PRO GRADU -TUTKIELMA

Tiia Friberg

Elinkaarialijäämän ajallinen kehitys sekamallin avulla tarkasteltuna

TAMPEREEN YLIOPISTO

Informaatiotieteiden yksikkö

Tilastotiede

Kesäkuu 2011

Tampereen yliopisto

Informaatiotieteiden yksikkö

FRIBERG TIIA: Elinkaarialijäämän ajallinen kehitys sekamallin avulla tarkasteltuna

Pro gradu -tutkielma 38 s., 3 liites.

Tilastotiede

Kesäkuu 2011

Tiivistelmä

Tulot ja kulutus ei jakaudu tasaisesti elinkaaren aikana. Nuorena ja vanhana ansiotuloja ei ole ja julkisen kulutuksen osuus on suurimmillaan. Nuoruuden ja vanhuuden välissä ollaan parhaassa työiässä, jolloin ansiotuloja kertyy ja vastaavasti julkisen kulutuksen osuus vähenee. Yksityistä kulutusta tapahtuu kaikissa ikäryhmissä. Kokonaiskulutuksesta vähennetään tulot, jolloin saadaan elinkaarialijäämä. Elinkaarialijäämässä on tapahtunut muutosta vuosien aikana, jota tässä tutkimuksessa mallinnetaan tasoittavalla kuutiosplinillä. Kuutiosplinin avulla muodostetaan aineistolle sekamalli. Sekamallin avulla testataan ajallista kehitystä seitsemäntoista vuoden aikana.

Ajallisen kehityksen tarkasteluun käytetään sekamallin kiinteän osan testaamista, jonka avulla voidaan testata ajan lineaarista vaikutusta elinkaarialijäämään, kulutukseen ja ansiotuloihin. Testauksessa käytettävä aineisto on tyypiltään kasvikäyräaineisto, joka on pitkittäisaineiston erikoistapaus. Aineisto on tällöin tasapainoinen; siitä ei puutu havaintoja ja mittaukset ovat samoina aikapisteinä suoritettuja.

Tutkielmassa esitellään tasoittavan kuutiosplinin teoriaa ja tarkastellaan tasoitusparametrin valinta eri valintakriteerien avulla. Esitellään myös lineaarisen sekamallin sekä tasoittavan kuutiosplinin välinen yhteys. Käydään lävitse myös sekamalliin perustuvan testauksen teoriaa, jota sitten lopuksi sovelletaan elinkaarialijäämäaineistoon.

Asiasanat kasvukäyrämalli, LCD, luonnollinen tasoittava kuutiosplini, tasoitusparametri, yleistetty ristiinvalidointi

Sisältö

1 Johdanto

1.1 Taustatietoa

Yksilön elinkaaren aikana kulutustarpeet ja näiden kulutustarpeiden rahoittaminen eivät ole tasaisesti jakautuneita. Kulutuksen ja tulojen eriaikaisuutta voidaan tasapainottaa säästämällä ja sijoittamalla varallisuutta tai lainaamalla ja luotottamalla. Toinen tapa tasoittaa kulutuksen ja ansioiden eriaikaisuutta ovat tulonsiirrot yli elinkaaren tai sukupolvien välillä. Sukupolvien välisiä taloudellisia suhteita voidaan arvioida tarkastelemalla ikäryhmien palkkatuloja ja kulutusta (Mason et al., 2009). Tavaroiden ja palveluiden kokonaiskulutusta ikäryhmässä a merkitään C(a) sekä ansiotuloja Y(a):lla. Määritellään elinkaarialijäämä (Life Cycle Deficit, LCD) seuraavasti:

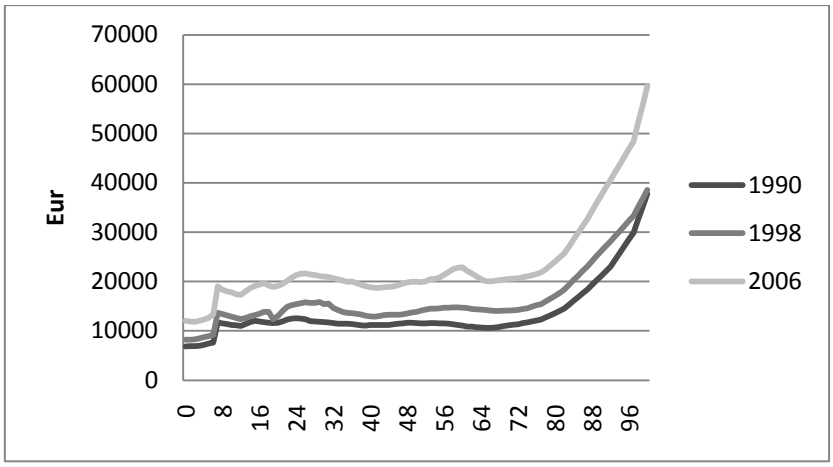
$$LCD(a) = C(a) - Y(a),$$

$$C(a) = C_j(a) + C_y(a),$$

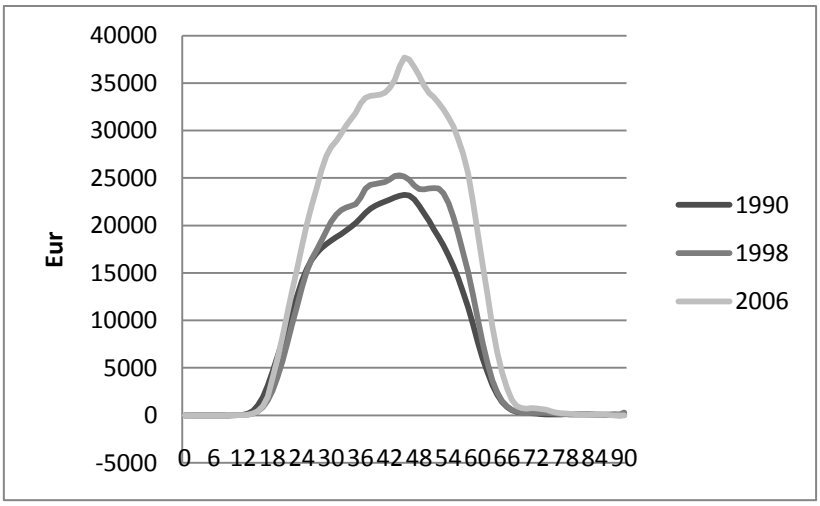
missä $C_j(a)$ on julkinen kulutus ja $C_y(a)$ on yksityinen kulutus.

Elinkaarialijäämää (LCD) sekä kansantalouden tilinpitoa (National Transfer Account, NTA) ovat tutkineet artikkeleissaan muun muassa Vaittinen & Vanne National Transfer Accounts for Finland in 2004 (2010), Riihelä, Vaittinen & Vanne Changing Patterns of Intergenerational Resource Allocation in Finland (2010), Lee, Lee & Mason Charting the Economic Life Cycle (2008).

Kuviossa 1.1 on kokonaiskulutus vuodessa jaettuna ikäluokkiin kuuluvalla väestöllä. Näin saadaan kulutus henkilöä kohti kolmena vertailuvuotena. Kuviossa 1.2 on vastaavasti vuoden kokonaisansiot jaettuna ikäluokkiin kuuluvalla väestömäärällä. Kuvion mukaan näyttäisi, että ansiotulot ovat kasvaneet eniten noin 30–55-vuotiaiden kohdalla. Kulutuksen ja palkkatulojen erotuksena saadaan elinkaarialijäämä, joka on tämän tutkimuksen pääkiinnostuksen kohde. Elinkaarialijäämä on nuoruudessa ja vanhuudessa positiivinen.



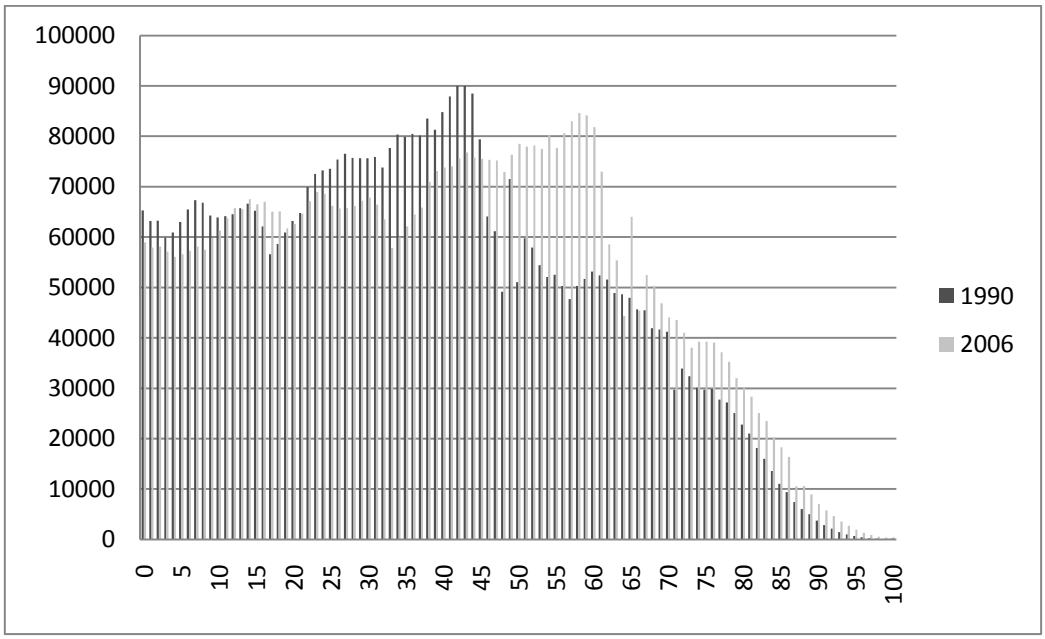
Kuvio 1.1. Kokonaiskulutus (miljoona eur) henkilöä kohti vuosina 1990, 1998, 2006



Kuvio 1.2. Ansiotulot (miljoona eur) henkilöä kohti vuosina 1990, 1998, 2006.

Suomessa tulonsiirtoja sukupolvien välillä leimaa kansainvälisesti vertailtuna kaksi erityispiirrettä. Julkisista menoista noin kaksi kolmasosaa on ikäriippuvaisia, mikä merkitsee noin 30 prosenttia Suomen bruttokansantuotteesta. Toinen erityispiirre on Suomen väestön rakenne eli meillä on poikkeuksellisen suuret sodan jälkeen syntyneet ikäluokat (kuvio 1.3). Väestön rakenteen muutoksen vuoksi on edetty jo siihen vaiheeseen, että työikäisen väestön suhteellinen osuus on pienenemässä, ja vuosina 2008–2013, kun eläkeiän saavuttavat suuret ikäluo-

kat, työikäisten osuus pienenee huomattavasti (Riihelä, Vaittinen & Vanne, 2010). Ikäryhmien kulutus- ja ansaintarakenteessa on tapahtunut muutosta. Muuttuneet kulutuksen ja tuotannon ikärakenteet ovat kasvattaneet elinkaarialijäämää koko talouden tasolla. Alijäämä suhteutettuna talouden palkkasummaan on kasvanut vuoden 1990 kolmesta prosentista vuoden 2006 seitsemääntoista prosenttiin (Riihelä, Vaittinen & Vanne, 2010). Väestön rakenteessa tapahtuneet muutokset selittävät kaksi viidesosaa alijäämän kehityksestä, vaikkakin vanhemman väestön parantunut työmarkkinatilanne on tasoittanut alijäämän kasvua. Yksityinen kulutus suhteessa työikäisten palkkoihin vaikuttaa eniten alijäämän kasvuun.



Kuvio 1.3. Suomen väestörakenne vuosina 1990, 2006.

1.2 Tavoitteet ja rakenne

Tässä tutkimuksessa päämielenkiinto on elinkaarialijäämän aikakehityksen tarkastelussa. Tavoitteena on kuvata ja testata aikakehityksen muutosta tilastollisella menetelmällä. Tässä tutkimuksessa käytettäviksi menetelmiksi aikakehityksen tarkasteluun on valittu tasoittava kuutiosplini sekä sekamalli, jota käytetään testauksessa. Mallinnetaan tasoittavalla kuutiosplinillä seitsemäntoista vuoden ajalta kerättyä aineistoa. Tarkastellaan myös erikseen elinkaarialijäämään vaikuttavia

ansiotuloja ja kulutusta, sekä sovitetaan myös näihin splinikäyriä. Käyrien sovituksen jälkeen muodostetaan aineistosta sekamalli. Sekamalli muodostetaan kasvukäyräaineistolle, pitkittäisaineiston erikoistapaukselle luodun mallin (GMANOVA) avulla, jota muokataan siten, että sekamallin satunnaisessa osassa hyödynnetään kuutiosplinejä. Tällä kuutiosplinien avulla muodostetun sekamallin kiinteällä osalla testataan lineaarista aikakehitystä.

Tutkimusaineisto esitellään luvussa kaksi, sekä esitellään muuttujat ja aineiston muokkaus. Kolmannessa luvussa esitellään tässä tutkimuksessa käytettyjen tutkimusmenetelmien teoriaa. Neljännessä luvussa teoriaa sovelletaan aineistoon, jota käsitellään kasvukäyräaineistona. Lopuksi tarkastellaan saatuja tutkimustuloksia.

2. Tutkimusaineisto

2.1 Aineiston esittely ja muuttujat

Tässä tutkimuksessa käytetty aineisto sisältää suomalaisten kulutus ja palkkatulotietoja seitsemäntoista (1990–2006) vuoden ajalta. Aineiston on saatu Eläketurvakeskus, jossa aineistoa on hyödynnetty elinkaarialijäämään liittyviin tutkimuksiin. Kulutus on jaettu sekä yksityiseen että julkiseen kulutukseen. Aineistossa selitettävät muuttujat ovat kokonaiskulutus, joka sisältää sekä yksityisen että julkisen kulutuksen, ja ansiotulot. Näiden erotuksena saadaan elinkaarialijäämä eli LCD (Life cycle deficit). Elinkaarialijäämää selitetään iällä (1–100) ja ajalla (vuodet 1990–2006).

Aineisto on kerätty viranomaisrekisteristä sekä Tilastokeskuksen toteuttamasta Kotitalouksien Kulutus – tutkimuksesta (http:// www.stat.fi/til/ktutk/). Tutkimus on suoritettu otantatutkimuksena kotitalouksille (noin 4000 kotitaloutta vuonna 2006). Kotitalouksiksi ovat laskettu kaikki kotimaiset kotitaloudet. Kaikkia tietoja kulutuksesta ei saada suoraan rekistereitä yksilötasolle vaan ne on laskettu kotitalouksittain, yrityksittäin tai julkisella tasolla. Luvut on sen jälkeen suhteutettu yksilöitä ja ikäryhmiä koskeviksi keskimääräisiksi luvuiksi.

Tutkimuksissa käytetyt tulotiedot on viranomaisrekistereistä saatuja palkkatuloja. Kotitalouksien tulot koostuvat eri tulolähteistä. Näitä ovat eläkkeet ja pääomatulot, joista taulukko löytyy liitteestä 2. Tässä tutkimuksessa tuloina käsitellään ainoastaan palkkatuloja, jotka ovat noin 60 – 85 % kotitalouksien tulorakenteesta vuosina 1990 – 2006 (liite 3). Yksityisen kulutuksen tiedot on pääasiassa saatu Tilastokeskuksen tekemästä Kotitalouksien Kulutus - tutkimuksesta. Tutkimusta on tehty viiden vuoden välein (vuosina 1990, 1995, 2001 ja 2006). Puuttuvien vuosien arvot on arvioitu näiden tutkimustulosten perusteella. Yksityinen kulutus sisältää seuraavat kulutuskohteet: 1) elintarvikkeet, 2) juomat ja tupakka, 3) vaatteet ja jalkineet, 4) asuminen ja energia, 5) kodin kalusteet ja koneet, 6) terveys, 7) liikenne, 8) tietoliikenne, 9) kulttuuri- ja vapaa-aika, 10) koulutus sekä 11) hotellit, ravintolat ja kahvilat. Julkiseen kulutukset tiedot on saatu viranomaisrekistereistä ja niihin lasketaan mukaan 1) koulutus, 2) sosiaalipalvelut sekä 3) terveyspalvelut.

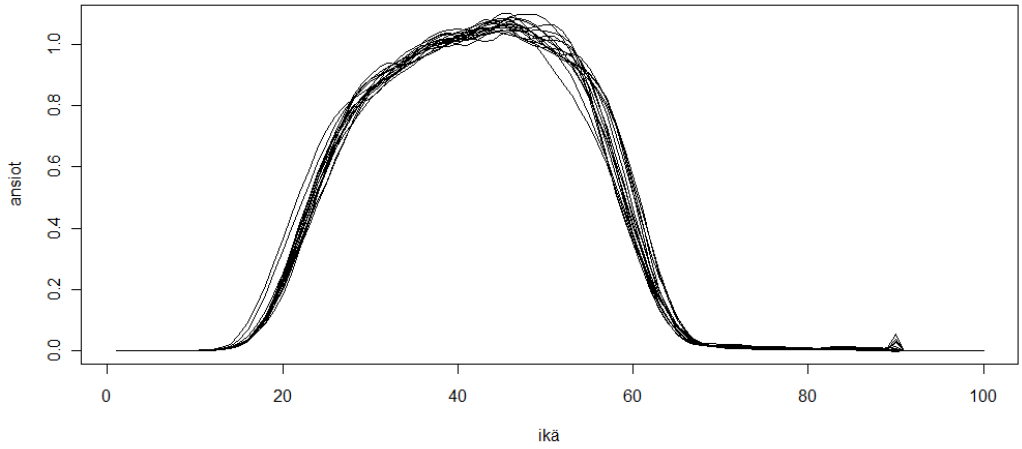
2.2 Aineiston muokkaus

Ajallisia vertailuja varten kulutukset ja ansiotulot on suhteutettu 30–49-vuotiaiden keskimääräisiin palkkoihin, jotta mahdollinen inflaatio sekä kasvanut työn tuottavuus on saatu rajattua pois ja näin saatu luvuista vertailukelpoisia keskenään. Jokainen selitettävänä oleva luku on vielä jaettu sitä vastaavalla henkilömäärällä, jotta on saatu muodostettua arvot henkilöä kohti. Aineistoissa on tiedot julkisesta ja yksityisestä kulutuksesta, ansiotuloista ja näiden erotuksena saaduista elinkaarialijäämistä. Aineistot on esitetty matriisimuodossa, missä n on vuosiluku ja q on ikäryhmä:

$$\begin{bmatrix} y_{11} & \cdots & y_{n1} \\ \vdots & \ddots & \vdots \\ y_{1q} & \cdots & y_{nq} \end{bmatrix}, \qquad n = 1, \dots, 17; q = 1, \dots, 100.$$

Ansiotulot

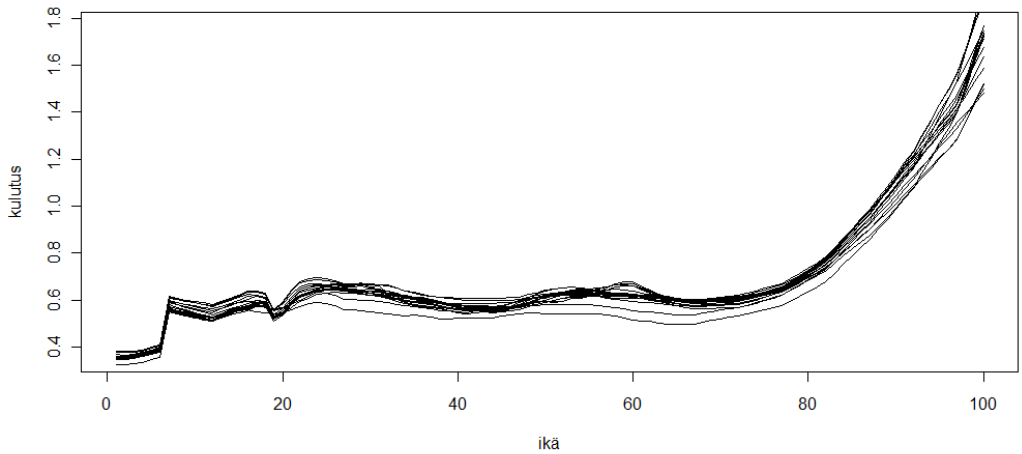
Jokaista seitsemäätoista vuotta kohti on määritelty ansiotulot ikäryhmittäin. Kun nämä kyseisen vuoden ikäryhmittäiset tulot jaetaan ikäryhmään kuuluvalla lukumäärällä, saadaan ansiotulot henkilöä kohti. Kun nämä henkilöä kohti lasketut tulot normeerataan vielä 30–49-vuotiaiden palkoilla, saadaan aineisto, joka on vertailukelpoinen eri ikäryhmien sekä vertailuvuosien välillä. Kun käytössä on normeerattu muokattu aineisto, vertailuluvut ovat väliltä (0,0-1,2). Kuviossa 2.1 on havaitut ansiotulot vuosilta 1990–2006.



Kuvio 2.1. Normeeratut ansiotulot henkilöä kohti vuosina 1990 – 2006.

Kulutus

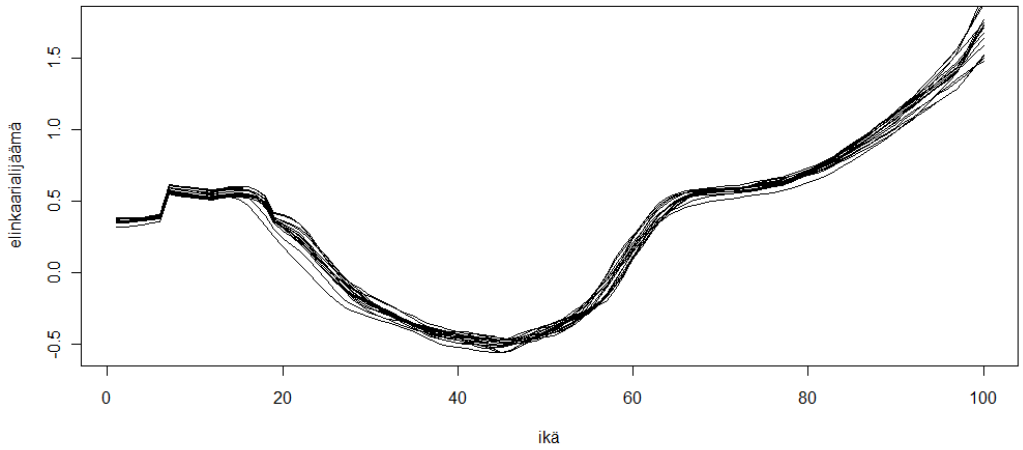
Kulutusaineisto on myös normeerattu ja jaettu kyseisen vuoden ja ikäryhmän väestömäärällä kuten ansiotuloaineistokin. Tällöin kulutusaineiston luvut ovat väliltä 0,0 – 2,0. Kokonaiskulutus jaetaan kahteen eri luokkaan: julkiseen ja yksityiseen kulutukseen. Julkisen kulutuksen aineiston luvut vaihtelevat välillä 0,0 – 1,7 ja yksityisen kulutuksen aineiston luvut välillä 0,0 – 0,6. Kuviossa 2.2 on kokonaiskulutus vuosilta 1990 - 2006. Huomataan, että kokonaiskulutus kasvaa huomattavasti väestön ikääntyessä ja ansiotulojen vähentyessä.



Kuvio 2.2. Normeerattu kokonaiskulutus henkilöä kohti vuosina 1990 – 2006.

Elinkaarialijäämä (LCD)

Aineistossa yhtenä selitettävänä muuttujana ja päämielenkiinnon kohteena on elinkaarialijäämä, joka saadaan vähentämällä kokonaiskulutuksesta ansiotulot. Tässä tutkimuksessa tutkitaan myös elinkaarialijäämää normeerattuna ja henkilöä kohti laskettuna. Tarkasteltavat luvut ovat väliltä -0,6 – 2,0. Kuviossa 2.3 on saadut elinkaarialijäämät vuosille 1990 – 2006.



Kuvio 2.3. Normeerattu elinkaarialijäämä henkilöä kohti vuosina 1990 – 2006.

3. Tutkimusmenetelmät

Kun parametriset menetelmät eivät ole riittävän joustavia kuvaamaan aineistoa, voidaan aineistoa lähteä mallintamaan parametrittomien menetelmien avulla. Toisin kuin parametrisissa menetelmissä parametrittomissa menetelmissä ei tehdä oletuksia käyrän parametreistä. Parametrittomissä malleissa voidaan olettaa esimerkiksi funktion kuuluvan johonkin funktioavaruuteen. Funktioavaruudesta voidaan sitten valita havaitun aineiston perusteella se käyrä, joka parhaiten kuvaa estimoitavaa käyrää. Jos mallin parametrinen muoto tunnetaan, on parametrinen menetelmä yleensä parametritonta tehokkaampi. Usein kuitenkaan mallin parametrinen muoto ei ole tunnettu, joten siitä joudutaan tekemään erilaisia oletuksia. Oletusten ollessa vääriä, voi parametriton menetelmä osoittautua tehokkaammaksi.

3.1 Tasoittavan kuutiosplinin sovittaminen

Tässä tutkielmassa odotusarvokäyrää mallinnetaan tasoittavan kuutiosplinin avulla, jolla saadaan arvoja tarkemmin myötäilevä käyrä kuin lineaarisilla malleilla. Tasoittavalla splinillä jokainen mittapiste on ns. solmukohta eli käyrä myötäilee aineiston mittauspisteitä tarkasti. Nyt kun jokainen mittauspiste on solmukohta, on käyrä hyvin epätasainen. Tätä epätasaisuutta säädellään ns. sakkotermin avulla, jonka ainoa määriteltävä parametri on tasoitusparametri α (Wu & Zhang, 2006). Tasoitusparametrin α valintaa käydään lävitse luvussa 3.1.3.

3.1.1 Tasoittava splini

Tarkastellaan epäparametrista mallia (3.1), jossa n on vertailtavien yksiköiden lukumäärä ja q_i on mittauspisteiden lukumäärä i:nessä yksikössä. Oletetaan, että mittauspisteet (solmukohdat) $x_1, x_2, \ldots, x_{qi}$ ovat välillä [a,b] siten, että $a < x_1 < \ldots < x_{qi} < b$. Oletetaan lisäksi, että havainnot y_{ij} toteuttavat yhtälön

$$y_{ij} = g_i(x_j) + \varepsilon_{ij}, \quad i = 1, \ldots, n; \quad j = 1, \ldots, q_{i,} \tag{3.1}$$

Missä virhetermit ε_{ij} ovat normaalijakautuneita odotusarvolla 0 ja kovarianssimatriisilla $\sigma^2\mathbf{R}_j$.

Käyrän $g_i(.)$ estimaattori $\hat{g}_i$ minimoi seuraavan pienimmän neliösumman kriteerin (penalized least squared (PLS) criterion):

$$\sum_{j=1}^{q}\left[y_{ij} - g_i(x_{ij})\right]^2 + \alpha \int_a^b \left[g_i{}^{(k)}(x)\right]^2 dx \qquad (3.2)$$

Ensimmäinen termi kuvaa sovitteen hyvyyttä (3.3). Nopeaa vaihtelua tasoitetaan rosoisuuden sakkotermillä (roughness penalty term) (3.4). Tasoitusparametrin α (> 0) avulla kontrolloidaan käyrän yhteensopivuutta aineistoon sekä rosoisuutta. Sen avulla pyritään löytämään tasapaino ensimmäisen ja toisen termin välille. Kun α on pieni, sovitettu käyrä myötäilee mittauspisteitä. Kun α kasvaa, käyrä lähenee suoraa.(Green & Silverman 1994.)

$$\sum_{j=1}^{q}\left[y_{ij} - g_i(x_{ij})\right]^2 \qquad (3.3)$$

$$\int_a^b \left[g_i{}^{(k)}(x)\right]^2 dx \qquad (3.4)$$

3.1.2 Luonnollinen kuutiosplini

Asettamalla k=2 pienimmän neliösumman kriteerissä (3.2), saadaan kuutiosplini, jolloin (3.2) voidaan lausua seuraavasti:

$$\sum_{j=1}^{q}\left[y_{ij} - g_i(x_{ij})\right]^2 + \alpha \int_a^b [g_i''(x)]^2 dx \qquad (3.5)$$

Kriteerin (3.5) minivoivalla estimaattorilla $\hat{g}_i(x)$ on seuraavat ominaisuudet:

1) Funktio on kullakin väleillä (x_j, x_{j+1}), $j = 1,\ldots, q_i-1$ kolmannen asteen polynomi

2) Mittauspisteessä x_j käyrä itse ja sen kaksi seuraavaa derivaatta ovat jatkuvia.

3) Estimaattori $\hat{g}(x)$ on määrittelyalueensa ulkopuolella lineaarinen eli välillä$(-\infty, x_1)$ ja (x_{q_i}, ∞) toinen ja kolmas derivaatta on nolla.

Jokainen käyrä, joka toteuttaa yllä mainitut kaksi ensimmäistä ehtoa (1, 2) ovat kuutiosplinejä. Lisäksi välillä (x_j, x_{j+1}) määritelty funktio g on luonnollinen kuutiosplini, jos lisäksi ehto 3) toteutuu. (Reinsch, 1964).

Malli (3.1) voidaan esittää myös vektoriesityksenä

$$\mathbf{y}_i = \mathbf{g}_i + \boldsymbol{\varepsilon}_i, \tag{3.6}$$

missä $\mathbf{g}_i = (g_i(x_1), g_i(x_2),\ldots, g_i(x_{qi}))'$ ja $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2},\ldots, \varepsilon_{iq_i})'$ noudattaa $N(\mathbf{0}, \sigma^2\mathbf{R}_i)$-jakaumaa.

Spliniestimaattorin $\mathbf{g}_i$ löytämiseksi käytetään sakotettua logaritmoitua uskottavuusfunktiota (penalized log-likelihood function)

$$2l = -\log[\det(\sigma^2\mathbf{R}_i)] - \frac{1}{\sigma^2}\left[(\mathbf{y}_i - \mathbf{g}_i)'\mathbf{R}_i^{-1}(\mathbf{y}_i - \mathbf{g}_i) + \alpha\int\{g_i''(x)\}^2dx\right] \tag{3.7}$$

Valitaan $\mathbf{g}_i$ siten, että uskottavuusfunktio maksimoituu. (Nummi & Koskela, 2008). Valinta uskottavuusfunktion avulla on vastaava kuin valinta pienimmän neliösumman menetelmän avulla.

Tasoittava kuutiosplini voidaan määritellä matriisioperaatioin. Määritellään apumatriisien Δ ja Q avulla rosoisuusmatriisi (roughness matrix) $K = Q\Delta^{-1}Q'$, missä Q on q_i x $(q_i - 2)$ ja Δ on $(q_i - 2)$ x $(q_i - 2)$ matriisi. Matriisien Q ja Δ alkiot saavat arvon 0 lukuun ottamatta seuraavia alkioita:

$$Q_{i;kk} = \frac{1}{h_k}, \quad Q_{i;k+1,k} = -\left(\frac{1}{h_k} + \frac{1}{h_{k+1}}\right), \quad Q_{i;k+2,k} = \frac{1}{h_{k+1}}$$

$$\Delta_{i;k,k+1} = \Delta_{i;k+1,k} = \frac{h_{k+1}}{6}, \quad \Delta_{i;kk} = \frac{h_k + h_{k+1}}{3}$$

missä $h_j = x_{j+1} - x_j$, j = 1, 2,…, q-1 ja k = 1, 2, …, (q_i – 2). (Green&Silverman, 1994). Uskottavuusfunktiosta (3.7) saadaan tasoittava kuutiosplini

$$\tilde{\mathbf{g}}_\mathbf{i} = (\mathbf{R}_\mathrm{i}^{-1} + \alpha\mathbf{K}_\mathrm{i})^{-1}\mathbf{R}_\mathrm{i}^{-1}\mathbf{y}_\mathrm{i.} \tag{3.8}$$

Tämä voidaan esittää myös seuraavalla lausekkeella:

$$\tilde{\mathbf{g}}_\mathbf{i} = (\mathbf{I} + \alpha\mathbf{R}_\mathrm{i}\mathbf{Q}_\mathrm{i}\Delta_\mathrm{i}^{-1}\mathbf{Q}_\mathrm{i}^{'})^{-1}\mathbf{y}_\mathrm{i.} \tag{3.9}$$

Jos oletetaan, että $\mathbf{R}_\mathrm{i}$ toteuttaa ehdon $\mathbf{R}_\mathrm{i}\mathbf{Q}_\mathrm{i} = \mathbf{Q}_\mathrm{i}$, lauseke (3.9) yksinkertaistuu muotoon

$$\hat{\mathbf{g}}_\mathbf{i} = (\mathbf{I} + \alpha\mathbf{K}_\mathrm{i})^{-1}\mathbf{y}_\mathrm{i}, \tag{3.10}$$

missä

$$\mathbf{S}_\boldsymbol{\alpha} = (\mathbf{I} + \alpha\mathbf{K}_\mathrm{i})^{-1} \tag{3.11}$$

on tasoittavan kuutiosplinin tasoittajamatriisi (cubic smoothing spline smoothing matrix).

Tällaisia ehdon $\mathbf{R}_\mathrm{i}\mathbf{Q}_\mathrm{i} = \mathbf{Q}_\mathrm{i}$ toteuttavia rakenteita ovat riippumaton $\mathbf{R}_\mathrm{i} = \mathbf{I}$,, tasainen $\mathbf{R}_\mathrm{i} = \mathbf{I} + \sigma_\mathrm{d}^2\mathbf{1}\mathbf{1}'$ ja lineaarinen rakenne $\mathbf{R}_\mathrm{i} = \mathbf{I} + \mathbf{X}_\mathrm{i}\mathbf{D}_\mathrm{i}\mathbf{X}_\mathrm{i}'$, missä $\mathbf{X}_\mathrm{i}$ on sellainen q_i x 2 matriisi, että matriisin ensimmäinen sarake on ykkösiä ja toisessa on yksilön mittauspisteet (Nummi & Koskela, 2008).

3.1.3 Tasoitusparametrin α valinta

Tasoitusparametrin valinnalla on merkittävä rooli tasoittavan splinikäyrän sovituksessa. Tasoitusparametrin valintaperusteet voidaan karkeasti jakaa kahteen lähestymistapaan. Tasoitusparametri voidaan valita subjektiivisesti eli valitsemalla vapaasti tasoitusparametri ja tarkastelemalla splinikäyrän sopivuutta. Toinen lähestymistapa on käyttää niin sanottuja automaattisia valintamenetelmiä. Auto-

maattisista menetelmistä seuraavaksi tutustutaan lähemmin ristiinvalidontiin (Cross-validation, CV) sekä yleistettyyn rintiinvalidointiin (Generalized Cross-validation, GCV). Muita automaattisia valintamenetelmiä ovat muun muassa Akaiken informaatiokriteeri (Akaike Information Criterion, AIC), (Akaike 1973) ja Bayesin informaatiokriteeri (Bayesian Information Critetion, BIC), (Schwarz 1978). Pääsääntöisesti voi olla hyvä käyttää samaan valittua tasoitusparametria eri aineistoissa, jotta aineistojen vertaaminen olisi helpompaa tai vain, koska valittu parametri on koettu hyväksi aiempien kokemuksien kautta (Green & Silverman, 1994).

Vapausasteiden määrästä (df) voidaan päätellä, kuinka monta estimoitua parametriä (p) mallissa on (df = n-p). Splinimallin tapauksessa määritellään tehokas vapausasteiden lukumäärä tasoittajamatriisin $\mathbf{S}_\alpha$ (3.11) diagonaalielementtien s_{ii} summana

$$\mathrm{df}_\alpha = \mathrm{tr}(\mathbf{S}_\alpha) = \sum_{i=1}^{n} s_{ii}. \qquad (3.12)$$

Hyvä tasoitusparametrin α arvo saadaan, kun mallin monimutkaisuus sekä mallin sopivuus ovat tasapainossa.

Ristiinvalidointi (CV)

Ristiinvalidointi eli CV on laajasti käytetty menetelmä yksinkertaisuutensa vuoksi. Menetelmä ei suoranaisesti liity mallin hyvyyden ja monimutkaisuuden suhteeseen, mutta idea on vastaavanlainen (Wu & Zhang, 2006). Menetelmässä minimoidaan ristiinvalidoinnin pistemääräfunktiota

$$\mathrm{CV}(\alpha) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{f}_\alpha^{(-i)}(t_i))^2, \qquad (3.13)$$

missä $\hat{f}_\alpha^{(-i)}(t_i)$ on sovitettu arvo mittauspisteessä t_i, kun estimoinnissa on käytetty koko aineistoa lukuun ottamatta havaintoa (y_i), kun i = 1,…, n. CV-menetelmällä valittu α:n arvo se, joka minimoi funktion (3.13). Funktion laskenta vaatii jokaisen lokaalin sovitteen $\hat{f}_\alpha^{(-i)}(t_i)$, minkä vuoksi funktion laskenta on työlästä (Wu & Zhang, 2006). Voidaan kuitenkin osoittaa, että funktio yksinkertaistuu, kun se määritellään seuraavasti:

$$\mathrm{CV}(\alpha) = n^{-1}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}_\alpha(t_i)}{1 - s_{ii}}\right\}^2 = n^{-1}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{y}_i}{1 - s_{ii}}\right\}^2, \qquad (3.14)$$

missä s_{ii} on i:nnes diagonaalielementti tasoittajamatriisista S_α. Kun funktio määritellään yllä olevalla tavalla (3.14), tarvitaan vain yksi lokaalinen sovite, jonka avulla voidaan määritellä CV(α) jokaiselle α:n arvolle (Wu & Zhang, 2006).

Yleistetty ristiinalidointi (GCV)

Yleistetty ristiinvalidointi (GCV) saadaan yksinkertaistetusta ristiinvalidoinnin (CV) tuloksesta korvaamalla jokainen s_{ii} niiden keskiarvolla

$$n^{-1}\sum_{i=1}^{n} s_{ii} = n^{-1}\mathrm{tr}(S_\alpha) = \frac{df}{n}.$$

Siis

$$\mathrm{GCV}(\alpha) = \frac{n^{-1}\sum_{i=1}^{n}[y_i - \hat{y}_i]^2}{\{1 - \mathrm{tr}(S_\alpha)/n\}^2} \qquad (3.15)$$

missä osoittaja kuvastaa mallin hyvyyttä ja nimittäjä mallin monimutkaisuutta. Valitaan α siten, että hyvyys ja monimutkaisuus ovat tasapainossa (Wu & Zhang, 2006).

3.2 Tasoittavan splinin yhteys sekamalleihin

Tasoittavalla splinillä on todettu olevan yhteys lineaariseen sekamalliin. Ensimmäisenä yhteyden esitti Speed (1991). Seuraavissa kappaleissa esitellään ensin lineaarinen sekamalli pääpiirteittäin, seuraavaksi sekamallin kiinteiden sekä satunnaisten vaikutusten estimointi ja lopuksi tasoittavan splinin yhteys sekamalleihin kasvukäyräaineiston tilanteessa.

3.2.1 Lineaarinen sekamalli

Lineaarinen malli määritellään seuraavasti:

$$y = X\beta + \epsilon,$$

missä y (n x 1) on havaintovektori, X (n x p) on kiinteiden vaikutusten suunnittelumatriisi, β (p x 1) on kiinteiden vaikutusten parametrivektori ja ϵ (n x 1) on satunnaisvirheiden vektori. Lineaarinen sekamalli (LME) on lineaarisen mallin laajennus, jonka ensimmäisenä esittivät Harville (1976, 1977) ja Laird & Ware (1982). Lineaarinen sekamalli sisältää sekä kiinteitä että satunnaisia vaikutuksia. Lineaarisen sekamallin yleinen muoto määritellään seuraavasti:

$$y_i = X_i\beta + Z_i u_i + \epsilon_i \qquad (3.16)$$

missä Z_i (n x p) on satunnaisvaikutusten suunnittelumatriisi ja u_i (n x 1) satunnaisvaikutusten vektori. Satunnaisvirheet ϵ_i ja satunnaisvaikutukset u_i oletetaan normaalijakautuneiksi sekä keskenään riippumattomiksi, jolloin

$$\mathrm{Cov}(\mathrm{u}, \epsilon) = 0.$$

Havaintojen y_i kovarianssimatriisi on muotoa

$$\mathrm{Var}(\mathrm{y}) = \mathrm{ZDZ}' + \mathrm{R},$$

missä R ja D ovat kovarianssimatriiseja, Var(u) = D ja Var(ϵ) = R. Sekamallissa määritellään myös, että E(y) = Xβ, E(u) = 0 ja E(ϵ) = 0.

3.2.2 Sekamallin ja tasoittavan splinin yhteys

Tasoittavan kuutiosplinin ratkaisuyhtälö (3.8) voidaan ilmaista myös parhaana lineaarisena ja harhattoman ennusteena (Best Linear Unbiased Prediction (BLUP)) sekamallin estimaateille. Kiinteä osuus on suora ja satunnainen osuus kuvastaa spliniä (Nummi & Koskela, 2008). Jos $\mathbf{X}_i = (\mathbf{1}, \mathbf{x}_i)$, missä $\mathbf{x}_i = (x_1, x_2, \ldots, x_{qi})'$, $\mathbf{Z}_i = \mathbf{Q}_i(\mathbf{Q}_i'\mathbf{Q}_i)^{-1}$ ja $\mathbf{V}_i = \sigma^2(\mathbf{R}_i + \alpha^{-1}\mathbf{Z}_i\mathbf{\Delta}_i\mathbf{Z}_i')$, voidaan yhtälö (3.8) kirjoittaa seuraavasti:

$$\tilde{\mathrm{g}}_\mathrm{i} = \mathrm{X}_\mathrm{i}\tilde{\beta}_\mathrm{i} + \mathrm{Z}_\mathrm{i}\tilde{\mathrm{u}}_\mathrm{i}, \qquad (3.17)$$

missä

$$\tilde{\beta}_\mathrm{i} = (\mathrm{X}_\mathrm{i}'\mathrm{V}_\mathrm{i}^{-1}\mathrm{X}_\mathrm{i})^{-1}\mathrm{X}_\mathrm{i}'\mathrm{V}_\mathrm{i}^{-1}\mathrm{y}_\mathrm{i} \qquad (3.18)$$

ja

$$\tilde{u}_i = (Z_i' R_i^{-1} Z_i + \alpha \Delta_i^{-1})^{-1} Z_i' R_i^{-1} (y_i - X_i \tilde{\beta}_i) \quad (3.19)$$

Edellä kuvattuja estimaatteja (3.18) ja (3.19) vastaa sekamalli

$$y_i = X_i \beta_i + Z_i u_i + \epsilon_i, \quad (3.20)$$

Tasoitusparametri on varianssien suhde $\alpha = \sigma^2 / \sigma_{u_i}^2$. Yhtälössä (3.20) varianssi-kovarianssirakenne u_i:lle ja ϵ_i:lle on seuraavanlainen:

$$Var \begin{pmatrix} u_i \\ \epsilon_i \end{pmatrix} = \begin{pmatrix} \sigma_{u_i}^2 \Delta_i & 0 \\ 0 & \sigma^2 R_i \end{pmatrix}. \quad (3.21)$$

3.2.3 Sekamalli kasvukäyräaineistolle

Tässä kappaleessa esitellään sekamalli tasoittavan kuutiosplinin avulla kasvukäyräaineistolle. Tämän alaluvun esitykset pohjautuvat Nummi & Koskela (2008) artikkeliin. Kasvukäyräaineistolla tarkoitetaan aineistoa, jossa mittaukset on kaikille yksilöille suoritettu samoissa aikapisteissä sekä puuttuvia havaintoja ei sallita. Yksi tapa analysoida tällaista kasvukäyräaineistoa on GMANOVA (Potthoff & Roy, 1964). Malli esitetään seuraavasti:

$$\mathbf{Y} = \mathbf{TBA}' + \mathbf{E} \quad (3.22)$$

missä $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n)$ on q x n havaintomatriisi, **T** (q x p) on yksilöiden sisäinen suunnittelumatriisi, ($p \leq q$, r(**T**) = p), **B** (p x m) on tuntemattomien parametrien matriisi, **A** (n x m) on yksilöiden välinen suunnittelumatriisi, (r(**A**) = m) ja **E** (q x n) on satunnaisvirheiden matriisi.

Mallia (3.22) vastaa tasoittaviin kuutiosplineihin perustuva malli

$$\mathbf{Y} = \mathbf{G}\mathbf{A}' + \mathbf{E} \tag{3.23}$$

missä $\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_m)$ on q x m splinikäyrienmatriisi. Matriisin **G** splinit voidaan estimoida maksimoimalla kaksinkertainen rankaiseva log-uskottavuusfunktio

$$2l = -\frac{1}{\sigma^2} tr[(\mathbf{Y}' - \mathbf{A}\mathbf{G}')\mathbf{R}^{-1}(\mathbf{Y}' - \mathbf{A}\mathbf{G}')' + \alpha(\mathbf{A}\mathbf{G}')\mathbf{K}(\mathbf{A}\mathbf{G}')']$$

$$- n\, log|\sigma^2\mathbf{R}|c \tag{3.24}$$

missä c = nq log(2π). Tuloksena saadaan ratkaisu

$$\widehat{\mathbf{G}} = (\mathbf{I} + \alpha\mathbf{K})^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1}, \tag{3.25}$$

jossa toteutuu aiemmin jo mainittu ehto **RQ** = **Q**.

Vektoroimalla yhtälö (3.25) saadaan $\hat{\mathbf{g}} = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\otimes(\mathbf{I} + \alpha\mathbf{K})^{-1}]\mathbf{y}$. Tasoittavan kuutiosplinin ja lineaarisen sekamallin välillä on todistettu olevan yhteys. Tämä yhteys on myös osoitettu kasvukäyräaineiston tilanteessa. Olkoon X q x 2 –matriisi, jonka ensimmäinen sarake on ykkösiä ja toisessa on mittauspisteet. Tällöin voidaan osoittaa, että

$$\hat{\mathbf{g}} = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\otimes\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y} + [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\otimes\mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \alpha\Delta^{-1})^{-1}\mathbf{Z}']\mathbf{y} \tag{3.26a}$$

$$= (\mathbf{I}_\mathbf{m} \otimes \mathbf{X})\widehat{\boldsymbol{\beta}}_* + (\mathbf{I}_\mathbf{m} \otimes \mathbf{Z})\hat{\mathbf{u}}_*, \tag{3.26b}$$

missä

$$\widehat{\boldsymbol{\beta}}_* = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\otimes(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\mathbf{y}$$

ja

$$\hat{\mathbf{u}}_* = [(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\otimes(\mathbf{Z}'\mathbf{Z} + \alpha\Delta^{-1})^{-1}\mathbf{Z}']\mathbf{y}.$$

Malli (3.23) voidaan uudelleen kirjoittaa seuraavasti:

$$\mathbf{y} = (\mathbf{A} \otimes \mathbf{I}_\mathbf{q})\mathbf{g} + \boldsymbol{\epsilon}_\mathrm{i} \tag{3.27}$$

missä g = vec(**G**) ja $\boldsymbol{\epsilon}$ = vec(**E**). Yhtälö (3.27) voidaan kirjoittaa sekamallina seuraavasti:

$$\mathbf{y} = \big(\mathbf{A} \otimes \mathbf{I}_\mathrm{q}\big)[(\mathbf{I}_\mathrm{m} \otimes \mathbf{X})\boldsymbol{\beta}_* + (\mathbf{I}_\mathrm{m} \otimes \mathbf{Z})\mathbf{u}_*] + \boldsymbol{\epsilon} \tag{3.28}$$

$$= (\mathbf{A} \otimes \mathbf{X})\boldsymbol{\beta}_* + (\mathbf{A} \otimes \mathbf{Z})\mathbf{u}_* + \boldsymbol{\epsilon},$$

missä $\mathbf{u}_* \sim \mathrm{N}\left(\mathbf{0}, \sigma_u^2 \tilde{\Delta}_*\right), \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{R}_*)$ ja $\tilde{\Delta}_* = [(\mathbf{A}'\mathbf{A})^{-1} \otimes \Delta], \mathbf{R}_* = (\mathbf{I}_n \otimes \mathbf{R})$.

Yllä esitetyn perusteella lineaarisella sekamallilla ja kuutiosplineillä on yhteys kasvukäyräaineiston tilanteessa. Malli (3.23) voidaan esittää sekamallimuodossa seuraavasti:

$$\mathbf{Y} = (\mathbf{X}\mathbf{B}_* + \mathbf{Z}\mathbf{U}_*)\mathbf{A}' + \mathbf{E}\,, \tag{3.29}$$

missä

$$\begin{pmatrix} \mathbf{u}_* \\ \boldsymbol{\varepsilon} \end{pmatrix} \sim \mathrm{N}\left(\mathbf{0}, \begin{pmatrix} \sigma_u^2 (\mathbf{A}'\mathbf{A})^{-1} \otimes \Delta & 0 \\ 0 & \mathbf{I}_n \otimes \sigma^2 \mathbf{R} \end{pmatrix}\right),$$

missä $\mathbf{u}_* = \mathrm{vec}(\mathbf{U}_*)$, $\varepsilon = \mathrm{vec}(\mathbf{E})$, $\mathbf{Z} = \mathbf{Q}(\mathbf{Q}'\mathbf{Q})^{-\mathbf{1}}$, $\mathbf{X} = (\mathbf{1}, \mathbf{x})$ ja $\mathbf{x} = (x_1, \ldots, x_q)'$. Tässä mallissa kiinteä osa $\mathbf{X}\mathbf{B}_*\mathbf{A}'$ on havaintoja kuvaavia suoria ja satunnainen osa kuvastaa splinikäyrien vaihtelua suorien ympärillä. Splinikäyrät lähestyvät lausekketta $\mathbf{X}\mathbf{B}_*\mathbf{A}'$, kun α lähestyy ääretöntä, vaikka α:n valinnalla ei ole suoranaista vaikutusta kiinteään osaan.

3.3 Hypoteesin testaus sekamallin avulla

Tässä tutkimuksessa sovelletaan Nummen ja Koskelan (2008) käyttämää lineaarisen hypoteesin testausta sekamallin kiinteän osan avulla kasvukäyräaineiston tilanteessa, josta kerrottiin tarkemmin luvussa 3.2.3.

3.3.1 Aikakehityksen testaaminen kasvukäyräaineistolle

Edellisessä kappaleessa esiteltiin sekamalli kasvukäyräaineiston tilanteessa, jossa satunnainen osa saadaan käyttämällä kuutiosplinejä ja kiinteä osa on lineaarinen. Kasvukäyräaineistolle on tyypillistä, että mittauksia on tehty säännöllisesti ajan kuluessa. Seuraavaksi esiteltävän testin ajatuksena on testata, onko aineistossa

tapahtunut muutosta ajan kuluessa. Testaus perustuu sekamallin kiinteään osaan, jota koskeva yleinen lineaarinen hypoteesi voidaan asettaa seuraavasti:

$$\mathrm{H}_0 : \mathbf{CB}_*\mathbf{L} = \mathbf{0}, \tag{3.30}$$

missä **C** on r x 2-matriisi ja **L** m x c-matriisi. Matriisi $\mathbf{B}_*$ voidaan estimoida kaavalla

$$\widehat{\mathbf{B}}_* = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \tag{3.31}$$

Nollahypoteesin vallitessa voidaan määritellä testisuure Q seuraavasti:

$$Q = \mathrm{tr}\left[\{\sigma^2\mathbf{C}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-1}\mathbf{C}'\}^{-1} \cdot \mathbf{C}\widehat{\mathbf{B}}_*\mathbf{L} \cdot \{\mathbf{L}'(\mathbf{A}'\mathbf{A})^{-1}\mathbf{L}\}^{-1} \cdot \left(\mathbf{C}\widehat{\mathbf{B}}_*\mathbf{L}\right)'\right] \sim \mathcal{X}^2_{cr} \tag{3.32}$$

Satunnaisjäännösten kovarianssimatriisi $\sigma^2\mathbf{R}$ on tuntematon ja pitää estimoida aineistosta. Tässä tutkimuksessa **R** oletetaan identiteettimatriisiksi.

4. Tutkimusmenetelmien soveltaminen aineistoon

Tässä luvussa sovitetaan edellä esiteltyjä menetelmiä tutkimuksen alussa esitettyyn aineistoon. Aineisto ei sisällä puuttuvia havaintoja ja sen mittauspisteet on mitattu samoina ajan hetkinä, joten aineistoa voidaan käsitellä kasvukäyräaineistona. Ensin mallinnetaan aineistoa splinifunktion avulla ja pyritään löytämään sopiva tasoitusparametri α. Tämän jälkeen sovitetaan aineistoon sekamalli splinejä hyödyntäen sekamallin satunnaisena vaikutuksena. Sekamallin muodostamisen jälkeen keskitytään lineaarisen aikakehityksen testaamiseen sekamallin kiinteän osan avulla.

4.1 Tasoittavan kuutiosplinin ja sekamallin sovitus

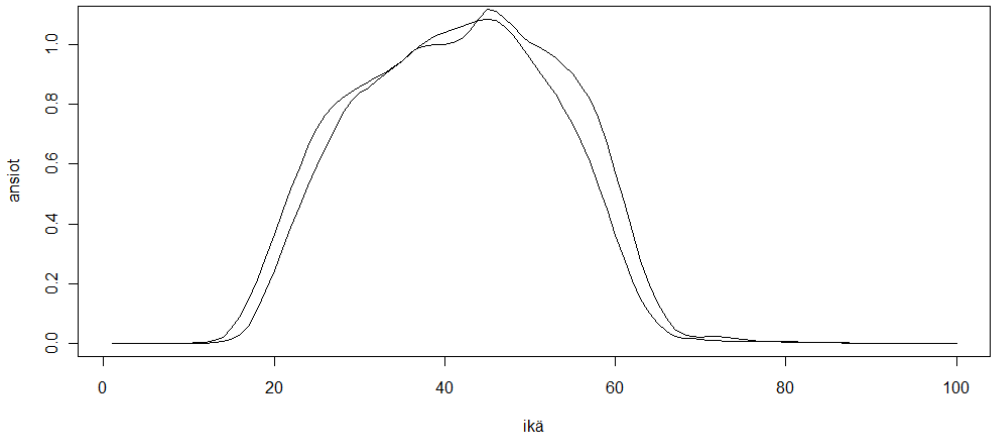
Menetelmien soveltamisessa käytetään kappaleessa kaksi esiteltyä elinkaarialijäämäaineistoa. Tarkastellaan mallien sopivuutta erikseen ansiotuloille, kulutuksille ja niiden erotuksena saaduille elinkaarialijäämille. Tässä tutkimuksessa käytetyt ja luvussa 3.2.3 määritellyt **X** ja **A** matriisit ovat seuraavanlaiset, missä q on ikä vuosina ja n on tarkasteltavat vuodet:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & q \end{bmatrix}, q = 1, 2, \dots, 100; \quad \mathbf{A} = \begin{bmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & n \end{bmatrix}, n = 1, 2, \dots, 17$$

4.1.1 Kuvallisia tarkasteluita

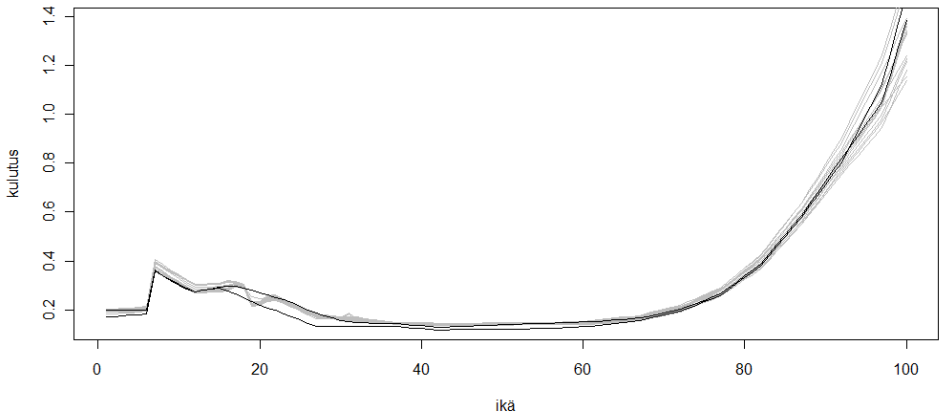
Kuviossa 4.1 on vertailtu ansiotulojen aikasarjan alku- ja loppupäitä eli vuosia 1990 ja 2006. Huomataan, että vuoden 1990 käyrä on alkaa aikaisemmassa vai-

heessa nousta ja laskea varhaisemmin kuin myöhemmät käyrät. Voidaan päätellä, että työurat alkavat nykyään myöhemmin kuin ennen, mutta myös päättyvät myöhemmin kuin aiemmin.



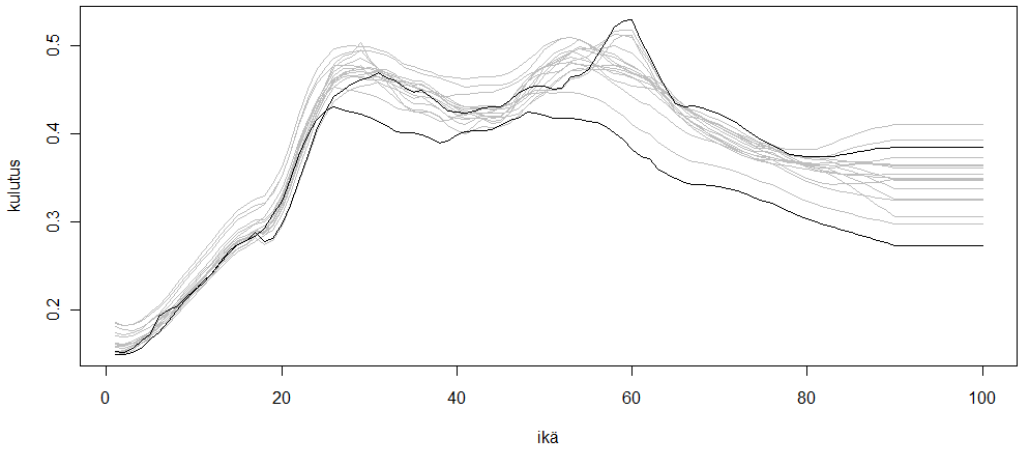
Kuvio 4.1. Normeeratut ansiotulot henkilöä kohti vuosina 1990 ja 2006. Vuoden 1990 käyrä alkaa nousta ja laskea aiemmin.

Kokonaiskulutus on jaettu erikseen julkiseen sekä yksityiseen kulutukseen. Kuviossa 4.2 on tarkasteltu julkista kulutusta ja aikasarjan alkuvuosi 1990 ja loppuvuosi 2006 on kuviossa mustilla viivoilla. Ylempi musta viiva kuvaa vuotta 2006. Huomataan, että julkinen kulutus ei ole juuri kasvanut. Noin 15 – 25-vuotiaiden kohdalla on havaittavissa hieman kasvua vertailuvuosien välillä. Vastaavasti ansiotuloissa on nähtävissä laskua 15 – 25-vuotiailla, kun verrataan vuotta 2006 vuoteen 1990.



Kuvio 4.2. Julkinen normeerattu kulutus henkilöä kohti. Ylempi musta viiva kuvaa vuoden 2006 ja alempi vuoden 1990 tilannetta.

Seuraavassa kuviossa 4.3 voidaan tarkastella yksityisen kulutuksen eroja aikasarjan aikana. Kuviosta huomataan selkeä ero aikasarjan alku- ja loppupuolen yksityisessä kulutuksessa. Selkein ero seitsemäntoista vuoden aikana on nähtävissä eläkeikäisten yksityisen kulutuksessa.

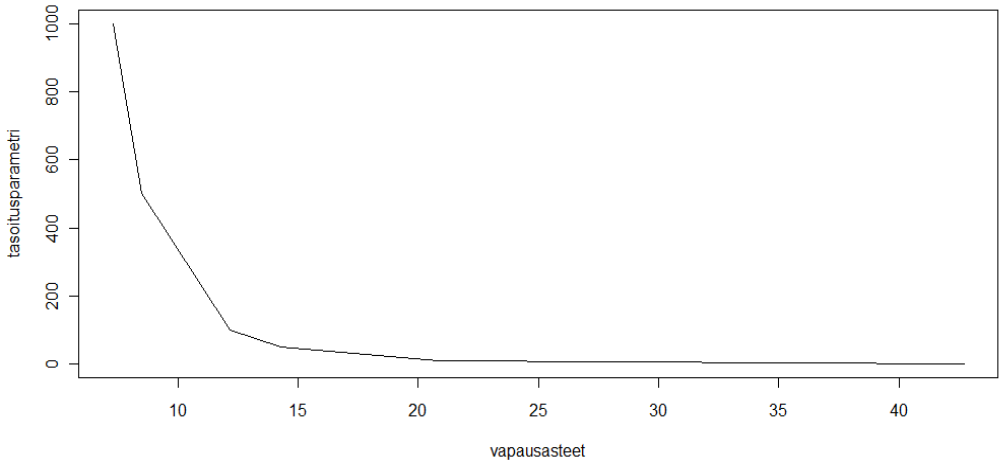


Kuvio 4.3. Normeerattu yksityinen kulutus henkilöä kohti. Musta ylempi viiva koskee vuotta 2006 ja alempi vuotta 1990.

4.1.2 Splini- sekä sekamallikäyrien sovitus aineistoon

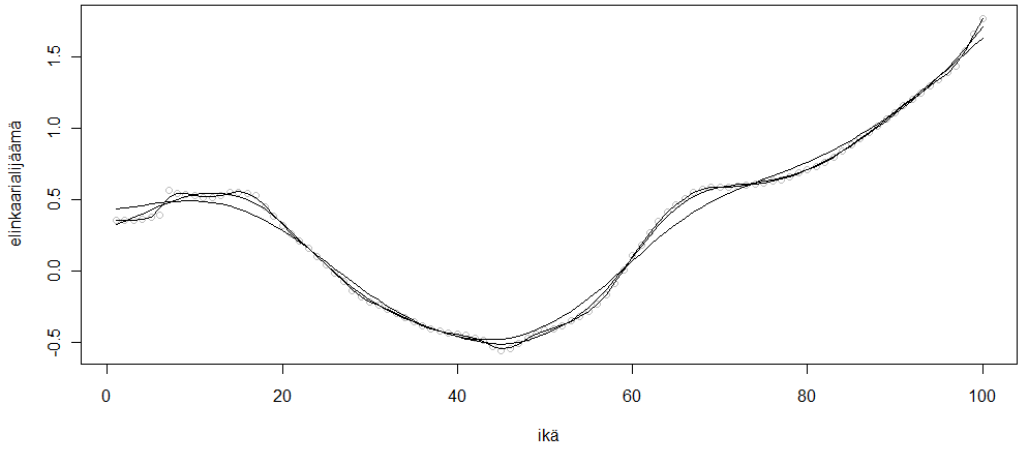
Tasoitusparametrin valinta

Seuraavaksi sovitetaan aineistoihin tasoittavia kuutiosplinejä. Splinien sovituksessa täytyy määrittää ensin tasoitusparametri α, joka määrää mallin tehokkaat vapausasteet (3.12). Kuviossa 4.4 on nähtävissä tasoitusparametrin sekä vapausasteiden riippuvuus tässä tutkimuksessa käytetyllä aineistolla. Kun malliin lisätään vapausasteita eli malli monimutkaistuu ja tasoittava splinikäyrä myötäilee tarkemmin havaintopisteitä, niin tasoitusparametri pienenee.

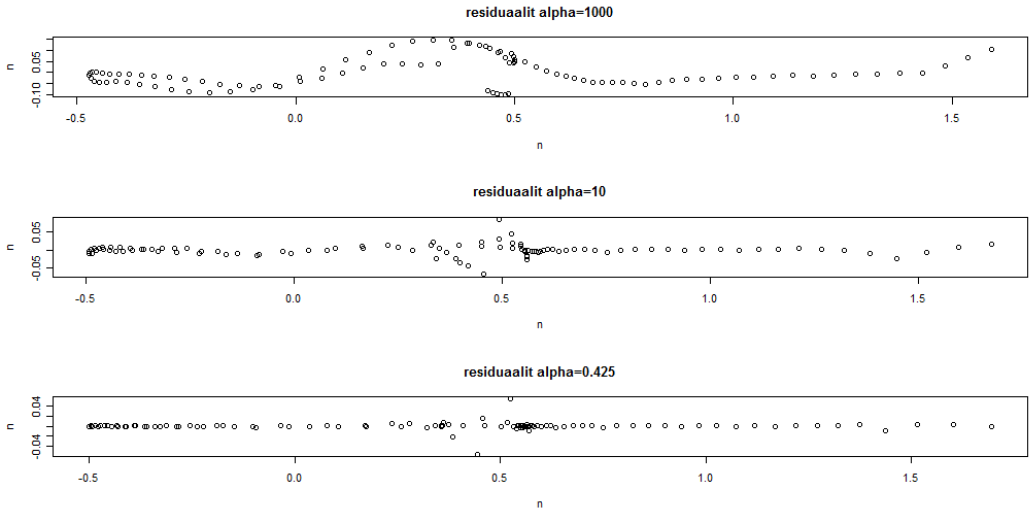


Kuva 4.4. Tasoitusparametrin sekä tehokkaiden vapausasteiden riippuvuus.

Tarkastellaan elinkaarialijäämäaineistoon sovitettuja splinikäyriä eri tasoitusparametreillä. Tasoitusparametrin valinnassa käytetään kappaleessa 3.1.3 esitettyä yleistettyä ristiinvalidointia (GCV) sekä kuvallisia tarkasteluita. Käytetään tasoitusparametrin määritykseen aineiston havaintojen keskiarvoja. Kuviossa 4.5 on sovitettu splinikäyrät kolmella eri tasoitusparametrillä (α = 1000, α = 10, α = 0.425). Tasoitusparametria α = 1 000 vastaava vapausasteluku (df) on 7.3 ja GCV:n arvoksi saadaan 0.004774. Vastaavasti tasoitusparametriä α = 10 vastaa vapausasteluku 20.9 ja GCV pienenee 0.000406:een. Kuviosta 4.5 voidaan huomata, että nyt splinikäyrä myötäilee paremmin havaintopisteitä. Yleistetyllä ristiinvalidointimenetelmällä tasoitusparametriksi valitaan 0.425, jolloin GCV on 0.000226 ja vapausasteluku 44.4. Tätä pienempi tasoitusparametrin arvo nostaa GCV:n arvoa sekä vapausasteita ja näin ollen malli tulee liian monimutkaiseksi mallin sopivuuteen nähden (Liite 1). Kuviossa 4.6 on kolmen eri mallin residuaalit. Residuaaleista huomataan, että pienimmän tasoitusparametrin avulla mallinnettu splini tarjoaa kaikista täsmällisimmän sopivuuden.



Kuvio 4.5. Keskiarvohavaintopisteen sekä kolme splinikäyrää eri tasoitusparametreillä (α = 1000, 10, 0.425).

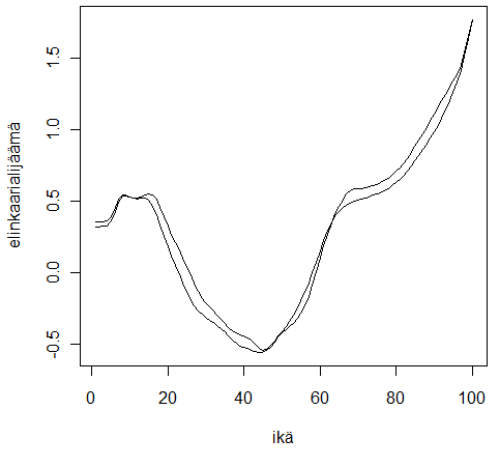


Kuvio 4.6. Residuaalit eri tasoitusparametreillä (α = 1000, 10, 0.425).

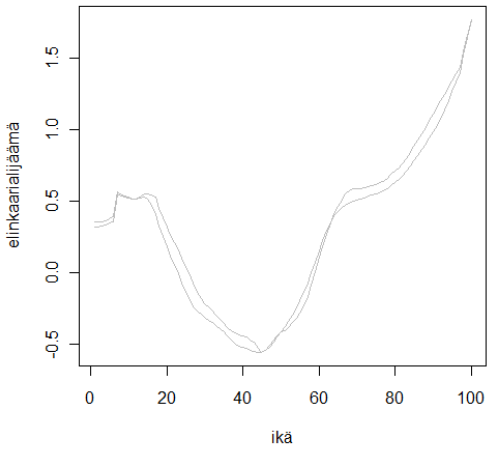
Tasoittavien kuutiosplinien ja sekamallien sovitus

Elinkaarialijäämäaineistoa mallinnetaan nyt tasoittavalla kuutiosplinillä, jonka tasoitusparametriksi valittiin 0.425. Kuviossa 4.7 on muodostetut splinit aikasarjan alku- ja loppupään vuosille. Ylempi käyrä on vuoden 2006 ja alempi vuoden 1990. Huomataan, että elinkaarialijäämä on suurempi noin 15–45-vuotialla sekä

yli 65-vuotiailla vuonna 2006. Tarkoittaen siis, että kulutusta on suhteessa tuloihin enemmän kuin vuonna 1990. Muodostetun tasoittavan kuutiosplinin avulla muodostetaan sekamalli (kuvio 4.8).

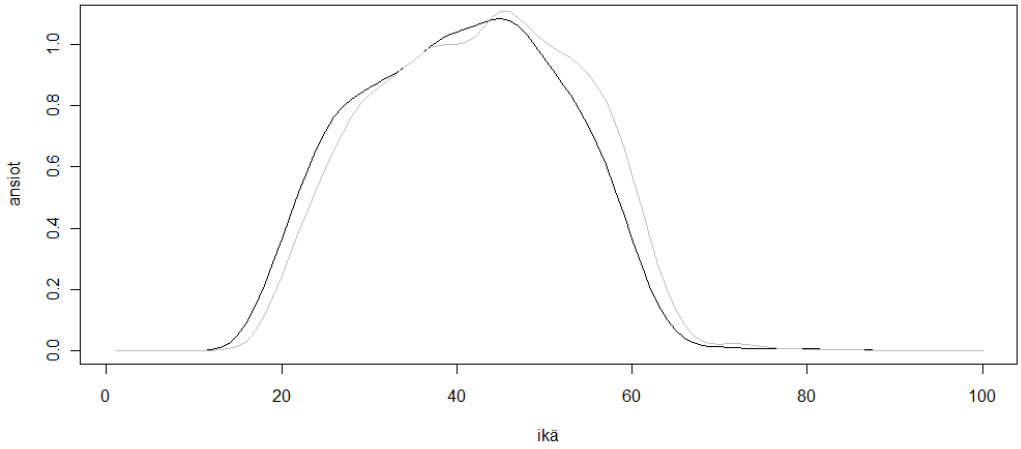


Kuvio 4.7. Tasoittavat splinit vuosille 1990 ja 2006.



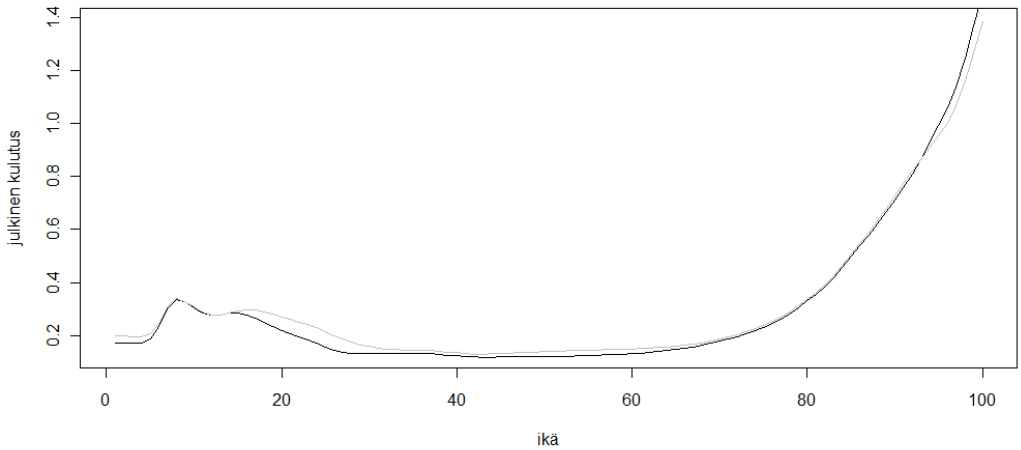
Kuvio 4.8. Sekamallin avulla lasketut sovitteet vuosille 1990 ja 2006.

Seuraavaksi tasoittavat kuutiosplinit sovitetaan ansiotuloaineistoon. Tasoitusparametrina käytetään samaa lukua 0.425, jotta aineistot ja niistä saatavat tulokset vastaavat toisiaan. Kuviossa 4.9 on sovitettu splinit vuosille 1990 ja 2006. Vuoden 2006 splinikäyrä on nuoremmilla alempana kuin vuoden 1990 splinikäyrä, jolloin siis vuoden 2006 ansiotulot ovat vuoden 1990 ansiotuloihin verrattuna pienemmät. Vastaavasti vuoden 2006 splinikäyrä on korkeammalla noin yli 45–vuotiailla kuin vuoden 1990 splinikäyrä.



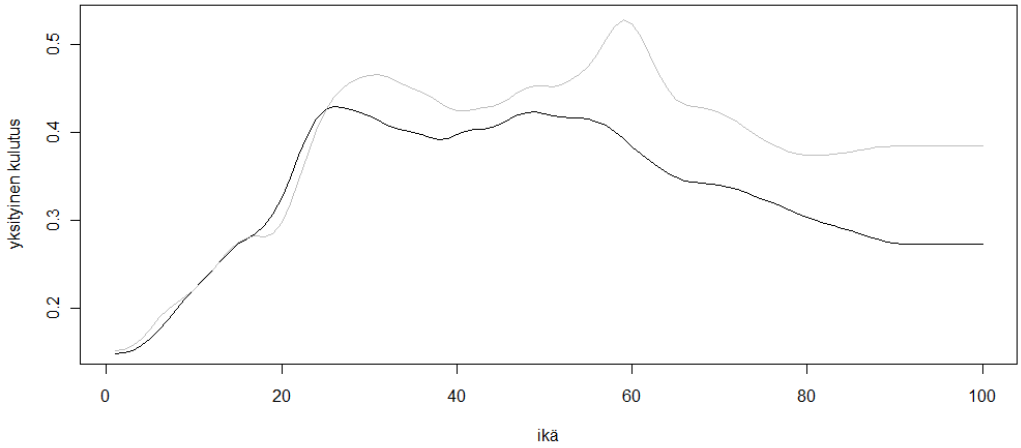
Kuvio 4.9. Tasoittavat kuutiosplinit vuosien 1990 (musta), 2006 (harmaa) ansiotuloille.

Mallinnetaan myös kulutusta splinien avulla. Tutkitaan tästä eteenpäin julkista sekä yksityistä kulutusta erikseen. Kuviossa 4.10 on piirretty splinisovitteet vertailuvuosien 1990 ja 2006 julkisen kulutuksen osalta. Huomataan, että julkisessa kulutuksessa ei suuria muutoksia näiden vertailuvuosien välillä ole. Tosin pientä kulutuksen kasvua on havaittavissa 15–30 vuotiaiden osalta ja laskua yli 90-vuotiaiden osalta.



Kuvio 4.10. Tasoittavat kuutiosplinit julkiselle kulutukselle vuosina 1990 (musta) ja 2006 (harmaa).

Seuraavana tarkastellaan vielä tasoittavia kuutiosplinejä yksityisen kulutuksen aineistolle. Yksityisessä kulutuksessa vuoden 2006 (harmaa) sekä vuoden 1990 (musta) käyrät eroavat selvästi toisistaan (kuvio 4.11). Selvin eroavaisuus vuosien välillä näkyy etenkin vanhemmalla väestöllä. Alle 25-vuotialla eroavaisuus vuosien välillä on jopa paikoitellen toisin päin eli yksityinen kulutus on pienentynyt.



Kuvio 4.11. Yksityisen kulutuksen tasoittavat kuutiosplinit vuosille 1990 (musta) sekä 2006 (harmaa).

4.2 Aikakehityksen testaaminen aineistolla

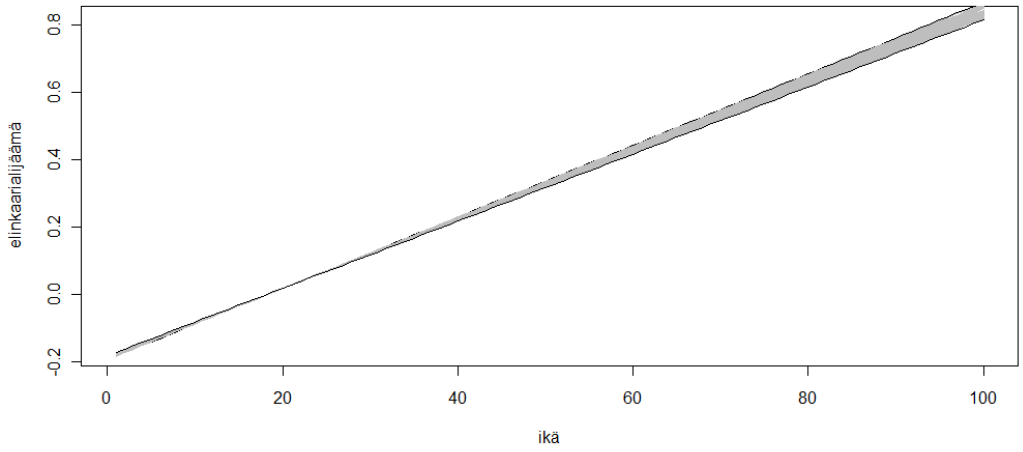
Tutkimuksen päätarkoituksena on tarkastella sekä testata aikakehitystä tilastollisin menetelmin. Aineistolla tutkittiin onko 17 vuoden aikana tapahtunut muutosta eli onko mallissa aika merkitsevä tekijä. Aikakehitystä tarkastellaan ja testataan lineaarisena kehityksenä luvussa 3.3 esitetyllä tavalla. Tutkittava hypoteesi on luvussa 3.3.1 esitetty (3.30). Matriisit **C**, **L** ja $\mathbf{B}_*$ ovat tälle aineistolle ja hypoteesille seuraavat:

$$\mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \mathbf{L} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \text{ja} \quad \mathbf{B}_* = \begin{bmatrix} b_{01} & b_{02} \\ b_{11} & b_{12} \end{bmatrix}.$$

Siis testataan hypoteesia $b_{02} = b_{12} = 0$.

4.2.1 Elinkaarialijäämäaineisto

Elinkaarialijäämäaineistosta muodostettiin sekamalli, jonka kiinteää osaa testataan. Siis testataan, onko elinkaarialijäämää kuvaavan käyrän lineaarisessa sovitteessa tapahtunut merkittävää lineaarista muutosta 17 vuoden aikana. Kuviosta 4.12 voidaan huomata, että elinkaarialijäämän lineaarinen sovite on kasvanut hieman seitsemäntoista vuoden aikana. Ylempi musta viiva kuvastaa vuotta 2006 ja alempi vuotta 1990 sekä harmaat vuosia 1991–2005. Lineaarinen sovite ei ole kasvanut kaikilla ikäryhmillä vaan eniten kasvua on tapahtunut vanhemmalla väestöllä, mikä johtuu yksityisen kulutuksen kasvusta.



Kuvio 4.12. Elinkaarialijäämän lineaarinen aikakehitys seitsemäntoista vuoden aikana.

Testataan aikakehitystä eli onko ajalla vaikutusta elinkaarialijäämäaineistosta muodostetussa mallissa. Kiinteän osan estimaateiksi saadaan

$$\widehat{\mathbf{B}}_* = \begin{bmatrix} -0.19663252 & 0.0008168679 \\ 0.01067231 & -0.0000407815 \end{bmatrix}$$

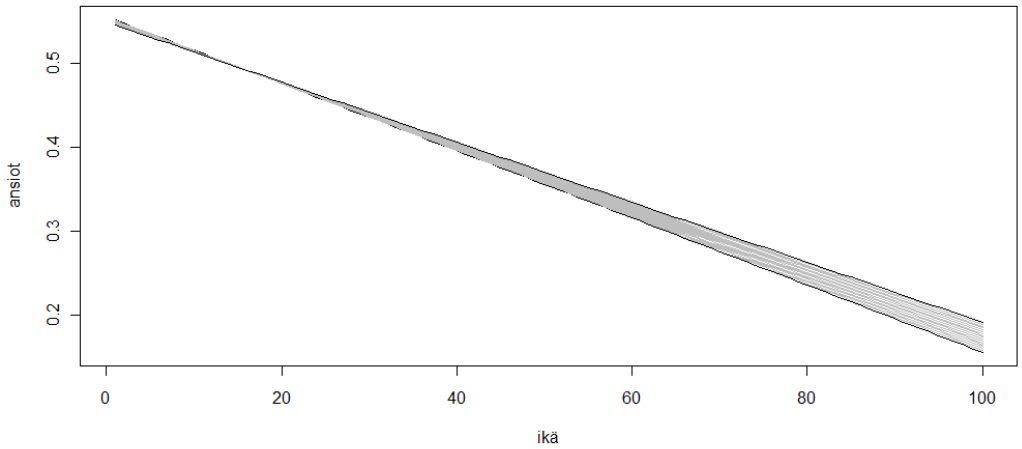
Testi noudattaa $\mathcal{X}^2$-jakaumaa vapausasteilla neljä. Tutkitaan aikakehityksen merkitsevyyttä 5 % merkitsevyystasolla. Tulokseksi saadaan $Q = 134.3512 > \mathcal{X}^2_{4;0.05} = 9.488$. Testin tulos on selvästi yli kriittisen tason ja näin ollen nollahypoteesi aikakehityksen merkitsemättömyydestä hylätään. Voidaan todeta, että ajallista kehitystä elinkaarialijäämässä on tapahtunut vanhemmalla väestöllä.

4.2.2 Ansioaineisto

Lineaarista aikakehitystä kuvaava estimaattimatriisi on ansiotulojen osalta

$$\widehat{\mathbf{B}}_* = \begin{bmatrix} 0.555930397 & -4.165587\text{e}-04 \\ -0.004020463 & 2.638058\text{e}-05 \end{bmatrix}.$$

Kuviossa 4.13 voidaan nähdä, että eroavaisuutta on ansiotuloissakin vanhemmalla väestöllä, kuten edellisessä elinkaarialijäämä tarkastelussa. Mustilla viivoilla on merkitty aikajakson alku- ja loppupää 1990 ja 2006 sekä harmailla vuodet 1991-2005. Testin tulokseksi saadaan nyt $Q = 152.8255 > \mathcal{X}^2_{4;0.05} = 9.488$, joka on hieman korkeampi kuin elinkaarialijäämässä. Ansiotulojen lineaarisella aikakehityksellä on merkitystä elinkaarialijäämän kehitykseen siten, että tulot ovat hieman kasvaneet vanhemmalla väestöllä. Tulojen kasvu johtunee pidentyneestä työiästä eli eläkkeelle siirrytään myöhemmin kuin ennen.



Kuvio 4.13. Ansiotulojen lineaarinen aikakehitys vuosina 1990–2006.

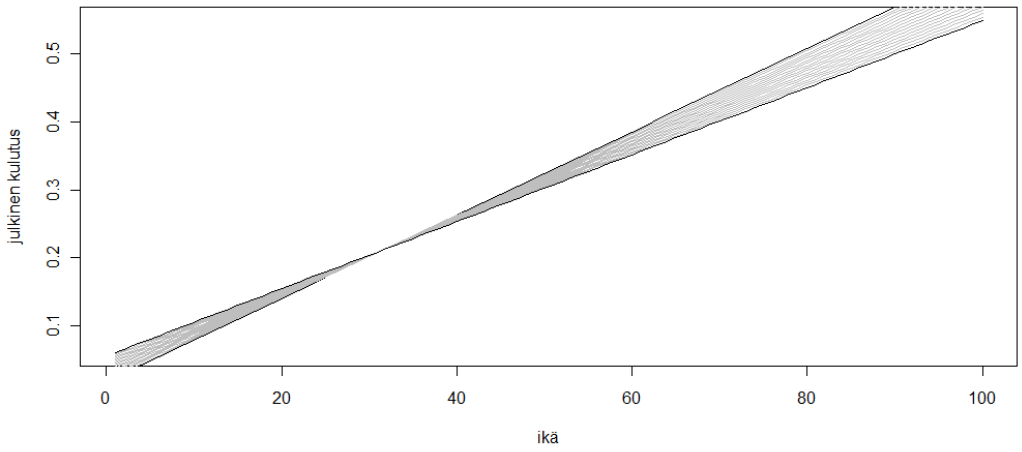
4.2.3 Kulutusaineistot

Julkinen kulutus

Seuraavaksi tutkitaan kulutusaineistojen aikakehityksen vaikutusta. Julkisen kulutuksen aikavaikutus voidaan nähdä kuviosta 4.14. Vuosi 2006 on ylimpänä noin alle 35-vuotialla eli julkinen kulutus olisi noussut seitsemäntoista vuoden aikana lapsilla ja nuorilla. Julkisen kulutuksen lineaarisen sovitteen estimaatit ovat

$$\widehat{\mathbf{B}}_* = \begin{bmatrix} 0.014758554 & 2.397750\text{e}-03 \\ 0.006212037 & -7.499131\text{e}-05 \end{bmatrix}$$

Lineaarisen aikakehityksen testauksen tulokseksi saadaan $Q = 303.389 > \mathcal{X}^2_{4;0.05} = 9.488$, joka on selvästi yli kriittisen arvon ja on jo selkeästi suurempi kuin elinkaarialijäämän testi arvo. Julkisen kulutuksen lineaariset sovitteet eivät siis ole yhdensuuntaiset, kuten kuviosta 4.14 huomataan vuoden. Voidaan siis todeta, että muutos nuorilla on erisuuntainen kuin vanhemmalla väestöllä.



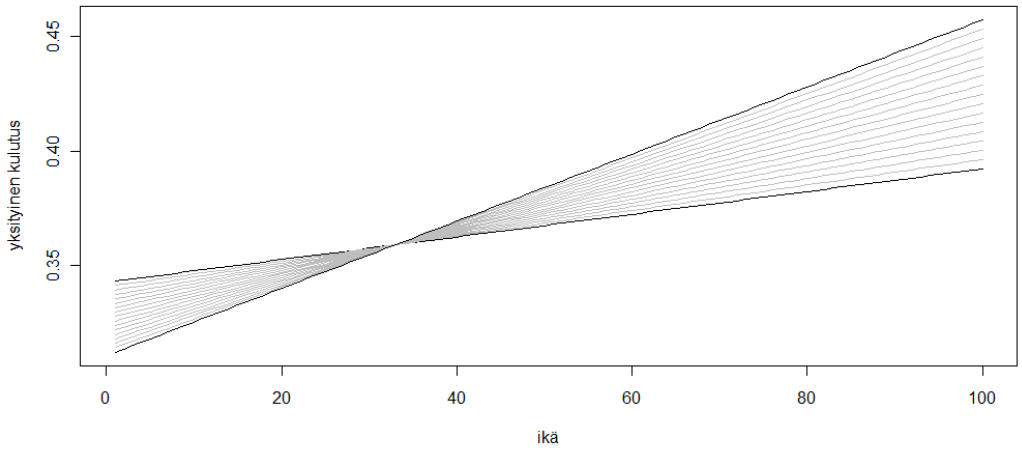
Kuvio 4.14. Julkisen kulutuksen lineaarinen aikakehitys.

Yksityinen kulutus

Yksityinen kulutus on aiemmissa tarkasteluissa selkeästi ollut se, jolla on kaikista suurin vaikutus elinkaarialijäämän kehitykseen. Tämä voidaan todeta kuviosta 4.15 myös lineaarisen sovitteenkin osalta. Vuosia 1990 ja 2006 kuvaavat suorat ovat huomattavan erisuuntaiset. Testin tulokseksi saadaan nyt $Q = 1247.633 > \mathcal{X}^2_{4;0.05} = 9.488$, joka on huomattavasti kriittistä arvoa suurempi. Kiinteän osan estimaateiksi saadaan

$$\widehat{\mathbf{B}}_* = \begin{bmatrix} 0.3446969064 & -2.000400\text{e}-03 \\ 0.0004344477 & 6.069102\text{e}-05 \end{bmatrix}$$

Eroavaisuutena julkiseen kulutukseen on suorien kehitys seitsemäntoista vuoden aikana. Yksityinen kulutus on vähentynyt alle 25-vuotiailla nuorilla aikuisilla, mutta kasvanut huomattavasti yli 25-vuotialla henkilöillä (kuvio 4.11). Koska yksityisen sekä julkisen kulutuksen kehitys on ollut erisuuntaista, elinkaarialijäämän aikakehityksen erot ovat maltillisempia ja elinkaarialijäämän testiarvo jää pienemmäksi.



Kuvio 4.15. Yksityisen kulutuksen aikakehitys vuosina 1990- 2006. Mustat viivat ovat vuodet 1990 ja 2006. Vuotta 2006 kuvaava suora on nuoremmilla alempana.

4.2.4 Ajallisen kehityksen testaamisen pohdintaa

Jokaiselle aineiston muuttujalle tehdyissä aikakehityksen testauksessa, tuli erittäin merkitsevä testiarvo. Aiemmista kuvioista voidaan huomata, että muutosta on tapahtunut, etenkin kulutuksen osalta. Testien avulla testattiin onko ajalla vaikutusta tähän muutokseen ja kaikissa testeissä vaikutusta näyttäisi olevan. Huomioitavaa on kumminkin, että otoskoon kasvaessa pienetkin erot vaikuttavat testin merkitsevyyteen ja nostavat testisuureen Q arvoa.

Yksityisen kulutuksen aineiston testiarvo oli lähes kymmenkertainen elinkaarialijäämän testiarvoon verrattuna, joten yksityisen kulutuksen ajallista kehitystä voidaan pitää selvänä. Julkisen kulutuksen ajallinen kehitys oli vastakkaissuuntaista yksityisen kulutuksen kehitykselle. Kun elinkaarialijäämän kehitykseen vaikuttaa kokonaiskulutus eli yksityinen ja julkinen kulutus yhteensä, pienenee elinkaarialijäämän ajallinen kehityksen eroavaisuus, koska kulutusten kehityssuunnat kumoavat toisiaan. Kokonaiskulutuksen testauksen arvoksi saadaan 21.06557, joka sekin ylittää kriittisen arvon, mutta on selvästi muita testin arvoja matalampi.

Ajallista kehitystä on myös tapahtunut ansiotuloissa, mikä johtuu pidentyneistä työurista työuran loppupuolella. Lineaarisen sovitteen suorat ovat laskevia, mikä johtuu siitä, että ansiotuloja ei ole nuorilla noin alle 15-vuotialla eikä eläkeikäisillä, ja siitä, että eläkeikäisten osuus 1-100-vuotiaista on suurempi (noin 35 %) kuin nuorten osuus (noin 15 %)

5. Loppusanat

Tutkielmassa elinkaarialijäämiä mallinnetaan tasoittavilla kuutiosplineillä ja testataan lineaarista aikakehitystä sekamallin avulla. Tasoittavat kuutiosplinit antavat joustavan tavan kuvata sellaista aineistoa, johon tavallinen parametrinen menetelmä ei sovellu. Aluksi aineistoon sovitetaan kuutiosplinit valitun tasoitusparametrin avulla. Spliniratkaisulla on yhteys sekamalliin. Sekamallin kiinteä osa on mallissa lineaarinen ja satunnainen osa kuvaa aineiston vaihtelua lineaarisen osan ympärillä.

Sekamallin muodostamisen jälkeen kiinnitetään huomio sekamallin kiinteään osaan ja tarkastellaan lineaarista aikakehitystä. Ajallisen kehityksen testaaminen perustuu sekamallin testaukseen kasvukäyräaineiston tilanteessa (Nummi & Koskela, 2008). Ajallista tarkastelua tehdään elinkaarialijäämäaineiston lisäksi tulo- ja kulutusaineistoille. Testauksen seurauksena saadaan erittäin merkitsevät tulokset kaikille selitettäville muuttujille. Kuten kappaleessa 4.2.4 mainitaan, testien merkitsevyyttä lisää ja testiarvoa nostaa aineiston suuri koko. Oleellisempaa kuin testin merkitsevyys on kuitenkin havaitun muutoksen suuruus. Muutos on ollut erityisen voimakas kulutuksen osalta. Kehitys ei ollut kaikille muuttujille samansuuntaista: yksityinen ja julkinen kulutus ovat kehittyneet vastakkaisiin suuntiin. Julkinen kulutus on vähentynyt ja yksityinen kulutus lisääntynyt vanhemmalla väestöllä ja nuorilla kulutukset ovat muuttuneet vastakkaiseen suuntaan.

Tutkimuksen tarkoituksena oli tarkastella elinkaarialijäämän kehitystä ja voidaan todeta siinäkin tapahtuneen muutosta seitsemäntoista vuoden aikana. Tosin muutos koskettaa ainoastaan vanhempaa väestöä. Yksityisessä kulutuksessa näkyy selkeä muutos vanhemman väestön kohdalla, ja voidaankin olettaa, että vanhemman väestön varallisuus on lisääntynyt. Varallisuutta ei tässä tutkimuksessa varsinaisesti oteta huomioon, mutta oletettavasti tämän hetken vanhemman väestön varallisuus siirtyy perintöinä seuraavalle sukupolvelle. Nuorilla aikuisiän kynnyksellä olevilla julkinen kulutus näyttäisi olevan lisääntynyt ja ansiotulot vähentyneet. Tämä näyttäisi viittaavan siihen, että opiskelua jatketaan nykyään pidempään ennen työelämään siirtymistä kuin ennen. Joten vaikka työikää on saatu pidennettyä loppupäästä eläkeikää nostamalla, on sitä lyhennetty työiän alkupäästä.

Lopuksi haluan lausua kiitoksen ohjaajalleni Arto Luomalle, joka ehdotti minulle tätä mielenkiintoista lähestymistapaa elinkaarialijäämien tilastolliseen tarkasteluun ja on edesauttanut työni edistymistä arvokkailla neuvoillaan. Kiitos professori Erkki Liskille ohjeista ja kannustuksesta. Kiitokset myös Eläketurvakeskuksesta Risto Vaittiselle, jolta sain tämän kiinnostavan aiheen ja aineiston käsiteltäväksi. Erityiskiitoksen haluan sanoa kodin tukijoukoilleni: aviomiehelleni ja kolmelle lapselleni tuesta, kannustuksesta ja kärsivällisyydestä läpi koko opiskeluaikani.

Lähteet

Akaike H. (1973*).* Information theory and an extension of the entropy maximization principle. *2nd International Symposium of Information Theory, (eds.) Akademia* 267 – 281.

Green, P., Silverman, B. (1994). Nonparametric Regression and Generalized Linear Models, *Chapman & Hall*, London.

Harville D. (1976). Extension of the Gauss-Markov theorem include the estimation of random effects. *Annals of Statistics*, 4, 384 – 395.

Harville D. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of American Statistical Association*, 72, 320 – 340.

Laird N., Ware J. (1982). Random effects models for longitudinal data. *Biometrics*, 38, 963 – 974.

Lee R., Lee S-H., Mason A. (2008). Charting the Economic Life Cycle.

Mason A., Lee R., Donehover G., Lee S-H., Miller T., Tung A-C., Chawla A. (2009). National Transfer Accounts Manual Draft version 1.0.

Nummi, T., Koskela, L. (2008*).* Analysis of growth data by using cubic smoothing splines. *Journal of Applied Statistics*, Vol. 35, No. 6, June 2008, 681 – 691.

Nummi, T., Möttönen, J. (2000). On the analysis of multivariate growth curves, *Metrika*, 52, pp, 77 – 89.

Potthoff, R., Roy S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 86, 677 – 690.

Robinson G. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistics Science*, 6, 15 – 32.

Reinsch, C. (1967). Smoothing by spline functions, *Numerische Mathematic*, 10, 177 – 183.

Riihelä M., Vaittinen R., Vanne R. (2010). Changing Patterns of Intergenerational Resource Allocation in Finland.

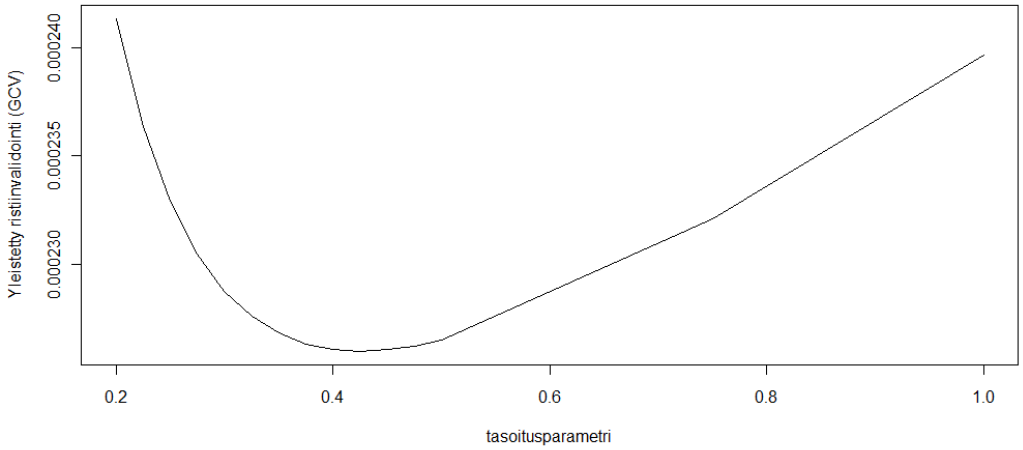
Schwarz G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461 – 464.

Speed T. (1991). Discussion of That BLUP is a good thing: The estimation of random effects by Robinson. *Statistics Science*, 6, 42 – 22.

Vaittinen R., Vanne R. (2010). National Transfer Accounts for Finland in 2004.

Wu H., Zhang J.-T. (2006), Nonparametric Regression Methods for Longitudinal Data Analysis. *New Jersey: John Wiley & Sons*.

Liite 1: Tasoitusparametrin vertailu



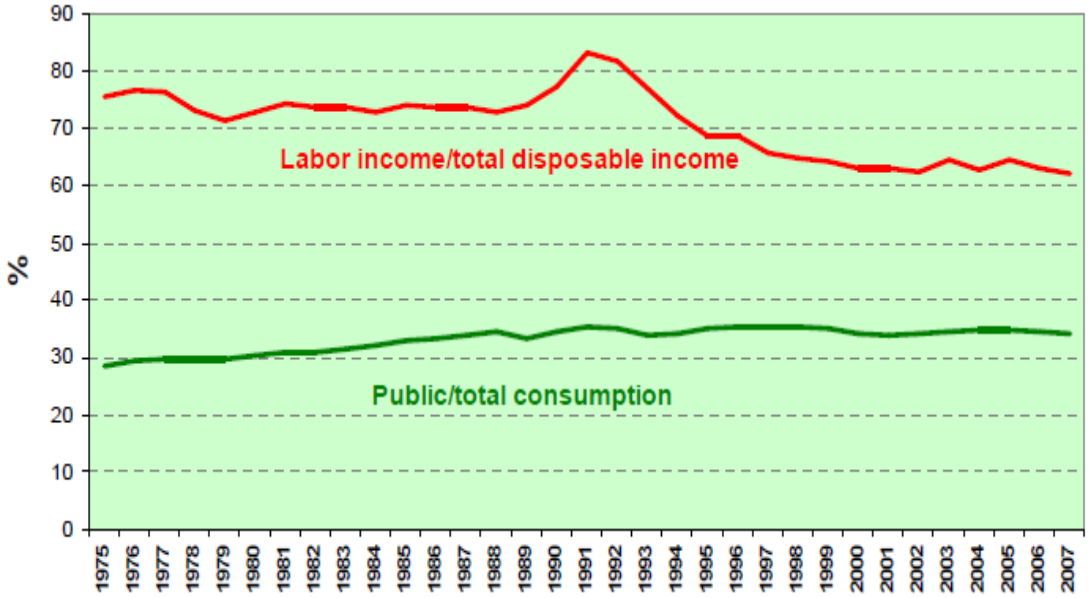
Tasoitusparametri	Tehokkaat vapausasteet	Yleistetty ristiinvalidointi, GCV
1.000	36.18634	0.00023964
0.750	38.77038	0.00023210
0.500	42.71265	0.00022650
0.475	43.23709	0.00022623
0.450	43.79622	0.00022605
0.425	**44.39440**	**0.00022598**
0.400	45.03682	0.00022607
0.375	45.72973	0.00022633
0.350	46.48074	0.00022682
0.325	47.29924	0.00022760
0.300	48.19695	0.00022877
0.275	49.18881	0.00023046
0.250	50.29415	0.00023289
0.225	51.53865	0.00023633
0.200	52.95724	0.00024131

Liite 2: Kotitalouksien tulonjako

Income source
Earned income
Wages and salaries
Entrepreneurial income
+Income from capital
Interest income and dividends
Imputed net rents of owner-occupied dwellings
Other capital income
=Factor income
+Current transfers received
(Public transfers received)
Pensions
Income maintenance during illness
Family policy transfers
Unemployment security
Other age-related transfers
Other transfers
Inter households' transfers received
=Gross income
+Current transfers paid
(Public transfers paid)
State income tax and municipal tax
Tax on capital
Wealth and property tax
Social security contribution
Employment pension contribution
Inter households' transfers paid
=Disposable income

Lähde: Riihelä M., Vaittinen R., Vanne R. (2010) & Tulonjakotutkimus, Tilastokeskus.

Liite 3: Palkkatulojen ja julkisen kulutuksen osuudet kokonaistuloista ja -kulutuksesta



Lähde: Riihelä M., Vaittinen R., Vanne R. (2010).