

Menetelmä (d) käyttää myös MIXED-proseduuria ja poikkeaa edellisestä vain siten, että otantapainojen sijaan malliin sisällytetään selittäjiksi asetelmamuuttujat, eli muuttujat, joita on hyödynnetty otoksen poiminnassa. Näitä ovat PPS-otannan perustana oleva yhdeksäsluokkalaisten määrä ja ositemuuttuja, joka jakaa oppilaat kahdeksaan ryhmään sen mukaan käykö koulua Etelä-, Länsi-, Itä- vai Pohjois-Suomessa ja sijaitseeko koulu kaupunki- vai maaseutualueella. Menetelmää (d) ei sovelleta ositetiedon puuttuessa Saksan aineistoon.

Menetelmässä (f) sovitetaan kiinteiden vaikutusten regressiomalli ilman otantapainoja MIXED-proseduurilla (ilman satunnaisvaikutuksia). Tämä malli ei ota otanta-asetelmaa millään tavalla huomioon ja on mukana vain vertailun vuoksi. Näin nähdään kuinka suuri ero tuloksissa on, jos otanta-asetelma jätetään analyysissä kokonaan huomiotta, ja näiden erojen valossa voidaan pohtia, kuinka merkittäviä erot ovat muilla menetelmillä saatujen tulosten välillä.

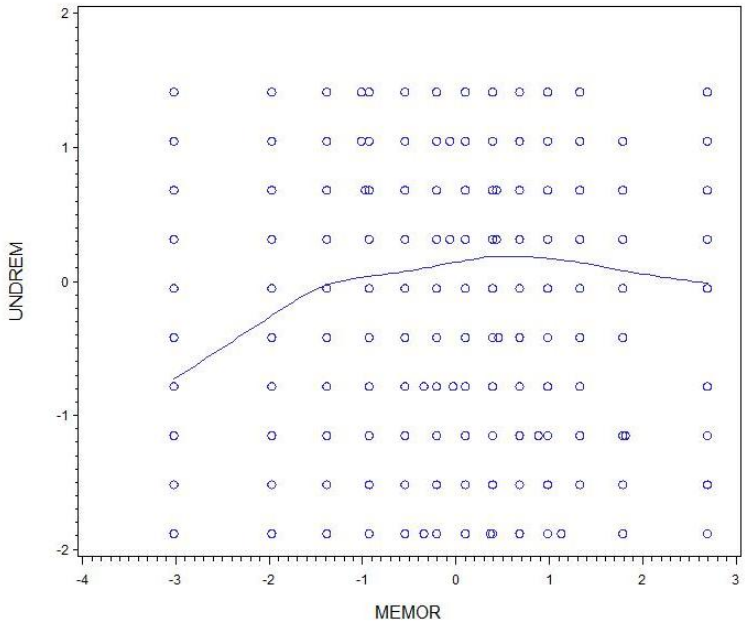
5.2 Mallinvalinta eri menetelmillä Suomen PISA 2009 -aineistolla

Toisin kuin Saksan aineistolla, Suomen aineistolla tehtäviin analyyseihin liittyy menetelmien vertailun lisäksi sisällöllinen tutkimusongelma. On kiinnostavaa selvittää, johtavatko eri menetelmät mallinvalinnassa sisällöllisesti huomattavasti erilaisiin malleihin vai päädytäänkö kaikilla samaan malliin.

Mallinvalinta tehdään jokaisella menetelmällä siten, että ensin selittäjinä ovat kaikki taulukon 4 muuttujat ja vastemuuttujana *UNDREM*. Lisäksi se-

littäjinä ovat kvadraattiset termit muuttujista, joiden vaikutusta vasteeseen ei voida kuvan perusteella pitää lineaarisena. Esimerkkinä tällaisesta on kuva 3, jossa suuren havaintomäärän ja havaintojen päällekkäisyyden takia on piirretty spline-käyrä havainnollistamaan muuttujien yhteyttä. Kvadraattiset termit tehdään muuttujille *JOYREAD*, *MEMOR*, *CSTRAT* ja *STIMREAD*.

Muuttujien valinta perustuu yksinkertaisesti p-arvoihin. Ei-merkitsevistä selittäjistä jätetään yksi kerrallaan pois se, jonka p-arvo on suurin, kunnes mallin kaikki selittäjät ovat merkitseviä. Työn myöhemmässä vaiheessa heräsi sisällöllinen mielenkiinto tutkia myös muuttujien interaktioita sukupuolen kanssa. Työmäärän rajoittamiseksi päätettiin tutkia näitä interaktioita vain niiden muuttujien kanssa, jotka edellä jäivät merkitseviksi. Jatketaan siis lisäämällä merkitseviksi jääneiden selittäjien parittaiset interaktiot sukupuolen kanssa. Jälleen edetään jättämällä pois ei-merkitseviä termejä kuten edellä, kunnes saadaan malli, jossa kaikki selittäjät ja mahdolliset interaktiot ovat merkitseviä. Tällainen p-arvoihin perustuva muuttujien valinta on tämän työn kannalta kiinnostava, koska mahdolliset menetelmien väliset erot regressiokertoimien estimaateissa ja niiden keskivirheissä johtavat eroihin myös p-arvoissa, ja näkyvät sitä kautta mallinvalinnan lopputuloksessa.



Kuva 3: Muuttujien *MEMOR* ja *UNDREM* hajontakuvio, johon on piirretty spline-käyrä havainnollistamaan muuttujien välistä yhteyttä.

Merkitseviksi jääneet muuttujat on esitetty taulukossa 6. Menetelmillä (a), (b), (c) ja (f) päädytään samaan malliin. Selittäjinä siinä ovat *GENDER*, *ESCS*, *JOYREAD* – lineaarinen, *JOYREAD* – kvadraattinen, *CSTRAT*, *MEMOR* – lineaarinen, *MEMOR* – kvadraattinen, *RFSFUMAT*, *RFSNCONT*,

STIMREAD – kvadraattinen ja interaktio (*STIMREAD* – kvadraattinen)**GENDER*.

Menetelmällä (d) saatu malli poikkeaa edellisestä siten, että *ESCS* jää pois ja *DIVREAD* jää malliin. Menetelmällä (e) saatava malli eroaa eniten muista. Millään muulla menetelmällä *LIBUSE* ei jää malliin. Se on myös ainoa menetelmä, jolla *STIMREAD* jää kokonaan pois.

Taulukko 6: Mallinvalinnan tulos kuudella eri menetelmällä Suomen aineistolla.

Menetelmät (a), (b), (c) ja (f)	Menetelmä (d)	Menetelmä (e)
GENDER	GENDER	GENDER
ESCS	JOYREAD – lin.	JOYREAD – lin.
JOYREAD – lin.	JOYREAD – kvadr.	JOYREAD – kvadr.
JOYREAD – kvadr.	CSTRAT	CSTRAT
CSTRAT	MEMOR – lin.	MEMOR – lin.
MEMOR – lin.	MEMOR – kvadr.	MEMOR – kvadr.
MEMOR – kvadr.	RFSFUMAT	RFSFUMAT
RFSFUMAT	RFSNCONT	RFSNCONT
RFSNCONT	STIMREAD – kvadr.	DIVREAD
STIMREAD – kvadr.	(STIMREAD – kvadr.)*	LIBUSE
(STIMREAD – kvadr.)*GENDER	GENDER	
	DIVREAD	

Menetelmät johtavat hieman erilaisiin malleihin, mutta syitä eroihin on vaikea keksiä. Näiden tulosten perusteella ei voida esimerkiksi sanoa, että ero johtuisi suoraan siitä, käyttääkö menetelmä kiinteiden vaikutusten mallia vai sekamallia. Kiinnostava huomio tosin on, että kaikki kiinteiden vaikutusten mallit johtavat samaan malliin, kun taas sekamalleilla päädytään erilaisiin lopputuloksiin. Itse asiassa kaikki kolme sekamallia johtavat erilaiseen malliin. Ehkä hieman yllättäen samaan malliin kuin kaikilla kiinteiden vaikutusten malleilla päädytään sekamalleista menetelmällä (c), jossa otantapainot käsitetään virheellisesti havaintokohtaisten varianssien käänteislukuina.

Erojen ei myöskään voida katsoa johtuvan siitä, lasketaanko keskivirheet BRR-menetelmällä vai ei. Menetelmällä (d) tehdyssä mallinvalinnassa muuttujan *ESCS* eli sosioekonomisen aseman poisjäämiseen saattaa vaikuttaa se, että mallissa on ositemuuttuja, joka on yhteydessä sosioekonomiseen asemaan. Tämä ei kuitenkaan selitä sitä, miksi *ESCS* jää pois myös menetelmällä (e).

5.3 Saman mallin vertailu eri menetelmillä

Menetelmiä vertaillaan seuraavaksi siten, että kullakin menetelmällä sovitetaan malli samoilla muuttujilla. Tämä tehdään sekä Suomen että Saksan aineistolla. Kiinnostava ero aineistojen välillä on se, että Saksassa oppilaiden klusteroituminen kouluihin on huomattavasti voimakkaampaa kuin Suomessa, joten Saksan analyyseissä tämän huomioon ottaminen on tärkeämpää.

Mallit sovitetaan Suomen aineistolla niillä muuttujilla, jotka jäävät malliin edellisen luvun mallinvalinnassa menetelmällä (a). Tätä menetelmää voidaan pitää kelvollisena vertailukohtana, koska sen käyttö on vakiintunut PISA-aineistojen analysoinnissa, ja se käyttää BRR-menetelmää, jota suositellaan käytettävän PISA-tutkimuksissa keskivirheiden laskemiseen. Myös Saksan aineistoon sovitettavien mallien muuttujat valitaan käyttäen menetelmää (a). Mallinvalinnassa ei kuitenkaan tutkita interaktioita. Saksan aineistolla tehtävistä vertailuista jätetään menetelmä (d) kokonaan pois, koska ositemuuttuja ei ollut helposti saatavilla.

5.3.1 Vertailu Suomen PISA 2009 -aineistolla

Taulukoissa 7-9 on esitetty eri tavoilla saadut regressiokertoimien estimaatit, keskivirheet ja p-arvot. Regressiokertoimissa on vain pieniä eroja, eikä mikään menetelmä näyttäisi antavan systemaattisesti joko suurempia tai pienempiä estimaatteja kuin jokin toinen. Menetelmän (d) muista poikkeava vakioparametri johtuu siitä, että mallissa on mukana selittävänä muuttujana ositemuuttuja, jonka vertailuryhmänä ovat Pohjois-Suomen maaseutukoulut. Ositemuuttuja ei ole tilastollisesti merkitsevä selittäjä ($F = 0.47, df = (7, 127), p = 0.856$), kuten ei myöskään yhdeksäsluokkalaisten määrä ($F = 2.42, df = (1, 135), p = 0.122$), joka on toinen asetelmamuuttuja.

Nähdään, että kiinteiden vaikutusten malleja käyttävillä menetelmillä (a) ja (b) sekä sekamalleja käyttävillä menetelmillä (c) ja (e) saadaan täsmälleen samat regressiokertoimien estimaatit, ja näin kuuluukin olla, koska menetelmien välillä on eroa vain keskivirheen laskemisessa. Tästä syystä menetelmien (a) ja (f) välisistä eroista nähdään, miten regressiokertoimet muuttuvat, kun kiinteiden vaikutusten regressiomalliin lisätään otantapainot.

Menetelmiin liittyvä kiinnostava huomio on se, että sekamallit antavat keskenään samanlaisia regressiokertoimien estimaatteja ja samoin käy kiinteiden vaikutusten mallien kohdalla. Sekamalli antaa keskimääräisen koulukohtaisen mallin, kun taas kiinteiden vaikutusten malli ei ota regressiokertoimissa huomioon aineiston hierarkkista rakennetta, eli se on malli kaikille perusjoukon yksilöille. Näin ollen mallit eivät edes mallinna samaa asiaa, mutta kuten tuloksista huomataan, niin Suomen aineisto on rakenteeltaan sellainen, ettei mallien välillä ole eroja. Tätä havainnollistetaan hieman jäljempänä kuvassa 4.

Kun vertaillaan keskivirheitä taulukosta 8, huomataan, että menetelmien väliset erot ovat hyvin pieniä, eikä muista poikkeavia keskivirheitä saada edes menetelmällä (f), joka ei ota otanta-asetelmaa millään tavalla huomioon. Itse asiassa sekamallia käyttävä menetelmä (c) ja menetelmä (f) antavat vakioparametria lukuunottamatta kaikille kertoimille kolmen desimaalin tarkkuudella saman keskivirheen. Näyttäisi siis siltä, että keskivirheiden osalta ei ole juurikaan väliä, otetaanko oppilaiden klusteroitumista kouluihin ollenkaan huomioon.

Taulukon 9 p-arvoissa ei myöskään ole huomattavia eroja menetelmien välillä. Klusteroituneen aineiston tapauksessa menetelmä (f) olettaa liian suuren

tehokkaan otoskoon. On oletettavaa, että tämä malli antaa liian pieniä keskivirheitä ja edelleen liian pieniä p-arvoja, mutta näin ei nyt näytä käyvän. Selitys lienee siinä, että Suomessa kouluihin klusteroituminen ei ole kovin voimakasta, eli tehokas otoskoko on lähes sama kuin aineiston havaintomäärä. Kun sovitetaan tyhjä sekamalli, jossa ei ole kiinteitä vaikutuksia ja vasteena on tietoisuus lukemisen strategioista, saadaan sisäkorrelaatioksi 0.039. Tämä tarkoittaa, että koulujen välinen vaihtelu kattaa kokonaisvaihtelusta vain noin 4 %, eikä koulujen välillä näin ollen juurikaan ole eroja. Jos sisäkorrelaatio lasketaan silloin, kun mallissa on merkitsevät selittäjät mukana, saadaan 0.030, joten mallin selittäjät selittävät osan koulujen välisestä vaihtelusta. Jos sisäkorrelaatio olisi 0, niin sekamalli ja kiinteiden vaikutusten malli olisivat matemaattisesti yhtäpitäviä.

Taulukko 7: Regressiokertoimien estimaatit kuudella eri menetelmällä Suomen aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaik. malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(d) Sekamalli asetelma-muuttujien kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiinteiden vaikutusten malli
vakiotermi	-0.036	-0.036	-0.043	-0.196	-0.043	-0.037
GENDER(tyttö)	0.331	0.331	0.338	0.341	0.338	0.333
ESCS	0.049	0.049	0.047	0.043	0.047	0.049
JOYREAD – lin.	0.193	0.193	0.185	0.187	0.185	0.193
JOYREAD – kvadr.	-0.024	-0.024	-0.022	-0.022	-0.022	-0.024
CSTRAT	0.208	0.208	0.203	0.200	0.203	0.208
MEMOR – lin.	-0.095	-0.095	-0.090	-0.090	-0.090	-0.096
MEMOR – kvadr.	-0.035	-0.035	-0.035	-0.036	-0.035	-0.036
RFSFUMAT	-0.069	-0.069	-0.073	-0.074	-0.073	-0.070
RFSNCONT	0.048	0.048	0.051	0.051	0.051	0.049
STIMREAD – kvadr.(tyttö)	0.000	0.000	-0.001	-0.001	-0.001	-0.002
STIMREAD – kvadr.(poika)	-0.050	-0.050	-0.049	-0.049	-0.049	-0.049

Taulukko 8: Keskivirheet kuudella eri menetelmällä Suomen aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaik. malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(d) Sekamalli asetelma-muuttujien kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiinteiden vaikutusten malli
vakiotermi	0.035	0.036	0.032	0.102	0.030	0.029
GENDER(tyttö)	0.041	0.041	0.038	0.038	0.040	0.038
ESCS	0.021	0.022	0.020	0.020	0.020	0.020
JOYREAD – lin.	0.019	0.019	0.018	0.018	0.019	0.018
JOYREAD – kvadr.	0.008	0.008	0.009	0.009	0.007	0.009
CSTRAT	0.021	0.022	0.021	0.021	0.021	0.021
MEMOR – lin.	0.024	0.023	0.023	0.023	0.024	0.023
MEMOR – kvadr.	0.012	0.013	0.011	0.011	0.011	0.011
RFSFUMAT	0.016	0.016	0.017	0.017	0.016	0.017
RFSNCONT	0.019	0.017	0.017	0.018	0.018	0.017
STIMREAD – kvadr.(tyttö)	0.014	0.014	0.018	0.018	0.015	0.018
STIMREAD – kvadr.(poika)	0.019	0.016	0.015	0.015	0.019	0.015

Taulukko 9: p-arvot kuudella eri menetelmällä Suomen aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaik. malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(d) Sekamalli asetelma-muuttujien kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiinteiden vaikutusten malli
vakiotermi	0.293	0.319	0.168	0.057	0.142	0.203
GENDER(tyttö)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
ESCS	0.018	0.028	0.020	0.035	0.022	0.013
JOYREAD – lin.	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
JOYREAD – kvadr.	0.002	0.003	0.016	0.016	0.003	0.010
CSTRAT	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
MEMOR – lin.	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
MEMOR – kvadr.	0.003	0.007	0.002	0.001	0.002	0.001
RFSFUMAT	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
RFSNCONT	0.011	0.004	0.004	0.003	0.005	0.005
STIMREAD – kvadr.(tyttö)	0.993	0.993	0.968	0.949	0.964	0.917
STIMREAD – kvadr.(poika)	0.008	0.002	0.001	0.001	0.010	0.001

5.3.2 Vertailu Saksan PISA 2009 -aineistolla

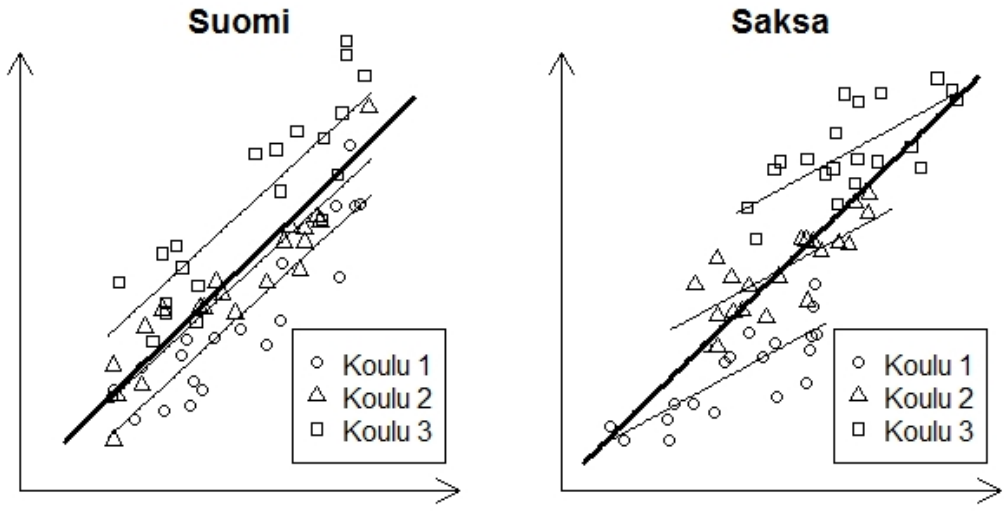
Regressiokertoimien estimaatteja vertailtaessa taulukosta 10 nähdään jälleen, että sekamalleilla eli menetelmillä (c) ja (e) saadaan täsmälleen samat estimaatit, niin kuin kuuluukin olla. Kiinteiden vaikutusten malleista menetelmillä (a) ja (b) pitääkin saada samat estimaatit, ja huomataan, että menetelmällä (f) ne ovat myös suunnilleen samat. Sekamalleja käyttävillä menetelmillä saadaan pääasiassa itseisarvoltaan pienempiä estimaatteja kuin kiinteiden vaikutusten malleja käyttävillä menetelmillä.

Taulukko 10: Regressiokertoimien estimaatit viidellä eri menetelmällä Saksan aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaikutusten malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiint. vaikutusten malli ilman painoja
vakiotermi	0.331	0.331	0.287	0.287	0.318
GENDER(tyttö)	0.097	0.097	0.120	0.120	0.093
ESCS	0.107	0.107	0.057	0.057	0.110
JOYREAD	0.129	0.129	0.104	0.104	0.130
DIVREAD – kvadr.	-0.016	-0.016	-0.013	-0.013	-0.013
CSTRAT – lin.	0.259	0.259	0.226	0.226	0.253
CSTRAT – kvadr.	-0.044	-0.044	-0.035	-0.035	-0.041
STIMREAD – kvadr.	-0.043	-0.043	-0.039	-0.039	-0.043
MEMOR	-0.152	-0.152	-0.136	-0.136	-0.151
ELAB	-0.061	-0.061	-0.047	-0.047	-0.060
LIBUSE – lin.	-0.162	-0.162	-0.128	-0.128	-0.164
LIBUSE – kvadr.	-0.080	-0.080	-0.061	-0.061	-0.084
RFSFUMAT – lin.	-0.076	-0.076	-0.070	-0.070	-0.068
RFSFUMAT – kvadr.	-0.028	-0.028	-0.026	-0.026	-0.026
RFSINTRP	0.087	0.087	0.060	0.060	0.091
RFSTRLIT	-0.038	-0.038	-0.012	-0.012	-0.050

Saksan tuloksissa tulee selvemmin esille, että sekamalli ja kiinteiden vaikutusten malli mallintavat hieman eri asioita, ja itse asiassa vastaavat eri kysymyksiin. Sekamalli sallii sen mahdollisuuden, että mallin vakioparametri vaihtelee koulujen välillä, ja näin jokainen koulu saa oman mallin. Tulokset kertovat, millainen tämä koulukohtainen malli on. Kiinteiden vaikutusten mallissa taas regressiokertoimien estimaatit on laskettu ottamatta huomioon aineiston hierarkkista rakennetta, jolloin malli kertoo selittäjien vaikutuksen koko valtion tasolla.

Se, että sekamallissa regressio on loivempi kuin kiinteiden vaikutusten mallissa, kertoo siitä, että Saksassa koulujen sisällä on vähemmän vaihtelua kuin valtion tasolla. Toisin sanoen Saksassa koulujen välillä on selviä eroja vastemuuttujan osalta. Suomessa taas selittävien muuttujien vaikutus on samanlaista koulu- ja valtiotasoilla, eikä Saksan kaltaisia eroja ole havaittavissa.



Kuva 4: Suomen ja Saksan aineistojen rakenteiden hahmottelua. Paksut suorat kuvaavat mallia koko valtion tasolla (kiinteiden vaikutusten malli), ohuet koulukohtaisia malleja (sekamalli).

Kuva 4 on hahmotelma Suomen ja Saksan aineistojen rakenteellista erosta, joka paljastuu vertailtaessa kiinteiden vaikutusten mallin ja sekamallin estimaatteja. Paksu suora edustaa kiinteiden vaikutusten mallia eli mallia koko valtion tasolla, ja ohuet suorat kuvaavat sekamalleja eli koulukohtaisia malleja. Kuvaa on yksinkertaistettu jättämällä muuttujien nimet ja asteikot pois ja esittämällä vain kolme koulua, jotta tilanne tulisi selkeämmin esille.

Suomen kuvassa koulukohtaisten suorien välillä on hyvin pieniä eroja, ja ne ovat yhdensuuntaisia kiinteiden vaikutusten mallin kanssa, mikä havaitaan siitä, että regressiokertoimien estimaateissa ei ole menetelmien välisiä eroja. Suomessa sisäkorrelaatio on pieni, joten koulukohtaiset suorat myös kattavat melkein koko vaihteluvälin vastemuuttujan osalta. Saksan kuvassa taas näkyy kiinteiden vaikutusten mallin ja sekamallin välillä todettu regressiokertoimien ero. Koulutason mallissa regressiokertoimet ovat pienempiä kuin koko valtiota koskevassa mallissa. Tyhjällä sekamallilla Saksan sisäkorrelaatioksi saadaan 0.192, eli toiset koulut ovat vastemuuttujan osalta selvästi parempia kuin toiset, mikä näkyy myös kuvasta. Koska Saksassa koulujen sisäinen vaihtelu on pientä suhteessa kokonaisvaihteluun, rajoittuvat koulukohtaiset mallit vastemuuttujan osalta melko homogeeniseen joukkoon, minkä takia sekamallin kulmakertoimet ovat pienempiä kuin kiinteiden vaikutusten mallin kertoimet.

Taulukko 11: Keskivirheet viidellä eri menetelmällä Saksan aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaikutusten malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiint. vaikutusten malli ilman painoja
vakiotermi	0.032	0.034	0.037	0.030	0.031
GENDER(tyttö)	0.033	0.034	0.032	0.031	0.032
ESCS	0.015	0.017	0.017	0.017	0.016
JOYREAD	0.012	0.014	0.014	0.013	0.015
DIVREAD – kvadr.	0.006	0.007	0.007	0.006	0.007
CSTRAT – lin.	0.023	0.022	0.020	0.023	0.020
CSTRAT – kvadr.	0.009	0.009	0.009	0.008	0.009
STIMREAD – kvadr.	0.010	0.010	0.009	0.010	0.009
MEMOR	0.020	0.022	0.019	0.020	0.020
ELAB	0.020	0.019	0.017	0.020	0.018
LIBUSE – lin.	0.024	0.024	0.022	0.025	0.023
LIBUSE – kvadr.	0.017	0.016	0.015	0.017	0.015
RFSFUMAT – lin.	0.020	0.021	0.018	0.017	0.018
RFSFUMAT – kvadr.	0.012	0.013	0.013	0.012	0.013
RFSINTRP	0.020	0.021	0.018	0.020	0.018
RFSTRLIT	0.019	0.020	0.018	0.017	0.018

Taulukon 11 keskivirheissä ei ole selviä eroja ja erot joidenkin p-arvojen välillä taulukossa 12 johtuvat pääasiassa edellä todetuista regressiokertoimien eroista. Kuitenkin klusteroituneen aineiston tapauksessa menetelmän (f) voisi olettaa antavan liian pieniä keskivirheitä. Yksi syy siihen, miksi näin ei käy, on se, että mallin selittävät muuttujat selittävät huomattavan osan koulujen välisestä vaihtelusta. Toisin sanoen, kun havainnot on vakioitu merkitsevien selittäjien suhteen, niin koulujen välille jää vain vähän selittämätöntä vaihtelua. Tyhjällä sekamallilla sisäkorrelaatioksi saadaan 0.192, mutta kun selittävät muuttujat ovat mallissa, niin sisäkorrelaatio on enää vain 0.095. Tämä on noin puolet alkuperäisestä sisäkorrelaatiosta, ja tulosten perusteella on käytännössä yhdentekevää, otetaanko kouluihin klusteroitumista ollenkaan huomioon laskettaessa keskivirheitä.

Taulukko 12: p-arvot viidellä eri menetelmällä Saksan aineistolla.

	(a) REG-makro (BRR-menetelmä)	(b) Kiint. vaikutusten malli, painot ja Taylor-lin.	(c) Sekamalli painojen kanssa	(e) MIXED-makro (BRR-menetelmä)	(f) Kiint. vaikutusten malli ilman painoja
vakiotermi	<0.001	<0.001	<0.001	<0.001	<0.001
GENDER(tyttö)	0.003	0.004	<0.001	<0.001	0.004
ESCS	<0.001	<0.001	0.001	0.001	<0.001
JOYREAD	<0.001	<0.001	<0.001	<0.001	<0.001
DIVREAD – kvadr.	0.015	0.031	0.056	0.032	0.052
CSTRAT – lin.	<0.001	<0.001	<0.001	<0.001	<0.001
CSTRAT – kvadr.	<0.001	<0.001	<0.001	<0.001	<0.001
STIMREAD – kvadr.	<0.001	<0.001	<0.001	<0.001	<0.001
MEMOR	<0.001	<0.001	<0.001	<0.001	<0.001
ELAB	0.003	0.002	0.006	0.017	0.001
LIBUSE – lin.	<0.001	<0.001	<0.001	<0.001	<0.001
LIBUSE – kvadr.	<0.001	<0.001	<0.001	<0.001	<0.001
RFSFUMAT – lin.	<0.001	<0.001	<0.001	<0.001	<0.001
RFSFUMAT – kvadr.	0.025	0.037	0.037	0.022	0.043
RFSINTRP	<0.001	<0.001	0.001	0.002	<0.001
RFSTRLIT	0.045	0.056	0.513	0.501	0.005

5.4 Mallin tulkinta

Esitetään seuraavaksi tulkinta Suomen aineistolla menetelmällä (a) saaduille tuloksille. Ensin käsitellään parametriestimaattien tulkintaa sellaisenaan, ja jäljempänä omassa alaluvussaan syvennytään tulosten merkitykseen käytännön kannalta. Taulukosta 7 nähdään mallin regressiokertoimien estimaatit. Kaikki estimaatit, paitsi muuttujan *STIMREAD* estimaatti tytöillä, ovat tilastollisesti merkitseviä, kuten muillakin menetelmillä.

Selittävillä muuttujilla *ESCS*, *JOYREAD*, *CSTRAT* ja *RFSNCONT* on positiivinen vaikutus ja muuttujalla *RFSFUMAT* negatiivinen vaikutus vastemuuttujaan *UNDREM*. Muuttujien *MEMOR* ja *STIMREAD* kasvu vaikuttaa vasteeseen ensin positiivisesti ja sitten negatiivisesti. Tulkinnoissa on hyvä muistaa, että sukupuolta lukuunottamatta muuttujat on standardoitu siten, että OECD-maiden keskiarvo on 0 ja keskihajonta 1. Vakiotermi -0.036 kuvaa vastemuuttujan tasoa pojilla, kun muiden selittävien muuttujien arvot ovat 0. Tytöillä *UNDREM* on keskimäärin 0.331 yksikköä korkeampi. Kun *ESCS* kasvaa yhden yksikön (ja muiden selittäjien arvot pysyvät ennallaan), niin vasteen arvo kasvaa keskimäärin 0.049 yksikköä.

Muuttujan *JOYREAD* vaikutus on käyräviivaista, joten sillä on mallissa myös kvadraattinen termi. Tulkinta menee siten, että kun muuttujan *JOYREAD* arvo kasvaa nollasta ykköseen, niin vastemuuttujan arvo kasvaa

$$0.193 \cdot JOYREAD - 0.024 \cdot (JOYREAD)^2 = 0.193 \cdot 1 - 0.024 \cdot 1^2 = 0.169$$

yksikköä. Kun *JOYREAD* kasvaa ykkösestä kakkoseen, nousee muuttujan *UNDREM* arvo

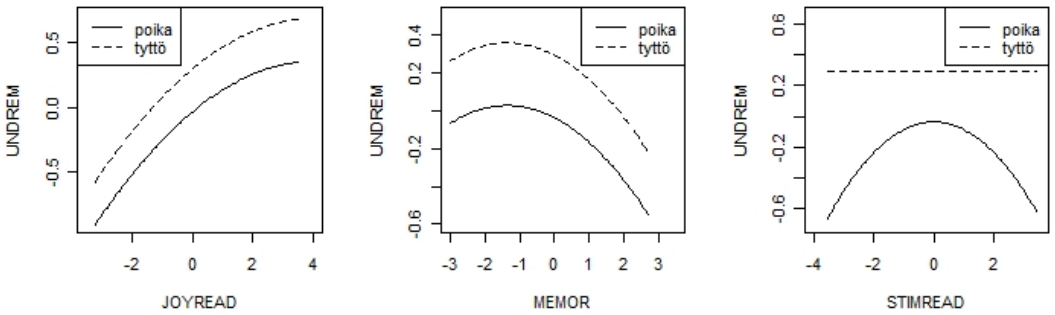
$$0.193 \cdot 2 - 0.024 \cdot 2^2 - 0.169 = 0.121$$

yksikköä. Muuttujan *JOYREAD* kasvaessa sen vaikutus vasteeseen siis vähenee.

Opiskelustrategioista muuttujalla *CSTRAT* on positiivinen vaikutus vasteeseen ja muuttujan *MEMOR* vaikutus on käyräviivaista. Kun *CSTRAT* kasvaa yhden yksikön, niin *UNDREM* kasvaa keskimäärin 0.208 yksikköä. Mallin mukaan *MEMOR* vaikuttaa vasteeseen positiivisesti, kun sen arvo on pienempi kuin -1.36, jonka jälkeen sen kasvu vaikuttaa negatiivisesti vasteeseen.

Koulun lukemistehtäviin liittyvistä muuttujista muuttujalla *RFSFUMAT* on negatiivinen ja muuttujalla *RFSNCONT* positiivinen vaikutus vastemuuttujaan. Muuttujan *RFSFUMAT* kasvaessa yhdellä *UNDREM* vähenee 0.069 yksikköä. Muuttujassa *RFSNCONT* yhden yksikön lisäys aiheuttaa keskimäärin 0.048 yksikön lisäyksen vasteessa.

STIMREAD on ainoa muuttuja, jolla on merkitsevä interaktio sukupuolen kanssa. Tytöillä muuttujalla *STIMREAD* ei ole mitään vaikutusta vastemuuttujaan, mutta pojilla sen vaikutus on käyräviivaista. Mallin mukaan se vaikuttaa pojilla positiivisesti, kun sen arvo on pienempi kuin nolla, jonka jälkeen vaikutus vasteeseen on negatiivista. Kuvasta 5 nähdään, millaista käyräviivaisten muuttujien vaikutus mallin mukaan on, kun muiden muuttujien arvot on kiinnitetty nollaksi. Muuttujilla *JOYREAD* ja *MEMOR* ei ole merkitsevää interaktiota sukupuolen kanssa, joten niiden vaikutus on samanlaista tytöillä ja pojilla; tyttöjä koskeva käyrä kulkee vain 0.331 yksikköä poikien käyrää korkeammalla.



Kuva 5: Muuttujien *JOYREAD*, *MEMOR* ja *STIMREAD* vaikutus vasteeseen.

5.4.1 Sisällöllistä pohdintaa

Suomalaisten nuorten lukutaito on heikentynyt 2000-luvun aikana (Sulkunen ym. 2010), ja siksi onkin tärkeää etsiä keinoja lukutaidon tason parantamiseksi. Erityisen huolestuttavaa on se, että heikkojen lukijoiden osuus on kasvanut, mikä tarkoittaa, että Suomessa on jo aivan liikaa nuoria, joilla on ongelmia tekstipainotteisen tietoyhteiskunnan arjen haasteista selviytymisessä ja myös kohonnut syrjäytymisriski (Sulkunen ja Nissinen 2012). Näin ollen heikon lukutaidon vaikutukset näkyvät niin yksittäisen nuoren tulevaisuudessa kuin koko yhteiskunnan kilpailukyvyssä.

On todettu, että muun muassa sukupuolen ja sosioekonomisen aseman ohella tietoisuus tekstin ymmärtämisen ja muistamisen strategioista on vahvasti yhteydessä lukutaitoon (Sulkunen ym. 2010). On kiinnostavaa tutkia, mitkä tekijät puolestaan vaikuttavat tietoisuuteen ymmärtämis- ja muistamisstrategioista, koska tähän opettajien on mahdollista vaikuttaa. Voidaan siis ajatella, että tietoisuuteen tekstin ymmärtämisen ja muistamisen strategioista vaikuttavat tekijät vaikuttavat välillisesti myös yleiseen lukutaitoon.

Tekstin ymmärtämis- ja muistamisstrategioiden tietoisuuden selittäminen yleisen lukutaidon sijaan on kiinnostavaa erityisesti siksi, että strategiat ovat hyvin konkreettisia ja eksplisiittisesti opetettavissa olevia asioita. Luetun ymmärtämisen ja tulkinnan opettaminen muulla tavalla kuin strategioita opettamalla taas on vaikeaa. Yleisen lukutaidon ja tietoisuuden tekstin ymmärtämisen ja muistamisen strategioista välinen korrelaatio on 0.44, joka on kohtalaista suuruusluokkaa. Kaikkien muuttujien parittaiset korrelaatiot on esitetty liitteessä B.

Tietoisuus tekstin ymmärtämisen ja muistamisen tehokkaista strategioista on kokonaisuus, jonka taustalla on luonnollisesti useita eri tekijöitä, jotka ovat myös yhteydessä toisiinsa. Jotkut näistä tekijöistä liittyvät suoraan siihen, millaiset edellytykset koti ja koulu tarjoavat oppimiselle. Jotkut tekijät taas liittyvät oppilaan opiskelustrategioihin ja -motivaatioon, joihin kodilla ja koululla voi tosin olla vahva vaikutus. Käytännössä muuttujien väliset kausaalisuhteet eivät välttämättä ole puhtaan yksisuuntaisia, mikä olisi tilastollisen mallintamisen kannalta ihanteellista.

Tulokset ovat pääasiassa saman suuntaisia kuin PISAn lukutaitotutkimuksissa aikaisemminkin. Sukupuolieron on todettu olevan Suomessa kansainvälisesti vertailtuna erityisen suuri tyttöjen hyväksi, ja sosioekonomisen taustan (*ESCS*) on huomattu vaikuttavan positiivisesti tietoisuuteen ymmärtämis- ja muistamisstrategioista (OECD 2010a). Muiden muuttujien vaikutusta ymmärtämis- ja muistamisstrategioiden tiedostamiseen ei tiettävästi ole aiemmin tutkittu Suomen aineistolla.

Merkitsevistä selittäjistä jotkut ovat myös lukutaidon merkitseviä selittäjiä. Muuttujat *kiinnostus lukemista kohtaan* (*JOYREAD*) ja *kontrollistrategioiden hyödyntäminen* (*CSTRAT*) vaikuttavat positiivisesti lukutaitoon (Sulkunen ym. 2010), joten on uskottavaa, että niillä on saman suuntainen vaikutus myös tietoisuuteen ymmärtämis- ja muistamisstrategioista. Kontrollistrategioi-

hin kuuluu oman oppimisen ja oppimistavoitteiden säätely ja tiedostaminen, joten niitä voidaankin pitää yleisesti hyvinä strategioina. Opetuksessa tulisi siis kiinnittää huomiota siihen, että oppilaat kykenevät hahmottamaan ja itse asettamaan oppimistavoitteita sekä osaavat myös seurata tavoitteiden saavuttamista. Strategioiden konkreettisen opettamisen tehostaminen hyödyttäisi ennen kaikkea heikoimpia lukijoita (Sulkunen ja Nissinen 2012). Kontrollistrategioiden hallinta saattaa olla osittain jopa päällekkäinen muuttuja vastemuuttujan kanssa, koska jos oppilas hallitsee opiskelun säätelyn, on hyvin johdonmukaista, että oppilas hallitsee myös tekstin ymmärtämisen ja muistamisen strategioiden säätelyn. Lukemistilanteet ovat aina eräänlaisia opiskelutilanteita ja edellyttävät lukemisen kontrollointia, jotta teksti tulee ymmärretyksi.

Tämän tutkielman mallissa merkitseväksi jäänyt muuttujan *kiinnostus lukemista kohtaan* (*JOYREAD*) kvadraattinen termi on myös tulkinnallinen. Kiinnostuksen lisääntyminen vaikuttaisi vasteeseen mallin mukaan voimakkaammin niillä, joilla kiinnostusta on vähän, kuin niillä, jotka ovat hyvin kiinnostuneita lukemisesta. Käytännössä lukumotivaatiota pitäisi ensisijaisesti pyrkiä lisäämään niillä oppilailla, joilla sitä ei ennestään juuri ole.

Mieleenpainamisstrategioiden hyödyntäminen (*MEMOR*) ei ole merkitsevä lukutaidon selittäjä (OECD 2010a), eikä sen yhteyttä tietoisuuteen ymmärtämis- ja muistamisstrategioista voida pitää lineaarisena. Kuitenkin kun sille otetaan mukaan kvadraattinen termi, saadaan hyvin mielekkäät estimaatit. Kuten kuvasta 5 nähdään, niin mallin mukaan mieleenpainamisstrategioiden käyttäminen pienissä määrin voi olla hyödyllistä, mutta sen suuremmalla käytöllä on negatiivinen vaikutus vasteeseen. Voi olla, että mieleenpainamisstrategioiden liiallinen käyttö alkaa dominoida opiskelustrategioita, eikä oppilas tällöin käytä hyödyllisempiä strategioita.

Muuttuja *epälineaarisia elementtejä sisältävien tekstien lukeminen koulua varten* (*RFSNCONT*) tarkoittaa tekstejä, jotka sisältävät kaavioita, karttoja, taulukoita tai kuvaajia. Nämä ovat pääasiassa haastavia asiatekstejä, joten on uskottavaa, että niiden lukeminen vaikuttaa positiivisesti tietoisuuteen tekstin ymmärtämisen ja muistamisen strategioista. Sen sijaan muuttuja *painetut lehti- ja ohjetekstit* (*RFSFUMAT*) kuvaa esimerkiksi aikakauslehtiartikkeleiden, käyttöoppaiden ja mainostekstien käyttöä. Tällaisten monesti lyhyiden ja helpohkojen tekstien runsas käyttö opetuksessa ei ehkä ole tarpeeksi kehittävää, mikä selittäisi kyseisen muuttujan negatiivisen vaikutuksen vasteeseen.

Sillä on siis suuri merkitys, minkä tyyppisiä tekstejä koulussa luetaan. Vaikeat tekstit ovat tärkeitä taitojen kehittymisen kannalta, mutta on huomattava, että helpohkoilla teksteillä on keskeinen rooli lukemismotivaation herättelemisessä ja onnistuneiden lukukokemusten tarjoamisessa (Sulkunen ja Nissinen 2012). Opettajat saattavat säädellä lukemistehtävien vaikeustasoa oppilaiden osaamisen perusteella, joten koulun lukemistehtäviin liittyvien muuttujien todellinen vaikutus voi myös jossain määrin olla päinvastainen kuin malli olettaa.

Lukemiseen kannustamisen ja innostamisen voisi luulla vaikuttavan positiivisesti tietoisuuteen ymmärtämis- ja muistamisstrategioista, mutta sen yhteys

vasteeseen on melko erikoinen, kuten kuvasta 5 nähdään. On vaikea löytää selkeää tulkintaa sille, miksi pojilla muuttujan *lukemiseen sitouttaminen* (*STIM-READ*) vaikutus on pienillä arvoilla positiivista ja suurilla negatiivista. On myös vaikea keksiä selitystä tyttöjen ja poikien väliselle erolle. Yksi selitys voisi olla, että poikien kohdalla runsas lukemiseen sitouttaminen olisi pikemminkin heikon lukutaidon seurausta kuin sen syy. Ehkäpä heikot pojat ovat tyttöjä alttiimpia opettajan itsepintaiselle lukemiseen sitouttamiselle. Liitteestä A nähdään, että muuttuja ei välttämättä mittaakaan juuri strategiaosaamisen lisääntymiseen liittyvää opettajan toimintaa, mikä näkyy tuloksissa sitouttamisen tehottomuutena. Vaikka opettaja pyytäisi selittämään tekstin merkitystä ja esittäisi kysymyksiä, se ei kehitä strategiaosaamista, jos ei opeteta, millä keinoilla vastaukset tekstiä koskeviin kysymyksiin löytyvät.

6 Yhteenveto

Erilaiset otanta-asetelmat johtavat erinäisiin haasteisiin regressiomallin estimoinnissa. Tutkielmassa esitettiin vaihtoehtoisia menetelmiä näistä haasteista selviytymiseen ja tutkittiin empiirisesti, kuinka yhdenmukaisia tuloksia ne antavat. Sovellusaineistona käytettiin vuoden 2009 Suomen ja Saksan PISA-aineistoja, joiden taustalla oleva otanta-asetelma on hyvinkin kompleksinen.

Vertailtavissa menetelmissä regressiokertoimien estimaatit laskettiin kiinteiden vaikutusten mallilla tai sekamallilla käyttäen analyysissä otantapainoja tai korvaamalla ne asetelmamuuttujilla. Keskivirheet laskettiin malliperusteisesti lineaarisella sekamallilla tai asetelmaperusteisesti Fay-modifioidulla BRR-menetelmällä tai Taylor-linearisoinnilla.

Vertailuissa tarkasteltiin menetelmien antamia estimaatteja regressiokertoimille ja niiden keskivirheille sekä näiden avulla saatuja p-arvoja. Näissä ei ollut huomattavia menetelmien välisiä eroja, paitsi kiinteiden vaikutusten mallin ja sekamallin regressiokertoimissa Saksan aineistolla. Mallit vastaavat kuitenkin hieman eri kysymyksiin, joten ero kertookin Saksan aineiston rakenteesta. Saksassa koulukohtaisen mallin regressio on loivempi kuin koko valtiota koskevalla mallilla. Suomessa tällaista eroa ei ole. Tulokset siis osoittivat, että aineistoa koskeva ymmärrys voi lisääntyä, jos analyyseissä käytetään erilaisia menetelmiä. Vaikka aineisto on hierarkkinen, ei ole järkevää käyttää automaattisesti sekamallia, koska tutkimusongelman kannalta saattaa olla kiinnostavampaa käyttää kiinteiden vaikutusten mallia. Tällöin hierarkkisuus otetaan huomioon keskivirheiden laskemisessa esimerkiksi BRR-menetelmällä tai Taylor-linearisoinnilla.

Muiden menetelmien kanssa yhdenmukaisia tuloksia saatiin jopa menetelmällä, joka ei ota otanta-asetelmaa millään tavalla huomioon. Yhtenä selityksenä tälle nähtiin se, että selittävät muuttujat selittävät huomattavan osan koulujen välisestä vaihtelusta. Sopivilla kovariaateilla voidaan siis ratkaisevasti selittää asetelman kompleksisuutta. Jos vertailevat analyysit osoittavat, että kovariaateilla saadaan eliminoitua aineiston kompleksisuutta, voidaan jatkoanalyyseissä käyttää perustellusti yksinkertaisempia menetelmiä. Huomattiin, että vaikka menetelmien väliset erot ovat hyvin pieniä, niin niillä voi kuitenkin olla hieman vaikutusta mallinvalinnan lopputulokseen.

Estimaattien yhdenmukaisuus on lohdullinen tulos, koska tässä työssä vertailtuja menetelmiä käytetään samaan tarkoitukseen. Tulosten yleistämisessä täytyy kuitenkin olla varovainen, koska PISA-aineistot ovat vain yksi esimerkki otanta-aineistosta. Jos aineiston klusteroituminen olisi vielä voimakkaampaa kuin tässä työssä käytetyssä Saksan aineistossa tai jos klusteroituminen ei selittyisikään selittävillä muuttujilla, saattaisivat tulokset olla erilaisia. On myös mahdollista, että selvästi pienemmillä aineistoilla saataisiin toisenlaisia tuloksia. Saatujen tulosten voidaan kuitenkin katsoa olevan yleistettävissä PISA-aineistojen lisäksi myös joihinkin muihin kansainvälisiin arviointitutkimuksiin, joissa käytetään pitkälti samanlaista metodologiaa.

Suomen aineistolla selvitettiin lisäksi, miten eri tekijät selittävät oppilaan tietoisuutta tekstin ymmärtämisen ja muistamisen strategioista. Tulokset vahvistivat osaltaan käsitystä kontrollistrategioiden hyödyllisyydestä ja oppilaan oman lukumotivaation merkityksestä. Sen sijaan mieleenpainamisstrategian voimakkaalla hyödyntämisellä nähtiin olevan negatiivinen vaikutus selitettävään muuttujaan, mutta myös tätä strategiaa on tulosten mukaan parempi käyttää pienissä määrin kuin ei ollenkaan. Huomattiin myös, että haastavien asiatekstien käyttö koulussa parantaa oppilaiden tietoisuutta tekstin ymmärtämisen ja muistamisen strategioista, kun taas helpohkojen tekstien kuten mainostekstien tai aikakauslehtiartikkelien käyttö heikentää sitä.

Lähteet

Gelman, A. & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge.

Goldstein, H. (2011). *Multilevel Statistical Models*. 4th ed. Wiley, Chichester.

Heeringa, S. G., West, B. T. & Berglund, P. A. (2010). *Applied Survey Data Analysis*. Chapman & Hall, Lontoo.

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove.

McCulloch, C. E. & Searle, S. R. (2001). *Generalized, Linear, and Mixed Models*. Wiley, New York.

OECD (2009). *PISA Data Analysis Manual: SAS Second Edition*. OECD, Pariisi.

OECD (2010a). *PISA 2009 Results: Learning to Learn – Student Engagement, Strategies and Practices (Volume III)*. OECD, Pariisi.

OECD (2010b). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*. OECD, Pariisi.

OECD (2012). *PISA 2009 Technical Report*. OECD, Pariisi.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H. & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society B*, **60**, 23–40.

Rabe-Hesketh, S. & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society A*, **169**, 805–827.

Rao, J. N. K. (2003). *Small Area Estimation*. Wiley, New Jersey.

Rust, K. & Krawchuk, S. (2002). Survey Weighting and the Calculation of Sampling Variance. Teoksessa OECD. *PISA 2000 Technical Report*. OECD, Pariisi.

Sulkunen, S. & Nissinen, K. (2012). Heikot lukijat Suomessa. Teoksessa Sulkunen, S. & Välijärvi, J. (toim.). *PISA09. Kestääkö osaamisen pohja?*. Opetus- ja kulttuuriministeriön julkaisuja 2012:12. Opetus- ja kulttuuriministeriö, Helsinki.

Sulkunen, S., Välijärvi, J., Arffman, I., Harju-Luukkainen, H., Kupari, P., Nissinen, K., Puhakka, E. & Reinikainen, P. (2010). *PISA 2009 Ensituloksia*. Opetus- ja kulttuuriministeriön julkaisuja 2010:21. Opetus- ja kulttuuriministeriö, Helsinki.

Särndal, C.-E., Swensson, B. & Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag, New York.

Valliant, R., Dorfman, A. H. & Royall, R. M. (2000). *Finite Population Sampling and Inference: A Prediction Approach*. Wiley, New York.

Wolter, K. M. (2007). *Introduction to Variance Estimation*. 2nd ed. Springer, New York.

Liite A: Muuttujien taustalla olevat kysymykset

Seuraavassa on esitetty muuttujittain koottuna kysymykset, joiden avulla tässä työssä käytetyt muuttujat on muodostettu. Opettajien vaikutus koulun ilmapiiriin (*TEACBEHA*) on koulukyselystä, muut oppilaskyselystä.

Tietoisuus tekstin ymmärtämisen ja muistamisen strategioista (UNDREM)

Lukemistehtävä: Sinun on ymmärrettävä ja muistettava tekstin sisältämät tiedot.
Kuinka hyödyllisiä seuraavat strategiat mielestäsi ovat tekstin ymmärtämisen ja muistamisen kannalta?
(Kuusi vaihtoehtoa: Ei lainkaan hyödyllinen (1),..., Erittäin hyödyllinen (6))

a) Keskityn niihin tekstin osiin, jotka on helppo ymmärtää.

b) Luen tekstin nopeasti kaksi kertaa läpi.

c) Kun olen lukenut tekstin, keskustelen sen sisällöstä muiden kanssa.

d) Alleviivaan tärkeitä kohtia tekstistä.

e) Teen tekstin sisällöstä omin sanoin lyhyen tiivistelmän.

f) Luen tekstin ääneen jollekulle muulle.

Kontrollistrategioiden hyödyntäminen (CSTRAT)

Kun opiskelet, kuinka usein teet seuraavaa?
(Neljä vaihtoehtoa: Tuskin koskaan, Toisinaan, Usein, Melkein aina)

a) Aloitan opiskelun selvittämällä ensin itselleni, mitä minun tarkalleen ottaen pitää oppia.

b) Varmistan, että ymmärrän lukemani asiat.

c) Yritän selvittää ne käsitteet, joita en ole vielä kunnolla ymmärtänyt.

d) Varmistan, että muistan tekstin tärkeimmät kohdat.

e) Jos en opiskellessani ymmärrä jotakin asiaa, etsin lisätietoja selvittääkseni sen.

Mieleenpainamisstrategioiden hyödyntäminen (MEMOR)

Kun opiskelet, kuinka usein teet seuraavaa?
(Neljä vaihtoehtoa: Tuskin koskaan, Toisinaan, Usein, Melkein aina)

a) Yritän painaa mieleeni kaiken, mitä tekstissä sanotaan.

b) Yritän opetella ulkoa mahdollisimman paljon.

c) Luen tekstin niin monta kertaa, että osaan toistaa sen ulkoa.

d) Luen tekstin yhä uudelleen ja uudelleen.

Elaborointistrategioiden hyödyntäminen (ELAB)

Kun opiskelet, kuinka usein teet seuraavaa?
(Neljä vaihtoehtoa: Tuskin koskaan, Toisinaan, Usein, Melkein aina)

a) Yritän liittää uudet tiedot muissa aineissa oppimiini asioihin.

b) Pohdin, mitä hyötyä kyseisestä tiedosta voisi olla koulun ulkopuolella.

c) Yritän ymmärtää aineiston paremmin liittämällä sen asiat omiin kokemuksiini.

d) Koetan selvittää, kuinka tekstin sisältö sopii yhteen todellisen elämän kanssa.

Kaunokirjallisuuden tulkinta koulua varten (RFSINTRP)

Kuinka usein sinun on täytynyt lukea seuraavanlaisia tekstejä tai tehdä seuraavanlaisia tehtäviä koulua varten (oppitunneilla tai kotitehtävänä) viimeksi kuluneen kuukauden aikana?
(Neljä vaihtoehtoa: Useita kertoja, Kaksi tai kolme kertaa, Kerran, Ei kertaakaan)

a) Kaunokirjallisuutta (esim. romaanit, novellit)

b) Selittää syy tekstin tapahtumiin

c) Selittää tekstin henkilöiden käyttäytymistä

d) Kertoa tekstin tarkoitus

Epälineaarisia elementtejä sisältävien tekstien lukeminen koulua varten (RFSNCONT)

Kuinka usein sinun on täytynyt lukea seuraavanlaisia tekstejä tai tehdä seuraavanlaisia tehtäviä koulua varten (oppitunneilla tai kotitehtävänä) viimeksi kuluneen kuukauden aikana?
(Neljä vaihtoehtoa: Useita kertoja, Kaksi tai kolme kertaa, Kerran, Ei kertaakaan)

a) Tekstejä, joihin sisältyy kaavioita tai karttoja

b) Tekstejä, joihin sisältyy taulukkoja tai graafisia kuvaajia

c) Etsiä tietoja graafisesta kuvaajasta, kaaviosta tai taulukosta

d) Kuvailla, miten taulukon tai graafisen kuvaajan tiedot on jäsennetty

Kaunokirjallisuuden perinteiset oppisisällöt koulua varten (RFSTRLIT)

Kuinka usein sinun on täytynyt lukea seuraavanlaisia tekstejä tai tehdä seuraavanlaisia tehtäviä koulua varten (oppitunneilla tai kotitehtävänä) viimeksi kuluneen kuukauden aikana?
(Neljä vaihtoehtoa: Useita kertoja, Kaksi tai kolme kertaa, Kerran, Ei kertaakaan)

a) Tietotekstejä kirjailijoista tai kirjoista

b) Runoutta

c) Perehtyä kirjoittajan elämään

d) Opetella teksti (esim. runo tai osa näytelmästä) ulkoa

e) Perehtyä siihen, millainen paikka tekstillä on kirjallisuuden historiassa

Painettujen lehti- ja ohjetekstien lukeminen koulua varten (RFSFUMAT)

Kuinka usein sinun on täytynyt lukea seuraavanlaisia tekstejä koulua varten (oppitunneilla tai kotitehtävänä) viimeksi kuluneen kuukauden aikana?
(Neljä vaihtoehtoa: Useita kertoja, Kaksi tai kolme kertaa, Kerran, Ei kertaakaan)

a) Sanomalehtijuttuja ja aikakauslehtiartikkeleita

b) Ohjeita tai käyttöoppaita, joissa kerrotaan, miten jokin asia tehdään (esim. miten joku laite toimii)

c) Mainostekstejä (esim. lehti-ilmoitukset, mainosjulisteet)

Opettajien vaikutus koulun ilmapiiriin (TEACBEHA)

Missä määrin seuraavat tekijät haittaavat oppilaiden oppimista koulussanne? (Neljä vaihtoehtoa: Ei lainkaan, Hyvin vähän, Jossain määrin, Paljon)

a) Opettajien vähäiset oppilaisiin kohdistuvat odotukset

b) Huonot suhteet oppilaiden ja opettajien välillä

c) Se, että opettajat eivät ota huomioon yksittäisten oppilaiden tarpeita

d) Opettajien poissaolot

e) Henkilöstön muutosvastarinta

f) Se, että opettajat ovat liian ankaria oppilaille

g) Se, että oppilaita ei rohkaista yrittämään parastaan

Osiot on koodattu käänteisesti, eli muuttujan suuret arvot tarkoittavat, että opettajat vaikuttavat positiivisesti koulun ilmapiiriin.

Sosioekonominen asema (ESCS)

Sosioekonominen asema koostuu muuttujista kodin varallisuus, kodin kulttuuriesineet, kodin koulutusresurssit, kirjojen lukumäärä kotona (kolmiluokkaisena: korkeintaan 25, 26-100, yli 100 kirjaa), vanhempien korkein ammatillinen asema ja vanhempien korkein koulutus vuosissa.

Kiinnostus lukemista kohtaan (JOYREAD)

Missä määrin olet samaa tai eri mieltä seuraavista väittämistä, jotka koskevat lukemista?
(Neljä vaihtoehtoa: Täysin eri mieltä, Eri mieltä, Samaa mieltä, Täysin samaa mieltä)

a) Luen vain jos on pakko.

b) Lukeminen on yksi mieliharrastuksistani.

c) Keskustelen mielelläni kirjoista toisten kanssa.

d) Minun on vaikea lukea kirjoja loppuun.

e) Olen iloinen, jos saan kirjan lahjaksi.

f) Minusta lukeminen on ajanhaaskausta.

g) Käyn mielelläni kirjakaupassa tai kirjastossa.

h) Luen ainoastaan saadakseni tietoja, joita tarvitsen.

i) En pysty keskittymään lukemiseen kauempaa kuin muutaman minuutin.

j) Kerron mielelläni mielipiteitäni kirjoista, joita olen lukenut.

k) Vaihtelen mielelläni kirjoja ystävieni kanssa.

Osiot a), d), f), h) ja i) on koodattu käänteisesti. Muuttujan suuret arvot tarkoittavat siis suurta kiinnostusta lukemista kohtaan.

Lukemisen monipuolisuus (DIVREAD)

Kuinka usein luet seuraavanlaisia tekstejä omasta halustasi?
(Viisi vaihtoehtoa: En koskaan tai tuskin koskaan, Muutaman kerran vuodessa, Noin kerran kuussa, Useita kertoja kuussa, Useita kertoja viikossa)

a) Aikakauslehtiä

b) Sarjakuvalehtiä

c) Kaunokirjallisuutta (romaanit, tarinat, kertomukset)

d) Tietokirjallisuutta

e) Sanomalehtiä

Lukemiseen sitouttaminen (STIMREAD)

Kuinka usein seuraavanlaisia asioita tapahtuu äidinkielen tunneillasi?
(Neljä vaihtoehtoa: Ei koskaan tai tuskin koskaan, Joillakin tunneilla, Useimmilla tunneilla, Kaikilla tunneilla)

a) Opettaja pyytää oppilaita selittämään tekstin merkitystä.

b) Opettaja esittää vaikeita kysymyksiä saadakseen oppilaat miettimään ja ymmärtämään tekstin paremmin.

c) Opettaja antaa oppilaille tarpeeksi aikaa vastausten miettimiseen.

d) Opettaja suosittelee oppilaille jotakin kirjaa tai kirjailijaa.

e) Opettaja kannustaa oppilaita ilmaisemaan mielipiteensä tekstistä.

f) Opettaja auttaa oppilaita näkemään, miten heidän lukemansa tekstit liittyvät heidän omaan elämäänsä.

g) Opettaja osoittaa, miten tekstien sisältämät asiat rakentuvat sille, mitä oppilaat tietävät entuudestaan.

Kirjaston käyttö (LIBUSE)

Kuinka usein käyt kirjastossa seuraavista syistä?
(Viisi vaihtoehtoa: En koskaan, Muutaman kerran vuodessa, Noin kerran kuussa, Useita kertoja kuussa, Useita kertoja viikossa)

a) Lainaat kirjoja, joita luet omaksi iloksesi

b) Lainaat kirjoja koulutehtäviä varten

c) Opiskelet tai teet koulutehtäviä

d) Luet aikakaus- tai sanomalehtiä

e) Luet kirjoja omaksi iloksesi

f) Hankit tietoja asioista, jotka eivät liity kouluun (esim. urheilusta, harrastuksista, ihmisistä tai musiikista)

g) Käytät Internetiä

Liite B: Korrelaatiot Suomen aineistossa

Taulukko 13: Muuttujien väliset korrelaatiot Suomen aineistossa (osa 1/2).

	Lukutaito	UNDREM	CSTRAT	MEMOR	ELAB	RFSINTRP	RFSNCONT	RFSTRLIT
Lukutaito	1	0.44	0.33	0.05	0.17	0.06	0.18	-0.02
UNDREM	0.44	1	0.29	0.11	0.14	0.07	0.10	0.04
CSTRAT	0.33	0.29	1	0.54	0.55	0.21	0.24	0.15
MEMOR	0.05	0.11	0.54	1	0.35	0.15	0.12	0.14
ELAB	0.17	0.14	0.55	0.35	1	0.19	0.21	0.14
RFSINTRP	0.06	0.07	0.21	0.15	0.19	1	0.40	0.60
RFSNCONT	0.18	0.10	0.24	0.12	0.21	0.40	1	0.32
RFSTRLIT	-0.02	0.04	0.15	0.14	0.14	0.60	0.32	1
RFSFUMAT	-0.18	-0.05	0.12	0.13	0.16	0.45	0.33	0.38
TEACBEHA	0.03	0.02	0.04	0.00	0.01	0.06	0.02	0.05
ESCS	0.28	0.12	0.20	0.09	0.15	0.10	0.14	0.07
JOYREAD	0.56	0.36	0.41	0.22	0.27	0.11	0.14	0.09
DIVREAD	0.38	0.21	0.34	0.19	0.28	0.16	0.18	0.12
STIMREAD	0.06	0.06	0.24	0.16	0.27	0.23	0.14	0.20
LIBUSE	0.13	0.16	0.26	0.21	0.22	0.13	0.08	0.15

Taulukko 14: Muuttujien väliset korrelaatiot Suomen aineistossa (osa 2/2).

	RFSFUMAT	TEACBEHA	ESCS	JOYREAD	DIVREAD	STIMREAD	LIBUSE
Lukutaito	-0.18	0.03	0.28	0.56	0.38	0.06	0.13
UNDREM	-0.05	0.02	0.12	0.36	0.21	0.06	0.16
CSTRAT	0.12	0.04	0.20	0.41	0.34	0.24	0.26
MEMOR	0.13	0.00	0.09	0.22	0.19	0.16	0.21
ELAB	0.16	0.01	0.15	0.27	0.28	0.27	0.22
RFSINTRP	0.45	0.06	0.10	0.11	0.16	0.23	0.13
RFSNCONT	0.33	0.02	0.14	0.14	0.18	0.14	0.08
RFSTRLIT	0.38	0.05	0.07	0.09	0.12	0.20	0.15
RFSFUMAT	1	0.04	0.03	-0.05	0.08	0.15	0.10
TEACBEHA	0.04	1	0.02	0.03	0.01	0.02	0.01
ESCS	0.03	0.02	1	0.17	0.19	0.07	0.02
JOYREAD	-0.05	0.03	0.17	1	0.51	0.15	0.43
DIVREAD	0.08	0.01	0.19	0.51	1	0.17	0.33
STIMREAD	0.15	0.02	0.07	0.15	0.17	1	0.12
LIBUSE	0.10	0.01	0.02	0.43	0.33	0.12	1