Monen selittäjän regressiomenetelmien vertailu: Alzheimerin tautiin liittyvien veren proteiinien yhteys kognitioon

Jasmine Hakala

Pro gradu -tutkielma
Kesäkuu 2024

Tarkastajat:
Prof. Kari Auranen
Erikoistutkija Marja Heiskanen

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Jasmine Hakala: Monen selittäjän regressiomenetelmien vertailu: Alzheimerin tautiin liittyvien veren proteiinien yhteys kognitioon
Pro gradu -tutkielma, 38 s., 13 liites.
Tilastotiede
Kesäkuu 2024

Tutkielmassa tarkasteltiin Alzheimerin tautiin liittyvien veren beta-amyloidien 40 (Aβ40) ja 42 (Aβ42), hapan säikeisen gliaproteiinin (GFAP) sekä neurofilamentin kevytketjun (NfL) pitoisuuksien yhteyttä kognitiotestien tuloksiin terveillä henkilöillä. Yhteyttä tutkittiin vertailemalla erilaisia menetelmiä, jotka sopivat monen selittäjän tilanteeseen. Selittäjinä käytettiin yksittäisten kognitiotestien tuloksia. Menetelmiksi valikoitui osittaisen pienimmän neliösumman regressio, pääkomponentti-, LASSO- ja harjaregressio.

Osittaisen pienimmän neliösumman regressio ja pääkomponenttiregressio ovat dimension pienentämismenetelmiä, jotka muodostavat korreloimattomia komponentteja alkuperäisten selittäjien tilalle. Tämä vähentää selittäjien määrää ja pienentää korrelaatiota muuttujien välillä. LASSO- ja harjaregressio kutistavat regressiokertoimien suuruutta kohti nollaa, mikä pienentää kertoimien varianssia ja minimoi ennusteharhaa.

Tutkielmassa käytettiin Turun yliopiston Väestötutkimuskeskuksen Lasten Sepelvaltimotaudin Riskitekijät (LASERI) -tutkimuksen vuosina 2018—2020 kerättyä osakohorttia, johon kuului 814 iältään 59–88-vuotiasta henkilöä (G_0-sukupolvi) ja 1 237 heidän iältään 41–58-vuotiasta lastaan (G_1-sukupolvi). Aineistossa oli kultakin yksilöltä 68 kognitiivisen testin tulosta, verinäytteistä otetut proteiinipitoisuudet sekä taustamuuttujia, kuten sukupuoli, ikä, koulutusvuodet, nukutut tunnit ja vireystila.

Kognitiotestien ja proteiinipitoisuuksien yhteyttä tutkittiin G_0- ja G_1-sukupolvissa sekä yhdessä että erikseen. Tulokset osoittivat heikon yhteyden kognitiotestien ja proteiinien välillä G_0-sukupolven aineistossa. Sen sijaan G_1-sukupolven aineistossa kognitiotestien arvojen suurentuessa NfL-pitoisuus kasvaa, mutta GFAP-pitoisuus laskee. Koko aineistossa kognitiotestien arvojen kasvaessa NfL- ja GFAP-pitoisuudet kasvavat. Tulokset eivät kuitenkaan ole yksiselitteisiä, koska pääkomponenttien latauksset olivat sekä positiivisia että negatiivisia. Proteiinipitoisuuksien ja iän välinen yhteys oli selkeämpi ja ikääntyessä pitoisuudet kasvavat. Vertailun kohteena olleet menetelmät toimivat toisiinsa nähden yhtä hyvin.

Asiasanat: osittaisen pienimmän neliösumman regressio, pääkomponentti-, harja- ja LASSO-regressio.

Sisällys

1 Johdanto

Suomessa on arviolta 193 000 muistisairasta ja lisäksi 200 000 ihmistä kärsii kognitiivisen toiminnan heikentymisestä [1]. Muistisairaudet saattavat kehittyä jopa vuosikymmeniä ennen kliinisten oireiden ilmenemistä [2]. Tästä syystä varhaiseen diagnosointiin sekä sairauksien ennaltaehkäisyyn on tärkeä panostaa. Varhainen diagnosointi voi vähentää muistisairauden riskiä, hidastaa oireiden etenemistä, ylläpitää toimintakykyä ja parantaa elämänlaatua. Ennaltaehkäisyllä voidaan viivästyttää muistisairauden puhkeamista jopa viidellä vuodella, mikä saattaa vähentää Alzheimerin taudin ilmaantuvuutta 50 prosentilla yhden sukupolven aikana [3].

Tämän tutkielman tavoitteena on tutkia Alzheimerin tautiin liittyvien veren proteiinien yhteyttä kognition neljään osa-alueeseen: muisti ja oppiminen, työmuisti, informaation käsittely sekä reaktionopeus. Tämän tiedon avulla voitaisiin kehittää uusia ennaltaehkäisyn ja varhaisen diagnostiikan menetelmiä Alzheimerin tautiin. Tutkielman tarkoituksena ei ole löytää syy-seuraussuhdetta vaan tutkia proteiinien ja kognitiotestien välistä yhteyttä lineaaristen regressiomallien avulla. Tarkasteltavia proteiineja ovat beta-amyloidit 40 (*beta-amyloid* 40, Aβ40) ja 42 (*beta-amyloid* 42, Aβ42), hapan säikeinen gliaproteiini (*glial fibrillary acidic protein*, GFAP) sekä neurofilamentin kevytketju (*neurofilament light chain*, NfL).

Aβ40- ja Aβ42-proteiinit ovat osana amyloidiplakkien muodostumista aivoissa, mikä on yksi Alzheimerin taudin tunnusmerkeistä [4]. GFAP-proteiini on puolestaan astrotyyppisten solujen pääasiallinen säie, jonka pitoisuus terveillä ihmisillä on matala [5]. Korkea GFAP-proteiinipitoisuus on yhdistetty esimerkiksi aivovammoihin. NfL-proteiini on puolestaan hermosolujen tärkein tukirangan komponentti, ja sen korkeita pitoisuuksia on havaittu neurodegeneratiivisissa sairauksissa, jotka heikentävät hermosolujen toimintaa [6]. Proteiinipitoisuuksia on aikaisemmin mitattu PET-kuvantamisella tai aivoselkäydinnestenäytteestä. Nämä menetelmät ovat sekä kalliita että aikaa vieviä, ja aivoselkäydinnesteen näytteenotto voi olla potilaalle epämiellyttävä kokemus. Nykytutkimuksissa on havaittu, että verinäytteillä voidaan saada yhtä luotettavia tuloksia kuin PET-kuvantamisella tai aivoselkäydinnestenäytteestä [7]. Tämä teknologinen edistysaskel on kiihdyttänyt biomarkkereiden tutkimusta, ja nykyään näiden proteiinipitoisuuksien tutkimus on mahdollista myös terveiltä henkilöiltä, joilla pitoisuudet ovat matalia.

Kyseisten proteiinien yhteydestä muistisairauksiin ja etenkin Alzheimerin tautiin on runsaasti kansainvälistä tutkimustietoa. NfL-proteiinin nousu heijastaa aksonivaurioita, joiden on todettu olevan yleisiä Alzheimerin taudin yhteydessä [8]. Tutkimuksissa on havaittu korkeamman NfL-proteiinipitoisuuden olevan yhteydessä myös heikentyneeseen kognitiiviseen toimintaan useissa neurologisissa häiriöissä [9, 10]. GFAP-proteiinipitoisuuden on havaittu olevan kohonnut Alzheimer-diagnoosin saaneilla [8]. Myös GFAP-proteiinilla on havaittu olevan yhteys kognition heikkenemiseen muistisairailla [11]. Aβ40- sekä Aβ42- proteiinien kerääntymisen aivoihin on havaittu olevan yksi Alzheimerin taudin tunnusmerkeistä [12], ja alhaisen Aβ42- ja Aβ40-proteiinien välisen suhteen on havaittu olevan yhteydessä kognition heikkenemiseen [13]. Lisäksi Aβ40-proteiinipitoisuuden on havaittu olevan yhteydessä kognition heikkenemiseen Parkinsonin tautiin sairastuneilla [14].

Aikaisemmissa tutkimuksissa yhteys kognition ja proteiinien välillä on löytynyt

eri sairauksiin diagnosoiduilla. Tämän tutkielman tavoitteena on tarjota uutta tietoa yhteyksistä terveillä yksilöillä. Uuden tutkimustiedon avulla voidaan kehittää keinoja muistisairauksien ennaltaehkäisyyn ja varhaiseen diagnostiikkaan. Tutkimuskysymyksenä on, onko veren NfL-, Aβ40-, Aβ42- ja GFAP-proteiinipitoisuuksilla yhteys yksilön kognitiotestien tuloksiin.

Tässä tutkielmassa käytetään Turun yliopiston Väestötutkimuskeskuksen keräämää Lasten Sepelvaltimotaudin Riskitekijät (LASERI) -tutkimuksen aineistoa [15]. LASERI on pitkittäistutkimus, jonka tavoitteena on selvittää sairauksien syntyyn vaikuttavia tekijöitä. Tutkimusaineiston kerääminen on aloitettu vuonna 1980 ja aineisto koostuu useammasta sukupolvesta. Tutkimuskerrat toistuvat muutaman vuoden välein ja viimeisin aineistonkeruu tapahtui 2018–2020. Tässä tutkielmassa hyödynnetään tämän viimeisimmän tutkimuskerran aineiston osakohorttia, johon kuuluu 814 iältään 59–88-vuotiasta henkilöä (G_0-sukupolvi) ja 1 237 heidän iältään 41–58-vuotiasta lastaan (G_1-sukupolvi).

Tutkimuskohderyhmän ikähaarukassa voidaan jo havaita muutoksia kognitiossa ja veren proteiinipitoisuuksissa, mitkä saattavat ennakoida Alzheimerin taudin puhkeamista myöhemmässä vaiheessa. Tukittavilta on mitattu NfL-, Aβ40-, Aβ42- sekä GFAP-proteiinipitoisuudet verinäytteestä Simoa-menetelmällä (*single molecule array HD-X Analyzer*), joka mahdollistaa pienten konsentraatioiden mittaamisen. Tämän lisäksi tutkittavilta on testattu elektronisella alustalla kognitiivisia toimintoja, kuten muistia ja oppimista. Aineistossa on yhteensä 68 kognitiivisen testin tulosta.

Vaikka varsinainen tutkimuskysymys on, kuinka Alzheimerin tautiin liittyvät veren proteiinit ovat yhteydessä kognitioon, tässä tutkielmassa tarkastellaan asiaa toisin päin. Koska tutkielmassa ei ole tavoitteena löytää syy-seuraussuhdetta vaan tutkia yhteyksiä, koettiin helpommaksi tutkia suhdetta päinvastoin. Tällöin yhtä proteiinipitoisuutta selitetään kognitiotestien tuloksilla sekä taustamuuttujilla. Tavoitteena on löytää sopiva malli vertailemalla erilaisia tilastollisia menetelmiä. Tutkielmaan valikoitui menetelmiä, joilla pystytään hallitsemaan useita selittäjiä. Näitä menetelmiä ovat osittaisen pienimmän neliösumman regressio, pääkomponentti-, harja- ja LASSO-regressio.

Tutkielmassa perehdytään ensin tilastollisiin menetelmiin ja niiden teoriataustoihin. Luvussa 2 tarkastellaan dimension pienentämismenetelmiä pääkomponenttiregressio ja osittaisen pienimmän neliösumman regressio. Luvussa 3 esitellään kutistamismenetelmät harja- ja LASSO-regressio. Tämän jälkeen perehdytään ristiinvalidointiin luvussa 4. Luvussa 5 vertaillaan valittuja menetelmiä ja tutustutaan harhavarianssikompromissiin. Näiden jälkeen luvussa 6 esitellään aineistoa ja sovelletaan menetelmiä aineistoon. Lopuksi luvussa 7 pohditaan tuloksien merkitystä.

2 Dimension pienentämismenetelmät

Tässä luvussa esitellään dimension pienentämismenetelmät pääkomponenttiregressio (*principal component analysis*, PKR) ja osittaisen pienimmän neliösumman regressio (*partial least square regression*, OPNR). Näitä menetelmiä sovelletaan monen selittäjän lineaarisissa regressiomalleissa, jotka ovat muotoa

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \cdots + \tilde{\beta}_m x_{im} + \epsilon_i, \tag{1}$$

jossa y_i on yksilön i vastemuuttuja, $x_{i1}, \ldots, x_{im}$ ovat jatkuvia selittäviä muuttujia, $\tilde{\beta}_0$ on vakiotermi, $\tilde{\beta}_1, \ldots, \tilde{\beta}_p$ ovat selittävien muuttujien regressiokertoimet ja ϵ_i mallin virhetermi. Regressiomalleissa vastemuuttujaa pyritään selittämään tai ennustamaan selittävien muuttujien avulla. Monissa aineistoissa esiintyy paljon selittäjiä, jotka ovat vahvasti korreloituneita. Vahva korrelaatio eli multikollineaarisuus voi aiheuttaa ongelmia ja rajoittaa joidenkin menetelmien käyttämistä. Lisäksi on oleellista identifioida mahdollisimman pieni määrä tärkeitä selittäjiä. Dimension pienentämismenetelmillä pyritään hallitsemaan näitä ongelmia. [16]

PKR-menetelmä on ohjaamatonta oppimista, jossa luodaan pääkomponentteja eli korreloimattomia lineaarikombinaatioita selittävistä muuttujista [17]. Ohjaamattomassa oppimisessa analyysi ei ota huomioon yhteyttä vastemuuttujaan. OPNR-menetelmä on ohjattu vaihtoehto PKR-menetelmälle [18]. Toisin sanoen OPNR-menetelmä ottaa huomioon vasteen komponenttien muodostuksessa [16].

Menetelmät tarjoavat monipuolisia sovellusmahdollisuuksia. Niiden avulla voidaan minimoida ennustevirhettä, vähentää selittäjien määrää, pienentää korrelaatiota muuttujien välillä, helpottaa aineiston visualisointia sekä imputoida puuttuvia arvoja. Menetelmät mahdollistavat mataladimensioisen esityksen alkuperäisestä aineistosta ilman, että menetetään tärkeää tietoa. [19, 16, 17]

Alkuperäiset selittäjät vaihtelevat usein eri mittayksiköissä, mikä voi vaikuttaa aineistosta luodun mallin suorituskykyyn sekä muuttujien vertailukelpoisuuteen. Tällöin mitta-arvoltaan suuremmat muuttujat saavat suuremman painon analyysissa. Ongelma voidaan korjata muuntamalla eli skaalaamalla muuttujien arvot samaan mittayksikköön. Tällöin jokaisen muuttujan keskiarvoksi tulee nolla ja keskihajonnaksi yksi. PKR- ja OPNR-menetelmissä on tärkeää skaalata muuttujat analyysin onnistumisen kannalta. [16]

PKR-menetelmän yhteydessä puhutaan usein pääkomponenteista ja OPNR-menetelmän yhteydessä pelkistä komponenteista. Pääkomponentit ja komponentit ovat kuitenkin molemmat korreloimattomia varianssin mukaan järjestettyjä projektioita $z_{i1}, \ldots, z_{im}$ selittävistä muuttujista $x_{i1}, \ldots, x_{im}$ [16]. Selkeyden vuoksi tässä puhutaan pelkistä komponenteista. Yleisesti yksilön i p:nnes komponentti z_{ip} voidaan kirjoittaa muodossa

$$z_{ip} = \phi_{1p} x_{i1} + \cdots + \phi_{mp} x_{im} = \boldsymbol{\phi}_m^T \boldsymbol{x}_i \quad i = 1, \ldots, n.$$

Menetelmät generoivat yhtä monta komponenttia kuin alkuperäisiä selittäviä muuttujia on [16]. Komponentit selittävät kaiken alkuperäisten selittäjien varianssista. Kaikki komponentit eivät ole välttämättömiä jatkoanalyysien kannalta. Tavoitteena on valita mahdollisimman pieni määrä p komponenttia kuitenkin niin, että riittävä osuus alkuperäisten selittäjien varianssista on selitetty [17]. Tämä pienentää

selittäjien dimensiota ja vähentää multikollineaarisuuden ongelmaa [17]. Menetelmät lieventävät myös ylisovittamisen ongelmaa, joka voi syntyä käytettäessä suurta joukkoa alkuperäisiä korreloituneita selittäjiä. Mitä enemmän komponentteja valitaan malliin, sitä suurempi osa varianssista on selitetty, mutta tällöin myös harhan suuruus kasvaa [16]. Tämä on huomioitava komponenttien valinnan yhteydessä.

Komponentit voidaan valita esimerkiksi taluskuvion tai histogrammin avulla, jossa y-akselilla on selitetyn varianssin määrä ja x-akselilla komponenttien lukumäärä [17]. Kuvaajan perusteella pyritään tunnistamaan kohta, jossa peräkkäisten komponenttien selittämä varianssin osuus laskee selkeästi. Mikäli komponentteja on tarkoitus käyttää ohjatussa jatkoanalyysissa, kuten lineaarisessa regressiossa, voidaan komponenttien lukumäärä valita ristiinvalidoinnin avulla [16]. Ristiinvalidointia käsitellään tämän tutkielman luvussa 4.

Valitut p kappaletta pääkomponentteja $z_{i1}, \ldots, z_{ip}$ sijoitetaan yhtälöön (1) alkuperäisten selittäjien $x_{i1}, \ldots, x_{im}$ tilalle. Tällöin monen selittäjän lineaarinen regressiomalli on muotoa [16]

$$y_i = \beta_0 + \beta_1 z_{i1} + \cdots + \beta_p z_{ip} + \epsilon_i.$$

Yhteensä n riippumattoman yksilön aineistoon perustuen alkuperäisten (skaalattujen) selittäjien kokonaisvarianssi on

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{1}{n} x_{ij}^2.$$

Tällöin p:nnen komponentin z_{ip} selittämä varianssi on

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{m} \phi_{jp} x_{ij} \right)^2.$$

Tällöin kokonaisvarianssista on selitetty osuus,

$$\frac{\sum_{i=1}^{n} \left(\sum_{j=1}^{m} \phi_{jp} x_{ij} \right)^2}{\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}^2},$$

joka on välillä $[0, 1]$. Komponenttien kumulatiivisen varianssin selitysosuuden laskemiseksi voidaan laskea yhteen haluttujen pääkomponenttien varianssin selitysosuudet. Kaikkein pääkomponenttien kumulatiivinen selitysosuus on yhtä kuin yksi.

Tutkielman analyysit on toteutettu R-ohjelmistolla, jonka *prcomp*-funktiota käytettiin pääkomponenttiregressioon ja *plsr*-funktiota osittaiseen pienimmän neliösumman regressioon *pls*-paketista.

2.1 Pääkomponenttien muodostaminen

Oletetaan, että selittävät muuttujat x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$, jossa n on yksilöiden lukumäärä ja m alkuperäisten selittäjien lukumäärä. Oletetaan, että alkuperäiset selittäjät on skaalattu. Ensimmäiset pääkomponentit z_{i1}, $i = 1, .., n$, ovat

lineaarikombinaatioita alkuperäisistä selittäjistä $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})^T$,

$$z_{i1} = \phi_{11}x_{i1} + \cdots + \phi_{m1}x_{im} = \boldsymbol{\phi}_1^T\boldsymbol{x}_i, \quad i = 1, \ldots, n,$$

jotka toteuttavat ehdon $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = 1$ ja maksimoivat komponenttien $z_{i1}, ..., z_{n1}$ otosvarianssin

$$Var(z_{i1}, ..., z_{n1}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\phi}_1^T x_{i1}x_{i1}^T\boldsymbol{\phi}_1 = \boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1.$$

Yllä $\boldsymbol{\Sigma} = \sum_{i=1}^{n} x_{i1}x_{i1}^T$ on vektoreiden $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ otoskovarianssimatriisi. Toiset pääkomponentit z_{i2}, $i = 1, .., n$, ovat sellaisia ensimmäisiin pääkomponentteihin nähden kohtisuorassa olevia lineaarikombinaatioita

$$z_{i2} = \phi_{12}x_{i1} + \cdots + \phi_{m2}x_{im} = \boldsymbol{\phi}_2^T\boldsymbol{x_i}, \quad i = 1, \ldots, n,$$

jotka maksimoivat varianssin $Var(z_{i2}, ..., z_{n2}) = \boldsymbol{\phi}_2^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2$ ja jossa $\boldsymbol{\phi}_2^T\boldsymbol{\phi}_2 = 1$. Toiset komponentit ovat korreloimattomia ensimmäisten komponenttien kanssa, mikä on yhtäpitävää sen kanssa, että ns. latausvektori $\boldsymbol{\phi}_2$ on ortogonaalinen eli kohtisuorassa ensimmäisen latausvektorin $\boldsymbol{\phi}_1$ kanssa. Yleisesti p:nnet pääkomponentit z_{ip}, $i = 1, \ldots, n$, voidaan kirjoittaa muodossa

$$z_{ip} = \phi_{1p}x_{i1} + \cdots + \phi_{mp}x_{im} = \boldsymbol{\phi}_m^T\boldsymbol{x_i}, \quad i = 1, \ldots, n,$$

jotka maksimoivat varianssin $Var(z_{ip}, ..., z_{np}) = \boldsymbol{\phi}_p^T\boldsymbol{\Sigma}\boldsymbol{\phi}_p$ ja jossa $\boldsymbol{\phi}_p^T\boldsymbol{\phi}_p = 1$. Myös nämä komponentit ovat korreloimattomia edellisten pääkomponenttien kanssa.

Ensimmäisten pääkomponenttien varianssin maksimointi voidaan ratkaista Langrangen kerrointen avulla. Aluksi määritellään $g(\boldsymbol{\phi}_1) = \boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 - 1$, jolloin Langrangen funktio saa muodon

$$L(\boldsymbol{\phi}, \lambda) = \boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1 - \lambda(\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 - 1),$$

jossa λ on Langrangen kerroin. Yhtälö maksimoidaan derivoimalla $\boldsymbol{\phi}_1$:n suhteen ja asettamalla yhtälö yhtä suureksi kuin nolla. Tämä johtaa ehtoon

$$\boldsymbol{\Sigma}\boldsymbol{\phi}_1 - \lambda\boldsymbol{\phi}_1 = 0 \Leftrightarrow (\boldsymbol{\Sigma} - \lambda\boldsymbol{I}_p)\boldsymbol{\phi}_1 = 0,$$

jossa $\boldsymbol{I}_p$ on $(p \times p)$-identiteettimatriisi. Havaitaan siis, että λ on kovarianssimatriisin $\boldsymbol{\Sigma}$ ominaisarvo ja $\boldsymbol{\phi}_1$ vastaava ominaisvektori. Kertomalla yllä oleva ehto vasemmalta $\boldsymbol{\phi}_1$:llä seuraa myös, että

$$\boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1 = \lambda.$$

Koska ensimmäisten komponenttien otosvarianssi on yhtä kuin λ, $\boldsymbol{\phi}_1$:ksi pitää valita suurinta ominaisarvoa λ vastaava ominaisvektori, jolloin varianssi on yhtä kuin λ. Tämä osoittaa, että $\boldsymbol{\phi}_1$ vastaa suurinta ominaisvektoria ja $Var(\boldsymbol{\phi}_1^T\boldsymbol{x}_i) = \boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1$ on suurin ominaisarvo λ_1.

Toiset pääkomponentit $z_{i2} = \boldsymbol{\phi}_2^T\boldsymbol{x}_i$ $i = 1, \ldots, n$, ovat, kuten aikaisemmin on todettu, korreloimattomia ensimmäisten pääkomponenttien z_{i1}, $i = 1, \ldots, n$ kanssa. Toisin sanoen

$$Cov((z_{i1}, ..., z_{n1}), (z_{i2}, ..., z_{n2})) = Cov((\boldsymbol{\phi}_1^T\boldsymbol{x}_i, ..., \boldsymbol{\phi}_1^T\boldsymbol{x}_n), (\boldsymbol{\phi}_2^T\boldsymbol{x}_i, ..., \boldsymbol{\phi}_2^T\boldsymbol{x}_n)) = 0.$$

Tällöin pätee, että

$$Cov((z_{i1},...,z_{n1}),(z_{i2},...,z_{n2})) = \boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2 = \boldsymbol{\phi}_2^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1 = \boldsymbol{\phi}_2^T\lambda_1\boldsymbol{\phi}_1^T = \lambda_1\boldsymbol{\phi}_2^T\boldsymbol{\phi}_1 = \lambda_1\boldsymbol{\phi}_1^T\boldsymbol{\phi}_2.$$

Näin ollen mitä tahansa seuraavista neljästä yhtälöstä,

$$\begin{aligned} \boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2 &= 0, & \boldsymbol{\phi}_2^T\boldsymbol{\Sigma}\boldsymbol{\phi}_1 &= 0, \\ \boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 &= 0, & \boldsymbol{\phi}_2^T\boldsymbol{\phi}_1 &= 0, \end{aligned}$$

voidaan käyttää edustamaan ehtoa $Cov((z_{i1},...,z_{n1}),(z_{i2},...,z_{n2})) = 0$. Valitaan viimeisin vaihtoehto. Tällöin Langrangen funktio on muotoa

$$L(\boldsymbol{\phi}_2,\lambda,\alpha) = \boldsymbol{\phi}_2^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2 - \lambda(\boldsymbol{\phi}_2^T\boldsymbol{\phi}_2 - 1) - \alpha\boldsymbol{\phi}_2^T\boldsymbol{\phi}_1,$$

jossa λ ja α ovat Langrangen kertoimia. Yhtälö maksimoidaan derivoimalla $\boldsymbol{\phi}_2$ suhteen ja asettamalla yhtälö yhtä suureksi kuin nolla, mikä johtaa ehtoon

$$\boldsymbol{\Sigma}\boldsymbol{\phi}_2 - \lambda\boldsymbol{\phi}_2 - \alpha\boldsymbol{\phi}_1 = 0.$$

Kertomalla yhtälö vasemmalta puolelta $\boldsymbol{\phi}_1^T$ saadaan,

$$\boldsymbol{\phi}_1^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2 - \lambda\boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 - \alpha\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = 0.$$

Koska kaksi ensimmäistä termiä ovat yhtä suuria kuin nolla ja $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = 1$, Langrangen kertoimen α on oltava yhtä suuri kuin nolla. Tällöin

$$\boldsymbol{\Sigma}\boldsymbol{\phi}_2 - \lambda\boldsymbol{\phi}_2 = 0,$$

joten λ on kovarianssimatriisin $\boldsymbol{\Sigma}$ ominaisarvo ja $\boldsymbol{\phi}_2$ vastaava ominaisvektori. Koska $Var(\boldsymbol{\phi}_2^T\boldsymbol{x}_i) = \boldsymbol{\phi}_2^T\boldsymbol{\Sigma}\boldsymbol{\phi}_2 = \lambda$, ominaisarvon λ on oltava suurin mahdollinen. Kuitenkin jos $\lambda = \lambda_1$, niin $\boldsymbol{\phi}_2 = \boldsymbol{\phi}_1$, mikä rikkoo oletusta $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 = 0$, koska $\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 = 1$. Tästä syystä ominaisarvon λ on oltava toiseksi suurin λ_2. Seuraavat pääkomponentit lasketaan samalla periaatteella, mutta laskut monimutkaistuvat. Yleisesti yksilön i p:nnes pääkomponentti on kuten edellä on todettu $\boldsymbol{\phi}_p^T\boldsymbol{x}_i$ ja tällöin kyseisen komponentin otosvarianssi on λ_p, jossa λ_p on p:nneksi suurin kovarianssimatriisin $\boldsymbol{\Sigma}$ ominaisarvo ja $\boldsymbol{\phi}_p$ on vastaava ominaisarvovektori. [20]

Algoritmi 1 Pääkomponenttien muodostaminen

$\boldsymbol{\Sigma} \leftarrow \frac{1}{n-1}\boldsymbol{x}_i^T\boldsymbol{x}_i$
$\boldsymbol{\Sigma}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$

Algoritmilla 1 muodostetaan pääkomponentit. Oletetaan, että selittäjät x_i, $i = 1,..,n$, on skaalattu. Algoritmissa lasketaan ensin kovarianssimatriisi $\boldsymbol{\Sigma}$. Tämän jälkeen algoritmi laskee kovarianssimatriisin ominaisarvohajotelman. Algoritmin $\boldsymbol{v}_i$ on i:nnes omaisvektori ja λ_i vastaava i:nnes ominaisarvo. Muodostetut ominaisvektorit vastaavat pääkomponentteja.

2.2 Osittaisen pienimmän neliösumman regressio

OPNR-menetelmä tuottaa PKR-menetelmän tavoin riippumattomia projektioita $z_{i1}, \ldots, z_{ip}$, jotka selittävät mahdollisimman paljon alkuperäisten selittäjien variansista [19]. Samaan tapaan kuin PKR-menetelmässä, myös OPNR-menetelmässä on tärkeää skaalata selittäjät, jotta muuttujat pysyvät vertailukelpoisina [16]. Lisäksi menetelmän käyttämisen yhteydessä täytyy keskistää vaste. Menetelmä sopii erityisesti tilanteisiin, joissa selittäjiä on enemmän kuin havaintoja, koska menetelmä ei käytä kovarianssirakennetta kuten PKR-menetelmä. [21]

Ensimmäiset komponentit z_{i1} ovat aivan kuten PKR-menetelmässä alkuperäisten selittäjien $x_{i1}, ..., x_{im}$ lineaarikombinaatioita,

$$z_{i1} = \phi_{11}x_{i1} + \cdots + \phi_{m1}x_{im} = \boldsymbol{\phi}_1^T \boldsymbol{x}_i, \quad i = 1, \ldots, n,$$

Yleisesti p:nnet komponentit z_{ip} voidaan kirjoittaa muodossa,

$$z_{ip} = \phi_{1p}x_{i1} + \cdots + \phi_{mp}x_{im} = \boldsymbol{\phi}_m^T \boldsymbol{x_i}, \quad i = 1, \ldots, n,$$

jotka ovat korreloimaton muiden komponenttien kanssa. OPNR-menetelmä projisoi kaikki selittäjät komponenteiksi käyttämällä korrelaatioita vasteen ja selittäjien välillä [21]. Tällöin OPNR-menetelmä painottaa eniten niitä muuttujia, jotka ovat eniten yhteydessä vasteen kanssa [16]. Komponentit voidaan muodostaa algoritmilla 2, jossa $\boldsymbol{X}$ on selittäjistä koostuva $n \times p$ matriisi, $\boldsymbol{y}$ on vasteesta koostuva keskistetty vektori, $\boldsymbol{z}_k$ on k:nnes komponentti ja $\boldsymbol{g}$ on painokerroinmatriisi. Algoritmissa 2 alustetaan ensin muuttuja $\boldsymbol{x}_0$ alkuperäisellä selittävien muuttujien matriisilla $\boldsymbol{x}$. Seuraavaksi algoritmi käy läpi komponenttien muodostamisen. Ensimmäiset komponentit z_{i1} ovat matriisin $\boldsymbol{x}_0$ tulo transpoosinsa ja vastevektorin kanssa. Seuraavaksi lasketaan seuraavien komponenttien painokerroinmatriisi $\boldsymbol{g}_k$. Viimeisenä päivitetään selittävien muuttujien matriisi $\boldsymbol{X}_k$ kertomalla selittävien muuttujien matriisi painokerroinmatriisilla $\boldsymbol{g}_k$. Algoritmista saadaan tuloksena p kappaletta komponentteja.

Algoritmi 2 Komponenttien muodostaminen OPNR-menetelmässä

$\boldsymbol{X}_0 \leftarrow \boldsymbol{X}$
for $k \leftarrow (1 : p)$ **do**
  $z_{ik} \leftarrow \boldsymbol{X}_{k-1}\boldsymbol{X}_{k-1}^T\boldsymbol{y}$
  $\boldsymbol{g}_k \leftarrow \boldsymbol{i}_n - z_{ik}(z_{ik}^T z_{ik})^{-1} z_{ik}^T$
  $\boldsymbol{X}_k \leftarrow \boldsymbol{g}_k \boldsymbol{X}_{k-1}$
end for,

3 Kutistamismenetelmät

Vaihtoehtona luvussa 2 esitellyille dimension pienentämismenetelmille voidaan usean selittäjän hallintaan käyttää erilaisia kutistamismenetelmiä. Tässä luvussa esitellään kutistamismenetelmistä harja- (*ridge regression*) ja LASSO-regressio (*least absolute shrinkage and selection operator regression*) [17]. Menetelmät pyrkivät siirtämään regressiokertoimien arvoja kohti nollaa, mikä vähentää kertoimien varianssia ja minimoi ennusteharhaa. Harjaregressiossa säilyvät kaikki alkuperäiset selittäjät, kun taas LASSO-regressiossa osa regressiokertoimista pienenee nolliksi, jolloin kaikki selittäjät eivät jää malliin [16].

Jäännösneliösumma

$$JNS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2$$

kertoo, kuinka paljon mallin ennustamat arvot $\hat{y}_i$ poikkeavat todellisista havaituista y_i arvoista. Pienimmän neliösumman (PNS) estimaattori estimoi regressiokertoimet $\beta_0, \ldots, \beta_p$, minimoimalla jäännösneliösumman (JNS), toisin sanoen

$$\hat{\beta} = \text{argmin}_\beta\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2\right\} = \text{argmin}_\beta\Big\{JNS\Big\}.$$

Analyysissa on käytetty R-ohjelmistoa ja *cv.glmnet*-funktiota (paketissa *glmnet*).

3.1 Harjaregressio

Harjaregressio minimoi jäännösneliösumman ja kutistusrangaistustermin summan:

$$\hat{\beta} = \text{argmin}_\beta\left\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right\} = \text{argmin}_\beta\left\{JNS + \lambda\sum_{j=1}^{p}\beta_j^2\right\},$$

jolloin menetelmä sakottaa suurista regressiokertoimien arvoista. Kutistusrangaistustermi $\lambda\sum_{j=1}^{p}\beta_j^2$ ohjaa sakkokertoimen λ avulla jäännösneliösumman ja kutistusrangaistustermin yhteisvaikutusta regressiokertoimiin $\beta_1, \ldots, \beta_p$. Kun $\lambda = 0$, sakkotermillä ei ole vaikutusta ja jäljelle jää pelkkä jäännösneliösumma. Toisaalta jos $\lambda \to \infty$, niin sakkotermin vaikutus kasvaa ja kertoimet lähestyvät nollaa. [19]

Toisin kuin PNS-menetelmä, joka tuottaa vain yhden joukon regressiokertoimia, harjaregressio tuottaa regressiokertoimet erikseen jokaiselle λ-parametrille [16]. Tällöin keskiössä on se, kuinka valitaan λ-parametrin arvo. Yksi keino optimoida parametrin valintaa on käyttää ristiinvalidointia, jota käsitellään luvussa 4. Harjaregressio vaatii skaalatut selittäjät [19]. Tämän lisäksi on huomioitava, onko mallissa vakiotermi, jota ei ole tarpeen kutistaa. Jos selittäjät on keskistetty ennen harjaregressiota, vakiotermi on muotoa $\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

3.2 LASSO-regressio

LASSO-regressio on harjaregression tapaan kutistamisrangaistusmenetelmä, mutta se käyttää eri sakkotermiä. LASSO-regressio sakottaa mallin monimutkaisuudesta

tavoitteena luoda mahdollisimman yksinkertainen malli [17]. Harjaregression heikkous on, että menetelmä säilyttää kaikki alkuperäiset muuttujat, mikä voi vaikuttaa ennustuksien tarkkuuteen. Tätä ongelmaa ei esiinny LASSO-regressiossa, sillä menetelmä kutistaa joidenkin muuttujien regressiokertoimet nolliksi, jolloin osa selittäjistä jää pois mallista. [16]

LASSO-regressio minimoi jäännösneliösumman ja kutistusrangaistustermin summan seuraavasti:

$$\hat{\beta} = \text{argmin}_\beta \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\} = \text{argmin}_\beta \left\{ JNS + \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Lasso-regression rangaistustermiä $\sum_{j=1}^{p} |\beta_j|$ kutsutaan L_1-termiksi, koska sakkotermeissä olevat normit ovat L_1-normeja. Vastaavasti harjaregressiossa termiä $\sum_{j=1}^{p} \beta_j^2$ kutsutaan L_2-termiksi, koska normit sakkotermissä ovat L_2-normeja. [19]. Samoin kuin harjaregressiossa, LASSO-regressio kutistaa kertoimia kohti nollaa. L_1-termi pakottaa osan kertoimista nolliksi, erityisesti silloin kun λ-parametrin arvo on suuri. Tämä ominaisuus tekee LASSO-regressiosta tehokkaan muuttujavalinnassa, mikä helpottaa tuloksien tulkitsemista. [16]

Kuten harjaregressiossa myös LASSO-regressiossa on otettava huomioon skaalaus. Jos selittäjät on skaalattu, vakiotermi $\hat{\beta}_0$ on havaintojen keskiarvo $\bar{y}$ ja tällöin malli sovitetaan ilman vakiotermiä [19]. LASSO- ja harjaregression keskiössä on λ-parametrin valinta, jota voidaan optimoida ristiinvalidoinnilla. [16]

3.3 Harja- ja LASSO-regression bayesläinen tulkinta

Harja- ja LASSO-regressiota voidaan tarkastella myös Bayes-malleina. Oletetaan, että regressiokerroinvektorilla $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)^T$ on priorijakauma $p(\boldsymbol{\beta})$. Uskottavuusfunktio voidaan kirjoittaa muodossa $f(\boldsymbol{y}|\boldsymbol{\beta}; \boldsymbol{x})$, jossa $\boldsymbol{x} = (\boldsymbol{x}_1^T, ..., \boldsymbol{x}_n^T)^T$. Posteriorijakauma saadaan verrannollisuuskerrointa vailla kertomalla priorijakauma uskottavuusfunktiolla:

$$p(\boldsymbol{\beta}|\boldsymbol{x}, \boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\beta}; \boldsymbol{x}) p(\boldsymbol{\beta}).$$

Oletetaan, että $p(\boldsymbol{\beta}) = \prod_{j=1}^{p} g(\beta_j)$, jossa $g(\cdot)$ on normaalijakauman $N(0, \lambda)$ tiheysfunktio. Tällöin posteriorijakauman moodi on sama kuin harjaregression antama piste-estimaatti. Toisaalta, jos oletetaan, että $g(\cdot)$ noudattaa kaksoiseksponenttijakaumaa (Laplace-jakauma) keskiarvolla nolla ja skaalaparametrin arvolla λ, posteriorijakauman moodi on sama kuin LASSO-regression antama piste-estimaatti. [16]

4 Ristiinvalidointi

Tässä luvussa perehdytään ristiinvalidointiin, joka on toisto-otantamenetelmä. Menetelmässä aineisto jaetaan opetus- ja testausaineistoihin [16]. Opetusaineistoa käytetään mallin sovittamiseen, ja testausaineiston selittäjien avulla ennustetaan vastemuuttujan arvoja, joita verrataan testausaineiston todellisiin havaintoihin. Aineiston jako ja vertailu toistetaan useita kertoja, ja menetelmä, esimerkiksi pääkomponenttiregressio, sovitetaan kuhunkin opetusaineistoon [17]. Ristiinvalidoinnilla tutkitaan sovitetun mallin sopivuutta esimerkiksi estimoimalla ennustevirhettä [16].

Ristiinvalidointimenetelmiä ovat esimerkiksi yksi-pois-ristiinvalidointi ja k-kertainen ristiinvalidointi. K-kertainen ristiinvalidointi antaa usein tarkempia estimaatteja kuin yksi-pois -ristiinvalidointi. Tässä tutkielmassa käytetään k-kertaista ristiinvalidointia. [16]

Ristiinvalidointi on analyysissa toteutettu R-ohjelmistolla itse tehdyllä funktiolla dimension pienentämismenetelmissä ja *cv.glmnet*-funktion avulla kutistamismenetelmissä.

4.1 K-kertainen-ristiinvalidointi

K-kertaisessa ristiinvalidoinnissa havainnot jaetaan satunnaisesti k yhtä suureen osaan [16]. Yleensä k:ksi valitaan joko viisi tai kymmenen. Yleisesti viisi- ja kymmenkertaiset ristiinvalidoinnit johtavat malliin perustuvissa ennusteissa sekä maltilliseen varianssiin että harhaan. Jos opetusaineistoon sovitetussa lineaarisessa mallissa on huomattava kulmakerroin, viisi- ja kymmenkertainen ristiinvalidointi kuitenkin yliarvioi helposti ennustevirheen [19].

K-kertaisessa-ristiinvalidoinnissa jokainen osa-aineisto on vuorollaan testausaineisto ja haluttu malli sovitetaan lopuista $K-1$ osasta koostuvaan opetusaineistoon. Tämän jälkeen lasketaan usein joko keskineliövirhe (*Mean Square Error*, KNV) tai jäännösvirhehajonta (*Root Mean Square Error*, JVR) testausaineiston havaintojen ja ennusteiden välillä. Keskineliövirhe (KNV) lasketaan kaavalla

$$KNV = \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{n} = \frac{JNS}{n}, \tag{2}$$

jossa jäännösneliösumma (JNS) jaetaan havaintojen lukumäärällä n. Jäännösvirhehajonta (JVR) lasketaan kaavalla

$$JVR = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{n}} = \sqrt{KNV}. \tag{3}$$

Keskineliövirhe ja jäännösvirhehajonta tuottavat pieniä arvoja, mikäli ennusteet ovat lähellä todellisia arvoja. Ristiinvalidointi toistetaan k kertaa niin, että jokainen osa on vuorollaan testausaineistona. Lopulta ristiinvalidoinnin keskimääräinen virhe (*Cross Validation*, CV) on

$$CV = \frac{1}{k}\sum_{i=1}^{k} JVR_i.$$

Pienempi CV-arvo osoittaa parempaa mallin sovitusta ja yleistettävyyttä. [16]

Ristiinvalidoinnin avulla voidaan etsiä keskineliövirheen minimikohta eri parametrin arvoilla. PKR- ja OPNR-menetelmissä komponenttien lukumäärä voidaan optimoida ristiinvalidoinnin avulla. Tällöin funktio laskee jokaisella komponentin arvolla jäännösvirhehajonnan. Komponenttien lukumääräksi valitaan se, joka minimoi virheen. Harja- ja LASSO-regressioissa λ-parametrin valinta on keskiössä. Ristiinvalidointi tarjoaa tehokkaan keinon λ-parametrin arvon optimoimiseen. Ristiinvalidoinnilla lasketaan eri λ-parametrin arvoilla jäännösvirhehajonnan arvoja, ja λ-parametrin arvoksi valitaan se, joka minimoi virheen. [16]

5 Menetelmien vertailu

Tässä luvussa vertaillaan edellisissä luvuissa esiteltyjä menetelmiä ja tutustutaan menetelmien harha-varianssikompromissiin. Sopivan menetelmän löytäminen vaatii aineiston ja tutkimusongelman tuntemista. Erilaisia menetelmiä voidaan vertailla toisiinsa laskemalla mallien kyky ennustaa tulevia arvoja. Ennustettuja arvoja voidaan verrata todellisiin arvoihin laskemalla näiden välisen virheen suuruus keskineliövirheen kaavalla (2) tai jäännösvirhehajonnan kaavalla (3). Lisäksi voidaan verrata mallien selitysastetta R^2, joka kertoo kuinka hyvin malli selittää vastetta. [16]

Selitysaste R^2 lasketaan kaavalla

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2}{\sum_{i=1}^{n}(y_i - \bar{y}^2)} = 1 - \frac{JNS}{KNS},$$

jossa jälkimmäisen termin osoittajassa on jäännösneliösumma ja nimittäjässä kokonaisneliösumma (KNS). Selitysaste suurenee, mitä suurempi osuus kokonaisneliösummasta voidaan selittää regressiomallilla. Tällöin R^2-luku suosii monimutkaisempia malleja, mikä voi johtaa ylisovittamiseen ja multikollineaarisuuteen. Ongelman korjaa muokattu selitysaste R_m^2,

$$R_m^2 = 1 - \frac{JNS/(n-p-1)}{KNS/(n-1)},$$

jossa p on selittäjien lukumäärä ja n havaintojen lukumäärä. Muokattu selitysaste sakottaa suuresta määrästä selittäjiä ja tarjoaa paremman arvion mallin yleistettävyydestä. [16]

5.1 Harha-varianssikompromissi

Aineisto voidaan jakaa opetus- ja testausaineistoihin, joilla testataan mallin sopivuutta. Opetusaineistoa käytetään mallin sovittamiseen, ja testausaineiston selittäjiä ennusteiden luomiseen. Ennusteita verrataan testausaineiston todellisiin havaintoihin. [16]

Harha-varianssikompromissilla tarkoitetaan ennustemallin monimutkaisuuden ja suorituskyvyn tasapainottamista. Tällä tarkoitetaan sitä, että usein malleissa harhan suuruus kasvaa, kun varianssi pienenee, ja päinvastoin. Harha kuvaa ennustemallin keskimääräistä poikkeamaa todellisista arvoista. Esimerkiksi jos valitaan lineaarinen regressiomalli kuvaamaan monimutkaista ilmiötä, malli ei ota huomioon kaikkia aineiston yksityiskohtia, mikä johtaa mallin harhaan. Varianssi puolestaan kertoo, kuinka paljon ennusteet vaihtelevat eri opetusaineiston otoksilla. Monimutkaisemmalla mallilla voidaan oppia monimutkaisiakin aineistoja, mutta tämä voi johtaa suureen vaihteluun ennustuksissa, mikä lisää varianssia. [16]

Ennustettuja arvoja voidaan verrata todellisiin arvoihin laskemalla näiden välisen virheen suuruus keskineliövirheen kaavalla (2). Hyvässä mallissa keskineliövirhe on pieni, jolloin varianssi ja harha ovat kohtuullisia. Tällöin sovitettu menetelmä kuvaa hyvin havaintoaineistoin säännönmukaisuuksia mutta yleistyy myös ennustuksiin. Joissain tilanteissa on syytä luopua parametriestimaattorien harhattomuudesta ja käyttää estimaattoreita, joiden varianssi on suhteessa pienempi. Esimerkiksi

monimutkainen malli voi tuottaa pienen harhan mutta suuren varianssin, kun taas liian yksinkertainen malli voi johtaa pieneen varianssiin mutta suureen harhaan. Tavoitteena on löytää sopiva kompromissi harhan ja varianssin välillä. [16]

PKR-menetelmässä harhan suuruus pienenee, mitä enemmän pääkomponentteja valitaan. Tällöin varianssi kasvaa, koska suurin osa alkuperäisestä varianssista on selitetty. Sama ilmiö tapahtuu OPNR-menetelmässä. Harjaregressiossa, kun λ-parametrin arvo kasvaa, harjaregression sovituksen joustavuus laskee, mikä johtaa varianssin pienenemiseen, mutta samalla harhan kasvuun. Tämä sama ilmiö tapahtuu LASSO-regressiossa. Ongelmana on löytää sellainen λ-parametrin arvo, jolla ennustevirhe on pienin mahdollinen. Tähän voidaan käyttää ristiinvalidoinnilla. [16]

5.2 Menetelmien hyödyt ja haitat

Regressioanalyysin yhteydessä voidaan käyttää menetelmiä, kuten osittaisen pienimmän neliösumman regressiota, pääkomponentti-, harja- tai LASSO-regressiota. Menetelmät tarjoavat erilaisia tapoja etsiä vasteen ja monen selittäjän välistä yhteyttä.

PKR-menetelmän etuna on kyky käsitellä useita selittäjiä ja niiden korrelaatiota. Haittapuolena on, ettei menetelmä ota huomioon komponentteja muodostaessaan selittäjien yhteyttä vasteeseen. OPNR-menetelmä käsittelee samanaikaisesti sekä selittäviä muuttujia että vastemuuttujaa. Tästä voisi päätellä, että OPNR-menetelmän toimivan paremmin verrattuna PKR-menetelmään. On kuitenkin havaittu, että OPNR-menetelmä ei yleisesti ottaen toimi sen paremmin kuin PKR-menetelmä. PKR toimii parhaiten tilanteessa, jossa muutama ensimmäinen komponentti selittää suurimman osan varianssista. Molemmat menetelmät vaativat komponenttien lukumäärän valintaa, joka tapahtuu subjektiivisesti. Komponenttien lukumäärä vaikuttaa tuloksiin, jolloin niiden valinta on riskialtista. Komponentteja on myös vaikea tulkita. [16]

PKR- ja OPNR-menetelmien haittana on se, että malliin jää komponentteja, joita on vaikea sanallisesti tulkita. LASSO- ja harjaregressiossa tuloksien tulkitseminen käy helposti, koska lopullisessa mallissa on alkuperäisiä muuttujia. Harjaregression haittapuolena on kuitenkin se, että menetelmä säilyttää kaikki alkuperäiset muuttujat, mikä saattaa johtaa ylisovittamiseen [17]. Tätä ongelmaa ei ole LASSO-regressiossa, jossa usein malliin jää vain osa alkuperäisistä muuttujista. Tästä voisi päätellä LASSO-regression toimivan paremmin verrattuna harjaregressioon, mutta näin ei kuitenkaan aina ole. LASSO-regressiossa varianssin pienentyessä harhan suuruus kasvaa. Näin ollen LASSO-regressio saattaa tuottaa aivan samankaltaisia tuloksia harjaregressioon verrattuna. Yleisesti voidaan odottaa, että LASSO-regressio toimii harjaregressiota paremmin tilanteissa, joissa on pieni määrä selittäjiä [16]. Harjaregressio taas toimii paremmin tilanteissa, joissa on useita selittäjiä, joiden kertoimet ovat lähellä toisiaan [17]. Etukäteen ei voida kuitenkaan arvioida merkittävien selittäjien määrää, jolloin on hyvä kokeilla molempia menetelmiä. Lisäksi joissakin tapauksissa LASSO-regressio saattaa asettaa kaikki selittäjät yhtä suuriksi kuin nolla, jolloin ennustaminen on mahdotonta. Tällöin harjaregressio, PKR- tai OPNR-menetelmä ovat parempia vaihtoehtoja. [16]

Harja- ja LASSO-regressiossa sakkoparametri λ vaihtelee jatkuvana arvona, joka

vaikuttaa kutistamisen suuruuteen. PKR- ja OPNR-menetelmissä komponenttien lukumäärä valitaan diskreetillä asteikolla. Harjaregressio kutistaa regressiokertoimia, kunnes ne lähestyvät PNS-menetelmän kertoimia. OPNR- ja PKR-menetelmät toimivat samoin muuten, paitsi että ne toimivat diskreetillä asteikolla. Harjaregressio kutistaa joka suuntaan, mutta se kutistaa matala-varianssi-tilanteessa enemmän. PKR-menetelmässä malliin jää pieni määrä komponentteja, jotka selittävät varianssista suurimman osan. OPNR-menetelmä toimii PKR-menetelmän tavoin, mutta pystyy myös kutistamaan matalavarianssisiin suuntiin. Tämä tekee OPNR-menetelmän käytöstä epävakaata, joka johtaa harjaregressiota suurempaan ennustevirheeseen. Ennustevirheen minimoimisen näkökulmasta harjaregressiota suositaan ennemmin kuin PKR- ja OPNR-menetelmiä. [19]

Menetelmät eroavat toisistaan ja sopivat näin ollen ennustamaan erilaisia aineistoja. Etukäteen on kuitenkin vaikea arvioida, mikä menetelmistä sopisi kulloinkin käsiteltävään aineistoon. Tällöin ristiinvalidoinnilla voidaan arvioida esimerkiksi ennustevirheen näkökulmasta parasta mallia [16].

6 Aineiston analysointi

Tässä luvussa tarkastellaan aineistoa, sovelletaan menetelmiä aineistoon ja esitellään tulokset. Tutkielmassa käytetään LASERI-tutkimuksen vuosien 2018–2020 tutkimuskäynneillä kerättyä aineistoa [15]. Aineistossa on 814 yksilöä, jotka ovat 59–88-vuotiaita (G_0-sukupolvi) ja 1 237 yksilöä, jotka ovat 41–58-vuotiaita (G_1-sukupolvi). Eri sukupolvien G_0 ja G_1 muodostamia aineistoja tutkittiin sekä yhdessä että erikseen.

6.1 Aineisto

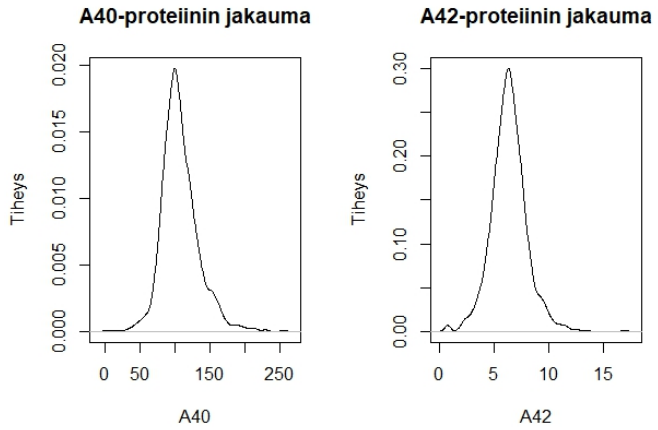
LASERI-tutkimuksen aineistoissa on arvioitu kognitiivista toimintaa CANTAB-testillä [15]. CANTAB-testi on kulttuurisesti neutraali testi, joka suoritetaan elektronisella alustalla. Kulttuurisesti neutraalilla testillä tarkoitetaan sitä, että siinä on mahdollisimman vähän käytetty esimerkiksi kielellisiä ohjeita, jotka rajoittavat käyttöä eri maissa. Testin avulla mitataan eri kognition osa-alueita, kuten muistia ja oppimista. Kognitiotestejä on viittä alatyyppiä:

- motorinen seulontatesti (*motor screening Test*, MOT), joka mittaa psykomotorista nopeutta ja tarkkuutta;
- oppimis- ja muistitesti (*learning and memory*, PAL), jossa mitataan visuaalista ja episodista muistia sekä visuaalista avaruudellista oppimista;
- työmuistitesti (*working memory*, SWM), joka mittaa lyhytaikaista ja spatiaalista työmuistia sekä ongelmanratkaisukykyä;
- informaation käsittelytesti (*information processing*, RTI), joka mittaa visuaalista käsittelyä, tunnistusta ja pysyvää huomiota;
- reaktioaikatesti (*reaction time*, RVP), joka mittaa reaktioaikaa sekä liikkeen nopeutta.

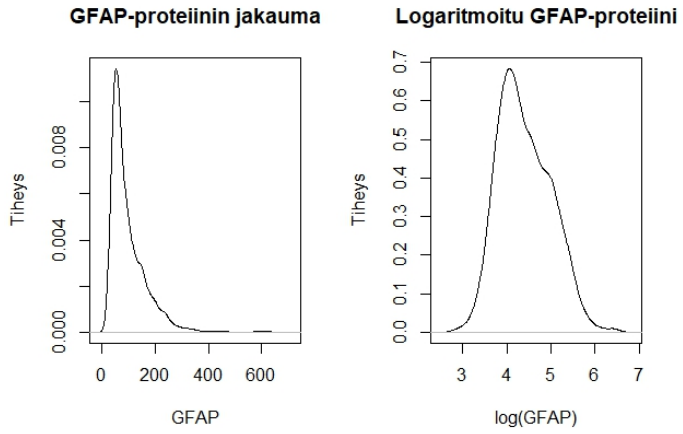
Jokaiseen testiluokkaan kuuluu useita kumulatiivisesti vaikeutuvia testejä, ja yhteensä testejä on 24. Aineistoon on muodostettu näiden testien tuloksista 68 muuttujaa esimerkiksi laskemalla testien virheiden lukumäärää, yrityksien lukumäärää ja latenssiaikoja. Aineistossa on 4 MOT-testin tulosta, 21 PAL-testin tulosta, 10 RTI-testin tulosta, 24 SWM-testin tulosta ja 9 RVP-testin tulosta. LASERI-tutkimuksessa MOT-testiä käytetään elektronisen alustaan tutustumiseen eikä se erottele terveitä tutkittavia. Tästä syystä tässä tutkielmassa ei myöskään oteta huomioon MOT-testin tuloksia. Tällöin kognitiotestien tuloksia on yhteensä 64. SWMPR eli esitettyjen tehtävien lukumäärä on kaikille tutkittaville sama, jolloin tämä tulos on mielenkiinnoton. Tästä syystä tämä testi poistetaan aineistosta, ja jäljelle jää 63 kognitiotestin tulosta jokaiselta yksilöltä. Tämän lisäksi tutkittavilta on mitattu NfL-, Aβ40-, Aβ42- ja GFAP-proteiinipitoisuudet verinäytteestä Simoa-menetelmällä (*single molecule array HD-X Analyzer*), joka mahdollistaa pienten konsentraatioiden mittaamisen.

Proteiinipitoisuuksissa oli jonkin verran oudokkeja, jotka tunnistettiin laskemalla jokaiselle mittaukselle z-arvot (normalisoidut poikkeamat). Oudokit päädyttiin

poistamaan aineistosta. G_0-sukupolven muodostamassa aineistossa oudokkeja oli 26, jolloin aineistoon jäi 788 yksilöä. G_1-sukupolven muodostamassa aineistoissa oudokkeja oli 39, jolloin aineistoon jäi 1 198 yksilöä. Molemmat sukupolvet yhdistävässä aineistossa oudokkeja oli 52, jolloin aineistoon jäi 1 999 yksilöä.

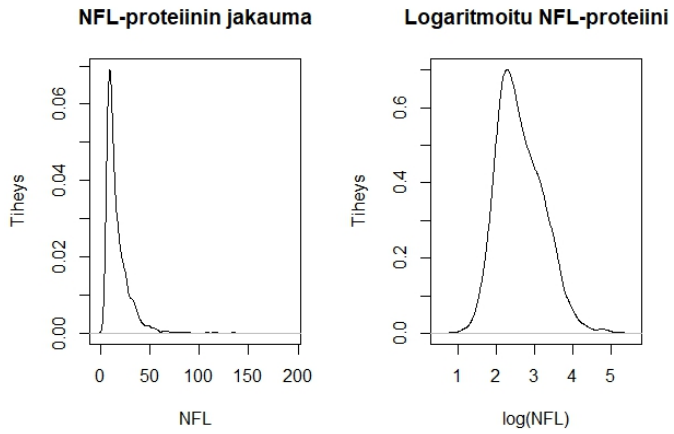


Kuva 1: Aβ40- ja Aβ42-proteiinien siloitetut tiheysjakaumat. Havaintoja on 1 999.



Kuva 2: GFAP-proteiinin siloitettu tiheysjakauma ja logaritmimuunnettujen pitoisuuksien jakauma. Havaintoja on 1 999.

Jokaista proteiinia tutkittiin erikseen niin, että proteiinipitoisuutta selitettiin kognitiotestin tuloksilla sekä taustamuuttujilla. Kuva 1 esittää koko aineiston Aβ40- sekä Aβ42-proteiinien siloitetut tiheysjakaumat, jotka noudattivat likimain normaalijakaumaa. Tällöin näihin proteiineihin ei ollut tarpeen tehdä muunnoksia jakaumiin. Tutkittiin myös jatkokysymyksenä vaikuttaako ikä näiden proteiinien pitkähäntäisyyteen. Kun nämä proteiinit jaettiin erilaisiin ikäluokkiin, jakaumat vaikuttivat edelleen likimain normaaleilta. Tällöin todettiin, ettei ikä selitä kyseistä ilmiötä. Kuvat 2 ja 3 esittävät GFAP- ja NfL-proteiinipitoisuuksien siloitettuja tiheysjakaumia ja logaritmimuunnettujen pitoisuuksien jakaumia. Logaritmimuunnokset



Kuva 3: NfL-proteiinin siloitettu tiheysjakauma ja logaritmimuunnettujen pitoisuuksien jakauma. Havaintoja on 1 999.

paransivat GFAP- ja NfL-proteiinien jakaumia. Näin ollen päätettiin jatkossa käyttää GFAP- ja NfL-proteiineja logaritmoituina. Vastaavat kuvaajat suoritettiin G_0- ja G_1-sukupolvien aineistoihin erikseen ja kuvaajat näyttivät näissä samoilta.

Verinäytteissä ei ollut puuttuvia havaintoja, mutta kognitiotesteissä oli puuttuvaa tietoa molemmissa aineistoissa. Tällöin oli tarpeen tutkia, vaikuttaako kognitiotestien puuttuminen eri vasteisiin eli proteiinipitoisuuksiin. Laskettiin pitoisuuksien erotuksen 95 %:n luottamusvälit niiden väliltä, joilta puuttui havaintoja kognitiotestien tuloksissa, ja niiden, joilta ei puuttunut havaintoja.

Sukupolvi	**Proteiini**	**Ei puutu**	***N***	**Puuttuu**	***N***	**Erotus**	**95 %**
$\boldsymbol{G_0}$	NFL	3.12	561	3.21	227	-0.09	[-0.16, -0.03]
	Aβ42	6.76	561	6.81	227	-0.05	[-0.32, 0.22]
	Aβ40	121.90	561	124.35	227	-2.45	[-6.37, 1.46]
	GFAP	4.89	561	4.99	227	-0.10	[-0.17, -0.03]
$\boldsymbol{G_1}$	NFL	2.27	1 105	2.28	93	-0.01	[-0.08, 0.06]
	Aβ42	6.10	1 105	6.01	93	0.09	[-0.22, 0.39]
	Aβ40	95.22	1 105	98.45	93	-3.23	[-6.94, 0.47]
	GFAP	4.04	1 105	4.09	93	-0.05	[-0.12, 0.04]

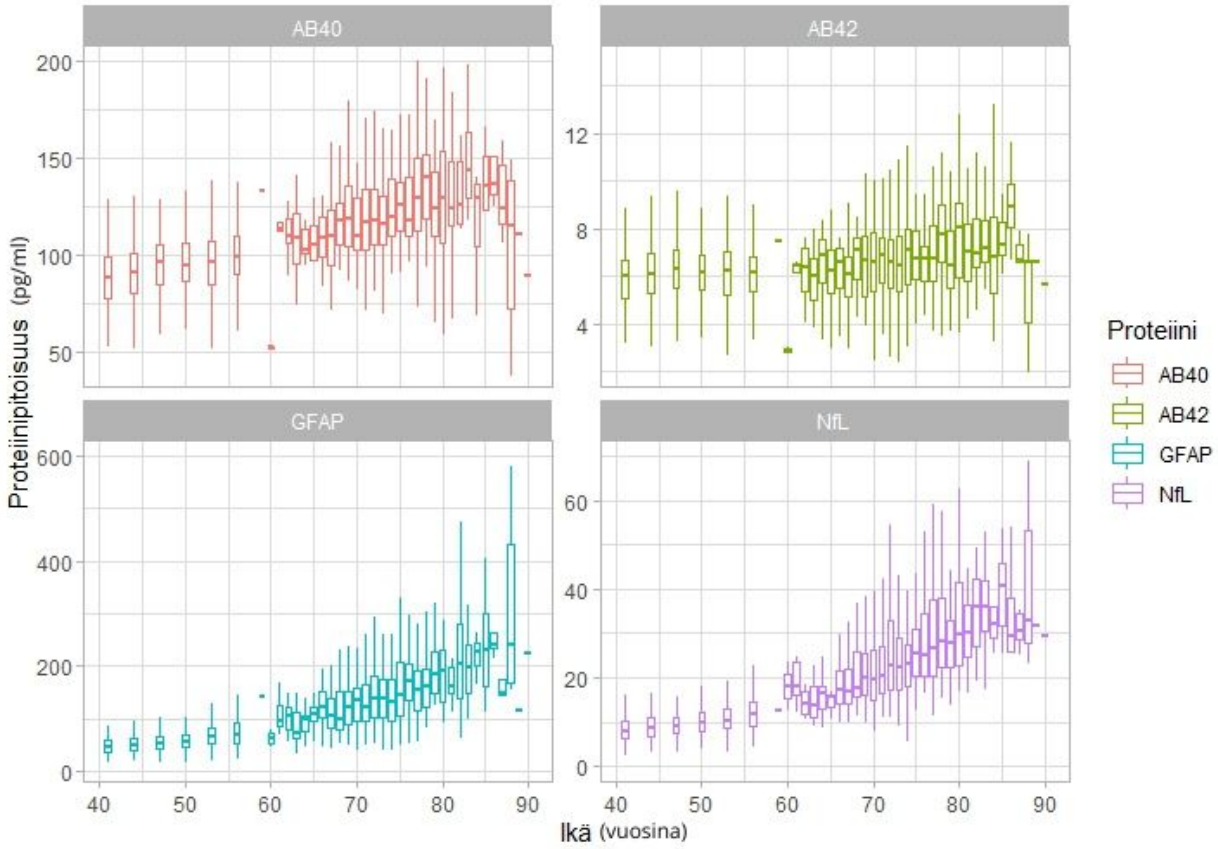
Taulukko 1: Proteiinipitoisuuksien (pg/ml) keskiarvot eri sukupolvien muodostamissa aineistossa sen mukaan, puuttuuko havaintoja vaiko ei. Taulukossa on esitetty myös havaintojen lukumäärät (N) ja pitoisuuksien erotus (ei puutu - puuttuu) ja sen 95 %:n luottamusväli. GFAP- ja NfL-proteiinit on logaritmoitu.

Taulukko 1 kertoo keskimääräiset proteiinipitoisuudet niillä, joilta havaintoja puuttui, ja niillä, joilta havaintoja ei puuttunut kognitiotestien tuloksissa eri osaaineistoissa. Näiden lisäksi taulukossa on näiden yksilöiden lukumäärät (N), keskiarvojen erotus ja 95 %:n luottamusvälit. Taulukosta nähdään, että G_0-sukupolven aineistossa logaritmoiduissa NfL- ja GFAP–pitoisuuksissa 95 %:n luottamusvälit eivät sisällä nollaa. Näin ollen näiden proteiinien keskiarvot poikkeavat niiden välil-

lä, joilta havaintoja puuttui, ja niiden, joilta havaintoja ei puuttunut. Niillä, joilla oli puuttuvaa tietoa, oli keskimäärin korkeammat proteiinien pitoisuudet. NfL- ja GFAP-proteiinien kohdalla ei siis voitu tehdä oletusta, että puuttuminen olisi täysin satunnaista. Tästä syystä kognitiotesteistä muodostettuihin selittäjiin päädyttiin käyttämään lähinaapuri-imputointimenetelmää, jolla imputoitiin puuttuvia arvoja.

G_1-sukupolven aineistossa kunkin proteiinin 95 %:n luottamusvälit sisälsivät nollan (Taulukko 1). Näin ollen kunkin proteiinin keskiarvot eivät poikkenneet niiden välillä, joilta havaintoja puuttui, ja niiden, joilta havaintoja ei puuttunut. Puuttumisen voitiin siis olettaa olevan satunnaista.

Taustamuuttujina analyyseissa käytettiin ikää, sukupuolta, nukuttuja tunteja, vireystilaa sekä opiskeluvuosia. Vireystila oli subjektiivisesti määritelty asteikolla 1–10 matalimmasta korkeimpaan koettuun vireystilaan. Vireystila oli luokiteltu uudestaan niin, että jäljelle jäi kolme kategoriaa. Ikä jaettiin neljään luokkaan, opiskeluvuodet jaettiin viiteen luokkaan ja nukutut tunnit säilytettiin jatkuvana muuttujana. Sukupuoli ilmaistiin indikaattorimuuttujana, jossa oletussukupuoli oli nainen.



Kuva 4: Aβ40-, Aβ42-, GFAP- ja NfL-proteiinipitoisuuksia (pg/ml) eri ikäisillä koko aineistossa (G_1- ja G_0-sukupolvet). Laatikot osoittavat, mille välille 50 % proteiinipitoisuuksista sijoittuu eri ikäisillä. Viiva laatikon sisällä kertoo havaintojen mediaanin ja laatikoiden pystyviivat osoittavat, kuinka 95 % havainnoista jakautuvat. Oudokit on poistettu aineistosta kuvaajien selkeyttämiseksi.

Kuva 4 esittää laatikkokuvaaja, joka kuvaa eri proteiinipitoisuuksia (pg/ml) eri

ikäisillä. Aβ40-, GFAP- ja NfL-proteiinipitoisuudet kasvavat keskimäärin iäänttyessä. Nuoremmissa ikäluokissa kasvu ei ole niin selkeää kuin vanhemmissa ikäluokissa. Aβ42-proteiinissa ei havaita samankaltaista selkeää kasvua iän myötä.

Taulukko 2: Keskimääräiset proteiinipitoisuudet (pg/ml) eri ikäisillä G_0- ja G_1-sukupolvien aineistoissa. N kertoo yksilöiden lukumäärän kyseisessä luokassa. Viimeisessä sarakkessa on esitetty ikäluokkaa vastaava numeerinen kategoria. NfL- ja GFAP-pitoisuudet on logaritmoitu.

Ikä (G_0)	**NfL**	**Aβ40**	**Aβ42**	**GFAP**	***N***	**Kategoria**
alle 67	2.79	110.97	6.21	4.62	73	1
67-75	3.04	118.91	6.62	4.81	404	2
75-83	3.33	129.46	7.08	5.10	271	3
yli 83	3.55	134.72	7.33	5.34	40	4
Yhteensä					788	
Ikä (G_1)						
alle 45	2.12	91.14	6.01	3.88	295	1
45–49	2.19	95.06	6.21	3.97	228	2
49–53	2.30	96.47	6.03	4.06	286	3
yli 53	2.40	98.25	6.12	4.21	389	4
Yhteensä					1 198	

Taulukko 2 esittää keskimääräiset proteiinipitoisuudet eri ikäisillä G_0- ja G_1-sukupolvien aineistoissa erikseen. Kuvaajasta nähdään, että pitoisuudet ovat keskimäärin korkeampia G_0-sukupolven aineistossa verrattuna G_1-sukupolven aineistoon. Kaikkien proteiinien pitoisuudet keskimääräiset kasvavat vanhemmissa ikäluokissa. Ainoa poikkeus on Aβ42-proteiinipitoisuus G_1-sukupolven aineistossa. Tämän proteiinin keskimääräinen suurin arvo on 45–49-vuotiailla.

Taulukko 3 esittää keskimääräiset proteiinipitoisuudet koko aineistossa, yksilöiden lukumäärän G_0- (N_0) ja G_1-sukupolven aineistossa (N_1) sekä sukupolvet yhdistävässä aineistossa (N) eri taustamuuttujien mukaisissa luokissa. Taulukosta nähdään, että kunkin proteiinin keskiarvot kasvoivat ikäluokkien mukaan.

Naisilla oli keskimäärin korkeampaa kunkin proteiinin pitoisuus miehiin verrattuna (taulukko 3). Alimmassa vireystilassa kunkin proteiinin pitoisuudet olivat keskimäärin matalampia verrattuna muihin vireystiloihin. Keskimmäisessä vireystilassa pitoisuudet olivat keskimäärin korkeimpia.

Alle 10 vuotta opiskelleilla oli keskimäärin suurimmat proteiinipitoisuudet koko aineistoissa (taulukko 3). Alle 10 vuotta opiskelleiden ikäjakaumasta nähdään kuitenkin, että suurin osa oli 65–77-vuotiaita, eli ikäjakauma oli painottunut vanhempiin yksilöihin. Tämä saattaisi vaikuttaa siihen, että alle 10 vuotta opiskelleilla oli keskimäärin korkeimmat proteiinipitoisuudet. NfL-, Aβ40-, GFAP-pitoisuudet pienenivät opiskeluvuosiin 20–25 asti, jolloin proteiinipitoisuudet olivat keskimäärin matalimmat. Näiden proteiinien pitoisuudet olivat keskimäärin korkeampia yli 25 vuotta opiskelleilla kuin 15–25 vuotta opiskelleilla. Yli 25 vuotta opiskelleita oli keskimäärin vähiten ja he olivat 41–81-vuotiaita. Ikä ei siis vaikuta taustalla keskiarvon suuruuteen yli 25 vuotta opiskelleilla. Kuitenkin vähäinen määrä henkilöitä saattoi aiheuttaa satunnaisuutta keskiarvoon. Keskimääräiset proteiinipitoisuudet vaihteli-

Taulukko 3: Keskimääräiset proteiinipitoisuudet (pg/ml) eri taustamuuttujilla koko aineistossa. N_0 on yksilöiden määrää G_0-aineistossa. N_1 on yksilöiden lukumäärää G_1-aineistossa. N on yksilöiden lukumäärää koko aineistossa. Viimeisessä sarakkessa on esitetty ryhmää vastaava numeerinen kategoria.

Ikä (vuosina)	**NfL**	**Aβ40**	**Aβ42**	**GFAP**	$\boldsymbol{N_0}$	$\boldsymbol{N_1}$	$\boldsymbol{N}$	**Kategoria**
alle 53	2.21	94.19	6.08	3.97	0	827	827	1
53–65	2.44	99.82	6.14	4.24	39	398	437	2
65–77	3.05	118.23	6.57	4.83	524	0	524	3
yli 77	3.39	130.32	7.10	5.17	211	0	211	4
Yhteensä					774	1 225	1 999	
Sukupuoli								
Naiset	2.62	105.77	6.37	4.45	485	707	1 192	0
Miehet	2.59	105.19	6.26	4.29	289	518	807	1
Yhteensä					774	1 225	1 999	
Vireystila								
1–4	2.43	101.23	6.14	4.26	5	19	24	1
4–7	2.62	106.12	6.35	4.39	480	756	1236	2
7–10	2.60	104.69	6.30	4.38	289	450	739	3
Yhteensä					774	1 225	1 999	
Opiskeluvuodet								
alle 10	3.15	123.56	6.78	4.88	275	13	287	1
10–15	2.61	105.72	6.33	4.40	344	457	798	2
15–20	2.44	99.97	6.15	4.22	134	582	721	3
20–25	2.35	98.09	6.27	4.14	16	145	160	4
yli 25	2.52	102.20	6.58	4.23	5	28	33	5
Yhteensä					774	1 225	1 999	
Nukutut tunnit								
alle 6	2.69	107.10	6.33	4.48	180	191	371	
6–8	2.59	105.27	6.33	4.35	441	757	1 198	
8–10	2.59	104.92	6.33	4.38	153	277	430	
Yhteensä					774	1 225	1 999	

vat satunnaisesti sen mukaan, miten henkilö oli nukkunut testejä edeltävänä yönä koko aineistossa.

6.2 Tulokset

Ristiinvalidoinnin avulla optimoitiin PKR- ja OPNR-menetelmissä komponenttien lukumäärät ennustevirheen näkökulmasta. Pääkomponentti-, osittaista pienimmän neliösumman , LASSO- ja harjaregressiota vertailtiin R^2-luvun sekä jäännösvirhehajonnan (JVR) avulla. Kutistamismenetelmien sakkotermin λ suuruutta optimoitiin myös ristiinvalidoinnin avulla. Tulokset esitellään tässä luvussa aineistojen mukaan, ensin G_0-sukupolven, sen jälkeen G_1-sukupolven ja viimeisenä koko aineiston tulokset. Liitteessä A on tarkempi kuvaus kognitiotesteistä, joita käytettiin analyysissa.

6.2.1 G_0-sukupolven tulokset

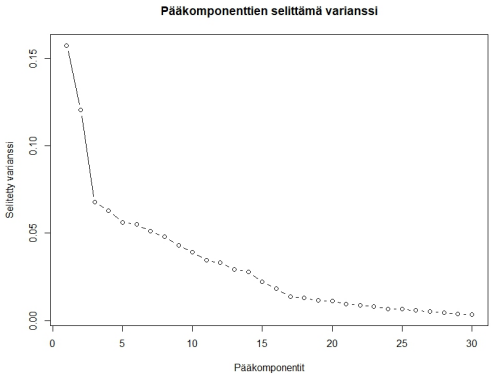
Kuva 5 esittää kolmenkymmenen ensimmäisen pääkomponentin selittämää varianssia, kumulatiivista varianssia ja ristiinvalidoinnin ennustevirhettä G_0-sukupolven aineistossa, kun vasteena oli NfL-proteiinipitoisuus ja menetelmänä pääkomponenttiregressio. Kuva 5a esittää pääkomponenttien selittämän varianssin osuutta. Kuvasta nähdään, että ensimmäiset komponentit selittävät suurimman osan varianssista, kun taas myöhempien komponenttien selitysvoima pienenee merkittävästi. Tämän kynnyskohdan perusteella voidaan valita komponenttien lukumäärä jatkoanalyysiin. Tässä tapauksessa kolme ensimmäistä komponenttia selittävät suuriman osan varianssista. Tästä syystä voitaisiin valita esimerkiksi kolme pääkomponenttia jatkoanalyysiin. Kuva 5b esittää pääkomponenttien kumulatiivisen varianssin selitysosuutta. Kuvan perusteella kolmekymmentä ensimmäistä komponenttia selittää melkein koko varianssin. Jos haluttaisiin esimerkiksi, että 70 % varianssista on selitetty, voitaisiin valita yksitoista ensimmäistä pääkomponenttia malliin. Kuva 5c esittää ristiinvalidoinnilla tuotettujen pääkomponenttien ennustevirheiden keskiarvoja. Kuvaajan perusteella voitaisiin valita neljä pääkomponenttia malliin, koska se minimoi ennustevirheen. Ennustevirhe kasvaa keskimäärin, mitä enemmän malliin otetaan mukaan komponentteja. Ennustevirhe kuitenkin pysyy suhteellisen pienenä, valitaan malliin yksi tai kolmekymmentä pääkomponenttia. Pääkomponenttien lukumäärä valittiin tässä tutkielmassa ristiinvalidoinnin avulla. Tästä syystä NfL-proteiinipitoisuutta selitettäessä regressiomalliin valittiin neljä pääkomponenttia, jotka minimoivat ennustevirheen.

Taulukko 4: Selittävien muuttujien lukumäärät eri menetelmillä G_0-sukupolven muodostamassa aineistossa. Lukumäärä perustuu kussakin tapauksessa ristiinvalidointiin. Harjaregressiossa selittäjien lukumäärä on 63.

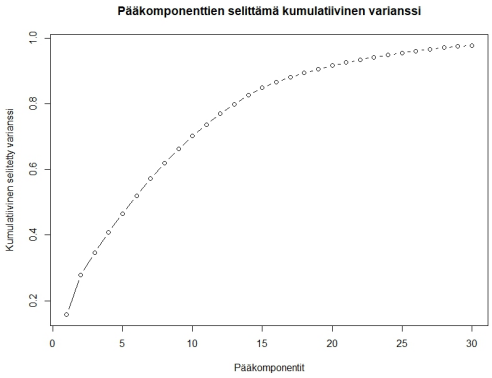
Menetelmä	**NfL**	**Aβ40**	**Aβ42**	**GFAP**
Pääkomponenttien lukumäärä **(PKR)**	4	1	6	2
Komponenttien lukumäärä **(OPNR)**	5	12	3	6
Selittäjien lukumäärä **(LASSO)**	8	12	2	12

Taulukko 4 esittää ristiinvalidoinnilla ennustevirheen minimoivien komponenttien lukumäärät PKR- ja OPNR-menetelmissä ja selittäjien lukumäärät LASSO-regressiossa G_0-sukupolven aineistossa. PKR-menetelmällä NfL-proteiinipitoisuutta selitettäessä malliin jäi neljä pääkomponenttia ja OPNR-menetelmällä viisi komponenttia. LASSO-regressiossa malliin jäi kahdeksan selittäjää, joista yksi oli vakio. Tämän lisäksi malliin jäi ikä, PAL-testin ensimmäisen yrityksen muistitulos (PALFAMS28) ja yrityksien lukumäärä (PALTA8), RTI-testin mediaanireaktioaika (RTIFMDRT) ja liikkumisen keskiarvoaika (RTIFMMT) sekä RVP-testin sensitiivisyys- (RVPA) ja kokonaisvirhemuuttuja (RVPTM).

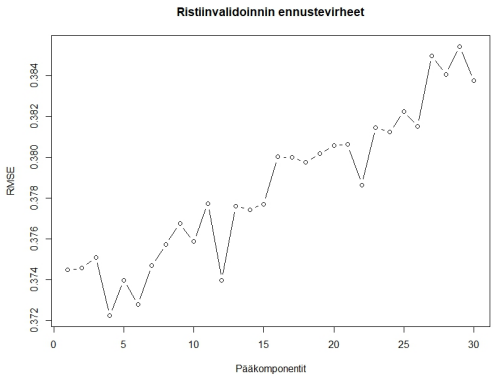
Kun vastemuuttujana oli Aβ40-proteiini PKR-menetelmällä vain yksi komponentti jäi malliin ja OPNR-menetelmällä kaksitoista komponenttia (taulukko 4). LASSO-regressiolla malliin jäi kaksitoista selittäjää, joista yksi oli vakio. Tämän lisäksi malliin jäi sukupuoli, ikä, opiskeluvuodet, PAL-testin yrityksien lukumäärä (PALTA28) ja virheiden lukumäärä (PALTE8), RTI-testin mediaaniaikamuuttu-



(a) Pääkomponenttien selittämä varianssi.



(b) Pääkomponenttien selittämä kumulatiivinen varianssi.



(c) Ristiinvalidoinnin ennustevirheet.

Kuva 5: G_0-sukupolven kolmenkymmenen ensimmäisen pääkomponentin selittämä varianssi, kumulatiivinen varianssi ja ennustevirhe, kun vasteena on logaritmoitu NfL-proteiinipitoisuus (pg/ml). Menetelmänä on käytetty pääkomponenttiregressiota.

jat (RTI-FM-DMT ja RTIFMDRT) ja virheiden lukumäärä (RTIFTES) sekä RVP-testin latenssiaika- (RVPMDL ja RVPML) ja kokonaisvirhemuuttujat (RVPTM).

Kun vastemuuttujana oli Aβ42-proteiini PKR-menetelmällä malliin jäi kuusi pääkomponenttia ja kolme komponenttia OPNR-menetelmällä (taulukko 4). LASSO-regressiossa malliin jäi vain kaksi selittäjää, jotka olivat vakio ja ikä.

Kun vastemuuttujana oli GFAP-proteiini PKR-menetelmällä malliin jäi kaksi pääkomponenttia ja kuusi komponenttia OPNR-menetelmällä (taulukko 4). LASSO-regressiossa malliin jäi kaksitoista selittäjää, joista yksi oli vakio. Tämän lisäksi malliin jäi sukupuoli, vireystila, ikä, opiskeluvuodet, PAL-testin virheiden keskiarvo (PALMETS28) ja yrityksien lukumäärä (PALTA2), RTI-testin virheiden lukumäärä (RTIFESI ja RTIFESPR), RVP-testin latenssiajan keskihajonta (RVPLSD) ja SWM-testin virheiden väli (SWMBE8). NfL-, Aβ40- ja GFAP-proteiinipitoisuuksia selitettäessä PKR-menetelmällä pienempi määrä komponentteja minimoi ennustevirheen verrattuna OPNR-menetelmään.

Taulukko 5 kertoo eri menetelmien ennustevirheet ja R^2-selitysasteet G_0-sukupolven aineistossa. NfL-proteiinipitoisuutta selitettäessä ennustevirhe sekä selitysas-

Taulukko 5: Ennustevirhe (jäännösvirhehajonta, JVR) ja R^2-selitysaste G_0-sukupolvessa eri menetelmiä käyttäen.

	NfL		**Aβ40**		**Aβ42**		**GFAP**	
Menetelmät	**JVR**	**R^2**	**JVR**	**R^2**	**JVR**	**R^2**	**JVR**	**R^2**
PKR	0.38	0.20	22.86	0.04	1.71	0.02	0.41	0.18
OPNR	0.36	0.20	23.36	0.04	1.73	-0.02	0.42	0.19
LASSO	0.38	0.19	24.81	0.03	1.65	0.02	0.42	0.12
Harja	0.39	0.16	25.15	-0.002	1.67	-0.01	0.42	0.13

teet olivat kohtuullisia. Noin 16–20 % aineiston vaihtelusta pystyttiin selittämään kyseisten mallin avulla. Ennustevirheet eri menetelmillä olivat kohtuullisen pieniä (jäännösvirhehajonta, JVR 0.36–0.38). Aβ40-proteiinipitoisuutta selitettäessä selitysasteet olivat heikkoja ja ennustevirheet suuria jokaisella menetelmällä. Harjaregressiolla selitysaste oli negatiivinen, mikä viittaa siihen, että ennuste oli huonompi kuin keskiarvo. Aβ42-proteiinipitoisuutta selitettäessä ennustevirheet olivat korkeita ja selitysasteet heikkoja. OPNR-menetelmällä ja harjaregressiolla selitysasteet olivat negatiivisia. GFAP-proteiinipitoisuutta selitettäessä ennustevirhe ja selitysaste olivat kohtuullisia. Noin 12–19 % aineiston vaihtelusta pystyttiin selittämään kyseisten mallien avulla. PKR-menetelmä toimi hieman muita menetelmiä paremmin Aβ40- ja GFAP-pitoisuuksia selitettäessä. OPNR-menetelmä toimi hieman muita menetelmiä paremmin NfL-proteiinipitoisuutta selitettäessä, kun taas, LASSO-regressio toimi hieman paremmin Aβ42-proteiinipitoisuutta selitettäessä. Erot kuitenkin olivat pieniä ja menetelmät toimivat tässä aineistossa yhtä hyvin toisiinsa nähden. Tästä syystä menetelmistä valittiin jatkotarkasteluun pääkomponenttiregressio (PKR).

Taulukko 6 kertoo pääkomponenttiregressiomallissa muuttujat, joilla oli yhteys proteiinipitoisuuksien kanssa G_0-sukupolven muodostamassa aineistossa. Yhteyden katsottiin olevan muuttujilla, joiden 95 %:n luottamusvälit eivät sisältäneet nollaa. Taulukko kertoo regressiokertoimet sekä niiden 95 %:n luottamusvälit. Tulokset osoittavat, että ikä oli positiivisesti yhteydessä NfL-proteiinipitoisuuden kanssa eli ikääntyessä proteiinipitoisuus kasvaa. Aβ40-proteiinipitoisuuden kanssa positiivisesti yhteydessä olivat sukupuoli, ikäluokat 75–83-vuotiaat ja yli 83-vuotiaat. Miehillä Aβ40-proteiinipitoisuus on korkeampi verrattuna naisiin ja ikääntyessä proteiinipitoisuus kasvaa. Aβ42-proteiinipitoisuuden kanssa positiivisessa yhteydessä olivat ikäluokat. Tällöin ikääntyessä proteiinipitoisuus kasvaa. GFAP-proteiinipitoisuuden kanssa yhteydessä olivat sukupuoli, ikäluokat ja yli 25 vuotta opiskelleiden luokka. Sukupuoli oli negatiivisessa yhteydessä GFAP-proteiinin kanssa eli miehillä proteiinipitoisuus on matalampaa naisiin verrattuna. Ikä ja yli 25 vuotta opiskelleilla oli positiivinen yhteys GFAP-proteiinin kanssa.

Tässä aineistossa yhteys kognitiotestien ja proteiinien välillä oli heikko. Vaikka selitysasteet olivat NfL- ja GFAP-pitoisuuksia selitettäessä kohtalaiset, eivät kognitiotestien tulokset olleet yhteydessä proteiinien kanssa pääkomponenttiregressiossa. Kaiken kaikkiaan menetelmät toimivat toisiinsa nähden yhtä hyvin.

Taulukko 6: G_0-sukupolvessa pääkomponenttiregressiomallien selittäjät, joilla oli yhteys proteiinipitoisuuksien kanssa, niiden regressiokertoimet ja 95 %:n luottamusvälit.

NfL	Muuttuja	Estimaatti	Luottamusväli
	Vakio	2.67	[2.22, 3.11]
	67–75-vuotiaat	0.21	[0.10, 0.32]
	75–83-vuotiaat	0.49	[0.38, 0.61]
	yli 83-vuotiaat	0.71	[0.53, 0.91]
Aβ40			
	Vakio	100.58	[74.32, 126.84]
	Sukupuoli	5.32	[1.12, 9.52]
	75–83-vuotiaat	15.89	[7.75, 24.02]
	yli 83-vuotiaat	26.54	[14.50, 38.58]
Aβ42			
	Vakio	4.33	[2.51, 6.15]
	67–75-vuotiaat	0.53	[0.01, 1.05]
	75–83-vuotiaat	0.89	[0.33, 1.44]
	yli 83-vuotiaat	1.00	[0.18, 1.81]
GFAP			
	Vakio	4.63	[4.16, 5.10]
	Sukupuoli	-0.13	[-0.20, -0.06]
	67–75-vuotiaat	0.17	[0.05, 0.30]
	75–83-vuotiaat	0.46	[0.33, 0.59]
	yli 83-vuotiaat	0.70	[0.51, 0.89]
	yli 25 vuotta opiskelleet	0.37	[0.03, 0.72]

6.2.2 G_1-sukupolven tulokset

Taulukko 7 esittää ristiinvalidoinnilla ennustevirheen minimoivien komponenttien lukumäärät PKR- ja OPNR-menetelmissä ja selittäjien lukumäärät LASSO-regressiossa G_1-sukupolven aineistossa. NfL-proteiinipitoisuutta selitettäessä malliin jäi PKR-menetelmällä neljä pääkomponenttia ja OPNR-menetelmällä kahdeksan komponenttia. LASSO-regressiossa malliin jäi 23 selittäjää, joista yksi oli vakio. Vakion lisäksi malliin jäi sukupuoli, vireystila, ikä, opiskeluvuodet, PAL-testin virheiden keskiarvo (PALMETS28) ja yrityksien lukumäärä (PALTA28, PALTA4 ja PALTEA6), RTI-testin virheiden lukumäärä (RTIFESI ja RTIFESNR), mediaanireaktioaika (RTIFMDRT) ja keskiliikkumisaika (RTIFMMT), RVP-testin latenssiaika (RVPMDL ja RVPML) ja väärien hälytysten lukumäärä (RVPPFA) sekä SWM-testin virheitä kuvaavia muuttujia (SWMBE8, SWMDE12, SWMDE6, SWMDE8, SWMWE4 ja SWMWE6) ja strategiamuuttuja (SWMSX).

Aβ40-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kolme pääkomponenttia ja OPNR-menetelmällä kaksitoista komponenttia (taulukko 7). LASSO-regressiolla malliin jäi kolmetoista selittäjää, joista yksi oli vakio. Vakion lisäksi malliin jäi nukutut tunnit, ikä, PAL-testin virheiden keskiarvo (PALMETS28), yrityksien lukumäärä (PALTA12 ja PALTA8) ja virheitä kuvaavia muuttujia (PAL-

Taulukko 7: Selittävien muuttujien lukumäärät eri menetelmillä G_1-sukupolven muodostamassa aineistossa. Lukumäärä perustuu kussakin tapauksessa ristiinvalidointiin. Harjaregressiossa selittäjien lukumäärä on 63.

Menetelmä	**NfL**	**Aβ40**	**Aβ42**	**GFAP**
Pääkomponenttien lukumäärä **(PKR)**	4	3	1	8
Komponenttien lukumäärä **(OPNR)**	8	12	5	3
Selittäjien lukumäärä **(LASSO)**	23	13	1	14

TE12 ja PALTE2), RTI-testin virhettä kuvaavia muuttujia (RTIFESI ja RTIFTES) sekä SWM-testin virhettä kuvaavia muuttujia (SWMBE6, SWMDE468 ja SWMDE8).

Aβ42-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi yksi pääkomponentti ja OPNR-menetelmällä viisi komponenttia (taulukko 7). LASSO-regressiossa malliin jäi vain vakio.

GFAP-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kahdeksan pääkomponenttia ja OPNR-menetelmällä kolme komponenttia (taulukko 7). LASSO-regressiossa malliin jäi 14 selittäjää, joista yksi oli vakio. Vakion lisäksi malliin jäi sukupuoli, ikä, opiskeluvuodet, PAL-testin virheiden keskiarvo (PALMETS28), kuvioiden lukumäärä (PALNPR28), yrityksien lukumäärä (PALTA8) ja virheitä kuvaavia muuttujia (PALTE2, PALTEA2 ja PALTEA4), RTI-testin virhettä kuvaava muuttuja (RTIFESNR), RVP-testin todennäköisyys osumasta (RVPPH) ja osumien lukumäärä (RVPTH) sekä SWM-testin virhettä kuvaava muuttuja (SWMBE12). NfL-, Aβ40- ja Aβ42-proteiinipitoisuuksia selitettäessä PKR-menetelmällä pienempi määrä komponentteja minimoi ennustevirheen verrattuna OPNR-menetelmään.

Taulukko 8: Ennustevirhe (JVR) ja R^2-selitysaste G_1-sukupolven aineistossa eri menetelmiä käyttäen.

	NfL		**Aβ40**		**Aβ42**		**GFAP**	
Menetelmät	**JVR**	**R^2**	**JVR**	***R^2***	**JVR**	***R^2***	**JVR**	***R^2***
PKR	0.33	0.11	16.83	0.02	1.20	0.03	0.35	0.09
OPNR	0.34	0.12	16.49	0.04	1.28	-0.02	0.34	0.10
LASSO	0.33	0.09	17.16	0.002	1.25	> 0.00	0.34	0.14
Harja	0.34	0.07	17.25	-0.01	1.25	> 0.00	0.35	0.10

Taulukko 8 esittää eri menetelmien ennustevirheet (jäännösvirhehajonta, JVR) ja R^2-selitysasteet G_1-sukupolven aineistossa. NfL-proteiinipitoisuutta selitettäessä ennustevirhe oli kohtuullinen ja hieman pienempi kuin G_0-sukupolven analyysissa. Kuitenkin selitysaste R^2 oli matalampi kuin G_0-sukupolven analyysissa. Vain noin 7–11 % aineiston vaihtelusta pystyttiin selittämään kyseisten menetelmien avulla. Saman suuntaisia tuloksia saatiin, kun vasteena oli GFAP-proteiini. Noin 9–14 % aineiston vaihtelusta pystyttiin selittämään. Aβ40-proteiinia selitettäessä ennustevirheet olivat suuria ja selitysasteet matalia. Aβ42-proteiinia selitettäessä ennustevirheet olivat kohtuullisen suuria ja selitysasteet olivat huonoja. Sekä Aβ40- ja Aβ42-proteiinipitoisuuttaa selitettäessä esiintyi negatiivisia selitysasteita, mikä viittaa sii-

hen, että mallit ennustavat tuloksia huonommin kuin keskiarvo. Tulokset osoittavat, että taustamuuttujilla ja kognitiotestien tuloksilla oli heikko yhteys proteiinipitoisuuksien kanssa. OPNR-menetelmä toimii hieman muita menetelmiä paremmin NfL- ja Aβ40-pitoisuuksia selitettäessä. PKR-menetelmä toimii hieman paremmin Aβ42-pitoisuutta selitettäessä. Ja LASSO-regressio toimii GFAP-pitoisuutta selitettäessä hieman paremmin kuin muut menetelmät. Kuitenkin eri menetelmien tuloksien välillä erot olivat pieniä, eikä yksikään menetelmä nouse ylitse muiden tässä aineistossa. Tästä syystä jatkotarkasteluun valikoitui pääkomponenttiregressio (PKR) aivan kuten G_0-sukupolven analyysissa.

Taulukko 9: G_1-sukupolvessa pääkomponenttiregressiomalleihin selittäjät, joilla oli yhteys proteiinien kanssa, niiden estimaattien arvot ja 95 %:n luottamusvälit.

NfL	**Muuttuja**	**Estimaatti**	**Luottamusväli**
	Vakio	2.04	[1.72, 2.36]
	1. pääkomponentti	0.01	[0.003, 0.02]
	49–53-vuotiaat	0.16	[0.10, 0.24]
	yli 53-vuotiaat	0.25	[0.18, 0.32]
Aβ40			
	Vakio	99.02	[83.16, 114.87]
	45–49-vuotiaat	3.58	[0.14, 7.02]
	49–53-vuotiaat	5.18	[1.96, 8.40]
	yli 53-vuotiaat	6.29	[3.16, 9.41]
Aβ42			
	Vakio	6.47	[5.30, 7.64]
GFAP			
	Vakio	3.87	[3.49, 4.25]
	Sukupuoli	-0.13	[-0.17, -0.08]
	5. pääkomponentti	-0.01	[-0.03 -0.002]
	45–49-vuotiaat	0.10	[0.03, 0.17]
	49–53-vuotiaat	0.19	[0.13, 0.26]
	yli 53-vuotiaat	0.35	[0.28, 0.42]

Taulukko 9 esittää pääkomponenttiregressiomallin ne muuttujat, joidet 95 %:n luottamusvälit eivät sisältäneet nollaa. NfL-proteiinipitoisuuden kanssa positiivisessa yhteydessä olivat ensimmäinen pääkomponentti, ikäluokat 49–53-vuotiaat ja yli 53-vuotiaat. Ensimmäisen pääkomponentin arvon kasvaessa myös NfL-proteiinipitoisuus kasvaa. Aβ40-proteiinipitoisuuden kanssa positiivisessa yhteydessä olivat kaikki ikäluokat. Tällöin ikääntyessä proteiinipitoisuus kasvaa. Aβ42-proteiinipitoisuuttaa selitettäessä malliin jäi ainoastaan vakio. GFAP-proteiinin kanssa negatiivisessa yhteydessä olivat sukupuoli ja viides pääkomponentti. Miehillä on matalampi proteiinipitoisuus naisiin verrattuna ja viidennen pääkomponentin arvon kasvaessa GFAP-proteiinipitoisuus laskee. Näiden lisäksi positiivisessa yhteydessä proteiinin kanssa olivat kaikki ikäluokat. Näiden tuloksien pohjalta yhteys kognitiotestien ja proteiinipitoisuuksien välillä oli olemassa vain NfL- ja GFAP-proteiinipitoisuuksia selitettäessä.

Taulukko 10 esittää ensimmäisen pääkomponentin muuttujat ja niiden lataukset.

Taulukko 10: Ensimmäisen pääkomponentin lataukset G_1-sukupolven aineistossa. Ensimmäinen pääkomponentti oli positiivisessa yhteydessä NfL-pitoisuuden kanssa.

PAL-testi	**Muuttuja**	**Lataus**	**RTI-testi**	**Muuttuja**	**Lataus**
	PALFAMS28	-0.19		RTIFESI	-0.02
	PALMETS28	0.05		RTIFESNR	0.02
	PALNPR28	-0.17		RTIFESPR	0.03
	PALTA12	-0.15		RTIFMDMT	0.06
	PALTA2	0.04		RTIFMDRT	0.04
	PALTA28	0.14		RTIFMMT	0.06
	PALTA4	0.14		RTIFMRT	0.05
	PALTA6	0.12		RTIFMTSD	0.05
	PALTA8	-0.01		RTIFRTSD	0.06
	PALTE12	-0.09		RTIFTES	0.001
	PALTE2	0.04	**SWM-testi**		
	PALTE28	0.17		SWMBE12	0.16
	PALTE4	0.14		SWMBE4	0.13
	PALTE6	0.14		SWMBE468	0.22
	PALTE8	0.07		SWMBE6	0.19
	PALTEA12	0.18		SWMBE8	0.20
	PALTEA2	0.04		SWMDE12	0.05
	PALTEA28	0.20		SWMDE4	0.03
	PALTEA4	0.12		SWMDE468	0.11
	PALTEA6	0.16		SWMDE6	0.07
	PALTEA8	0.18		SWMDE8	0.09
RVP-testi				SWMS	0.19
	RVPA	-0.15		SWMS6	0.18
	RVPLSD	0.10		SWMSX	0.19
	RVPMDL	0.08		SWMTE12	0.16
	RVPML	0.09		SWMTE4	0.13
	RVPPFA	0.06		SWMTE468	0.22
	RVPPH	-0.14		SWMTE6	0.18
	RVPTFA	0.06		SWMTE8	0.20
	RVPTH	-0.14		SWMWE12	0.03
	RVPTM	0.14		SWMWE4	0.03
				SWMWE468	0.10
				SWMWE6	0.07
				SWMWE8	0.08

Ensimmäinen pääkomponentti oli yhteydessä NfL-pitoisuuden kanssa. PAL-testin muuttujissa oli sekä positiivisesti että negatiivisesti latauksia. PAL-testin suurin lataus oli virhettä kuvaavassa muuttujassa (PALTEA28, lataus 0.20).

RTI-testin muuttujissa lataukset olivat sekä negatiivisia että positiivisia ja lataukset olivat pienempiä verrattuna muiden testien latauksiin (taulukko 10). RTI-testillä oli siis pienin vaikutus ensimmäiseen pääkomponenttiin. RTI-testin suurin lataus oli keskiliikkumisajassa (RTIFMMT, lataus 0.06) ja reaktioajan keskihajonnassa (RTIFRTSD, lataus 0.06).

RVP-testin lataukset olivat sekä negatiivisia että positiivisia ja suurin lataus oli sensitiivisyysmuuttujassa (RVPA, lataus -0.15) (taulukko 10). SWM-testin lataukset olivat kaikki positiivisia, mutta latauksien suuruudet vaihtelivat 0.03 ja 0.22 välillä. Suurin lataus oli virhettä kuvaavassa muuttujassa (SWMBE468, lataus 0.22).

PAL-testin ensimmäisen yrityksen pistemäärä -muuttujalla (PALFAMS28), RVP-testin sensitiivisyys lukujonolle -muuttujalla (RVPA), oikean osuman todennäköisyys -(RVPPH) ja oikeiden osumien lukumäärä -muuttujilla (RVPTH) oli suuret negatiiviset lataukset (taulukko 10). Muuten suurilla latauksilla oli positiivinen yhteys ensimmäiseen pääkomponenttiin. Näitä muuttujia olivat PAL-testissä virhettä kuvaavat muuttujat (PALTEA12, PALTEA28 ja PALTEA8) ja SWM-testissä virhettä kuvaavat muuttujat (SWMTE468, SWMTE6 ja SWMTE8).

Taulukko 11 esittää viidennen pääkomponentin muuttujat ja niiden lataukset. Viides pääkomponentti oli yhteydessä GFAP-pitoisuuden kanssa. Viidennen pääkomponentin virhemuuttujissa oli sekä positiivisia että negatiivisia latauksia. PAL-testin suurin lataus oli virheiden keskiarvo (PALMETS28, lataus 0.37). Suurin lataus oli samassa testissä kuin ensimmäisessä pääkomponentissa PAL-testien tuloksissa. RVP-testin suurin lataus oli sensitiivisyysmuuttujassa (RVPA, lataus 0.12). Suurin lataus oli tässäkin testissä samassa muuttujassa kuin ensimmäisessä pääkomponentissa. RTI-testin lataukset olivat huomattavasti pienempiä verrattuna muihin muuttujiin, aivan kuten ensimmäisessäkin pääkomponentissa. Suurin lataus oli mediaanireaktioaikamuuttujassa (RTIFMDRT, lataus -0.07). Näin ollen RTI-testillä oli heikoin yhteys viidenteen pääkomponenttiin. SWM-testin muuttujien suurin lataus oli virhettä kuvaavassa muuttujassa (SWMDE8, lataus -0.20).

Ne muuttujat, joissa oli suurimmat lataukset viidennessä pääkomponentissa, olivat suurimmaksi osaksi positiivisesti latautuneita. Näitä muuttujia oli PAL-testin virheiden keskiarvo (PALMETS28), onnistuneiden kuvioiden lukumäärä (PALNPR-28), yrityksien lukumäärää kuvaavat muuttujat (PALTA12, PALTA28, PALTA4 ja PALTA8), virheiden lukumäärää kuvaavat muuttujat (PALTE12, PALTE4 ja PALTEA4) ja virheiden lukumäärä SWM-testissä (SWMTE4). Suuresti negatiivisesti latautuneita muuttujia oli SWM-testissä virhettä kuvaavissa muuttujissa (SWMDE468, SWMDE8 ja SWMWE468).

6.2.3 Yhdistettyyn aineistoon ($G_0 + G_1$) perustuvat tulokset

Taulukko 12 esittää ristiinvalidoinnilla ennustevirheen minimoivien komponenttien lukumäärät PKR- ja OPNR-menetelmissä ja selittäjien lukumäärät LASSO-regressiossa sukupolvet yhdistävässä aineistossa. NfL-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kymmenen pääkomponenttia ja OPNR-menetelmällä seitsemän komponenttia. LASSO-regressiossa malliin jäi kolmetoista selit-

Taulukko 11: Viidennen pääkomponentin lataukset G_1-sukupolven aineistossa. Viides pääkomponentti oli negatiivisessa yhteydessä GFAP-pitoisuuden kanssa.

PAL-testi	**Muuttuja**	**Lataus**	**RTI-testi**	**Muuttuja**	**Lataus**
	PALFAMS28	-0.05		RTIFESI	-0.004
	PALMETS28	0.37		RTIFESNR	-0.05
	PALNPR28	0.22		RTIFESPR	0.04
	PALTA12	0.22		RTIFMDMT	-0.06
	PALTA2	0.01		RTIFMDRT	-0.07
	PALTA28	0.26		RTIFMMT	-0.06
	PALTA4	0.19		RTIFMRT	-0.07
	PALTA6	0.03		RTIFMTSD	-0.02
	PALTA8	0.20		RTIFRTSD	-0.04
	PALTE12	0.23		RTIFTES	0.004
	PALTE2	0.01	**SWM-testi**		
	PALTE28	0.12		SWMBE12	0.09
	PALTE4	0.18		SWMBE4	0.17
	PALTE6	-0.04		SWMBE468	0.05
	PALTE8	0.11		SWMBE6	0.11
	PALTEA12	-0.12		SWMBE8	-0.01
	PALTEA2	0.01		SWMDE12	-0.002
	PALTEA28	-0.08		SWMDE4	0.11
	PALTEA4	0.17		SWMDE468	-0.19
	PALTEA6	-0.05		SWMDE6	-0.05
	PALTEA8	-0.14		SWMDE8	-0.20
RVP-testi				SWMS	0.09
	RVPA	0.12		SWMS6	0.11
	RVPLSD	-0.07		SWMSX	0.09
	RVPMDL	-0.08		SWMTE12	0.08
	RVPML	-0.08		SWMTE4	0.18
	RVPPFA	0.04		SWMTE468	0.03
	RVPPH	0.13		SWMTE6	0.10
	RVPTFA	0.04		SWMTE8	-0.03
	RVPTH	0.13		SWMWE12	-0.01
	RVPTM	-0.10		SWMWE4	0.10
				SWMWE468	-0.19
				SWMWE6	-0.04
				SWMWE8	-0.21

Taulukko 12: Selittävien muuttujien lukumäärät eri menetelmillä koko aineistossa. Lukumäärä perustuu kussakin tapauksessa ristiinvalidointiin. Harjaregressiossa selittäjien lukumäärä on 63.

Menetelmä	**NfL**	**Aβ40**	**Aβ42**	**GFAP**
Pääkomponenttien lukumäärä **(PKR)**	10	2	3	6
Komponenttien lukumäärä **(OPNR)**	7	5	11	9
Selittäjien lukumäärä **(LASSO)**	13	20	10	15

täjää, jotka olivat vakion lisäksi ikä, PAL-testin ensimmäisen yrityksen pistemäärän -muuttuja (PALFAMS28), yrityksien lukumäärää kuvaavia -muuttujia (PALTA2 ja PALTA8) ja virhettä kuvaava -muuttuja (PALTE4), RTI-testin virhettä kuvaava -muuttuja (RTIFESPR) keskiliikkumisaika (RTIFMMT) ja keskireaktioaika (RTIFMRT), RVP-testin sensitiivisyysmuuttuja (RVPA) ja mediaanilatenssiaika (RVPMDL) sekä SWM-testin strategiamuuttuja (SWMSX) ja keskivirhe (SWMTE6).

Aβ40-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kaksi pääkomponenttia ja OPNR-menetelmällä viisi komponenttia (taulukko 12). LASSO-regressiossa malliin jäi 20 muuttujaa, joista yksi oli vakio. Vakion lisäksi malliin jäi vireystila, ikä, opiskeluvuodet, PAL-testin virheiden keskiarvo (PALMETS28), yrityksien lukumäärä (PALTA2), virhettä kuvaavat muuttujat (PALTE6, PALTEA12 ja PALTEA6), RTI-testin virhettä kuvaava muuttuja (RTIFESI), mediaaniliikkumisaika (RTIFMDMT), keskireaktioaika (RTIFMRT) ja reaktioajan keskihajonta (RTIFRTSD), RVP-testin mediaanilatenssiaika (RVPMDL), osumien todennäköisyys (RVPPH), väärän hälytyksen todennäköisyys (RVPTFA) ja ohimenevien lukumäärä (RVPTM) sekä SWM-testin strategiamuuttuja (SWMS) ja virheitä kuvaavia muuttujia (SWMTE12 ja SWMTE6).

Aβ42-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kolme pääkomponenttia ja yksitoista komponenttia OPNR-menetelmällä (taulukko 12). LASSO-regressiossa malliin jäi kymmenen muuttujaa, joista yksi oli vakio. Vakion lisäksi malliin jäi ikä, PAL-testin virheiden keskiarvo (PALMETS28) ja yrityksien lukumäärä (PALTA28), RTI-testin keskireaktioaika (RTIFMRT), RVP-testin keskireaktioaika (RVPML) ja väärän hälytyksen todennäköisyys (RVPPFA) sekä SWM-testin virhettä kuvaavat muuttujat (SWMDE12 ja SWMWE6).

GFAP-proteiinipitoisuutta selitettäessä PKR-menetelmällä malliin jäi kuusi pääkomponenttia ja yhdeksän komponenttia OPNR-menetelmällä (taulukko 12). LASSO-regressiossa malliin jäi 15 muuttujaa, joista yksi oli vakio. Vakion lisäksi malliin jäi ikä, PAL-testin oikeiden kuvioiden lukumäärä (PALNPR28), yrityksien lukumäärä (PALTA2) ja virhettä kuvaavat muuttujat (PALTE12 ja PALTEA8), RTI-testin virhettä kuvaava muuttuja (RTIFESPR), RVP-testin sensitiivisyysmuuttuja (RVPA), latenssiajan keskihajonta (RVPLSD) ja latenssiajan mediaani (RVPMDL) sekä SWM-testin virhettä kuvaavat muuttujat (SWMBE12, SWMBE8, SWMDE12 ja SWMTE12). Aβ40-, Aβ42 ja GFAP-pitoisuuksia selitettäessä pienempi määrä komponentteja minimoi ennuste virheen PKR-menetelmässä verrattuna OPNR-menetelmään.

Taulukko 13 esittää eri menetelmien tuloksia koko aineiston analyysissa, jossa

Taulukko 13: Ennustevirhe (JVR) ja R^2-selitysaste koko aineistossa eri menetelmiä käyttäen.

	NfL		**Aβ40**		**Aβ42**		**GFAP**	
Menetelmät	**JVR**	**R^2**	**JVR**	***R*^{2}**	**JVR**	***R*^{2}**	**JVR**	***R*^{2}**
PKR	0.37	0.59	18.91	0.32	1.51	0.04	0.40	0.55
OPNR	0.38	0.57	19.21	0.32	1.40	0.04	0.38	0.57
LASSO	0.38	0.59	21.11	0.18	1.44	0.04	0.38	0.58
Harja	0.38	0.58	21.41	0.15	1.46	0.02	0.38	0.57

on yhdistetty G_0- ja G_1-sukupolvet. NfL- ja GFAP-proteiinipitoisuuksia selitettäessä ennustevirheet olivat kohtuullisia ja selitysaste oli hyvä, noin 60 % vasteen vaihtelusta pystyttiin selittämään kyseisten mallien avulla. Ennustevirheet pysyivät yhtä suurina kuin erikseen G_0- ja G_1-sukupolvien analyyseissa, mutta selitysaste kasvoi huomattavasti. Myös Aβ40-proteiinipitoisuutta selitettäessä selitysaste kasvoi, kun aineistot yhdistettiin ja noin 15–32 % vasteen vaihtelusta pystyttiin selittämään kyseisten mallien avulla. Kuitenkin ennustevirhe oli edelleen suuri. Aβ42-proteiinille selitysaste parani koko aineiston analyysissa, mutta jäi kuitenkin huonoksi. GFAP-proteiinipitoisuutta selitettäessä ennustevirhe pysyi yhtä suurena, kun aineistot yhdistettiin. Selitysaste kasvoi huomattavasti aivan kuten NfL-proteiinipitoisuutta selitettäessä. PKR-menetelmä toimi hieman muita menetelmiä paremmin NfL- ja Aβ40-proteiinipitoisuuksia selitettäessä. Aβ42-proteiinipitoisuutta selitettäessä OPNR-menetelmä toimi hieman muita paremmin. GFAP-proteiinipitoisuutta selitettäessä LASSO-regressio toimi muita menetelmiä paremmin. Kaiken kaikkiaan menetelmät toimivat tässä aineistossa yhtä hyvin.

Taulukko 14 kertoo koko aineiston (G_0- ja G_1-sukupolvi) pääkomponenttiregressiomallin tilastollisesti merkitsevät selittäjät, jotka olivat yhteydessä proteiinipitoisuuksien kanssa. Tällöin näiden muuttujien 95 %:n luottamusvälit eivät sisältäneet nollaa. NfL-proteiinipitoisuuden kanssa negatiivisessa yhteydessä olivat ensimmäinen ja neljäs pääkomponentti. Tällöin pääkomponenttien arvojen kasvaessa NfL-proteiinipitoisuus laskee. Pääkomponenttien lisäksi ikä oli positiivisessa yhteydessä NfL-proteiinipitoisuuden kanssa. Tällöin ikääntyessä proteiinipitoisuus kasvaa. Aβ40-proteiinipitoisuuden kanssa positiivisessa yhteydessä olivat kaikki ikäluokat. Aβ42-proteiinipitoisuuden kanssa positiivisessa yhteydessä olivat ikäluokat 63–77- ja yli 77-vuotiaat. Näiden lisäksi negatiivisessa yhteydessä oli 15–20 vuotta opiskelleiden luokka. Ikääntyessä proteiinipitoisuus kasvaa, mutta kun opiskeluvuosia on 15–20 vuotta proteiinipitoisuus laskee. GFAP-proteiinipitoisuuden kanssa yhteydessä olivat ensimmäinen ja viides pääkomponentti, sukupuoli, kaikki ikäluokat ja 15–20 vuotta opiskelleiden luokka. Ensimmäinen pääkomponentti ja sukupuoli olivat negatiivisessa yhteydessä proteiinin kanssa, muut olivat positiivisessa yhteydessä. Tällöin miehillä proteiinipitoisuus on matalampi verrattuna naisiin ja ensimmäisen pääkomponentin arvon kasvaessa GFAP-proteiinipitoisuus laskee. Ikääntyessä proteiinipitoisuus kasvaa. 15–20 vuotta opiskelleilla on korkeampi GFAP-pitoisuus. Liitteessä B on ensimmäisen pääkomponentin lataukset, liitteessä C on neljännen pääkomponentin lataukset ja liitteessä D on viidennen pääkomponentin lataukset koko aineistossa.

Taulukko 14: Pääkomponenttiregressiomalleihin jäävät selittäjät, niiden estimaattien arvot ja luottamusvälit koko aineistossa.

NfL	Muuttuja	Estimaatti	Luottamusväli
	Vakio	2.28	[2.06, 2.50]
	1. pääkomponentti	-0.02	[-0.03, -0.01]
	4. pääkomponentti	-0.02	[-0.03, -0.01]
	53–65-vuotiaat	0.20	[0.15, 0.26]
	65–77-vuotiaat	0.74	[0.67, 0.81]
	yli 77-vuotiaat	1.03	[0.93, 1.12]
Aβ40			
	Vakio	95.62	[84.16, 107.07]
	53–65-vuotiaat	6.11	[3.28, 8.93]
	65–77-vuotiaat	22.54	[18.93, 26.14]
	yli 77-vuotiaat	33.26	[28.39, 38.13]
Aβ42			
	Vakio	5.84	[5.01, 6.67]
	65–77-vuotiaat	0.58	[0.33, 0.84]
	yli 77-vuotiaat	1.07	[0.74, 1.41]
	15–20 vuotta opiskelleet	-0.33	[-0.60, -0.05]
GFAP			
	Vakio	4.04	[3.84, 4.25]
	1. pääkomponentti	-0.01	[-0.02, -0.004]
	5. pääkomponentti	0.01	[0.001, 0.02]
	Sukupuoli	-0.13	[-0.17, -0.09]
	53–65-vuotiaat	0.25	[0.19, 0.30]
	65–77-vuotiaat	0.80	[0.74, 0.87]
	yli 77-vuotiaat	1.12	[1.04, 1.21]
	15–20 vuotta opiskelleet	0.08	[0.01, 0.16]

7 Pohdinta

Tutkielman tavoitteena oli tutkia kognitiotestien ja Alzheimerin tautiin liittyvien veren proteiinien yhteyttä terveillä henkilöillä. Yhteyttä tutkittiin selittämällä proteiinipitoisuuksia kognitiotestien tuloksilla ja taustamuuttujilla. Menetelmällisenä tavoitteena oli vertailla erilaisia malleja, joilla hallitaan usean selittäjän ongelmia. Proteiineja olivat beta-amyloidit 40 ja 42 (Aβ40 ja Aβ42), hapan säikeinen gliaproteiini (GFAP) ja neurofilamentin kevytketju (NFL). Analyysimenetelmiä olivat osittaisen pienimmän neliösumman regressio (OPNR), pääkomponentti- (PKR), LASSO- sekä harjaregressio. Valitut menetelmät sopivat aineistoon, jossa on paljon selittäjiä. Aineistossa oli 63 kognitiotestin tulosta, jolloin oli tarpeenmukaista pyrkiä pienentämään selittävien muuttujien dimensiota. Lisäksi analyysissa käytettiin taustamuuttujia selittämään proteiinipitoisuuksia.

Taustamuuttujat ikä ja opiskeluvuodet olivat analyysissa kategorisina muuttujina. Alun perin analyyseissa kokeiltiin ikää ja opiskeluvuosia jatkuvina muuttujina, mutta tuloksien varmistamiseksi ikä ja opiskeluvuodet kategorisoitiin. Tulokset ei-

vät kuitenkaan eronneet, kun ikä oli jatkuvana tai kategorisena muuttujana. Tulokset kuitenkin muuttuivat, kun opiskeluvuosia käsiteltiin kategorisena muuttujana. GFAP-proteiinia selitettäessä, kun opiskeluvuodet oli jatkuvana selittäjänä, opiskeluvuosilla oli negatiivinen yhteys GFAP-proteiinin kanssa. Kun opiskeluvuodet kategorisoitiin, yhteys löytyi yli 25 vuotta opiskelleilla ja GFAP-proteiinin välillä G_0-sukupolven aineistossa ja 15–20 vuotta opiskelleilla ja proteiinin välillä koko aineistossa. Tällöin opiskeluvuodet jatkuvana muuttujana olisi johtanut harhaanjohtaviin tuloksiin. Tästä syystä lopullisissa tuloksissa ikä ja opiskeluvuodet olivat kategorisina muuttujina.

NfL- ja GFAP-proteiinia selitettäessä jäännösvirhehajonta oli maltillinen, jolloin ennusteet olivat lähellä todellisia arvoja. Nämä proteiinit olivat logaritmoituina, jolloin vasteet olivat pienempiä kuin alkuperäiset NfL- ja GFAP-proteiinipitoisuudet. Logaritmin käytölle ei ollut perustetta Aβ40- ja Aβ42-proteiineja selitettäessä. Aβ40- ja Aβ42-proteiinien havainnot olivat suurempia verrattuna logaritmoituihin NfL- ja GFAP-pitoisuuksiin, mikä selittää suurempia jäännösvirhehajontalukuja Aβ40- ja Aβ42-proteiineissa. Tämä ei kuitenkaan suoraan tarkoita, että virheet olisivat olleet huonompia Aβ40- ja Aβ42-proteiineissa.

7.1 G_0-sukupolvi

G_0-sukupolven aineistossa selitysasteet olivat kohtalaiset jokaisella menetelmällä NfL-proteiinipitoisuutta selitettäessä. Menetelmät toimivat yhtä hyvin ennustevirheiden ja selitysasteiden näkökulmasta. Tästä syystä tarkempaan tarkasteluun valikoitui pääkomponenttiregressio (PKR). Yhteyttä pääkomponenttiregressiossa ei havaittu kognitiotestien ja NfL-proteiinin välillä. Vain ikä oli selvästi yhteydessä proteiinin kanssa. Ikääntyessä NfL-pitoisuus kasvaa. Myös aikaisemmissa tutkimuksissa on havaittu, että ikääntyessä NfL-proteiinipitoisuus kasvaa [22].

Aβ40-proteiinipitoisuuden kanssa yhteydessä olivat sukupuoli ja suuri ikä (yli 75-vuotiaat). Miehillä on korkeampi Aβ40-proteiinipitoisuus naisiin verrattuna. Suurella ikäluokalla on korkeampi Aβ40-pitoisuus muihin ikäluokkiin verrattuna. Aβ42-proteiinipitoisuuden kanssa yhteydessä olivat ikäluokat. Ikääntyessä Aβ42-pitoisuus kasvaa. Kuitenkin selitysaste jäi hyvin matalaksi ja yhteys oli heikko. LASSO-regressiossa Aβ42-proteiinipitoisuutta selitettäessä malliin ei jäänyt ainuttakaan kognitiotestin tulosta. Aikaisemmissa tutkimuksessa on havaittu, että Aβ42-pitoisuus kasvaa terveillä henkilöillä ikääntyessä [23]. Tässä aineistossa ei kuitenkaan havaittu vahvaa yhteyttä.

GFAP-proteiinipitoisuuden kanssa yhteydessä olivat sukupuoli, ikä ja pitkä opiskeluaika (yli 25 vuotta). Ikääntyessä GFAP-proteiinipitoisuus kasvaa, mikä on linjassa aikaisempien tutkimuksien kanssa [24]. Naisilla GFAP-proteiinipitoisuus on keskimäärin korkeampi verrattuna miehiin. Tämä tulos on myös linjassa aikaisempien tutkimuksien kanssa [25]. Niillä, joilla on pitkä opiskeluaika (yli 25 vuotta) on korkeampi GFAP-pitoisuus muihin nähden. Yli 25 vuotta opiskelleita oli keskimäärin vähiten ja he olivat 41–81-vuotiaita. Ikä ei siis vaikuta taustalla siihen, että yli 25 vuotta opiskelleilla oli positiivinen tilastollinen yhteys GFAP-proteiinin kanssa. Kuitenkin vähäinen määrä henkilöitä saattoi aiheuttaa vääristymän tähän.

Ikä oli ainoa tilastollisesti selvästi Nfl-, Aβ40-, ja GFAP-proteiineja selittävä te-

kijä. Matalasta selitysasteesta voitiin päätellä, että selittävät muuttujat eivät olleet vahvasti yhteydessä proteiinipitoisuuksien kanssa. NfL- ja GFAP-proteiineja selitettäessä selitysasteet olivat parempia kuin muissa proteiineissa. Aβ42- ja Aβ42-proteiinipitoisuuksia selitettäessä muutamalla menetelmällä oli negatiivinen selitysaste, mikä viittaa siihen, että menetelmien ennustus on huonompi kuin keskiarvo. Tämä vahvistaa havaintoa siitä, että yhteys proteiinien ja kognition välillä oli heikko.

LASSO-malleissa yhteisiä selittäjiä olivat RTI-testin mediaanireaktioaika (RTIFMDRT) ja RVP-testin ohimenevien lukujonojen lukumäärä (RVPTM) NfL- ja Aβ40-proteiinipitoisuuksia selitettäessä. SWM-testin eli työmuistitestin tuloksista mikään ei jäänyt LASSO-regressiossa malleihin NfL-, Aβ40- ja Aβ42-proteiinipitoisuuksia selitettäessä. Tällöin työmuistitestin tuloksilla oli heikoin yhteys proteiineihin tässä aineistossa. Kuitenkin matalien selitysasteiden takia yhteys kaiken kaikkiaan proteiinien ja kognitiotestien välillä oli heikko. Tutkielmassa oli tarkoituksena vertailla käytettyjä menetelmiä ja löytää sopiva menetelmä tähän aineistoon, menetelmien välillä ei kuitenkaan löytynyt suuria eroavaisuuksia.

7.2 G_1-sukupolvi

G_1-sukupolven aineistossa PKR-menetelmällä NfL-proteiinipitoisuuden kanssa yhteydessä olivat ensimmäinen pääkomponentti ja vanhemmat ikäluokat (yli 49-vuotiaat). Vanhemmassa ikäluokassa pitoisuus on korkeampaa. Myös aikaisemmissa tutkimuksissa on havaittu, että ikääntyessä NfL-proteiinipitoisuus kasvaa [22]. Ristiinvalidoinnilla neljä kognitiotesteistä muodostettua pääkomponenttia minimoivat ennustevirheen. Kuitenkin vain yksi (ensimmäinen) pääkomponentti oli yhteydessä NfL-proteiinipitoisuuden kanssa. Tällöin muut komponentit eivät tuoneet lisäarvoa mallin selitysasteelle.

Ensimmäisessä pääkomponentissa suurin osa latauksista oli positiivisia ja ensimmäinen pääkomponentti oli positiivisessa yhteydessä NfL-proteiinin kanssa. Tällöin pääkomponentin arvojen suurentuessa NfL-proteiinipitoisuus kasvaa. Suurimmat positiiviset lataukset olivat PAL-testin yrityksien lukumäärässä (PALTEA12) ja virheiden lukumäärässä (PALTEA28 ja PALTEA8) sekä SWM-testin virheiden lukumäärää kuvaavissa muuttujissa (SWMTE468, SWMTE6 ja SWMTE8).

PAL-testin ensimmäisen yrityksen pistemäärä -muuttujalla (PALFAMS28), RVP-testin sensitiivisyys lukujonolle -muuttujalla (RVPA), oikean osuman todennäköisyys -muuttujalla (RVPPH) ja oikeiden osumien lukumäärä -muuttujalla (RVPTH) oli suuret negatiiviset lataukset ensimmäisessä pääkomponentissa. Näiden muuttujien arvojen kasvaessa NfL-proteiinipitoisuus laskee.

RTI-testin muuttujien lataukset olivat heikoimpia muiden testien latauksiin verrattuna, tällöin tällä testillä oli pienin vaikutus ensimmäiseen pääkomponenttiin. Näin ollen yhteys RTI-testin ja NfL-proteiinin välillä oli heikoin. LASSO-regressiossa malliin jäi RTI-testin mediaanireaktioaika ja keksiliikkumisaika (RTIFMDRT ja RTIFMMT) nämä olivat samat, jotka jäivät malliin myös G_0-sukupolven aineistossa NfL-proteiinipitoisuutta selitettäessä. Tämä on ristiriitaista sen kanssa, että pääkomponenttiregressiossa havaittiin, että yhteys RTI-testin ja NfL-proteiinin välillä oli heikoin verrattuna muihin testeihin.

Aβ40-proteiinin kanssa yhteydessä oli ikä. Ikääntyessä proteiinipitoisuus kasvaa,

mikä on linjassa aikaisempien tutkimuksien kanssa. Aβ42-proteiinin kanssa yhteydessä eivät olleet mitkään muuttujat tässä aineistossa.

GFAP-proteiinin kanssa yhteydessä olivat sukupuoli, viides pääkomponentti ja ikäluokat. Ristiinvalidoinnilla kahdeksan pääkomponenttia minimoivat ennustevirheen. Kuitenkin vain yksi (viides) oli negatiivisessa yhteydessä proteiinipitoisuuden kanssa. Tällöin viidennen pääkomponentin arvon kasvaessa GFAP-proteiinipitoisuus laskee. Ne muuttujat, joissa oli suurimmat lataukset viidennessä pääkomponentissa, olivat suurimmaksi osaksi positiivisesti latautuneita PAL-testien tuloksia: virheiden keskiarvo (PALMETS28), kuvioiden lukumäärä (PALNPR28), yrityksien lukumäärä (PALTA12, PALTA28, PALTA4 ja PALTA8) ja virheiden lukumäärä (PALTE12, PALTE4 ja PALTEA4). Lisäksi positiivisesti suuresti latautunut oli virheiden lukumäärä SWM-testissä (SWMTE4). Näin ollen näiden muuttujien arvojen kasvaessa GFAP-pitoisuus laskee.

Viidennessä pääkomponentissa suuresti negatiivisesti latautuneita muuttujia oli SWM-testin virhettä kuvaavat muuttujat (SWMDE468, SWMDE8 ja SWMWE468). Näin ollen näiden arvojen suurentuessa GFAP-pitoisuus kasvaa. Myös tässä pääkomponentissa RTI-testien lataukset olivat pienimmät ja näin ollen tämä testi vaikuttaa vähiten viidenteen pääkomponenttiin. Tästä syystä RTI-testin yhteys GFAP-proteiinin kanssa oli heikoin. GFAP-proteiinia selitettäessä LASSO-regressiolla PAL-testin virhettä kuvaava muuttuja (PALMETS28) jäi malliin samoin kuin G_0-sukupolven analyysissa GFAP-proteiinia selitettäessä.

G_1-sukupolven selitysasteet olivat huonommat verrattuna G_0-sukupolven analyysiin. G_0-sukupolven aineisto koostui G_1-sukupolven vanhemmista, jolloin G_0-sukupolven henkilöt olivat vanhempia verrattuna G_1-sukupolveen. Vanhemmassa sukupolvessa saattaa näkyä paremmin muutoksia proteiinipitoisuuksissa, joihin ikä on yhteydessä. Tämä saattoi selittää huonontuneen selitysasteen G_1-sukupolven analyysissa verrattuna G_0-sukupolven analyysiin. Tulokset osoittavat, että vain pieni osa varianssista oli selitetty, jolloin yhteys proteiinien ja selittäjien välillä oli heikko. Kognitiotestit olivat yhteydessä vain NfL- ja GFAP-proteiineihin. LASSO-regressiossa PAL-testin virhettä kuvaava muuttuja (PALMETS28) jäi malliin Aβ40-, NfL- ja GFAP-proteiinipitoisuuksia selitettäessä.

7.3 Koko aineisto

NfL- ja GFAP-proteiinia selitettäessä selitysaste kasvoi selkeästi, kun G_1- ja G_0-sukupolven aineistot yhdistettiin. Myös Aβ40-proteiinipitoisuutta selitettäessä selitysaste kasvoi aikaisempaan nähden. Näissä proteiineissa todettiin jo kuvan 4 perusteella, että ikääntyessä proteiinipitoisuudet kasvoivat. Kun aineistot yhdistettiin, ikäjakauma leveni. Tämä saattoi olla taustasyynä sille, että selitysasteet paranivat, kun aineistot yhdistettiin. Aβ42-proteiinia selitettäessä selitysasteet olivat edelleen huonoja.

NfL-pitoisuutta selitettäessä malliin oli ristiinvalidoinnilla optimoitu viisi pääkomponenttia ja näistä kaksi (ensimmäinen ja neljäs) olivat yhteydessä proteiinipitoisuuden kanssa. Tämän lisäksi ikä oli yhteydessä proteiinipitoisuuden kanssa, mikä on linjassa aikaisempien tutkimuksien kanssa [22]. Pääkomponentit olivat negatiivisessa yhteydessä proteiinin kanssa. Suuremmat pääkomponenttien arvot pienensivät

proteiinipitoisuutta. Koko aineiston ensimmäisessä pääkomponentissa suurimmat positiiviset lataukset olivat ensimmäisen yrityksien lukumäärässä (PALFAMS28), kuvioiden lukumäärässä (PALNPR28) ja sensitiivisyysmittarissa (RVPA). Näiden muuttujien arvojen kasvaessa NfL-pitoisuus laskee. Suurimmat negatiiviset lataukset olivat virheiden lukumäärässä PAL-testissä (PALTEA28, PALTEA6, PALTEA8) ja virheitä ilmaisevissa muuttujissa SWM-testissä (SWMBE46, SWMBE6, SWMBE8, SWMTE468, SWMTE6 ja SWMTE8). Tällöin näiden muuttujien arvojen kasvaessa NfL-pitoisuus kasvaa.

Neljännen pääkomponentin lataukset olivat sekä positiivisia että negatiivisia. Suurimmat lataukset olivat RTI-testin tuloksilla, mikä on ristiriidassa G_1-sukupolven tuloksien kanssa. RTI-testillä oli heikoin yhteys NfL- ja GFAP-proteiinin kanssa G_1-sukupolven aineistossa. Suurimmat positiiviset lataukset neljännessä pääkomponentissa olivat PAL-testin muuttujissa, jotka ilmaisivat kokonaisvirhettä (PALTEA12, PALTEA28 ja PALTEA8). Tällöin näiden muuttujien arvojen kasvaessa NfL-pitoisuus kasvaa.

Suurimmat negatiiviset lataukset neljännessä pääkomponentissa olivat RTI-testin virhemuuttujassa (RTIFESPR), mediaaniluvuissa (RTIFMDMT ja RTIFMDRT), keksiarvoissa (RTIFMMT ja RTIFMRT), keskihajonnoissa (RTIFMTSD ja RTIFRTSD) ja kokonaisvirheessä (RTIFTES). Näiden lisäksi PAL-testin kuvioiden lukumäärässä (PALNPR28), kokeilujen lukumäärässä (PALTA12) ja kokonaisvirheiden lukumäärässä (PALTE12) oli suuret negatiiviset lataukset. Tällöin näiden muuttujien arvojen kasvaessa NfL-pitoisuus laskee. Neljännessä pääkomponentissa SWM-testillä oli matalimmat lataukset. LASSO-regressiossa RTI-testin keskiliikkumisaika (RTIFMMT) jäi malliin, kuten G_0- ja G_1-sukupolvien analyyseissa erikseen.

Aβ40-proteiinipitoisuutta selitettäessä ikä oli yhteydessä proteiinipitoisuuden kanssa ja ikääntyessä proteiinipitoisuus kasvoi. LASSO-regressiossa malliin jäi PAL-testin virhettä ilmaiseva muuttuja (PALMETS28), joka jäi myös G_1-sukupolven analyysissa malliin. Aβ42-proteiinipitoisuutta selitettäessä ikäluokat 65–77- ja yli 77-vuotiaat olivat yhteydessä proteiinipitoisuuden kanssa. Aikaisemmissa tutkimuksissa on havaittu, että Aβ42-proteiinilla ja iällä olisi selkeämpi positiivinen yhteys verrattuna Aβ40-proteiiniin [4]. Tämän tutkielman tulokset olivat ristiriitaiset aikaisempiin tutkimuksiin ja tässä aineistossa Aβ40-proteiinipitoisuuden kasvulla oli selkeämpi positiivinen yhteys iän kanssa verrattuna Aβ42-proteiiniin selitysasteiden perusteella.

GFAP-proteiinia selitettäessä kognitiotesteistä muodostettu ensimmäinen ja viides pääkomponentti olivat yhteydessä proteiinipitoisuuden kanssa. Ristiinvalidoinnilla kuusi pääkomponenttia minimoivat ennustevirheen, mutta vain kaksi (ensimmäinen ja viides) olivat yhteydessä proteiinipitoisuuden kanssa. Ensimmäinen pääkomponentti oli negatiivisessa yhteydessä proteiinin kanssa. Tällöin samoin kuin NfL-proteiinia selitettäessä PALFAMS28-, PALNPR28- ja RVPA-muuttujien arvojen kasvaessa GFAP-pitoisuus laskee. Lisäksi PALTEA28-, PALTEA6-, PALTEA8-, SWMBE46-, SWMBE6- ja SWMBE8-, SWMTE468-, SWMTE6- ja SWMTE8-muuttujien arvojen kasvaessa GFAP-pitoisuus kasvaa.

Viides komponentti oli positiivisessa yhteydessä proteiinin kanssa. Viidennen pääkomponentin lataukset olivat sekä negatiivisia että positiivisia. RTI-testin lataukset olivat suurimmat, kuten neljännessä pääkomponentissa. Suurimmat posi-

tiiviset lataukset olivat RTI-testin mediaanissa (RTIFMDMT), keskiarvossa (RTIFMMT ja RTIFMRT) ja keskihajonnassa (RTIFMTSD). Näiden lisäksi suuria positiivisia latauksia oli SWM-testin virhettä kuvaavissa muuttujissa (SWMDE468 ja SWMDE8). Näin ollen näiden muuttujien arvojen kasvaessa myös proteiinipitoisuus suurenee. Negatiivisia suuria latauksia oli SWM-testin strategiamuuttujissa (SWMS, SWMS6 ja SWMSX) ja kokonaisvirhettä kuvaavassa muuttujassa (SWMTE4). Näiden muuttujien arvojen kasvaessa proteiinipitoisuus pienenee.

Pääkomponenttien lisäksi sukupuoli oli negatiivisessa yhteydessä GFAP-proteiinin kanssa. Tällöin miehillä proteiinipitoisuus on matalampi naisiin verrattuna. Tämä tulos oli sama, kun tutkittiin G_1- ja G_0-sukupolvea erikseen. Lisäksi tulos on linjassa aikaisempien tutkimuksien kanssa [25]. Tämän lisäksi ikä oli yhteydessä proteiinipitoisuuden kanssa. Ikääntyessä pitoisuus kasvaa, mikä on linjassa aikaisempien tuloksien kanssa [24]. LASSO-regressiossa malliin jäi SWM-testin virhettä kuvaava muuttuja (SWMBE12), kuten G_1-sukupolven analyysissa.

NfL- ja GFAP-proteiinia selitettäessä kognitiotesteillä oli yhteys proteiinipitoisuuksiin pääkomponenttiregressiossa. Aβ40- ja Aβ42-proteiinipitoisuuksia selitettäessä kognitiotesteillä ja proteiinipitoisuuksilla ei löytynyt yhteyttä. Menetelmistä mikään ei noussut ylitse muiden tässäkään aineistoissa. LASSO-regressiossa PALtestin yrityksien lukumäärää kuvaava muuttuja (PALTA2) jäi malliin NfL-, Aβ40- ja GFAP-proteiineja selitettäessä.

7.4 Lopuksi

Analyysin tarkoituksena oli tutkia yhteyksiä proteiinipitoisuuksien ja kognitiotestien välillä. Lisäksi menetelmällisenä tavoitteena oli tutkia, mikä menetelmä sopisi yhteyksien tutkimiseen parhaiten tässä aineistossa. Menetelmiä vertailtiin ennusteiden näkökulmasta, vaikka pääasiallisena tavoitteena ei ollut luoda ennusteita. Ennustamisen tavoitteena on luoda malli, joka pystyy mahdollisimman tarkasti ennustamaan tulevia havaintoja. Ennustekyvyn optimoiminen on ensisijaista, ja yhteyksien löytäminen jää toissijaiseksi. Ennustusten luominen on tehokas tapa vertailla erilaisia menetelmiä, koska se tarjoaa mitattavia tuloksia, joiden perusteella menetelmiä voidaan objektiivisesti arvioida. Menetelmien välillä ei kuitenkaan löytynyt eroavaisuuksia ennustuskyvyn näkökulmasta. Jatkotarkasteluun valikoitui pääkomponenttiregressio, jonka avulla tarkasteltiin yhteyksiä proteiinipitoisuuksien ja kognitiotestien välillä. Tässä lähestymistavassa tavoitteena on ymmärtää vasteen ja selittäjien välisiä yhteyksiä.

Selkeää yhteyttä proteiinipitoisuuksien ja kognitiotestien välillä ei löytynyt. Kognitiotestien ja proteiinien välillä löytyi yhteys vain NfL- ja GFAP-proteiinipitoisuuksia selitettäessä G_1-sukupolven aineistossa ja koko aineistossa. Pääkomponenttiregressiota tutkittiin tarkemmin ja yhteys löytyi pääkomponenttien ja proteiinien välillä. Positiivisesti ja negatiivisesti latautuneet muuttujat hankaloittavat kuitenkin yhteyksien tulkintaa. Selkein yhteys löytyi iän ja NfL-, Aβ40- ja GFAP-proteiinien välillä.

Tutkielman tarkoituksena oli myös selvittää, miten eri menetelmät soveltuvat proteiinipitoisuuksien ja kognitiotestien yhteyksien löytämiseen. Pääkomponenttiregressio on tehokas menetelmä vähentämään aineiston korrelaatiota, mutta tämä

menetelmä ei ota huomioon vastetta. Tällöin menetelmä ei välttämättä ole paras, kun pyritään selittämään tai ennustamaan vastetta. OPNR-menetelmä ottaa huomioon vasteen sekä selittäjät komponentteja muodostaessaan, jolloin menetelmän käyttäminen on hyvä valinta, kun tavoitteena on löytää lineaarinen yhteys selittäjien ja vasteen välillä. Näissä menetelmissä kuitenkin tuloksien sanallinen avaaminen on hankalaa. LASSO-regressio on sopiva, kun suuresta määrästä selittäjiä halutaan löytää ennustekyvyn kannalta tärkeimmät. Se voi kuitenkin jättää huomiotta eilineaarisia suhteita. Harjaregressio säilyttää mallissa kaikki alkuperäiset selittäjät pienentämällä näiden regressiokertoimia, jolloin korrelaatio vähenee. Tässä aineistossa menetelmien hyödyt ja haitat eivät nousseet esiin ja menetelmät toimivat aineistoissa samankaltaisesti.

Lisäksi tutkittiin, kuinka PAL-testeistä muodostetut pääkomponentit olivat yhteydessä proteiinipitoisuuksiin. Yhteys oli selkeämpi kuin tässä tutkielmassa muodostetuissa pääkomponenteissa, jotka oli muodostettu kaikkien kognitiotestien tuloksista. Jatkotutkimuksena voitaisiin tutkia erikseen jokaisesta testiluokasta muodostettujen pääkomponenttien yhteyttä proteiinipitoisuuksiin.

Aikaisemmissa tutkimuksissa on havaittu yhteys Aβ42- ja Aβ40-proteiinipitoisuuksien suhteella heikentyneen kognition kanssa [13]. Tätä tutkittiin lisänä myös tässä tutkielmassa, mutta yhteys oli heikko.

G_0- ja G_1-sukupolvien yksilöillä oli sukulaissuhde keskenään. Tätä ei otettu huomioon tässä tutkielmassa. Jatkotutkimuksena voitaisiin perehtyä, miten tämä sukulaisuussuhde vaikuttaa yhteyksiin.

Viitteet

[1] Muistisairauksien yhteiskunnalliset vaikutukset. (2021). Rud Pedersen Public Affairs Oy:n tuottama raportti Muistiliitto ry:n ja Biogen Oy:n toimeksiannosta. Biogen-128250.

[2] Jack, C., Knopman, D., Jagust, W., Petersen, R., Weiner, M., Aisen, P., Shaw, L., Vemuri, P., Wiste, H., Weigand, S., Lesnick, T., Pankratz, V., Donohue, M. & Trojanowski, J. (2013). Update on hypothetical model of Alzheimer's disease biomarkers. Lancet Neurol Neurology. Vol. 12(2), s. 207–216.

[3] Eloniemi-Sulkava, U., Erkinjuntti, T., Huhtamäki-Kuoppala, M., Jolkkonen, J., Kontturi, J., Lupsakko, T., Malmivaara, A., Olkkonen-Nikula, A., Palomäki, H., Sarlio-Lähteenkorva, A., Soininen, H., Strandberg, T., Suhonen, J. & Virnes, E. (2012). Kansallinen muistiohjelma 2012–2020 Tavoitteena muistiystävällinen Suomi. Sosiaali- ja terveysministeriö. Suomi.

[4] Fukumoto, H., Asami-Odaka, A., Suzuki, N., Shimada, N., Shimada, Y., Ihara Y. & Iwatsubo, T. (1996). Amyloid beta protein deposition in normal aging has the same characteristics as that in Alzheimer's disease. Predominance of A beta 42(43) and association of A beta 40 with cored plaques. American Journal of Pathology. Vol. 148, s. 259—265.

[5] Mayer, C., Brunkhorst, R., Niessner, M., Pfeilschifter, W., Steinmetz, H. & Foerch, C. (2013). Blood levels of glial fibrillary acidic protein (GFAP) in patients with neurological diseases. PloS one, 8(4), e62101.

[6] Byrne, L., Rodrigues, F., Blennow, K., Durr, A., Leavitt, B., Roos, R., Scahill, R., Tabrixi, S., Zetterberg, H., Langbehn, D. & Wild, E. (2017). Neurofilament light protein in blood as a potential biomarker of neurodegeneration in Huntington's disease: a retrospective cohort analysis. Lancet Neurol Neurology. Vol. 16(8), s. 601-609.

[7] Barthélemy, N., Salvadó, G., Schindler, S., He, Y., Janelidze, S., Collij, L., Saef, B., Henson, R., Chen, C., Gordon, B., Li, Y., Joie, R., Benzinger, T., Morris, J., Mattsson-Carlgren, N., Palmqvist, S., Ossenkoppele, R., Rabinovici, G., Stomrud, E., Bateman, R. & Hansson, O. (2024). Highly accurate blood test for Alzheimer's disease is similar or superior to clinical cerebrospinal fluid tests. Nature Medicine. Vol. 30, s. 1085–1095.

[8] Verberk, I., Thijssen, E., Koelewijn, J., Mauroo, K., Vanbrabant, J., Wilde, A., Zwan,, M., Verfaillie, S., Ossenkoppele, R., Barkhof, F., Berckel, B., Scheltens, P., Flier, W., Stoops, E., Venderstichele, H. & Teunissen, C. (2020). Combination of plasma amyloid beta(1-42/1-40) and glial fibrillary acidic protein strongly associates with cerebral amyloid pathology. Alzheimer's research & therapy. Vol.12 (1). s. 1–118.

[9] Bridel C., van Wieringen W. N., Zetterberg H., Tijms B. M. & Teunissen C. E. (2019). Diagnostic value of cerebrospinal fluid neurofilament light protein in neurology. JAMA Neurology. Vol. 76(9), s. 1035–1048.

[10] Mattsson N., Andreasson U., Zetterberg H. & Blennow K.(2017). Association of plasma neurofilament light with neurodegeneration in patients with Alzheimer disease. JAMA Neurology. Vol. 74(5), s. 557–566.

[11] Parvizi, T., König, T., Wurm, R., Silvaieh, S., Altmann, P., Klotz, S., Rommer, P., Furtner, J., Regelsberger, G., Lehrner, J., Traub-Weidinger, T., Gelpi, E. & Stögmanncorresponding, E. (2022) Real-world applicability of glial fibrillary acidic protein and neurofilament light chain in Alzheimer's disease. Frontiers in Aging Neuroscience. Vol. 14. s. 887498.

[12] Gu, L. & Guo, Z. (2013). Alzheimer's AB42 and AB40 peptides form interlaced amyloid fibrils. Journal Neurochemistry. Vol. 126(3), s. 305–311.

[13] Graff-Radford, N., Crook, J., Lucas, J., Boeve, B., Knopman, D., Ivnik, R., Smith, G., Younkin, L., Petersen, R. & Younkin, S. (2007). Association of low plasma Abeta42/Abeta40 ratios with increased imminent risk for mild cognitive impairment and Alzheimer disease. Arch Neurology. Vol. 64(3), s. 354-362.

[14] Tsai, H.-H., Tsai, L.-K., Lo, Y.-L. & Lin, C.-H. (2021). Amyloid related cerebral microbleed and plasma Aβ40 are associated with cognitive decline in Parkinson's disease. Scientific Reports. Vol. 11, s. 1-9.

[15] Rovio, P., Pahkala, K., Nevalainen, J., Juonala, M., Salo, P., Kähönen, M., Hutri-Kähönen, N., Lehtimäki, T., Jokinen, E., Laitinen, T., Taittonen, L., Tossavainen, P., Viikari, J., Rinne, J. & Raitakari, O. Cognitive Performance in Young Adulthood and Midlife: Relations With Age, Sex, and Education—The Cardiovascular Risk in Young Finns Study. (2016). Neuropsychology. Vol. 30(5). s. 532-542.

[16] James, G., Witten, D., Hastie, T. & Tibshirani, R. (2023). Introduction to Statistical Learning with applications in R. Springer.

[17] Taddy, M. (2019) Business Data Science. McGraw-Hill Education. New York.

[18] Vinzi, V., Chin, W., Henseler, J. & Wang, H. (2010). Handbook of Partial Least Squares. Springer. Berlin.

[19] Hastie, T., Tibshirani, R. & Friedman, J. (2009). The Elements of Statistical Learning. Springer.

[20] Jolliffe, I. (2002). Principal Component Analysis, Second Edition. Springer. New York.

[21] Lin, J-L. & Tsay, R. (2005). Comparison of Forecasting Methods with many predictors.

[22] Capo, X., Galmes-Panades, A., Navas-Enamorado, C., Ortega-Moral, A., Marin, S., Cascante, M., Sanchez-Polo, A., Masmiquel, L., Torrens-Mas, M. & Conzalez-Freire, M. (2023) Circulating Neurofilament Light Chain Levels Increase with Age and Are Associated with Worse Physical Function and Body

Composition in Men but Not in Women. International Journal of Molecular Sciences. Vol. 24(16). s. 12751.

[23] Zecca, C., Pasculli, G., Tortelli, R., Dell'Abate, M., Capozzo, R., Barulli, M., Barone, R., Accogli, M., Arima, S., Pollice, A., Brescia, V. & Logroscino, G. (2021). The Role of Age on Beta-Amyloid1–42 Plasma Levels in Healthy Subjects. Front Aging Neurosci. Vol. 13. s. 698571.

[24] Teter, B. (2009) Rodent Aging. Encyclopedia of Neuroscience. s. 397-406.

[25] Sass, D., Guedes, V., Smith, E., Vorn, R., Devoto, C., Edwards, K., Mithani, S., Hentig, J., Lai, C., Wagner, C., Dunbar, K., Hyde, D., Saligan, L., Roy, M. & Gill, J. (2021). Sex Differences in Behavioral Symptoms and the Levels of Circulating GFAP, Tau, and NfL in Patients With Traumatic Brain Injury. Front Pharmacol. Vol. 12. s. 76491.

A Liite: Kognitiotestit

CANTAB	Name	Short label
MOT	MOTML	The mean latency time from stimulus to response
MOT	MOTSDL	The standard deviation of latency time from stimulus to response
MOT	MOTTC	The total number of correct response.
MOT	MOTTE	The total number of failed response.
PAL	PALFAMS28	First Attempt Memory Score
PAL	PALMETS28	Mean Errors to Success
PAL	PALNPR28	Number of Patterns Reached
PAL	PALTA12	Total Attempts (12 Patterns)
PAL	PALTA2	Total Attempts (2 Patterns)
PAL	PALTA28	Total Attempts (2-8 Patterns)
PAL	PALTA4	Total Attempts (4 Patterns)
PAL	PALTA6	Total Attempts (6 Patterns)
PAL	PALTA8	Total Attempts (8 Patterns)
PAL	PALTE12	Total Errors (12 Patterns)
PAL	PALTE2	Total Errors (2 Patterns)
PAL	PALTE28	Total Errors (2-8 Patterns)
PAL	PALTE4	Total Errors (4 Patterns)
PAL	PALTE6	Total Errors (6 Patterns)
PAL	PALTE8	Total Errors (8 Patterns)
PAL	PALTEA12	Total Errors (12 Shapes, adjusted)
PAL	PALTEA2	Total Errors (2 Shapes, adjusted)
PAL	PALTEA28	Total Errors (2-8 Shapes, adjusted)
PAL	PALTEA4	Total Errors (4 Shapes, adjusted)
PAL	PALTEA6	Total Errors (6 Shapes, adjusted)
PAL	PALTEA8	Total Errors (8 Shapes, adjusted)
RTI	RTIFESI	Error Score (inaccurate)
RTI	RTIFESNR	Error Score (no response)
RTI	RTIFESPR	Error Score (premature)
RTI	RTIFMDMT	Median Movement Time
RTI	RTIFMDRT	Median Reaction Time
RTI	RTIFMMT	Mean Movement Time
RTI	RTIFMRT	Mean Reaction Time
RTI	RTIFMTSD	Standard Deviation of Movement Time
RTI	RTIFRTSD	Standard Deviation of Reaction Time
RTI	RTIFTES	Total Error Score
SWM	SWMBE12	Between errors (12 boxes)
SWM	SWMBE4	Between errors (4 boxes)
SWM	SWMBE468	Between errors (4-8 boxes)
SWM	SWMBE6	Between errors (6 boxes)
SWM	SWMBE8	Between errors (8 boxes)
SWM	SWMDE12	Double errors (12 boxes)
SWM	SWMDE4	Double errors (4 boxes)
SWM	SWMDE468	Double errors (4-8 boxes)
SWM	SWMDE6	Double errors (6 boxes)
SWM	SWMDE8	Double errors (8 boxes)
SWM	SWMPR	Problem Reached
SWM	SWMS	Strategy (6-8 boxes)
SWM	SWMS6	Strategy (6 box only)
SWM	SWMSX	Strategy (6-12 boxes)

SWM	SWMTE12	Total errors (12 boxes)
SWM	SWMTE4	Total errors (4 boxes)
SWM	SWMTE468	Total errors (4-8 boxes)
SWM	SWMTE6	Total errors (6 boxes)
SWM	SWMTE8	Total errors (8 boxes)
SWM	SWMWE12	Within errors (12 boxes)
SWM	SWMWE4	Within errors (4 boxes)
SWM	SWMWE468	Within errors (4-8 boxes)
SWM	SWMWE6	Within errors (6 boxes)
SWM	SWMWE8	Within errors (8 boxes)
RVP	RVPA	A' (sensitivity to target sequence)
RVP	RVPLSD	The standard deviation of response latency
RVP	RVPMDL	Median Response Latency
RVP	RVPML	Mean Response Latency
RVP	RVPPFA	Probability of False Alarm
RVP	RVPPH	Probability of Hit
RVP	RVPTFA	Total False Alarms
RVP	RVPTH	Total Hits
RVP	RVPTM	Total Misses

B Liite: Ensimmäinen pääkomponentti koko aineistossa

PAL-testi	Muuttuja	Lataus	RTI-testi	Muuttuja	Lataus
	PALFAMS28	0.19		RTIFESI	-0.03
	PALMETS28	-0.03		RTIFESNR	-0.04
	PALNPR28	0.18		RTIFESPR	-0.08
	PALTA12	0.15		RTIFMDMT	-0.08
	PALTA2	-0.09		RTIFMDRT	-0.10
	PALTA28	-0.06		RTIFMMT	-0.09
	PALTA4	-0.14		RTIFMRT	-0.11
	PALTA6	-0.06		RTIFMTSD	-0.07
	PALTA8	0.09		RTIFRTSD	-0.10
	PALTE12	0.10		RTIFTES	-0.05
	PALTE2	-0.08	**SWM-testi**		
	PALTE28	-0.12		SWMBE12	-0.16
	PALTE4	-0.14		SWMBE4	-0.14
	PALTE6	-0.11		SWMBE468	-0.20
	PALTE8	0.01		SWMBE6	-0.17
	PALTEA12	-0.16		SWMBE8	-0.18
	PALTEA2	-0.08		SWMDE12	-0.08
	PALTEA28	-0.19		SWMDE4	-0.03
	PALTEA4	-0.15		SWMDE468	-0.10
	PALTEA6	-0.17		SWMDE6	-0.05
	PALTEA8	-0.18		SWMDE8	-0.08
RVP-testi				SWMS	-0.16
	RVPA	0.18		SWMS6	-0.15
	RVPLSD	-0.16		SWMSX	-0.16
	RVPMDL	-0.13		SWMTE12	-0.16
	RVPML	-0.14		SWMTE4	-0.14
	RVPPFA	-0.11		SWMTE468	-0.20
	RVPPH	0.14		SWMTE6	-0.17
	RVPTFA	-0.09		SWMTE8	-0.18
	RVPTH	0.14		SWMWE12	-0.06
	RVPTM	-0.14		SWMWE4	-0.03
				SWMWE468	-0.08
				SWMWE6	-0.05
				SWMWE8	-0.07

C Liite: Neljäs pääkomponentti koko aineistossa

PAL-testi	Muuttuja	Lataus	RTI-testi	Muuttuja	Lataus
	PALFAMS28	-0.15		RTIFESI	-0.08
	PALMETS28	-0.06		RTIFESNR	-0.16
	PALNPR28	-0.20		RTIFESPR	-0.18
	PALTA12	-0.21		RTIFMDMT	-0.21
	PALTA2	-0.005		RTIFMDRT	-0.27
	PALTA28	0.04		RTIFMMT	-0.21
	PALTA4	0.07		RTIFMRT	-0.28
	PALTA6	0.09		RTIFMTSD	-0.22
	PALTA8	-0.07		RTIFRTSD	-0.22
	PALTE12	-0.20		RTIFTES	-0.17
	PALTE2	-0.0003	**SWM-testi**		
	PALTE28	0.13		SWMBE12	0.04
	PALTE4	0.07		SWMBE4	-0.07
	PALTE6	0.13		SWMBE468	-0.02
	PALTE8	0.03		SWMBE6	-0.01
	PALTEA12	0.18		SWMBE8	-0.01
	PALTEA2	-0.001		SWMDE12	0.08
	PALTEA28	0.17		SWMDE4	0.02
	PALTEA4	0.06		SWMDE468	0.08
	PALTEA6	0.13		SWMDE6	0.08
	PALTEA8	0.20		SWMDE8	0.06
RVP-testi				SWMS	-0.03
	RVPA	0.15		SWMS6	-0.03
	RVPLSD	-0.13		SWMSX	-0.03
	RVPMDL	-0.17		SWMTE12	0.05
	RVPML	-0.18		SWMTE4	-0.07
	RVPPFA	0.003		SWMTE468	-0.01
	RVPPH	0.15		SWMTE6	-0.01
	RVPTFA	0.01		SWMTE8	0.002
	RVPTH	0.15		SWMWE12	0.08
	RVPTM	-0.15		SWMWE4	0.02
				SWMWE468	0.09
				SWMWE6	0.08
				SWMWE8	0.07

D Liite: Viides pääkomponentti koko aineistossa

PAL-testi	Muuttuja	Lataus	RTI-testi	Muuttuja	Lataus
	PALFAMS28	-0.05		RTIFESI	0.08
	PALMETS28	0.03		RTIFESNR	0.15
	PALNPR28	-0.05		RTIFESPR	0.16
	PALTA12	-0.05		RTIFMDMT	0.18
	PALTA2	0.05		RTIFMDRT	0.16
	PALTA28	0.01		RTIFMMT	0.19
	PALTA4	0.05		RTIFMRT	0.17
	PALTA6	-0.005		RTIFMTSD	0.21
	PALTA8	-0.03		RTIFRTSD	0.16
	PALTE12	-0.05		RTIFTES	0.13
	PALTE2	0.04	**SWM-testi**		
	PALTE28	0.04		SWMBE12	-0.09
	PALTE4	0.06		SWMBE4	-0.19
	PALTE6	0.02		SWMBE468	-0.10
	PALTE8	0.003		SWMBE6	-0.16
	PALTEA12	0.04		SWMBE8	-0.03
	PALTEA2	0.05		SWMDE12	0.09
	PALTEA28	0.06		SWMDE4	-0.05
	PALTEA4	0.07		SWMDE468	0.25
	PALTEA6	0.04		SWMDE6	0.06
	PALTEA8	0.05		SWMDE8	0.26
RVP-testi				SWMS	-0.20
	RVPA	0.10		SWMS6	-0.20
	RVPLSD	-0.01		SWMSX	-0.20
	RVPMDL	0.004		SWMTE12	-0.06
	RVPML	-0.01		SWMTE4	-0.19
	RVPPFA	0.15		SWMTE468	-0.08
	RVPPH	0.17		SWMTE6	-0.15
	RVPTFA	0.14		SWMTE8	-0.01
	RVPTH	0.17		SWMWE12	0.10
	RVPTM	-0.17		SWMWE4	-0.06
				SWMWE468	0.24
				SWMWE6	0.07
				SWMWE8	0.25

E Liite: R-koodi

```
# Monen selittajan regressiomenetelmien vertailu
# Paakomponenttiregressio, osittaisen pienimman neliosumman
# regressio, harjaregressio ja LASSO-regressio
# Aineistona kaytetaan LASERI-tutkimuksen aineistoa vuodelta # 2018-2020
# Esimerkkina kaytetaan NfL-proteiinia
# Tekija: Jasmine Hakala
# jasmine.hakala@hotmail.com
# Paivays: 11.4.2024

#################################################################
# Tarvittavat kirjastot
#################################################################

library(caret)
library(tidyverse)
library(pls)
library(glmnet)

#################################################################
# Ristiinvalidointi
#################################################################

# Funktio, joka toteuttaa ristiinvalidointia
# Parametrin arvoina aineisto (data),
# ristiinvalidointienmaara (fold)
# ja tutkittava proteiini (protein)
# Palauttaa keskiarvon JVR-virheesta

cross_validate <- function(data, fold=10, protein){
  # Yksiloiden lukumaara aineistoissa
  n <- nrow(data)
  # Aineiston satunnainen jako ryhmiin
  fold_vector <- replicate(n, sample(1:fold, 1, replace = TRUE))

  # Tulosvektori, johon tallennetaan ennustevirheet
  errors <- c(numeric(0))
  # Kaydaan lapi ryhmat
  for (i in unique(fold_vector)){
    # Opetusjoukko
    train_set <- data[fold_vector != i,]
    # Testausjoukko
    test_set <- data[fold_vector == i,]

    if (protein == "NFL"){
      # Mallin sovitus opetusjoukolla, kun proteiini on NFL
      model <- lm(train_set$NFL ~., data = train_set)
```

```
    }
    else if (protein == "A40"){
      # Mallin sovitus opetusjoukolla, kun proteiini on A40
      model <- lm(train_set$A40 ~., data = train_set)
    }
    else if (protein == "A42"){
      # Mallin sovitus opetusjoukolla, kun proteiini on A42
      model <- lm(train_set$A42 ~., data = train_set)
    }
    else{
      # Mallin sovitus opetusjoukolla, kun proteiini on GFAP
      model <- lm(train_set$GFAP ~., data = train_set)
    }

    # Ennusteet testijoukolle
    pred <- predict(model, test_set[,-1])
    # Lasketaan ennustevirhe JVR ja lisataan errors-tulosvektoriin
    errors <- c(errors, sqrt(mean((pred - test_set[,1])^2)))
  }

  # Muodostetaan keskiarvo
  mean_errors <- mean(errors)


  # Palautetaan keskiarvo JVR-virheesta
  return(mean_errors)
}

#################################################################
# Menetelmien vertailu JVR ja R^2 avulla
#################################################################

# Funktio, joka laskee ennusteet ja vertaa naita todellisiin arvoihin
# Palauttaa JVR ja R^2 selitysasteen

result_function <- function(model, test_set, protein, method){

  if (method == "dim_reduction"){
    # Ennuste jos dimension pienentamismenetelma
    prediction <- predict(model, test_set)
  }
  else{
    # Ennuste jos kutistamismenetelma
    prediction <- predict(model, newx = as.matrix(test_set[, 2:69]),
                          s = "lambda.min")
  }


  # Jos proteiini on NFL
```

```
  if (protein == "NFL"){
    # Ennustevirhe JVR
    error <- sqrt(mean((prediction - test_set$NFL)^2))
    SSE <- sum((prediction - test_set$NFL)^2)
    SST <- sum((test_set$NFL - mean(test_set$NFL))^2)
  }
  # Jos proteiini on A40
  else if (protein == "A40"){
    # Ennustevirhe JVR
    error <- sqrt(mean((prediction - test_set$A40)^2))
    SSE <- sum((prediction - test_set$A40)^2)
    SST <- sum((test_set$A40 - mean(test_set$A40))^2)
  }
  # Jos proteiini on A42
  else if (protein == "A42"){
    # Ennustevirhe JVR
    error <- sqrt(mean((prediction - test_set$A42)^2))
    SSE <- sum((prediction - test_set$A42)^2)
    SST <- sum((test_set$A42 - mean(test_set$A42))^2)
  }
  # Jos proteiini on GFAP
  else{
    # Ennustevirhe JVR
    error <- sqrt(mean((prediction - test_set$GFAP)^2))
    SSE <- sum((prediction - test_set$GFAP)^2)
    SST <- sum((test_set$GFAP - mean(test_set$GFAP))^2)
  }

  # Selitysaste R^2
  R_square <- 1 - SSE / SST

  # Palautetaan ennustevirhe ja selitysaste
  return(c(error, R_square))
}


#################################################################
# Paakomponenttiregressio
#################################################################

# Paakomponenttianalyysi muodostaa kognitiotesteista paakomponentit
# Oletetaan, etta data aineistossa nelja ensimmaista
# saraketta ovat proteiinit
# sarakkeet 4-10 taustamuuttujia ja 10-72 kognitiotestin tuloksia
pca <- prcomp(data[,10:72])

# Tulosmatriisi
results <- data.frame(k = numeric(0), err = numeric(0))
```

```
# Kaydaan lapi kaikki muodostetut paakomponentit
for (i in 1:63){
  if  (i==1){
    components <- pca$x[,1]
  }
  else{
    components <- pca$x[,1:i]
  }
  # Muodostetaan uusi aineisto komponenteista, proteiinista
  # ja taustamuutujista
  data_new <- data.frame(data[, 1:6], components)

  # Kutsutaan ristiinvalidointifunktiota
  # Parametrinarvoina aineisto, ristiinvalidointien lukumaara
  # ja proteiini
  err <- cross_validate(data_new, fold=10, "NFL")

  # Lisataan tulokset tulosmatriisiin
  results <- rbind(results, c(i, err))
}

# Kuinka monta paakomponenttia minimoi ennustevirheen
which.min(results[,2])

# Kuvaaja
plot(results[,1], results[,2], type="b", xlab="Paakomponentit",
     ylab="Ennustevirhe", main="Paakomponenttien␣ennustevirhe")

# Valitaan jatkoanalyysiin ne komponentit, jotka minimoivat
# ennustevirheen
components <- pca$x[,1:which.min(results[,2])]

# Muodostetaan uusi aineisto proteiinista, komponenteista ja
# taustamuuttujista
data_pca <- data.frame(data$NFL, components, data[, 5:9])
# Korjataan sarakkeen nimi
colnames(data_pca)[colnames(data_pca) == "data.NFL"] ="NFL"

# Jako train ja test joukkoihin
index <- createDataPartition(data_pca$NFL, p = 0.7, list = FALSE)
train_pca <- data_pca[index, ]
test_pca <- data_pca[-index, ]

# Mallin sovitus opetusjoukolla
model_pca <- lm(train_pca$NFL ~., data = train_pca)
summary(model_pca)

# Ennuste ja jaannosvirhehajonta, R^2
results_PKR <- result_function(model_pca, test_pca, "NFL",
```

```
                "dim_reduction")

#################################################################
# Osittaisen pienimman neliosumman regressio
#################################################################

# PLS
pls <- plsr(NFL~., data=data, validation = "none")

# Tulosmatriisi
results <- data.frame(k = numeric(0), err = numeric(0))

# Kaydaan lapi komponentit ja niiden muodostamat ennustevirheet
for (i in 1:63){
  if  (i==1){
    components <- pls$scores[,1]
  }
  else{
    components <- pls$scores[,1:i]
  }
  # Muodostetaan aineisto
  data_new <- data.frame(data[, 1:6], components)

  # Kutsutaan ristiinvalidointi-funktiota
  # Parametrien arvoina aineisto, ristiinvalidointien
  # lukumaara ja proteiini
  err <- cross_validate(data_new, fold=10, "NFL")

  # Lisataan tulosmatriisiin
  results <- rbind(results, c(i, err))
}

# Kuinka monta komponenttia minimoi ennustevirheen
which.min(results[,2])

# Kuvaaja
plot(results[,1], results[,2], type="b", xlab="Komponentit",
     ylab="Ennustevirhe", main="Komponenttien␣ennustevirhe")

# Valitaan komponentit, jotka minimoivat ennustevirheen
components_pls <- pls$scores[,1:which.min(results[,2])]

# Muodostetaan uusi aineisto, jossa proteiini, komponentit ja
# taustamuuttujat
data_pls <- data.frame(data$NFL, components_pls, data[, 5:9])
# Korjataan sarakkeen nimi
colnames(data_pls)[colnames(data_pls) == "data.NFL"] ="NFL"

# Jako train ja test joukkoihin
```

```
index <- createDataPartition(data_pls$NFL, p = 0.7, list = FALSE)
train_pls <- data_pls[index, ]
test_pls <- data_pls[-index, ]

# Mallin sovitus opetusjoukolla
model_pls <- lm(train_pls$NFL ~., data = train_pls)
summary(model_pls)

# Ennuste ja jaannosvirhehajonta, R^2
results_plsr <- result_function(model_pls, test_pls, "NFL",
"dim_reduction")


#################################################################
# Harjaregressio
#################################################################

# Jako opetus- ja testausaineistoihin kutistamismenetelmissa
index <- createDataPartition(data$NFL, p = 0.7, list = FALSE)
train <- data[index, ]
test <- data[-index, ]

# Ridge malli
model_ridge <- cv.glmnet(as.matrix(train[, 2:69]), train$NFL,
            alpha = 0,
                        nfolds = 10)
# Kuvaaja
plot(model_ridge, main="Harjaregressio", ylab="Ennustevirhe")

# Ennuste ja jaannosvirhehajonta, R^2
results_ridge <- result_function(model_ridge, test, "NFL", "shrinkage")


#################################################################
# LASSO-regressio
#################################################################


# LASSO-malli
model_lasso <- cv.glmnet(as.matrix(train[, 2:69]), train$NFL,
            alpha = 1,
                        nfolds = 10)
plot(model_lasso, ylab="Ennustevirhe", main="LASSO")

# Ennuste ja jaannosvirhehajonta, R^2
results_lasso <- result_function(model_lasso, test, "NFL", "shrinkage")

#################################################################
# Menetelmien vertailu
```

```
#################################################################

data.frame(method = c( "PKR", "PLSR", "LASSO", "Ridge"),
           JVR = round(c(results_PKR[1], results_plsr[1],
                         results_lasso[1], results_ridge[1]), 2),
           R_square = round(c(results_PKR[2], results_plsr[2],
                              results_lasso[2], results_ridge[2]), 2))
```