Lineaarinen sekamalli rekisteripohjaisen lasten ja nuorten neuvola- ja kouluterveysaineiston analyysivälineenä

Petteri Mäntymaa

Helsingin yliopisto
Valtiotieteellinen tiedekunta
Tilastotiede
Pro Gradu -tutkielma
15. toukokuuta 2020

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Valtiotieteellinen tiedekunta	

Tekijä — Författare — Author
Petteri Mäntymaa

Työn nimi — Arbetets titel — Title
Lineaarinen sekamalli rekisteripohjaisen lasten ja nuorten neuvola- ja kouluterveysaineiston analyysivälineenä

Oppiaine — Läroämne — Subject
Tilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro Gradu -tutkielma	15. toukokuuta 2020	64 sivua + 3 liitesivua

Tiivistelmä — Referat — Abstract

Terveyden ja hyvinvoinnin laitoksen FinLapset-rekisteri tutkii lasten ja nuorten ylipainon ja lihavuuden yleisyyttä Suomessa. Tiedot perustuvat valtakunnalliseen rekisteriaineistoon neuvola- ja kouluterveydenhuollon paino- ja pituusmittauksista. Tuloksia on raportoitu poikkileikkausasetelmassa raportointivuosittain, mutta aineisto mahdollistaa myös samoilta yksilöiltä kertyneiden toistettujen mittausten analyysin pitkittäistutkimusasetelmassa.

Tutkielmassa arvioimme lineaaristen sekamallien soveltuvuutta FinLapset-rekisteriaineistosta muodostetun pitkittäisaineiston analyysivälineeksi. Teoriaosassa esittelemme lineaaristen sekamallien keskeiset ominaisuudet ja estimointimenetelmät sekä tarkastelemme hyviä mallinarvioinnin käytäntöjä. Soveltavassa vaiheessa sovitamme aineistoon kaksitasoisen lineaarisen sekamallin, jolla tutkimme lasten ja nuorten painoindeksin yhteyttä ikään ja biologiseen sukupuoleen sekä arvioimme mallin kykyä selittää aineistossa esiintyvää yksilökohtaista ja yksilöiden välistä painoindeksin vaihtelua. Mallin suoriutumista tarkastellaan erityisesti rekisteriaineiston analyysin muodostamien haasteiden näkökulmasta.

Lineaariset sekamallit muodostavat luontevan analyysikehikon FinLapset-rekisteriaineiston kaltaisen pitkittäisaineiston analyysiin. Yksinään iän kiinteä populaatiovaikutus, yhdessä yksilö- ja ikäkohtaisten satunnaisvaikutusten kanssa selittää mallin vaihtelua erittäin hyvin. Painoindeksin ja iän yhteyden lineaarisuusoletus jää kuitenkin epäilyksen alaiseksi ja yksilökohtaisten residuaalien autokorrelaatio sekä varianssin heteroskedastisuus osoittautuvat merkittäviksi haasteiksi.

Rekisteriaineistolle tyypilliset ominaisuudet, kuten passiivisesta kertymistavasta seuraava populaatiokehikon täsmällisen määrittelyn puute ja aineistoa tuottavien prosessien tuntemattomuus vaikeuttavat mallin estimaatteihin liittyvien epävarmuustekijöiden arviointia. Suuresta havaintomäärästä seuraten estimaattien keskivirheet ovat hyvin pieniä, mikä antaa virheellisen kuvan mallin hyvyydestä, vaikka estimaatteihin liittyvä harha jää osin tunnistamatta.

Tutkielmassa näytetään, että lineaarisille sekamallelle löytyy joustavia laajennoksia, joilla osa tutkielmassa esitetyn mallin haasteista on mahdollista ylittää. Osa laajennetuista malleista ovat suoraan yhteensopivia tutkielman frekventistisen lähestymistavan kanssa, mutta useat vaihtoehtoiset menetelmät suosivat bayesiläistä ajattelutapaa. Myös näkökulmia rekisteriaineiston epävarmuuslähteiden tunnistamiseksi ja edustavuuden parantamiseksi punnitaan.

Avainsanat — Nyckelord — Keywords
Lineaarinen sekamalli, pitkittäisaineisto, rekisteriaineisto, FinLapset, Terveyden ja hyvinvoinnin laitos

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Sisältö

1 Johdanto

Terveyden ja hyvinvoinnin laitoksen FinLapset-rekisteri (THL, 2019) kerää ajankohtaista tietoa lasten ja nuorten terveydestä ja hyvinvoinnista. Eräänä tietolähteenä ovat lasten ja nuorten neuvola- ja kouluterveydenhuollon käyntien yhteydessä kirjatut paino- ja pituustiedot, joista muodostetaan ajantasaista ja kattavaa tietoa lasten ja nuorten ylipainosta.

Tiedot raportoidaan vuositasolla valiten kultakin yksilöltä yksi edustavaksi katsottu mittaus raportointivuotta kohden. Aineisto muodostaa vuosittaisen poikkileikkauksen lasten- ja nuorten ylipainosta ja lihavuudesta. Samoilta yksilöiltä voidaan kuitenkin saada useita mittauksia useilta vuosilta. Raportoinnissa menetetään siis merkittävä pitkittäisulottuvuus, joka voisi tarjota hyödyllistä lisätietoa ylipainon kehityksestä.

Toistettuja mittauksia sisältävän aineiston analyysiin on kehitetty useita tilastollisia menetelmiä. Osa menetelmistä asettaa aineistolle tiukkoja vaatimuksia mm. kontrolloidusta koeasetelmasta, mutta toiset ovat joustavampia aineiston ominaisuuksien suhteen. Erään monipuolisen menetelmäperheen muodostavat lineaariset sekamallit, jotka tarjoavat lupaavan analyysivälineen toistettuja mittauksia sisältävälle aineistolle.

Rekisteriaineistoja ei tyypillisesti ole kerätty tutkimus- tai analyysinäkökulmasta ja aineistot voivat olla myös huomattavasti otosaineistoja suurempia. Tästä syystä katsomme menetelmien joustavuuden suureksi eduksi.

Tässä tutkielmassa tutustumme lineaarisiin sekamalleihin ja arvioimme niiden soveltuvuutta FinLapset-rekisteriaineistosta muodostetun pitkittäisaineiston analyysivälineeksi. Luvussa 2 tarkastelemme pitkittäistutkimusta tutkimusasetelmana ja esittelemme tutkielman aineiston. Luvussa 3 muodostamme teoreettisen pohjan lineaaristen sekamallien yleiselle muodolle, mallin ominaisuuksille ja estimointimenetelmille. Luvun 4 keskiössä on tutkielman aineiston analyysiin sopivan lineaarisen sekamallin muodostaminen sekä lopullisen mallin arviointi. Lopuksi, luvussa 5 tuomme yhteen tutkielman keskeiset havainnot ja johtopäätökset.

2 Pitkittäistutkimus tutkimusasetelmana

Pitkittäistutkimuksen määrittelevänä ominaispiirteenä pidetään yleisesti yhden tai useamman yksilön seurantaan perustuvaa asetelmaa, jossa samasta yksilöstä saadaan havaintoja eri aikoina (Diggle *ym.*, 2013; Fitzmaurice *ym.*, 2011; Twisk, 2013; Laird ja Ware, 1982). Toistuvat havainnot samasta yksilöstä mahdollistavat ajassa

tapahtuvien muutosten, muutokseen vaikuttaneiden tekijöiden, kuin myös yksilön yksilöllisistä ominaisuuksista koostuvien tekijöiden tunnistamisen ja analysoinnin.

Diggle *ym.* (2013) mukaan taloustieteessä ja yhteiskuntatieteissä voidaan pitkittäistutkimuksesta käyttää myös termiä paneelitutkimus, mutta sovellettujen tilastollisten menetelmien kannalta, erityisesti terveystieteellisissä ja epidemiologisissa tutkimuskysymyksissä, käsitteet *pitkittäistutkimus* ja sen yhteydessä havaittu *pitkittäisaineisto* muodostavat luontevamman viitekehyksen.

Twisk (2013) jakaa epidemiologiset tutkimusasetelmat karkeasti kahteen tyyppiin, *havainnoiviin* ja *kokeellisiin* tutkimusasetelmiin (Kuva 1). Havainnoivat tutkimusasetelmat puolestaan jakautuvat *tapaus-verrokkitutkimukseen* ja *kohorttitutkimukseen*. Lisäksi havainnoivat kohorttitutkimukset voivat olla alatyypiltään eteneviä (*prospektiivinen tutkimus*), takautuvia (*retrospektiivinen tutkimus*) tai poikkileikkaustutkimuksia. (Diggle *ym.*, 2013; Twisk, 2013).

Twisk (2013) ottaa voimakkaamman kannan salliessaan pitkittäistutkimuksen käsitteen käytön edellämainituista vain prospektiivisten kohorttitutkimusten tapauksessa, kun Diggle *ym.* (2013) tyytyvät suosittelemaan prospektiivista aineistonkeruutapaa, lähinnä retrospektiivisen aineiston laatuun kohdistuvan kritiikin kannalta. Mainittakoon, että poikkileikkaustutkimukset eivät kuulu lainkaan pitkittäistutkimusten piiriin, sillä samaa yksilköä tarkastellaan siinä vain yhdessä aikapisteessä.

Tämän tutkielman asetelma ei ole ristiriidassa edellämainittujen näkökulmien kanssa, joten tutkielman piirissä omaksumme pitkittäistutkimuksen ja pitkittäisaineistojen käsitteiden käytön.

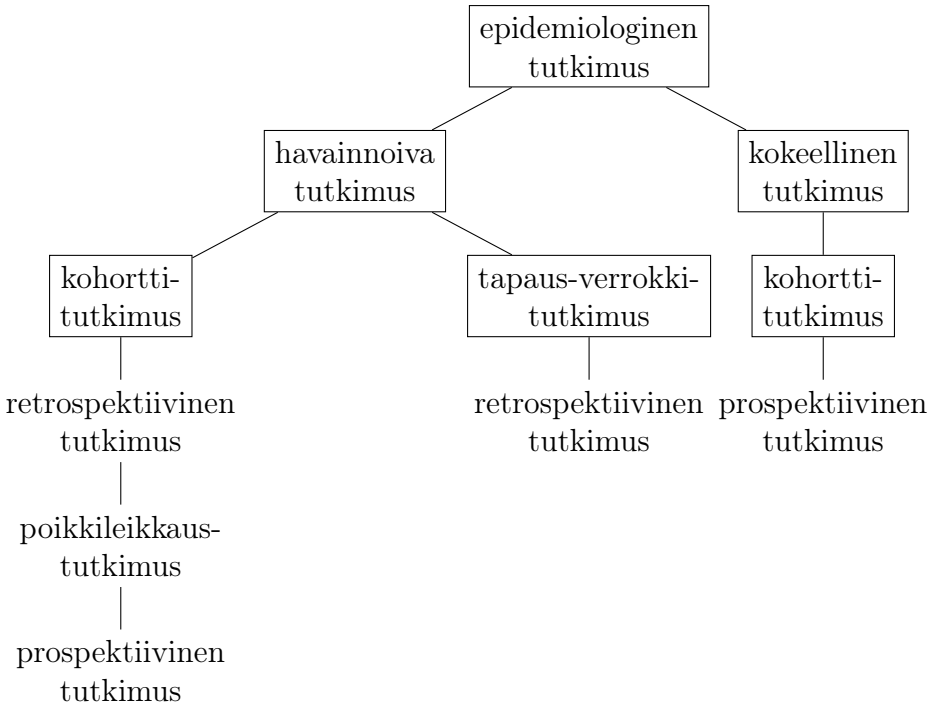
2.1 Pitkittäisaineiston analyysi

Kuten sanottu, pitkittäisaineistoille on ominaista yhden tai useamman yksilön seuranta usean havaintokerran ajan. Käytännössä useinkaan kaikilta yksilöiltä ei kyetä saamaan samaa määrää havaintoja ja havaintoajankohdat voivat vaihdella. Tällöin puhumme epätasapainoisesta (*unbalanced*) pitkittäisaineistosta .Laird ja Ware (1982).

Mm. Verbeke ja Molenberghs (2000) sekä Goldstein (2011) mukaan epätasapainoisen pitkittäisaineiston analyysissa ei suoraan voida hyödyntää yleistä monimuuttujamenetelmien kehikkoa, vaan malli tulee jakaa kahdelle tai useammalle tasolle.

Samoilta yksilöiltä kerättyjen toistettujen mittausten muodostamassa pitkittäisaineistossa yksilökohtaiset mittausvektorit muodostavat hierarkian ensimmäisen tason ja itse yksilöt toisen tason. (Goldstein, 2011).

Lineaaristen sekamallien kirjallisuudessa kaksitasoisen lähestymistavan perusteoksena viitataan Laird ja Ware (1982) artikkeliin *Random-Effects Models for Longitudinal Data*. Tosin Laird ja Ware (1982) esittävät tasot yhtenä lineaarikombinaationa ja vasta myöhemmässä kirjallisuudessa mm. Verbeke ja Molenberghs (2000) ja Talbott (2006) on katsottu luontevaksi esitellä ensin kaksitasoinen malli ja johtaa siitä



Kuva 1: Epidemiologiset tutkimusasetelmat (Twisk, 2013).

Laird ja Ware (1982) yleinen lineaarisen sekamallin muoto.

Palaamme kaksitasoisen mallin ja yleisen lineaarisen sekamallin määritelmään myöhemmin (Luku 3).

2.2 FinLapset-aineisto

Tutkielmassa hyödyntämämme aineisto perustuu Terveyden ja hyvinvoinninlaitoksen FinLapset-rekisterihankkeen (THL, 2019) ohessa kerättyihin tietoihin lastenneuvoloiden, kouluterveydenhuollon ja opiskelijaterveydenhuollon terveydenhoitokäynneillä suoritetuista paino- ja pituusmittauksista. Tiedot on kerätty osana perusterveydenhuollon avohoidon hoitoilmoitusrekisterin (Avohilmo) tietojen keräystä. Terveystarkastuksissa kerättävistä tiedoista pituus- ja painotiedot ovat olleet osa Avohilmon tietosisältöä vuodesta 2010 lähtien, mutta tietopuutteiden vuoksi tutkielman aineisto on rajattu vuosiin 2013–2020.

FinLapset-rekisterihankkeen eräitä keskeisiä tutkimuskysymyksiä ovat lasten ja nuorten ylipaino ja lihavuus sekä pituus- ja painotietojen valtakunnallinen kattavuus (THL, 2019). Tutkielman aineistoa hyödynnämme Terveyden ja hyvinvoinnin laitoksen luvalla. Tarkastelemme seuraavaksi aineiston keskeisiä rajauksia ja teknisiä määrityksiä.

2.2.1 FinLapset-aineiston muodostus

FinLapset-rekisterihankkeen yhteydessä aineistoon on tehty seuraavat rajaukset:

- Käynti on tehty vastaanotolla
- Käynti on luonteeltaan terveydenhoitokäynti
- Käynti on luokiteltu lastenneuvola-, kouluterveydenhuolto-, opiskelijaterveydenhuoltokäynniksi
- Lapsen tai nuoren ikä mittaushetkellä on välillä [1,75;20)

Rajauksilla on pyritty poistamaan mahdollisia harhan lähteitä. Esimerkiksi rajaamalla tarkastelu vain terveydenhoitokäynteihin voidaan sulkea pois sellaisia käyntejä, jotka liittyvät jonkin sairauden hoitoon tai seurantaan. Näillä lapsilla käyntejä voi olla huomattavasti enemmän kuin vertaisillaan ja joihinkin sairauksiin voi myös liittyä epätavallisia painon muutoksia.

Paino- ja pituustietojen kirjaaminen tapahtuu sähköisesti potilastietojärjestelmiin, joista ne siirretään osaksi Avohilmoa. Kirjaamiskäytännöt vaihtelevat potilastietojärjestelmittäin ja pituus- ja painomittaukset voivat olla kirjattu eri mittayksiköissä sekä niissä voi olla inhimillisestä kirjausvirheitä.

Edellämainittuja on pyritty karsimaan seuraavilla yksinkertaisilla skaalaussäännöillä:

- Jos paino yli 1000 $\longrightarrow$ yksikkönä g? $\longrightarrow$ jaetaan 1000:lla kunnes alle 1000
- Jos pituus on yli 300 $\longrightarrow$ yksikkönä mm? $\longrightarrow$ jaetaan 10:llä kunnes alle 300
- Jos pituus pienempää kuin 2,3 $\longrightarrow$ yksikkönä m? $\longrightarrow$ kerrotaan 100

Aineistosta rajattiin pois myös biologisesti mahdottomiksi tulkitut mittaukset. Tämä suoritettiin laskemalla painosta (w) ja pituudesta (h) ensin niiden keskinäistä suhdetta kuvaava painoindeksi (*Body Mass Index, BMI*) tavanomaisella kansainvälisesti vakiintuneella menetelmällä

$$\mathbf{BMI} = \frac{w_{kg}}{h_m^2}.$$

Lapsen ruumiinrakenne, ja siten painon ja pituuden suhde, vaihtelee huomattavasti iän mukaan, joten lapsen iän mukaisen painoindeksin jakauman ääriarvoja voidaan arvioida Colen LMS-menetelmällä (Cole, 1990). LMS-menetelmä, joskus myös *BMI z-score function*, tarjoaa potenssimuunnoksella normalisoidun ja standardisoidun keskihajontapistemäärän, perustuen lapsen kuukausi-iän ja sukupuolen perusteella muodostettuihin parametreihin. Pistemäärä määritellään

$$\mathrm{BMI}_z = \frac{\left(\left(\frac{\mathrm{BMI}}{\mu}\right)^{\lambda} - 1\right)}{\lambda\ \sigma_{\mathrm{CV}}},$$

jossa λ on jakaumaa normalisoivan potenssimuunnoksen (*Box Cox*) parametri, μ BMI jakauman odotusarvo ja σ_{CV} jakauman variaatiokerroin. Taulukoidut parametrit ovat saatavilla esimerkiksi Maailman terveysjärjestö WHO:lta ja Yhdysvaltain tautikeskus CDC:ltä.

Esimerkiksi 11-vuotiaan pojan taulukoiduilla parametreilla, $\lambda = -1,7862$, $\mu = 16,9392$ ja $\sigma_{\mathrm{CV}} = 0,11070$, painoindeksiä 34 $\frac{kg}{m^2}$ vastaava BMI_z olisi

$$\frac{\left(\left(\frac{34}{16,9392}\right)^{-1,7862} - 1\right)}{16,9392 \cdot 0,11070} \approx 3,6,$$

kun vaikkapa kirjauksessa tapahtuneen näppäilyvirheen takia kirjattu painoindeksi 43 tuottaisi pistemääräksi noin 4,1.

FinLapset -hankkeessa kriittiseksi rajaksi valittiin $|\mathrm{BMI}_z| = 4$ ja sitä poikkeavammat mittaukset rajattiin poiminnan ulkopuolelle.

LMS-menetelmää on kritisoitu, kuten yllä havaittiin, mm. siitä, että se kuvaa leveän painoindeksiarvojoukon hyvin kapealle välille. Lisäksi vanhemmilla lapsilla äärimmäisenkin korkeat painoindeksiarvot kuvautuvat hyvin lähelle hyväksyttäviksi katsottuja painoindeksin arvoja. (Flegal ja Cole, 2013; CDC, 2013).

Flegal ja Cole (2013) suosittelevat käyttämään modifioitua pistemäärää, jossa etäisyys jakauman odotusarvosta kuvataan takaisin painoindeksiavaruuteen. Tätä on syytä punnita kehitystarpeena myös FinLapset-hankkeen tulevissa lasten ja nuorten pituus- ja painotietoja käsittelevissä julkaisuissa.

FinLapset-hankkeessa havaintoja rajattiin lisäksi siten, että lapsilta ja nuorilta valittiin kunkin tutkimuksessa mukana olleen kalenterivuoden syntymäpäivää lähinnä ollut mittaus. Mikäli yhtäkään mittausta ei ollut 180 vuorokauden absoluuttisella etäisyydellä kalenterivuoden syntymäpäivästä, ei kyseiselle lapselle otettu mittausta mukaan kyseiseltä kalenterivuodelta.

Tulokset julkaistiin vuosilta 2014–2018 sellaisilta lapsilta ja nuorilta, jotka olivat mittaushetkellä 2–16-vuotiaita. (THL, 2019).

Tässä tutkielmassa noudatamme FinLapset-aineiston rajauksia muilta osin, lukuu-

nottamatta rajausta yhteen mittaukseen kalenterivuotta kohden tai julkaisun yhdeydessä suoritettua kalenterivuoden ja iän perusteella tehtyä rajausta. Tutustumme seuraavaksi aineistoon yksinomaan tämän tutkielman piirissä.

2.2.2 Rekisteripohjainen lasten ja nuorten neuvola- ja kouluterveysaineisto pitkittäisaineistona

Esittäessämme FinLapset-aineiston formaalisti pitkittäisaineiston muodossa, voimme, Fitzmaurice *ym.* (2011) esitystapaa mukaillen, aloittaa määrittelemällä painoindeksin satunnaismuuttujana Y_{ij} ja aineistosta havaitun arvon y_{ij}, jossa $i = 1, \ldots, N$ on yksilöön viittaava indeksi ja $j = 1, \ldots n_i$ yksilön i mittaukseen viittava indeksi. On tärkeää huomioida, että FinLapset-aineistossa lasten mittausmäärät vaihtelevat yksilöittäin, eli on mahdollista, että $n_i \neq n_k$, kun $i, k = 1, \ldots, N$ ja $i \neq k$. FinLapset-aineisto on siten epätasapainoinen pitkittäisaineisto.

Yksilön i painoindeksejä vastaavan satunnaismuuttujan vektori voidaan kirjoittaa $n_i \times 1$ matriisina

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix}$$

Mittausajankohta FinLapset-aineistossa on käynnin päivämäärä, jolla pituus- ja painomittaus on tehty. Tästä seuraa, että yksilöitä ei ole välttämättä mitattu samoina ajankohtina. Siten eräs mittausajankohdan luonteva määritelmä on yksilön i j:nnen mittauksen päivämäärä $t_{\mathrm{PVM}_{ij}}$.

Pitkittäisaineiston aikaskaala voidaan määritellä muillakin tavoin, esimerkiksi lapsen mittausajankohdan desimaali-iän mukaan, jolloin $t_{\mathrm{IKÄ}_{ij}} = \frac{t_{\mathrm{PVM}_{ij}} - t_{\mathrm{SPVM}_i}}{365{,}25}$, jossa t_{SPVM_i} on yksilön i syntymäpäivä. Jakajana käytetään lukua 365,25, jolla pyritään huomioimaan karkausvuoden vaikutus.

Mittauksiin voi liittyä myös taustatietoja, jotka voivat olla aika-invariantteja (*time invariant*), kuten biologinen syntymäsukupuoli ja -kunta tai ajassa muuttuvia (*time variant*), kuten aika ensimmäisestä mittauksesta tai mittaushetken asuinkunta.

Vasteeseen Y_{ij} liittyvät taustatiedot voidaan kirjoittaa $p \times 1$ taustamuuttujavektorina.

$$\boldsymbol{X}_{ij} = \begin{bmatrix} X_{ij_1} \\ X_{ij_2} \\ \vdots \\ X_{ij_p} \end{bmatrix},$$

jossa $i = 1, \ldots, N$ ja $j = 1, \ldots n_i$.

Yksilön i taustamuuttujavektorit voidaan kirjoittaa siistimmin matriisina

$$\boldsymbol{X}_i = \begin{bmatrix} X_{i_1}^\top \\ X_{i_2}^\top \\ \vdots \\ X_{i_{n_i}}^\top \end{bmatrix} = \begin{bmatrix} X_{i_{11}} & X_{i_{12}} & \ldots & X_{i_{1p}} \\ X_{i_{21}} & X_{i_{22}} & \ldots & X_{i_{2p}} \\ \vdots & \vdots & \ddots & \vdots \\ X_{i_{n_i1}} & X_{i_{n_i2}} & \ldots & X_{i_{n_ip}} \end{bmatrix}$$

Muodostamme seuraavaksi FinLapset-aineistosta tämän tutkielman näkökulmasta rajatun aineiston.

2.3 Tutkielman aineiston rajaus

Tutkielman keskeinen kohde on lineaaristen sekamallien soveltuvuuden tarkastelu analyysin kohteen ollessa rekisteriaineistosta muodostettu pitkittäisaineisto. Siten kysymykset liittyvät aineistossa esiintyviin ilmiöihin tilastollisten mallien sovellettavuuden ja toiminnan näkökulmasta.

Rekisteriaineistona FinLapset-aineisto sisältää useita tunnettuja ja tuntemattomia harhan ja mittausvirheen lähteitä, mm. tietojen puuttuminen eräiden kuntien osalta ja aineiston edustavuuden vahvistamattomuus. Tällaiset tarkastelut ja korjaavat toimenpiteet on tässä tutkielmassa pitkälti sivuutettu ja siten tämän tutkielman piirissä aineiston taustalla vallitsevista ilmiöistä tehtävien johtopäätösten suhteen tulee suorittaa varovaisuutta.

Alkuperäinen FinLapset-aineisto käsittää painoindeksihavaintoja 2.1.2013 ja 9.4.2020 väliseltä ajalta. Alkuperäisen aineiston tunnuslukuja on tulevissa tarkasteluissa merkitty selvyyden vuoksi tähdellä (*).

Havaintoja on kaikkiaan $N* = 718080$ yksilöltä ja mittausten kokonaismäärä $n* = \sum_{i=1}^{N*} n_i = 2768012$.

Yksilökohtaisissa mittausten määrissä n_i on merkittävää vaihtelua. Mittausten yksilökohtaisten määrien mediaani on 4, kvartiiliväli $[2, 5]$ ja vaihteluväli $[1, 116]$.

n_i	N*
1	136319
2	103017
3	103836
4	107896
5	100783
6	79459
7	46876
8	21040
9	8420
10-29	10333
30-	101

Taulukko 1: FinLapset-aineiston yksilökohtaisten mittauslukumäärien jakauma.

Mittausten määrien jakaumista (Taulukko 1) huomaamme, että suurimman yksittäisen ryhmän muodostavat yksilöt, joilla on vain yksi mittaus. Tämä on syytä huomioida, sillä yleisesti pitkittäisaineiston oletetaan sisältävän yksilöä kohden vähintään kaksi mittausta (West *ym.*, 2014). Yksilöitä, joilla on yli 30 mittausta on koko ainestossa vain 101 kappaletta.

Täyttääksemme pitkittäisaineiston vaatimukset, tutkielman aineistoon valittiin sellaiset yksilöt, joilla on 2–9 mittausta. Lisäksi, alkuperäisen FinLapset-aineiston sijaan tutkielman aineisto rajoittuu yksilöihin, joiden kotikuntana on kunkin kalenterivuoden lopussa ollut Helsinki. Yksilöiltä, joiden kotikunta on muuttunut seuranta-aikana, otetaan mukaan kaikki saatavilla olevat mittaukset.

Tulee kuitenkin painottaa, että rajattunakin aineisto sisältää huomattavan määrän havaintoja suurelta määrältä yksilöitä ja se mahdollistaa samankaltaisten haasteiden kohtaamisen kuin alkuperäinen FinLapset-aineisto.

Tutkielman rajatussa aineistossa havaintoja on kaikkiaan $N = 101239$ yksilöltä ja mittausten kokonaismäärä $n = \sum_{i=1}^{N} n_i = 421660$.

Vaikka yksilökohtaisten mittausmäärien vaihteluväli pieneni (Taulukko 2), mediaani ja kvartiiliväli pysyivät samana.

Tutkielman aineston muuttujat

n_i	N
2	20168
3	21115
4	20186
5	17070
6	12144
7	6586
8	2811
9	1159

Taulukko 2: Tutkielman rajatun aineiston yksilökohtaisten mittauslukumäärien jakauma.

Tutkielmassa hyödynnämme seuraavia FinLapset-aineiston muuttujia

- Käynnin yhteydessä kirjattu painoindeksi
- Käynnin yhteydessä kirjattu ikä
- Biologinen sukupuoli

Sukupuolijakauma (Taulukko 3) vaikuttaa kuvaavaan likimain Suomen väestöä, jossa tyttöjen osuus on hieman poikien osuutta pienempi.

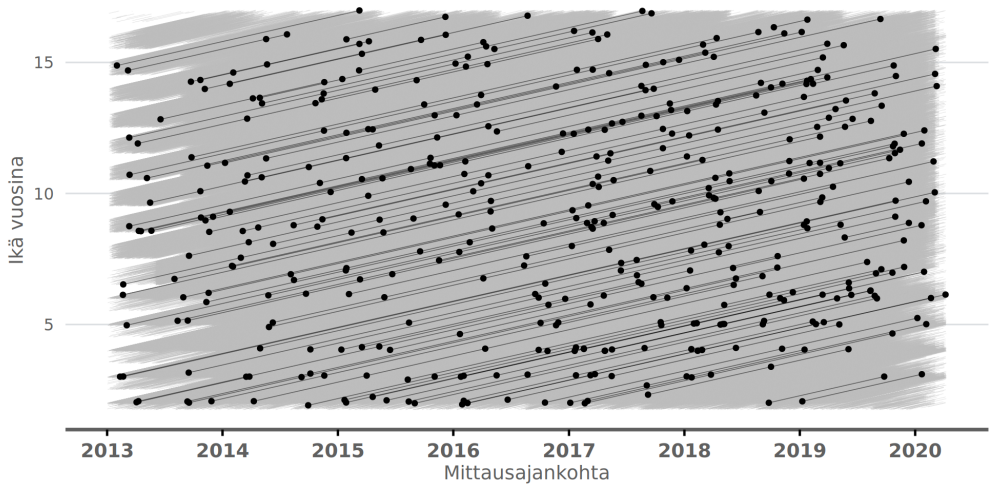
Sukupuoli	N
Tytöt	49837
Pojat	51402

Taulukko 3: FinLapset-aineiston sukupuolijakauma.

Tutkielman aineistossa esiintyy kolme kiinnostavaa aikadimensiota. Ensimmäinen näistä on yksilön ikä mittaushetkellä, toinen on aika, jolloin mittaus on suoritettu ja kolmas on syntymäkohortti. Näillä dimensioilla on muihin muuttujiin verrattuna poikkeuksellinen ominaisuus, sillä kunkin dimension voi yksikäsitteisesti päätellä kahdesta muusta. Nämä dimensiot muodostavat ikä-periodi-kohortti-ilmiön, jota on analysoitu myös lineaaristen sekamallien kirjallisuudessa. (Yang ja Land, 2006).

Tämän tutkielman osalta oleellisia ovat dimensiot mittausaika ja yksilön ikä mittaushetkellä. Näitä voi luontevasti havainnollistaa esimerkiksi Lexis-diagrammeilla

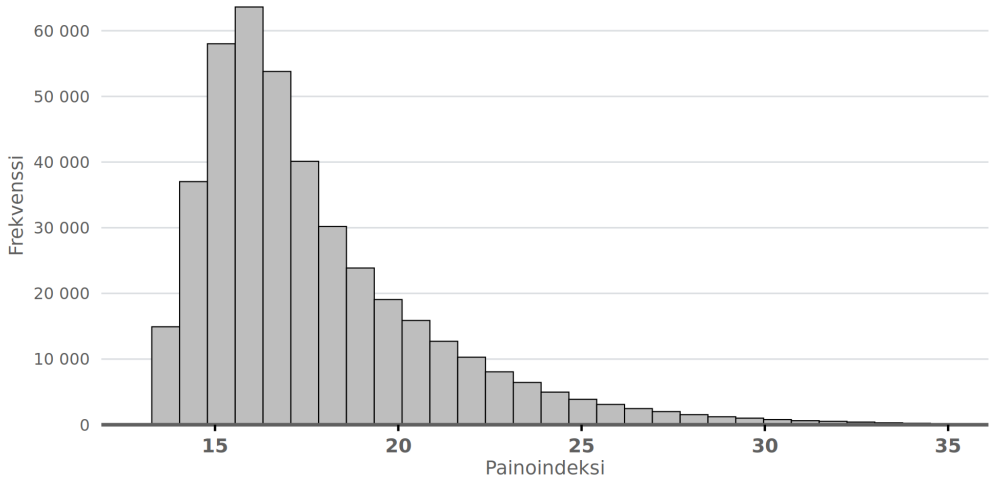
(Kuva 2), joissa pitkittäisaineisto kuvataan yksilökohtaisina polkuina, x-akselin kuvatessa aikadimensiota ja y-akselin ikää mittaushetkellä.



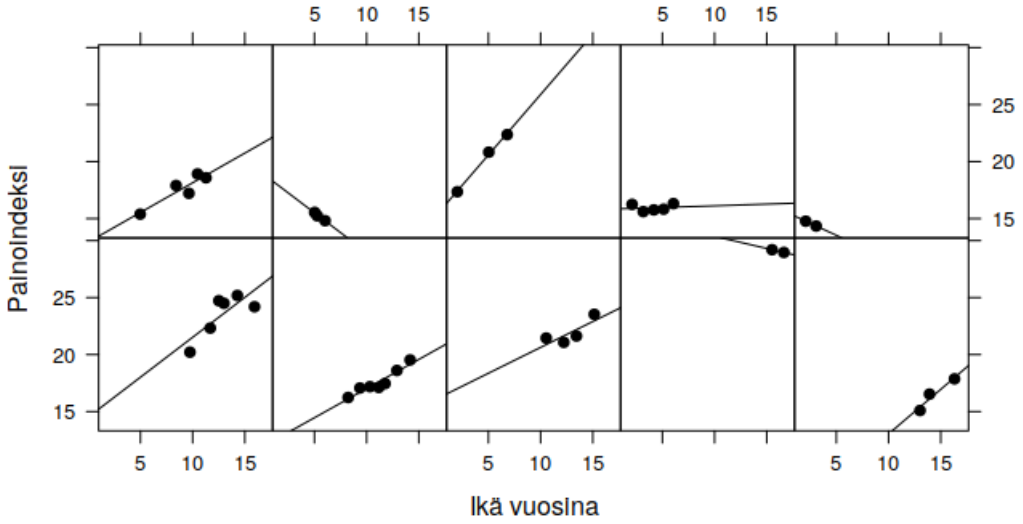
Kuva 2: Mittausajankohdan ja iän suhdetta havainnollistava Lexis-diagrammi. Mahdollisia seurantapolkuja ja havaintoja kuvattu mustilla viivoilla ja palloilla.

Tärkeimpänä muuttujana tarkastelemme vastemuuttuja painoindeksiä. Painoindeksin jakauma (Kuva 3) on positiivisesti vino. Tämä ei kuitenkaan välttämättä ole ongelma, sillä normaalisuusoletus koskee yksinomaan jäännösvirheitä (West *ym.*, 2014). Vaikka vastemuuttujan ja jäännösvirheiden jakaumat ovat tavallisten lineaaristen mallien tapauksessa analogisia (Fitzmaurice *ym.*, 2011), lineaaristen sekamallien tapauksessa tarkastelemme yksilön toistettujen mittausten jäännösvirheitä.

Toisin sanoen, kunkin yksilön havaintoihin sovitetun lineaarisen mallin (Kuva 4) jäännösvirheiden tulisi noudattaa yksiulotteista normaalijakaumaa. Esimerkkikuva havainnollistaa tämänkaltaista yksilöiden välistä vaihtelua ja lineaarisuusoletusta.



Kuva 3: Painoindeksihavaintojen empiirinen jakauma.



Kuva 4: Mahdollisia regressiosuoria ja havaintoja kuvattu mustilla viivoilla ja palloilla.

3 Lineaarinen sekamalli

Lineaariset sekamallit (LSM) ovat tilastollisten mallien joukko lähtökohtaisesti jatkuville vastemuuttujille, joiden jäännösvirheet (residuaalit) ovat normaalisti jakautuneita, mutta eivät välttämättä riippumattomia tai niiden varianssi ei ole vakio (West *ym.*, 2014). Kyseisten mallien joukkoa kuvaava, yleisesti käytössä oleva englan-

ninkielinen nimi *linear mixed (effect) model*, juontuu mallien ominaisuuksista siten, että mallit ovat parametreiltään lineaarisia ja niihin voi liittyä taustamuuttujien muodossa sekä *kiinteitä* (*fixed*) ja *satunnaisia* (*random*) vaikutuksia (*effects*). (West *ym.*, 2014).

Laird ja Ware (1982) kuvaavat kiinteiden ja satunnaisvaikutusten yhteyttä seuraavalla tavalla: Olkoon toistettujen mittausten yksilökohtaiset todennäköisyysjakaumat samaa muotoa (*form*) kaikille yksilöille, mutta sallittakoon näiden todennäköisyysjakaumien parametrien vaihtelu yksilöiden välillä. Näiden populaation *satunnaisvaikutusten* parametrien jakauma muodostaa mallin toisen tason.

Siirtääksemme Laird ja Ware (1982) esittämän esimerkin FinLapset-aineiston kehikkoon oletamme siis, että yksittäisen lapsen painoindeksin ja valittujen taustamuuttujien (esimerkiksi ikä, sukupuoli jne.) välinen yhteys on lineaarinen, mutta kutakin lasta kohden sovitetun lineaarisen regression parametrit voivat vaihdella. Siten, populaation regressioparametrien noudattaessa 2-ulotteista normaalijakaumaa, yksilökohtaisten toistettujen mittausten reunajakauma noudattaa moniulotteista normaalijakaumaa asetelmaa vastaavalla kovarianssirakenteella.

3.1 Lineaarisen sekamallin määritelmä

Jotta voimme perusteellisesti ymmärtää lineaarisen sekamallin luonteen, aloitamme tarkastelun hyvin yleisestä tilastollisen mallin esitysmuodosta, yksinkertaisesta lineaarisesta regressiomallista.

3.1.1 Lineaarinen regressio

Olkoon y_i satunnaismuuttujan Y_i havaittu arvo, x_i satunnaismuuttujan X_i havaittu arvo, β_0 (vakiotermi) ja β_1 (taustamuuttujan x_i regressiokerroin) kiinteitä, mutta tuntemattomia regressiokertoimia ja satunnaismuuttuja ϵ_i mallin selittämätön osa (jäännöstermi) havainnoille $i = 1, \ldots, n$.

Satunnaismuuttujan Y_i ja taustamuuttujan x_i yhteyttä kuvaa yksinkertainen lineaarinen regressiomalli

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

Laajennettuna usean taustamuuttujan lineaariseen regressioon edellinen voidaan esittää muodossa

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Tiivistääksemme edellistä muotoa, voimme määritellä sarakevektorit $\boldsymbol{x_i} = [x_{i1}, \ldots, x_{ip}]^\top$ ja $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_p]^\top$ ja kirjoittaa mallin muodossa

$$Y_i = \boldsymbol{x_i}^\top \boldsymbol{\beta} + \epsilon_i$$

havainnoille $i = 1, \ldots, n$.

Seuraavaksi hyödynnämme lineaarisen mallin määritelmää esitellessämme lineaariselle sekamallille ensin kaksitasoisen mallin, jonka jälkeen etenemme luontevasti tasot yhdistävään Laird ja Ware (1982) yleiseen lineaariseen sekamalliin.

3.1.2 Kaksitasomalli

Monitasomalleille keskeistä on aineiston hierarkkisen rakenteen huomioiminen mallissa *satunnaisvaikutusten* avulla (Talbott, 2006). Pitkittäisaineistoa kuvaavassa kaksitasomallissa hierarkian ensimmäinen taso käsittää yksilön mittauskertojen välisen vaihtelun ja toinen taso yksilöiden välisen vaihtelun. Monitasomallit voidaan yleistää myös useammalle tasolle (Goldstein, 2011; Burzykowski ja Galecki, 2013).

Ensimmäinen taso

Ensimmäisellä tasolla oletamme kunkin yksilön keskimääräisen vasteen (*Mean response*) noudattavan lineaarista regressiomallia samoilla taustamuuttujilla, mutta yksilöllisillä regressiokertoimilla (Fitzmaurice *ym.*, 2011).

Edellistä mukaillen, voimme esittää ensimmäisen tason muodossa

$$\boldsymbol{Y}_i = \boldsymbol{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i,$$

jossa $\boldsymbol{Y}_i$ on yksilön i n_i-vastevektori, $\boldsymbol{\beta}_i$ tuntemattomat regressiokertoimet sisältävä p-vektori ja $\boldsymbol{Z}_i\boldsymbol{\beta}_i$ yksilön i tuntematon keskimääräinen vasteen kehitys (*Response trajectory*). Ensimmäisen tason kontekstissa matriisi $\boldsymbol{Z}_i$ määrittelee kuinka yksilön keskimääräinen vaste muuttuu ajassa, mutta määritelmä tulee vaatimaan tarkennusta.

Niin sanottujen yksilöiden välillä vaihtelevien kiinteiden, mutta tuntemattomien vaikutusten lisäksi ensimmäiseen tasoon liittyy myös yksilön i mittauksia koskeva satunnaisvirhe $\boldsymbol{\epsilon_i}$, jossa $\boldsymbol{\epsilon_i} \sim N(0, \boldsymbol{R}_i)$, jossa $\boldsymbol{R}_i$ on jokin $n_i \times n_i$ kovarianssimatriisi. Yksinkertaisuuden vuoksi voimme olettaa, että $\boldsymbol{R}_i = \sigma^2\boldsymbol{I}_{n_i}$.

Toinen taso

Toisella tasolla tuomme malliin oletuksen, että yksilökohtaiset vaikutukset $\boldsymbol{\beta}_i$ ovat satunnaismuuttujia jollakin odotusarvolla ja kovarianssilla.

Keskeistä on $\boldsymbol{\beta}_i$ yksilöiden välisen vaihtelun mallintaminen mittausajankohdasta riippumattomien vasteen yksilöllisiä eroja selittävien taustamuuttujien funktiona (Fitzmaurice *ym.*, 2011).

Siten $\boldsymbol{\beta}_i$ odotusarvoksi saamme

$$\mathrm{E}(\boldsymbol{\beta}_i) = \boldsymbol{K}_i\boldsymbol{\beta},$$

jossa $\boldsymbol{K}_i$ on yksilöiden välistä vaihtelua selittävien muuttujien $(q \times p)$ matriisi.

Yksilöiden välistä residuaalivaihtelua, jota $\boldsymbol{K}_i$ ei selitä, merkitsemme

$$\mathrm{Cov}(\boldsymbol{\beta}_i) = \boldsymbol{G}.$$

Lopuksi voimme yhdistää ensimmäisen ja toisen tason komponentit esittääksemme $\boldsymbol{Y}_i$:lle lineerisen sekamallin yleisen muodon. (Fitzmaurice *ym.*, 2011; Verbeke ja Molenberghs, 2000).

3.1.3 Yleinen lineaarinen sekamalli

Yhdistääksemme ensimmäisen ja toisen tason, hyödyntäen toisen vaiheen oletuksia, voimme kirjoittaa $\boldsymbol{\beta}_i$ uudelleen muodossa

$$\boldsymbol{\beta}_i = \boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i,$$

jossa $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$.

Tässä $\boldsymbol{b}_i$ kuvaa i:nnen yksilön poikkeamaa keskimääräisestä vasteesta, kun yksilöiden yhteiset taustamuuttujat on huomioitu. (Fitzmaurice *ym.*, 2011).

Yhdistetty muoto saadaan sijoittamalla

$$\begin{aligned}
\boldsymbol{Y}_i &= \boldsymbol{Z}_i\boldsymbol{\beta_i} + \boldsymbol{\epsilon}_i \\
&= \boldsymbol{Z}_i(\boldsymbol{K}_i\boldsymbol{\beta} + \boldsymbol{b}_i) + \boldsymbol{\epsilon}_i \\
&= (\boldsymbol{Z}_i\boldsymbol{K}_i)\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\
&= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i,
\end{aligned}$$

jossa $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{K}_i$.

Yleisesti muotoiltuna, malli joka tyydyttää vaatimukset

$$\begin{cases}
\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i \\
\boldsymbol{b}_i \sim N(0, \boldsymbol{G}) \\
\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i) \\
\boldsymbol{b}_i, \ldots, \boldsymbol{b}_N \perp \boldsymbol{\epsilon}_i, \ldots, \boldsymbol{\epsilon}_N
\end{cases}$$

on lineaarinen sekamalli. (Verbeke ja Molenberghs, 2000; Laird ja Ware, 1982).

Fitzmaurice *ym.* (2011) alleviivaavat, että kaksitasomalliin, sen ollessa pedagogisesti hyödyllinen lineaarisen sekamallin yksilöiden välisten ja yksilökohtaisten vaikutusten esitystapa, liittyy merkittävä rajoite, sillä muoto $\boldsymbol{X}_i = \boldsymbol{Z}_i\boldsymbol{K}_i$ edellyttää, että $\boldsymbol{K}_i$ sisältää vain yksilöiden välisiä vaikutuksia selittäviä muuttujia ja $\boldsymbol{Z}_i$ vain yksilökohtaisia ajassa muuttuvia vaikutuksia selittäviä muuttujia.

Tämä rajoite on kuitenkin kierrettävissä olettamalla, että $\boldsymbol{X}_i$ on lineaarisen sekamallin mielivaltainen koematriisi ja, että $\boldsymbol{Z}_i$ muodostuu sarakkeista, jotka ovat $\boldsymbol{X}_i$ sarakkeiden osajoukko. Tällöin vasteen $\boldsymbol{Y}_i$ odotusarvo yli kaikkien yksilökohtaisten vaikutusten

$$\mathrm{E}(\boldsymbol{Y}_i) = \boldsymbol{X}_i\boldsymbol{\beta},$$

määrittelee vasteen keskimääräisen kehityksen koko tarkasteltavana olevalle joukolle ja $\boldsymbol{Z}_i\boldsymbol{b}_i$ kuvaa selittämättä jäänyttä osaa, eli yksilökohtaisia poikkeamia tarkastelujoukon vasteen keskimääräisestä kehityksestä. (Fitzmaurice *ym.*, 2011).

Täsmällisemmin muotoiltuna, tarkastellaan ehdollista odotusarvoa

$$\mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i.$$

Koska $\mathrm{E}(\boldsymbol{b}_i) = 0$, saamme reunajakauman $\boldsymbol{Y}_i$ odotusarvoksi

$$\begin{aligned}\mathrm{E}(\boldsymbol{Y}_i) &= \mathrm{E}(\mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{b}_i)) \\ &= \mathrm{E}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathrm{E}(\boldsymbol{b}_i) \\ &= \boldsymbol{X}_i\boldsymbol{\beta}.\end{aligned}$$

Siten vahvistamme tulkinnan, että parametrit $\boldsymbol{\beta}$ kuvaavat koko tarkastelujoukon yhteisiä, niin kutsuttuja *kiinteitä* vaikutuksia ja ottaen huomioon kiinteät vaikutukset, $\boldsymbol{b}_i$ kuvaa yksilön i poikkeamaa koko tarkastelujoukon keskimääräisesta vasteesta, toisin sanoen mallin *satunnaisvaikutuksia*.

Fitzmaurice *ym.* (2011) mukaan lineaarisen sekamallin kaksitasoesitys asettaa erityisvaatimuksia myös mallin kovarianssille. Koska kovarianssirakenteella on monin tavoin keskeinen asema, sekä yleisen lineaarisen sekamallin teoriassa, että pitkittäisaineiston erityistapauksessa, erotamme sen käsittelyn omaksi alaluvukseen.

3.1.4 Kovarianssirakenteet

Tarkastellaan seuraavaksi lineaarisen sekamallin kovarianssirakenteita Fitzmaurice *ym.* (2011) esittämässä yleisessä muodossa. Lähtökohtana on erottaa yksilökohtaisen ja yksilöiden välisen vaihtelun lähteet ja mahdollistaa näiden yksikäsitteinen

analyysi.

Yleisen lineaarisen sekamallin

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

tapauksessa voimme, edellisessä luvussa esitetyn määritelmän mukaan, esittää yksilön i keskimääräisen vasteen ehdollisena odotusarvona

$$\mathrm{E}(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i,$$

jossa otamme huomioon yksilön i poikkeaman $\boldsymbol{b}_i$ koko tutkimusjoukon keskimääräisestä vasteesta.

Pyrkimyksessämme erottaa yksilön i mittausten keskinäinen vaihtelu koko tutkimusjoukon mittausten vaihtelusta, määritellään yksilön i mittausten kovarianssi samaan tapaan ehdollisena kovarianssina

$$\mathrm{Cov}(\boldsymbol{Y}_i|\boldsymbol{b}_i) = \mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{R}_i,$$

jolloin Fitzmaurice *ym.* (2011) mukaan reunajakauman $\boldsymbol{Y}_i$ kovarianssiksi saamme

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{Y}_i) &= \mathrm{Cov}(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i)\\ &= \mathrm{Cov}(\boldsymbol{Z}_i\boldsymbol{b}_i) + \mathrm{Cov}(\boldsymbol{\epsilon}_i)\\ &= \boldsymbol{Z}_i\mathrm{Cov}(\boldsymbol{b}_i)\boldsymbol{Z}_i^\top + \mathrm{Cov}(\boldsymbol{\epsilon}_i)\\ &= \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top + \boldsymbol{R}_i.\end{aligned}$$

Näin ollen, lineaarinen sekamalli mahdollistaa yksilöiden välisten varianssilähteiden $\boldsymbol{G}$ ja yksilökohtaisten varianssilähteiden $\boldsymbol{R}_i$ yksikäsitteisen analyysin.

Koska mallin kovarianssi esitetään mittausajankohtien funktiona, ei satunnaisvaikutusten kovarianssirakenne vaadi tasapainoista asetelmaa pitkittäisaineistossa, kovarianssiparametrien määrä ei riipu mittausajankohdista tai niiden lukumäärästä ja varianssi sekä kovarianssi voivat myös muuttua mittausajankohtien funktiona, mistä on merkittävää hyötyä epätasapainoisen pitkittäisaineiston analyysissä. (Fitzmaurice *ym.*, 2011).

Kirjoitetaan yksilöiden välisten satunnaisvaikutusten q-vektori $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$ yksilölle i matriisimuodossa

$$\boldsymbol{b}_i = \begin{bmatrix} \mathrm{b}_{1i} \\ \vdots \\ \mathrm{b}_{qi} \end{bmatrix},$$

jolloin kovarianssi $\mathrm{Cov}(\boldsymbol{b}_i) = \boldsymbol{G}$ voidaan esittää muodossa

$$\boldsymbol{G} = \begin{bmatrix} \mathrm{Var}(\mathrm{b}_{1i}) & \dots & \mathrm{Cov}(\mathrm{b}_{1i}, \mathrm{b}_{qi}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathrm{b}_{1i}, \mathrm{b}_{qi}) & \dots & \mathrm{Var}(\mathrm{b}_{qi}) \end{bmatrix}.$$

Lisäksi kirjoitetaan yksilön i mittausten n_i satunnaisvirhe, eli residuaali $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$ yksilölle i matriisimuodossa

$$\boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{1i} \\ \vdots \\ \epsilon_{n_i i} \end{bmatrix},$$

jolloin kovarianssi $\mathrm{Cov}(\boldsymbol{\epsilon}_i) = \boldsymbol{R}_i$ voidaan esittää muodossa

$$\boldsymbol{R}_i = \begin{bmatrix} \mathrm{Var}(\epsilon_{1i}) & \dots & \mathrm{Cov}(\epsilon_{1i}, \epsilon_{n_i i}) \\ \vdots & \ddots & \vdots \\ \mathrm{Cov}(\epsilon_{1i}, \epsilon_{n_i i}) & \dots & \mathrm{Var}(\epsilon_{qi}). \end{bmatrix}$$

Burzykowski ja Galecki (2013) sekä West *ym.* (2014) mukaan matriisin $\boldsymbol{G}$ alkiot voidaan kuvata jonkin kovarianssiparametrivektorin $\boldsymbol{\theta}$ alkioiden funktioina $\boldsymbol{G} = \sigma(\boldsymbol{\theta}_G)$.

Jos emme aseta muita vaatimuksia, kuin että $q \times q$ matriisi $\boldsymbol{G}$ on symmetrinen ja positiivisesti definiitti, käytetään nimitystä *rakenteeton* kovarianssirakenne. Tässä tapauksessa $\boldsymbol{G}$ symmetrisyyden johdosta $\boldsymbol{\theta}_G$ sisältää $(q(q+1))/2$ parametria. (Burzykowski ja Galecki, 2013; West *ym.*, 2014).

Tarkastellaan tarkemmin matriisin $\boldsymbol{G}$ mahdollisia kovarianssirakenteita. Esimerkiksi jos rakenteettoman kovarianssirakenteen tapauksessa oletetamme malliin kaksi satunnaisvaikutusta, satunnaisen vakiotermin ja satunnaisen regressiokertoimen. Tällöin kovarianssiparametrien määrä olisi $q = 2(2+1)/2 = 3$, jolloin $\boldsymbol{\theta}_G$ voitaisiin kirjoittaa muodossa

$$\boldsymbol{\theta}_G = \begin{bmatrix} \sigma^2_{\mathrm{b}_{1i}} \\ \sigma_{\mathrm{b}_{1i}, \mathrm{b}_{2i}} \\ \sigma^2_{\mathrm{b}_{2i}} \end{bmatrix}$$

ja $\boldsymbol{G}$ muodossa

$$\begin{aligned}\boldsymbol{G} &= \begin{bmatrix} \mathrm{Var}(\mathrm{b}_{1i}) & \mathrm{Cov}(\mathrm{b}_{1i}, \mathrm{b}_{2i}) \\ \mathrm{Cov}(\mathrm{b}_{1i}, \mathrm{b}_{2i}) & \mathrm{Var}(\mathrm{b}_{2i}) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2_{\mathrm{b}_{1i}} & \sigma_{\mathrm{b}_{1i},\mathrm{b}_{2i}} \\ \sigma_{\mathrm{b}_{1i},\mathrm{b}_{2i}} & \sigma^2_{\mathrm{b}_{2i}} \end{bmatrix}\end{aligned}$$

Matriisin $\boldsymbol{G}$ rakennetta voidaan yksinkertaistaa vaihtoehtoisilla parametrisoinneilla kuten esimerkiksi *varianssikomponenttirakenteessa*, jossa jokaisella satunnaisvaikutuksella $\boldsymbol{b_i}$ on yksilöllinen varianssi ja kovarianssit ovat rajoitettu nollaan. Tässä tapauksessa $\boldsymbol{\theta}_G$

$$\boldsymbol{\theta}_G = \begin{bmatrix} \sigma^2_{\mathrm{b}_{1i}} \\ \sigma^2_{\mathrm{b}_{2i}} \end{bmatrix}$$

määrittelee matriisin $\boldsymbol{G}$ diagonaalin

$$\boldsymbol{G} = \begin{bmatrix} \sigma^2_{\mathrm{b}_{1i}} & 0 \\ 0 & \sigma^2_{\mathrm{b}_{2i}} \end{bmatrix}$$

Rakenteeton kovarianssirakenne ja varianssikomponenttirakenne ovat yleisimmin sovelletut matriisin $\boldsymbol{G}$ rakenteet, mutta myös muita rakenteita on esitetty. (West *ym.*, 2014).

Siirrymme tarkastelemaan matriisin $\boldsymbol{R}_i$ mahdollisia kovarianssirakenteita. Määrittelemme $\boldsymbol{R}_i$ alkiot kovarianssiparametrivektorin $\boldsymbol{\theta_R}$ alkioiden funktioina.

Yksinkertaisin kovarianssirakenne matriisille $\boldsymbol{R}_i$ on diagonaalirakenne, jossa yksilön i mittauksiin liittyvät residuaalit oletetaan keskenään korreloimattomiksi ja niiden varianssit vakioiksi. Diagonaalimatriisi $\boldsymbol{R}_i$ on muotoa

$$\boldsymbol{R}_i = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}$$

ja parametrivektori $\boldsymbol{\theta}_R = [\sigma^2]$ sisältää vain yhden parametrin. Yksilöiden i mittausten välillä ei siis oleteta olevan korrelaatiota, mikä pitkittäisaineiston tapauksessa on hyvin epärealistinen oletus.

Eräänä yksinkertaisena vaihtoehtona on esitetty *tasakorrelaatiorakennetta*, jossa yksilön i mittausten residuaaleille oletetaan vakiokorrelaatio.

$$\boldsymbol{R}_i = \begin{bmatrix} \sigma^2 + \sigma_1 & \dots & \sigma_1 \\ \vdots & \ddots & \vdots \\ \sigma_1 & \dots & \sigma^2 + \sigma_1 \end{bmatrix}$$

jonka parametrit muodostavat parametrivektorin

$$\boldsymbol{\theta}_R = \begin{bmatrix} \sigma^2 \\ \sigma_1 \end{bmatrix}$$

Vaikka tasakorrelaatiorakenteen esitetään soveltuvan tilanteisiin, joissa toistetut mittaukset suoritetaan identtisissä olosuhteissa, sitä pidetään yleisesti liian optimistisena pitkittäisaineiston kannalta. (Pinheiro ja Bates, 2000; Fitzmaurice *ym.*, 2011; West *ym.*, 2014).

On luontevaa ajatella pitkittäisaineiston yksilön i mittausten välillä vallitsevan jokin korrelaatio ja ajallisesti lähellä olevien mittaukset olevan samankaltaisempia kuin ajallisesti etäämmällä olevien. Yksilön mittausten residuaaleja kuvaavan kovarianssirakenteen tulisi siis heijastaa tällaista asetelmaa.

Tasapainoisen pitkittäisaineiston tapauksessa *autoregressiivinen* kovarianssirakenne on usein sopiva valinta. Tällöin kovarianssimatriisi $\boldsymbol{R}_i$ on muotoa

$$\boldsymbol{R}_i = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \dots & \sigma^2\rho^{n_i-1} \\ \sigma^2\rho & \sigma^2 & \dots & \sigma^2\rho^{n_i-2} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma^2\rho^{n_i-1} & \sigma^2\rho^{n_i-2} & \dots & \sigma^2 \end{bmatrix}$$

parametrivektorilla

$$\boldsymbol{\theta}_R = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix}$$

Käytännön tapauksissa, koskien vaikkapa epätasapainoista pitkittäisaineistoa, on usein tarve joustavammalle matriisin $\boldsymbol{R}_i$ kovarianssirakenteelle. Sopivan rakenteen valinta vaatii kuitenkin havaitun aineiston ja tutkittavan ilmiön ominaisuuksien tarkastelua. Tähän prosessiin palaamme myöhemmin tutkielmassa (Kappale 4.2.3).

3.2 Estimointi

Estimoinnin näkökulmasta kiinnostuksen kohteena ovat kiinteiden vaikutusten parametrit $\boldsymbol{\beta}$ ja kovarianssiparametrit $\boldsymbol{\theta}_G$ ja $\boldsymbol{\theta}_R$. Yleisimmät menetelmät edellä mainittujen parametrien estimointiin ovat suurimman uskottavuuden (*maximum likelihood, ML*) ja rajoitettu suurimman uskottavuuden (*restricted maximum likelihood, REML*) menetelmät .(West *ym.*, 2014).

Suurimman uskottavuuden menetelmässä tavoitteena on uskottavuusfunktion maksimointi. Uskottavuusfunktio on mallin parametrien funktio ja sen globaali maksimi, parametrin suurimman uskottavuuden estimaatti, on parametrin arvo jolla havaittu aineisto on uskottavin. (Casella ja Berger, 2002).

Lineaarisen sekamallin tapauksessa uskottavuuspäättely perustuu parametrivektoreista $\boldsymbol{\beta}$ ja $\boldsymbol{\theta} = [\boldsymbol{\theta}_G, \boldsymbol{\theta}_R]^\top$ ja $\boldsymbol{Y}_i$ reunajakaumasta muodostettuun uskottavuusfunktioon.

Reunajakauman uskottavuusfunktion muodostamiseksi mm. West *ym.* (2014) ehdottavat lineaarisen sekamallin yleiseen muotoon läheisesti liittyvää marginaalimallia. West *ym.* (2014) mukaan lineaarisen sekamallin yleisestä muodosta voidaan päätellä (*imply*) seuraava marginaalimalli

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i, \tag{1}$$

jossa

$$\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{V}_i)$$

ja kovarianssimatriisi $\boldsymbol{V}_i$ määritellään

$$\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top + \boldsymbol{R}_i$$

West *ym.* (2014) painottavat, että marginaalimalli ei ole ekvivalentti lineaarisen sekamallin yleisen muodon kanssa, sillä marginaalimalli ei sisällä satunnaisvaikutuksia. Satunnaisvaikutuksia vastaava kovarianssirakenne on kuitenkin mahdollista sisällyttää malliin matriisin $\boldsymbol{V}_i$ kautta.

Päättellyn marginaalimallin avulla voimme määritellä vasteen $\boldsymbol{Y}_i$ reunajakauman moniulotteisena normaalijakaumana

$$\boldsymbol{Y}_i \sim N(\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top + \boldsymbol{R}_i)$$

Huomioitavaa on, että $\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top$ kuvastaa yksilöiden välistä vaihtelua ja $\boldsymbol{R}_i$ yksilön i mittausten sisäistä vaihtelua.

Koska $\boldsymbol{V}_i = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top + \boldsymbol{R}_i$ on funktio $\boldsymbol{V}_i(\boldsymbol{\theta})$ kovarianssiparametreillä $\boldsymbol{\theta}$, voimme määritellä marginaalimallia vastaavan tiheysfunktion $f(\boldsymbol{Y}_i\,;\,\boldsymbol{\beta},\boldsymbol{\theta})$ muodossa

$$f(\boldsymbol{Y}_i;\boldsymbol{\beta},\boldsymbol{\theta}) = (2\pi)^{-\frac{n_i}{2}}\det(\boldsymbol{V}_i)^{-\frac{1}{2}}\exp\big(-\frac{1}{2}[\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{Y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]\big),$$

jolloin havaitun aineiston $\boldsymbol{Y}_i = \boldsymbol{y}_i$ uskottavuusfunktioksi saamme N riippumattoman kontribuution tulona

$$\begin{aligned}
L(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) &= \prod_{i=1}^{N} f(\boldsymbol{y}_i;\boldsymbol{\beta},\boldsymbol{\theta}) \\
&= \prod_{i=1}^{N}(2\pi)^{-\frac{n_i}{2}}\det(\boldsymbol{V}_i)^{-\frac{1}{2}}\exp\big(-\frac{1}{2}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]\big),
\end{aligned}$$

jossa $i = (1,\ldots,N)$. Vastaava log-uskottavuusfunktio saadaan luonnollisella logaritmilla

$$\begin{aligned}
l(\boldsymbol{\beta},\boldsymbol{\theta};\boldsymbol{y}) &= \log(\prod_{i=1}^{N} f(\boldsymbol{y}_i\,;\,\boldsymbol{\beta},\boldsymbol{\theta})) \\
&= \log\left(\prod_{i=1}^{N}(2\pi)^{-\frac{n_i}{2}}\det(\boldsymbol{V}_i)^{-\frac{1}{2}}\exp\big(-\frac{1}{2}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]\big)\right), \\
&= \log\left(\prod_{i=1}^{N}(2\pi)^{-\frac{n_i}{2}}\right) + \log\left(\prod_{i=1}^{N}\det(\boldsymbol{V}_i)^{-\frac{1}{2}}\right) \\
&\quad + \log\left(\prod_{i=1}^{N}\exp\big(-\frac{1}{2}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]\big)\right) \\
&= -\frac{1}{2}\sum_{i=1}^{N} n_i\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log(\det(\boldsymbol{V}_i)) - \frac{1}{2}\sum_{i=1}^{N}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}] \\
&= -\frac{1}{2}n\log(2\pi) - \frac{1}{2}\sum_{i=1}^{N}\log(\det(\boldsymbol{V}_i)) - \frac{1}{2}\sum_{i=1}^{N}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}]^\top\boldsymbol{V}_i^{-1}[\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}],
\end{aligned}$$

jossa $n = \sum_{i=1}^{N} n_i$.

West *ym.* (2014) mukaan olisi mahdollista estimoida $\boldsymbol{\beta}_i$ ja $\boldsymbol{\theta}$ samanaikaisesti, mutta käytännössä algoritminen optimointi suoritetaan usein *profiloimalla* $\boldsymbol{\beta}$ ulos log-uskottavuusfunktioista.

Vaikka yleisessä tapauksessa $\boldsymbol{\theta}$ on tuntematon, sellaisen erityistapauksen käsittely, että $\boldsymbol{\theta}$ on tunnettu, tarjoaa tärkeän välivaiheen lopullisen profiili-log-uskottavuusfunktion muodostamiseksi.

Parametrivektorin $\boldsymbol{\theta}$ ollessa tunnettu, estimoitavaksi jäävät ainoastaan kiinteät vaikutukset $\boldsymbol{\beta}_i$. Log-uskottavuusfunktio on tällöin vain parametrin $\boldsymbol{\beta}_i$ funktio ja sen optimointi on yhteensopivaa tappiofunktion (*loss function*) $q(\boldsymbol{\beta})$ minimoinnin kanssa, joka määritellään

$$q(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^{N}[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}]^\top \boldsymbol{V}_i^{-1}[\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}].$$

Funktion $q(\boldsymbol{\beta})$ optimointi voidaan suorittaa yleistetyllä pienimmän neliösumman menetelmällä.

Derivoimalla log-uskottavuusfunktio $\boldsymbol{\beta}$ suhteen saamme

$$\frac{\partial L(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})$$

ja asettamalla $\sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) = 0$ saamme normaaliyhtälömuodon

$$\begin{aligned}
\sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}) &= 0 \\
\sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{y}_i - \sum_{i=1}^{N}(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i)\boldsymbol{\beta} &= 0 \\
\sum_{i=1}^{N}(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i)\boldsymbol{\beta} &= \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{y}_i
\end{aligned}$$

jolloin parametrivektorin $\boldsymbol{\beta}$ optimaaliselle arvolle löydetään suljetun muodon ratkaisu

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{y}_i,$$

joka, $\boldsymbol{y}_i$ ollessa normaalijakautunut, on ominaisuuksiltaan *paras lineaarinen harhaton estimaattori*. (West *ym.*, 2014).

Nyt voimme siirtyä tarkastelemaan kovarianssiparametrivektorin $\boldsymbol{\theta}$ ja kiinteiden parametrien vektorin $\boldsymbol{\beta}$ estimointia yleisessä tapauksessa, kun $\boldsymbol{\theta}$ on tuntematon.

Muodostetaan profiili-log-uskottavuusfunktio $l(\boldsymbol{\theta}; \boldsymbol{y})$ sijoittamalla $\hat{\boldsymbol{\beta}}$ log-uskottavuusfunktioon $l(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{y})$

$$\begin{aligned} l(\boldsymbol{\theta}; \boldsymbol{y}) &= -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log(\det(\boldsymbol{V}_i)) - \frac{1}{2} \sum_{i=1}^{N} [\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}]^\top \boldsymbol{V}_i^{-1} [\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}] \\ &= -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log(\det(\boldsymbol{V}_i)) - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i, \end{aligned}$$

jossa $\boldsymbol{r}_i = \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}} = \boldsymbol{y}_i - \boldsymbol{X}_i (\sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{y}_i$

Kun estimaatit $\hat{\boldsymbol{G}}$ ja $\hat{\boldsymbol{R}}_i$ ovat löytyneet, ratkaistaan $\hat{\boldsymbol{V}}_i$ sijoittamalla

$$\hat{\boldsymbol{V}}_i = \boldsymbol{Z}_i \hat{\boldsymbol{G}} \boldsymbol{Z}_i^\top + \hat{\boldsymbol{R}}_i.$$

Siten $\hat{\boldsymbol{\beta}}_i$ estimaattoriksi saamme sijoittamalla $\hat{\boldsymbol{V}}_i$ yleistetyn PNS-estimaattorin yhtälöön

$$\hat{\boldsymbol{\beta}} = (\sum_{i=1}^{N} \boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{y}_i,$$

joka on ominaisuuksiltaan *empiirinen paras lineaarinen harhaton estimaattori* West *ym.* (2014).

Kovarianssimatriisi estimaattorille $\hat{\boldsymbol{\beta}}$ on

$$Cov(\hat{\boldsymbol{\beta}}) = (\sum_{i=1}^{N} \boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_i^{-1} \boldsymbol{X}_i)^{-1}$$

West *ym.* (2014) mukaan suuriman uskottavuuden kovarianssiparametrien estimaatit ovat kuitenkin harhaisia, sillä niissä ei oteta huomioon p kiinteiden vaikutusten parametrien $\boldsymbol{\beta}$ estimoinnista johtuvaa vapausasteiden menetystä. Diggle *ym.* (2013) esittävät myös, että suurimman uskottavuuden menetelmä on erittäin herkkä väärin spesifioidulle matriisille $\boldsymbol{X}_i$ ja voi johtaa ei-konsistenttiin kovarianssiparametrin estimaattoriin, jolloin estimaattori ei tarkennu asymptoottisesti kohti kovarianssiparametrin todellista arvoa.

Lineaaristen sekamallien tapauksessa vaihtoehdoksi suurimman uskottavuuden menetelmälle suositellaan REML-menetelmää. (Diggle *ym.*, 2013; Pinheiro ja Bates,

2000; Verbeke ja Molenberghs, 2000).

REML-menetelmän estimaattori on suurimman uskottavuuden estimaattori sellaiselle lineaarimuunnokselle $\boldsymbol{K}_i^\top \boldsymbol{y}_i$, jossa $\boldsymbol{K}_i$ on $n_i \times (n_i - p)$ matriisi ja jolle $\boldsymbol{K}_i^\top \boldsymbol{X}_i = 0$ ja

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

Sopiva ehdokas matriisille $\boldsymbol{K}_i$ on projektio $\boldsymbol{Q}_i = \boldsymbol{I}_i - \boldsymbol{X}_i(\boldsymbol{X}_i^\top \boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top$, jolla $\mathrm{E}(\boldsymbol{K}_i^\top \boldsymbol{y}_i) = 0$ riippumatta parametrien $\boldsymbol{\beta}$ arvoista. (Diggle *ym.*, 2013).

Nissinen (2009) mukaan REML-menetelmän uskottavuusfunktio voidaan muotoilla määritelemällä sellainen $n_i \times (n_i - p)$ matriisi $\boldsymbol{A}_i$, jolle $\boldsymbol{A}_i \boldsymbol{A}_i^\top = \boldsymbol{Q}_i$ ja $\boldsymbol{A}_i^\top \boldsymbol{A}_i = \boldsymbol{I}_i$. Silloin $\boldsymbol{A}_i$ on sopiva valinta matriisille $\boldsymbol{K}_i$ ja

$$\boldsymbol{A}_i^\top \boldsymbol{y}_i \sim N(0, \boldsymbol{A}_i^\top \boldsymbol{V}_i \boldsymbol{A}_i).$$

Nissinen (2009) näyttää, kuinka $\boldsymbol{A}_i^\top \boldsymbol{y}_i$ muodostavat tiheysfunktion

$$\begin{aligned} f(\boldsymbol{A}_i^\top \boldsymbol{y}_i; \boldsymbol{\theta}) &= (2\pi)^{-\frac{n_i - p}{2}} \det(\boldsymbol{A}_i^\top \boldsymbol{V}_i \boldsymbol{A}_i)^{-\frac{1}{2}} \\ &\quad \exp\big(-\frac{1}{2} [\boldsymbol{A}_i^\top \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}]^\top \boldsymbol{A}_i^\top (\boldsymbol{A}_i^\top \boldsymbol{V}_i \boldsymbol{A}_i)^{-1} \boldsymbol{A}_i [\boldsymbol{A}_i^\top \boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}] \big) \\ &= (2\pi)^{-\frac{n_i - p}{2}} \det(\boldsymbol{X}_i^\top \boldsymbol{X}_i)^{\frac{1}{2}} \det(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)^{-\frac{1}{2}} \det(\boldsymbol{V}_i)^{-\frac{1}{2}} \\ &\quad \exp\big(-\frac{1}{2} [\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}]^\top \boldsymbol{V}_i^{-1} [\boldsymbol{y}_i - \boldsymbol{X}_i \hat{\boldsymbol{\beta}}] \big) \end{aligned}$$

jossa $\hat{\boldsymbol{\beta}}$ yleistetty PNS-estimaattori.

Nissinen (2009) osoittaa myös, että $\boldsymbol{A}_i$ liittyy tiheysfunktioon vain epäsuorasti $\boldsymbol{X}_i^\top \boldsymbol{X}_i$ kautta, joka puolestaan ei riipu matriisista $\boldsymbol{V}_i$, eikä myöskään vaikuta uskottavuusfunktion optimointiin ja REML-menetelmän log-uskottavuusfunktio voidaan siten kirjoittaa

$$l(\boldsymbol{V}_i; \boldsymbol{y}_i) = -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \log(\det(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i)) - \frac{1}{2} \sum_{i=1}^{N} \log(\det(\boldsymbol{V}_i)) - \frac{1}{2} \sum_{i=1}^{N} \boldsymbol{r}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{r}_i,$$

jossa $n = \sum_{i=1}^{N} n_i$

Huomioitavaa on, että ML- ja REML-menetelmien log-uskottavuusfunktiot eroavat vain niin kutsutun sakkotermin $\sum_{i=1}^{N} \log(\det(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1} \boldsymbol{X}_i))$ perusteella. Siten käytännössä samat optimointialgoritmit ovat sovellettavissa molempien numeeriseen estimointiin.

Diggle *ym.* (2013) tarkentavat, että ML- ja REML-menetelmät ovat asymptoottisesti ekvivalentteja ja siten yleisiä teoreettisia eroja on menetelmistä vaikea osoittaa, mutta sovelluksissa REML-menetelmän on osoitettu suoriutuvan tehokkaammin tilanteissa, joissa esimerkiksi käännettävä kovarianssimatriisi on lähes singulaarinen. Koska sakkotermi on $p \times p$ matriisi, ML- ja REML-menetelmien erot korostuvat vain, jos p on suuri suhteessa aineiston mittausten kokonaismäärään $n = \sum_{i=1}^{N} n_i$

Sijoittamalla estimaattori $\hat{\boldsymbol{V}}_{i_{\text{REML}}}$ yleistetyn PNS-estimaattorin $\hat{\boldsymbol{\beta}}$ kaavaan, saadaan

$$\hat{\boldsymbol{\beta}}_{\text{REML}} = (\sum_{i=1}^{N} \boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_{i_{\text{REML}}}^{-1} \boldsymbol{X}_i)^{-1} \sum_{i=1}^{N} \boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_{i_{\text{REML}}}^{-1} \boldsymbol{y}_i.$$

3.2.1 Satunnaisvaikutusten ennustaminen

Koska satunnaisvaikutukset $\boldsymbol{b}_i$ eivät frekventistisen uskottavuuspäättelyn diskurssissa ole mallin parametreja vaan satunnaismuuttujia, mm. West *ym.* (2014) sekä Pinheiro ja Bates (2000) suosittelevat estimoinnin sijaan puhuttavan satunnaisvaikutusten ennustamisesta.

Toisin kuin kiinteiden vaikutusten tapauksessa, johtuen siitä, että $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, ei kiinnostuksen kohteena ole itse satunnaismuuttujan odotusarvo, vaan ehdollinen odotusarvo $\mathrm{E}(\boldsymbol{b}_i|\boldsymbol{y}_i)$.

Nissinen (2009) näyttää, että $\boldsymbol{b}_i$ ja $\boldsymbol{y}_i$ normaalisuusoletuksesta seuraten, $\boldsymbol{b}_i$ ja $\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$ noudattavat moniulotteista normaalijakaumaa

$$\begin{bmatrix} \boldsymbol{b}_i \\ \boldsymbol{y}_i \end{bmatrix} \sim N_{q+n_i}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{X}_i\boldsymbol{\beta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G} & \boldsymbol{G}\boldsymbol{Z}_i^\top \\ \boldsymbol{Z}_i\boldsymbol{G} & \boldsymbol{V}_i \end{bmatrix} \right),$$

josta moniulotteisen normaalijakauman ehdollisen jakauman määritelmän mukaisesti saadaan ehdolliseksi odotusarvoksi

$$\begin{aligned} \mathrm{E}(\boldsymbol{b}_i|\boldsymbol{y}_i) &= \mathrm{E}(\boldsymbol{b}_i) + \boldsymbol{G}\boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \mathrm{E}(\boldsymbol{y}_i)) \\ &= \boldsymbol{G}\boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}). \end{aligned}$$

Kun tuntematon $\boldsymbol{\beta}$ tilalle sijoitetaan yleistetty PNS-estimaattori $\hat{\boldsymbol{\beta}}$ saadaan

$$\hat{\boldsymbol{b}}_i = \boldsymbol{G}\boldsymbol{Z}_i^\top \boldsymbol{V}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}),$$

joka on vektorin $\boldsymbol{b}_i$ *paras lineaarinen harhaton ennuste (BLUP).*

Harhattomuudella viitataan tässä yhteydessä siihen, että ennusteella $\hat{\boldsymbol{b}}_i$ ja satunnaismuuttujalla $\boldsymbol{b}_i$ on sama odotusarvo.

Nissinen (2009) huomauttaa, että ennusteella on pienempi varianssi kuin satunnaismuuttujalla, sillä

$$\begin{aligned} Var(\boldsymbol{b}_i) &= Var(\mathrm{E}(\boldsymbol{b}_i|\boldsymbol{y}_i)) + \mathrm{E}(Var(\boldsymbol{b}_i|\boldsymbol{y}_i)) \\ &= Var(\hat{\boldsymbol{b}}_i) + c, \end{aligned}$$

jossa $c \geq 0$.

Parasta lineaarista harhatonta ennustetta $\hat{\boldsymbol{b}}_i$ voidaan kutsua myös *kutistusestimaattoriksi* (*shrinkage estimator*) (Robinson, 1991).

Koska usein $\boldsymbol{G}$, $\boldsymbol{V}_i$ sekä $\boldsymbol{R}_i$ ovat tuntemattomia, korvataan ne vastaavilla ML- tai REML-estimaateilla, jonka seurauksena saamme *empiirisen parhaan lineaarisen harhattoman ennusteen (EBLUP)*.

$$\hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}} = \hat{\boldsymbol{G}}\boldsymbol{Z}_i^\top \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}).$$

Nissinen (2009) johtaa ennusteen kovarianssimatriisin määrittelemällä ensin

$$\begin{aligned} \hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}} &= \hat{\boldsymbol{G}}\boldsymbol{Z}_i^\top \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}) \\ &= \hat{\boldsymbol{G}}\boldsymbol{Z}_i^\top \hat{\boldsymbol{P}}\boldsymbol{y}_i, \end{aligned}$$

jossa

$$\hat{\boldsymbol{P}} = \hat{\boldsymbol{V}}_i^{-1}(\boldsymbol{I}_{n_i} - \boldsymbol{X}_i(\boldsymbol{X}_i^\top \boldsymbol{V}_i^{-1}\boldsymbol{X}_i)^{-1}\boldsymbol{X}_i^\top \hat{\boldsymbol{V}}_i^{-1}).$$

Ennusteen kovarianssimatriisiksi saadaan

$$Cov(\hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}}) = \hat{\boldsymbol{G}}\boldsymbol{Z}_i^\top \hat{\boldsymbol{P}}\boldsymbol{Z}_i\hat{\boldsymbol{G}}.$$

Verbeke ja Molenberghs (2000) mukaan $Cov(\hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}})$ on aliarvio $\hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}} - \boldsymbol{b}_i$ vaihtelusta ja usein päättely perustuu

$$Cov(\hat{\boldsymbol{b}}_{i_{\mathrm{EBLUP}}} - \boldsymbol{b}_i).$$

Verbeke ja Molenberghs (2000) kuitenkin korostavat, että edellä mainittu lähestymistapa tulisi tulkita bayesiläisessä viitekehyksessä, sillä $\boldsymbol{b}_i$ tulkitaan satunnaisparametriksi.

Nissinen (2009) sekä Verbeke ja Molenberghs (2000) esittelevät vaihtoehtoiseksi lähestymistavaksi Henderson *ym.* (1959) sekamalliyhtälöt, jossa $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{b}}_i$ estimoidaan samanaikaisesti lineaarisen yhtälöryhmän ratkaisuna, mutta Verbeke ja Molenberghs (2000) tarjoavat laajemman tilastotieteenfilosofisen pohjan päättelylle.

Verbeke ja Molenberghs (2000) mukaillen Hendersonin (1959) sekamalliyhtälöiden voidaan osoittaa olevan yhteensopivia edellisen esitystavan kanssa määrittelemällä lineaarinen sekamalli muodossa

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon},$$

jossa yksilökohtaiset vektorit ja matriisit ovat yhdistetty vektoreiksi $\boldsymbol{Y}$, $\boldsymbol{b}$, lohkomatriisiksi $\boldsymbol{X}$ ja lohkodiagonaalimatriiseiksi $\boldsymbol{G}$ ja $\boldsymbol{R}$ seuraavalla tavalla

$$\begin{aligned}
\boldsymbol{y} &= [\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_N^\top]^\top \\
\boldsymbol{b} &= [\boldsymbol{b}_1^\top, \ldots, \boldsymbol{b}_N^\top]^\top \\
\boldsymbol{X} &= [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_N^\top]^\top \\
\boldsymbol{Z} &= diag(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N) \\
\boldsymbol{G} &= diag(\boldsymbol{G}_1, \ldots, \boldsymbol{G}_N) \\
\boldsymbol{R} &= diag(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_N),
\end{aligned}$$

jolloin

$$\begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{X} & \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^\top \boldsymbol{R}^{-1} \boldsymbol{y} \\ \boldsymbol{Z}^\top \boldsymbol{R}^{-1} \boldsymbol{y} \end{bmatrix}.$$

Sekamalliyhtälöjen ratkaisut ovat $\hat{\boldsymbol{\beta}}$ ja $\hat{\boldsymbol{b}}$, jotka ovat parametrien $\boldsymbol{\beta}$ ja $\boldsymbol{b}$ paras lineaarinen harhaton estimaattori (BLUE) ja ennuste (BLUP).

Laird ja Ware (1982) ovat osoittaneet, että niin ML- kuin REML-estimaatit voidaan laskea numeerisesti EM-algoritmilla, mutta huomauttavat konvergointiongelmista erityisesti ML-estimoinnissa, jos suurin uskottavuus saavutetaan lähellä parametriavaruuden rajaa.

Verbeke ja Molenberghs (2000) viittaavat Newton-Raphson-menetelmään aikansa yleisimpänä parametrien estimointimenetelmänä. Tarkan kuvauksen näiden algoritmien implementointiin antavat Lindstrom ja Bates (1988) sekä Lindstrom ja Bates (1990), joista jälkimmäiseen teokseen mm. R-ohjelmiston (R Core Team, 2013) **nlme**-paketin dokumentaatiossa viitataan (Pinheiro *ym.*, 2013). Keskustelemme lyhyesti optimointialgoritmien valinnasta liitteessä B.

3.3 Mallidiagnostiikka

Mallin valinnassa ja arvioinnissa keskeistä on valita sopiva malli, joka vastaa parhaiten esitettyyn tutkimuskysymykseen. Haasteena on useiden kilpailevien mallien määrä, josta tutkijan tulisi pystyä valitsemaan malli, joka samaan aikaan parhaiten selittää vastemuuttujan vaihtelua, mutta mahdollistaa haluttujen tutkimushypoteesien testaamisen. (West *ym.*, 2014).

Keskeisiä työkaluja sopivan mallin etsintään ovat mallin residuaalijakauman ja estimoitujen parametrien tarkastelu sekä mallien keskinäisen vertailun mahdollistavat kriteerit.

Uskottavuusosamäärän testi on hyvin tunnettu menetelmä kahden mallin vertailuun. Menetelmässä vertaillaan kahden sisäkkäisen (*nested*) mallin, rajoittamattoman vertailumallin ja rajoitetun nollahypoteesimallin uskottavuusfunktioita.

Tarkemmin muotoiltuna, rajoitetun mallin parametriavaruuden tulee olla rajoittamattoman mallin parametriavaruuden osajoukko (West *ym.*, 2014).

Uskottavuusosamäärän testi (*Likelihood ratio test, LRT*) määritellään

$$-2\log(\frac{L_0}{L}),$$

jossa L on rajoittamattoman mallin ja L_0 rajoitetun mallin uskottavuus.

Uskottavuusosamäärän testin testisuure noudattaa asymptoottisesti χ^2-jakaumaa vapausastein $p - p_0$, p ja p_0 ollessa rajoittamattoman ja rajoitetun mallin parametrien lukumäärät.

Uskottavuusosamäärän testi soveltuu lineaaristen sekamallien vertailuun sekä ML- että REML-menetelmillä, mutta jälkimmäisessä tapauksessa molempien mallien tulee olla estimoitu REML-menetelmällä ja kiinteiden vaikutusten parametrien identtisiä. (Pinheiro ja Bates, 2000).

Kiinteiden vaikutusten parametrien mukaan rajoitetun ja rajoittamattoman mallin vertailuun tulee siis perustua ainoastaan ML-menetelmään. Vertaillessa malleja kiinteiden vaikutusten parametrien perusteella, tulee vuorostaan kovarianssiparametrien olla samoja rajoitetussa ja rajoittamattomassa mallissa. (West *ym.*, 2014).

Kovarianssiparametrien perusteella rajoitetun ja rajoittamattoman mallin vertailun tulisi perustua vuorostaan REML-menetelmään, sillä kovarianssiparametrien REML-estimaattien harhan on osoitettu olevan ML-estimaatteja pienempää. (West *ym.*, 2014).

Toisen mallien vertailun kannalta hyödyllisen menetelmäjoukon muodostavat informaatiokriteerit. Informaatiokriteerit arvioivat mallin ja aineiston yhteensopivuutta optimoidun log-uskottavuusfunktion perusteella lisäten arvioon parametrien lukumäärään perustuvan sakkotermin.

Informaatiokriteerien keskeinen ominaisuus on, että niillä voidaan vertailla kaikkia samaan aineistoon sovitettuja malleja. (West *ym.*, 2014).

West *ym.* (2014) mukaan Akaiken informaatiokriteeri (AIC) voidaan laskea ML- ja REML-menetelmillä optimoiduille log-uskottavuusfunktioille $l(\boldsymbol{\beta}, \boldsymbol{\theta})$

$$\text{AIC} = -2l(\boldsymbol{\beta}, \boldsymbol{\theta}) + 2p,$$

jossa p on mallin yhteenlaskettu parametrien määrä. West *ym.* (2014) huomauttavat, että tämä pitää paikkaansa R- ja Stata-ohjelmistoilla, mutta SAS- ja SPSS-ohjelmistot huomioivat p:ssä kovarianssiparametrien lukumäärän.

Toinen yleinen, usein AIC:n kanssa esiintyvä informaatiokriteeri on Bayesin informaatiokriteeri (BIC)

$$\text{BIC} = -2l(\boldsymbol{\beta}, \boldsymbol{\theta}) + p\log(n).$$

BIC soveltaa suurempaa sakkotermiä parametrien kokonaislukumäärään p nähden kertomalla parametrien lukumäärän aineiston havaintomäärän $n = \sum_{i}^{N} n_i$ luonnollisella logaritmilla.

West *ym.* (2014) argumentoivat, ettei informaatiokriteereistä tähänastisen lineaaristen sekamallien kirjallisuuden valossa kumpikaan nouse toisen yläpuolelle ja lisää tutkimusta informaatiokriteerien merkityksestä lineaaristen sekamallien valinnassa kaivataan.

Myös Verbeke ja Molenberghs (2000) ilmaisevat voimakkaan kannan, että informaatiokriteerit eivät tarjoa muuta kuin *peukalosääntöjä* sopivan mallin valintaan, eikä niihin tule suhtautua formaalisti määriteltyinä tilastollisina testeinä.

Mallin residuaalien tarkastelu tarjoaa tärkeän näkökulman mallin sopivuuden arviointiin. Yksilön i ehdollinen residuaalivektori määritellään

$$\hat{\boldsymbol{\epsilon}}_i = \boldsymbol{y}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}} - \boldsymbol{Z}_i\hat{\boldsymbol{b}}_i.$$

Niin kutsutut *raa'at* residuaalit eivät sovellu hyvin mallin arviointiin, sillä ne ovat,

huolimatta todellisista residuaaleista, usein keskenään korreloituneita ja heteroskedastisia. (West *ym.*, 2014).

Vaihtoehtoisia ratkaisuja ongelmaan ovat skaalatut residuaalit, mm. standardisoidut, Studentin sisäiset residuaalit. Burzykowski ja Galecki (2013) mukaan standardisoidut residuaalit saadaan skaalaamalla estimoidut residuaalit vastemuuttujan todellisella keskihajonnalla, mutta käytännössä todellinen keskihajonta on harvoin tunnettu. Siten, estimoidut residuaalit voidaan skaalata myös estimoidulla keskihajonnalla, jolloin saadaan Studentin sisäiset residuaalit (*internally studentized residuals*)

$$\frac{\hat{\boldsymbol{\epsilon}}_i}{\sqrt{\widehat{Var}(\boldsymbol{y})_i}}.$$

Burzykowski ja Galecki (2013) suosittelevat Studentin sisäisiä residuaaleja, sillä ne vähentävät heteroskedastisuutta poistamatta sitä kokonaan. Myös residuaalien välinen korrelaatio säilytetään. Burzykowski ja Galecki (2013) huomauttavat, että R-ohjelmistossa Studentin residuaaleja kutsutaan yleisesti Pearsonin residuaaleksi, joten otamme myös käyttöön kyseisen nimeämiskäytännön.

4 Tutkielman rajatun FinLapset-aineiston analyysi lineaarisilla sekamalleilla

Tavoitteena on, edelliset luvut yhdistäen, löytää sopiva malli FinLapset-aineiston analysointiin. Pyrimme ensin sovittamaan aiemmin FinLapset-aineistosta tunnistetut pitkittäisaineiston piirteet lineaaristen sekamallien kehikkoon.

Seuraavassa vaiheessa sovellamme West *ym.* (2014) sekä Verbeke ja Molenberghs (2000) mukaillen ylhäältä alas -strategiaa sopivan mallin etsimiseksi. Tätä seuraa mallin kriitinen arviointi ja dialogi mallia puoltavien ja sitä vastustavien argumenttien kesken.

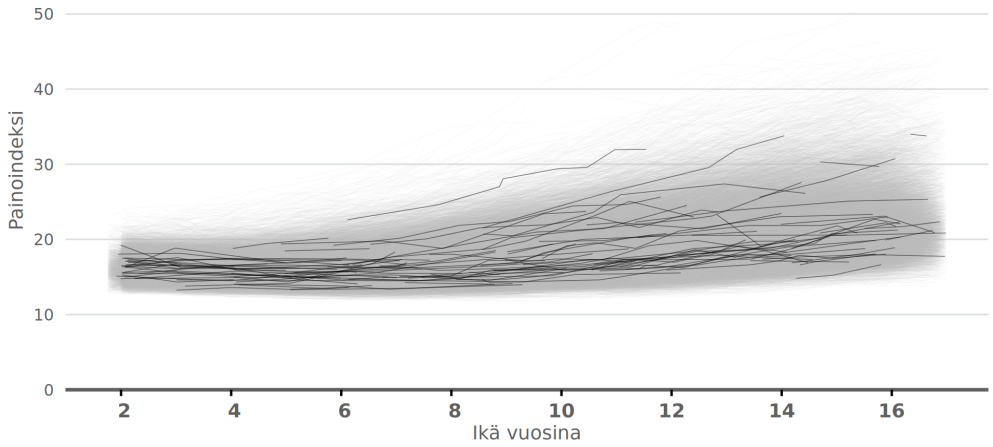
Tästä johdamme, ja lopuksi kiteytämme, keskeiset löydökset ja peilaamme näitä mallin tunnistettuihin puutteisiin sekä avoimeksi jääneisiin kysymyksiin.

4.1 Operationalisointi

Tässä vaiheessa suuntaamme kiinnostuksemme siihen, minkälainen yhteys painoindeksin ajallisella muutoksella on yksilön ikään ja sukupuoleen. Näiden yhteyksien tutkimiseksi pyrimme määrittelemään millä tavoin tutkielman aineiston muuttujat kuvaavat vastaavia reaalimaailman ilmiöitään.

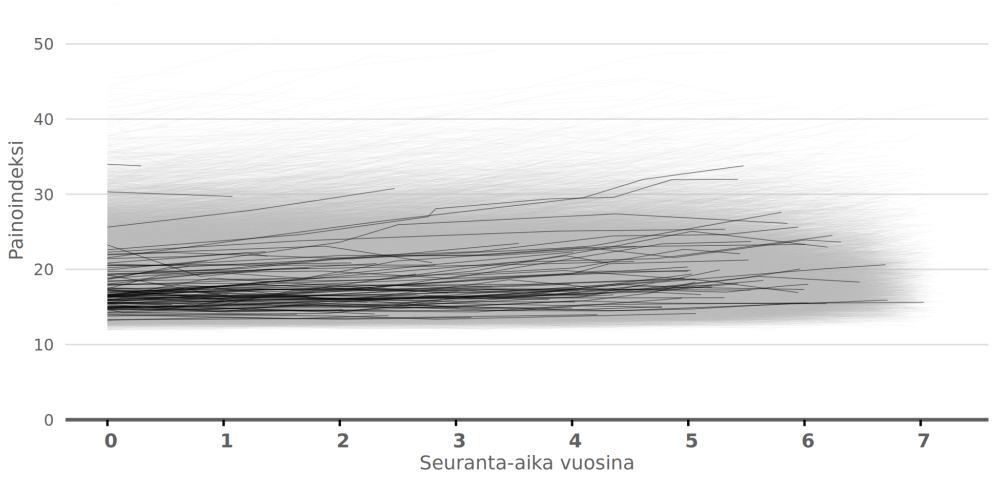
Kuten jo alustavassa aineiston tarkastelussa huomioimme, aikadimensio vaatii erityistä huomiota. Fitzmaurice *ym.* (2011) varoittavat, että iän käyttäminen ajan *metamittarina* tuo malliin kaksi informaation lähdettä.

Ensimmäinen informaation lähde on yksilöiden välinen poikkileikkausvaikutus, joka seuraa siitä, että yksilöt voidaan mitata eri ikäisinä. Aineiston tarkastelu tästä näkökulmasta paljastaa myös mahdollisen huolen painoindeksin ja iän suhteen lineaarisuudesta (Kuva 5).



Kuva 5: Yksilökohtaiset painoindeksiprofiilit iän mukaan (harmaa). Yksittäisiä painoindeksiprofiileja korostettu mustalla.

Toinen lähde on yksilökohtainen pitkittäisvaikutus, jonka taustalla on yksilöiden vanheneminen seurannan aikana. Tästä näkökulmasta voimme vakioida ajan alkamaan ensimmäisestä mittauksesta (Kuva 6). Vaikka, lapsen koko kehityskaari huomioiden, painoindeksin kehitys voi hyvin olla epälineaarista, havaituilla lyhyillä seurantajaksoilla voimme pitää likimain kiinni lineaarisuusoletuksesta.



Kuva 6: Yksilökohtaiset painoindeksiprofiilit seuranta-ajan mukaan (harmaa). Yksittäisiä painoindeksiprofiileja korostettu mustalla.

Iän mukana malliin sisällytettyjen, mahdollisesti ristiriitaisten informaation lähteiden erottamiseksi Fitzmaurice *ym.* (2011) suosittelevat ajan operationalisoimista kahdeksi eri muuttujaksi, joista yksilön ikä ensimmäisellä mittauksella IKÄ_{0i} (*baseline age*) tuo malliin iän poikkileikkausvaikutuksen ja ikä mittaushetkellä IKÄ_{ij} pitkittäisvaikutuksen. Palaamme ajan operationalisointiin mallinvalintaluvun kappaleessa 4.2.1).

4.2 Mallinvalinta

West *ym.* (2014) sekä Verbeke ja Molenberghs (2000) lineaarisille sekamalleille esittämässään mallinvalintastrategiassa tavoitteena on ensin muodostaa kiinteitä vaikutuksia lisäämäällä *yliparametrisoitu* malli, jolla voidaan selittää mahdollisimman hyvin aineistossa vallitsevaa keskimääräistä vastetta.

Toisena askeleena on löytää mallille sopiva satunnaisvaikutusten rakenne, jolla pyritään huomioimaan aineiston hierarkkinen rakenne. Tutkielman aineiston tapauksessa hierarkia syntyy ensisijaisesti satunnaisvaihtelun sallimisessa yksilökohtaisten vakiotermien ja regressiokertoimien välillä.

Kolmannessa vaiheessa tutkitaan satunnaisvaikutusten lisäämisen jälkeen selittämättä jäänyttä vaihtelua. Huomio kohdistuu erityisesti varianssin heteroskedastisuuteen sekä yksilön mittausten väliseen autokorrelaatioon.
Viimeinen vaihe käsittää mallin parametrien kriittisen tarkastelun, varsinkin kiin-

teiden vaikutusten osalta ja lopullisen mallin arvioinnin.

Aloitamme kuitenkin tarkastelemalla painoindeksin ja iän yhteyttä, sekä lisäksi mahdollista biologisen sukupuolen vaikutusta ja iän sukupuolen yhdysvaikutusta painoindeksiin naiivissa asetelmassa lineaarisen regressiomallin avulla, sivuuttaen hetkeksi pitkittäisaineiston hierarkkisen luonteen.

Alustavien havaintojen (Taulukko 4) mukaan keskimääräinen painoindeksi kasvaa huomattavasti iän mukana, mutta sukupuolen vaikutus ilmenee monisyisempänä.

Sukupuoli	Ikäluokka	Keskiarvo	n	Keskihajonta	Minimi	Maksimi
Poika	$[0, 2)$	16,37	3113	1,320	12,92	23,47
	$[2, 7)$	15,98	80952	1,449	12,27	29,14
	$[7, 13)$	17,81	85244	3,186	12,12	48,81
	$[13, 17)$	20,73	43291	4,012	12,74	55,58
Tyttö	$[0, 2)$	16,11	2980	1,401	12,71	23,23
	$[2, 7)$	15,90	79334	1,549	11,93	30,61
	$[7, 13)$	17,73	84368	3,131	11,77	40,70
	$[13, 17)$	21,09	42378	3,729	12,77	45,55

Taulukko 4: Painoindeksijakauman tunnuslukuja luokiteltuna biologisen sukupuolen ja ikäluokan perusteella.

Poikien keskimääräinen painoindeksi on kolmessa ensimmäisessä ikäluokassa suurempi, mutta tyttöjen painoindeksi nousee viimeisessä ikäluokassa poikien ohi. Tytöillä painoindeksin keskihajonta on poikia suurempaa kahdessa ensimmäisessä ikäluokassa, kun taas pojilla vaikuttaa olevan huomattavasti suurempi keskihajonta viimeisessä ikäluokassa. Tosin, tutkielman alussa tehdyssä jakaumatarkakastelussa havaitsimme jakaumassa vinoutta ja jakauman oikeassa hännässä merkittävästi poikkeavilla, mahdollisesti virheellisillä havainnoilla voi olla suuri vaikutus.

Mallin kannalta niin sukupuolen, kuin iän ja sukupuolen yhdysvaikutuksen tarkastelulle voi olla perusteita.

Muodostetaan ensin yksinkertainen lineaarinen regressiomalli havainnolle $i = 1, \ldots, n$, jossa $n = \sum_{i=1}^{N} n_i$ on aineiston havaintojen kokonaismäärä

$$\text{BMI}_i = \beta_0 + \beta_1 \times \text{IKÄ}_i + \epsilon_i \tag{2}$$

Regressiokertoimen β_0 estimaatiksi (Taulukko 5) saamme likimain 13,94. Tämä niin kutsuttu vakiotermi voidaan tulkita keskimääräiseksi painoindeksiksi, kun muuttu-

ja IKÄ_i saa arvon 0. Itse estimaatin arvo ei siis itsessään ole kovin mielekäs, sillä aineston havainnot alkavan noin kahdesta ikävuodesta. Tarvittaessa tilanne voitaisi korjata skaalaamalla ikämuuttujaa siirtäen vakiotermin estimaatin vastaamaan haluttua ikää.

	Estimaatti	Keskivirhe (se)	P	$\text{LV}_{2,5}$	$\text{LV}_{97,5}$
$\hat{\beta}_0$	13,939	0,0098	0,0000	13,920	13,958
$\hat{\beta}_1$	0,4310	0,0010	0,0000	0,4291	0,4330

Taulukko 5: Yksinkertaisen lineaarisen regressiomallin (LM 2) parametrien estimaatit, keskivirheet, tilastollinen merkitsevyys ja luottamusväli.

Regressiokertoimen β_1 estimaatti 0,4310, voidaan tulkita puolestaan PNS-menetelmällä sovitetun regressiosuoran kulmakertoimeksi. Toisin sanoen jokaista ikävuoden yksikkömuutosta kohden keskimääräinen painoindeksi kasvaa likimain 0,43 yksikköä.

Huomioitavaa on, että yksinkertaisen regressiomallin parametrien keskivirheen määritelmästä

$$se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(\frac{1}{n}\sum_{i=1}^{n} \text{IKÄ}_i)^2}{\sum_{i=1}^{n} (\text{IKÄ}_i - \frac{1}{n}\sum_{i=1}^{n} \text{IKÄ}_i)^2} \right)}$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_i (\text{IKÄ}_i - \frac{1}{n}\sum_{i=1}^{n} \text{IKÄ}_i)^2}},$$

jossa $\hat{\sigma}^2 = \frac{1}{n-2}\sum_i \hat{\epsilon}_i^2$,

seuraa, että havaintomäärän n ollessa hyvin suuri keskivirheet jäävät aina pieniksi. Teknisessä mielessä voimme siis luottaa, että estimaattorimme tuottamat estimaatit ovat täsmällisiä, mutta ne voivat silti olla hyvinkin harhaisia.

Tutkitaan seuraavaksi sukupuolen vaikutusta lisäämällä malliin muuttuja SUKUPUOLI_i

$$\text{BMI}_i = \beta_0 + \beta_1 \times \text{IKÄ}_i + \beta_2 \times \text{SUKUPUOLI}_i + \epsilon_i \tag{3}$$

Vaikuttaisi siltä, että sukupuolella ei ole merkitsevää vaikutusta keskimääräiseen painoindeksiin (Taulukko 6).

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\beta}_0$	13,940	0,0107	0,0000	13,919	13,961
$\hat{\beta}_1$	0,4310	0,0010	0,0000	0,4291	0,4330
$\hat{\beta}_2$	-0,0025	0,0086	0,7761	-0,0194	0,0145

Taulukko 6: Lineaarisen regressiomallin (LM 3) parametrien estimaatit ja keskivirheet kun malliin on lisätty biologisen sukupuolen vaikutus.

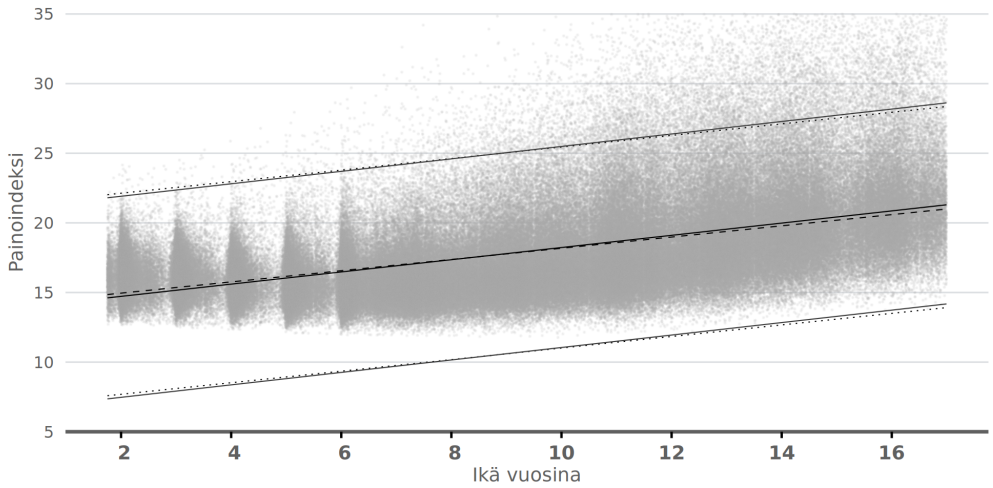
Tarkastellaan kolmatta mallia, jossa lisäämme iän ja sukupuolen yhdysvaikutuksen.

$$\text{BMI}_i = \beta_0 + \beta_1 \times \text{IKÄ}_i + \beta_2 \times \text{SUKUPUOLI}_i + \beta_3 \times \text{IKÄ}_i \times \text{SUKUPUOLI}_i + \epsilon_i \quad (4)$$

Nyt sekä sukupuoli, että iän ja sukupuolen yhdysvaikutukset ovat merkitseviä. Mallin tuloksista (Taulukko 7) voimme nähdä, että iän estimaatti on hieman edellistä pienempi. Nyt iän estimaatti viittaa päävaikutukseen, kun sukupuolena on poika, kun taas iän ja sukupuolen yhdysvaikutus viittaa muutokseen iän päävaikutuksesta, kun sukupuolena tyttö.

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\beta}_0$	14,079	0,0138	0,0000	14,052	14,1060
$\hat{\beta}_1$	0,4151	0,0014	0,0000	0,4123	0,4179
$\hat{\beta}_2$	-0,2831	0,0196	0,0000	-0,3215	-0,2447
$\hat{\beta}_3$	0,0322	0,0020	0,0000	0,0283	0,0362

Taulukko 7: Lineaarisen regressiomallin (LM 4) parametrien estimaatit, keskivirheet, tilastollinen merkitsevyys ja luottamusväli. kun malliin on lisätty biologisen sukupuolen ja iän yhdysvaikutus.



Kuva 7: Riippumattomiin mittauksiin sovitettu lineaarinen malli iän ja sukupuolen yhdysvaikutus huomioituna ennusteväleineen. Tytöt yhtenäisillä viivoilla, pojat katkoviivalla.

Iän ja sukupuolen yhdysvaikutus vaikuttaa kuitenkin melko pieneltä ja sovitteen sekä aineiston visuaalinen tarkastelu (Kuva 7) viittaa samaan. Lisäksi on selvää, että kyseinen malli on hyvin puuttellinen tämän kaltaiseen aineistoon. Jo visuaalinen tarkastelu paljastaa, että naiivissa asetelmassa mallin oletukset lineaarisuudesta, residuaalien normaalisuudesta ja varianssin homoskedastisuudesta eivät täyty.

On kuitenkin muistettava, että tavoitteenamme on muodostaa yliparametrisoitu kiinteiden vaikutusten malli ja teemme varsinaisen kriittisen mallinvalinnan myöhemmässä vaiheessa.

Edellä sovitettujen mallien vertailu (Taulukko 8) osoittaa, että pelkän sukupuolen lisäämisellä malliin ei ole suotuisaa vaikutusta, mutta iän ja sukupuolen yhdysvaikutus parantaa mallin yhteensopivuutta maltillisesti.

Malli	df	$l(\boldsymbol{\beta})$	AIC	BIC
LM 2	3	-1033025	2066056	2066089
LM 3	4	-1033025	2066058	2066102
LM 4	5	-1032898	2065806	2065861

Taulukko 8: Lineaaristen mallien vapausasteiden, log-uskottavuuksien ja informaatiokriteerien vertailu.

Bayesin informaatiokriteeri, erityisesti mallien LM 2 ja LM 4 välillä, näyttää, että

parametrien ja havaintojen määrä huomioiden malli on vain hieman paremmin sopusoinnussa aineiston kanssa.

Yliparametrisoidun kiinteiden vaikutusten mallin tarpeet täyttääksemme, voimme silti ottaa lineaarisen mallin LM 4.

4.2.1 Yliparametrisoitu kiinteiden vaikutusten malli

Siirrämme lähestymistavan nyt pitkittäisaineiston kehikkoon. Muotoilemme kiinteiden vaikutusten mallin, noudattaen Fitzmaurice *ym.* (2011) suositusta ajan operationalisoinnista, seuraavalla tavalla

$$\begin{aligned} \mathrm{BMI}_{ij} = \beta_0 + \beta_1 \times \mathrm{IKÄ}_{0i} + \beta_2 \times \mathrm{IKÄ}_{ij} + \beta_3 \times \mathrm{SUKUPUOLI}_{ij} \\ + \beta_4 \times \mathrm{IKÄ}_{ij} \times \mathrm{SUKUPUOLI}_{ij} + \mathrm{b}_{0i} + \epsilon_{ij} \end{aligned} \tag{5}$$

West *ym.* (2014) kuvaavat kiinteiden vaikutusten mallia keskiarvorakenteeksi (*mean structure*), sillä se pyrkii kuvaamaan lineaarisen regression tavoin keskimääräistä yhteyttä vastemuuttujan ja taustamuuttujien välillä.

Kiinteiden vaikutusten ML-estimaateilla (Taulukko 9) onkin vastaava tulkinta kuin lineaarisen regressiomallin (Taulukko 4) estimaateilla. Huomaamme, että estimaatit vastaavat likimain toisiaan, joskin aikadimension uudelleenoperationalisointi on lisännyt malliin uuden parametrin.

	Estimaatti	Keskivirhe (se)	P	$\mathrm{LV}_{2,5}$	$\mathrm{LV}_{97,5}$
$\hat{\beta}_0$	14,124	0,0189	0,0000	14,087	14,161
$\hat{\beta}_1$	0,0194	0,0022	0,0000	0,0152	0,0236
$\hat{\beta}_2$	0,4044	0,0015	0,0000	0,4016	0,4073
$\hat{\beta}_3$	-0,3285	0,0231	0,0000	-0,3737	-0,2833
$\hat{\beta}_4$	0,0404	0,0019	0,0000	0,0367	0,0442

Taulukko 9: Keskiarvorakenteen kiinteiden vaikustusten ML-estimaatit, keskivirheet, tilastollinen merkitsevyys ja luottamusväli.

Keskeinen ero on kuitenkin yksilöiden välinen satunnaisvaikutus b_{0i}, joka kiinteiden vaikutusten mallissa kuvaa yksilöiden vakiotermin vaihtelua. Koska $\boldsymbol{b}_i \sim N(0, \boldsymbol{G})$, ainoastaan vakiotermin vaihtelun sisältävän mallimme satunnaisvaikutusten kovarianssimatriisi on siten

$$\boldsymbol{G} = \left[\mathrm{Var}(\mathrm{b}_{0i})\right] = \left[\sigma^2_{\mathrm{b}_{0i}}\right].$$

Oletamme vielä toistaiseksi residuaalit $\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{R}_i)$ korreloimattomiksi, jolloin matriisilla $\boldsymbol{R}_i$ on diagonaalirakenne

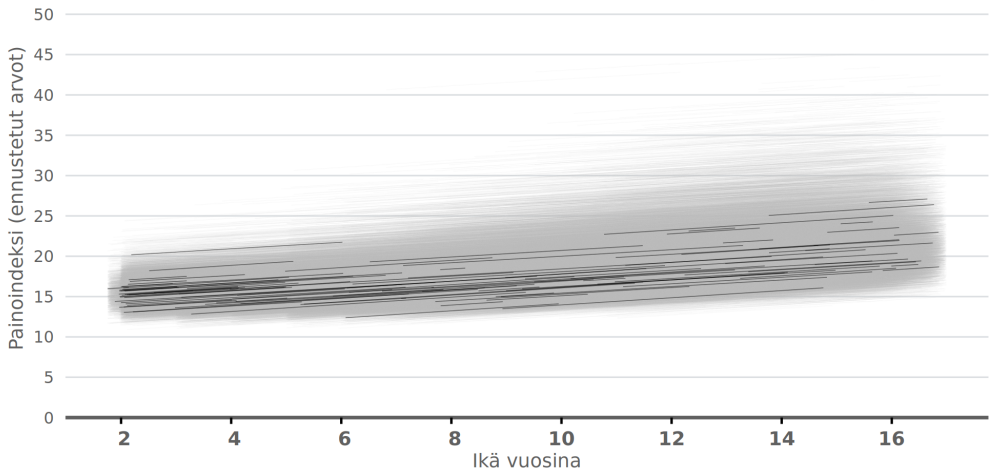
$$\boldsymbol{R}_i = \begin{bmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{bmatrix}.$$

Estimaateista $\hat{\boldsymbol{G}}$ ja $\hat{\boldsymbol{R}}_i$ (Taulukko 10) saamme alustavia, joskaan ei yllättäviä viitteitä, että yksilöiden keskimääräisten painoindeksien varianssi on suurempaa, kuin yksilökohtaisten residuaalien varianssi.

	Estimaatti	Keskivirhe (se)	P	$\text{LV}_{2,5}$	$\text{LV}_{97,5}$
$\hat{\sigma^2}_{\text{b}0i}$	6,224	2,495	0	6,167*	6,280*
$\hat{\sigma^2}$	1,382	1,176	0	1,376*	1,387*

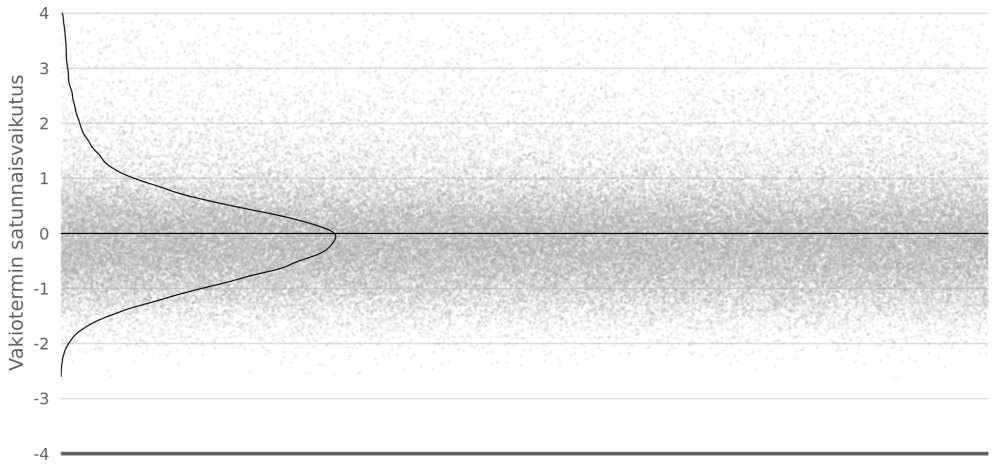
Taulukko 10: Keskiarvorakenteen satunnaisvaikutusten ja residuaalien varianssi-kovarianssiestimaatit. *Luottamusvälit arvioitu normaaliapproksimaatiolla (Pinheiro ja Bates, 2000).

Kiinteiden vaikutusten mallin lähtöaineistolle ennustetuista arvoista (Kuva 8) vahvistamme vakiotermin satunnaisvaikutusten intuitiivisen tulkinnan. Yksilöillä on sama *keskiarvorakenne*, mutta vakiotermit vaihtelevat.

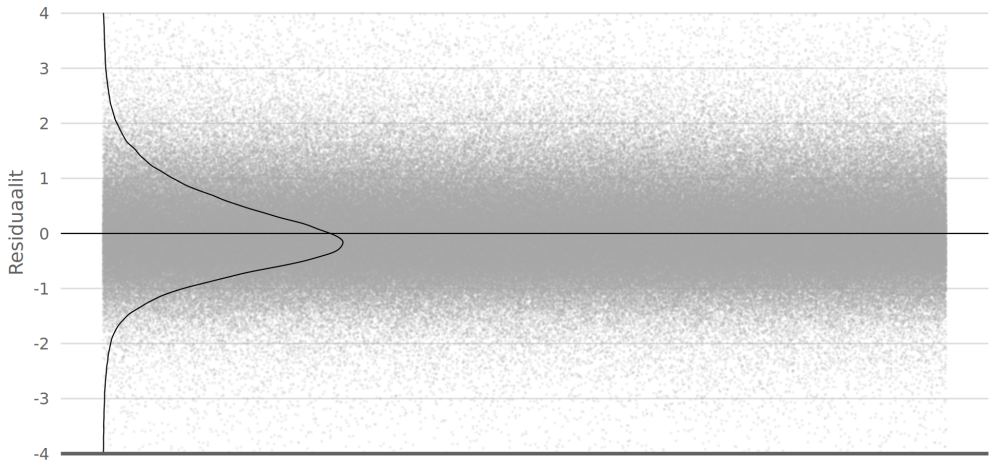


Kuva 8: Kiinteiden vaikutusten mallin lähtöaineistolle ennustetut painoindeksin arvot. Yksittäisiä ennusteita korostettu mustalla.

Vakiotermin satunnaisvaikutusten b_{0i} (Kuva 9) ja yksilökohtaisten residuaalien $\boldsymbol{\epsilon}_i$ (Kuva 10) jakaumien tarkastelu paljastavat ongelmia molempien jakaumaoletuksissa. b_{0i} jakaumassa on havaittavissa vinoutta ja $\boldsymbol{\epsilon}_i$ jakaumassa ilmenee alaspäin suuntautuvaa harhaa.



Kuva 9: Kiinteiden vaikutusten mallin vakiotermin ennustetut satunnaisvaikutukset ja tiheysfunktion Gaussin ydinestimaatti.



Kuva 10: Kiinteiden vaikutusten mallin yksilökohtaisten mittausten Pearsonin residuaalit ja tiheysfunktion Gaussin ydinestimaatti.

Kiinteiden vaikutusten malli ei sellaisenaan vaikuta riittävältä tutkielman aineiston analyysivälineeksi. Tarkastelemme seuraavaksi, kuinka satunnaisvaikutusten rakennetta tarkentamalla voisimme parantaa mallin yhteensopivuutta.

4.2.2 Satunnaisvaikutusten rakenteen valinta

Aineiston tarkastelun perustella (Kuva 5) on uskottavaa ajatella, että painoindeksin ja iän suhteessa, keskimääräisten erojen lisäksi myös kehityksessä iän suhteen olisi yksilöiden välillä eroa.

Tutkimme seuraavaksi millainen seuraus iän satunnaisvaikutuksen lisäämisellä on malliin. Kirjoitetaan malli yksilön i painoindeksihavainnolle j

$$\begin{aligned} \mathrm{BMI}_{ij} = \beta_0 + \beta_1 \times \mathrm{IKÄ}_{0i} + \beta_2 \times \mathrm{IKÄ}_{ij} + \beta_3 \times \mathrm{SUKUPUOLI}_{ij} \\ +\beta_4 \times \mathrm{IKÄ}_{ij} \times \mathrm{SUKUPUOLI}_{ij} + \mathrm{b}_{0i} + \mathrm{b}_{1i} \times \mathrm{IKÄ}_{ij} + \epsilon_{ij}, \end{aligned} \tag{6}$$

jossa b_{0i} vakiotermin satunnaisvaikutus ja b_{1i} iän satunnaisvaikutus.

Merkitään satunnaisvaikutusten parametrivektoria

$$\boldsymbol{b}_i = \begin{bmatrix} \mathrm{b}_{0i} \\ \mathrm{b}_{1i} \end{bmatrix},$$

jolloin kovarianssimatriisi $\boldsymbol{G}$ on

$$\boldsymbol{G} = \begin{bmatrix} \mathrm{Var}(\mathrm{b}_{0i}) & \mathrm{Cov}(\mathrm{b}_{0i}, \mathrm{b}_{1i}) \\ \mathrm{Cov}(\mathrm{b}_{0i}, \mathrm{b}_{1i}) & \mathrm{Var}(\mathrm{b}_{1i}) \end{bmatrix} = \begin{bmatrix} \sigma^2_{\mathrm{b}_{0i}} & \sigma_{\mathrm{b}_{0i},\mathrm{b}_{1i}} \\ \sigma_{\mathrm{b}_{0i},\mathrm{b}_{1i}} & \sigma^2_{\mathrm{b}_{1i}} \end{bmatrix}.$$

Matriisille $\boldsymbol{R}_i$ oletamme edelleen korreloimattoman diagonaalirakenteen.

Satunnaisvaikutuksia vertallaksemme sovitamme mallin REML-menetelmällä. Koska kiinteiden vaikutusten ja satunnaisvaikutusten mallien kiinteät vaikutukset ovat samoin määritelty, voidaan aiemmin ML-menetelmällä estimoituja kiinteitä vaikutuksia (Taulukko 9) kuitenkin verrata satunnaisvaikutusten REML-estimaatteihin (Taulukko 11). (West *ym.*, 2014).

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\beta}_0$	14,616	0,0175	0,0000	14,582	14,651
$\hat{\beta}_1$	-0,1722	0,0020	0,0000	-0,1762	-0,1682
$\hat{\beta}_2$	0,3832	0,0024	0,0000	0,3784	0,3879
$\hat{\beta}_3$	-0,3614	0,0225	0,0000	-0,4055	-0,3173
$\hat{\beta}_4$	0,0432	0,0034	0,0000	0,0366	0,0497

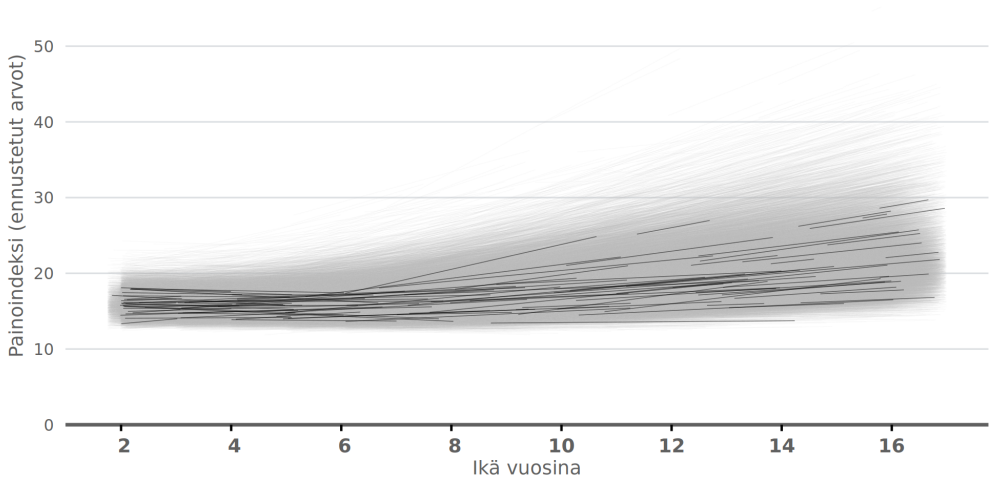
Taulukko 11: Satunnaisvaikutusten mallin LSM 6 kiinteiden vaikustusten REML-estimaatit.

Taulukosta 12 huomaamme vakiotermin varianssin kasvaneen hieman, mutta toisaalta residuaalivarianssi on pienentynyt huomattavasti. Malli vaikuttaa siis selittävän yksilökohtaisten mittausten vaihtelua paremmin.

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\sigma^2}_{b_{0i}}$	8,4467	2,9063	0	8,2955*	8,6007*
$\hat{\sigma^2}_{b_{1i}}$	0,2116	0,4600	0	0,2091*	0,21413*
$\hat{\sigma}_{b_{0i},b_{1i}}$	-1,0566	-	-	-	-
$\hat{\sigma^2}$	0,5926	1,176	0	0,5889*	0,5964*

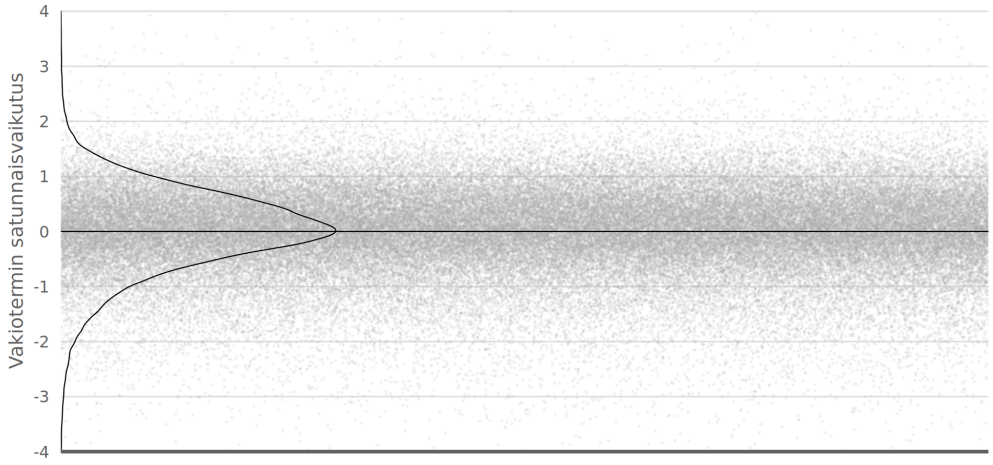
Taulukko 12: Satunnaisvaikutusmallin LSM 6 satunnaisvaikustusten ja residuaalien varianssi-kovarianssiestimaatit. Kovarianssille ei saatavilla keskivirhettä tai luottamusvälejä (-). *Luottamusvälit arvioitu normaaliapproksimaatiolla (Pinheiro ja Bates, 2000).

Aineistolle ennustettujen painoindeksiarvojen tarkastelu (Kuva 11) viittaa samankaltaiseen johtopäätökseen. Yksilöiden mittausjaksot vaikuttavat eroavan toisistaan iästä riippuen. Nuoremmilla lapsilla kehityksen vaihtelu on maltillisempaa, mutta vanhemmilla lapsilla ja nuorilla painoindeksin kehitys vaihtelee iän satunnaisvaikutuksien perusteellaa huomattavasti enemmän.

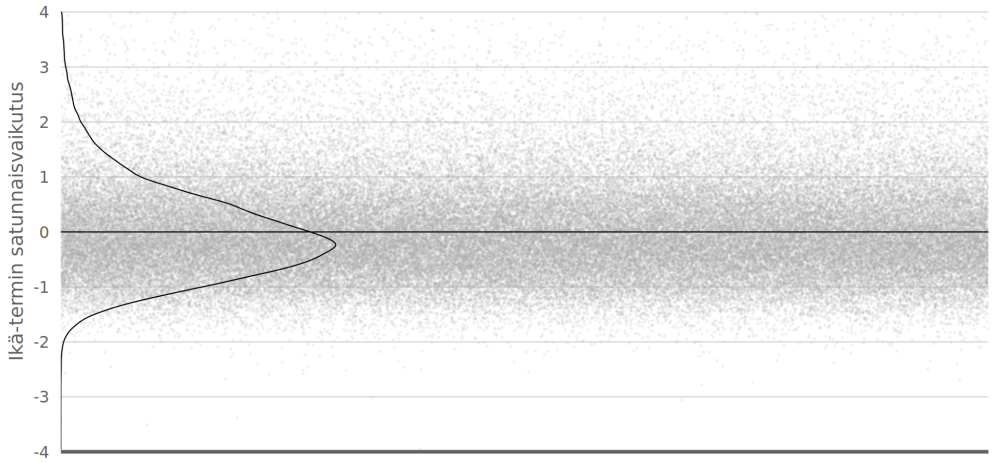


Kuva 11: Satunnaisvaikutusten mallin LSM 6 lähtöaineistolle ennustetut painoindeksin arvot. Yksittäisiä ennusteita korostettu mustalla.

Vakiotermin ennustettujen satunnaisvaikutusten jakauma (Kuva 12) on silmämääräisesti symmetrisempi kuin kiinteiden vaikutusten mallissa, mutta iän ennustetuissa satunnaisvaikutukset (Kuva 13) ovat mahdollisesti harhaisia.

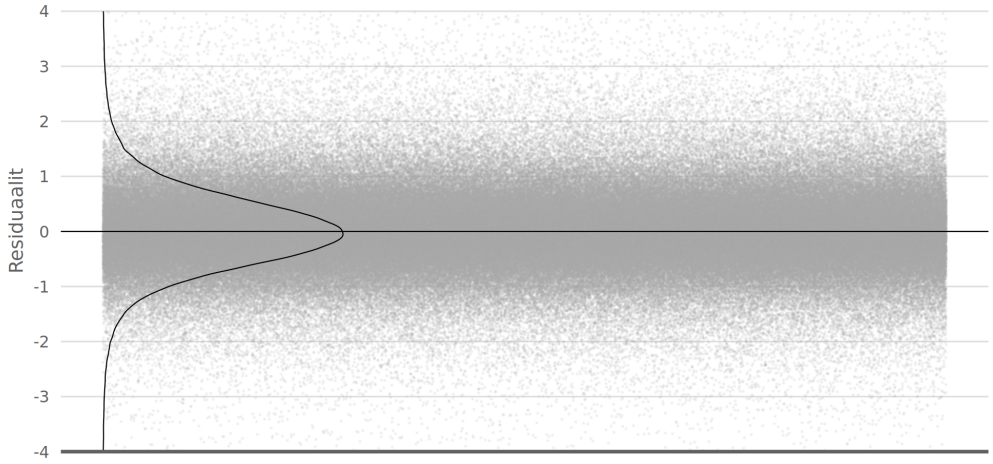


Kuva 12: Satunnaisvaikutusten mallin vakiotermin ennustetut satunnaisvaikutukset ja tiheysfunktion Gaussin ydinestimaatti.



Kuva 13: Satunnaisvaikutusten mallin LSM 6 iän ennustetut satunnaisvaikutukset ja tiheysfunktion Gaussin ydinestimaatti.

Pienemmän varianssin lisäksi residuaalien jakauma (Kuva 14) näyttää vastaavan jakaumaoletuksia huomattavasti edeltävää mallia paremmin, joskin lopullinen arvio vaatii tarkempaa tarkastelua.



Kuva 14: Satunnaisvaikutusten mallin LSM 6 yksilökohtaisten mittausten Pearsonin residuaalit ja tiheysfunktion Gaussin ydinestimaatti.

Satunnaisvaikutusten malli on osoittanut positiivista parannusta kiinteiden vaiku-

tusten malliin verrattuna, joskin joitakin perustavia kysymyksiä mallin oletusten toteutumisesta on vielä jäänyt vastaamatta.

Taulukon 13 Bayesin informaatiokriteerin (BIC) perusteella satunnaisvaikutusten malli (LSM 6), parametrien määrä huomioiden, osoittaa huomattavasti parempaa sopusointua aineiston suhteen verrattuna kiinteiden vaikutusten malliin (LSM 5). Vertailuna taulukossa 13 myös alustavan lineaarisen mallin (LM 4) tuloksia, joihin nähden molemmat lineaaristen sekamallien ryhmään kuuluvat mallit osoittavat merkittävää parannusta.

Malli	*df*	$l(\boldsymbol{\beta})$	AIC	BIC
LM 4	5	-1032898	2065806	2065861
LSM 5	7	-813634	1627282	1627359
LSM 6	9	-729237	1458492	1458591

Taulukko 13: Lineaarisen mallin ja lineaaristen sekamallien vapausasteiden, loguskottavuuksien ja informaatiokriteerien vertailu.

Valitsemme mallin LSM 6 seuraavan vaiheen tarkasteluun, jossa tutkimme tarkemmin residuaalien kovarianssirakennetta.

4.2.3 Residuaalien kovarianssirakenteen valinta

Pinheiro ja Bates (2000) tuovat esiin, että Laird ja Ware (1982) mukainen lineaarisen sekamallin määrittely tuo huomattavaa joustavuutta satunnaisvaikutusten rakenteen määrittelyyn, jos yksilökohtaiset residuaalit ovat korreloimattomia ja niiden varianssi on vakio. Ristiriitaisia tuloksia voi kuitenkin syntyä tapauksissa, joissa yksilön residuaalit ovat korreloituneita, niiden varianssissa esiintyy heteroskedastisuutta tai molempia samanaikaisesti.

Alaluvussa 3.1.4 totesimme, että matriisiin $\boldsymbol{R}_i$ voidaan korreloimattoman diagonaalirakenteen sijaan sisällyttää myös muita rakenteita. Pinheiro ja Bates (2000) mukaan tämä käytännössä vaatii kuitenkin lineaarisen sekamallin laajentamista.

Pinheiro ja Bates (2000) näyttävät, että residuaalien kovarianssimatriisille $\boldsymbol{R}_i$ löytyy hajotelma

$$\boldsymbol{R}_i = \boldsymbol{W}_i \boldsymbol{C}_i \boldsymbol{W}_i,$$

jossa $\boldsymbol{W}_i$ on diagonaalimatriisi, jonka diagonaalialkiot ovat positiivisia ja $\boldsymbol{C}_i$ on positiividefiniitti korrelaatiomatriisi, jonka kaikki diagonaalialkiot ovat ykkösiä. Pinheiro ja Bates (2000) mukaan hajotelma on residuaalien kovarianssimatriisin tarkastelun

kannalta hyödyllinen, sillä sen avulla voimme määritellä yksilön i residuaaleille sekä varianssirakenteen

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2(\boldsymbol{W}_i)^2_{jj}$$

ja korrelaatiorakenteen

$$\mathrm{cor}(\epsilon_{ij}, \epsilon_{ik}) = (\boldsymbol{C}_i)_{jk}.$$

Pinheiro ja Bates (2000) mukaan laajennettuun lineaariseen malliin liittyy kaksi keskeistä huomiota. Ensinnäkin laajennetun mallin estimointi eroaa mm. tämän tutkielman määrittelystä. Toiseksi, koska laajennetun lineaarisen sekamallin vasteen kovarianssimatriisissa

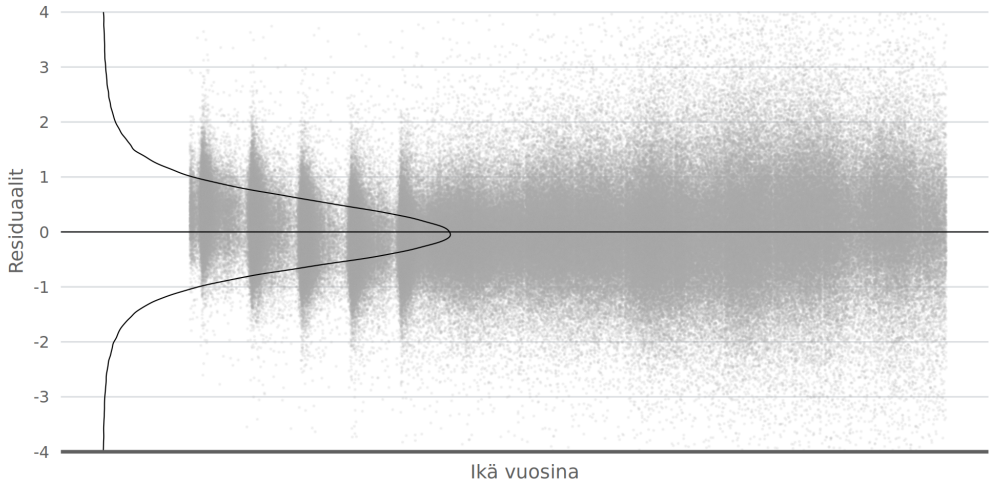
$$\mathrm{Cov}(\boldsymbol{y}_i) = \boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top + \boldsymbol{W}_i\boldsymbol{C}_i\boldsymbol{W}_i,$$

satunnaisvaikutusten komponentti $\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top$ ja yksilökohtaisten residuaalien komponentti $\boldsymbol{W}_i\boldsymbol{C}_i\boldsymbol{W}_i$ kilpailevat osin keskenään, tulee komponenttien määrittely suunnitella huolella.

Esimerkiksi monimutkaisen satunnaisvaikutusten rakenteen $\boldsymbol{Z}_i\boldsymbol{G}\boldsymbol{Z}_i^\top$ yhdistäminen yksinkertaiseen yksilökohtaisen varianssin komponenttirakenteeseen $\boldsymbol{W}_i\boldsymbol{C}_i\boldsymbol{W}_i$ voi Pinheiro ja Bates (2000) mukaan tuottaa hyvin samankaltaisen kovarianssimatriisin $\mathrm{Cov}(\boldsymbol{y}_i)$ kuin päinvastainen yhdistely yksinkertaisen satunnaisvaikutusten rakenteen ja monimutkaisen yksilökohtaisen varianssin komponenttirakenteen välillä.

Näitä huomioita noudattaen, varianssin heteroskedastisuuden ja residuaalien korreloituneisuuden tarkastelu on tärkeä vaihe, mutta suhtaudumme suuntaa antavasti mahdollisiin ehdotuksiin niiden ratkaisemiseksi ja rajaamme Pinheiro ja Bates (2000) tarjoaman laajennetun lineaarisen sekamallin määrittelmän tämän tarkastelun ulkopuolelle.

Vaikka satunnaisvaikutusmallin residuaalien jakauma vaikutti lupaavalta, mikäli tarkastelemme residuaaleja iän suhteen (Kuva 15) paljastuu, ettei oletus varianssin homoskedastisuudesta ole pitävä. Silmämääräisesti varianssi vaikuttaa olevan pienempi nuorempana ja kasvavan iän myötä, tosin aivan pienten lasten kohdalla on havaittavissa myös merkittävää harhaa.



Kuva 15: Satunnaisvaikutusten mallin yksilökohtaisten mittausten Pearsonin residuaalit iän suhteen ja tiheysfunktion Gaussin ydinestimaatti.

Yksilön mittausten välistä autokorrelaatiota voimme Pinheiro ja Bates (2000) mukaan tarkastella autokorrelaatiofunktion avulla. Oletamme, että virhetermit $\epsilon_{ij}, \epsilon_{ik}$ ovat riippuvat sijainneista p_{ij}, p_{ik}, jonkin etäisyyden $d(p_{ij}, p_{ik})$ mukaan. Voimme siten määritellä autokorrelaatiofunktion h

$$\text{cor}(\epsilon_{ij}, \epsilon_{ik}) = h(d(p_{ij}, p_{ik}), \boldsymbol{\rho})),$$

jossa d on jokin etäisyysfunktio ja $\boldsymbol{\rho}$ korrelaatioparametrien vektori.

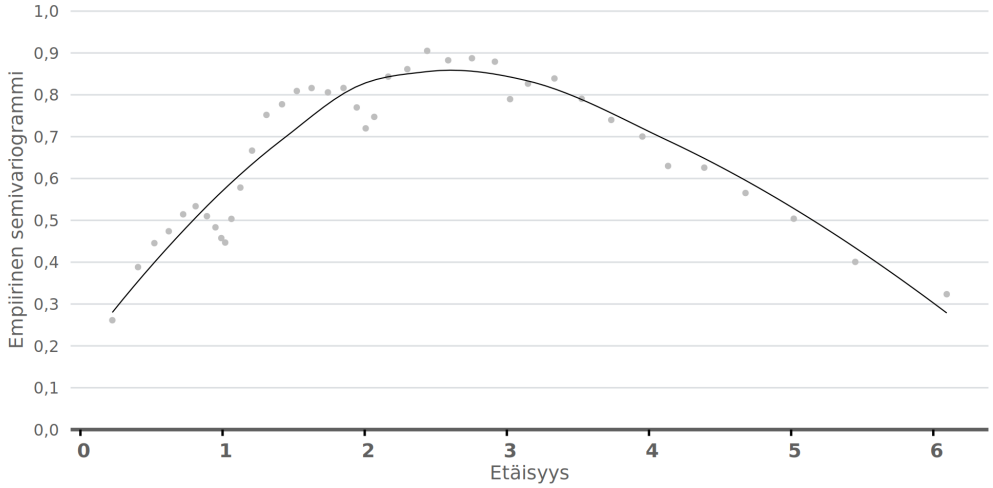
Tarkastellaksemme autokorrelaatiota esimerkiksi iän perusteella, voimme määritellä sijainneiksi p_{ij}, p_{ik} yksilön i iän mittauksilla j ja k, sekä etäisyydeksi d euklidisen etäisyyden, eli yksiulotteisessa tapauksessa erotuksen itseisarvon $|\text{IKÄ}_{ij} - \text{IKÄ}_{ik}|$ mittausten j ja k välillä.

Pinheiro ja Bates (2000) osoittavat, että mittausten samankaltaisuutta perustuen määriteltyyn etäisyyteen s korrelaatioparametreillä $\boldsymbol{\rho}$ voidaan tarkastella semivariogrammilla

$$\gamma(s, \rho) = 1 - h(s, \boldsymbol{\rho}).$$

Semivariogrammin (Kuva 16) pisteet kuvaavat kukin noin 20 000 Pearsonin residuaaliparin perustuvaa samankaltaisten havaitoparien klusteria ja käyrä kuvaa LOESS-tasoitettua trendiä suhteessa residuaaliparin sijaintien etäisyyteen. Matalammat semivariogrammin arvot viittaavat samankaltaisuuteen ja korkeammat arvot erilai-

suuteen.



Kuva 16: Satunnaisvaikutusten mallin empiirinen semivariogrammi ($\hat{\gamma}$).

Tarkastelun perusteella aineistossa ilmenee kohtalaisen voimakasta autokorrelaatiota yksilön mittausten välillä suhteessa mittausten väliseen ajalliseen etäisyyteen. Noin kolmeen vuoteen saakka tarkastelu tukee intuitiota, että lähellä olevat arvot ovat samankaltaisempia.

Kolmen vuoden jälkeen samankaltaisuus vaikuttaisi vastoin odotuksia lisääntyvän, mikä Pinheiro ja Bates (2000) mukaan voi usein johtua estimaattien epäluotettavuudesta pitkillä etäisyyksillä, mutta se voi kertoa myös ongelmista mallin oletusten suhteen.

On selviä viitteitä, että varianssin heteroskedastisuuteen ja residuaalien autokorrelaatioon olisi syytä kiinnittää huomiota. Tarkastelemme molempia ongelmia esimerkkien kautta ja vertaamme lopullisen mallin suhteen mahdollisia ratkaisuja, mutta syvällinen analyysi on rajattu tutkielman ulkopuolelle.

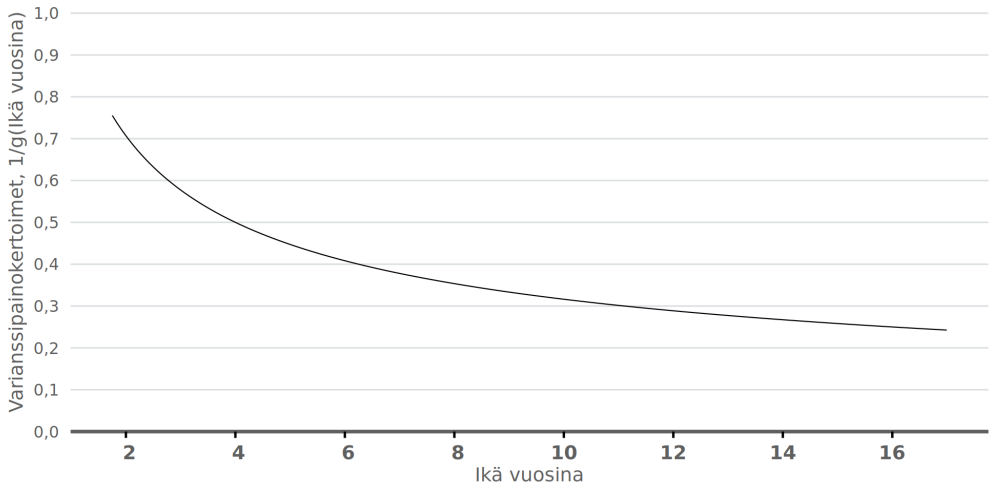
Kuten aiempi tarkastelu antoi viiteitä (Kuva 15), varianssia voi olla perusteltua tarkastella iän funktiona. Mikäli oletamme varianssin iän lineaariseksi funktioksi, voimme Pinheiro ja Bates (2000) mukaillen määritellä varianssiksi

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2 \mathrm{IKÄ}_{ij}$$

ja siten varianssifunktioksi

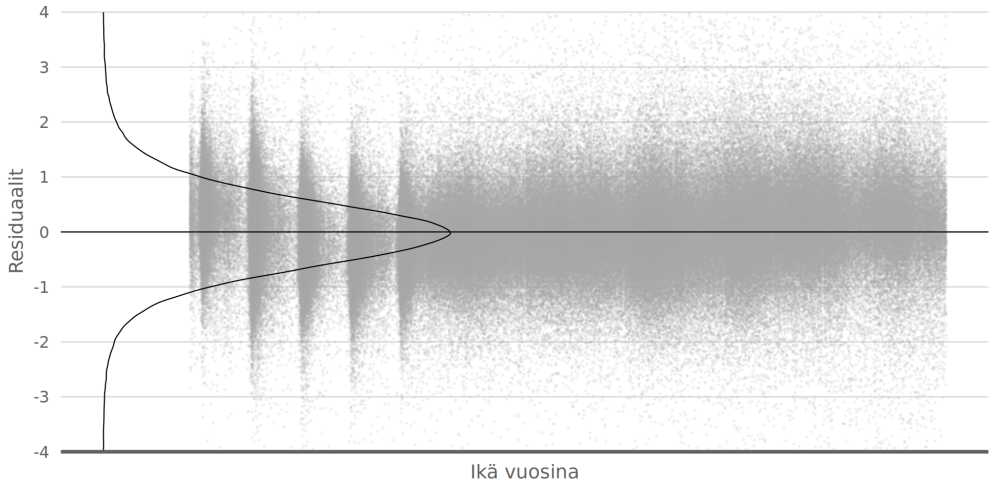
$$g(\text{IKÄ}_{ij}) = \sqrt{\text{IKÄ}_{ij}}.$$

Tällöin saamme varianssipainoiksi $1/\sqrt{\text{IKÄ}_{ij}}$, joiden avulla voimme sovittaa uudelleen satunnaisvaikusten mallin pyrkien painottamaan varianssia heteroskedastisuuden vähentämiseksi. Kuvassa 17 varianssipainokertoimien muutos kuvattu iän suhteen.

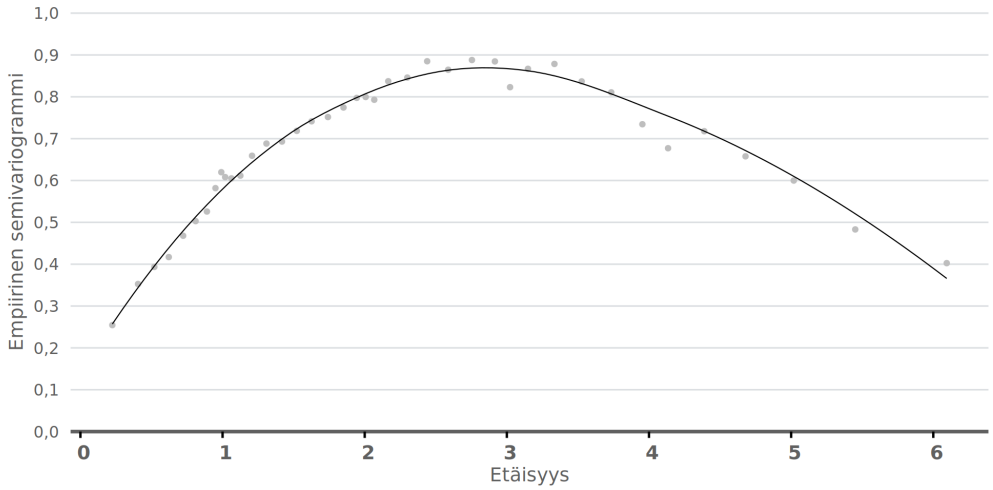


Kuva 17: Varianssipainokertoimien $1/g(\text{IKÄ}_{ij})$ muutos iän suhteen.

Satunnaisvaikutusten varianssipainotetun mallin residuaalien vaihtelu (Kuva 18) on hieman parantunut suuremmilla iän arvoilla, erityisesti kuvaajan alkupäässä residuaalivaihtelu on vääristynyttä. Heteroskedastisuuden tasoittuminen on havaittavissa myös semivariogrammin estimaateissa (Kuva 19), mutta autokorrelaatio vaikuttaa pysyneen muuttumattomana.



Kuva 18: Satunnaisvaikutusten varianssipainotetun mallin yksilökohtaisten mittausten Pearsonin residuaalit iän suhteen ja tiheysfunktion Gaussin ydinestimaatti.

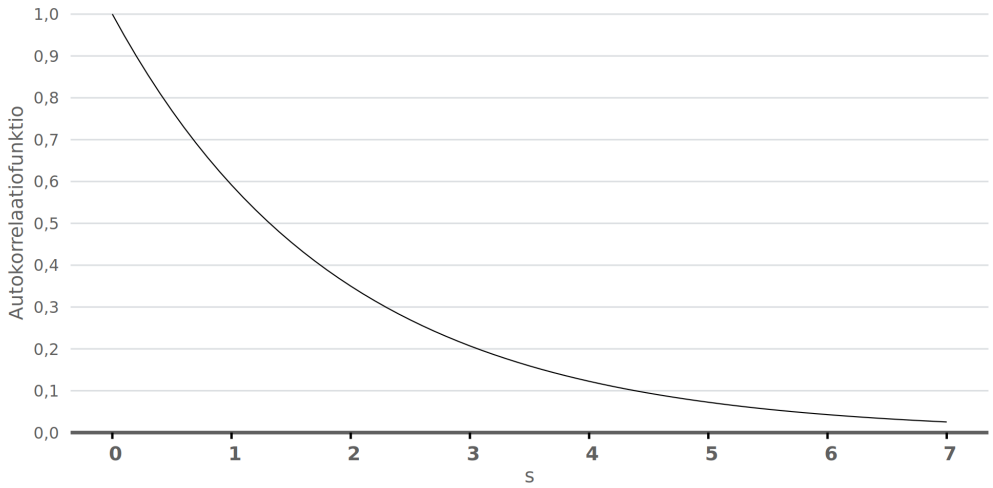


Kuva 19: Satunnaisvaikutusten varianssipainotetun mallin empiirinen semivariogrammi ($\hat{\gamma}$).

Kovarianssirakenteita käsittelevässä alaluvussa 3.1.4 esittelimme ensimmäisen asteen autoregressiivisen mallin AR(1) kovarianssirakenteen olevan riittämätön epätasapainoiseen aineistoon. Pinheiro ja Bates (2000) määrittelevät ensimmäisen asteen autoregressiiviseksi autokorrelaatiofunktioksi $h(k, \phi) = \phi^k$, jossa korrelaatioparametri ϕ on *lag-1*-korrelaatiokerroin ja $k = 0, 1, \ldots n_i$ Tämä voidaan heidän mu-

kaansa yleistää jatkuvan autoregressiivisen mallin CAR(1) autokorrelaatiofunktioksi $h(s, \phi) = \phi^s$, jossa $s > 0$ ja $\phi > 0$.

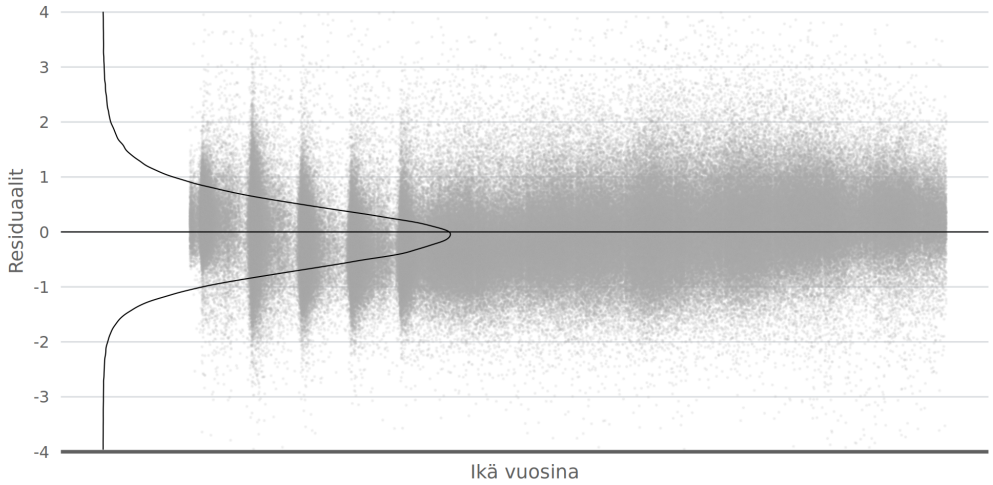
Korrelaatioparametrin estimaatiksi saamme $\hat{\phi} = 0,591$, mikä viittaa korkeaan autokorrelaatioon (Pinheiro ja Bates, 2000). Kuvan 20 mukaan korrelaatio laskee jyrkemmin noin kahden vuoden etäisyyteen saakka, jonka jälkeen se suppenee kohti nollaa. Tämä on ensimmäisten kolmen vuoden osalta likimain yhtäpitävää empiirisen semivariogrammin tulosten suhteen, mutta ei vastaa tilannetta yli kolmen vuoden etäisyyksillä.



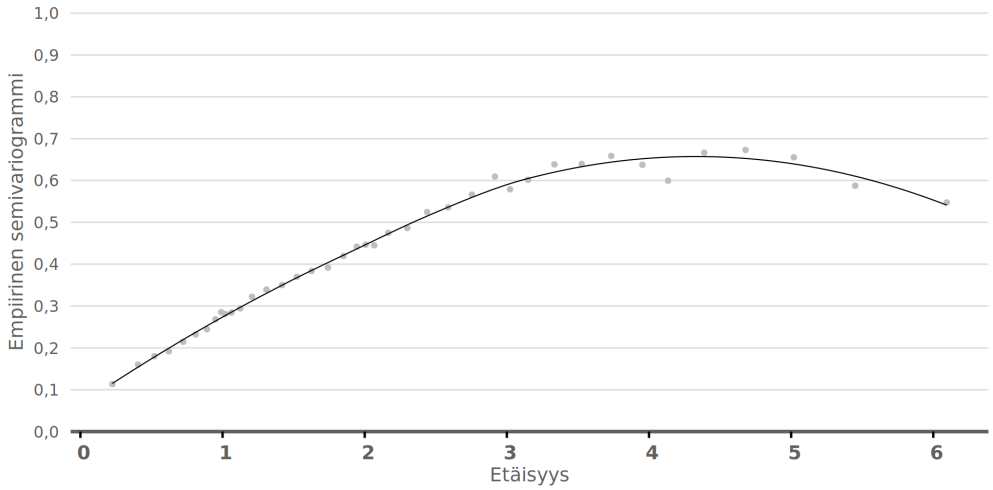
Kuva 20: CAR(1)-mallin autokorrelaatiofunktion $\hat{\phi}^s$ muutos s suhteen, jossa $s > 0$ on mittausten etäisyys vuosina ja $\hat{\phi} = 0,591$.

Tarkastellaan satunnaisvaikutusten varianssipainotettua mallia, kun mallissa huomioidaan residuaalien autokorrelaatio, olettaen residuaaliparien etäisyydelle CAR(1)-mallin mukainen autokorrelaatiofunktio estimoidulla korrelaatioparametrilla $\hat{\phi} = 0,591$. Kutsumme varianssipainotettua residuaalien autokorrelaation huomioivaa mallia jatkossa laajennetuksi malliksi.

Residuaaleissa on havaittavissa merkittävää parannusta, vaikka varianssissa ilmeneekin iän suhteen aaltoilua (Kuva 21). Semivariogrammin perusteella autokorrelaatiota ei ole saatu täysin eliminoitua, mutta sen arvot ovat tasoittuneet huomattavasti (Kuva 22). Myös kaukana toisistaan sijaitsevien mittausten epäilyttävä korreloituneisuus on lieventynyt.



Kuva 21: Satunnaisvaikutusten laajennetun mallin yksilökohtaisten mittausten Pearsonin residuaalit iän suhteen ja tiheysfunktion Gaussin ydinestimaatti.



Kuva 22: Satunnaisvaikutusten laajennetun mallin empiirinen semivariogrammi ($\hat{\gamma}$).

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\beta}_0$	15,428	0,0132	0,0000	15,402	15,454
$\hat{\beta}_1$	-0,1564	0,0023	0,0000	-0,1609	-0,1519
$\hat{\beta}_2$	0,2841	0,0025	0,0000	0,2792	0,2889
$\hat{\beta}_3$	-0,3179	0,0175	0,0000	-0,3522	-0,2835
$\hat{\beta}_4$	0,0447	0,0032	0,0000	0,0383	0,0510

Taulukko 14: Laajennetun satunnaisvaikutusmallin LSM_{VK} 6 kiinteiden vaikustusten REML-estimaatit.

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\sigma^2}_{b_{0i}}$	3,7878	1,9462	0	3,7047*	3,8729*
$\hat{\sigma^2}_{b_{1i}}$	0,1915	0,4376	0	0,1884*	0,1946*
$\hat{\sigma}_{b_{0i},b_{1i}}$	-0,6898	-	-	-	-
$\hat{\sigma^2}$	0,1611	0,4014	0	0,1566*	0,1659*

Taulukko 15: Laajennetun satunnaisvaikutusmallin LSM_{VK} 6 satunnaisvaikustusten ja residuaalien varianssi-kovarianssiestimaatit. Kovarianssille ei saatavilla keskivirhettä tai luottamusvälejä (-). *Luottamusvälit arvioitu normaaliapproksimaatiolla.

Kiinteiden vaikutusten parametrien estimaateissa (Taulukko 14) on havaittavissa pieniä muutoksia satunnaisvaikutusten korreloimattoman ja vakiovarianssisen mallin estimaatteihin, mutta merkittävin ero luonnollisesti löytyy satunnaisvaikutusten ja residuaalien varianssi- ja kovarianssiestimaateista (Taulukko 15). Vakiotermin satunnaisvaikutuksen ja residuaalivarianssin estimaatit ovat huomattavasti pienempiä.

Tulokset antavat vahvoja viitteitä varianssi- ja autokorrelaatiofunktioiden määrittelylle, mutta vakavastiotettavien johtopäätösten tuottamiseksi malli tulisi tarkistaa Pinheiro ja Bates (2000) esittämän laajennetun lineaarisen sekamallin määritelmän mukaiseksi.

Lopullista mallia etsiessämme tulemme kuitenkin hyödyntämään laajennetun mallin tuloksia viitteellisinä kontrafaktuaaleina, joihin peilaamme lopullisesta mallista saatuja tuloksia.

4.2.4 Mallin parametrien karsinta

Palaamme tarkastelemaan satunnaisvaikuten mallia LSM 6 ja arvioimme kriittisesti kiinteiden vaikutusten perusteita vertaamalla täyttä mallia kiinteille vaikutuksille

kriittisiin malleihin. Koska vertaamme malleja, joissa kiinteät vaikutukset eivät ole samoja, sovitamme mallit ML-menetelmällä.

Kriittisiksi malleiksi muodostamme satunnaisvaikutusten mallin ilman kiinteää sukupuolen ja iän yhdysvaikutusta (LSM 7)

$$\begin{aligned} \text{BMI}_{ij} = \beta_0 + \beta_1 \times \text{IKÄ}_{0i} + \beta_2 \times \text{IKÄ}_{ij} + \beta_3 \times \text{SUKUPUOLI}_{ij} \\ + \text{b}_{0i} + \text{b}_{1i} \times \text{IKÄ}_{ij} + \epsilon_{ij}, \end{aligned} \tag{7}$$

sekä täysin ilman kiinteää sukupuolen vaikutusta (LSM 8)

$$\text{BMI}_{ij} = \beta_0 + \beta_1 \times \text{IKÄ}_{0i} + \beta_2 \times \text{IKÄ}_{ij} + \text{b}_{0i} + \text{b}_{1i} \times \text{IKÄ}_{ij} + \epsilon_{ij}. \tag{8}$$

West *ym.* (2014) mukaan yliparametrisoidun mallin on tarkoitus tarjota perusteet satunnaisvaikutusten määrittämiseen, mutta lopuksi tavoitteena on löytää kompromissi mahdollisimman yksinkertaisen, mutta riittävän selitysvoimaisen mallin välillä.

Vaikka täysi malli (LSM 6) tarjoaa viitteellisesti parhaan yhteensopivuuden aineiston suhteen, informaatiokriteerien (Taulukko 16) perusteella ylimääräisten kiinteiden vaikutusten säilyttämiseen ei näytä olevan perusteita.

Malli	*df*	$l(\boldsymbol{\beta})$	AIC	BIC
LSM 8	7	-729345.8	1458706	1458782
LSM 7	8	-729297.8	1458612	1458699
LSM 6	9	-729214.8	1458448	1458546

Taulukko 16: Kriittisten mallien ja täyden mallin log-uskottavuudet ja informaatiokriteerit.

Suuresta havaintomäärän n suhteesta parametrien määrään p seuraten perinteiset mm. F- ja χ^2 -testisuureisiin perustuvat tilastolliset testit tulkitsevat hyvinkin pienet erot mallien uskottavuusosamäärissä merkitseviksi, ottamatta kantaa siihen onko malleilla ilmiön kannalta merkittävää eroa.

	Estimaatti	Keskivirhe (se)	P	$\text{LV}_{2,5}$	$\text{LV}_{97,5}$
$\hat{\beta}_0$	14,440	0,0135	0,0000	14,413	14,466
$\hat{\beta}_1$	-0,1730	0,0020	0,0000	-0,1769	-0,1690
$\hat{\beta}_2$	0,4045	0,0018	0,0000	0,4010	0,4080

Taulukko 17: Kriittisen mallin LSM 8 kiinteiden vaikustusten ML-estimaatit.

	Estimaatti	Keskivirhe (se)	P	$LV_{2,5}$	$LV_{97,5}$
$\hat{\sigma^2}_{b_{0i}}$	8,4515	2,9071	0	8,2993*	8,6065*
$\hat{\sigma^2}_{b_{1i}}$	0,2134	0,4620	0	0,2109*	0,2159*
$\hat{\sigma}_{b_{0i},b_{1i}}$	-1,0658	-	-	-	-
$\hat{\sigma^2}$	0,5938	0,7706	0	0,5900*	0.5977*

Taulukko 18: Kriittisen mallin LSM 8 satunnaisvaikustusten ja residuaalien varianssi-kovarianssiestimaatit. Kovarianssille ei saatavilla keskivirhettä tai luottamusvälejä (-). *Luottamusvälit arvioitu normaaliapproksimaatiolla.

Burzykowski ja Galecki (2013) mukaan tilanteissa, joissa hypoteesien testausta ei katsota mielekkääksi, voidaan mallinvalinnassa hyödyntää informaatiokriteerien arviota mallin ja aineiston yleisestä yhteensopivuudesta.

Erääksi vaihtoehdoksi Burzykowski ja Galecki (2013) ehdottavat myös kiinteiden vaikutusten parametrien empiiristen jakaumien simulointia esimerkiksi LOO-menetelmällä (*Leave-One-Out*), jossa jakaumaa estimoidaan poistamalla yksittäisiä havaintoja, tai siirtymällä kokonaan bayesiläisiin menetelmiin.

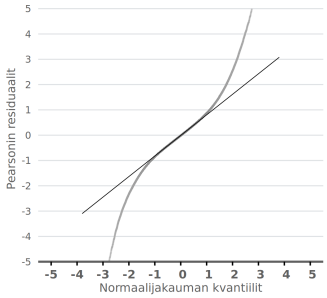
Simulaatiomenetelmille ei kuitenkaan kyetty tämän tutkielman piirissä löytämään laskennallisesti riittävän tehokasta toteutusta ja yleisesti Bayes-kehikon yhteensovittamista tutkielmassa rajattuun näkökulmaan lineaarisista sekamalleista ei katsottu tutkielman kokonaisuuden kannalta mielekkääksi.

Vahvistaaksemme päätöksen valita kiinteille vaikutuksille kriittinen malli LSM 8, pyrimme varmistumaan mallin satunnaisvaikutusten ja residuaalien jakaumaoletuksista, verraten niitä täyteen malliin LSM 6. Peilaamme näitä malleja myös laajennettuihin vertailumallehin LSM_{VK} 6 ja LSM_{VK} 8, joissa on huomioitu varianssipainot ja autokorreloituneet residuaalit.

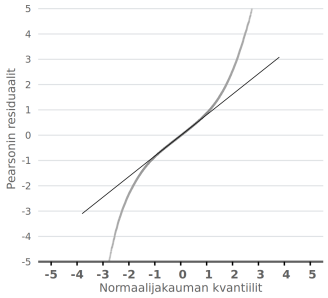
Yksilökohtaisten residuaalien normaalisuusoletuksen tarkistamiseen Burzykowski ja Galecki (2013) suosittelevat kvantiili-kvantiilikuvaajaa (*Q-Q plot*), jossa Pearsonin residuaaleja, järjestettynä pienimmästä arvosta suurimpaan, verrataan oletusjakauman teoreettisiin kvantiileihin. Kuvaajassa havaittujen arvojen tulisi asettua likimain suoralle apuviivalle, joka kulkee ensimmäisen ja kolmannen kvantiilin läpi, jotta havaitun jakauman katsotaan noudattavan oletettua jakaumaa.

Kuvan 23 mallien symmetrinen, mutta hännistä kaareutuva kuvaaja viittaa jakauman paksuhäntäisyyteen. Tämä tarkoittaa sitä, että empiirinen jakauma sisältää normaalijakaumaan nähden äärimmäisempiä arvoja (Pinheiro ja Bates, 2000). Silmämääräisesti ei ole havaittavissa, että mallit LSM_{VK} 6 ja LSM_{VK} 8 tarjoaisivat

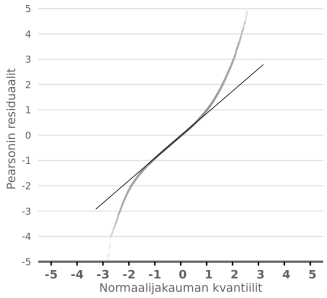
huomattavaa lisäarvoa.



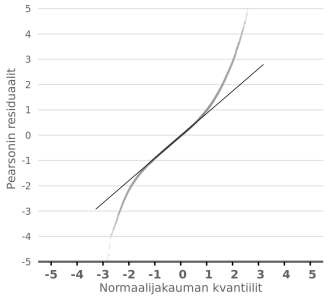
(a) Kriittinen malli LSM 8.



(b) Täysi malli LSM 6.



(c) Kriittinen malli LSM_{VK} 8.

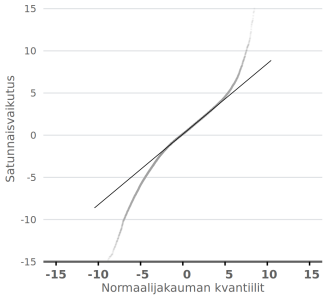


(d) Täysi malli LSM_{VK} 6.

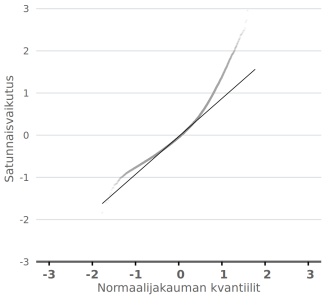
Kuva 23: Kriittisen (a) ja täyden mallin (b) sekä vastaavien laajennettujen mallien (c) ja (d) kvantiili-kvantiilikuvaajat.

Kvantiili-kvantiilikuvaajia voi varovaisin johtopäätöksin hyödyntää myös ennustettujen satunnaisvaikutusten normaalisuusoletusten tarkistamiseen, mutta Burzykowski ja Galecki (2013) huomauttavat, että määritelmäänsä nojaten $\hat{\boldsymbol{b}}_i$ ei välttämättä vastaa $\boldsymbol{b}_i$ todellista jakaumaa.

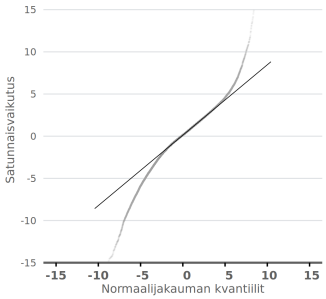
Kuvan 24 perusteella vakiotermin ennustetut satunnaisvaikutuksen vaikuttavat residuaalien tapaan likimain symmetriseltä kaikkien mallien kohdalla, mutta iän osalta jakauma on merkittävän vino.



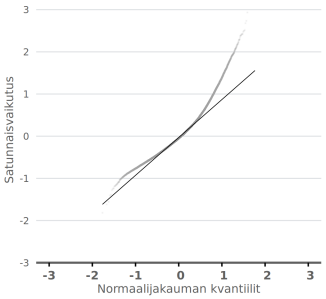
(a) LSM 8 (Vakiotermi).



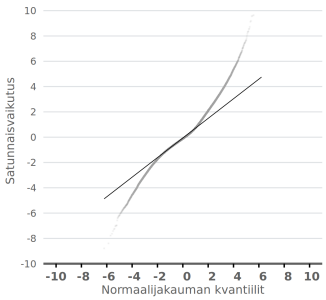
(b) LSM 8 (Ikä).



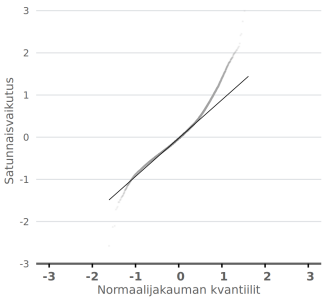
(c) LSM 6 (Vakiotermi).



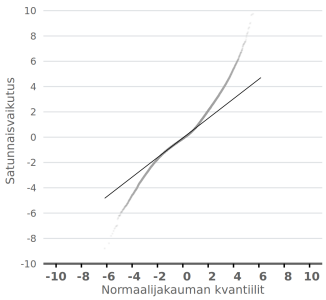
(d) LSM 6 (Ikä).



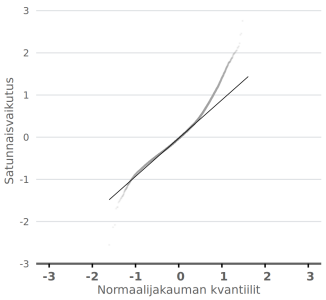
(e) LSM_{VK} 8 (vakiotermi).



(f) LSM_{VK} 8 (Ikä.)



(g) LSM_{VK} 6 (Vakiotermi).

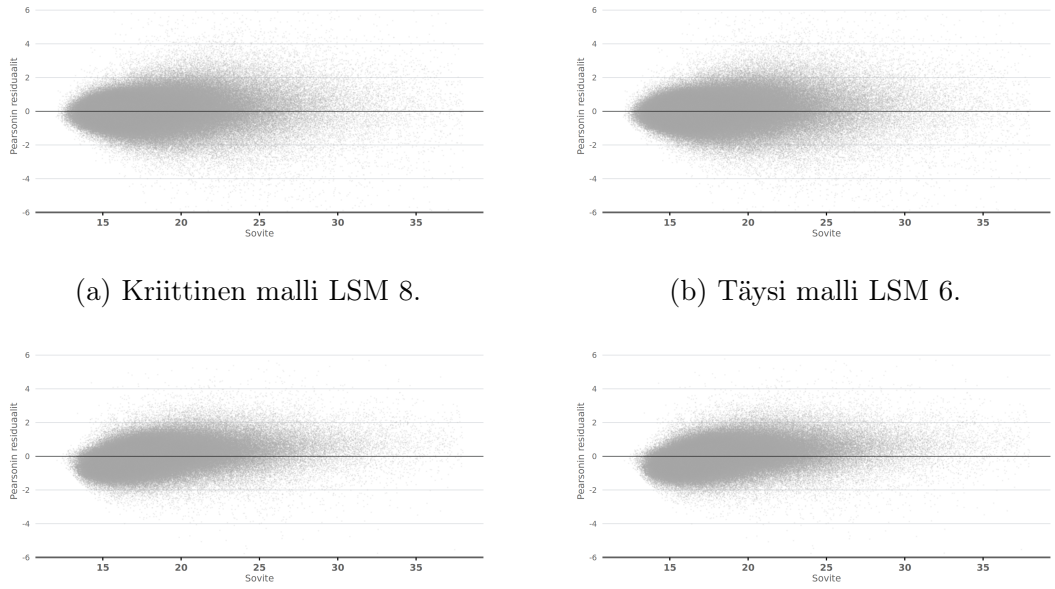


(h) LSM_{VK} 6 (Ikä).

Kuva 24: Kriittisen (a), (b) ja täyden mallin (c), (d) sekä vastaavien laajennettujen mallien (e),(f) ja (g),(h) vakiotermin ja iän ennustettujen satunnaisvaikutusten kvantiili-kvantiilikuvaajat.

Viimeiseksi tarkastelemme Pearsonin residuaaleja suhteessa vastemuuttujan sovitteeseen. West *ym.* (2014) sekä Pinheiro ja Bates (2000) mukaan Pearsonin residuaalit vähentävät residuaalien heteroskedastisuutta ja residuaalien tulisi olla symmetrisesti jakautuneita suhteessa vastemuuttujan sovitteeseen. Burzykowski ja Galecki (2013) suosittelevat kuitenkin varovaisempaan suhtautumiseen, sillä kuvaaja näyttää kaikkien aikapisteiden havainnot päällekäin ja tiedämme, että yksilökohtaiset mittaukset ovat ajan suhteen korreloituneita.

Kuvan 25 mukaan kriittisen ja täyden mallin LSM 8 ja LSM 6 Pearsonin residuaaleissa on havaittavissa lievää kasvavaa trendiä suurilla vastemuuttujan arvoilla. Tämä voi olla selitettävissä Burzykowski ja Galecki (2013) huomiolla mittausten korreloituneisuudesta, sillä painoindeksi saa suurempia arvoja vanhemmilla lapsilla. Mallien LSM_{VK} 8 ja LSM_{VK} 6 osalta paljastuu, että vaikka laajennetun mallin yksilökohtainen residuaalivarianssi on huomattavasti pienempää, ovat residuaaliestimaatit hyvin vääristyneitä suhteessa vastemuuttujan sovitteeseen ja siten laajennetun mallin määrittelyssä voi olla vakavia puutteita.



(a) Kriittinen malli LSM 8.

(b) Täysi malli LSM 6.

(c) Kriittinen malli LSM_{VK} 8.

(d) Täysi malli LSM_{VK} 6.

Kuva 25: Kriittisen (a) ja täyden mallin (b) sekä vastaavien laajennettujen mallien (c) ja (d) skaalaamattomat residuaalit suhteessa vastemuuttujan sovitteeseen.

Tavoitteenamme on ollut valita mahdollisimman yksinkertainen, mutta riittävä malli. Huomioiden kriittisen mallin LSM 8 ja täyden mallin LSM 6 tulosten häviävän pienet eroavaisuudet, sekä laajennetun mallin heikko määrittely ja vakavat puutteet, valitsemme lopulliseksi malliksi kiinteille vaikutuksille kriittisen mallin LSM 8.

4.3 Lopullinen malli

Tarkastelemme seuraavaksi tarkemmin lopulliseksi malliksi valittua kriittistä mallia LSM 8. Tiivistämme mallin tuloksia ja pyrimme arvioimaan millaisia johtopäätöksiä voimme tehdä mallin perusteella tutkielman aineistosta.

Kuten alustavasti olemme huomioineet, suuresta havaintomäärästä seuraten parametrien estimaattien tilastollisesti merkitsevä eroaminen nollasta ei välttämättä tarjoa mielekästä lisätietoa, sillä hyvin pienetkin poikkeamat voivat olla merkitseviä. Ilman aineiston edustavuuden validointia, voimme jopa kyseenalaistaa liittyykö seurannan alun iän IKÄ_{0i} estimoidun kiinteän vaikutuksen $\hat{\boldsymbol{\beta}}_1 = -0,1730$ tuoma tieto siitä, että alun iän muutos vuodella vähentäisi keskimääräistä painoindeksiä alle viidenneksellä, todella tutkittavaan ilmiöön vai aineiston edustavuuden puutteisiin.

Toisin sanoen, uskomme mallin LSM 8 estimaattorien tuottavan täsmällisiä estimaatteja, mutta ilman tietoa aineiston edustavuudesta, emme voi olla varmoja niiden mielekkyydestä tai harhattomuudesta.

Esimerkiksi Sullivan ja Feinn (2012) korostavat artikkelissaan, että lääketieteellisen tutkimuksen näkökulmasta, tilastollisen mallin pohjalta tehty mielekäs tulkinta tulisi perustaa itse vaikutusten voimakkuuteen (*effect size*), ei siihen poikkeavatko vaikutukset merkitsevästi annetun hypoteesin mukaisesta vertailuarvosta.

Nakagawa ja Schielzeth (2013) ja laajentaen Johnson (2014) esittävät artikkeleissaan lineaarisiin sekamalleihin soveltuvan marginaalisen pseudo-selitysasteen ($R^2_{\text{LSM}(m)}$), huomioiden ainoastaan residuaalivarianssin

$$R^2_{\text{LSM}(m)} = \frac{\sigma^2_f}{\sigma^2_f + \bar{\sigma^2_l} + \sigma^2_\epsilon},$$

sekä ehdollisen selitysasteen ($R^2_{\text{LSM}(e)}$), huomioiden residuaalien sekä satunnaisvaikutusten varianssin

$$R^2_{\text{LSM}(e)} = \frac{\sigma^2_f + \bar{\sigma^2_l}}{\sigma^2_f + \bar{\sigma^2_l} + \sigma^2_\epsilon},$$

jossa σ^2_f kiinteisiin vaikutuksiin liitetty varianssi, $\bar{\sigma^2_l}$ satunnaisvaikutusten varianssin yhdiste ja σ^2_ϵ residuaalivarianssi.

Selitysasteiden (Taulukko 19) mukaan kiinteät vaikutukset yksinään selittävät ainoastaan noin 11 % varianssikomponenttien yhteisvaihtelusta, kun kiinteät ja satunnaisvaikutukset yhdessä selittävät noin 95 % mallin vaihtelusta.

$R^2_{\text{LSM}(m)}$	0,11
$R^2_{\text{LSM}(e)}$	0,95

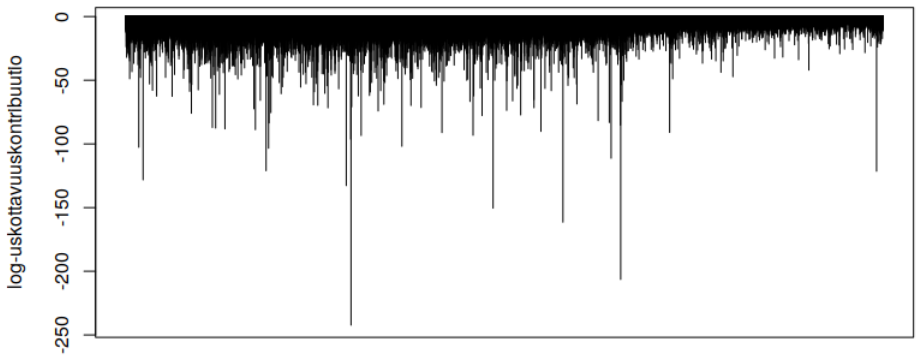
Taulukko 19: Johnson (2014) laajennettu marginaali- ja ehdollinen pseudo-R^2_{LSM}.

Kriittisen mallin LSM 8 tarkastelu osoittaa empiiristä tukea sille, että Laird ja Ware (1982) kaksitasoisen lineaarisen sekamallin kehikko tarjoaa potentiaalisen työkalun FinLapset-rekisteriaineiston kaltaisesta lähteestä muodostetun pitkittäisaineiston analyysiin.

Mallin satunnaisvaikutusten määrittely vaikuttaa parantavan huomattavasti mallin kykyä selittää yksilöiden välistä vaihtelua, mutta on pidettävä todennäköisenä, että olisi löydettävissä sellaisia painoindeksiin vaikuttavia taustatekijöitä, joiden avulla voisimme parantaa myös kiinteiden vaikutusten selitysvoimaa.

Tutkielmassa esitetyt esimerkit varianssin heteroskedastisuuden ja residuaalien autokorrelaation huomioimisesta eivät kestäneet tarkempaa tarkastelua. Esimerkit antoivat kuitenkin viitteitä näiden ominaisuuksien olemassaolosta ja tarpeesta tarkastella niitä tarkemmin.

Viimeiseksi, pienistä keskivirheistä seuraneet haasteet mallin piste-estimaattien epävarmuustekijöiden havainnollistamisessa muodostavat suuren puutteen kokonaisuuteen nähden. Esimerkiksi yksilökohtaisissa log-uskottavuuskontribuutioissa on huomattavaa vaihtelua (Kuva 26), johon parametrien jakaumien simulointi uudelleenotannalla, kuten Burzykowski ja Galecki (2013) esittämällä LOO-menetelmällä, voisi tuoda lisävaloa. On myös huomioitava, että tällaisen näkökulman tuominen frekventistisen uskottavuuspäättelyn piiriin vaikuttaa väkinäiseltä, sillä jo lähtökohtaisesti Bayes-menetelmät tarjoaisivat luontevamman lähestymistavan parametrien posteriorijakaumien tarkasteluun.



Kuva 26: Kriittisen mallin LSM 8 yksilökohtaiset log-uskottavuuskontribuutiot.

5 Johtopäätökset

Lineaariset sekamallit tarjoavat luontevan työkalun FinLapset-aineiston analyysiin. Lopullisen mallin residuaalien ja selitysvoiman tarkastelun perusteella jo hyvin yksinkertainen kiinteiden vaikutusten rakenne, yhdessä vakiotermin ja aikadimension regressiokertoimen satunnaisvaikutusten kanssa paransi huomattavasti mallin kykyä selittää aineiston vaihtelua. Tämä tarjoaa mielekkään jatkokysymyksen mielenkiintoisten kiinteiden vaikutusten löytämiselle, ja niiden vaikutusten tutkimiselle yksilöja populaatiotasoilla.

Havaitsimme myös puutteita joissakin mallia koskevissa oletuksissa mm. residuaalien jakaumasta ja niiden korreloimattomuudesta, sekä residuaalivarianssin homoskedastisuudesta. Laajennettu lineaarinen sekamalli voisi tarjota mahdollisia ratkaisuja, mutta välttääksemme mahdollisen konfliktin satunnaisvaikutusten kovarianssimatriisin ja residuaalimatriisin välillä, tulisi malli kokonaisuudessaan määritellä huolellisesti laajennetun mallin näkökulmasta.

FinLapset-aineiston huomattavasta havaintomäärästä seuraten, totesimme monien tavanomaisten mallin arviointikäytäntöjen olevan riittämättömiä malliin liittyvien epävarmuustekijöiden kuvaamiseen ja esitimme myös joitakin vaihtoehtoisia lähestymistapoja. Esimerkiksi Connelly *ym.* (2016) kuitenkin painottavat rekisteriaineistojen hyödyntämistä käsittelevässä artikkelissaan, että ennen rekisteriaineistosta johdettuja päätelmiä, tulisi suurin työ analyysissä sen sijaan suunnata aineiston synnyttäneiden prosessien syvälliseen ymmärtämiseen.

Fricke (2014) puolestaan tarkastelee tiedon ja suuren havaintomäärän yhteyttä tietoopin näkökulmasta ja osoittaa useita ongelmia passiivisesti kertyvien massiivisten aineistojen (*Big Data*) pohjalta tehtävässä päättelyssä. Hän yhdistää massiivisten aineistojen aikakaudella vallitsevan aineistolähtöisen (*data-driven*) päättelyn Popper (1986) induktion kritiikkiin. Fricke (2014) mukaan aineistolähtöinen päättely on luonteeltaan induktiivista ja siinä korostuu harha tiedon lisääntymisestä havaintomäärän kasvaessa. Vaikka FinLapset-aineisto ei ole varsinaisesti massiivinen aineisto, on se rekisteriaineistona luonteeltaan passiivista ja houkuttelee induktiiviseen päättelyyn.

Otostutkimuksen näkökulmasta, hyvin määritellystä otantakehikosta nostetun otoksen havaintomäärän lisääminen vähentää tutkimuksen mittareihin liittyvää epävarmuutta. Koska rekisteritutkimuksissa aineistoa käytetään usein muuhun, kuin sen alkuperäiseen käyttötarkoitukseen, voimme jopa kyseenalaistaa voiko kehikkoa, ja siten validia ja reliaabelia mittaria olla edes olemassa. Aineiston koko voi myös luoda induktiivisen harhan, että havaintomäärän lähestyessä populaation todellista kokoa, havaintojen perusteella ilmiötä koskevat päätelmämme lähestyisivät myös populaation todellisen ilmiön luonnetta, silloinkin kun meiltä puuttuu täysin kyky mittarimme falsifiointiin.

Päinvastoin, huolellisesti määritellyn kehikon turvin olisi suurten aineistojen parissa varaa aktiivisesti rajautua tutkimuskysymyksen kannalta vain kaikista luotettavimpiin ja edustavimpiin havaintoihin ja muodostaa itse ilmiön näkökulmasta validi ja reliaabeli mittari.

Viitteet

Ben Bolker. GLMM FAQ. 2020. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html.

Tomasz Burzykowski ja Andrzej Galecki. *Linear Mixed-Effects Models using R: A Step-by-Step Approach*. Springer, toinen painos, 2013.

George Casella ja L. Berger, Roger. *Statistical Inference*. Duxbury Press, toinen painos, 2002.

CDC. *Modified z-scores in the CDC growth charts*, 2013. https://www.cdc.gov/nccdphp/dnpao/growthcharts/.

Tim James Cole. The LMS Method for Constructing Normalized Growth Standards. *European journal of clinical nutrition*, **44**, no. January, 45–60., 1990.

Roxanne Connelly, Christopher J. Playford, Vernon Gayle ja Chris Dibben. The Role of Administrative Data in the Big Data Revolution in Social Science Research. *Social Science Research*, **59**, 1 – 12, 2016. Special issue on Big Data in the Social Sciences.

J. Diggle, Peter, J. Heagerty, Patrick, Kung-Yee Liang ja L. Zeger, Scott. *Analysis of Longitudinal Data*. Oxford Press, toinen painos, 2013.

M. Fitzmaurice, Garret, M. Laird, Nan ja H. Ware, James. *Applied longitudinal analysis*. Wiley, toinen painos, 2011.

M. Flegal, Katherine ja Tim James Cole. The LMS Method for Constructing Normalized Growth Standards. *National Health Statistics Report*, **63**, no. February, 2013.

Martin Fricke. Big Data and Its Epistemology. *Journal of The Association for Information Science and Technology*, **66**, no. 4, 651–661, 2014.

Harvey Goldstein. *Multilevel statistical models*. Wiley, neljäs painos, 2011. Dawsonera Ebook.

C. R. Henderson, Oscar Kempthorne, S. R. Searle ja C. M. von Krosigk. The Estimation of Environmental and Genetic Trends from Records Subject to Culling. *Biometrics*, **15**, no. 2, 192–218, 1959.

Marek Hlavac. *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Central European Labour Studies Institute (CELSI), Bratislava, Slovakia, 2018. R package version 5.2.2.

C. D. Johnson, Paul. Extension of Nakagawa and Schielzeth's R^2_{GLMM} to Random Slopes Models. *Methods in Ecology and Evolution*, **5**, no. 9, 944–946, 2014.

Nan M. Laird ja James H. Ware. Random-Effects Models for Longitudinal Data. *Biometrics*, **38**, no. 4, 963–974, 1982.

Mary J. Lindstrom ja Douglas M. Bates. Newton—Raphson and EM Algorithms for Linear Mixed-Effects Models for Repeated-Measures Data. *Journal of the American Statistical Association*, **83**, no. 404, 1014–1022, 1988.

Mary J. Lindstrom ja Douglas M. Bates. Nonlinear Mixed Effects Models for Repeated Measures Data. *Biometrics*, **46**, no. 3, 673–687, 1990.

Xian Liu. *Methods and Applications of Longitudinal Data Analysis*. Academic Press, 2016. Elsevier ScienceDirect Ebook.

Oskari Luomala. thlGraphs: Implementation of THL graphical guidelines in R. 2019. https://github.com/THLfi/thlGraphs.

S. Nakagawa ja H. Schielzeth. A General and Simple Method for Obtaining R^2 from Generalized Linear Mixed-Effects Models. *Methods in Ecology and Evolution*, **4**, no. 2, 133–142, 2013.

Kari Nissinen. *Small Area Estimation with Linear Mixed Models from Unit-level Panel and Rotating Panel Data*. Väitöskirja, Jyväskylän yliopisto, 2009.

Jorge Nocedal ja Stephen J. Wright. *Numerical Optimization*. Springer, New York, NY, USA, toinen painos, 2006.

Overleaf Digital Science. Overleaf. 2020. https://www.overleaf.com/.

C. Pinheiro, Jose ja Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, 2000.

C. Pinheiro, Jose, Douglas M. Bates, Saikat DebRoy, Deepayan Sarkar ja R Core Team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2013. R package version 3.1-147.

R. Popper, Karl. *The Logic of Scientific Discovery*. Hutchinson, 12. painos, 1986.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

G. K. Robinson. That BLUP is a Good Thing: The Estimation of Random Effects. *Statistical Science*, **6**, no. 1, 15–32, 1991.

G. Sullivan ja R. Feinn. Using Effect Size-or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, **4**, no. 3, 279–282, 2012.

J. Talbott, William. *Models for Intensive Longitudinal Data*. Oxford University Press USA, 2006. ProQuest Ebook Central.

THL. FinLapset-rekisteri. 2019. https://thl.fi/fi/tutkimus-ja-kehittaminen/tutkimukset-ja-hankkeet/finlapset-lasten-nuorten-ja-perheiden-terveys-ja-hyvinvointi/finlapset-rekisteri.

W. R. Twisk, Jos. *Applied Longitudinal Data Analysis for Epidemiology : A Practical Guide*. Cambridge University Press, 2013. ProQuest Ebook Central.

Geert Verbeke ja Geert Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer, 2000. ProQuest Ebook Central.

Brady T. West, Brenda W. Gillespie, Andrzej T. Gałecki ja Kathleen B. Welch. *Linear Mixed Models: A Practical Guide Using Statistical Software*. Chapman and Hall/CRC, toinen painos, 2014. eBook Collection (EBSCOhost).

WHO. *Computation Of Centiles and Z-Scores for Height-for-Age, Weight-for-Age and BMI-for-Age*, 2007. https://www.who.int/growthref/en/.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4.

Yihui Xie. knitr: A comprehensive tool for reproducible research in R. Teoksessa Victoria Stodden, Friedrich Leisch ja Roger D. Peng, toimittajat, *Implementing Reproducible Computational Research*. Chapman and Hall/CRC, 2014. ISBN 978-1466561595.

Yang Yang ja Kenneth C. Land. A Mixed Models Approach to the Age-Period-Cohort Analysis of Repeated Cross-Section Surveys, with an Application to Data on Trends in Verbal Test Scores. *Sociological Methodology*, **36**, 75–97, 2006.

Liite 1. Tutkielman työkalut

Tutkielma on toteutettu LaTeX-ladontakielellä käyttäen Overleaf-editoria (Overleaf Digital Science, 2020) ja tutkielman tilastoanalyysit on suoritettu R-ohjelmistolla (R Core Team, 2013).

R-ohjelmiston tulosteiden LaTeX-käännökset on tehty hyödyntäen R-ohjelmiston knitr- ja stargazer-paketteja. (Xie, 2014; Hlavac, 2018) ja tilastografiikat käyttäen ggplot2-pakettia (Wickham, 2016) sekä THL:n visuaalisen ohjeiston ggplot2-määrittelyjä (Luomala, 2019).

A Lineaarisen sekamallin sovittaminen lme()-funktiolla

Tutkielman mallit sovitettiin R-ohjelmiston (R Core Team, 2013) **nlme**-paketin **lme**-funktiolla. Kaikki tutkielman mallit konvergoivat käyttäen *Nelder-Mead* optimointialgoritmia. Tarkempaa keskustelua optimointialgoritmin valinnasta käymme liittessä B.

Alla esitämme R-koodit tutkielman mallien muodostamiseksi **lme**-funktiolla. REML-estimointi tapahtuu antamalla argumentin **method** arvoksi *"REML"*.

LM 4

```
lm(bmi~ ika*sukupuoli,
   data = aineisto)
```

LSM 5

```
lme(bmi~ ika_0i + ika*sukupuoli,
    random =~1 | yksilo,
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

LSM 6

```
lme(bmi~ ika_0i + ika*sukupuoli,
    random =~ika | yksilo,
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

LSM 7

```
lme(bmi~ ika_0i + ika + sukupuoli,
    random =~ika | yksilo,
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

LSM 8

```
lme(bmi~ ika_0i + ika,
    random =~ika | yksilo,
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

LSM_{VK} 6

```
lme(bmi~ ika_0i + ika*sukupuoli,
    random =~ika | yksilo,
    weights = varFixed(~ika),
    correlation = corCAR1(form =~ika | yksilo),
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

LSM_{VK} 8

```
lme(bmi~ ika_0i + ika*sukupuoli,
    random =~ika | yksilo,
    weights = varFixed(~ika),
    correlation = corCAR1(form =~ika | yksilo),
    data = aineisto,
    method = "ML",
    control = lmeControl(opt = "optim", optimMethod = "Nelder-Mead"))
```

B Optimointimenetelmät

Kaikki mallien kohdalla lme()-funktion oletusarvoinen optimointimenetelmä **nlminb** ei konvergoinut. Mm. West *ym.* (2014) mainitsevat vastaavista ongelmista, mutta eivät käsittele ratkaisuja tarkemmin R-ohjelmiston näkökulmasta.

Nocedal ja Wright (2006) tarjoavat syvällisen katsauksen mm. nlme-paketin hyödyntämiin optimointialgoritmeihin, mutta tämän tutkielman osalta perustellinen tarkastelu sivuutetaan.

Tyydymme McMasterin yliopiston matematiikan ja tilastotieteen laitoksen professorin, Ben Bolkerin (2020) blogi-kirjoituksessaan tarjoamaan *kultaiseen sääntöön* suorittaa estimointi kaikilla saatavilla olevilla optimointialgoritmeillä, tarkistaen ovatko estimaatit *käytännössä toisiaan vastaavia.*

Kaikista nlme-paketin tukemista optimointialgoritmeista (*nlminb, Nelder-Mead, BFGS, CG, L-BFGS-B, SANN, Brent*) vain Nelder-Mead konvergoi kaikkien mallien ML- ja REML-estimointien kohdalla. Taulukossa 20 esitetään lopullisen mallin LSM 8 ML-estimaatit nlminb, Nelder-Mead ja L-BFGS-B -menetelmillä.

Nelder-Mead -menetelmän estimaatit poikkeavat hieman nlminb- ja L-BFGS-B -menetelmistä, mutta ero on hyvin pieni, emmekä katso sen vaikuttavan mallin tulkintaan. Lisäksi, koska Nelder-Mead konvergoi kaikkien mallien kohdalla, ovat mallit

siten optimointimenetelmän suhteen vertailukelpoisia.

	Vastemuuttuja: bmi		
	(nlminb)	(Nelder-Mead)	(L-BFGS-B)
Vakiotermi ($\hat{\beta}_0$)	14,432802	14,439736	14,432899
	(0,013605)	(0,013543)	(0,013604)
ika_0i ($\hat{\beta}_1$)	−0,171405	−0,172975	−0,171421
	(0,002031)	(0,002024)	(0,002031)
ika ($\hat{\beta}_2$)	0,404348	0,404479	0,404351
	(0,001772)	(0,001776)	(0,001772)
Vakiotermi ($\hat{\sigma^2}_{b0i}$)	8,497261	8,451481	8,496399
	(2,915006)	(2,907143)	(2,914858)
ika ($\hat{\sigma^2}_{b1i}$)	0,212137	0,213400	0,212131
	(0,460584)	(0,461953)	(0,460577)
Resid.varianssi ($\hat{\sigma^2}$)	0,592408	0,593841	0,592420
	(0,769680)	(0,770611)	(0,769688)
Log-uskottavuus	−729343,0	−729346	−729343
AIC	1458700	1458706	1458700
BIC	1458777	1458782	1458777

Taulukko 20: Eri optimointimenetelmien kiinteiden vaikutusten parametrien ML-estimaatit sekä satunnaisvaikutusten ja residuaalin varianssiestimaatit sekä keskivirheet (suluissa) mallille LSM 8.