Tampereen yliopisto

Heli Oksanen

Kvasikokeellisen tutkimusasetelman luominen propensity score matching -menetelmällä

Esimerkkinä lasten saamisen vaikutus tuloihin

Informaatioteknologian ja viestinnän tiedekunta
Tilastollisen data-analyysin kandidaattitutkielma
Toukokuu 2021

Tiivistelmä

Heli Oksanen: Kvasikokeellisen tutkimusasetelman luominen propensity score matching -menetelmällä
Tilastollisen data-analyysin kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastollisen data-analyysin tutkinto-ohjelma
Toukokuu 2021

Tutkielma käsittelee kvasikokeellisen tutkimusasetelman luomista propensity score matching -menetelmällä. Esimerkkinä luodaan asetelma, jossa kiinnostuksen kohteena on lasten saamisen vaikutus tuloihin. Tavoitteena on luoda tutkimusasetelma, jossa kysymykseen lasten saamisen vaikutuksesta tuloihin voisi vastata.

Jotta voitaisiin päätellä tietyllä tekijällä olevan kausaalinen vaikutus johonkin ilmiöön, on tutkimusasetelman täytettävä tietyt ehdot. Satunnaistettu kokeellinen tutkimusasetelma yleisesti ottaen täyttää nämä ehdot, mutta aina sellaisen toteuttaminen ei ole mahdollista. Kvasikokeellisessa asetelmassa koe- ja kontrolliryhmät eroavat systemaattisesti toisistaan. Tilastollisilla menetelmillä voidaan tuottaa kaltaistettu otos, jolloin tutkimusasetelma on satunnaistetun kaltainen.

Propensity score matching -menetelmällä kaltaistetaan koe- ja kontrolliryhmät taustamuuttujiltaan mahdollisimman samankaltaisiksi. Menetelmä hyödyntää niin kutsuttuja propensiteetti- tai alttiuslukuja, jotka määritelmän mukaisesti ovat ehdollisia todennäköisyyksiä yksilön valikoitumiselle koe- tai kontrolliryhmään. Nämä luvut estimoidaan aineistosta käyttäen esimerkiksi logistista regressiota. Itse kaltaistus voidaan toteuttaa erilaisilla algoritmeilla riippuen otoksen ominaisuuksista.

Tässä tutkielmassa kaltaistukseen käytetty esimerkkiaineisto on peräisin Tilastokeskuksen FOLK-kokonaisrekisteriaineistosta. Aineisto sisältää vuodesta 1988 alkaen kunakin vuonna Suomessa työikäiseen väestöön tulleet henkilöt. Koe- ja kontrolliryhmät muodostettiin siten, että seurannan aikana lapsia saavat henkilöt muodostavat koeryhmän ja ne, jotka eivät saa lapsia, muodostavat kontrolliryhmän. Lisäksi aineisto jaettiin naisten ja miesten aineistoihin, jotta lasten saamisen vaikutusta voitaisiin vertailla myös sukupuolittain. Aineistossa havaittiin eroja koe- ja kontrolliryhmän välillä muutamien taustamuuttujien osalta.

Naisten ja miesten aineistoista otettiin 10 000 yksilön otokset. Kaltaistus toteutettiin R-ohjelmointiympäristössä. Tuloksista käy ilmi, että valittu menetelmä onnistui tuottamaan kaltaistetut otokset. Johtopäätöksenä todetaan, että näillä otoksilla voisi mahdollisesti estimoida lasten saamisen vaikutusta tuloihin tietyin rajoituksin.

Avainsanat: kaltaistaminen, kausaliteetti, datan esikäsittely.

Sisällys

1 Johdanto

Tutkittaessa kausaalivaikutuksia satunnaistettua kokeellista tutkimusasetelmaa pidetään pätevän kausaalipäättelyn edellytyksenä. Voidaan päätellä jollain tekijällä olevan kausaalinen vaikutus ilmiöön, mikäli tietty tulos havaitaan kyseisen tekijän läsnä ollessa ja mikäli tulosta ei havaita sen poissa ollessa silloin, kun muiden vaikuttavien tekijöiden osallisuus on pois suljettu. Kausaalipäättelyyn liittyy keskeisesti puuttuvan tiedon ongelma: yksilöllä voidaan havaita vaste vain joko vaikuttavan tekijän läsnä ollessa tai sen poissa ollessa. Satunnaistetussa tutkimusasetelmassa koe- ja kontrolliryhmien yksilöt ovat vertailukelpoisia vasteiltaan, mutta aina ei ole mahdollista toteuttaa satunnaistettua koeasetelmaa. (Holmes 2014.) Kvasikokeellinen tutkimusasetelma tarkoittaa asetelmaa, jossa koe- ja kontrolliryhmiin valikoituminen ei ole satunnaista. Jotta kausaalivaikutuksia voitaisiin estimoida kvasikokeellisessa asetelmassa, tilastollisin keinoin voidaan pyrkiä luomaan otos, joka on satunnaistetun kaltainen. (Leite 2017.)

Tämä tutkielma käsittelee kvasikokeellisen tutkimusasetelman luomista käyttäen propensity score matching -menetelmää. Esimerkkinä luodaan asetelma, jossa kiinnostuksen kohteena on lasten saamisen vaikutus tuloihin. Kysymyksenasettelu on peräisin Työsuojelurahaston Pirstoutuvatko työurat? -tutkimushankkeesta ja analyyseihin käytettiin Tilastokeskuksen (2020) FOLK -kokonaisrekisteriaineistoa. Tavoitteena on luoda tutkimusasetelma, jossa kysymykseen lasten saamisen vaikutuksesta tuloihin voisi vastata.

Tutkielman ensimmäisessä osassa syvennytään propensiteettiluvuilla kaltaistuksen taustateoriaan, esitellään määritelmät, estimointimenetelmät sekä kaltaistusalgoritmit. Lisäksi mainitaan muutama huomio menetelmän heikkouksista. Toisessa osassa esitellään kaltaistukseen käytetty aineisto ja käydään läpi kaltaistusprosessi esimerkin kautta. Lopuksi arvioidaan kaltaistuksen onnistumista ja tutkimusasetelman pätevyyttä.

2 Kaltaistus propensiteettiluvuilla

Propensiteetti- tai altistusluku (propensity score) on ehdollinen todennäköisyys yksilön valikoitumiselle koe- tai kontrolliryhmiin ehdolla selittävien muuttujien havaitut arvot (Rosenbaum & Rubin 1983). Estimoitujen propensiteettilukujen perusteella voidaan kaltaistaa (matching) koe- ja kontrolliryhmien yksilöt taustamuuttujiltaan mahdollisimman samankaltaisiksi. Kaltaistuksen tavoite kvasikokeellisessa tutkimusasetelmassa on vähentää valikoitumisharhaa. Satunnaistetussa koeasetelmassa populaation yksilöt valikoituvat satunnaisesti koe- tai kontrolliryhmiin, kun taas ei-satunnaistetussa kvasikokeellisessa asetelmassa valikoitumiseen vaikuttaa tietyt taustamuuttujat. Kun ryhmät kaltaistetaan taustaominaisuuksiensa perusteella, ryhmään valikoituminen voidaan ajatella satunnaiseksi. (Leite 2017, luku 1.)

2.1 Kausaalipäättelyn ongelma ja kausaalivaikutusten estimointi

Käsitellään kahden käsittelyryhmän tapausta, jossa merkitään koekäsittely numerolla 1 ja koeryhmään valikoituminen $z_i = 1$. Vastaavasti merkitään kontrollikäsittely numerolla 0 ja kontrolliryhmään valikoituminen $z_i = 0$. Jokaisella populaation yksilöllä i voidaan ajatella olevan potentiaalinen vaste r_{1i} koekäsittelyyn, sekä vaste r_{0i} kontrollikäsittelyyn. Nyt yksilön i käsittelyn kausaalivaikutus on erotus $r_{1i} - r_{0i}$. Kuitenkin tässä asetelmassa vasteista voidaan havaita vain toinen, sillä yksilö ei voi saada sekä koe- että kontrollikäsittelyä. Tämä puuttuva tieto tuottaa ongelman kausaalipäättelyyn. (Rosenbaum & Rubin 1983.)

Otoksesta useimmiten estimoitavat kausaalivaikutukset ovat keskimääräinen käsittelyn vaikutus, johon on kirjallisuudessa viitattu nimellä ATE eli average treatment effect ja joka määritellään

$$\mathsf{E}(\boldsymbol{r}_1) - \mathsf{E}(\boldsymbol{r}_0) \tag{2.1}$$

sekä keskimääräinen käsittelyn vaikutus koekäsittelyn yksilöille ATT eli average treatment effect on the treated:

$$\mathsf{E}(\boldsymbol{r}_1 | z = 1) - \mathsf{E}(\boldsymbol{r}_0 | z = 1). \tag{2.2}$$

ATE on siis potentiaalisten vasteiden odotusarvojen erotus koe-, sekä kontrolliryhmien yksilöille, kun taas ATT on potentiaalisten vasteiden odotusarvojen erotus koeryhmän yksilöille. (Leite 2017, luku 1.)

Käsittelyn vaikutuksen estimointi edellyttää tiettyjen oletusten täyttymistä. Oletetaan, että jokaista yksilöä i ja käsittelyä t vastaa yksiselitteinen arvo r_{ti}. Toisin sanoen yksilön i vaste käsittelyyn t on riippumaton yksilön j saamasta käsittelystä. Tämä oletus tunnetaan kirjallisuudessa nimellä SUTVA eli stable unit-treatment value assumption. Lisäksi oletetaan, että koe- tai kontrolliryhmään valikoituminen on riippumaton potentiaalisten vasteiden jakaumista havaittujen kovariaattien ehdolla,

$$(\boldsymbol{r}_1, \boldsymbol{r}_0) \perp\!\!\!\perp z|\boldsymbol{x}. \tag{2.3}$$

Oletus edellyttää, että jokaisella kovariaattien $\boldsymbol{x}$ arvolla koeryhmään valikoitumisen todennäköisyys on $0 < P(z = 1|\boldsymbol{x}) < 1$. Satunnaistetussa kokeellisessa tutkimusasetelmassa tämä ehto täyttyy, sillä ryhmään valikoituminen on satunnaista, jolloin yksilöiden ajatellaan olevan taustamuuttujiltaan samankaltaisia ryhmien välillä. (Rosenbaum & Rubin 1983.) Kvasikokeellisessa tutkimusasetelmassa tämä ehto ei yleensä täyty, vaan tietyt taustamuuttujat vaikuttavat yksilön valikoitumiseen käsittelyryhmään johtaen korreloitumiseen ryhmään valikoitumisen ja tutkittavan vasteen välillä. Tätä kutsutaan myös valikoitumisharhaksi ja se hankaloittaa kausaalipäättelyä, sillä ryhmät eroavat systemaattisesti toisistaan muiltakin ominaisuuksilta kuin käsittelystatukseltaan. (Holmes 2014, luku 1.)

2.2 Propensiteettiluku

2.2.1 Määritelmä

Rosenbaumin ja Rubinin (1983) määritelmän mukaan propensiteettiluku $e(\boldsymbol{x})$ on ehdollinen todennäköisyys yksilön valikoitumiselle käsittelyryhmään ehdolla selittävien muuttujien havaitut arvot ja koeryhmän tapauksessa $e(\boldsymbol{x}) = P(z = 1|\boldsymbol{x})$, missä

$$P(z_1, ..., z_n|x_1, ..., x_n) = \prod_{i=1}^{n} e(x_i)^{z_i}[1 - e(x_i)^{1-z_i}].$$

Koe- ja kontrolliryhmien kaltaistus taustamuuttujien perusteella voidaan yksinkertaistaa käyttäen propensiteettilukua, sillä se tiivistää taustamuuttujien informaation yhteen arvoon (Leite 2017, luku 1).

Propensiteettiluku on niin kutsuttu *balancing score*, joka Rosenbaumin ja Rubinin (1983) määritelmän mukaisesti on havaittujen selittävien muuttujien funktio $b(\boldsymbol{x})$, jolle muuttujien $\boldsymbol{x}$ ehdollinen todennäköisyysjakauma kyseisen funktion ehdolla on sama koe- ja kontrolliryhmässä:

$$\boldsymbol{x} \perp\!\!\!\perp z|b(\boldsymbol{x}).$$

Koska propensiteettiluvulla $e(\boldsymbol{x})$ on tämä tasapainottava ominaisuus, mikäli yksilöt on kaltaistettu ryhmittäin tai pareittain niiden perusteella, parien tai ryhmien taustamuuttujien $\boldsymbol{x}$ jakaumat ovat samat. Lisäksi mikäli kaavan (2.3) oletus on voimassa kovariaattien $\boldsymbol{x}$ ehdolla, se on myös todistetusti voimassa ehdolla $b(\boldsymbol{x})$. Näistä ominaisuuksista seuraa tulos, jonka mukaan yksilöt, joilla on sama propensiteettiluvun arvo, mutta eri käsittelystatukset, ovat vertailukelpoisia vasteiltaan ja näiden vasteiden odotusarvojen erotus vastaa kaavassa (2.1) esitettyä keskimääräistä käsittelyn vaikutusta. Näin ollen käsittelyn vaikutusta voidaan harhattomasti estimoida, kun aineisto on kaltaistettu propensiteettilukujen perusteella. (Rosenbaum & Rubin 1983.)

2.2.2 Estimointi

Käytännössä kvasikokeellisessa asetelmassa propensiteettiluvut $e(\boldsymbol{x})$ ovat tuntemattomia ja ne on estimoitava aineistosta. Yleisimmin käytettyjä parametrisia estimointimenetelmiä ovat logistinen regressio, diskriminanttianalyysi ja probit-regressio. Logistinen regressio soveltuu tapauksissa, joissa käsittelyryhmiä on kaksi ja useamman käsittelyryhmän asetelmassa voidaan käyttää diskriminanttianalyysia. (Holmes 2014, luku 1.) Parametristen estimointimenetelmien lisäksi voidaan käyttää muun muassa tiedonlouhintaa tai Bayesilaisia menetelmiä (Leite 2017, luku 2).

Logistinen regressio on yleisimmin käytetty menetelmä ja tuottaa samankaltaisia propensiteettilukuja verraten diskriminanttianalyysiin tai probit-regressioon. Propensiteettilukujen estimointiin käytettävä malli on seuraava:

$$logit(z_i = 1|\boldsymbol{x}) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki},$$

missä

$$logit(z_i = 1|\boldsymbol{x}) = log\left(\frac{P(z_i = 1)}{1 - P(z_i = 1)}\right).$$

Propensiteettiluvut saadaan estimoidusta mallista:

$$e_i(\boldsymbol{x}) = \frac{exp(logit(z_i = 1|\boldsymbol{x}))}{1 + exp(logit(z_i = 1|\boldsymbol{x}))}.$$

Estimointimenetelmän sopivuutta aineistoon arvioidaan estimoitujen propensiteettilukujen kykynä tasapainoittaa selittävien muuttujien jakaumat käsittelyryhmissä. (Leite 2017, luvut 1-2.)

2.3 Kaltaistusalgoritmit

Kaltaistuksessa luodaan estimoitujen propensiteettilukujen perusteella vertailukelpoiset koe- ja kontrolliryhmät. Kaltaistukseen on käytettävissä monenlaisia algoritmeja riippuen datan ominaisuuksista. Vain osa kaltaistusalgoritmeista käyttää propensiteettilukuja ja niiden käyttö kaltaistukseen ei kaikissa tapauksissa ole järkevintä. On syytä käyttää jotain muuta tunnuslukua, kuten Mahalanobisin etäisyysmittaa ja genetic matching -algoritmia, jos propensiteettiluvuilla ei saada riittävää tasapainotusta tai estimointi tuottaa lukuja arvoiltaan 0 tai 1. (Leite 2017, luku 5.) Seuraavassa on esitelty propensiteettilukuja hyödyntäviä algoritmeja.

2.3.1 Greedy matching

Jokaiselle koeryhmän yksilölle haetaan propensiteettiluvultaan mahdollisimman samankaltainen kontrolliryhmän yksilö. Algoritmi pyrkii siis yksilöittäin suurimpaan mahdolliseen samankaltaisuuteen, muttei huomioi kaltaistettujen ryhmien propensiteettilukujen kokonaisetäisyyttä. Greedy matching on soveltuva algoritmi siinä tapauksessa, kun koeryhmän koko on pienempi kuin mahdollisen kontrolliryhmän.

Kaltaistusalgoritmiin liittyy muutamia vaihtoehtoja. Ensinnäkin kaltaistus voidaan toteuttaa palauttaen tai palauttamatta. Palauttaen toteutetussa algoritmissa kullakin yksilöllä voidaan muodostaa useampia pareja toisten yksilöiden kanssa. Tällöin myös kaltaistuksen järjestyksellä ei ole merkitystä, toisin kuin palauttamatta toteutetussa algoritmissa. Toisekseen kaltaistus voidaan toteuttaa yksi yhteen (1-1) tai yksi moneen (1-k), eli joko jokaiselle koeryhmän yksilölle haetaan vain yksi vastaava kontrolliryhmän yksilö tai k kappaletta. Yksi yhteen kaltaistuksen heikkous on siitä seuraava otoskoon pieneminen etenkin silloin, kun toinen käsittelyryhmistä on huomattavasti pienempi. Yksi moneen kaltaistus taas hakee k kappaletta pareja parien laadusta huolimatta, eikä siksi useinkaan ole suositeltava vaihtoehto. Lisäksi voidaan säätää propensiteettilukujen etäisyyttä, jolla kaltaistus sallitaan. Lähimmän naapurin menetelmässä koeryhmän yksilön pariksi valikoituu yksilö, joka on lähin propensiteettiluvultaan. Toisaalta voidaan käyttää myös ikkunaa (caliper), jonka sisällä yksilöiden kaltaistus sallitaan. Ikkunan yksikkö on yleensä linearisoidun propensiteettiluvun standardoitu keskihajonta ja useimmiten 0.25 keskihajontaa. Ikkunan käyttö voi tosin johtaa yksilöiden poistamiseen otoksesta, mikäli sopivia pareja ei löydy.

2.3.2 Optimal matching

Algoritmi tuottaa kaltaistetut ryhmät pyrkien minimoimaan ryhmien välillä propensiteettilukujen kokonaisetäisyyden. Optimal matching voidaan myös toteuttaa yksi yhteen tai yksi moneen. Etenkin tapauksissa, joissa koeryhmän ja kontrolliryhmään valikoitavien yksilöiden ryhmien kokoero on suuri, yksi yhteen optimal matching tuottaa parempia tuloksia verrattuna yksi yhteen greedy matching -algoritmiin.

2.3.3 Full matching

Jokaiselle koeryhmän yksilölle haetaan ainakin yksi kontrolliyksilö ja päinvastoin. Algoritmi muodostaa ositteet jokaisesta propensiteettiluvultaan samankaltaisten yksilöiden ryhmistä, joissa on vähintään yksi koeryhmän yksilö ja yksi kontrolliryhmän yksilö ja maksimoi näiden ositteiden lukumäärän. Etenkin tapauksissa, joissa kaltaistukseen käytettäviä kovariaatteja on suuri määrä ja propensiteettilukujen jakaumissa on suuria eroja ryhmien välillä, full matching -algoritmi on soveltuva. Lisäksi otoskoon säilyminen alkuperäisenä on full matching -algoritmin etu. (Leite 2017, luku 5.)

2.4 Menetelmän heikkouksia

Kuten aiemmin on pariin otteeseen todettu, kaltaistus propensiteettilukujen perusteella ei aina ole ideaali vaihtoehto. King ja Nielsen (2019) ovat tuoneet esiin propensiteettiluvuilla kaltaistuksen heikkouksia: kaltaistus voi pahimmillaan lisätä kovariaattien jakaumien epätasapainoa, mikäli kaltaistusprosessi jatkuu yli vaiheen, jossa algoritmi saavuttaa optimaalisen kaltaistuksen ja jolloin koeasetelma on lähimmillään satunnaistettua. Tämä seuraa liiasta yksilöiden karsiutumisesta otoksesta.

Kun kaltaistus suoritetaan propensiteettiluvuilla, pyritään luomaan koeasetelma, joka on satunnaistetun koeasetelman kaltainen. Aiemmin todettiin, että taustamuuttujien informaatio tiivistyy propensiteettilukuun, jolloin voidaan käyttää yhtä lukua tasapainottamaan taustamuuttujien jakaumat, sen sijaan, että ryhmien taustamuuttujien jakaumat tasapainotettaisiin suoraan havaittujen arvojen perusteella (exact matching). Kuitenkin muilla kaltaistusmenetelmillä, kuten Mahalanobisin etäisyysmittaa käyttävällä kaltaistuksella (Mahalanobis distance matching) tai karkeasti havaittujen arvojen perusteella kaltaistuksella (coarsened exact matching) voitaisiin luoda koeasetelma, joka on satunnaistettu lohkoittain tiettyjen taustamuuttujien perusteella, näin taaten suuremman samankaltaisuuden koe- ja kontrolliryhmien välillä. Satunnaistetussa koeastelmassa oletetaan satunnaisen ryhmiin valikoitumisen takaavan suurin piirtein samankaltaiset ryhmät taustamuuttujiltaan, mutta asetelma ei kuitenkaan välttämättä tuota täydellistä tasapainoa jakaumissa.

Lisäksi on esitetty kritiikkiä propensiteettiluvuilla kaltaistuksen teorian perustelua kohtaan. Kuten aiemmin todettiin, Rosenbaum ja Rubin (1983) ovat todistaneet, että mikäli kaavassa (2.3) esitetty oletus on voimassa ehdolla $\boldsymbol{x}$, se on myös voimassa ehdolla $e(\boldsymbol{x})$. Sama ei kuitenkaan päde toisinpäin, jolloin teoria olisi käyttökelpoisempi. Tasapainotus propensiteettiluvun ehdolla ei vastaa tasapainotusta muuttujien $\boldsymbol{x}$ havaittujen arvojen ehdolla.

Suositeltavaa kaltaistusta tehdessä on kiinnittää erityistä huomiota aineiston ominaisuuksiin ja pyrkiä valitsemaan menetelmä, joka parhaiten tuottaa tasapainoa kovariaattien jakaumissa. Propensiteettiluvuilla kaltaistus ei välttämättä takaa tasapainotusta jakaumissa ja kaltaistuksen suorittamisen jälkeen on syytä tutkia mahdollisesti jäljelle jäävää epätasapainoa. Lisäksi on kiinnitettävä huomiota kaltaistetun otoksen kokoon ja välttää liikaa yksiköiden karsiutumista. (King & Nielsen 2019.)

3 Kaltaistus FOLK-rekisteriaineistossa

3.1 Kaltaistukseen käytetty aineisto

Käytetty aineisto on peräisin Tilastokeskuksen (2020) FOLK-rekisteriaineistosta. Aineisto sisältää vuodesta 1988 alkaen kunakin vuonna Suomessa työikäiseen väestöön tulleet henkilöt. Tällä hetkellä rekisteritietoa on vuosiin 2017-2018 saakka. Aineistoa käytettiin Tilastokeskuksen hallinnoiman FIONA-etäkäyttöpalvelun kautta käyttölupanumerolla TK/849/07.03.00/2020.

Kaltaistusta varten aineisto rajattiin kohortteihin 1970, 1975 ja 1980 riittävän tiedon takaamiseksi. Tutkimusasetelmassa kiinnostuksen kohteena on lasten saamisen vaikutus tuloihin. Erityisesti on kiinnostavaa tarkastella vaikutuksen eroa sukupuolittain, joten analyyseja varten aineisto jaettiin miesten ja naisten aineistoihin. Seuranta alkaa yksilön täyttäessä 30 vuotta ja tarkasteltava seuranta-ajanjakso on rajattu seitsemään vuoteen. Tässä tapauksessa koe- ja kontrolliryhmät muodostettiin siten, että seurannan aikana lapsia saavat henkilöt muodostavat koeryhmän ja ne, jotka eivät saa lapsia, muodostavat kontrolliryhmän.

3.1.1 Valitut taustamuuttujat

Taustamuuttujat, joiden perusteella kaltaistus suoritetaan tulisi käsittää sekoittavat tekijät sekä tutkittavaan vastemuuttujaan suoraan vaikuttavat tekijät. Sekoittava tekijä on sellainen muuttuja, jolla on vaikutusta sekä käsittelyryhmään valikoitumiseen, että suoraan tutkittavaan vastemuuttujaan. Sekoittavat tekijät on siis pyrittävä erottamaan välittävistä tekijöistä eli muuttujista, joihin käsittelyryhmään valikoituminen vaikuttaa ja siten välittää käsittelyn vaikutusta vastemuuttujaan. Välittävien tekijöiden sisällyttäminen poistaa osan käsittelynvaikutuksesta. Lisäksi muuttujat, jotka ovat yhteydessä tutkittavaan vasteeseen, mutta joilla ei ole yhteyttä käsittelyryhmään valikoitumiseen, tulee sisällyttää malliin, sillä niiden sisällyttäminen vähentää käsittelyn vaikutusten estimaattien varianssia. (Leite 2017, luku 2.)

Tässä tapauksessa valittiin kontrolloitaviksi taustamuuttujiksi muuttujat sukupuoli ja lapset ennen seurantaa, kohortti, koulutusaste seurannan alussa, työllisyys-, työttömyys- ja opiskeluvuodet ikävuosien 23-29 välisenä aikana, parisuhdestatus seurannan alussa sekä valtionverotuksen alaiset tulot 29 vuoden iässä korjattuna vuoden 2015 tasolle.

Taulukossa 3.1. on esitelty kaltaistuksessa käytettävien kategoristen muuttujien frekvenssit koko aineiston osalta. Lisäksi nähdään sukupuolen jakautuneisuus aineistossa: 51.2 % miehiksi- ja 48.8 % naisiksi määriteltyjä. Yhteensä yksilöitä aineistossa on 192 248 ja puuttuvia tietoja sisältävien rivien poistamisen jälkeen jää 181 596 yksilöä. Puuttuva tieto on pääsääntöisesti peräisin maasta muuttaneista tai kuolleista yksilöistä.

Sukupuolittain jaetun datan frekvenssitaulukko kategoristen muuttujien osalta on liitteen A taulukossa 1. Huomattavimmat erot näyttäisi olevan koulutusasteessa ja siinä, onko lapsia ennen seurantaa vai ei. Naisista 45.1 %:lla on lapsia ennen seurantaa ja miehistä 32.4 %:lla. Näiden lasten saamisen vaikutus tuloihin jää tämän asetelman ulkopuolelle.

Taulukko 3.1. Kaltaistuksessa käytettävät kategoriset muuttujat

		Lukumäärä	Osuus	Puuttuu
Sukupuoli	Mies	98 439	51.2 %	
	Nainen	93 809	48.8 %	
	Yhteensä	192 248		0
Kohortit	1970	62 305	32.4 %	
	1975	64 553	33.6 %	
	1980	65 390	34.0 %	
	Yhteensä	192 248		0
Parisuhdestatus	Sinkut ja yksinhuoltajat	65 666	34.2 %	
	Avio- ja avoparit	126 582	65.8 %	
	Yhteensä	192 248		0
Koulutusaste	Perus ja toinen aste	114 273	59.4 %	
	Alempi korkeakoulu	51 900	27.0 %	
	Ylempi korkeakoulu	26 075	13.6 %	
	Yhteensä	192 248		0
Lapsia ennen seurantaa	Ei	112 385	61.5 %	
	Kyllä	70 497	38.5 %	
	Yhteensä	182 882		9 366
Lapsia seurannan aikana	Ei	90 317	47.7 %	
	Kyllä	98 953	52.3 %	
	Yhteensä	189 270		2 978

Taulukossa 3.2. on esitetty kaltaistuksessa käytettävien numeeristen muuttujien jakaumat koko aineistossa. Nähdään, että puuttuvaa tietoa löytyy opiskelu-, työttömyys- ja työllisyysvuosimuuttujien osalta kustakin 9 303 yksilöä. Ryhmä, josta nämä ovat peräisin, vaikuttaa lähemmän tarkastelun perusteella satunnaisesti syntyneeltä. Tässä siis oletetaan puuttuvien arvojen syntyneen täysin satunnaisesti.

Taulukko 3.2. Kaltaistuksessa käytettävät numeeriset muuttujat

	Tulot 29v	Opiskeluvuodet 23-29v	Työttömyysvuodet 23-29v	Työllisyysvuodet 23-29v
Minimi	0.00	0.00	0.00	0.00
Mediaani	23 806.74	0.00	0.00	5.00
Keskiarvo	24 105.54	1.01	0.77	4.67
Maksimi	101 071.12	7.00	7.00	7.00
Keskihajonta	14 205.47	1.47	1.34	2.25
Lukumäärä	190 740	182 945	182 945	182 945
Puuttuu	1 508	9 303	9 303	9 303

Sukupuolittain jaetun datan numeeristen muuttujien jakaumataulukot ovat liitteen A taulukoissa 2 ja 3. Vuosituloissa 29 vuoden iässä on joitain eroavaisuuksia: miehillä tulojen keskiarvo on 27 418.86 € ja naisilla 20 628.31 €. Toisaalta keskihajonta miehillä on hieman suurempi: 15 273.96 € kun naisilla se on 12 046.91 €.

3.1.2 Taustamuuttujien jakaumatarkasteluja käsittelyryhmittäin

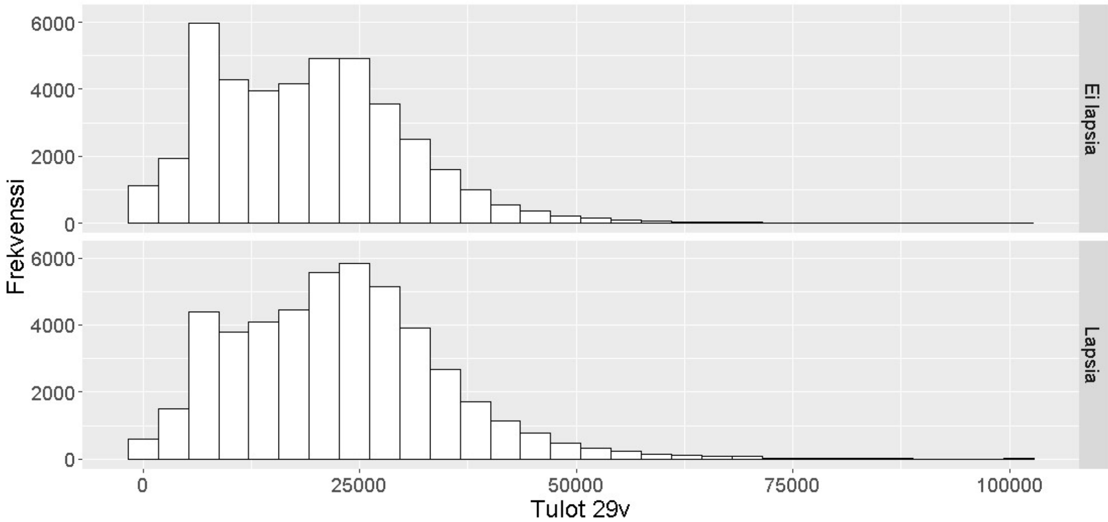
Koska kaltaistuksen tavoitteena on tasapainottaa taustamuuttujien jakaumat koe- ja kontrolliryhmien välillä on mielenkiintoista tarkastella ehdollisia jakaumia ennen kaltaistusprosessia. Kuvassa 3.1. on naisten aineistossa muuttujan ”tulot 29 vuoden iässä” jakaumat käsittelyryhmittäin. Jakaumissa on havaittavissa pieniä eroja: selkeä piikki erottuu jakauman vasemmassa päässä niillä, jotka eivät saa lapsia seurannassa. Miesten vastaavat jakaumat ovat kuvassa 3.2. Nähdään, että myös miesten jakaumat eroavat toisistaan käsittelyryhmittäin. Lisäksi jos tarkastellaan miesten ja naisten eroja näissä jakaumissa voidaan huomata, että miesten jakaumat ovat selkeästi levittäytyneet oikeammalle naisiin verrattuna. Voidaan siis päätellä tuloissa olevan eroja sukupuolittain 29 vuoden iässä.

Lisäksi liitteen A taulukossa 4 on naisten aineistossa frekvenssijakaumat käsittelyryhmittäin muuttujille koulutusaste, lapset ennen seurantaa ja parisuhdestatus. Nämä muuttujat on esitetty, sillä niiden osalta oli havaittavissa eroja. Naisista, jotka saavat lapsia seurannan aikana, 18.4 % on sinkkuja tai yksinhuoltajia ja 81.6 % lukeutuu avo- tai aviopareihin seurannan alkaessa. Vastaavat osuudet niillä, jotka eivät saa lapsia seurannan aikana on 41.9 % ja 58.1 %. Näistä luvuista tosin jää tavoittamatta ne parit, jotka eivät lukeudu avio- tai avopareihin.

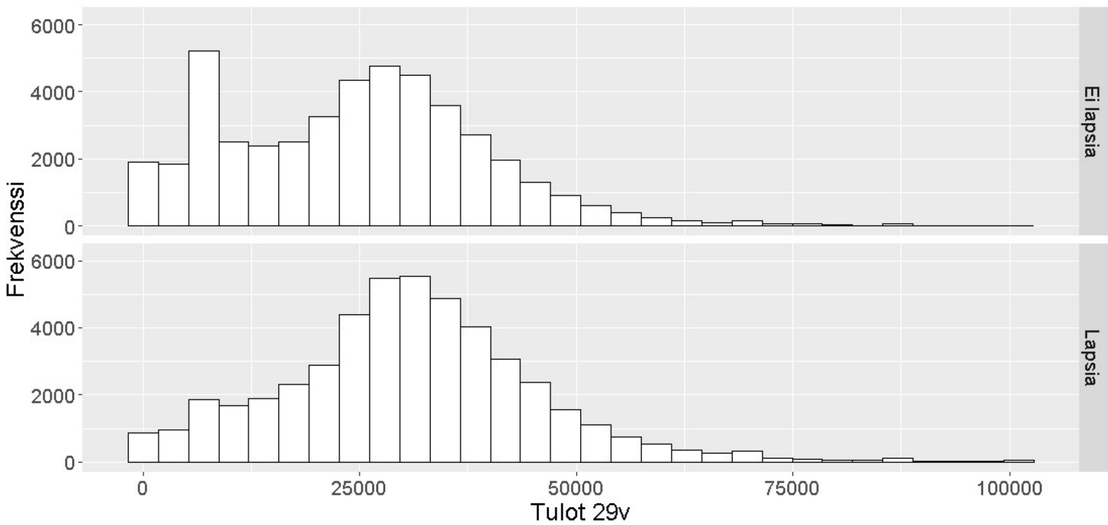
Muita huomattavia eroja löytyy koulutuasteen osalta. Naisista, jotka saavat lapsia seurannan aikana, 20.2 % on ylemmän korkeakoulun käyneitä, 36.8 % alemman korkeakoulun ja 43.0 % perus- tai toisen asteen. Vastaavasti niistä, jotka eivät saa lapsia on 10.8 % ylemmän korkeakoulun käyneitä, 29.6 % alemman korkeakoulun ja

59.6 % perus tai toisen asteen. Lapsia ennen seurantaa -muuttujan jakaumat näyttävät olevan melko tasapainossa ryhmien välillä.

Miesten vastaavat frekvenssijakaumat on liitteen A taulukossa 5. Naisiin verraten parisuhdestatus näyttää jakautuneen saman suuntaisesti. Ero naisiin nähden on se, että niillä miehillä, jotka saavat lapsia seurannan aikana on lapsia seurannan alussa 60.8 %:lla ja niillä, jotka eivät saa, luku on 25.3 %. Ennen seurantaa ja seurannan aikana lapsettomia miehiä on 74.7 % . Miehiä, joilla lapsia on ennen seurantaa sekä saavat lisää seurannassa on 39.3 %. Koulutusasteen jakaumaerot käsittelyryhmittäin näyttävät saman suuntaisilta kuin naisten aineistossa. Miesten aineistossa koulutuaste -muuttujan jakauma näyttää olevan painottuneempi alemmille koulutusasteille naisiin verrattuna.



Kuva 3.1. Tulot 29 vuoden iässä käsittelyryhmittäin naisten aineistossa



Kuva 3.2. Tulot 29 vuoden iässä käsittelyryhmittäin miesten aineistossa

3.2 Kaltaistuksen tulokset

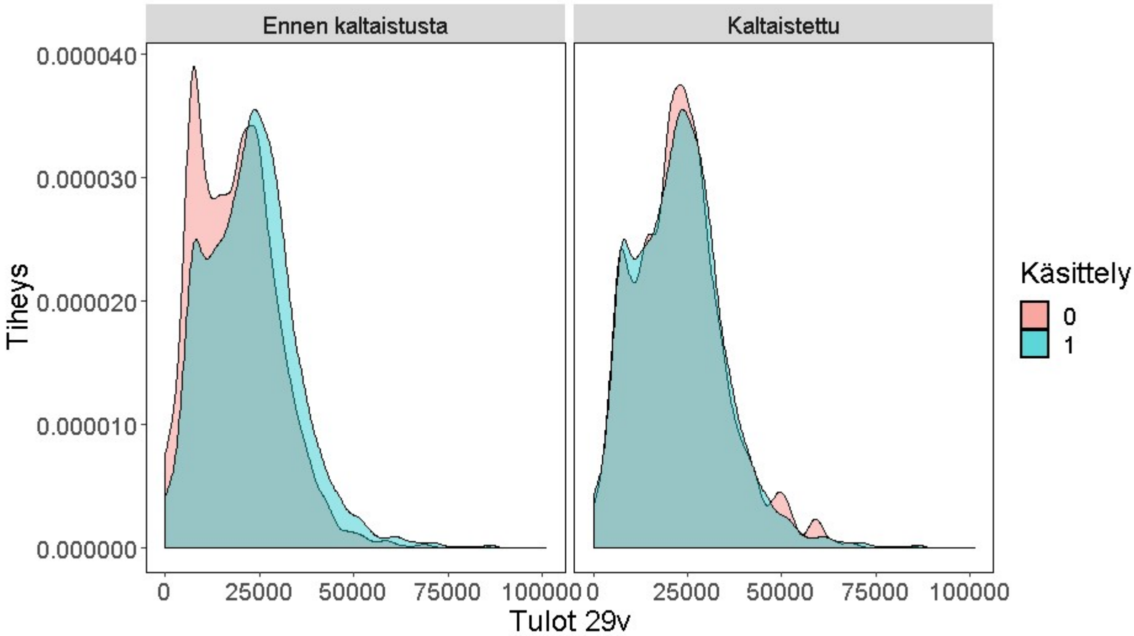
Analyyseihin käytettiin R-ohjelmointiympäristön versiota 4.0.3 (R Core Team, 2021). Kaltaistus toteutettiin `MatchIt`-paketin versiolla 4.1.0. Paketin `matchit` -funktiolla voidaan suorittaa propensiteettilukujen estimointi sekä kaltaistusprosessi. Estimointi suoritetaan oletusarvoisesti logistisella regressiolla. (Ho, Imai, King & Stuart 2011.) Lisäksi tulosten visuaaliseen arviointiin käytettiin `cobalt`-pakettia ja sen versiota 4.3.1 (Greifer 2021).

Kaltaistusta varten otettin 10 000 yksilön satunnaisotokset miesten sekä naisten aineistoista. Molemmissa on enemmän koeryhmän yksilöitä kuin kontrolliryhmän, mutta kokoero ryhmien välillä ei ole huomattavan suuri. Naisten otoksessa on 5 304 koeryhmän yksilöä ja 4 696 kontrolliryhmän kun taas miesten otoksessa 5 117 koeryhmän yksilöä ja 4 883 kontrolliryhmän. Koska kokoero ei ole suuri, kaltaistus tehtiin full matching -algoritmilla käyttäen `matchit`-funktion optiota `method = "full"`.

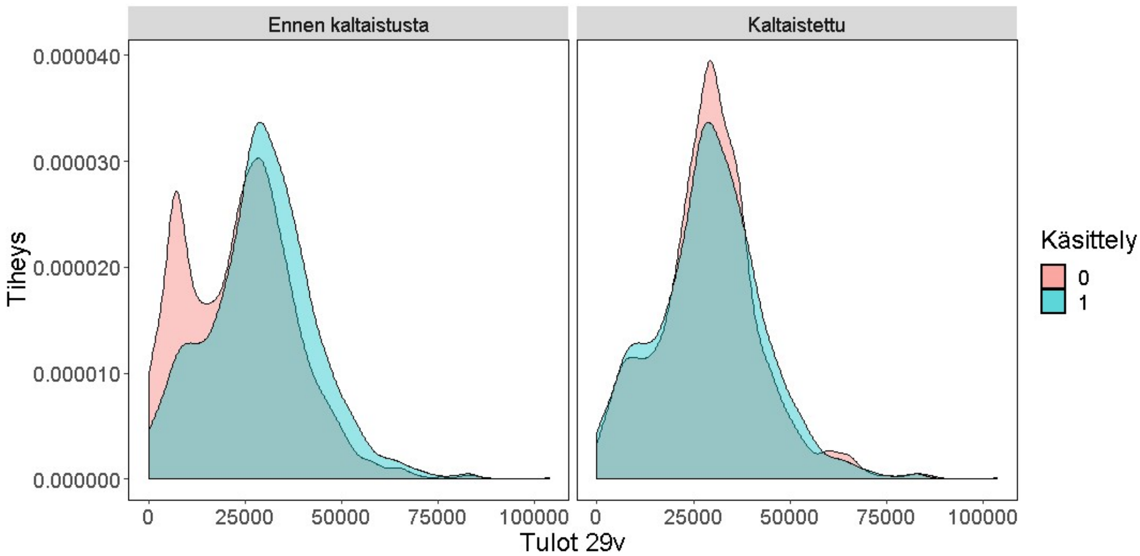
Liitteessä B on molempien aineistojen kaltaistuksen tulosteet. Tulosteissa on lasketut tunnusluvut ensin alkuperäisestä otoksesta ja sitten kaltaistetusta otoksesta. Tunnusluvut kovariaattien tasapainolle käsittelyryhmien välillä, jotka funktio `summary` tulostaa `matchit`-objektille, ovat standardoidut keskierot (standardized mean differences), varianssien suhdeluvut (variance ratio) sekä eCDF (empirical cumulative density functions) -arvojen erotukset. eCDF -arvojen erotusten maksimi vastaa Kolmogorov-Smirnov-testisuureiden arvoja. Kuitenkin hypoteesin testaukseen perustuvien arviointimenetelmien sopivuutta otoksen ominaisuuksia koskeviin kysymyksiin on kyseenalaistettu. (Greifer 2020.) Tästä syystä niitä ei käytetä arvioinnissa.

Riittävän tasapainon kriteeriksi standardoidulle keskierolle on suositeltu muun muassa itseisarvoltaan enintään 0.1 keskieroa (Leite 2017, luku 1). Naisten sekä miesten kaltaistettujen otosten tuloksista nähdään, että kummassakin standardoitujen keskierojen arvot ovat kaikilla selittäjillä alle 0.1. Voidaan myös tarkastella käsittelyryhmien varianssien suhdelukuja numeerisille muuttujille. Mikäli käsittelyryhmien varianssit ovat samankaltaiset, suhdeluku lähestyy lukua yksi (Greifer 2020). On ehdotettu, että suhdeluvun tulisi olla välillä 0.8 ja 1.2, jotta tasapainotuksen voi katsoa onnistuneeksi (Leite 2017, luku 1). Tulosteista nähdään, että molemmissa otoksissa varianssien suhdeluvut täyttävät tämän ehdon.

Lisäksi kuvassa 3.3. on naisten otoksessa muuttujan ”tulot 29 vuoden iässä” jakaumat käsittelyryhmissä ennen kaltaistusta ja kaltaistuksen jälkeen. Nähdään, että kaltaistetussa otoksessa muuttujan jakaumat ovat huipulta päällekkäisemmät kuin ennen kaltaistusta. Toisaalta kaltaistus näyttää tuottaneen hienoista epätasapainoa tulojen kasvaessa. Miesten vastaava kuvaaja on kuvassa 3.4. Jälleen kaltaistetussa aineistossa jakaumat ovat selkeästi päällekkäisemmät, kuin ennen kaltaistusta. Edelleen on pientä epätasapainoa havaittavissa jakaumien huipulla, jossa lapsettomien ryhmän jakauman huippu erottuu selkeästi.



Kuva 3.3. Jakaumien tasapaino naisten otoksessa muuttujalla tulot 29 vuoden iässä



Kuva 3.4. Jakaumien tasapaino miesten otoksessa muuttujalla tulot 29 vuoden iässä

3.3 Yhteenveto

Yhteenvetona voidaan todeta, että esimerkkiaineistossa kaltaistus propensiteettiluvuilla oli melko onnistunut sillä jokaisen kontrolloitavan muuttujan osalta arviointikriteerit täyttyivät. Ainoastaan jakaumatarkastelussa oli havaittavissa epätasapainoa vielä kaltaistuksen jälkeen. Avoimeksi kysymyksesi jää, olisiko jollain muulla kaltaistusmenetelmällä saanut aikaan vielä paremman tasapainon kaltaistuksessa.

Tällä kaltaistetulla aineistolla mahdollisesti voisi estimoida lasten saamisen vaikutusta tuloihin seitsemänä seurantavuotena. Käsittelyn vaikutuksista ATT:n (kaava 2.2) estimointia varten `matchit`-objektista saadaan kaltaistettu aineisto funktiolla `match.data`. Kaltaistettu aineisto sisältää alkuperäisen aineiston, sekä estimoidut propensiteettiluvut ja painotukset kaltaistetuille yksilöille. (Ho et al. 2011.)

Vaikka kaltaistus tässä esimerkissä onnistui, asetelmaan liittyy monia heikkouksia. Ensinnäkin tästä asetelmasta jää ulkopuolelle seurannan aikana syntyneiden lasten vaikutus tuloihin 37 ikävuoden jälkeisenä aikana sekä sen jälkeen vielä syntyvien lasten vaikutus. Lisäksi asetelma ei ota huomioon syntyvien lasten lukumäärän mahdollista vaikutusta. Mahdollisesti asetelmasta on myös jäänyt ulkopuolelle muuttujia, joilla on todellisuudessa vaikutusta käsittelyryhmään valikoitumiseen ja siten kaltaistuksen ei voi olettaa takaavan harhattomia tuloksia. Tässä valittiin mukaan saatavilla olevista muuttujista ne, jotka olivat käyttökelpoisia ja mahdollisesti yhteydessä ryhmään valikoitumiseen ja tutkittavaan vasteeseen.

Lähteet

Greifer, N. (2020). *Assessing Balance.* https://CRAN.R-project.org/web/packages/MatchIt/vignettes/assessing-balance.html (viitattu 22.04.2021)

Greifer, N. (2021). *cobalt: Covariate Balance Tables and Plots.* Versio 4.3.1. https://CRAN.R-project.org/package=cobalt

Ho, D. E., Imai, K., King, G. & Stuart, E. A. (2011). *MatchIt: Nonparametric Preprocessing for Parametric Causal Inference.* Journal of Statistical Software, 42(8), 1–28. https://www.jstatsoft.org/v42/i08/

Holmes, W. (2014). *Using propensity scores in quasi-experimental designs.* SAGE Publications, Ltd. https://doi.org/10.4135/9781452270098

King, G. & Nielsen, R. (2019). *Why propensity scores should not be used for matching*. Political Analysis, 27(4), 435-454. https://doi.org/10.1017/pan.2019.11

Leite, W. (2017). *Practical propensity score methods using R*. SAGE Publications, Inc. https://doi.org/10.4135/9781071802854

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosenbaum, P. R. & Rubin, D. B. (1983). *The Central Role of the Propensity Score in Observational Studies for Causal Effects.* Biometrika, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Tilastokeskus (2020). *FOLK perustieto -aineistokuvaus.* Taika-tutkimusaineistokatalogi. https://taika.stat.fi/fi/aineistokuvaus.html#!?dataid=FOLK_19872019_jua_perus20_002.xml

Liite A: Ehdolliset frekvenssitaulukot

Taulukko 1. Kaltaistuksessa käytettävät kategoriset muuttujat

		Sukupuoli					
		Mies			Nainen		
		Määrä	Osuus	Puuttuu	Määrä	Osuus	Puuttuu
Kohortit	1970	31638	32,1%		30667	32,7%	
	1975	33189	33,7%		31364	33,4%	
	1980	33612	34,1%		31778	33,9%	
	Yhteensä	98439		0	93809		0
Parisuhdestatus	Sinkut ja yksinhuoltajat	37991	38,6%		27675	29,5%	
	Avio- ja avoparit	60448	61,4%		66134	70,5%	
	Yhteensä	98439		0	93809		0
Koulutusaste	Perus tai toinen aste	66657	67,7%		47616	50,8%	
	Alempi korkeakoulu	20692	21,0%		31208	33,3%	
	Ylempi korkeakoulu tai tutkija	11090	11,3%		14985	16,0%	
	Yhteensä	98439		0	93809		0
Lapsia seurannan aikana	Ei	47243	48,8%		43074	46,6%	
	Kyllä	49629	51,2%		49324	53,4%	
	Yhteensä	96872		1567	92398		1411
Lapsia ennen seurantaa	Ei	63422	67,6%		48963	54,9%	
	Kyllä	30334	32,4%		40163	45,1%	
	Yhteensä	93756		4683	89126		4683

Taulukko 2. Kaltaistuksessa käytettävät numeeriset muuttujat naisten datassa

	Tulot 29v	Opiskeluvuodet 23-29	Työllisyysvuodet 23-29	Työttömyysvuodet 23-29
Minimi	0,00	0,00	0,00	0,00
Mediaani	20277,12	0,00	5,00	0,00
Keskiarvo	20628,31	1,07	4,42	0,72
Maksimi	101071,12	7,00	7,00	7,00
Keskihajonta	12046,91	1,48	2,27	1,19
Määrä	93809	93809	93809	93809
Puuttuu	741	4648	4648	4648

Taulukko 3. Kaltaistuksessa käytettävät numeeriset muuttujat miesten datassa

	Tulot 29v	Opiskeluvuodet 23-29	Työllisyysvuodet 23-29	Työttömyysvuodet 23-29
Minimi	0,00	0,00	0,00	0,00
Mediaani	27807,86	0,00	6,00	0,00
Keskiarvo	27418,86	0,95	4,90	0,83
Maksimi	101071,12	7,00	7,00	7,00
Keskihajonta	15273,96	1,46	2,22	1,47
Määrä	98439	98439	98439	98439
Puuttuu	767	4655	4655	4655

Taulukko 4. Ehdollisia frekvenssijakaumia käsittelyryhmittäin naisten datassa

		Lapsia seurannan aikana			
		Ei		Kyllä	
		Määrä	Osuus	Määrä	Osuus
Parisuhdestatus	Sinkut ja yksinhuoltajat	18059	41,9%	9069	18,4%
	Avio- ja avoparit	25015	58,1%	40255	81,6%
	Yhteensä	43074	100,0%	49324	100,0%
Lapsia ennen seurantaa	Ei	24247	58,5%	24277	51,6%
	Kyllä	17215	41,5%	22799	48,4%
	Yhteensä	41462	100,0%	47076	100,0%
Koulutusaste	Perus, toinen aste	25663	59,6%	21195	43,0%
	Alempi korkeakoulu	12739	29,6%	18165	36,8%
	Ylempi korkeakoulu tai tutkija	4672	10,8%	9964	20,2%
	Yhteensä	43074	100,0%	49324	100,0%

Taulukko 5. Ehdollisia frekvenssijakaumia käsittelyryhmittäin miesten datassa

		Lapsia seurannan aikana			
		Ei		Kyllä	
		Määrä	Osuus	Määrä	Osuus
Parisuhdestatus	Sinkut ja yksinhuoltajat	27370	57,9%	9944	20,0%
	Avio- ja avoparit	19873	42,1%	39685	80,0%
	Yhteensä	47243	100,0%	49629	100,0%
Lapsia ennen seurantaa	Ei	34052	74,7%	28837	60,8%
	Kyllä	11562	25,3%	18607	39,2%
	Yhteensä	45614	100,0%	47444	100,0%
Koulutusaste	Perus ja toinen aste	35390	74,9%	30283	61,0%
	Alempi korkeakoulu	8262	17,5%	12194	24,6%
	Ylempi korkeakoulu tai tutkija	3591	7,6%	7152	14,4%
	Yhteensä	47243	100,0%	49629	100,0%

Liite B: R-tulosteet

```
## Kaltaistus naisten aineistossa


Call:
matchit(formula = saalapsia ~ aste + paris + onlapsia + kohortit +
tyov + tyotonv + opiskv + tulot29v, data = fs, method = "full",
distance = "logit")

Summary of Balance for All Data:
                                Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                                 0.7019     0.8400    0.1857   0.2725
asteperus, toinen aste                  -0.3524          .    0.1738   0.1738
astealempi kk                            0.1586          .    0.0767   0.0767
asteylempi kk, tutkija                   0.2391          .    0.0972   0.0972
parissinkut ja yksinhuoltajat           -0.5771          .    0.2256   0.2256
parisavio- ja avoparit                   0.5771          .    0.2256   0.2256
onlapsia0                               -0.0948          .    0.0473   0.0473
onlapsia1                                0.0948          .    0.0473   0.0473
kohortit1970                            -0.0930          .    0.0430   0.0430
kohortit1975                             0.0412          .    0.0196   0.0196
kohortit1980                             0.0492          .    0.0234   0.0234
tyov                                     0.2817     0.7980    0.0747   0.1071
tyotonv                                 -0.2755     0.6213    0.0359   0.0959
opiskv                                   0.0395     0.9522    0.0107   0.0444
tulot29v                                 0.3112     1.2392    0.0951   0.1373


Summary of Balance for Matched Data:
                                Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                                 0.0006     1.0014    0.0006   0.0043
asteperus, toinen aste                   0.0113          .    0.0056   0.0056
astealempi kk                           -0.0013          .    0.0006   0.0006
asteylempi kk, tutkija                  -0.0121          .    0.0049   0.0049
parissinkut ja yksinhuoltajat           -0.0264          .    0.0103   0.0103
parisavio- ja avoparit                   0.0264          .    0.0103   0.0103
onlapsia0                                0.0614          .    0.0307   0.0307
onlapsia1                               -0.0614          .    0.0307   0.0307
kohortit1970                            -0.0234          .    0.0108   0.0108
kohortit1975                            -0.0148          .    0.0070   0.0070
kohortit1980                             0.0375          .    0.0178   0.0178
tyov                                    -0.0027     1.0146    0.0030   0.0091
tyotonv                                 -0.0016     0.9968    0.0019   0.0064
opiskv                                   0.0093     0.9918    0.0051   0.0131
tulot29v                                -0.0370     0.9599    0.0098   0.0296
```

```
Sample Sizes:
  Control Treated
All           4696.      5304
Matched (ESS) 1135.57    5304
Matched       4696.      5304
Unmatched        0.         0
Discarded        0.         0
```

```
## Kaltaistus miesten aineistossa
```

```
Call:
matchit(formula = saalapsia ~ onlapsia + aste + paris + kohortit +
tyov + tyotonv + opiskv + tulot29v, data = ms, method = "full",
distance = "logit")

Summary of Balance for All Data:
                               Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                                1.0524     0.6726    0.2430   0.3925
onlapsia0                              -0.2822          .    0.1376   0.1376
onlapsia1                               0.2822          .    0.1376   0.1376
asteperus, toinen aste                 -0.3104          .    0.1519   0.1519
astealempi kk                           0.1919          .    0.0839   0.0839
asteylempi kk, tutkija                  0.1956          .    0.0680   0.0680
parissinkut ja yksinhuoltajat          -0.9798          .    0.3869   0.3869
parisavio- ja avoparit                  0.9798          .    0.3869   0.3869
kohortit1970                           -0.0413          .    0.0192   0.0192
kohortit1975                            0.0655          .    0.0314   0.0314
kohortit1980                           -0.0259          .    0.0122   0.0122
tyov                                    0.4230     0.5974    0.1000   0.1513
tyotonv                                -0.3495     0.5451    0.0543   0.1102
opiskv                                 -0.0385     0.7846    0.0132   0.0234
tulot29v                                0.3922     0.9711    0.1151   0.1670
```

Summary of Balance for Matched Data:

	Std. Mean Diff.	Var. Ratio	eCDF Mean	eCDF Max
distance	0.0000	0.9989	0.0007	0.0066
onlapsia0	0.0627	.	0.0306	0.0306
onlapsia1	-0.0627	.	0.0306	0.0306
asteperus, toinen aste	0.0088	.	0.0043	0.0043
astealempi kk	-0.0205	.	0.0090	0.0090
asteylempi kk, tutkija	0.0134	.	0.0047	0.0047
parissinkut ja yksinhuoltajat	-0.0079	.	0.0031	0.0031
parisavio- ja avoparit	0.0079	.	0.0031	0.0031
kohortit1970	-0.0050	.	0.0023	0.0023
kohortit1975	0.0080	.	0.0038	0.0038
kohortit1980	-0.0032	.	0.0015	0.0015
tyov	-0.0315	0.9680	0.0102	0.0264
tyotonv	0.0102	1.0223	0.0019	0.0098
opiskv	0.0193	0.9659	0.0052	0.0269
tulot29v	0.0169	1.0988	0.0168	0.0539

Sample Sizes:

	Control	Treated
All	4883.	5117
Matched (ESS)	789.47	5117
Matched	4883.	5117
Unmatched	0.	0
Discarded	0.	0