Tampereen yliopisto

Aino Kumpumäki

Lasten pituuskasvun analysointi mixture regressiolla ja sekamalleilla

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Toukokuu 2021

Tiivistelmä

Aino Kumpumäki: Lasten pituuskasvun analysointi mixture regressiolla ja sekamalleilla
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Toukokuu 2021

Mixture regressio ja sekamalli ovat tilastollisia mallinnusmenetelmiä, jotka soveltuvat hyvin pitkittäistutkimuksessa kerättyyn aineistoon. Mixture regressiossa aineisto pyritään jakamaan klustereihin eli alipopulaatioihin. Alipopulaation jakaumia kutsutaan mixture-malleiksi. Jokaisen alipopulaation kohdalla voidaan tarkastella erikseen, kuinka hyvin malli asettuu niihin. Sekamallilla voidaan mallintaa aineistoa niin, että myös satunnaisvaikutukset otetaan huomioon. Sekamalli sopii erityisesti pitkittäisaineistoon, koska puuttuvat havainnot voidaan olettaa satunnaisiksi. Molemmilla menetelmillä voidaan useampien kuvaajien avulla pohtia, kuinka hyvin malli asettuu aineistoon.

Työssä kerrotaan aineistosta, esitellään menetelmät sekä sovelletaan menetelmiä aineistoon R-ohjelmiston avulla. Työssä käytetty aineisto on pitkittäisaineisto, jossa on mitattu lasten pituuskasvua usealla eri mittauskerralla. Aineisto on jaettu erikseen tyttöjen ja poikien aineistoon. Työssä käytetään mallia, jossa lapsen pituutta selitetään iällä eli mittausajankohdalla. R-ohjelmiston avulla tehdyllä mixture regressiolla saadaan, että tyttöjen ja poikien aineisto voidaan jakaa viiteen alipopulaatioon. Asetettu malli sopii melko hyvin eri alipopulaatioihin. Sekamallilla saadaan R-ohjelmiston avulla estimaattiarvoja, residuaalikuvaaja ja normaalisuusoletus. Residuaalikuvaajan mukaan malli asettuu hyvin varsinkin lapsen myöhemmillä mittauskerroilla. Aineiston normaalisuusoletus on kunnossa eli voidaan likimain olettaa, että aineisto olisi normaalisti jakautunut. Saadut tulkinnat pätevät sekä tyttöjen että poikien aineistoon.

Lopuksi työssä vertaillaan vielä mallin asettumista aineistoon kuvaajassa, jossa tarkastellaan sovitettuja ja havaittuja arvoja. Malli asettuu mixture regressiossa jaettuihin ryhmiin kaikkiin onnistuneesti, varsinkin samassa ryhmässä, jossa residuaalikuvaajassakin malli asettuu hyvin. Sekamallin tilanteessa malli asettuu myös onnistuneesti aineistoon. Mallin asettumista voidaan tarkastella mixture regressiossa ryhmittäin ja sekamallissa kokonaisuudessaan. Molemmilla menetelmillä malli istuu onnistuneesti aineistoon.

Avainsanat: mallintaminen, mixture regressio, mixture-malli, R, sekamalli

Sisällys

1 Johdanto

Tässä työssä käytetään mixture regressiota ja sekamalleja lasten pituuden analysointiin. Kahdella menetelmällä voidaan arvioida, kuinka hyvin aineistoon asetettu malli sopii. Aineiston analysointiin voidaan käyttää lukuisia menetelmiä, joista tässä työssä esitelläänkin mixture regressio ja sekamalli. Tässä työssä käytetty aineisto on pitkittäisaineisto, mihin mixture regressio ja sekamalli soveltuvat hyvin.

Mixture regressio on tilastollinen menetelmä, jossa valitaan, kuinka moneen alipopulaatioon aineisto jaetaan. Jaetusta aineistosta voidaan tuottaa residuaalikuvaajia, joista voidaan päätellä, kuinka hyvin asetettu malli istuu eri mittausaikapisteissä. Sekamallit taas tarkoittavat tilastollisia malleja, joilla voidaan mallintaa aineistoa. Sekamalleilla voidaan tarkastella residuaalikuvaajien avulla, kuinka satunnaisesti eri aikapisteissä residuaalit ovat jakautuneet. Sekamalliin liittyen voidaan myös tarkastaa normaalisuusoletukset. Tässä työssä painottuu idea siitä, miten mallinnus onnistuu eri menetelmiä käytettäessä.

Lasten pituuden kehitystä on tutkittu paljon. Sorvan ym. (1985) tutkimuksessa on tutkittu lasten varhaisen pituuskasvun ja painonkehityksen suuntaa. Tutkimuksessa on mallinnettu kasvun kehitystä kasvukäyrästöllä, jonka tavoitteena on pystyä toteamaan kasvun häiriöt varhaisessa vaiheessa. Toinen esimerkki pituuskasvun tutkimuksesta on Saarin ym. (2011) artikkeli heidän tutkimuksestaan. Tutkimuksessa esitellään uudet kasvuviitteet 0-20-vuotiaille lapsille ja nuorille. Tutkimuksessa huomattavaa on se, että 1-11.5-vuotiailla tytöillä ja 13-vuotiailla pojilla on selkeämpi tasainen kasvu keskimääräisessä pituudessaan vuosina 1983-2008 kuin vuosina 1959-1971 mitatuilla lapsilla ja nuorilla. Tässä työssä käytettyssä aineistossa lasten pituuskasvu on hyvin samansuuntaista kuin edellä mainituissa tutkimuksissa. Erityisesti Saarin ym. (2011) tutkimukseen verrattuna pituuskasvu on samansuuntaista, koska tässä työssä käytetty aineisto on kerätty lähellä samaa ajankohtaa kuin Saarin ym. (2011) tutkimukseen kerätty aineisto.

Työn rakenne koostuu aineiston esittelystä, menetelmien teoriasta ja käytännön tarkasteluista. Luvussa 2 esitellään työssä käytetty aineisto. Työssä on käytetty lapsen pituutta selitettävänä muuttujana. Useampaa lasta on mitattu usealla mittausajankohdalla, jota käytetään selittäjänä. Luvussa on myös havainnollistettu tyttöjen ja poikien pituuskasvua. Luvuissa 3 ja 4 käsitellään menetelmien teorioita. Luku 3 kattaa lyhyen katsauksen mixture regression teoriaan. Luvussa 4 esitellään toinen työssä käytetty menetelmä, sekamalli.

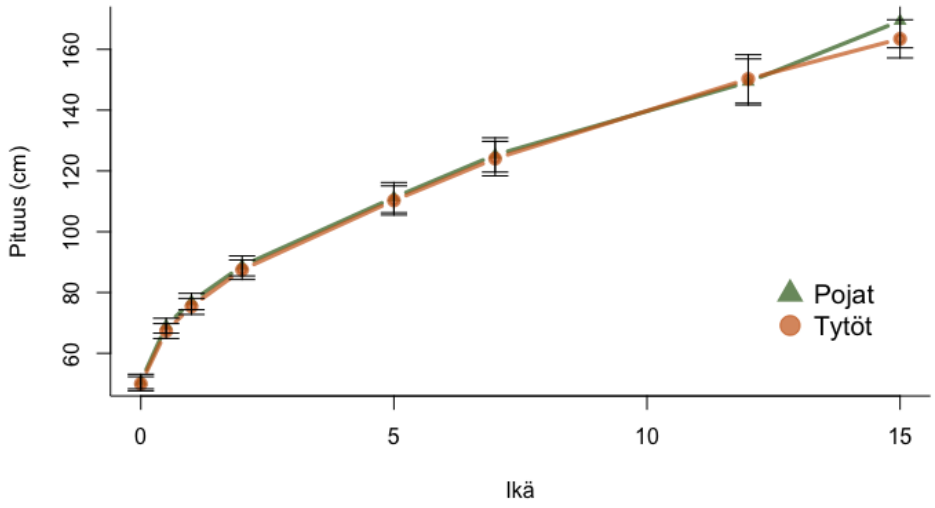
Luvussa 5 sovelletaan esitettyä aineistoa molemmilla menetelmillä. Ensin etsitään mixture regression avulla, kuinka moneen alipopulaatioon tyttöjen ja poikien aineisto voidaan jakaa. Tyttöjen aineiston alipopulaatioista esitellään residuaalikuvaajat, joiden avulla voidaan arvioida mallin istuvuutta aineistoon. Sekamallin käytännön osiossa avataan sekamalliin liittyviä tulosteita ja tarkastellaan mallin istuvuutta residuaalikuvaajien ja normaalisuusoletuksen avulla. Sekamalliin liittyen on vielä perusteltu, miksi työssä on käytetty juuri valittua mallia. Lopuksi on vielä vertailtu mixture regressiota ja sekamallia sekä sitä, kuinka hyvin malli on istunut aineistoon.

2 Aineisto

Työssä käytetty aineisto koostuu 4223 Pirkanmaan alueen lapsen mitatuista pituuksista ja painoista. Aineisto on alunperin kerätty Pirkanmaan alueelta. Suurin osa aineiston lapsista on Tampereelta, ja loput muista kunnista Pirkanmaalta. Tiedot ovat 1970-2000-luvuilta. Lapsia on mitattu useamman kerran eri ajankohtina syntymästä 15 ikään asti. Vuosina 1974, 1981, 1991 ja 1995 syntyneitä on seurattu 15 vuoden ikään saakka ja vuonna 2001 syntyneitä 11 ikävuoteen saakka.

Aineisto on jaettu kahteen osaan, jossa toisessa on pojat ja toisessa tytöt. Tyttöjen aineistossa mittauskertoja on yhteensä 15157 ja poikien aineistossa 17621. Mittauskertoja eli käytännössä koko aineiston havaintoja on yhteensä 32778.

Samaa aineistoa on käytetty artikkelissa *A trajectory analysis of body mass index for Finnish children* (Nummi ym. 2014). Tässä työssä vastemuuttujana käytetään lapsen pituutta. Vastemuuttujaa selitetään mittausajankohdalla eli lapsen iällä.



Kuva 2.1. Poikien ja tyttöjen keskimääräiset pituuskasvujen kehitykset hajontoineen

Kuten kuvassa 2.1 nähdään, mittausajankohdat ovat olleet 0, 0.5, 1, 2, 5, 7, 12 ja 15 vuoden iässä. Poikien ja tyttöjen keskimääräinen pituuden kasvun kehitys on hyvin samansuuntaista. Ainoa pieni ero on se, että pojat ovat 15 vuoden iässä keskimäärin hieman pidempiä kuin tytöt. Mittauspisteissä näkyvät janat kertovat, kuinka paljon hajontaa keskimääräisestä pituudesta on eri mittauspisteissä. Pidempi jana kertoo siitä, että pituuskasvussa on enemmän vaihtelua kyseisen mittauspisteen mittauskerroissa. Esimerkiksi kuvasta 2.1 nähdään, että 12- ja 15-vuotiailla pituus vaihtelee eniten.

Taulukko 2.1. Keskiarvot ja hajonnat keskimääräisestä pituudesta eri mittauspisteissä

		Ikä (mittauspisteet)							
		0	0.5	1	2	5	7	12	15
Pojat	Keskiarvo (cm)	50.7	69.1	77.0	88.7	111	125	149	169
	Hajonta (cm)	2.41	2.45	2.69	3.30	4.95	5.63	7.57	8.83
Tytöt	Keskiarvo (cm)	50.0	67.3	75.4	87.5	110	124	150	163
	Hajonta (cm)	2.33	2.45	2.64	3.21	4.79	5.65	7.96	6.28

Taulukossa 2.1 on vielä esitetty numeroin jokaisen mittauspisteen keskiarvo ja hajonta sukupuolittain. Taulukko vahvistaa sen, että hajonta on suurimmillaan 12- ja 15-vuoden iässä. Pojat ovat keskimäärin jokaisella mittauskerralla pidempiä paitsi 12-vuoden iässä tytöt ovat keskimäärin 1 cm pidempiä. Ero on kuitenkin hyvin pieni ja sukupuolien väliset pituuserot ovat muutenkin pieniä. Suurin pituusero onkin 15-vuoden iässä, jolloin pojat ovat keskimäärin 6 cm pidempiä kuin tytöt.

3 Mixture regressio

Mixture regressiossa jokaisella alipopulaatiolla on oma jakaumansa, joita mallinnetaan. Näiden alipopulaatioiden malleja kutsutaan mixture-malleiksi. Aineistosta pyritään etsimään näitä alipopulaatioita ja tutkimaan niiden jakaumia. Mixture-malleja käytetäänkin useilla eri tieteenaloilla, kuten biologiassa, lääketieteessä, fysiikassa, ekonomiassa ja markkinoinnissa (Leisch 2004, 1).

3.1 Määritelmä

Olkoon $\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n$ satunnaisotos, jossa on n havaintoa. Tässä $\boldsymbol{Y}_j$ on p-dimensiollinen satunnaisvektori tiheysfunktiolla $f(\boldsymbol{y}_j)$, joka on määritelty $\mathbb{R}^p$. Olkoon $\boldsymbol{Y} = (\boldsymbol{Y}_1^T, ..., \boldsymbol{Y}_n^T)^T$, missä T tarkoittaa transpoosia ja $\boldsymbol{Y}$ vastaa koko otosta.

McLachlanin ja Peelin (2000, 6) mukaan satunnaismuuttujien $\boldsymbol{Y_j}$ tiheysfunktio $f(\boldsymbol{y}_j)$ voidaan kirjoittaa muodossa

$$f(\boldsymbol{y}_j) = \sum_{i=1}^{K} \pi_i f_i(\boldsymbol{y}_j), \tag{3.1}$$

missä $f_i(\boldsymbol{y}_j)$ ovat tiheysfunktioita ja π_i ovat ei-negatiivisia ja summaantuvat ykköseksi;

$$0 \leq \pi_i \leq 1, \qquad (i = 1, ..., K), \tag{3.2}$$

$$\sum_{i=1}^{K} \pi_i = 1. \tag{3.3}$$

Suureita $\pi_1, ..., \pi_K$ kutsutaan sanalla *mixing proportion* tai painoiksi (*weights*). Koska funktiot $f_1(\boldsymbol{y}_j), ..., f_K(\boldsymbol{y}_j)$ ovat tiheyksiä, myös (3.1) määritellään tiheydeksi. Tiheyttä $f(\boldsymbol{y}_j)$ kutsutaankin mixturen komponentin tiheyksiksi (*component densities of the mixture*).

Mallinnettaessa aineistoa mixture-mallissa pitää määritellä komponenttien lukumäärä eli K (Frühwirth-Schnatter 2006, 24). Mixture-mallissa K (alipopulaatiot $g_1, ..., g_K$) määrittää alipopulaatioiden lukumäärän aineistossa. Usein K on tuntematon ja se täytyy päätellä aineistosta. Samalla estimoidaan painojen $\pi_1, ..., \pi_K$ osuudet ja tiheysfunktioiden $f_1(\boldsymbol{y}_j), ..., f_K(\boldsymbol{y}_j)$ parametrit.

3.2 Posterioritodennäköisyys

Jokaiselle yksilölle voidaan laskea posterioritodennäköisyys p_{ij}, mikä tarkoittaa todennäköisyyttä sille kuuluuko p_{ij} tiettyyn alipopulaatioon $g_1, ..., g_K$. Kun malli on

estimoitu, posterioritodennäköisyys voidaan laskea kaavalla

$$p_{ij} = \frac{\pi_i f_i(\boldsymbol{y}_j)}{f(\boldsymbol{y}_j)}. \tag{3.4}$$

Kaavan (3.4) avulla voidaan sitten esimerkiksi suurimman posteriori todennäköisyyden mukaan sijoittaa yksilöt yhteen alipopulaatioista $g_1, ..., g_K$.

3.3 Suurimman uskottavuuden menetelmä

Parametrien estimoimiseen voidaan käyttää log-uskottavuus funktiota (*log-likelihood function*)

$$logL(\boldsymbol{\theta}|\boldsymbol{y}_1, ..., \boldsymbol{y}_n) = \sum_{i=1}^{n} log f(\boldsymbol{y}_j), \tag{3.5}$$

jossa $\boldsymbol{y}_1, ..., \boldsymbol{y}_n$ ovat riippumattomia yksilöitä ja n on otoksen koko. Kaavassa (3.5) on käytetty suurimman uskottavuuden menetelmää (*maximum likelihood estimation, MLE*). Ideana on estimoida parametria $\boldsymbol{\theta}$. Suurimman uskottavuuden menetelmässä estimoidaan $\boldsymbol{\theta}$ etsimällä sellainen $\boldsymbol{\theta}$:n arvo, joka maksimoi uskottavuusfunktion. Ei kuitenkaan ole takuuta, että maksimit löytyvät aina (McLachlan & Peel 2000, 44), vaan algoritmi voi konvergoida joskus paikalliseen maksimiin.

4 Sekamalli

Sekamallin avulla voidaan esimerkiksi tarkastella, miten tietty ominaisuus muuttuu ajan myötä ja mallintaa sitä. Ongelmat, joita voi tulla, jos on puuttuvia havaintoja aineistossa, eivät tule esille sekamallien tilanteessa, koska puuttuvat havainnot voidaan olettaa satunnaiseksi (Brown & Prescott 2006, 24). Sekamalli sopiikin hyvin pitkittäisdatan analysointiin. Sekamallin etuna on kyky yhdistää data aidosti ottamalla käyttöön monitasoisia satunnaisia vaikutuksia (Demidenko 2013, 1). Sekamalleja sovelletaan useilla eri tieteenaloilla (esimerkiksi lääketiede ks. Brown & Prescott 2006).

4.1 Määritelmä

Sekamallin lähtökohtana on yleinen lineaarinen malli

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

jossa $\boldsymbol{y}$ on havaintovektori, $\boldsymbol{X}$ on suunnittelumatriisi ja $\boldsymbol{\epsilon}$ on virhetermi. Mallin kiinteä osa on $\boldsymbol{X\beta}$.

Olennainen ero yleiseen lineaariseen malliin on se, että sekamallissa on kiinteän osan lisäksi satunnaisosa. Sekamalli voidaan kirjoittaa muodossa

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}, \tag{4.2}$$

jossa $\boldsymbol{y}$, $\boldsymbol{X}$ ja $\boldsymbol{\epsilon}$ ovat samat kuin edellä. Mallin satunnaisosa on $\boldsymbol{Zu}$, joka tekee mallista nimenomaan sekamallin. Satunnaisosassa $\boldsymbol{Z}$ on suunnittelumatriisi satunnaisvaikutukselle $\boldsymbol{u}$.

Sekamallin vaikutuksille pätee

$$E(\boldsymbol{u}) = \boldsymbol{0} \qquad E(\boldsymbol{\epsilon}) = \boldsymbol{0}, \tag{4.3}$$

eli satunnaisvaikutuksen ja virhetermin odotusarvot oletetaan nolliksi. Voidaan myös olettaa, että $\boldsymbol{\epsilon}$ ja $\boldsymbol{u}$ ovat riippumattomia, ja siten

$$Cov(\boldsymbol{u}, \boldsymbol{\epsilon}) = \boldsymbol{0}. \tag{4.4}$$

Sekamallilla voidaan mallintaa dataa, jossa havainnot eivät ole riippumattomia (Brown & Prescott 2006, 22). Toisin sanoen sekamallit pystyvät mallintamaan aineiston kovarianssirakennetta. Voidaan määritellä, että $\mathrm{Var}(\boldsymbol{u}) = \boldsymbol{D}$ ja $\mathrm{Var}(\boldsymbol{\epsilon}) = \boldsymbol{R}$, jolloin $\boldsymbol{D}$ ja $\boldsymbol{R}$ ovat kovarianssimatriiseja. Havaintojen $\boldsymbol{y}$ kovarianssimatriisi on tällöin muotoa

$$Var(\boldsymbol{y}) = \boldsymbol{ZDZ'} + \boldsymbol{R}. \tag{4.5}$$

4.2 Parametrien estimointi

Kun $\boldsymbol{D}$ ja $\boldsymbol{R}$ tiedetään, voidaan estimoida parametrit $\boldsymbol{\beta}$ ja $\boldsymbol{u}$ sekamallien yhtälöiden avulla

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{D}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{pmatrix},$$

joista saadaan ratkaisuiksi

$$BLUE : \tilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y} \qquad BLUP : \tilde{\boldsymbol{u}} = \boldsymbol{D}\boldsymbol{Z}'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}).$$

Lyhenne BLUE tulee sanoista *best linear unbiased estimator*, jossa ideana on löytää paras $\tilde{\boldsymbol{\beta}}$, joka on lineaarinen ja harhaton estimaattori. Lyhenne BLUP tulee sanoista *best linear unbiased prediction*, missä ennustetaan parasta lineaarista ja harhatonta $\tilde{\boldsymbol{u}}$:ta.

5 Menetelmien soveltaminen käytännön aineistoon

Työssä käytetään R-ohjelmistoa menetelmien soveltamisessa käytännön aineistoon. Malli, jota käytetään molemmissa menetelmissä, on sama. Vastemuuttujana käytetään lapsen pituutta. Lapsen pituutta selitetään hänen iällään, sekä sen toisella ja kolmannella potenssilla. Identifioivana muuttujana käytetään yksittäistä lasta.

5.1 Alipopulaatioihin jakaminen mixture regressiolla

Mixture regressiossa valitaan klustereiden eli alipopulaatioiden lukumäärä. Alipopulaatioiden löytämiseksi voidaan muodostaa mixture regressioon liittyvä mixture-malli. R-ohjelmistosta löytyy siihen hyvin soveltuva stepFlexmix-funktio.

```
    Call:
stepFlexmix(height ~ ttime + I(ttime^2) + I(ttime^3) | child,
    data = growth_boys, k = 1:5, nrep = 10)

  iter converged k k0    logLik      AIC      BIC      ICL
1    2      TRUE 1  1 -56865.01 113740.0 113778.9 113778.9
2   12      TRUE 2  2 -54570.69 109163.4 109248.9 109556.5
3   17      TRUE 3  3 -53879.85 107793.7 107925.9 108468.9
4   27      TRUE 4  4 -53671.22 107388.4 107567.3 108343.5
5   61      TRUE 5  5 -53601.29 107260.6 107486.1 108499.4
```

Kuva 5.1. Poikien aineistoon sovitettu malli stepFlexmix-funktiolla.

Kuvassa 5.1 näkyy ensin aikaisemmin mainittu malli. Klustereiden lukumääriä on testattu yhdestä viiteen (*k=1:5*). Sama toistetaan 10 kertaa (*nrep = 10*). Kuvasta 5.1. nähdään myös, että hypoteettinen klustereiden lukumäärä (*k0*) on viisi, sama määrä, jonka algoritmi on pystynyt estimoimaan (*k*).

Suurimman uskottavuuden menetelmän avulla voidaan laskea AIC, BIC ja ICL, jotka ovat informaatiokriteerejä. Näitä tarkastelemalla voidaan päätellä, kuinka moneen alipopulaatioon otos on jakautunut. Yleisimmin näistä käytetty on BIC, jota tarkastellaan myös tässä. Huomataan, että pienin BIC:n arvo (107486.1) on viiden klusterin kohdalla. Tämän tarkastelun mukaan otoksessa olisi siis viisi alipopulaatiota.

Seuraavaksi tehdään analyysi tyttöjen aineistolle käyttämällä samaa mallia alipopulaatioiden löytämiseksi.

```
Call:
stepFlexmix(height ~ ttime + I(ttime^2) + I(ttime^3) | child,
    data = growth_girls, k = 1:5, nrep = 10)

  iter converged k k0    logLik      AIC      BIC      ICL
1    2      TRUE 1  1 -48304.82 96619.64 96657.77 96657.77
2    7      TRUE 2  2 -46633.25 93288.51 93372.39 93681.91
3   20      TRUE 3  3 -46180.89 92395.79 92525.43 93046.79
4   36      TRUE 4  4 -46045.35 92136.70 92312.11 93057.49
5   76      TRUE 5  5 -46012.85 92083.71 92304.87 93293.23
```

Kuva 5.2. Tyttöjen aineistoon sovitettu malli stepFlexmix-funktiolla.

Kuvan 5.2 analysoinnissa voidaan käyttää vastaavaa päättelyä kuin poikien aineistoon tehdyssä mixture regressiossa. Informaatiokriteerien tarkastelussa BIC:n pienin arvo olisi tässäkin tilanteessa viiden klusterin kohdalla. Yhtälailla kuin pojilla, tällä mallilla tyttöjen aineistossa olisi viisi alipopulaatiota.

Tarkastelun voisi vaihtoehtoisesti tehdä mallilla, jossa käytettään selittävinä muuttujina vain time-muuttujaa ja time-muuttujan toista potenssia, eli time-muuttujan kolmas potenssi jätetty pois. Tällä vaihtoehtoisella mallilla olisi saatu poikien sekä tyttöjen aineistossa neljän klusterin tilanne. Kolmannen asteen polynomin malli on kuitenkin parempi tähän aineistoon käytettynä, minkä takia tässä työssä on käytetty sitä mallina. Perustelu mallin valinnasta on esitetty luvussa 5.2.1.

Yhteenvetona voidaan siis päätellä, että molemmissa, poikien ja tyttöjen aineistoissa, on viisi alipopulaatiota.

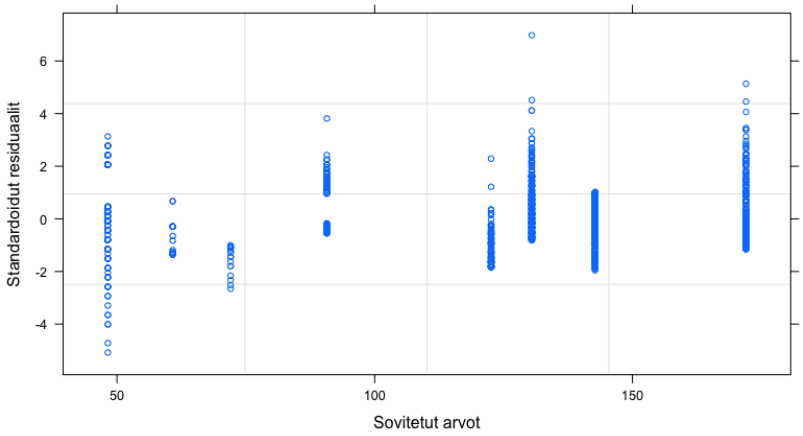
Seuraavaksi tyttöjen aineisto (vastaavasti voidaan tehdä myös poikien aineistolle) jaetaan viiteen alipopulaatioon. Jokainen mittauskerta sijoitetaan yhteen viidestä alipopulaatiosta.

Taulukko 5.1. Viiteen alipopulaatioon jaettu tyttöjen aineisto

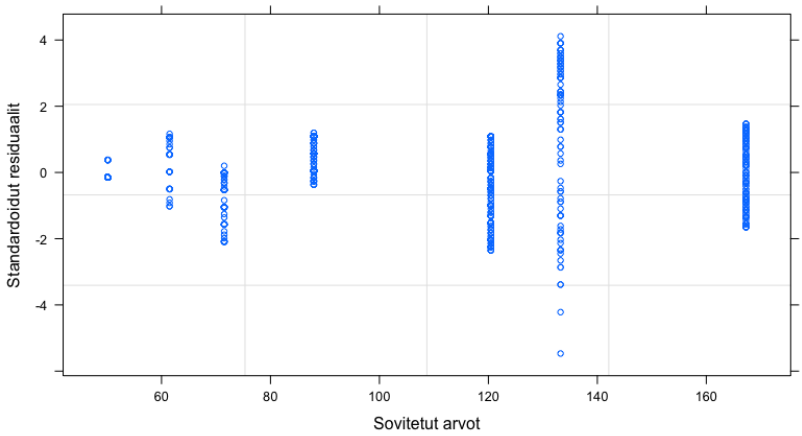
Ryhmä 1	Ryhmä 2	Ryhmä 3	Ryhmä 4	Ryhmä 5
2437	1673	1323	2438	7286
16%	11%	9%	16%	48%

Taulukosta 5.1 näkee jokaiseen ryhmään kuuluvien mittauskertojen lukumäärän. Esimerkiksi ryhmässä 1 on 2437 mittauskertaa. Mittauskertojen alapuolella on jokaisen ryhmän suhteellinen osuus. Taulukosta 5.1 nähdään, että melkein puolet havainnoista kuuluu ryhmään 5. Kuvassa 5.3 piirretään jokaisen ryhmän residuaalien kuvaajat.

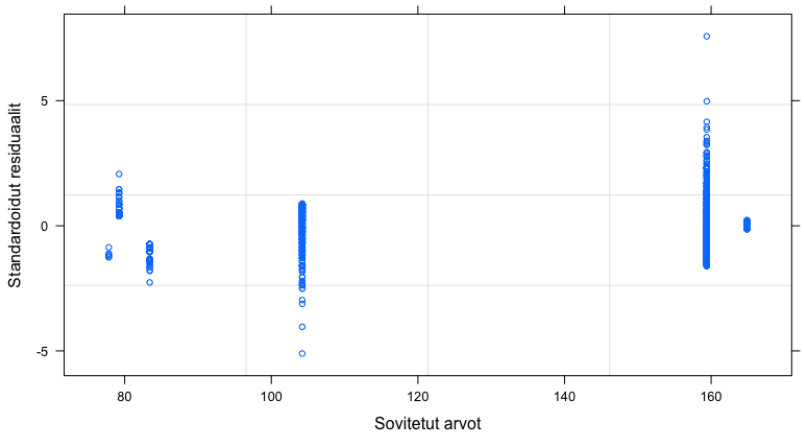
Ryhmä 1

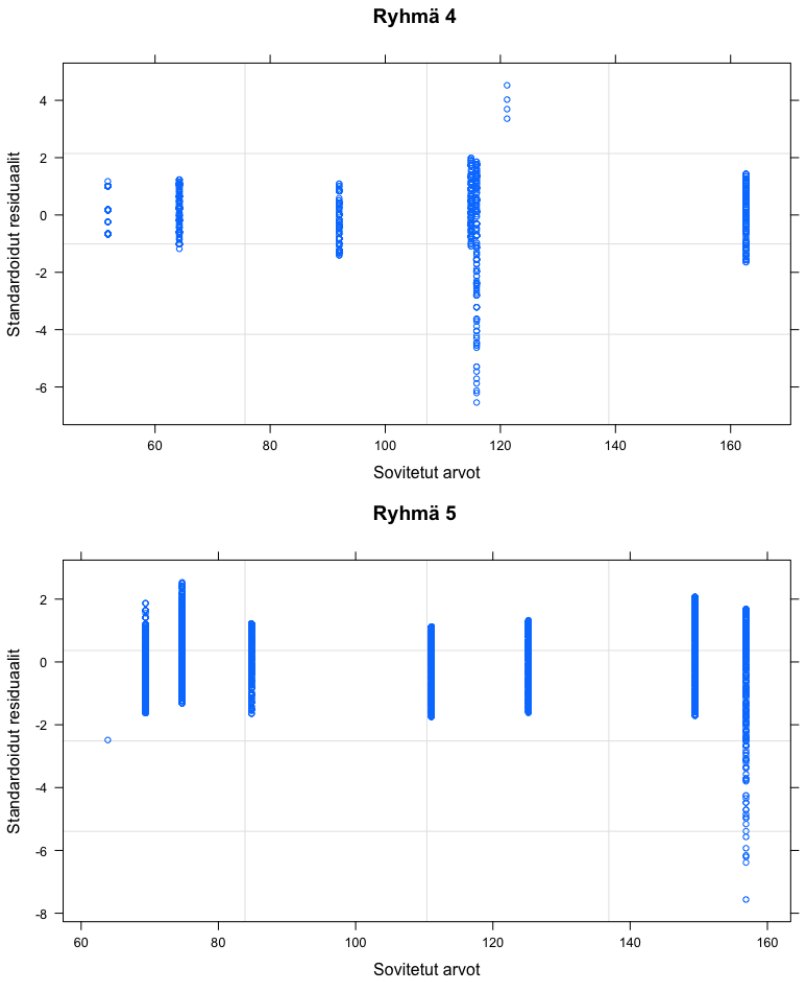


Ryhmä 2



Ryhmä 3





Kuva 5.3. Tyttöjen aineiston jokaisen ryhmän residuaalien jakautuneisuus. X-akselilla on sovitetut arvot ja y-akselilla standardoidut residuaalit.

Residuaalikuvaajista nähdään, että mittauskertoja on ollut yhdelle henkilölle enintään kahdeksan. Kuvaajista voidaan päätellä, kuinka hyvin asetettu malli sopii jokaisen ryhmän kohdalle.

Residuaalien perusteella ryhmässä 1 malli toimii paremmin myöhemmissä mittauskerroissa kuin ensimmäisellä mittauskerralla. Ryhmässä 2 kuudenteen mittauskertaan malli ei sovi kovin hyvin. Ryhmässä 3 mittauskertoja olisi nähtävästi vähemmän kuin muissa ryhmissä. Malli ei toimi varsinkaan ryhmän 4 neljännellä tai viidennellä mittauskerralla. Muissa kuin edellä mainituissa eri ryhmien mittauskerroissa mallissa on vain pieniä ongelmia.

Parhaiten malli sopii ryhmään 5, vaikka kylläkin kyseisen ryhmän viimeisessä mittauskerrassa malli ei toimi niin hyvin. Yleisesti malli sopii ihan hyvin eri ryhmiin, erityisesti ryhmään 5.

5.2 Sekamallin käytännön tarkastelu

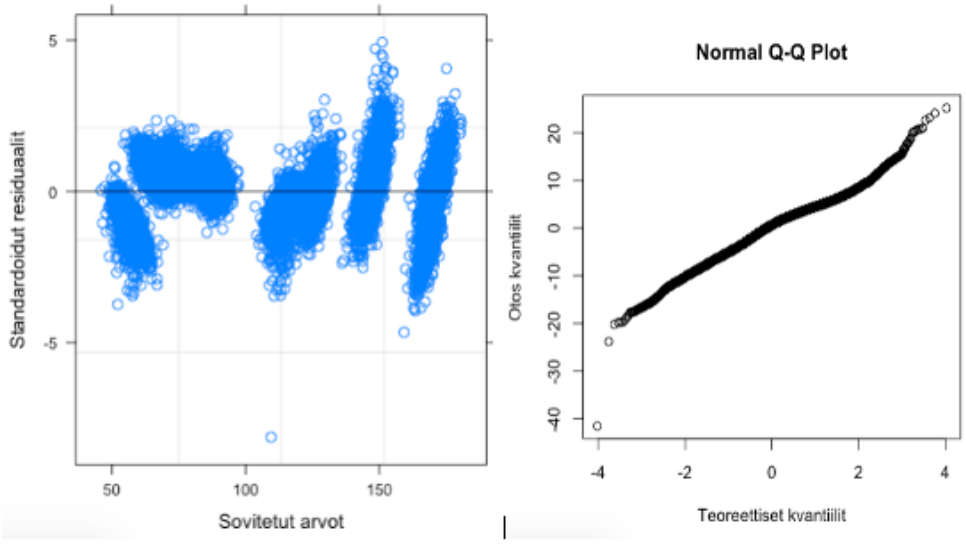
Seuraavassa R-tulosteessa on käytetty REML-menetelmää (*Restricted Maximum Likelihood*) lineaarisen sekamallin sovittamiseen. Vaihtoehtoinen tapa olisi ollut käyttää suurimman uskottavuuden menetelmää (MLE), mutta se ei ota huomioon vapausasteita (Nummi 2020, 151). REML-menetelmä onkin siksi useammin käytetty.

```
Linear mixed-effects model fit by REML
 Data: growth_boys
       AIC      BIC    logLik
  110789.9 110836.6 -55388.97

Random effects:
 Formula: ~1 | child
        (Intercept) Residual
StdDev:    3.335719 5.100814

Fixed effects: height ~ ttime + I(ttime^2) + I(ttime^3)
               Value  Std.Error    DF   t-value p-value
(Intercept) 56.94231 0.10358305 15348  549.7261       0
ttime       17.94988 0.06610768 15348  271.5249       0
I(ttime^2)  -1.57444 0.01183741 15348 -133.0057       0
I(ttime^3)   0.05880 0.00053968 15348  108.9595       0
 Correlation:
           (Intr) ttime  I(t^2)
ttime      -0.527
I(ttime^2)  0.425 -0.959
I(ttime^3) -0.371  0.904 -0.987

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-8.1381908 -0.6617645  0.1299068  0.6646950  4.9365515

Number of Observations: 17621
Number of Groups: 2270
```

Kuva 5.4. Sekamallin R-tuloste poikien aineistolle.

Ensinnäkin voidaan nähdä kuvan 5.4 satunnaisosan estimaatit (*Random effects*). Vakiotermin hajonta on 3.34 ja virhetermin hajonnan estimaatti 5.10 mallissa. Seuraavaksi kuvan 5.4. tulosteessa on mallin kiinteä osa (*Fixed effects*). Kiinteän osan vakiotermin estimaatti eli β_0 on 56.95. Aikamuuttujan estimaatti eli β_1 on 17.95. Aikamuuttujan toisen potenssin estimaatti eli β_2 on -1.57 ja kolmannen potenssin estimaatti eli β_3 on 0.06.

Mallin residuaalit vaikuttavat hyviltä, koska esimerkiksi mediaani on suhteellisen lähellä nollaa (0.13). Residuaaleja voidaan vielä tarkastella kuvaajan avulla (kuva 5.5).



Kuva 5.5. Vasemmalla poikien aineiston residuaalien jakautuneisuus. Oikealla poikien aineiston normaalisuusoletus.

Vasemmanpuoleisessa kuvaajassa nähdään, miten residuaalit ovat jakautuneet, kun sovitettuja arvoja verrataan suhteessa standardoituihin residuaaleihin. Residuaalikuvaajan pitäisi näyttää satunnaisesti jakautuneelta y-akselin nollakohdan ympärillä (Berridge & Crouchley 2011, 13). Vaikka residuaalit näyttävät ryhmittyneiltä, ne ovat kuitenkin suhteellisen symmetrisesti jakautuneita lukuunottamatta ensimmäistä mittausajankohtaa.

Vasemmanpuoleisimmassa residuaalipilvessä näkyy vastasyntyneiden pituuden kasvu, joka osuu hieman vaakatasossa olevaan viivaan (y-akselilla oleva kohta nolla). Yhden ikävuoden kohdalla viiva osuu jo vähän enemmän vaakatason viivaan. Jos siirrytään seuraaviin ikäluokkiin (residuaalipilviin), huomataan, että myöhemmässä iässä residuaalipilvet osuvat aina paremmin ja paremmin viivalle. Myöhemmissä iän vaiheissa residuaalipilvet kuvaavat siis hyvin kasvukehitystä.

Oikeanpuoleisessa kuvaajassa nähdään teoreettisten kvantiilien suhde otoskvantiileihin, missä tarkastellaan normaalisuusoletusta. Kuvaaja näyttää melko lineaariselta, joten voidaan päätellä, että aineisto olisi suurin piirtein normaalisti jakautunut.

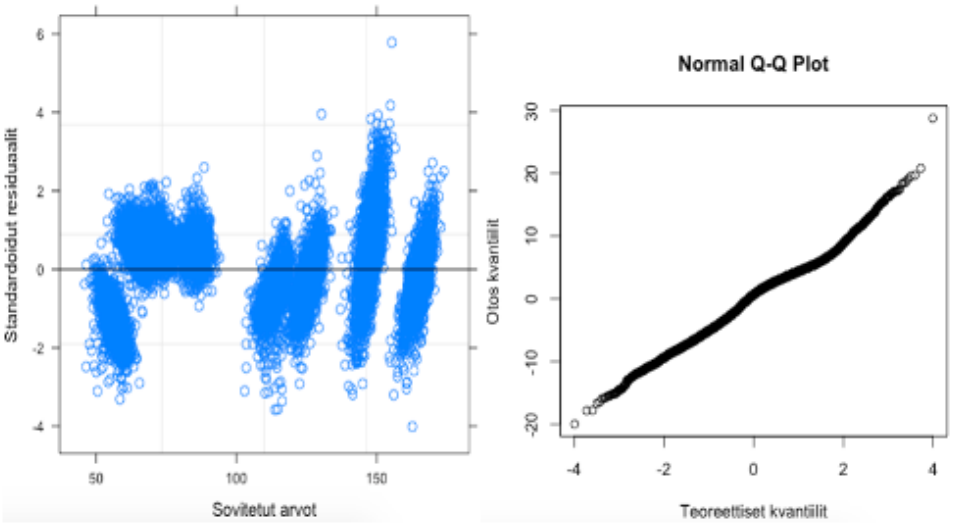
Saadaan samannäköinen R-tuloste kuin kuvassa 5.4, kun käytetään R-ohjelmiston lme-funktiota tyttöjen pituuden selittämiseen.

```
Linear mixed-effects model fit by REML
 Data: growth_girls
      AIC      BIC    logLik
  94354.9 94400.65 -47171.45

Random effects:
 Formula: ~1 | child
        (Intercept) Residual
StdDev:    3.109115 4.964431

Fixed effects: height ~ ttime + I(ttime^2) + I(ttime^3)
               Value  Std.Error    DF   t-value p-value
(Intercept) 56.27530 0.10657957 13200  528.0121       0
ttime       16.67593 0.06941571 13200  240.2327       0
I(ttime^2)  -1.26498 0.01243455 13200 -101.7308       0
I(ttime^3)   0.04239 0.00056696 13200   74.7631       0
 Correlation:
           (Intr) ttime  I(t^2)
ttime      -0.538
I(ttime^2)  0.434 -0.959
I(ttime^3) -0.378  0.904 -0.987

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-4.0197623 -0.7028587  0.1388788  0.6746184  5.7926465

Number of Observations: 15157
Number of Groups: 1954
```

Kuva 5.6. Sekamallin R-tuloste tyttöjen aineistolle.

Kuvasta 5.6 nähdään, että satunnaisosan vakiotermin hajonnaksi saadaan 3.11 ja virhetermin hajonnaksi 5.10.

Kiinteän osan vakiotermin estimaatti eli β_0 on 56.28 ja aikamuuttujan estimaatti eli β_1 on 16.68. Aikamuuttujan toisen potenssin estimaatti eli β_2 on -1.26 ja kolmannen potenssin estimaatti eli β_3 on 0.04. Kaikki luvut ovat hyvin lähellä poikien aineistoista saatuihin estimaatteihin.



Kuva 5.7. Vasemmalla tyttöjen aineiston residuaalien jakautuneisuus. Oikealla tyttöjen aineiston normaalisuusoletus.

Tyttöjen aineiston residuaaleja voidaan tarkastella kuvasta 5.7. Sekä residuaalien jakautuneisuus että normaalisuusoletus näyttää suhteellisen samanlaiselta kuin poikien aineistoon tehdyssä (kuva 5.5). Joten poikien aineiston residuaalitarkastelu ja normaalisuusoletukseen liittyvä päättely pätee myös tyttöjen aineistoon.

5.2.1 Perustelu mallin valinnasta

Tarkasteluissa käytettiin kolmannen asteen polynomin mallia. Vaihtoehtona olisi ollut käyttää mallia, jossa olisi toisen asteen polynomiin asti selittäjiä. R-ohjelmistolla voidaan vielä perustella, miksi tässä työssä on käytetty kolmannen asteen polynomin mallia.

```
     Model df      AIC      BIC    logLik   Test L.Ratio p-value
f_b      1  5 119552.0 119590.9 -59770.99
f2_b     2  6 110754.9 110801.6 -55371.46 1 vs 2 8799.06  <.0001
```

Kuva 5.8. Suppeamman ja laajemman mallin vertailu R-ohjelmistolla

Kuvan 5.8 tuloste saadaan R-ohjelmistolla asettamalla kaksi mallia anova-funktioon. On huomioitava vielä se, että tässä vertailussa molemmissa malleissa on käytetty ML-menetelmää REML-menetelmän sijasta. Tämä sen takia, koska REML vaatii mallien kiinteät osat samoiksi. ML-menetelmällä voidaan testata myös mallin kiinteää osaa, minkä takia se soveltuu tähän paremmin.

R-tulosteesta voidaan katsoa uskottavuusosamäärätestisuureeseen liittyvää p-arvoa, joka on merkitsevä (<.0001). Tämän perusteella suppea malli eli toisen asteen polynomin malli voidaan hylätä ja laajempi eli kolmannen asteen polynomin malli hyväksytään. Sama päätelmä voidaan tehdä informaatiokriteerien avulla, jotka ovat

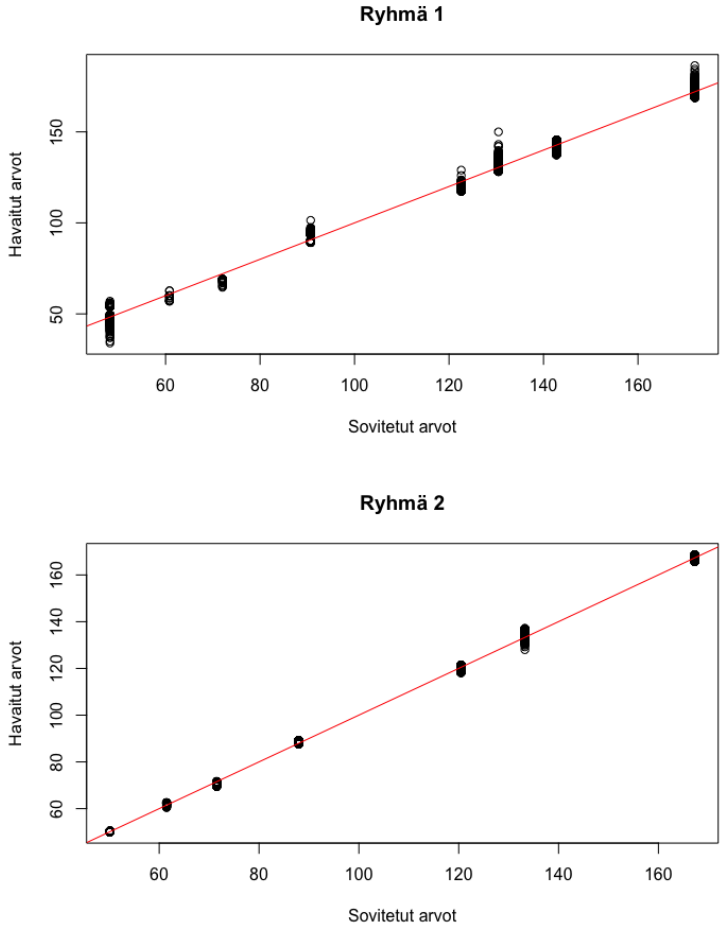
pienempiä laajemman mallin kohdalla. Tulosteessa on vertailtu malleja poikien aineistoon, mutta vastaavanlaiset tulokset oltaisiin saatu myös tyttöjen aineistolla.

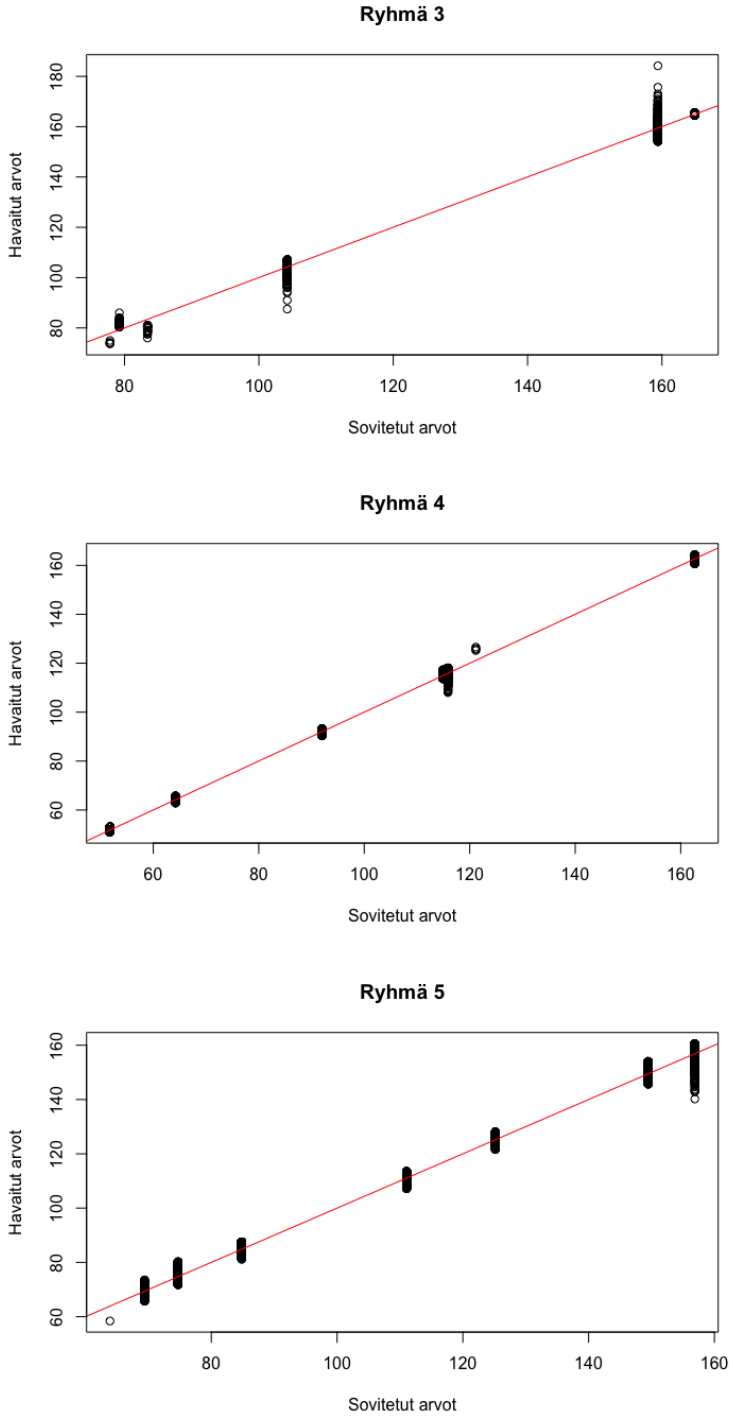
5.3 Vertailu

Mixture regressio ja sekamalli antavat molemmat havainnollistavia kuvaajia residuaalien jakautuneisuudesta. Mixture regressiossa voidaan tarkastella aineiston residuaaleja ryhmittäin ja sekamallissa voidaan tarkastella koko aineiston residuaaleja. Molempien menetelmien residuaalikuvaajista näkyy eri mittauskertoja ja kuinka hyvin niiden residuaalien mallinnus on onnistunut.

Mixture regressiossa riippuu, minkä ryhmän residuaaleja tarkastellaan, kun mietitään kuinka hyvin eri mittauskertojen residuaalit ovat jakautuneet. Sekamallissa sekä poikien että tyttöjen aineistossa malli sopii paremmin, kun tarkastellaan myöhempien mittauskertojen residuaaleja. Sekamallin avulla saadaan myös havainnollinen kuvaaja normaalisuusoletuksesta.

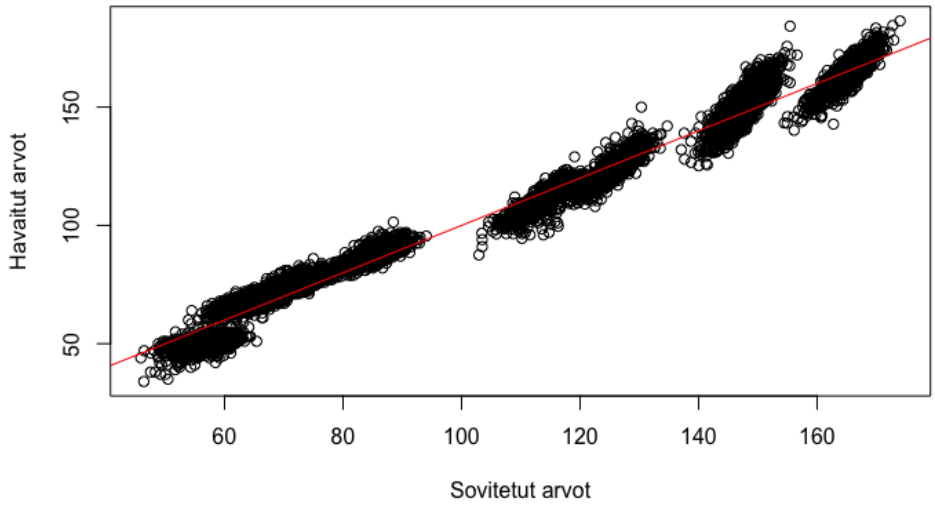
Lopuksi voidaan vielä kuvaajien avulla tarkastella mallin istuvuutta ensin mixture regression ja sitten sekamallin tilanteessa. Kuvassa 5.9 on aikaisemmin ryhmiin jaettu tyttöjen aineisto (edelleen yhtälailla oltaisiin voitu käyttää poikien aineistoa).





Kuva 5.9. Tyttöjen aineiston jokaisen ryhmän pistekaavio (*scatter plot*) sovitetuista ja havaituista arvoista

Jokaiseen ryhmään liittyen on tehty kuvaaja, josta näkee x-akselilla sovitetut arvot ja y-akselilla havaitut arvot. Kuvaajista nähdään, että parhaiten malli asettuu ryhmiin 2 ja 5. Ryhmässä 3 punainen viiva osuu huonoiten arvoihin. Kuitenkin muihin ryhmiin malli asettuu suhteellisen hyvin. Seuraavassa kuvaajassa nähdään vielä, miten malli asettuu aineistoon sekamallin tilanteessa, kun vertaillaan havaittuja ja sovitettuja arvoja tyttöjen aineistossa.



Kuva 5.10. Tyttöjen aineistosta tehty pistekaavio sovitetuista ja havaituista arvoista sekamallilla.

Kuvasta 5.10 nähdään, että malli asettuu hyvin aineistoon. Havaittujen arvojen ja sovitettujen (ennustettujen) arvojen korrelaatio on vahva. Kuvaajan oikeassa yläkulmassa havainnot ovat vähän enemmän laajemmalle jakautuneita, mutta ei niin paljon, etteikö malli istuisi hyvin aineistoon.

Molempien menetelmien kuvaajat antoivat hyvin tietoa siitä, että valittu malli istuu hyvin aineistoon. Kuten residuaalikuvaajissakin, eroja mixture regressiossa jaettuissa ryhmissä oli jonkin verran. Sekamalli antoi sen sijaan melko hyvät tulokset mallin istuvuudesta. Molemmat menetelmät ovat laajoja ja hyvin toimivia varsinkin pitkittäisdatan käsittelyssä. Mixture regression klusterointi antaa paremmat mahdollisuudet mallin yksityiskohtaisempaan tarkasteluun, jos aineisto halutaan jakaa alipopulaatioihin. Sekamalli toisaalta antaa hyvän kokonaistarkastelun mallin sopivuuden arvioimiseen.

Lähteet

Berridge, D. ja Crouchley R. (2011). *Multivariate Generalized Linear Mixed Models Using R*. CRC Press.

Brown, H. ja Prescott R. (2006). *Applied Mixed Models in Medicine*. 2. painos. Wiley.

Demidenko, E. (2013). *Mixed models: Theory and Applications with R*. 2. painos. Wiley.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.

Leisch, F. (2004). *FlexMix: A general framework for finite mixture models and latent glass regression in R*. Journal of Statistical Software, 11 (8), 1-18.

Nummi, T. (2020). *Multivariate analysis*, luentomoniste.
Tampereen yliopisto.

McLachlan, G. ja Peel D. (2000). *Finite mixture models*. Wiley.

Saari, A., Sankilampi, U., Hannila, M., Kiviniemi, V., Kesseli, K. ja Dunkel, L. (2011). *New Finnish growth references for children and adolescents aged 0 to 20 years: Length/height-for-age, weight-for-length/height, and body mass index-for-age*. Annals of medicine, 43 (3), 235-248.

Sorva, R., Tolppanen, E., Lankinen, S. ja Perheentupa, J. (1985). *Lasten kasvu ja sen arviointi*. Duodecim, 101 (5), 465-476.