Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät

Alueellisen terveys- ja hyvinvointitutkimuksen kyselyaineisto 2010

Oona Pentala

Helsingin Yliopisto

Valtiotieteellinen tiedekunta

Tilastotiede

Pro gradu - tutkielma

Marraskuu 2014

HELSINGIN YLIOPISTO - HELSINGFORS UNIVERSITET - UNIVERSITY OF HELSINKI

Tiedekunta/Osasto — Fakultet/Sektion — Faculty	Laitos — Institution — Department
Valtiotieteellinen tiedekunta	Sosiaalitieteiden laitos

Tekijä — Författare — Author
Oona Pentala

Työn nimi — Arbetets titel — Title
Väestötutkimusaineiston tilastolliset kadonhallintamenetelmät

Oppiaine — Läroämne — Subject
Tilastotiede

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
Pro gradu -tutkielma	Marraskuu 2014	77 s. + liitteet

Tiivistelmä — Referat — Abstract

Väestötutkimuksilla kerätään tietoja, joita ei rekistereistä saada. Tällaista tietoa ovat esimerkiksi väestön arviot terveydentilastaan, mielipiteet ja palveluiden tarpeen tyydyttyminen. Väestötutkimuksen taustalla on aina sopivalla otantamenetelmällä poimittu otos, jonka katsotaan edustavan tutkimuksen kohteena olevaa väestöjoukkoa. Valitettavasti tällaisten väestötutkimusten ja erityisesti tiedonkeruumenetelmänä kyselyä käyttävien tutkimusten vastauskato on ollut kuitenkin nousussa koko 2000-luvun alun. Tämä tarkoittaa, ettei edellä mainittu otos enää edustakaan tutkimuksen kohteena olevaa väestöä, jolloin tutkimuksesta saadut tulokset eivät välttämättä ole suoraan yleistettävissä alkuperäiseen perusjoukkoon.

Tässä pro gradu-työssä käsitellään väestötutkimusaineiston tilastollisia kadonhallintamenetelmiä, joissa tavoitteena on tuottaa mahdollisimman luotettavia väestöä edustavia tilastollisia tunnuslukuja mahdollisesta vastauskadosta huolimatta. Työssä käsitellään vastauskatoa ensin otannan ja tiedonkeruumenetelmien näkökulmasta, jolloin vastauskadon muodostumiseen voidaan vaikuttaa. Empiirisen aineiston avulla kuvaillaan, millaista vastauskatoa kyselynä toteutetussa väestötutkimusaineistossa esiintyy ja tarkastellaan, millaisia tuloksia tilastollisilla kadonhallintamenetelmillä saadaan vastauskatoa sisältävästä aineistosta. Aineistona käytetään Terveyden ja hyvinvoinnin laitoksen (THL) Alueellisen terveys- ja hyvinvointitutkimuksen (ATH) vuonna 2010 kerättyä aineistoa. Lisäksi kadonhallintamenetelmissä hyödynnetään Suomessa hyvin saatavilla olevaa rekisteriperäistä tietoa Väestörekisterikeskukselta, Tilastokeskukselta ja Kansaneläkelaitokselta, minkä avulla saadaan arvokasta tietoa myös katoon jääneistä vastaajista.

Tilastollisina kadonhallintamenetelminä tässä työssä käytetään Inverse Probability Weighting (IPW)- painotusmenetelmää, painotettua Hot Deck-imputointia ja moni-imputointia. Näillä menetelmillä saatuja tuloksia verrataan sekä keskenään, että estimaatteihin, jotka on tuotettu menetelmillä, jotka eivät huomioi vastauskatoa. Saatujen tulosten vertailukohtana käytetään myös rekisteriperäisiä tietoja tutkimusalueilta.

Työssä todetaan, että kadonhallintamenetelmillä saadaan erilaisia tuloksia kuin tavallisilla analyysimenetelmillä. Eri menetelmillä saadut tulokset ovat kaikki samansuuntaisia, mutta erityisesti moni-imputoinnilla saadaan merkitsevästi eriäviä tuloksia kuin muilla menetelmillä. Analysoitaessa mitä tahansa aineistoa, jonka tulokset on tarkoitus yleistää väestöön, vastauskadon tutkiminen ja sen aiheuttaman mahdollisen harhan huomioiminen tuloksissa olisi ensiarvoisen tärkeää. Vastauskadon huomioiminen ja sen hallintamenetelmät ovat tärkeä osa väestötutkimusaineistojen käyttöä, jolloin kadon huomioimisen tärkeys ja sen hallintamenetelmien käyttökelpoisuus olisi hyvä olla tiedossa kaikilla tutkismaineiston käyttäjillä. Selkeästi dokumentoidut otantamenetelmät, mahdollisen kadon vähentämiseen pyrkiminen jo tiedonkeruuvaiheessa ja hyvät koko otokselle saatavissa olevat rekisteritiedot omalta osaltaan edesauttavat tehokkaiden kadonhallintamenetelmien käyttöä, joilla voidaan luotettavia väestöön yleistettävissä olevia tuloksia.

Avainsanat — Nyckelord — Keywords
yksikkökato, erävastauskato, moni-imputointi, painotusmenetelmät, väestötutkimus

Säilytyspaikka — Förvaringsställe — Where deposited
Helsingin Yliopiston keskustakampuksen kirjasto

Muita tietoja — Övriga uppgifter — Additional information

Kiitokset ohjaajilleni Risto Kaikkoselle ja Tommi Härkäselle Terveyden ja hyvinvoinnin laitokselle kannustavasta ja opettavaisesta ohjauksesta kaiken kiireen keskellä. Suuri kiitos kuuluu myös kotijoukoilleni - tästä ei olisi tullut mitään ilman teidän tukeanne.

Sisältö

Johdanto

Väestötutkimuksilla kerätään tietoja, joita ei hallinnollisista rekistereistä saada. Tällaista tietoa ovat esimerkiksi väestön omat arviot terveydentilastaan, mielipiteet ja palveluiden tarpeen tyydyttyminen. Esimerkiksi lääkärikäyntien määriä voidaan seurata rekistereistä, mutta kyselyillä voidaan selvittää onko jokainen lääkäriä tarvinnut päässyt lääkärin vastaanotolle. Väestöä edustavista otoksista saadaan tärkeää tietoa esimerkiksi väestön terveydestä, elintavoista, riskitekijöistä, palveluiden tarpeesta ja tyytyväisyydestä. Oletuksena on, että jos kaikki otokseen valituiksi tulleet osallistuvat tutkimukseen, saadut tutkimustulokset ovat yleistettävissä koko väestöön. Edustavan otoksen katsotaan siis olevan väestö pienoiskoossa.

Väestötutkimusten ja erityisesti tiedonkeruumenetelmänä kyselyä käyttävien väestötutkimusten vastauskato on ollut kuitenkin nousussa koko 2000-luvun alun, eikä erityisesti nuorempaa sukupolvea ole enää moneen vuoteen saatu vastaamaan kyselyihin (Prättälä ja Tolonen (2007), Pohjanpää (2010)). Vastauskato aiheuttaa sen, ettei otos enää edustakaan alkuperäistä väestöä, jolloin tutkimuksesta saadut tulokset eivät välttämättä ole suoraan yleistettävissä alkuperäiseen perusjoukkoon eli väestöön.

Tilastollisilla kadon hallintamenetelmillä pyritään vähentämään vastauskadon aineiston edustavuuteen aiheuttamaa epävarmuutta. Aineistoja kerätään jatkuvasti ja niiden keräämiseen kuluu paljon useiden tutkimuslaitosten ja muiden kyselyjä toteuttavien tahojen resursseja. Näin ollen onkin tärkeää, että vastauskadosta huolimatta kerättyä aineistoa voidaan käyttää mahdollisimman monipuolisesti ja siten että kerätystä aineistosta saadut tulokset olisivat luotettavasti yleistettävissä koko tutkimuksen kohteena olevaan perusjoukkoon.

Aineiston tasolla vastauskato tarkoittaa puuttuvia arvoja, eli ”reikiä” aineistossa. Valitettavan usein puuttuvat arvot jätetään huomiotta aineistoa analysoidessa

ja saatuja tuloksia raportoidessa niiden yleisyydestä huolimatta. Tästä kertoo jo se, että useimmissa tilasto-ohjelmistoissa puuttuvien arvojen huomiotta jättäminen on oletusarvoista, jollei toisin määritetä. Jopa kansainvälisissä arvostetuissa tutkimusartikkeleissa suurin osa jättää puuttuvan tiedon analyysin kokonaan pois tutkimustuloksista (Carpenter et al., 2014). Tämä on hälyttävä tieto vastauskadon oletettavasti yhä kasvaessa, jolloin vastauskadon aiheuttaman epävarmuuden huomioiminen tutkimusta tehdessä olisi ensiarvoisen tärkeää luotettavien tulosten kannalta.

Tässä pro gradu-työssä käsitellään väestötutkimusaineiston tilastollisia kadon hallintamenetelmiä, joissa tavoitteena on saada aikaan mahdollisimman luotettavia väestöä edustavia tilastollisia tunnuslukuja vastauskadosta huolimatta. Aluksi luvussa yksi käydään läpi vastauskadon mahdollisia syitä ja muodostumista väestötutkimuksissa. Pääasiassa käsitellään väestötutkimuksia, jotka toteutetaan kyselytutkimuksina. Toisessa luvussa käsitellään kyselytutkimusten perusteita kuten otantamenetelmiä, tiedonkeruuta ja aineiston analyysia, sekä sitä miten mahdollinen vastauskato voidaan huomioida jo näissä tutkimuksen vaiheissa. Luvut yksi ja kaksi käsittelevät vastauskatoa ilmiönä, johon on mahdollista vaikuttaa jo aineiston keruuvaiheessa, sillä mikään tilastollinen kadonhallintamenetelmä ei korvaa aitoa vastaajalta saatua vastausta. Tämän lisäksi tutkimuksen aineistonkeruun tarkka dokumentointi helpottaa tilastollisten kadonhallintamenetelmien käyttöä.

Kolmannessa luvussa vastauskatoa tarkastellaan kerätyn aineiston näkökulmasta ja keskitytään vastauskadon analysoimiseen ja mallintamiseen. Vastauskato voidaan jakaa kahteen tyyppiin, yksikkövastauskatoon ja erävastauskatoon. Yksikkövastauskadolla tarkoitetaan, ettei vastaajalta saada vastauksia ollenkaan tutkimukseen ja erävastauskadolla sitä, että vastaajalta saadaan vastauksia vain osaan tutkimuksen kysymyksistä. Vastauskatoa analysoidessa on tärkeää kiinnittää huomiota sen määrän ja satunnaisuuteen. Mikäli vastauskato on satunnaista, eli ei riipu esimerkiksi vastaajan iästä tai sukupuolesta tai muusta kiinnostuksen kohteena olevasta muuttujasta, ei vastauskato välttämättä aiheuta aineistosta saatuihin tuloksiin harhaa. Vastauskato on väestötutkimuksissa harvoin kuitenkaan täysin satunnaista, vaan kato on suurempaa tietyissä tunnistettavissa olevissa ryhmissä. Tällöin vastauskatoa sanotaan valikoituneeksi. Suomessa tällä hetkellä vastauskato vaikuttaisi olevan korkeinta esimerkiksi matalan koulutuksen saaneet alle 30-vuotiaat miehet (Tolonen,

2005). Heitäkään ei voi jättää kuitenkaan tutkimatta, mikäli toivotaan koko väestöä edustavia tuloksia. Myös käytettävän tilastollisen kadonhallintamenetelmän valinta riippuu vastauskadon valikoituneisuudesta.

Neljännessä luvussa esitellään tässä työssä käytössä oleva aineisto, joka on Terveyden ja hyvinvoinnin laitoksen (THL) Alueellisen terveys- ja hyvinvointitutkimuksen (ATH) vuonna 2010 kerätty aineisto Turusta, Pohjois-Pohjanmaalta ja Kainuusta, sekä koko Suomen väestöä edustava aineisto. Tämän aineiston katoa analysoidaan tutkielmassa aiemmin esiteltyjen menetelmien avulla. Rekisteritietoja käyttämällä myös katoon jääneistä tutkittavista saadaan tietoa ja voidaan korjata väestöestimaatteja myös tämän tiedon avulla. Tärkeänä tietona kadon hallintamenetelmissä ja sen analysoinnissa käytetään tässä yhteydessä erilaisia rekisteritietoja mm. Väestörekisterikeskuksesta, Tilastokeskuksesta ja Kansaneläkelaitokselta, jotka ovat saatavilla koko alkuperäiselle otokselle.

Työn englanninkielisessä osiossa luvuissa viisi ja kuusi käsitellään tarkemmin erilaisia tilastollisia kadonhallintamenetelmiä ja testataan niitä empiirisellä aineistolla. Kadon hallintamenetelmistä tarkemmin esitellään yksikkökadon tapauksessa painotusmenetelmänä *Inverse probability weighting* (IPW) ja erävastauskadon tapauksessa *Hot Deck*- ja moni-imputointi (MI). Nämä menetelmät valikoituvat tähän työhön, sillä niillä on mahdollista korjata vastauskadon vaikutusta, vaikka se olisi valikoitunutta. Menetelmistä esitellään teoreettinen tausta ja käsitellään myös niiden käytettävyyttä sekä hyviä ja huonoja puolia. Menetelmillä saatuja tuloksia verrataan toisiinsa ja pelkästään korjaamattomalla aineistolla saatuihin tuloksiin. Vastemuuttujina tässä työssä käytetään aineiston taustamuuttujien jakaumia, kuten ikä- ja sukupuolijakauma, sekä itse raportoitua masennusta, josta on saatavilla vertailutiedoksi rinnastettavaa rekisteritietoa.

Lopuksi päätelmissä käydään läpi saadut tulokset suomeksi, pohditaan valittujen menetelmien toimivuutta käytetyn aineiston tapauksessa ja esitetään jatkotutkimuksen kohteeksi sopivia huomioita. Tämän tutkielman luvut 1-4 ja 7 ovat siis väestötutkimuksia, niiden katoa ja sen syitä kuvailevaa tietoa sekä tilastollisilla kadonhallintamenetelmillä saatujen tulosten esittelyä, joka on oleellista jokaiselle tutkimusta tekevälle tai sen tuloksia käyttävälle henkilölle. Empiirinen osa on syvällisempää tieteellistä analyysia englanniksi, sillä siinä esitellyistä tuloksista on tarkoitus julkaista

artikkeli myöhemmin.

Sanasto - Abbreviations

AIC Akaike Information Criterion; Akaiken informaatiokriteeri. Mallin valinnassa käytettävä kriteeri, joka ottaa huomioon mallin ja aineiston yhteensopivuuden sekä mallin monimutkaisuuden. 27

ATH Alueellinen terveys- ja hyvinvointitutkimus; Regional Health and Wellbeing study. 4, 23

BIC Bayesian Information Criterion; Bayesilainen informaatiokriteeri. Mallin valinnassa käytettävä kriteeri, joka ottaa huomioon mallin ja aineiston yhteensopivuuden, mallin monimutkaisuuden sekä otoskoon. 27, 77

CART Classification And Regression Tree modelling; Päätöspuu, luokittelualgoritmi, jossa aineistoa pilkotaan piirteiden avulla osajoukkoihin, kunnes jokainen osajoukko sisältää vain samaan luokkaan kuuluvia havaintoja. 44

CC Complete Case analysis; yksinkertainen analyysimenetelmä, jossa käytetään vain havaittuja arvoja eikä oteta kantaa puuttuviin arvoihin. 36, 56

EM Expectation Maximation-algorithm; EM-algoritmi, iteratiivinen menetelmä suurimman uskottavuuden estimaattien löytämiseksi tilastollisten mallien parametreille tilanteessa, jossa osa tiedosta puuttuu. 42

FCS Fully Conditional Specification; iteratiivinen menetelmä, jota käytetään moni-imputoinnissa. 42, 46

IPW Inverse Probability Weighting; Painotusmenetelmä, jossa käytetään hyväksi mallinnetun vastaamistodennäköisyyden käänteislukua. 4, 27, 34, 38, 76

JM Joint Modelling; Moni-imputoinnin menetelmä, jossa käytetään valittujen muuttujien yhteisjakaumaa. 42, 46

LOCF Last Observation Carried Forward, yksinkertainen pitkittäistutkimuksissa käytetty kadonkorjausmenetelmä, jossa puuttuva arvo korvataan viimeisimmällä vastaajalta saadulla arvolla. 37

MAR Missing At Random; puuttuneisuus on ei-valikoitunutta tulosmuuttujan suhteen, mutta saattaa riippua otosmuuttujista. 20, 22

MCAR Missing Completetly At Random; puuttuneisuus on täysin ei-valikoitunutta, eikä riipu tulos- tai otosmuuttujista. 20, 22

MCMC Markovian Chain Monte Carlo; iteratiivinen menetelmä, jota voidaan käyttää moni-imputoinnissa. 42, 46

MI Multiple Imputation; moni-imputointi. 4, 45, 76

MICE Multiple Imputation by Chained Equations; moni-imputoinnissa käytetty menetelmä, jossa voidaan määrittää oma imputointimalli jokaiselle puuttuvalle muuttujalle erikseen. 42, 46, 62

MNAR Missing Not At Random; puuttuneisuus on täysin valikoitunutta. 20, 22

OR Odds Ratio; Vetosuhde tai ristitulosuhde, tunnusluku suhteellisten osuuksien ja todennäköisyyksien vertailuun. 25, 28

PPS Probability Proportional to Size; Otannassa käytettävä poimintatodennäköisyys määrittyy kyseisen osajoukon koon perusteella. 12

SRS Simple Random Sampling; Yksinkertainen satunnaisotanta. 11

STR Stratified Sampling; Ositettu otanta. 11

SYS Systematic Sampling; Systemaattinen otanta. 12

THL Terveyden ja hyvinvoinnin laitos; National Institute for Health and Welfare. 4

WEE Weighted Estimation Equation; Analyysimenetelmä, jossa käytetään painokertoimia osana aineiston analyysia, jossa on puuttuvaa tietoa. 34, 40

WSHDI Weighted Sequential Hot Deck Imputation; Vastaajaluovuttajamenetelmään perustuva imputointimenetelmä, jossa käytetään painokertoimia rajoittamaan yksittäisen luovuttajan arvojen käytön määrää imputoidessa. 58

Luku 1

Vastauskadon muodostuminen

1.1 Syitä ja ratkaisuja

Syyt katoon väestötutkimuksissa, tai kyselytutkimuksissa ylipäätään, voidaan jakaa karkeasti kahteen luokkaan, tutkittavaan tavoittamattomuuteen ja tutkittavan kieltäytymiseen. Tavoittamattomuuteen liittyvät syyt voidaan jakaa vielä kahteen ryhmään sen mukaan, onko kyse siitä, ettei tutkittavaa tavoiteta vai onko tutkittava kykenemätön vastaamaan. Kykenemättömyys vastaamaan muodostaa otokseen pääasiassa alipeittoa ja kieltäytyminen yksikkökatoa, mutta tavoittamattomuus voi olla sekä alipeittoa että yksikkökatoa (kts. kuva 2.1).

Tavoittamattomuuteen on etsitty erilaisia ratkaisuja tiedonkeruutavasta riippuen. Esimerkiksi Dillman (2000) ehdottaa, että tiedonkeruussa otettaisiin huomioon kyselyiden aikatauluttaminen (esimerkiksi puhelinhaastatteluiden eri aaltojen toteuttaminen eri aikoihin päivästä), tiedonkeruuajan riittävä pituus ja mahdollisten haastattelijoiden työmäärän sopivuus. Useimmissa tutkimuksissa edellä mainittujen seikkojen huomioimisen on todettu parantavan tutkittavien tavoitettavuutta. Tavoittamattomuutta vähentävät myös ajan tasalla pidetyt yhteystiedot, eli varsinaisen tiedonkeruun aloittaminen mahdollisimman pian tutkittavien yhteystietojen saamisen jälkeen.

Tutkittavan kieltäytymistä ehkäiseviä tekijöitä on todettu olevan tiedonkeruun toteuttavan tahon hyvä maine, riittävä tiedottaminen tutkimuksesta ja sen tavoitteista, etukäteen tiedonkeruusta ilmoittaminen, kannustimien tarjoaminen tutkittaville ja vaihtoehtoisen tiedonkeruutavan tarjoaminen (Groves et al., 2002). Hyvällä

maineella voidaan tarkoittaa esimerkiksi sitä, että tiedon kerääjätaho on tunnettu viranomaislaitos, jonka tutkittavat luottavat käsittelevän antamiaan tietoja tietosuojalain mukaan. On myös todettu, että jos tutkittava ymmärtää kyselyn ja tutkimuksen tarkoituksen, on hän motivoituneempi vastaamaan (Pohjanpää, 2010).

Yksittäiseen kysymykseen vastaamiseen on todettu vaikuttavan esimerkiksi tiedonkeruutapa, mahdollisen haastattelijan koulutus ja asenne, kysymyksen aihe, kysymyksen rakenne ja vaikeus, sekä ohjeistus ja tutkittavan ominaispiirteet. Tiedonkeruutavan ollessa haastattelu, virheellisiä vastauksia jää aineistoon vähemmän ja tutkittavan kieltäytyessä tai epäröidessä haastattelija voi kysyä kysymyksen uudelleen. Hyvin koulutetut ja kokeneet haastattelijat toimivat tässä tapauksessa positiivisena vaikuttajana. Kysymysten asettelu ja selkeä ohjeistus edesauttavat vastauksen antamista. (Groves et al., 2002).

1.2 Yksikkövastauskato ja erävastauskato

Yksikkövastauskadon tapauksessa otokseen poimitulta tutkittavalta ei saada ollenkaan tietoja tutkimukseen. Yksikkövastauskadon tapauksessa tutkittavalta ei ole saatavilla kuin otoksessa olevat tiedot, joita on käytetty otoksen muodostamiseen, sekä mahdollisesti koko otokselle saatavissa olevat rekisteritiedot. Otosperäisiä tietoja väestötutkimuksissa ovat esimerkiksi sukupuoli ja ikä, rekisteriperäisiä esimerkiksi koulutus- ja ammattitiedot. Yksikkökadon tapauksessa saatavilla olevien tietojen perusteella pyritään arvioimaan, kuinka paljon yksikkökato vaikuttaa aineiston jakaumiin ja aineistosta tuotettaviin tuloksiin. Yksikkövastauskatoa pyritään korjaamaan usein painotusmenetelmillä, joissa jokaiselle lopulliseen aineistoon päätyneelle tutkittavalle lasketaan painokerroin, eli kerroin kuinka montaa tutkittavaa väestössä kyseinen tutkittava edustaa. Painokerrointa laskettaessa otoksen ulkopuolisista rekisteritiedoista on suurta hyötyä, sillä ne toimivat arvokkaana lisätietona väestön rakenteesta.

Erävastauskadolla tarkoitetaan, että tutkittava jättää vastaamatta yksittäiseen kysymykseen tai antaa epäkelvon arvon vastaukseksi. Aineistoa analysoitaessa erävastauskadon aiheuttamat aukot aiheuttavat ongelmia analyysimenetelmien käytössä. Useimmat analyysimenetelmät tarvitsevat toimiakseen täysin aukottoman ai-

neiston ja toiset taas saattavat erävastauskadon takia tuottaa virheellisiä arvoja tuloksina. Erävastauskatoa korjataan esimerkiksi erilaisilla imputointimenetelmillä, joissa puuttuvan arvon tilalle tuotetaan jollain tietyllä menetelmällä keinotekoinen muun aineiston avulla laskettu arvo.

Luku 2

Väestötutkimuksen lähtökohdat ja vastauskadon huomioiminen

2.1 Otantamenetelmät

Otantamenetelmällä tarkoitetaan sitä tapaa, jolla otos on muodostettu perusjoukosta, eli siitä väestöryhmästä, johon tulokset halutaan yleistää. Nykyään lähes jokainen tutkimus toteutetaan ottamalla otos kiinnostuksen kohteena olevasta perusjoukosta, sillä perusjoukkoa on usein lähes mahdotonta tutkia kokonaan, jollei kyse ole pelkästään rekisteriperusteisesta tutkimuksesta. Väestötutkimuksissa perusjoukkona on maan tai tietyn alueen väestö, joka voi olla rajattu esimerkiksi iän tai sukupuolen mukaan.

Useimmat väestötutkimuksissa käytetyt otantamenetelmät perustuvat todennäköisyysteoriaan ja satunnaisuuteen. Todennäköisyysteoriaan perustuvissa otantamenetelmissä jokaisella tutkittavalla on tietty poimintatodennäköisyys, jonka avulla otos voidaan palauttaa vastaamaan alkuperäistä perusjoukkoa. (Kuusela, 2009).

Satunnaisotannassa perusperiaatteena on, että jokaisella tutkittavalla on yhtä suuri mahdollisuus tulla poimituksi otokseen. Yksinkertaisimmillaan tämä tarkoittaa sitä, että koko perusjoukosta poimitaan satunnaisesti vastaajat (SRS-otanta). Monimutkaisemmissa satunnaisotantamenetelmissä poimintatodennäköisyys saattaa vaihdella tutkittavien välillä (Lehtonen ja Pahkinen, 2004).

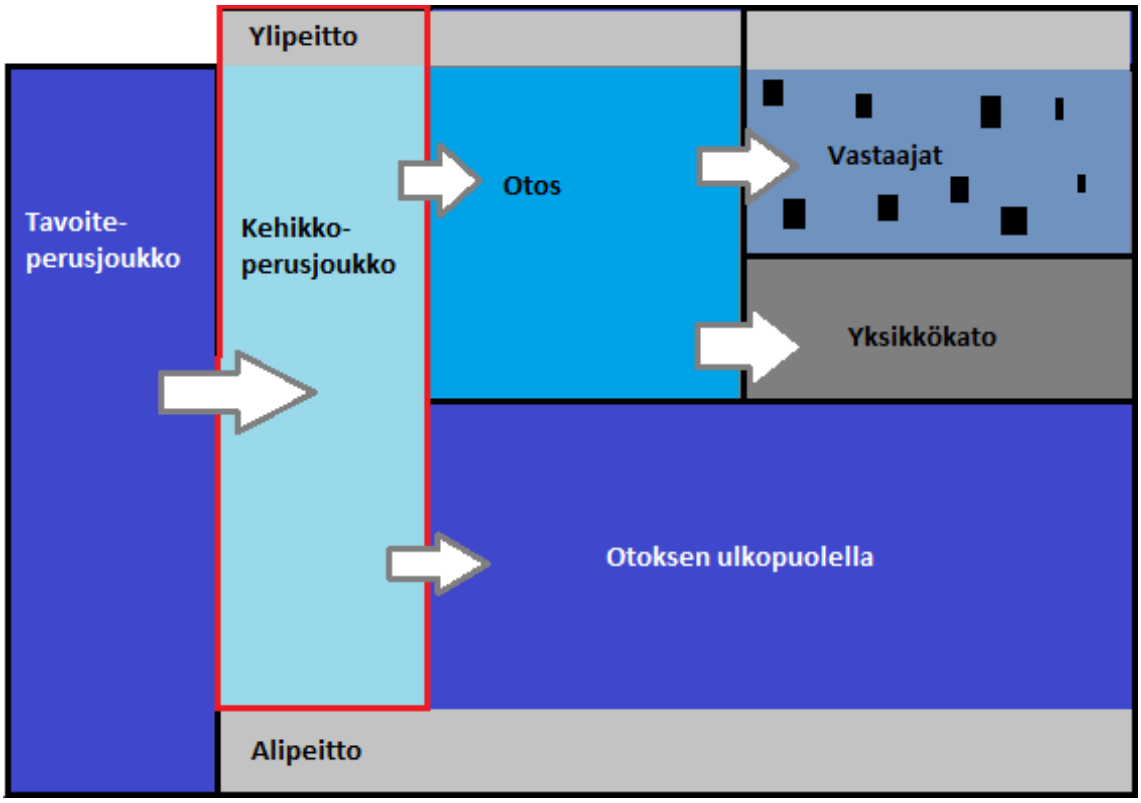
Esimerkkejä monimutkaisemmista satunnaisotantamenetelmistä ovat ositettu otanta, klusteriotanta ja systemaattinen otanta. Ositetussa otannassa (STR-otanta) muo-

dostetaan ensin ositteita joidenkin taustamuuttujien mukaan ja ositteiden sisällä poimintatodennäköisyys voi olla sama kaikille, mutta se usein vaihtelee ositteiden välillä. Poimintatodennäköisyys voidaan esimerkiksi määrittää ositteen koon mukaan (PPS-otanta) (Lehtonen ja Pahkinen, 2004). Klusteriotannassa poimitaan ensin klusteri eli suurempi ryhmä (esim. yritys tai koulu), jonka sisältä poimitaan satunnaisesti tutkittavat. Systemaattisessa otannassa (SYS-otanta) perusjoukko järjestetään jonkin muuttujan mukaiseen järjestykseen ja joka n:s tutkittava poimitaan otokseen, jossa n voi olla mikä tahansa perusjoukkoon sopiva positiivinen kokonaisluku. Näitä kaikkia otantamenetelmiä yhdistää se, että otantamenetelmiä varten on oltava listattu perusjoukko, josta tutkittavat voidaan poimia. Näin ollen jokaiselle otoksen tutkittavalle on laskettavissa poimintatodennäköisyys, joten otokset ovat suoraan perusjoukon suhteen edustavia (Kuusela, 2009). Tätä poimintatodennäköisyyttä voidaan hyödyntää myös mahdollisen vastauskadon hallinnassa.

Otantamenetelmän ei ole kuitenkaan pakko perustua poimintatodennäköisyyteen ja ennalta määriteltyyn perusjoukkoon. Itseohjautuvalla otannalla tarkoitetaan sitä, että tutkittavat valikoituvat tutkittaviksi omasta halustaan eikä tutkittavia poimita mistään ennalta määritellystä joukosta. Tällainen otanta on käytössä esimerkiksi monissa internetsivujen käyttäjäkyselyissä. Toinen esimerkki otannasta ilman poimintatodennäköisyyttä on kiintiöotanta. Kiintiöotannassa hankitaan tavalla tai toisella ennalta määritelty määrä tiettyyn joukkoon kuuluvia tutkittavia. Joukot ovat usein muodostettu perusjoukon pohjalta, mutta otos muodostuu vasta samalla kun tietoja kerätään tutkittavilta. Itseohjautuva otanta ja kiintiöotanta eivät kuitenkaan suoraan mahdollista todennäköisyysteorian käyttöä, joten näillä menetelmillä muodostettuja otoksia on usein vaikea saada edustamaan perusjoukkoa. (Kuusela, 2009).

Otokseen perustuvissa kyselytutkimuksissa otos pyritään muodostamaan niin, että se vastaisi mahdollisimman tarkasti otoksen perusjoukkoa, jolloin otosvirheen mahdollisuus on pieni. Otosta poimittaessa oletetaan, että otokseen kuuluvat edustavat otoksen ulkopuolelle jääneitä valittujen taustamuuttujien ja satunnaisvaihtelun puitteissa. Tavoiteperusjoukolla tarkoitetaan koko sitä tutkittavien joukkoa, jolle kyselytutkimuksen tulokset pyritään yleistämään. Taustamuuttujat valitaan tutkittavan ilmiön vaihtelun mukaan, eli taustamuuttujat valitaan oletuksena, että niillä

on vaikutusta tutkittavaan ilmiöön. Suomen väestöä koskevissa tutkimuksissa taustamuuttujia voisivat olla esimerkiksi ikä, sukupuoli ja koulutus.



Kuva 2.1: Kyselytutkimusaineiston muodostuminen perusjoukosta vastaajien aineistoon. Otoskehikko merkitty punaisella. Mukaillen Laaksonen et al. (2013).

Tavoiteperusjoukon määrittelyn jälkeen muodostetaan kehikkoperusjoukko, jossa ovat mukana käytettävissä olevat tutkittavat. Jos esimerkiksi tutkitaan Suomen väestöä, kehikkoperusjoukon muodostavat ne, joiden tiedot ovat väestörekisterissä. Alipeittoon kuuluvat henkilöt, jotka kuuluvat tavoiteperusjoukkoon mutta eivät kehikkoperusjoukkoon, sillä heitä ei ole mahdollista esimerkiksi tavoittaa valitulla tiedonkeruumenetelmällä. Alipeiton määrä riippuu tiedonkeruutavasta, esimerkiksi puhelinhaastatteluina toteutettavissa tutkimuksissa alipeiton muodostavat puhelimettomat henkilöt. Ylipeittoon kuuluvat ne henkilöt, jotka eivät kuulu tavoiteperusjoukkoon eivätkä kehikkoperusjoukkoon, mutta tulevat mukaan otokseen syystä tai toisesta. Esimerkkejä ylipeitosta ovat kuolleet, laitoksiin siirtyneet tai ulkomaille muuttaneet.

Varsinainen otos poimitaan kehikkoperusjoukosta. Otoksen koko riippuu usein tutkimuksen aikataulusta ja kustannuksista sekä otantamenetelmästä, mutta myös tutkimuksen tulosten halutusta kattavuudesta (Eurostat, 2008). Tarpeeksi suuri otos

varmistaa, että aineistossa on tarpeeksi ”voimaa” tarvittavien tulosten tuottamiseen. Painottamalla otoksen vastaajat väestöpainolla saadaan koko perusjoukkoa edustavia tuloksia. Väestöpaino on otostiedoista laskettava kerroin, joka kertoo kuinka montaa henkilöä väestössä kyseinen tutkittava edustaa (Laaksonen, 2009).

Vastauskato saattaa vaikeuttaa tutkimustulosten suoraa yleistettävyyttä kehikkoperusjoukkoon. Erityisesti tiettyihin vastaajaryhmiin painottunut kato, eli valikoitunut kato, voi aiheuttaa aineistosta saataviin tuloksiin haraa. Kuvassa 2.1 yksikkövastauskato näkyy yhtenäisenä laatikkona vastaajien alla ja erävastauskato yksittäisinä kysymysmerkkeinä vastaajien aineistossa. Mikäli yksikkökatoa tai erävastauskatoa on paljon ja se on valikoitunutta, ei saatu aineisto vastaa välttämättä alkuperäistä otosta eikä myöskään perusjoukkoa. Myös suuri valikoitumaton kato voi aiheuttaa ongelmia aineiston käytössä, sillä se vähentää aineiston ”voimaa” sitä käytettäessä.

2.2 Tiedonkeruumenetelmät

Olennainen osa kyselytutkimusta on tiedonkeruumenetelmä. Tiedonkeruumenetelmällä tarkoitetaan sitä tapaa, jolla kyselyn tiedot kerätään vastaajilta. Kyselytutkimuksissa yleisimpiä menetelmiä ovat postikyselyt ja haastattelut puhelimitse tai henkilökohtaisesti. Myös internet-pohjaiset kyselyt ovat yleistyneet muiden tiedonkeruumenetelmien ohella.

Tiedonkeruumenetelmällä on suora yhteys muodostuvaan katoon. Puhelinhaastatteluina toteutetuissa kyselyissä katoon jäävät automaattisesti puhelimattomat henkilöt, postikyselyissä taas vailla vakinaista osoitetta olevat. Internet-pohjaisissa kyselyissä internet-yhteyden puuttuminen tai heikot tietotekniset valmiudet saattavat aiheuttaa katoa erityisesti vanhemmissa ikäryhmissä. Useissa laajoissa tutkimushankkeissa hyödynnetäänkin useampaa tiedonkeruumenetelmää samanaikaisesti. Tiedonkeruuta suositellaankin muokattavan tutkittavan etukäteen tiedossa olevien ominaisuuksien mukaan esimerkiksi niin, että nuoremmille vastaajille tarjotaan internetissä vastaamista ja vanhemmille paperilomaketta (Groves et al., 2002).

Tiedonkeruumenetelmään liittyy myös nk. vastausharha. Haastatteluissa vastaajat ovat taipuvaisia vastaamaan ”sosiaalisesti hyväksyttävällä” tavalla, eli vastaajat

saattavat antaa vastauksen sen perusteella, jonka he ajattelevat olevan kyselyn kannalta oikea vaihtoehto. Itsenäisesti täytetyissä kyselyissä vastaukset saattavat olla totuudenmukaisempia. Haastattelumuotoisessa tiedonkeruussa yksittäiseen kysymykseen vastaamiseen on todettu vaikuttavan haastattelijan koulutus ja asenne. Hyvin koulutetut ja kokeneet haastattelijat toimivat tässä tapauksessa positiivisena vaikuttajana, sillä he voivat kannustaa vastaamaan ja ohjeistaa vastaajaa haastattelun edetessä (Groves et al., 2002). Itsenäisesti täytetyissä kyselyissä esiintyy enemmän puuttuvia vastauksia, sillä epämiellyttävät tai vaikeat kysymykset on helpompi jättää vastaamatta kuin haastattelutilanteessa. Itsenäisesti täytettävä kysely on myös helpompi jättää kokonaan täyttämättä kuin jättää vastaamatta esimerkiksi puhelimitse esitettyyn haastattelupyyntöön.

Kadon vähentämiseksi on etsitty erilaisia ratkaisuja tiedonkeruutavasta riippuen. Erityisesti tiedonkeruuajan pituuden ja tavoittelukertojen määrän on todettu olevan yhteydessä myös varsinaisiin tuloksiin, sillä aikaisten ja myöhäisten vastaajien välillä on monissa tutkimuksissa todettu olevan eroa (Voigt et al. (2003), Dunkelberg ja Day, (1973)). Tämä osoittaa sen, että jos tiedonkeruu olisi lopetettu aikaisemmin, kadon aiheuttama harha olisi suurempi (Peress, 2010). On myös osoitettu, että tavoittelukertojen vähentäminen saattaisi johtaa erilaisiin tutkimustuloksiin (Hawkins, 1975).

Tiedonkeruumenetelmän valintaan vaikuttaa olennaisesti siihen käytettävissä olevat resurssit ja tämä näkyy myös kadon minimoimisessa. Usein paljon resursseja vieviä tiedonkeruumenetelmiä, kuten haastatteluja, käytettäessä kadon minimoiminen on kallista, mutta sen tarve vähäisempää kuin kevyemmillä tiedonkeruumenetelmillä, kuten postikyselyillä. Taulukkoon 2.1 on kerätty yleisimpiä tiedonkeruumenetelmiä, niiden käytössä ilmeneviä syitä katoon ja kadon minimoimisen kustannus. Kahdessa ensimmäisessä, posti- ja internetkyselyissä ei päästä suoraan havainnoimaan kadon syytä, jolloin sitä ei voida varmasti luokitella. Haastatteluun perustuvissa tiedonkeruissa haastattelija havainnoi suoraan tutkittavaa, jolloin myös katoon liittyviä asioita on helppo huomioida ja päätellä erilaisia keinoja kadon vähentämiseksi. Kadon syyt vaihtelevat selkeästi tiedonkeruumenetelmien välillä, jolloin myös kadon määrän ja kadon aiheuttaman virheen yhteys riippuu tiedonkeruumenetelmästä (Groves et al., 2002).

Taulukko 2.1: Kyselytutkimusten mahdollisia tiedonkeruumenetelmiä, niissä mahdollisesti syntyvän kadon syitä ja kadon vähentämiseksi tarvittavien toimenpiteiden kustannukset

Tiedonkeruumenetelmä	**Kadon syy**	**Esimerkki**	**Kadon minimoimisen kustannus**
Postikysely	Ei tavoita	Väärä osoite	alhainen
	Kieltäytyminen	Posti luettu, mutta ei huomioitu	korkea
	Este	Lukutaidottomuus	korkea
Internetkysely	Ei tavoita	Linkki ei toimi	alhainen
	Kieltäytyminen	Linkki avattu, mutta ei huomioitu	korkea
	Este	Ei internet-yhteyttä	korkea
Puhelinhaastattelu	Ei tavoita	Ei vastausta puheluun	alhainen
	Kieltäytyminen	Katkaisee puhelun	korkea
	Este	Kuurous	korkea
Kasvokkain tehty haastattelu	Ei tavoita	Haastateltava ei saavu paikalle	alhainen
	Kieltäytyminen	Kieltäytyy suoraan	korkea
	Este	Kieliongelmat	korkea

2.3 Analyysimenetelmät

Tilastolliset analyysimenetelmät on kehitetty alun perin analysoimaan aineistoja, joissa ei ole puuttuvia arvoja. Kuten edellä on jo todettu, yhä useammin nykyään tutkimusaineistot ovat kuitenkin epätäydellisiä, eli joiltakin tutkimusyksiköiltä ei pystytä saamaan ollenkaan vastauksia tai jotkin yksittäiset vastaukset puuttuvat. Useissa tilasto-ohjelmistoissa puuttuvaa tietoa sisältävät yksilöt kuitenkin jätetään oletuksena analyyseissa käsittelemättä. Mikäli tutkittavalta ei saada ollenkaan vastauksia ja sellaiset tutkittavat jätetään huomiotta analyyseissa, aiheutuu aineistoon esimerkiksi valittujen taustamuuttujien suhteen harhaa. Esimerkkejä tästä ovat estimaatit onnellisuudesta ja köyhyydestä, sillä onnettomat ja pienituloiset ihmiset jättävät vastaamatta kyselyihin onnellisia ja hyvätuloisia ihmisiä useammin (Laaksonen, 2009). Tilastollisilla kadonhallintamenetelmillä näitä taustamuuttujien jakaumien eroja voidaan yrittää arvioida ja korjata niiden vaikutuksia tuloksissa.

Jos tutkimusyksiköltä on saatu vastaus vain osaan kysymyksistä, voi erilaisiin piste-estimaatteihin, kuten keskiarvoihin ja väestöosuuksiin, aiheutua harhaa tällaisen puuttuneisuuden takia. Harhat piste-estimaateissa ovat keränneet suurta huo-

miota 2000-luvulla, mutta jäävät melko usein raportoimatta tuloksia julkaistessa (Carpenter et al., 2014). Juuri piste-estimaattien harhoja pyritään kuitenkin useimmin korjaamaan, mutta huomiota tulisi kiinnittää myös estimaattien varianssiin ja tarkkuusestimaatteihin. Kadonhallintamenetelmät, jotka eivät huomioi puuttuneisuuden aiheuttamaa epävarmuutta johtavat usein varianssin aliarviointiin, liian pieniin luottamusväleihin ja liian suuriin testien p-arvoihin. (Groves et al., 2002).

Joissakin tapauksissa puuttuvat arvot voivat estää jonkin halutun tilastollisen analyysimenetelmän käytön kokonaan tai tehdä tuloksista niin epävarmoja, ettei menetelmää voida käyttää. Erityisesti tämä tilanne saattaa tulla eteen silloin, kun ohjelmisto automaattisesti jättää puuttuvia arvoja sisältävät tutkittavien rivit analyysin ulkopuolelle. Analyysimenetelmiä varten puuttuneisuuden valikoituneisuus tulee aina tutkia. Mikäli puuttuvia arvoja on vähän tai puuttuneisuus on täysin satunnaista (katso luku 3.2), voidaan se mahdollisesti jättää huomioimatta analyysimenetelmissä.

Luku 3

Vastauskadon analysoiminen ja mallintaminen

Kadon mallintaminen ja analysoiminen aloitetaan yleensä tarkastelemalla saadun aineiston määrää suhteessa alkuperäiseen otokseen ja sitä kautta kehikkoperusjoukkoon. Yksikkökatoa mallinnetaan muodostamalla osallistumisindikaattori, joka kertoo onko tutkittava osallistunut vai ei. Erävastauskatoa varten samankaltainen vastausindikaattori muodostetaan yksittäiselle kysymykselle, joka kertoo onko tutkittava vastannut kysymykseen vai ei. Tarkastelemalla erävastauskatoindikaattorien summia kysymyskohtaiseti, voidaan tarkastella puuttuneisuuden muodostumista kuviona. Puuttuneisuuden kuvio kertoo, missä muuttujissa puuttuneisuutta on ja puuttuneisuuden mekanismi taas sen, millä tavalla puuttuneisuus on valikoitunutta aineistossa.

3.1 Mallintaminen

Yksikkökatoa voidaan mallintaa muodostamalla osallistumisindikaattori alkuperäiselle otosjoukolle. Otokseen kuuluvista yksiköistä on usein rajallisesti tietoa tarjolla, joten mallintaminen perustuu usein vain otoksen poiminnassa käytettyihin tietoihin. Yksikkökadon mallintamisessa voidaan myös käyttää apuna rekisteritietoja, mikäli niitä on saatavilla. Jokaiselle otoksen tutkittavalle i muodostetaan osallistumisindikaattori seuraavasti:

$$r_i = \begin{cases} 1, & \text{jos } i \text{ on vastannut} \\ 0, & \text{jos } i \text{ ei ole vastannut} \end{cases}$$

Osallistumisindikaattoria voidaan mallintaa asettamalla selittäviksi tekijöiksi otoksesta saatavat tiedot, kuten sukupuoli, ikä ja asuinalue. Nämä selittävät tekijät ovat otosaineistosta saatavia otosmuuttujia (*auxiliary variables*), jotka ovat saatavilla kaikille otokseen kuuluville tutkittaville. Otosmuuttujat muodostavat vektorin $x_i = (x_{1i}, x_{2i}, ..., x_{pi}), p = 1, ..., P$, jossa P viittaa muuttujien määrään. Otosmuuttujavektorit muodostavat matriisin X, jossa ovat siis koko otokselle täysin havaitut muuttujat.

Tutkimuksessa kysytyt kysymykset eli aineiston tulosmuuttujat muodostavat vektorin $y_i = (y_{1i}, y_{2i}, ..., y_k), k = 1, ..., K$, jossa i viittaa tutkimusyksikköön ja K muuttujiin. Vektorit y_i muodostavat matriisin Y, johon kuuluvat siis kaikki tutkimuksen kysymyksistä muodostetut muuttujat. Matriisin Y muuttujissa saattaa siis olla puuttuvia arvoja.

Kun otostiedoista saatavat muuttujat X ja tutkimuksessa käytetyt muuttujat Y erotetaan toisistaan, nähdään selkeästi ero täysin havaittujen muuttujien ja mahdollista puuttuvuutta sisältävien muuttujien välillä kuvan 3.1 tilanteessa (ii).

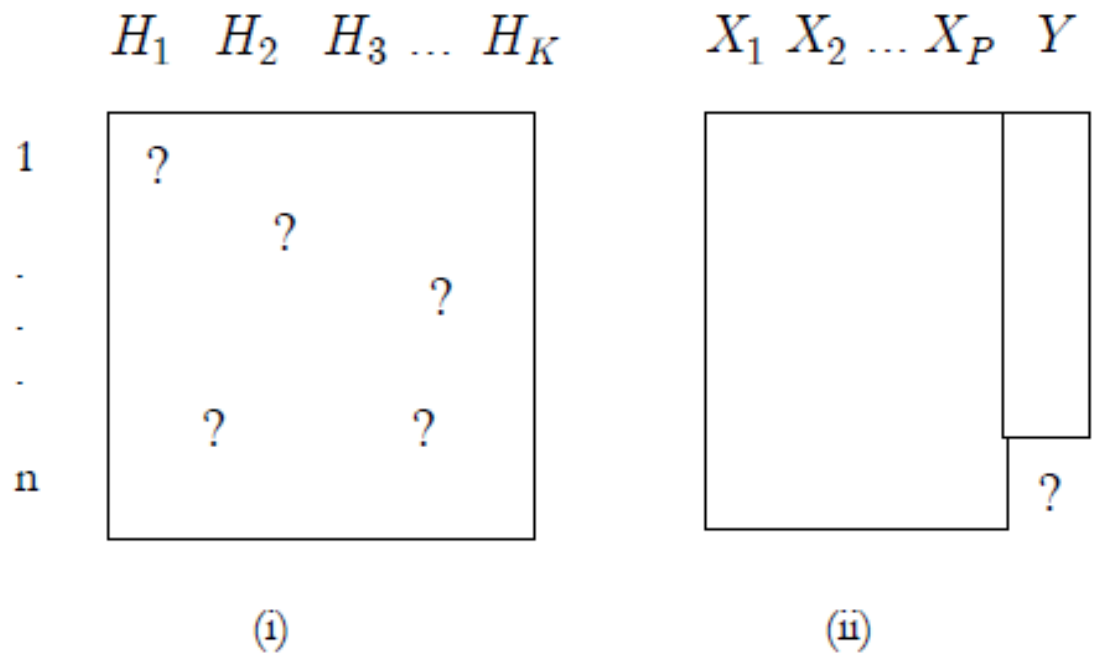
Erävastauskatoa varten yksittäiselle muuttujalle y_i voidaan muodostaa vastausindikaattori. Puuttuneisuutta tarkastellaan ja mallinnetaan usein luomalla jokaiselle tarkastettavalle muuttujalle y_i binäärinen vastausindikaattori $r_{1i}, i = 1, ..., Q$. Vastausindikaattorille voidaan muodostaa matriisi R, jossa

$$r_{ik} = \begin{cases} 1, & \text{jos } h_{ik} \text{ on havaittu arvo} \\ 0, & \text{jos } h_{ik} \text{ ei ole havaittu arvo} \end{cases}$$

Tällöin $h_{ik} = (x_i, y_i)$ ja H tarkoittaa kokonaista datamatriisia, jossa on $H_1 ... H_k$ muuttujaa eli saraketta. Tällöin erävastauskato voidaan havaita kuvan 3.1 matriisien H sisältämistä kysymysmerkeistä.

3.2 Puuttuneisuuden mekanismi ja kuvio

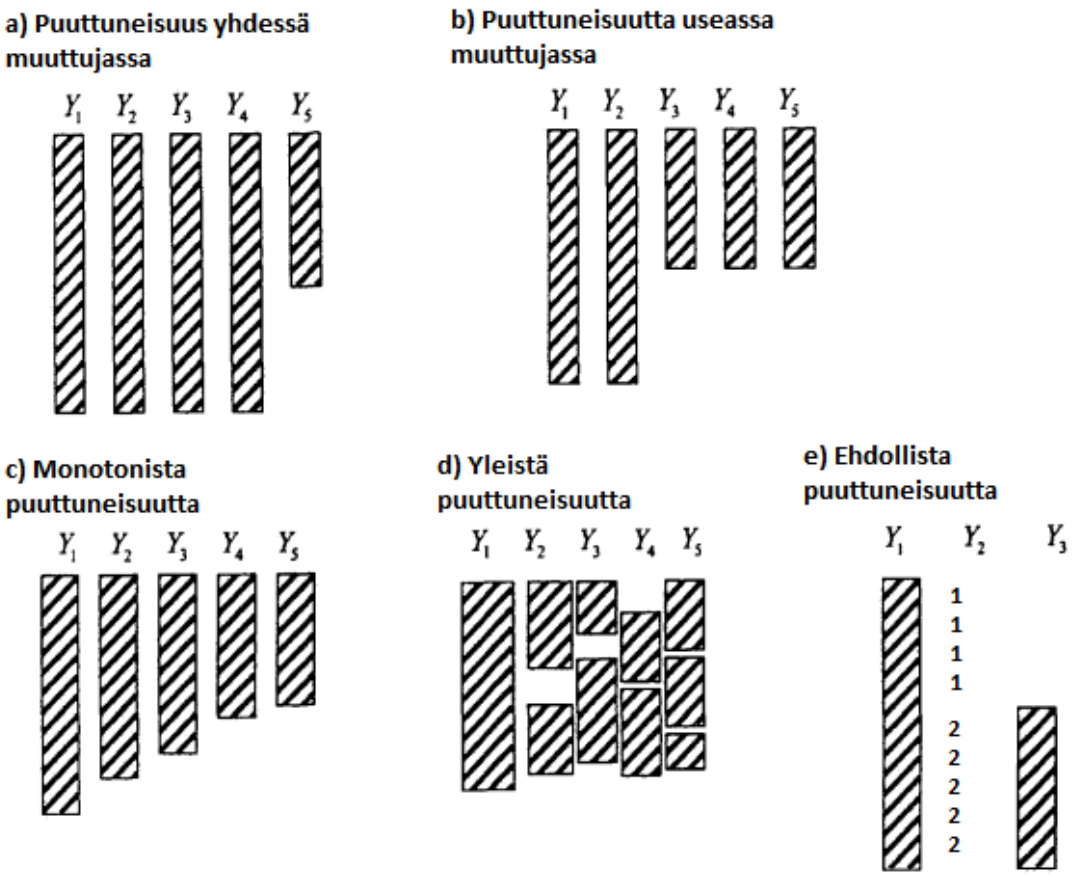
Puuttuneisuuden kuvio kertoo, mikä osa aineistosta on puuttuvaa ja puuttuneisuuden mekanismksi kutsutaan sitä tapaa, jolla tietoa puuttuu aineistosta. Tämä tapa



Kuva 3.1: (i) Monen muuttujan tilanne, jossa puuttuvuus saattaa olla satunnaista (ii) Monen muuttujan tilanne, jossa eroteltuna täysin havaitut muuttujat X ja puuttuvuutta sisältävät muuttujat Y (Durrant, 2005).

voi olla täysin satunnainen (*Missing Completely At Random*, MCAR), riippua havaituista arvoista (*Missing At Random*, MAR) tai olla täysin epäsatunnaista (*Missing Not At Random*, MNAR) (Rubin, 1976). Nämä Rubinin (1976) esittelemät termit ovat kuitenkin hieman ongelmallisia niin englanniksi kuin suomeksikin, sillä niillä on tarkoitus kuvata kadon valikoituvuutta, eli sitä miten kato on jakautunut erilaisissa väestöryhmissä, eikä varsinaista satunnaisuutta. Siksi puuttuneisuuden mekanismia on helpompi tarkastella riippuvuuksien kautta. Laaksosen (2010) mukaan MCAR tarkoittaa, että puuttuneisuus ei riipu mistään aineistossa olevista muista muuttujista, MAR taas tarkoittaa, että se saattaa riippua esimerkiksi otannassa käytetyistä muuttujista tai rekisteriperäisistä apumuuttujista.

Mallintamalla puuttuneisuutta luvussa 3.1 mainituilla tavoilla, voidaan muodostaa malli puuttuneisuudelle. Puuttuneisuuden oletettu mekanismi määrittää pitkälle käytettävän kadonhallintamenetelmän ja muodostettu malli toimii arvokkaana lisätietona hallintamenetelmässä. Puuttuneisuuden mekanismia lähdetään määrittämään määrittelemällä funktio $f(R|H)$, jossa f on tuntematon todennäköisyysjakauma, R puuttuneisuutta kuvaava matriisi ja H koko aineistoa kuvaava matriisi. Erilaisia puuttuneisuuden kaavoja on kuvattu kuvassa 3.2.



Kuva 3.2: Erilaisia mahdollisia puuttuneisuuden kuvioita aineistossa, jossa Y_i viittaa yhteen muuttujaan ja rivit havaintoihin. Mukaillen Little ja Rubin (2002).

Puuttuneisuuden ollessa täysin ei-valikoitunutta (MCAR) puuttuvien arvojen todennäköisyys ei riipu mistään otostiedoston tai tutkimusaineiston toisesta muuttujasta eikä kyseisten analyysimuuttujien arvoista. Yksittäisen muuttujan y_i tapauksessa tämä tarkoittaa sitä, ettei vastaustodennäköisyys ole riippuva muuttujasta y_i itsestään eikä mistään muustakaan muuttujasta x_{pi} (esimerkiksi kuva 3.1 (ii) ja kuva 3.2 a)) (Rubin, 2004). Oletus puuttuneisuuden satunnaisuudesta on erittäin vahva ja valitettavan usein paikkansapitämätön väestötutkimuksissa (Durrant, 2005). Puuttuneisuuden ollessa täysin satunnaista sitä ei välttämättä tarvitse huomioida aineistosta tehtävissä analyyseissa ja katoa voidaan paikata hyvin yksinkertaisilla menetelmillä.

Jos puuttuneisuus on riippuvaista vain havaituista arvoista, tällöin sanotaan puuttuvuuden olevan satunnaista ehdollisesti (MAR). Tällöin puuttuvan arvon todennäköisyys ei riipu arvosta itsestään kun jokin toinen muuttuja on vakioitu. Esimerkki tällaisesti tilanteesta voisi olla, että jokin kaikilta havaittu muuttuja assosioituu puuttuvuuteen (Howell, 2012).

Oletus, että puuttuneisuus on vain osittain ehdollista, on heikompi oletus kuin MCAR, mutta on vaikeampi todeta kuin MNAR, sillä aineiston muuttujat y_i ovat katoon jääneillä havaitsemattomia arvoja (Durrant, 2005). Käytännössä MAR tarkoittaa sitä, että puuttuneisuus on valikoitunutta ja se voidaan mallintaa havaitusta aineistosta. Puuttuneisuus voi riippua myös siitä, kuinka pitkällä kyselyssä kyseistä muuttujaa kuvaava kysymys on ollut, ts. puuttuneisuus riippuu monotonisesti sitä edeltävistä muuttujista (esimerkiksi mahdolliset hyppykäskyt, kuvan 3.2 tapaus c)).

Mikäli todetaan, että puuttuneisuus riippuu sekä havaituista että havaitsemattomista arvoista, puuttuneisuus on valikoitunutta ja ehdollista (MNAR). Tällöin ne muuttujat, joiden suhteen puuttuneisuus on ehdollista, on pakko huomioida aineiston analyysissa mahdollisimman luotettavien tuloksien saamiseksi. Näiden muuttujien avulla voidaan muodostaa vastaustodennäköisyyttä edustava malli, jota käytetään apuna kadon hallinnassa. Mahdollisesti ehdollista puuttuneisuutta on kuvan 3.2 tapauksissa a), b), c) ja e).

Luku 4

Alueellinen terveys- ja hyvinvointitutkimus (ATH)

Alueellinen terveys- ja hyvinvointitutkimus (ATH) on Terveyden ja hyvinvoinnin laitoksen, kuntien ja organisaatioiden yhteistyössä toteuttama väestötutkimus, jossa tavoitteena on tarjota kunnille ja muille alueille tarvittavaa tietoa asukkaidensa terveyden ja hyvinvoinnin seuraamiseen. Tutkimus toteutetaan postikyselynä (sisältäen myös mahdollisuuden vastaamiseen internetissä) ja tarkoituksena on kerätä kyselyllä tietoa, jota ei ole saatavissa rekistereistä. (Kaikkonen et al., 2010a).

Tutkimuslomake sisältää kysymyksiä elinoloista, koetusta hyvinvoinnista, terveydestä, toiminta- ja työkyvystä, elintavoista, erilaisista terveyteen ja hyvinvointiin vaikuttavista riskitekijöistä ja palveluiden saannista sekä laadusta. Kysely on suunnattu yli 20-vuotiaalle väestölle kolmessa eri ikäryhmässä, 20–54-vuotiaat, 55–74-vuotiaat sekä yli 75-vuotiaat ja se on saatavilla neljällä eri kielellä: suomeksi, ruotsiksi, englanniksi ja venäjäksi. Kyselyyn voi vastata sekä postitetulla paperilomakkeella että internetissä lomakkeella olevien tunnusten avulla.

ATH- tutkimuksen pilottivaiheessa vuonna 2010 kerättiin yhteensä 31 000 henkilön otos Turun, Kainuun ja Pohjois-Pohjanmaan alueilta sisältäen myös 5000 henkilön koko Suomen väestöä edustavan otoksen. Koko aineiston tasolla vastausprosentti oli 49 % vaihdellen ikäryhmittäin (20–54-vuotiaat: 40 %, 55–74-vuotiaat: 63 % ja yli 75-vuotiaat: 59 %). Aineistosta julkaistiin interaktiivisessa palvelussa tuloksia alueittain, ikäryhmittäin, sukupuolittain ja koulutusryhmittäin. (Kaikkonen et al., 2010c).

Aineiston otostiedostoon sisältyy poimintahetken tiedot vastaajan iästä, sukupuolesta, asuinpaikasta, äidinkielestä ja siviilisäädystä. Nämä tiedot löytyvät siis jokaiselle otokseen poimitulle tutkittavalle, joten näitä tietoja voidaan käyttää mallinnettaessa yksikkövastauskatoa. Kontaktikertoja per tutkittava oli yhteensä kolme. Otostiedostoa ja palautuvia lomakkeita hallinnoitiin tietokannan avulla niin, että jokaiselle palautuneelle vastaukselle ja katokoodi-ilmoitukselle on kirjattu myös ajankohta tutkittavakohtaisesti. Joidenkin katokoodien tapauksessa (esim. kuollut tai sairas) saadaan tietoa myös alipeitosta (katso kuva 2.1). Paperilomakkeiden osalta tietokantaan kirjautui myös tieto siitä, monenteenko kontaktikertaan tutkittava on reagoinut. Näiden kirjaustietojen avulla saadaan tietoa katoon jääneistä tutkittavista sekä voidaan tutkia, löytyykö aikaisten ja myöhäisten vastaajien välillä eroa tutkimustuloksissa. Myös tätä tietoa vastausaallosta käytetään mallintaessa erävastauskatoa.

Taulukko 4.1: Tunnettu kato ja lomakkeiden täyttöaste, koko aineisto 2010

Syy	**Lukumäärä**
Toimitettu tyhjä lomake	468
Sisältää lisäpapereita	6
Posti palauttanut	52
Kuollut	39
Osoite tuntematon	16
Kieltäytynyt	189
Ei pystynyt sairautensa vuoksi täyttämään	183
Ei halua osallistua jatkossa	23
Muu syy	72

Mallinnettaessa katoa ikää käsitellään kymmenvuotisikäluokittain mahdollisen epälineaarisuuden takia ja muut otosmuuttujat ovat luokitusasteikollisia. Siviilisääty on muutettu viisiluokkaiseksi eli käsitellään erikseen naimattomat, eronneet, lesket ja avioliitossa tai rekisteröidyssä parisuhteessa elävät liian pienten luokkien välttämiseksi. Kaikille ei ole kuitenkaan siviilisäätytietoa saatavilla, joten nekin käsitellään omana luokkanaan.

Tämän lisäksi otoksen ulkopuolisena lisätietona mallinnuksessa käytetään Tilastokeskuksen perusasteen jälkeistä koulutusta ja ammattitietoa sekä Kansaneläkelaitoksen lääkkeiden erityiskorvausoikeusrekisteriä. Koulutus on muutettu kolmiluokkaiseksi ja ammattitiedot ISCO-08-luokituksen 1-numerotason mukaan kym-

menluokkaiseksi. Myös ammattitiedoissa oli puuttuvuutta, joten ne, joille tietoa ei ollut saatavilla, käsitellään omana luokkanaan. Erityiskorvauksia lääkkeistä saaneet on jaettu kolmeen ryhmään: diabeteslääkkeet, psyykelääkkeet ja muut lääkkeet. Psyykelääkkeisiin luetaan vaikeaan psykoosiin ja muihin vaikeisiin mielenterveyden häiriöihin lukeutuvat lääkkeet ja muihin lääkkeisiin seuraaviin sairauksiin liittyvät lääkitykset: Myasthenia gravis, MS-tauti, Parkinsonin tauti, epilepsia ja älylliset kehitysvammat (Social Insurance Institute, 2014). Lääkkeiden erityiskorvattavuusoikeudesta ylipäätään tehtiin myös oma muuttuja, joka kertoo onko vastaajalla erityiskorvausoikeutta sisältäen myös muut sairaudet kuin edellä mainitut.

4.1 Yksikkövastauskadon analyysi

Yksikkökatoa, eli tutkittavien osallistumista, voidaan mallintaa kaikilla otostiedostoista ja rekistereistä saatavilla muuttujilla. Aloitetaan yksikkökadon tarkasteleminen otos- ja rekisterimuuttujien taustajakaumista. Osallistuneet ja katoon jääneet eroavat toisistaan erityisesti sukupuolen, siviilisäädyn, ikäryhmän ja koulutuksen osalta (taulukko 4.2). Lisäksi myös tutkimusalue, ammatti ja psyykelääkkeiden erityiskorvausoikeus ovat merkittäviä selittäjiä yksikkökatoa tutkiessa (taulukko 4.3). Kuvassa 4.1 on vielä tarkemmin kuvattuna ikäjakaumat otoksessa ja aineistossa viisivuotisikäryhmittäin. Kuvastakin huomataan, etteivät jakaumat vastaa toisiaan, vaan vanhemmat ikäluokat ovat yliedustettuina vastanneiden aineistossa.

Muodostetaan osallistumiselle indikaattori,jota mallinnetaan logistisella regressiomallilla. Mallissa koko otokselle saatavissa olevat muuttujat toimivat selittäjinä. Tässä mallissa jokainen näistä muuttujista on yksittäin iän ja sukupuolen kanssa huomataan, että kaikki edellä mainitut muuttujat näyttäisivät selittävän yksikkökatoa jokin verran (taulukko 4.3). Merkittävimmät erot yksikkökadossa löytyvät eri ikäryhmien ja koulutusryhmien välillä. Tämä merkitsee sitä, ettei yksikkökato voi olla täysin satunnaista (MCAR) ja sen huomiotta jättäminen tuottaa analyyseissa virheellisiä estimaatteja. Sen sijaan kieliryhmien välillä ei juuri löydy eroja, eikä diabeteslääkkeiden erityiskorvausoikeustieto tai tieto erityislääkekorvausoikeudesta ylipäätään selitä eroja merkitsevästi.

Tarkasteltaessa taustamuuttujittain tutkimukseen osallistumisen vetosuhteita (OR,

Taulukko 4.2: ATH 2010-tutkimuksen otoksen, vastaajien aineiston ja painotetun aineiston jakaumat taustamuuttujittain.

Tausta-muuttuja	**Luokka**	**Otos (%)**	**Aineis-to (%)**	**Painotettu aineisto (%)**
sukupuoli	Mies	48.8	42.9	47.8
	Nainen	51.2	57.1	52.2
siviilisääty	avioliitossa	47.9	54.2	48.0
	ei tietoa	0.4	0.2	0.4
	eronnut	11.6	11.7	11.4
	leski	11.1	12.1	12.4
	naimaton	29.0	21.8	27.8
ikäryhmä	20-54	52.6	42.7	50.3
	55-74	27.3	34.0	26.4
	75+	20.2	23.3	23.3
alue	Koko Suomi	16.1	15.5	15.5
	Kainuu	29.0	31.6	31.6
	Pohjois-Pohjanmaa	25.8	24.7	24.7
	Turku	29.0	28.2	28.3
kieli	suomi	96.8	97.0	96.8
	englanti	0.2	0.1	0.1
	ruotsi	2.0	2.0	2.0
	venäjä	1.1	0.9	1.1
koulutus	Perusaste	33.7	32.1	34.8
	Keskiaste	61.8	62.1	60.7
	Korkea-aste	4.5	5.8	4.6
ammatti	Johtajat ja ylimmät virkamiehet	2.1	2.0	1.9
	Asiantuntijat	8.6	9.1	8.4
	Ei tietoa	51.1	54.5	53.0
	Erityisasiantuntijat	8.1	8.8	7.8
	Maanviljelijät, metsätyöntekijät ym.	1.9	1.7	1.9
	Muut työntekijät	5.6	4.6	5.4
	Palvelu-, myynti- ja hoitotyöntekijät	8.2	7.9	8.0
	Prosessi- ja kuljetustyöntekijät	4.8	3.7	4.5
	Rakennus-, korjaus- ja valmistustyöntekijät	5.9	4.0	5.5
	Sotilaat	0.2	0.2	0.2
	Toimisto- ja asiakaspalvelutyöntekijät	3.5	3.5	3.4
lääkkeiden	Diabetes	6.9	7.7	7.2
erityis-	Psyykelääkkeet	2.3	1.7	2.3
korvaus-	Muut lääkkeet	2.6	2.3	2.6
oikeus	Erityiskorvattavat yhteensä	37.6	42.2	40.0

Taulukko 4.3: Yksittäisten taustamuuttujien ikä-sukupuolivakioidut Waldin testisuureet ja niiden p-arvot tutkimukseen osallistumiselle

Effect	**DF**	**Wald** χ^2	**P** $>\chi^2$
Kieliryhmä	3	13.1	0.0044
Sukupuoli	1	388.5	<.0001
Siviilisääty	4	816.7	<.0001
Ikäryhmä	2	1334.5	<.0001
Alue	3	112.6	<.0001
Koulutus	2	110.8	<.0001
Ammatti	10	461.5	<.0001
Erityiskorvausoikeus: psyykelääkkeet	1	46.3	<.0001
Erityiskorvausoikeus: muut lääkkeet	1	12.0	0.0005
Erityiskorvausoikeus: kaikki lääkkeet	1	0.0	0.842
Erityiskorvausoikeus: diabeteslääkkeet	1	1.8	0.1839

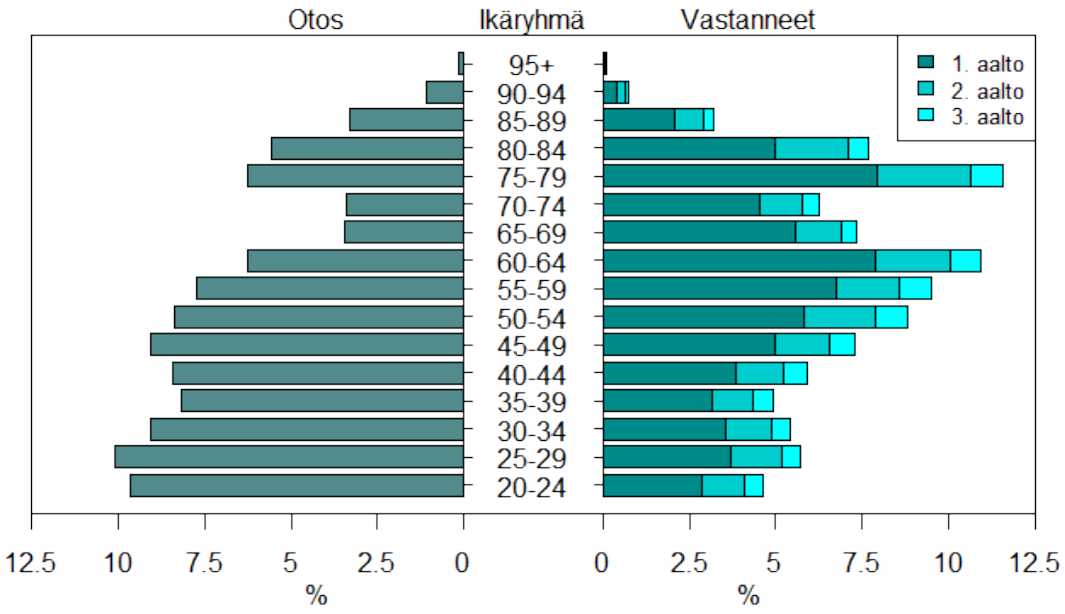
taulukko 4.4), huomataan että ainakin sukupuolten, ikäryhmien ja siviilisäätyluokkien välillä on selkeitä. Ammattiasemalla ei vetosuhteiden mukaan ole kovinkaan merkittävää vaikutusta, kun vertailuryhmänä käytetään johtavia ja ylimpiä virkamiehiä. Korkeasti koulutetut ovat kuitenkin vastanneet todennäköisemmin kuin perus- ja keskiasteen koulutuksen saaneet.

Muodostetaan merkittävistä muuttujista vastaustaipumusmalli (*propensity model*). Tutkitaan ensin näiden muuttujien toimintaa yhdessä mallissa vastauskatoa mallinnettaessa ja niiden ensimmäisen asteen interaktioita. Taulukossa 6.2 on näiden yhdistelmistä muodostettujen ikä-sukupuoli-vakioitujen mallien BIC- ja AIC-kertoimet, devianssi ja vapausasteet. *Main effects*-termi sisältää kaikki edellä mainitut muuttujat. BIC-kerroin kertoo mallin sopivuudesta aineistoon suhteessa selittävien muuttujien ja niiden luokkien määrään. Mitä pienempi BIC-kerroin, sen parempi ja tehokkaampi malli on. AIC-kerroin on samankaltainen, mutta siinä muuttujien määrä vaikuttaa hieman vähemmän kertoimen arvoon. Devianssi kertoo mallin sopivuudesta; mikäli malli sopisi täydellisesti ainestoon, devianssi olisi nolla. Näiden tunnuslukujen perusteella paras malli yksikkökadolle saatavissa olevilla muuttujilla olisi sellainen, jossa on päävaikutusten lisäksi myös sukupuolen ja koulutuksen yhdysvaikutus huomioituna.

Vastaustaipumusmallin avulla voidaan laskea IPW-menetelmällä (*Inverse Probability Weighting*) analyysipainot vastaajille. Painotuksen onnistuneisuutta voi tut-

Taulukko 4.4: ATH-tutkimukseen osallistumisen vetosuhteet taustamuuttujittain, niiden luottamusvälit (CL), Waldin testisuureet ja niiden p-arvot

	Ryhmä	**OR**	**95% Wald CL**		**Wald χ^2**	**P $> \chi^2$**
Suku-	Mies	1.0				
puoli	Nainen	1.5	1.4	1.5	267.0	<.0001
Siviili-	avioliitossa	1.0				
sääty	ei tietoa	0.4	0.3	0.6	9.1	0.00
	eronnut	0.7	0.7	0.8	10.2	0.00
	leski	0.5	0.5	0.6	13.1	0.00
	naimaton	0.7	0.6	0.7	2.6	0.10
Ikäryhmä	20-54	1.0				
	55-74	2.4	2.3	2.6	333.0	<.0001
	75+	2.2	2.1	2.3	142.5	<.0001
Alue	Koko Suomi	1.0				
	Kainuu	1.2	1.1	1.3	39.3	<.0001
	Pohjois-Pohjanmaa	1.0	1.0	1.1	2.0	0.16
	Turku	1.0	0.9	1.1	4.3	0.04
Kieli	suomi	1.0				
	englanti	0.9	0.5	1.7	0.0	0.92
	ruotsi	1.0	0.8	1.2	0.4	0.51
	venäjä	0.8	0.6	1.0	2.3	0.13
Koulutus	Perusaste	1.0				
	Keskiaste	1.9	1.8	2.0	35.4	<.0001
	Korkea-aste	2.3	2.0	2.5	74.6	<.0001
Ammatti	Johtajat ja ylimmät virkamiehet	1.0				
	Asiantuntijat	1.1	0.9	1.3	16.3	<.0001
	Ei tietoa	0.8	0.7	1.0	20.0	<.0001
	Erityisasiantuntijat	1.2	1.0	1.4	28.7	<.0001
	Maanviljelijät, metsätyöntekijät ym.	0.8	0.6	1.0	3.6	0.06
	Muut työntekijät	0.7	0.6	0.9	20.2	<.0001
	Palvelu-, myynti- ja hoitotyöntekijät	0.9	0.8	1.1	0.1	0.77
	Prosessi- ja kuljetustyöntekijät	0.8	0.6	0.9	13.9	0.00
	Rakennus-, korjaus- ja valmistustyöntekijät	0.7	0.5	0.8	46.2	<.0001
	Sotilaat	1.8	1.1	3.0	8.3	0.00
	Toimisto- ja asiakaspalvelutyöntekijät	0.9	0.8	1.1	0.0	0.96
Lääkkeiden	diabetes 1 vs 0	0.9	0.9	1.0	1.8	0.18
erityis-	psyykelääkkeet 1 vs 0	0.6	0.5	0.7	46.3	<.0001
korvaus-	muut lääkkeet 1 vs 0	0.8	0.7	0.9	12.0	0.00
oikeus	erityiskorvattavat yhteensä 1 vs 0	1.0	0.9	1.1	0.0	0.84

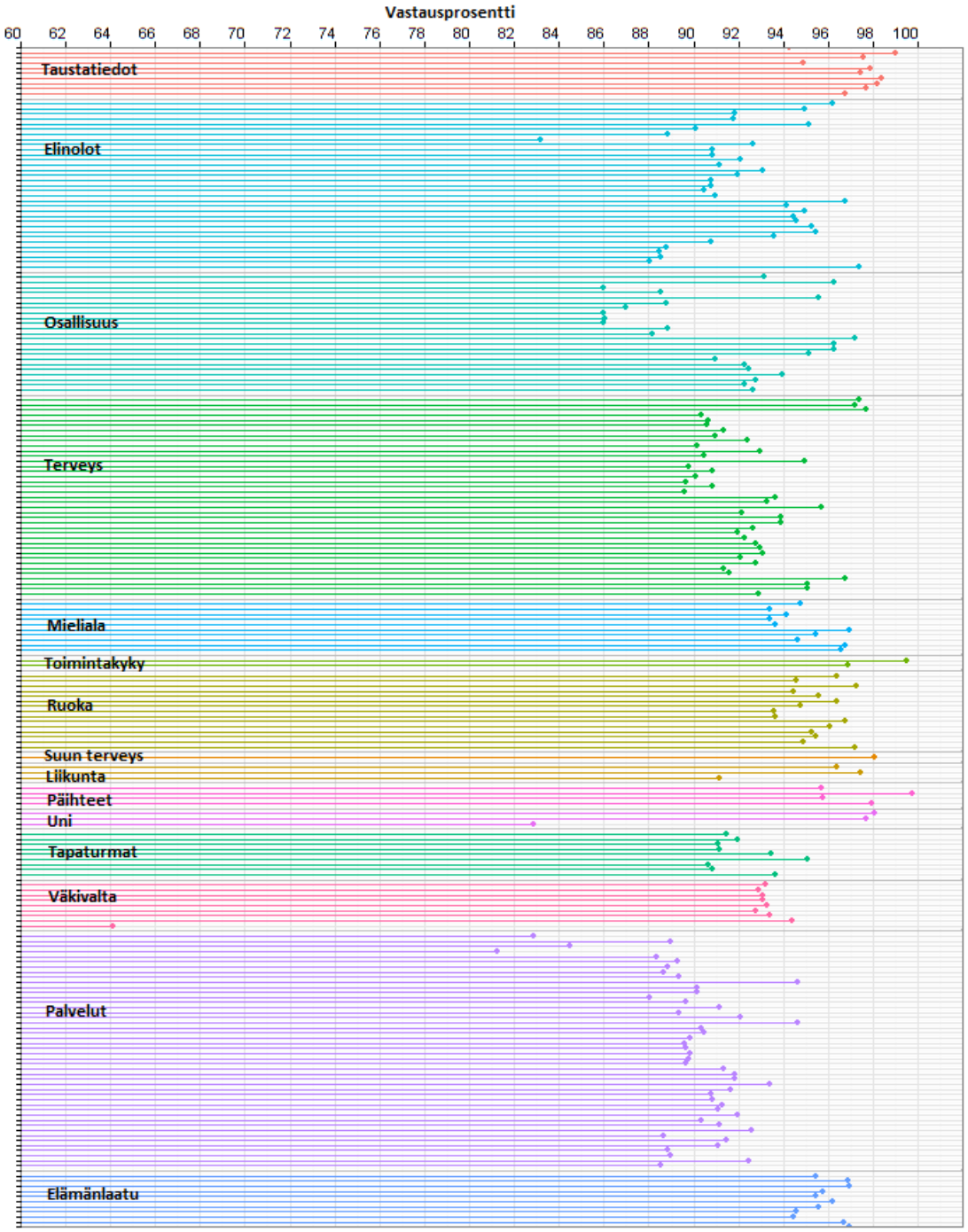


Kuva 4.1: Ikäjakauma otoksessa ja vastausaalloittain vastanneiden aineistossa. Otoksessa yli 75-vuotialla on kaksinkertainen poimintatodennäköisyys.

kia esimerkiksi tarkastelemalla taustamuuttujien jakaumia aineistossa painotuksen jälkeen. Nämä on kerätty taulukkoon 4.2 viimeiseksi sarakkeeksi. Tavoitteena on, että painotetun aineiston jakauma olisi mahdollisimman lähellä alkuperäisen otoksen jakaumaa, jonka puolestaan katsotaan edustavan koko väestön taustamuuttujien jakaumaa ainakin iän ja sukupuolen osalta. Tarkempi kuvaus IPW-menetelmästä on luvussa 5.3.

4.2 Erävastauskadon analyysi

Yksikkökadon lisäksi tarkastellaan myös erävastauskatoa, eli sitä, kuinka moni osallistuneista on vastannut kuhunkin kysymykseen. Kuvassa 4.2 on koottu kysymysten vastausprosentit aihealueittain niiden esiintymisjärjestyksessä. Kuvasta nähdään, että taustatietokysymysten vastausprosentit ovat lähellä sataa, mutta jo seuraavissa elinoloja koskevissa kysymyksissä on paljon puuttuvaa tietoa. Eniten puuttuvia vastauksia sisältävät osiot ovat osallisuus, terveys ja palvelut. Esimerkiksi palvelut- osio sisältää useita pitkiä kysymyspattereita, joihin vastaaja ei jaksa enää välttämättä keskittyä (Kaikkonen et al., 2010b). Lomakkeen viimeisiin kysymyksiin elämänlaa-



Kuva 4.2: ATH-kysymyslomakkeen kysymysten vastausosuudet vastanneiden aineistosta kysymysosioittain esiintymisjärjestyksessä. Yksi vaakaviiva vastaa yhtä osion kysymystä.

dusta on vastattu taas paremmin.

Taulukko 4.5: ATH-kysymyslomakkeen 20 kysymystä, jossa eniten ja vähiten puuttuvia arvoja ja puuttuvien arvojen osuudet tutkimukseen osallistuneiden aineistossa.

Kysymys	**puuttuvia(%)**
Nukutteko yleensä myös päivällä?	17.2
Esiintyykö asuntonne lähiympäristössä seuraavia tekijöitä, ja missä määrin ne haittaavat Teitä?	16.9
Kuinka useasti olette internetin välityksellä yhteydessä ystäviinne ja sukulaisiinne?	14.1
Kuinka usein olette osallistunut kulttuuriyhdistyksen tai -järjestön toimintaan 12 viime kk aikana?	14.1
Kuinka useasti olette kirjeitse yhteydessä ystäviinne ja sukulaisiinne?	11.5
Käytetättekö internetiä tietojen hakemiseen?	11.3
Saatteko taloudellista tukea ystäviltä tai naapureilta?	11.2
Oletteko käynyt 12 viime kuukauden aikana terveyskeskuksen hammaslääkärillä?	11.1
Oletteko mielestänne saanut riittävästi tietoa kuntanne kulttuuripalveluista 12 viime kk aikana?	10.8
Oletteko mielestänne saanut riittävästi talous- ja velkaneuvonnan palveluita 12 viime kk aikana?	10.7
Onko Teillä ollut lääkärin toteamaa tai hoitamaa aivohalvausta 12 viime kk aikana?	10.5
Onko Teillä ollut lääkärin toteamaa tai hoitamaa masennusta 12 viime kk aikana?	10.4
Onko Teillä ollut lääkärin toteamaa tai hoitamaa sydänveritulppaa 12 viime kk aikana?	10.3
Oletteko mielestänne saanut riittävästi sosiaalityöntekijän palveluita 12 viime kk aikana?	10.3
Oletteko mielestänne saanut riittävästi mielenterveyspalveluita 12 viime kk aikana?	10.2
Oletteko mielestänne saanut riittävästi sosiaaliasiamiehen palveluita 12 viime kk aikana?	10.2
Saatteko taloudellista tukea muilta sukulaisilta?	10.0
Onko Teillä ollut lääkärin toteamaa tai hoitamaa pitkäaikaista keuhkoputkentulehdusta 12 viime kk aikana?	10.0
Oletteko mielestänne saanut riittävästi tietoa kuntanne sosiaalipalveluista 12 viime kk aikana?	9.9
Onko Teillä ollut lääkärin toteamaa tai hoitamaa syöpää 12 viime kk aikana?	9.9
Onko teillä tarpeeksi rahaa tarpeisiinne nähden	3.1
Miten paljon painatte kevyissä vaatteissa?	2.9
Äänestittekö edellisissä kunnallisvaaleissa?	2.9
Syöttekö yleensä aamupalaa?	2.9
Kuinka usein olette viimeisen 7 päivän aikana syönyt tummaa leipää?	2.8
Kuinka pitkä olette?	2.7
Onko Teillä matkapuhelin?	2.7
Missä asutte tällä hetkellä?	2.6
Kuinka paljon liikutte ja rasitatte itseänne ruumiillisesti vapaa-aikana?	2.6
Siviilisääty	2.5
Täytittekö lomakkeen yksin vai auttoiko Teitä siinä joku muu?	2.4
Millaiseksi koette terveytenne?	2.4
Kuinka monta tuntia tavallisesti nukutte yöunta?	2.4
Asutteko...	2.2
Kuinka usein Teistä on 12 viime kk aikana tuntunut, että rahapelaaminen saattaa olla Teille ongelma?	2.1
Nukutteko mielestänne tarpeeksi?	2.0
Kuinka usein yleensä harjaatte hampaanne / hammasproteesinne?	2.0
Kuinka monta vuotta olette asunut yhtäjaksoisesti nykyisellä asuinpaikkakunnallanne?	1.9
Montako huonetta asunnossanne on?	1.7
Syntymävuosi	1.1

Myös kysymyksen muotoilu ja sen sisältämät aikamääreet voivat vaikuttaa vastaamiseen (kts. taulukko 4.5). Kysymyksiin, joissa puuttuvaa tietoa on vähän, on useimmiten helppo vastata (esim. syntymävuosi, asuinpaikka, paino, pituus). Sen sijaan paljon puuttuvaa tietoa sisältävät kysymykset sisältävät aikamääreitä kuten ”12 viime kuukauden aikana”, jolloin vastaaja voi kokea vaikeaksi muistella koko edellistä vuotta. Puuttuva tieto voi myös kertoa siitä, ettei vastaaja koe että kysymys koskettaa häntä, kuten tässä sairaus- ja palvelukysymykset, ja jättää vastaamatta vaikka ”ei” tai ”ei ole tarvittu” vaihtoehto on vastattavissa. Haluttomuus vastata kysymykseen voi myös johtua vastaajan muista ominaisuuksista, joita hän ei halua vastauksistaan paljastuvan. Esimerkiksi masentuneiden on huomattu jättävän vastaamatta masennusta koskeviin kysymyksiin (Suvisaari et al., 2009).

Selkeä ero löytyy myös erävastauskadon määrässä vastausaaltojen välillä. Kun vastaajat jaetaan vastauksen saapumisajankohdan mukaan kahteen ryhmään, aikaisiin ja myöhäisiin vastaajiin, huomataan esimerkiksi masennuskysymyksessä olevan enemmän puuttuvia vastauksia myöhäisten vastaajien välillä (taulukko 4.6). Mielen-

Taulukko 4.6: Itse raportoitua masennusta koskevaan kysymykseen vastanneiden osuus (%) vastausaalloittain

Vastausaalto	Osuus (%)	Keski-virhe	95% luottamusväli	
Aikaiset	91.5	0.3	91.0	92.1
Myöhäiset	87.8	0.5	86.9	88.8

kiintoista on pohtia, mitä tuloksia olisi saatu, mikäli olisi tavoitettu kaikki otokseen kuuluvat henkilöt ja he kaikki olisivat vastanneet esimerkiksi masennusta koskeviin kysymyksiin. Jotain käsitystä tästä voi saada tarkastelemalla saatuja tuloksia juuri vastausaalloittain, sillä usein kyselytutkimuksissa myöhäisimmät vastaajat edustavat katoon jääneitä paremmin kuin aikaiset vastaajat (Peress, 2010). Taulukossa 4.7 on tarkasteltu itse raportoidun masennuksen osuuksia vastausaalloittain ja tämä tukee ajatusta, että myöhäiset ja aikaiset vastaajat eroavat toisistaan. Jälkimmäisessä aallossa vastanneet ovat raportoineet masennusta suhteessa enemmän kuin ensimmäisessä aallossa vastanneet henkilöt. Erävastauskatoa voidaan korjata esimerkiksi imputoimalla. Masennuskysymyksen erävastauskadon imputointia on käsitelty luvussa 6.3.

Taulukko 4.7: Itse raportoituna masennus (%) vastausaalloittain

Vastausaalto	Osuus (%)	Keski-virhe	95% luottamusväli	
Aikaiset	10.6	0.3	9.9	11.2
Myöhäiset	13.1	0.5	12.0	14.1
Puuttuvia arvoja 1417				

Chapter 5

Statistical methods for missing data

Traditional approaches to missing data are deletion and substitution. Removing or omitting missing values is often set as a default in majority of statistical software. The more advanced methods can be divided into weighting, imputation and model-based methods (Little and Rubin, 2002). Weighting and imputation are attempts to address data not missing at random and they are widely used in systematic reviews. Model-based methods try to make assumptions about missing values' relationships with the available data and they require more statistical knowledge than the other approaches (Higgins and Green, 2011).

The choice of the method is closely linked with the type of missing data. If the data is missing completely at random, then deletion or omitting is an option. Unfortunately this is rarely the case with survey data (Molenberghs and Kenward, 2007). Weighting methods are usually used to correct the bias of unit nonresponse in sample surveys and they use information about the missingness probabilities themselves (Laaksonen, 2009; Molenberghs and Kenward, 2007). For example in design weights the aim is to adjust for nonresponse as if it was part of the sample design (Little and Rubin, 2002). Weighting methods are not only restricted for MCAR (*Missing Completely At Random*) missing data mechanisms and can be used also with MAR (*Missing At Random*) data (see section 5.1). Imputation-based methods use the available information to fill in the missing values in the data. Imputation works with MAR and MNAR (*Missing Not At Random*) missingness mechanisms but usually requires modifications to the standard analyses in order for valid inferences. Model-based methods define a model for the observed data and

base inferences on the likelihood or posterior distribution under that model. (Little and Rubin, 2002).

Traditional and simple methods are shortly reviewed in section 5.2. Weighting methods, like IPW and WEE are introduced in section 5.3. Imputation methods and especially multiple imputation is described in the section 5.4. Some these methods are compared using empirical population survey data in chapter 6. All the used methods are also compared to CC analysis based prevalences, where the estimates are calculated by using only sample weights (see section 5.2).

5.1 Missingness mechanisms and missingness patterns

Defining the missing-data mechanisms and patterns is an important part of choosing the method for handling the missingness. Especially the missingness mechanism has a crucial role in the analysis of data with missing values because the properties of missing data methods depend strongly on the nature of the dependencies in the mechanisms (Little and Rubin, 2002). Theoretical foundations of modern missing data analyses were first described by Rubin (1976). In his theory the missingness mechanism describes how the propensity for a missing value on a variable Y relates to other variables or to the possible values of Y itself. On an individual level the mechanism describes how the respondents propensity for missing data is related to other variables. (Enders, 2013). The basic taxonomy of missing data mechanisms is:

- Missing completely at random (MCAR)
- Missing at random (MAR)
- Missing not at random (MNAR)

There have been some discrepancies in the definitions of the terms here throughout the history of missing data modelling, especially in the MAR case (Seaman et al., 2013). If we start defining the missingness mechanism in the traditional way, the complete data is $Y = y_{ij}$ and $M = m_{ij}$ is the missing-data indicator matrix. The

missing-data mechanism is then defined by the conditional distribution of M given Y, say $f(M|Y,\phi)$, where ϕ denotes the unknown parameters. In MCAR case missingness does not depend on the missing or observed values of the data Y, which means

$$f(M|Y,\phi) = f(M|\phi) \text{ for all } Y,\phi \tag{5.1}$$

In the MAR case the missingness depends only on the observed components of Y and not on the missing components. Let Y_{obs} denote the observed components and Y_{mis} the missing components. Then MAR is defined

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \text{ for all } Y_{mis},\phi \tag{5.2}$$

Seaman et al. (2013) point out that this definition is problematic, since there is lack of detail whether the MAR condition is a statement only about the realized missingness pattern or about all possible patterns and also whether it is only about the realized values of the observed data or all possible observable data values. To clarify this, they created to definitions for MAR, realized and everywhere MAR.

Let a random variable $\mathbf{Y}$ denote the vector of the data values potentially observed on the data and $\mathbf{M}$ denote a vector of missingness indicators as described before. Both vectors are of the same length and the j^{th} element of $\mathbf{M}$ equals one if the j^{th} element of $\mathbf{Y}$ is observed and zero if it is missing. Let $o(\mathbf{Y},\mathbf{M})$, a function of $\mathbf{Y}$ and $\mathbf{M}$, denote a subvector of $\mathbf{Y}$ which contains the observed elements of $\mathbf{Y}$. K denotes the length of $o(\mathbf{Y},\mathbf{M})$ and is equal to the sum of the elements of $\mathbf{M}$. If $o(\mathbf{Y},\mathbf{M})$ is an empty set then $K = 0$ and no elements of $\mathbf{Y}$ are observed. Here $o(\mathbf{Y},\mathbf{M})$ is equivalent of Y_{obs}. Let $\overline{\mathbf{m}},\overline{\mathbf{y}}$ denote the realized values of the random variables $\mathbf{Y},\mathbf{M}$. The values of $\overline{\mathbf{m}}$ and $o(\overline{\mathbf{m}},\overline{\mathbf{y}})$ are known but that of $\overline{\mathbf{y}}$ is only known if all elements of $\overline{\mathbf{m}}$ equal one. Here $g_\phi(\mathbf{m}|\mathbf{y})$ (see before $f(\cdot|\cdot,\phi)$) denotes the probability that $\mathbf{M} = \mathbf{m}$ given that $\mathbf{Y} = \mathbf{y}$, where ϕ is an unknown parameter. According to Seaman et al. (2013),

$$\text{The data are realized MAR if } g_\phi(\overline{\mathbf{m}}|\mathbf{y}) = g_\phi(\overline{\mathbf{m}}|\overline{\mathbf{y}}) \quad \forall \mathbf{y},\phi, \tag{5.3}$$

$$\text{such that } o(\mathbf{y},\overline{\mathbf{m}}) = o(\overline{\mathbf{y}},\overline{\mathbf{m}})$$

Here the hypothesized missingness model assumes that the conditional missingness pattern $\mathbf{M}$ is its realized value $\overline{\mathbf{m}}$, given the realized values of $\mathbf{Y}$ that are observed when $\mathbf{M} = \overline{\mathbf{m}}$. MAR can also be a statement of about all the possible missingness patterns and values of the observed data.

$$\text{The data are everywhere MAR if } g_\phi(\mathbf{m}|\mathbf{y}) = g_\phi(\mathbf{m}|\mathbf{y}^*) \quad \forall \mathbf{m}, \mathbf{y}, \mathbf{y}^*, \phi, \qquad (5.4)$$

$$\text{such that } o(\mathbf{y}, \mathbf{m}) = o(\mathbf{y}^*, \mathbf{m})$$

where $\mathbf{y}$ and $\mathbf{y}^*$ represent a pair of values of $\mathbf{Y}$. This means that the probability of any possible missingness pattern does not depend on the values of the missing elements given the values of the corresponding observed elements and missing elements of the data.

These two definitions of MAR also lead to another way to define MCAR:

$$\text{The data is realized MCAR if } g_\phi(\overline{\mathbf{m}}|\mathbf{y}) = g_\phi(\overline{\mathbf{m}}|\mathbf{y}^*) \quad \forall \mathbf{y}, \mathbf{y}^*, \phi \qquad (5.5)$$

Realized MAR means that the realized missingness pattern does not depend on the data and it implies realized MAR. This definition is also equivalent to the combination where the missing data is realized MAR and the observed data is "observed at random" (Heitjan, 1994).

$$\text{The data is everywhere MCAR if } g_\phi(\mathbf{m}|\mathbf{y}) = g_\phi(\mathbf{m}|\mathbf{y}^*) \quad \forall \mathbf{m}, \mathbf{y}, \mathbf{y}^*, \phi \qquad (5.6)$$

Everywhere MCAR means that the probability of any missingness pattern given the data $\mathbf{M}$ is independent of $\mathbf{Y}$. (Seaman et al., 2013).

5.2 Simple methods

One of the oldest methods to treat missing data is to ignore the data with missing values. This is called complete case (CC) analysis. Until the 1980s getting rid of the incomplete cases was the only way to treat them in the data analysis (Schafer and Olsen, 1998). Complete case analysis can involve different type of deletion methods, like listwise or pairwise deletion. Listwise deletion simply discards all the observa-

tions with missing data and this is very often set as a default in many statistical software. Pairwise deletion is commonly used in correlations and modelling, there the deletion concerns all the cases where one or both of the values are unavailable for correlation matrix. To compute correlation matrix for regression models only complete data for pairs of correlated variables are used. Both of these deletion methods can lead to significant biases if the missing data is not MCAR and it also diminishes the analytic power if a large portion of data is missing. (Allison, 2001).

Other simple way to treat missing data is to use simple substitution methods. These involve single imputation such as the mean substitution and in longitudal designs last observation carried forward-method (LOCF). In single imputation, missing values are filled in with a predicted value based on for example a linear regression model of the available data. This may work well if the model is carefully chosen and the predictors are strong but often uncertainty inherent in the missing data may result in underestimation of standard errors (Rubin, 2004). In mean substitution missing values are filled with mean of the observed values, e.g. group mean. This method can lead to erroneous statistical inferences because the mean naturally depends on the study sample and there may be different reasons for individuals not to answer certain questions (Patrician, 2002). In longitudal studies a missing value can be replaced by the last observed value from the same individual.Very often this leads to distorted means and covariance structure even though the data is MCAR (Carpenter et al., 2014).

5.3 Weighting methods

Weighting methods in population surveys are used not only for nonresponse but also for unequal selection probabilites, overcoverage and sampling flunctuations from known population distributions (Kalton and Flores-Cervantes, 2003). In weighting methods usually population distributions are used to calibrate weights and in population surveys it is assumed that sample distributions represent the population. In this chapter the focus is on the inverse probability weighting (IPW) but also selection probabilities and overcoverage are accounted for in the empirical part in chapter 6. One of the simplest methods called base weighting is based on known selection

probabilites, which only accounts for unequal selection probabilities or equal sampling probabilites in case they do not represent the original population of interest. Complete case analysis (CC) based estimates are calculated with this method in chapter 6. These weights can be adjusted to account for nonresponse via different methods like cell weighting, poststratification, raking or inverse probability weighting. Adjusting is usually done by using auxiliary information of sample units and including that information in different modelling methods like generalized regression estimation or logistic regression modelling (Kalton and Kasprzyk, 1986).

Inverse Probability Weighting

In inverse probability weighting (IPW) individuals or single observations are weighted by the inverse of their probability of being completely observed. The model for that probability, *missingness model*, is based on variables in the data set, most often only completely observed ones (Seaman and White, 2013). Let's consider a case where we want to fit a model, *analysis model*, for a data set with non-complete observations. The outcome is a scalar Y on covariates X. Let Y_i and X_i denote the values for individual i and θ denote the model parameters. Parameter θ is estimated as the value of $\hat{\theta}$ that solves the score equations:

$$\sum_{i=1}^{n} U_i(\theta) = 0 \tag{5.7}$$

where $U_i(\theta)$ is the log likelihood function's first derivate with respect to θ. An individual is a complete case if both X and Y are observed. Let $R_i = 1$ if Y_i and X_i are observed and otherwise $R_i = 0$. In complete case analysis (CC analysis) estimating parameters means solving the CC score equations: $\sum_{i=1}^{n} R_i U_i(\theta) = 0$. In the IPW approach the model is fitted the same way but more weight is given to some complete cases than others. There the IPW score equations are:

$$\sum_{i=1}^{n} R_i w_i U_i(\theta) = 0 \tag{5.8}$$

where w_i is the weight given to individual i. If $w_i = 1$ for all i then IPW and CC score equations are the same. The weight w_i is the inverse of the probability of i being a

complete case. w_i can be estimated from the data available for all the individuals which means observed values of X, Y and Z, where Z means auxiliary variables not used in the analysis model. Only in the case of clear monotone missingness pattern or if a complicated Markov randomized monotone missingness model is used, also non fully observed data can be used (Seaman and White, 2013). The missingness model is commonly based on a logistic regression model for the outcome R_i where predictors come from the set X, Y, Z. When choosing the variables for the missingness model, the aim is to include sufficient variables from the set X, Y, Z where H denotes the predictors set such that

$$P(R = 1|\mathbf{X}, Y, Z, \mathbf{H}) = P(R = 1|\mathbf{H}) \tag{5.9}$$

is a credible assumption. Although completeness is not independent of Y given X, it is independent of Y and X given Z, i.e. $P(R = 1|X, Y, Z) = P(R = 1|Z)$. So, provided Z is fully observed, IPW can be used with $H = Z$. Variables for the model should be chosen so that the relation between them and the probability of being a complete case is correctly modelled (Seaman and White, 2013). All the variables which affect the probability that $R = 1$ should be included but also the possible interactions and nonlinearities between those variables should be examined and considered. Efficiency of the missingness model can also be increased by adding variables that do not predict missingness but are associated with Y and X (Tsiatis, 2004). Determining whether to include a new variable, say H^c, into the missingness model depends on how strongly the variable is associated with R given the variables in the H matrix and the different associationts between them but also on the sample size (Seaman and White, 2013). Including variables that do not associate enough with R usually decrease efficiency. A saturated model is not often possible to use in order to avoid unnecessary increase of variability of the weights. The fit should always be studied with model statistics such as goodness of fit and deviance, since poor fit can result in very small fitted probabilities for some individuals because of incorrect probability estimation. Large weights can also indicate a misspecified model.

Let $\pi_i = P(R = 1|\mathbf{H})$. If the model is right, $\frac{E(R_i)}{\pi_i} = 1$. IPW normal equations

are unbiased for known π_i and consistent for θ:

$$\sum_{observed} \frac{x_i(y_i - x_i'\hat{\theta})}{\pi_i} = 0 \tag{5.10}$$

IPW is applicable if the data are MAR, since then it is possible to obtain suitable estimates for the probabilities using for example logistic regression. (Molenberghs and Kenward, 2007).

Weighted estimation equations

IPW can be further used for estimating equation which is an approach to analyzing non-complete data. With weighted estimating equations (WEE), the contribution to the estimating equation from a complete observation is weighted by the IPW estimated weight (Lipsitz et al., 1999). Say we have an unbiased estimating equation for a parameter vector β and a vector of observations x_i (Molenberghs and Kenward, 2007):

$$\mathbf{S}(\beta) = \sum_{i=1}^{n} \mathbf{S_i}(\mathbf{x_i}, \hat{\beta}) = 0 \tag{5.11}$$

If $R_i = 1$ the following IPW estimating equation is:

$$\sum_{observed} \frac{\mathbf{S_i}(\mathbf{x_i}, \hat{\beta})}{P(R_i = 1)} = 0 \tag{5.12}$$

In its simplest form (with at least one predictor) the equation can be written as (see equation 5.10):

$$\sum_{i=1}^{n} \frac{R_i}{\pi_i} x_i(Y_i - x_i'\beta) = 0. \tag{5.13}$$

Inverse probability weights can also be applied for maximum likelihoods estimation (Rotnitzky and Robins, 1995). For examle, define a maximum likelihood estimator $\hat{\beta}_{MLE}$ of $E(Y_i)$ for binary X_i as following:

$$\hat{E}(Y_i|X_i = 1)\hat{f}(X_i = 1) + \hat{E}(Y_i|X_i = 0)\hat{f}(X_i = 0), \quad \text{where} \tag{5.14}$$

$$\hat{E}(Y_i|X_i = x) = \frac{\sum_i R_i Y_i I(X_i = x)}{\sum_i R_i I(X_i = x)} \tag{5.15}$$

and $\hat{f}(X_i = j)$ is $\sum I(W_i = j)/n, (j = 0 \times 1)$. Then $\hat{\beta}_{MLE}$ is a solution to the inverse probability weighted estimation equations as in equation 5.13, if it is written as

$$\sum_{i=1}^{n} \hat{\pi}_i^{-1} R_i (Y_i - x_i \beta) = 0, \quad \text{where}$$

$$\hat{\pi}_i = \hat{\pi}(X_i) = \frac{\sum_j R_j I(W_j = W_i)}{\sum_j I(W_j = W_i)}$$

5.4 Imputation

Imputation methods intend to replace the missing values with real values estimated from the data. The estimation can be based on different regression models or real donors from the data. For this, imputation model is constructed. The imputation model should account for the process that created the missing data, preserve the relations in the data, and preserve the uncertainty about these relations. van Buuren and Groothuis-Oudshoorn (2011) conclude constructing the imputation model into seven choices:

1. Defining the missingness mechanism. Does the data support MAR or MNAR assumption?

2. Form of the imputation model. What kind of model the dependent variable to be imputed needs? Does it have to preserve relations between other variables?

3. Predictor selection. Which variables to include as predictors?

4. Is there a need to impute variables that are functions of other possibly incomplete variables (e.g. sum scores, interactions)?

5. The order in which variables should be imputed. If there are multiple variables with missing values, which one is best to impute first so that the imputed values can be used in another variable's imputation?

6. How many imputations and iterations? Is the method to be used single imputation or multiple imputation?

7. If the choice is to do multiple imputation, the last choice is to decide the number of multiply imputed data sets.

Before or after these choices, I would like to add one concerning the software. Not all the software support all the possible imputation methods but all do support some. In table 5.1 there are listed some imputation methods and under which procedure or package they are available in the software. Which ever software and method is chosen it is vital to know how the software executes the imputation. The methods mentioned in the table are better introduced later in this chapter.

Table 5.1: Statistical software and their imputation methods available. See for example Van Buuren (2012).

Method	**SAS**	**SUDAAN**	**SPSS**	**R**	**Stata**
MI with JM	PROC MI (MCMC)		Amos, MI	mi, BaBoon	mi
MI with FCS or MICE	IVEware, PROC MI (FCS in version 9.3)		MI with MCMC method	mice, Hmisc, Amelia	ice
HOT DECK	Macros by Altmyer and Ellis	PROC HOTDECK		rpart, mirf, VIM	module hotdeck
EM	PROC MI (EM)				mi with EM optimization

Hot deck imputation

Hot deck imputation is a method for handling missing data in which each missing value is replaced with an observed response from a "similar" unit. This method is extensively used in correcting survey non-response yet there is no consensus on what would be the best way to apply the hot deck method. There are variations on how to define the similarity of the recipient with missing values and the donor from which the substituting value is borrowed. The choice of donor can be deterministic or stochastic. (Marker et al., 1999).

Based on this variation hot deck methods can be divided into two subgroups; random hot deck methods and deterministic hot deck methods. In random hot deck methods the donor of the value to replace the missing one is randomly chosen

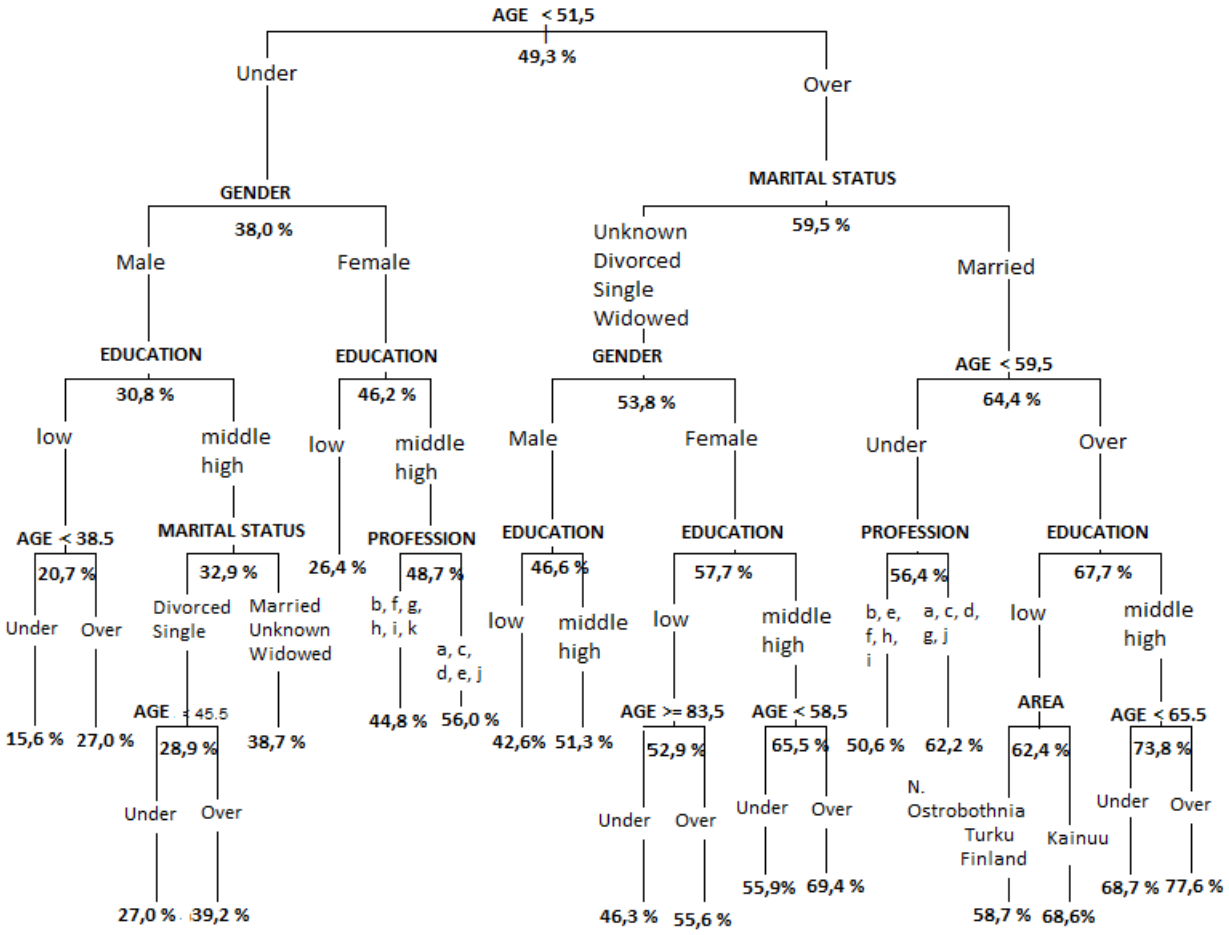


Figure 5.1: Regression tree of response rates in the ATH study 2010 (See chapter 6). In profession a=Experts, b=Unknown, c=Technicians, d=Managers, e=Agricultural workers, f=Other, g=Service and Sales workers, h=Craft related trades workers, i=Elementary workers, j=Armed forces, k=Clerical support workers.

from the pre-defined donor pool. Respondents are divided into different groups called donor pools and the substituting value for the missing value is borrowed from within the same donor pool. Donor pools are constructed using the characteristics observed for both recipient with missing values and the donor, based on the similarity of those characteristics. Those significant characteristics can be found by examining the covariates' correlations and other associations with the variable to be imputed (Marker et al., 1999) and also the response propensity to that item is a valuable part of adjusting donor pools (Andridge and Little, 2010). One example of this is to use CART-functions (*Classification And Regression Trees*) to construct a regression tree (one example with empirical data used here is figure 5.1) (Therneau et al., 2014) or simple regression (Little and Rubin, 2002). These regression models should be carefully chosen in order to avoid over-parameterization that may lead to too small donor pools.

In deterministic hot deck methods one single donor is chosen usually based on nearest neighbor methods. Very often the nearest neighbor donor is chosen based on the same characteristics as in donor pool formations. One method that combines both of these is selecting a single donor by random selection with probability inversely proportional to their distance from the recipient(Andridge and Little, 2010). Non-response adjusting weights can also be included in the hot deck imputation process to reduce the bias in case the weights are related to the imputed variable. This method is called weighted sequential hot deck imputation or weighted random hot decks.

The simplest hot deck procedure uses the entire group of respondents as a single donor pool, but this method only produces consistent estimates when data are missing completely at random (Little and Rubin, 2002). If the data are not MCAR, in random hot deck method it is essential that the response probability is allowed to depend on auxiliary variables that create the donor pools. In this case both random and deterministic hot deck methods are possible to use. The one most important requirement for the hot deck methods to yield consistent estimates is the existence of some donors for non-respondents at every value of the set of covariates that are related to missingness. Donor pools sizes should also be carefully considered in order to avoid exploitation of one single donor within a donor pool. It is also crucial

to remember that hot deck imputed data sets are often analyzed as if they had no missing values, which leads to deflated variance estimates. Variance estimation should always be considered separately via explicit variance formulae, resampling methods like jack-knife or bootstraps or using multiple hot deck imputation.

Good qualities of hot deck methods are that they do not rely on model fitting for the variable to be imputed. This means that the methods can be less sensitive to model misspecification than parametric imputation methods such as single regression imputation. Hot deck methods also always produce only plausible values since they come from real observed values. Also with hot deck methods the non-response bias can be reduced if there is association between the variables defining donor pools and both the propensity to respond and the variable to be imputed. (Andridge and Little, 2010).

Multiple imputation

In multiple imputation methods (MI) imputed values come from predictive multivariate distribution specified for the data. These imputation methods have proved to be valuable especially in survey context because of their ability to handle large datasets, to provide suitable values for parameter estimation often used for surveys and to account for the differences between respondents and nonrespondents. The three basic steps to multiple imputation according to Rubin (2004) are:

1. Introduce random variation into the process of imputing missing values, and generate several data sets using predictive distributions, each with slightly different imputed values.

2. Perform a standard complete case analysis on each of the data sets.

3. Combine the results into a single set of parameter estimates, standard errors, and test statistics.

Each missing data entry is replaced with plausible data values that come from a distribution specifically modelled for that data entry. This produces multiple data sets that are identical for the non-missing data entries but differ in the imputed values. The difference magnitude reflects the uncertainty about what value to impute.

These data sets can be then pooled together to form one estimate to use in analysis. It is also important to estimate the variance of the final estimate (Rubin, 2004).

There are two general approaches to multiple imputation: joint modeling (JM) and fully conditional specification (FCS), also known as multivariate imputation by chained equations (MICE). Both JM and FCS are iterative methods. In JM techniques a tractable multivariate distribution is specified for the missing data and the missing value replacement are drawn from their conditional distributions by Markov chain Monte Carlo (MCMC) techniques. FCS specifies the multivariate imputation model on variable-by-variable basis. This is done using a set of conditional densities, one for each incomplete variable. JM is often used when a multivariate distribution is a reasonable description of the data but FCS provides an alternative for the cases where no suitable multivariate distribution can be found. FCS might also be found more efficient since a low number of iterations is often sufficient (van Buuren and Groothuis-Oudshoorn, 2011). When using MCMC algorithm the data has to be MAR and usually based on multivariate normal distribution but if coded right by the user, it can handle different types of variables. FCS algorithm is a bit more flexible and it is able to impute both quantitative and categorical variables (Allison, 2012).

The aim is to create statistically correct imputations for a wide range of estimators, for example regression coefficents. Let Y_j with $j = (1, \ldots, p)$ denote one of the p incomplete variables where $Y = (Y_1, \ldots, Y_p)$ and Q denote the quantity of scientific interest. Imputing multivariate missing data we may face different problems:

- Predictors used for given Y_j in the imputation model may be incomplete
- Circular dependence can occur between incomplete variables
- If there is a large amount of incomplete variables, collinearity may be a problem and empty cells may occur
- Variables can be of different scale of measurement and thereby convienient models may be hard to find
- The relation between Y_j and its predictors may be complex or subject to censoring processes

- Imputation may create implausible combinations or be nonsensical
- Models for Q that will be applied are not properly defined

When facing these problems it is convenient to specify the imputation model for each variable separately which has led to using chained equations. There the posterior multivariate distribution for the complete data Y is believed to be completely specified by θ. With MICE algorithm it is possible to obtain the postierior distribution of θ by sampling iteratively from conditional distributions of the form

$$F(Y_1|Y_{-1}, \theta_1)$$

$$\vdots$$

$$F(Y_p|Y_{-1}, \theta_p)$$

where Y is a partially observed random sample from the p-variate multivariate distribution $F(Y|\theta)$ and the parameters $\theta_1 \dots \theta_p$ are specific to the respective conditional densities. Chained equations start from a simple draw from observed marginal distributions. The t^{th} iteration of chained equations is a Gibbs sampler that successively draws

$$\theta_1^{*(t)} \sim F(\theta_1|Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)})$$

$$Y_1^{*(t)} \sim F(Y_1|Y_1^{obs}, Y_2^{(t-1)}, \dots, Y_p^{(t-1)}, \theta_1^{*(t)})$$

$$\vdots$$

$$\theta_p^{*(t)} \sim F(\theta_p|Y_p^{obs}, Y_1^{(t)}, \dots, Y_{p-1}^{(t)})$$

$$Y_p^{*(t)} \sim F(Y_p|Y_p^{obs}, Y_1^{(t)}, \dots, Y_p^{(t)}, \theta_p^{*(t)})$$

where $Y_j^t = (Y_j^{obs}, Y_j^{*(t)})$ is the j^{th} imputed variable at iteration t. Previous imputations $Y_j^{*(t-1)}$ only enter $Y_j^{*(t)}$ through its relation with other variable (van Buuren and Groothuis-Oudshoorn, 2011). One of the flaws of basic imputation is that it often produces standard errors that are too low because it does not take into account that the parameters in the imputation equation are only estimates with their own sampling variability. This can be corrected by using different imputation parameters to create each imputation set as in chained equations. Different imputation

parameters can be chosen using Bayesian inference where the imputation parameters are all random draws from the posterior distribution of the imputation parameters. Using a different set of imputation parameters for each data set induces additional variability into the imputed values across data sets, leading to larger standard errors (Allison, 2012).

Chapter 6

Empirical analysis

In empirical analysis a unit level population survey data from Regional Health and Wellbeing study 2010 (ATH from its Finnish initials) is used to examine the use of inverse probability weighting in correcting the unit nonresponse bias and to examine the use of hot deck imputation and multiple imputation in correcting the item nonresponse bias.

ATH study was conducted by National Institute for Health and Welfare as a mail survey including both a representative sample of Finland (sample of 5,000) and area samples from city of Turku (sample of 9,000), Northern Ostrobothnia (sample of 8,000) and region of Kainuu (sample of 9,000). The aim of the study was to form a conception of respondents' health and welfare but also their use of public services (such as health and social stations etc.) and their rate of quality of the services used. The questionnaire covered questions for example about living conditions, well-being, health, functional capacity, food, exercise, risk living habit and services. The whole questionnaire is available at `www.thl.fi/ath`. The response rate varied between 37-65 % within different areas and agegroups. The variation of the nonresponse between age groups indicates that the data is not MCAR so the methods are chosen to work with MAR or MNAR data.

Nonresponse bias will be examined via both sample variable distributions and via self-reported depression. The question looked at was "Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months?" and depression was one of the conditions mentioned in this question. The original question is presented in figure 6.1. Because of its question structure and its disposi-

42. Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months?

	no	yes
high blood pressure, hypertension	1	2
(cerebral) stroke	1	2
high blood cholesterol	1	2
coronary thrombosis, myocardial infarction	1	2
coronary disease, angina pectoris (=chest pain under physical strain)	1	2
cancer	1	2
rheumatoid arthritis or other inflammatory arthritis	1	2
arthrosis of the back, sciatica, low back pain or other back condition	1	2
chronic bronchitis, emphysema	1	2
depression	1	2
other mental health problem	1	2
asthma	1	2
pollen allergy, hay fever	1	2
lactose intolerance (sensitivity to lactose = milk sugar)	1	2
food allergy to milk (not lactose intolerance), to egg, fish or wheat or other grain	1	2
food allergy to raw vegetables or fruit (e.g. peas, carrots, apples)	1	2

Figure 6.1: The original question of interest and its response possibilities. (Kaikkonen et al., 2010b)

tion also valuable information is gained from the response indicators of the previous and the following alternative. The overall item nonresponse rate for self-reported depression was 9.4 % which equals 1426 missing values. Both response indicator and the value of the variable itself was used in correcting the nonresponse bias.

In survey research it has been often found that respondents who belong to the lower socioeconomic group or have poor health do not answer the surveys (see for example Tilastokeskus (2009)). Also the item nonresponse rate seems to be higher for those who suffer from depression than those who do not (see for example Koyama et al. (2014)). The aim is to examine if the nonresponse affects the rate of people reporting depression and the results are reflected on the register statistics of people receiving reimbursements of depression medication in 2010 (register from Social Insurance Institution, published by SOTKAnet Statistics and Indicator Bank (2005 - 2013)). The focus is also on correcting the unit nonresponse bias in distributional level of sample variables such as age and gender to match the distribution of the original sample which represents the population of the research area.

The sample variables such as age, gender, living area and marital status are obtained from National Register Centre. The auxiliary information about education and profession come from Statistics Finland and the information of the right for special reimbursement of medication come from Social Insurance Institution. Education and profession are based on International Standard Classification of Occupations (ISCO-08). In Finland special reimbursement of medication is granted on the basis of a physician at National Social Insurance Institution confirming each diagnosis of internest. These special reimbursement medication include for example psyche related and diabetes medication (Social Insurance Institute, 2014).

6.1 Correcting unit nonresponse with IPW

Application

The response rate in ATH study varied between 37 % and 65 % and it was clear to see that the rate was higher in the age groups of 55-74-year-olds and over 75-year-olds than in the age group of 20-54-year-olds (Kaikkonen et al., 2010c). Overall in the whole data set (n= 31,000) the response rate was 49 %.

Table 6.1: All possible completely observed coefficents and their age and gender adjusted univariate estimates for unit non-response model

Coefficients:	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.04	0.32	-3.28	0.00	**
Agegoup (every 10 years)	0.25	0.01	27.52	< 2e-16	***
Gender: Female	0.54	0.03	16.82	< 2e-16	***
Area: Finland	-0.21	0.04	-5.74	0.00	***
Area: Northern Ostrobotnia	-0.21	0.03	-6.50	0.00	***
Area: Turku	-0.19	0.03	-5.95	0.00	***
Marital status: Unknown	-0.60	0.21	-2.83	0.00	**
Marital status: Divorced	-0.27	0.04	-7.12	0.00	***
Marital status: Widowed	-0.49	0.04	-11.04	< 2e-16	***
Marital status: Single	-0.36	0.03	-11.39	< 2e-16	***
Language: Swedish	-0.04	0.32	-0.11	0.91	
Language Finnish	0.02	0.31	0.06	0.95	
Language: Russian	-0.29	0.33	-0.87	0.38	
Profession: Unknown	-0.06	0.05	-1.23	0.22	
Profession: Technicians etc	0.04	0.06	0.77	0.44	
Profession: Managers	-0.14	0.09	-1.51	0.13	
Profession: Agricultural workers	-0.27	0.09	-2.86	0.00	**
Profession: Elementary workers	-0.25	0.07	-3.89	0.00	***
Profession: Service and Sales workers	-0.15	0.06	-2.66	0.01	**
Profession: Craft related trades workers	-0.22	0.07	-3.20	0.00	**
Profession: Plant and machine operators	-0.39	0.07	-5.89	0.00	***
Profession: Armed forces	0.51	0.25	2.00	0.05	*
Profession: Clerical support workers	-0.17	0.07	-2.23	0.03	*
Eduction: High	0.33	0.08	3.85	0.00	***
Education: Low	-0.35	0.04	-8.83	< 2e-16	***
Gender:Female * Education: High level	-0.32	0.12	-2.66	0.01	**
Gender:Female * Education: Basic level	-0.37	0.05	-7.10	0.00	***
Medicine reimbursement: Other	-0.24	0.07	-3.16	0.00	**
Medicine reimbursement: Psyche	-0.51	0.08	-6.12	0.00	***
Medicine reimbursement: Diabetes	-0.05	0.05	-0.91	0.36	
Medicine reimbursement: Overall	0.10	0.03	3.24	0.00	**

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Modelling the unit nonresponse indicator i.e. the fact that the unit has participated to the survey, with all sample and auxiliary variables that are available for the whole sample (see table 6.1), it is clear that respondents differ from the non-respondents. Most of the available variables are significant coefficents in the model, except for language information. Also the interaction of gender and education is significant, indicating that the rate of male respondents with only basic level education is the highest in this case. This can also be seen on the odds ratio values (table 4.4). Odds ratio for women to participate is 0.5 times higher than for men and between the educational levels the difference is even higher (2.3 with higher level education). If simplified, the respondent most likely to participate is 55-74-year-old married woman from Kainuu area who works as an expert or specialist and has no medical reimbursements. It is clear that this assumption cannot be implemented to cover neither the whole sample nor the population of Finland or Kainuu. Since there are significant coefficents to the nonresponse model, there is a reason to assume that there is some nonresponse bias in population estimates drawn from only complete case data.

Using the information of significant coefficients the model for the nonresponse can be formed and use those significant coefficients as predictors for the response. The final model should include all the variables and their significant interactions. Since there is a lot of deviation in response rate in the age group of 20-54-year-olds (see figure 5.1), age will be used factored into 10-year intervals (20-29-year-olds, 30-39-year-olds etc.). To find the best model we form all the possible combinations of the main effects and first order interactions and then study the models' AIC, BIC and deviation. The best models are introduced in table 6.2 sorted by BIC since it is strictest criterion on adding new variables and choosing the model with smallest BIC is equivalent to selecting the model with the maximum posterior probability (Posada and Buckley, 2004).

Here the best model according to BIC is the main effects model with gender-education interaction. Best model according to AIC and deviance contains the main effects and age-gender interaction, in which the degrees of freedom (df) is larger and that is why it is not the best model based on BIC. Overall the best AIC and deviance are not far apart from the AIC and the deviance of the best model based on BIC so

Table 6.2: Models for unit nonresponse arranged by the smallest Bayesian information criterion. Highlighted cells indicate the best value for AIC and deviance

AIC	**BIC**	**deviance**	**df**	**Model name**
40135	40435	40063	36	Main effects+Gender*Education
40167	40459	40097	35	Main effects+Gender*Meds:other
40178	40462	40110	34	Main effects
40112	40462	40028	42	Main effects+Age*Gender
40147	40464	40071	38	Main effects+Gender*Maritalstatus
40175	40467	40105	35	Main effects+Gender*Meds:psyche
40177	40469	40107	35	Main effects+Meds:other*Meds:psyche
40179	40479	40107	36	Main effects+Meds:other*Education
40182	40482	40110	36	Main effects+Meds:psyche*Education
40180	40489	40106	37	Main effects+Area*Meds:psyche
40181	40489	40107	37	Main effects+Gender*Area
40181	40489	40107	37	Main effects+Area*Meds:other
40181	40489	40107	37	Main effects+Language*Meds:other
40181	40490	40107	37	Main effects+Gender*Languagegroup
40184	40492	40110	37	Main effects+Language*Meds:psyche
40179	40496	40103	38	Main effects+Maritalstatus*Meds:psyche
40184	40501	40108	38	Main effects+Maritalstatus*Meds:other
40141	40508	40053	44	Main effects+Gender*Profession
40177	40510	40097	40	Main effects+Area*Education
40178	40512	40098	40	Main effects+Language*Education

Main effects include age, study area, gender, education, profession, marital status, language, answer wave (early/late), right to receive special reimbursement for psyche medication (Meds:psyche) and right to receive special reimbursement for medication other than psyche or diabetes medication (Meds:other).

it is safe to choose the model with smallest BIC. IPW weights are calculated based on the probability to participate predicted by the chosen model. Weights are also scaled to match the final number of respondents (14,799) and population sizes in each area. Here are the basic statistics of the weigths:

```
   Min.  1st Qu. Median  Mean    3rd Qu. Max.
0.2732  0.6649  0.9580  1.0000  1.2570  4.3710
```

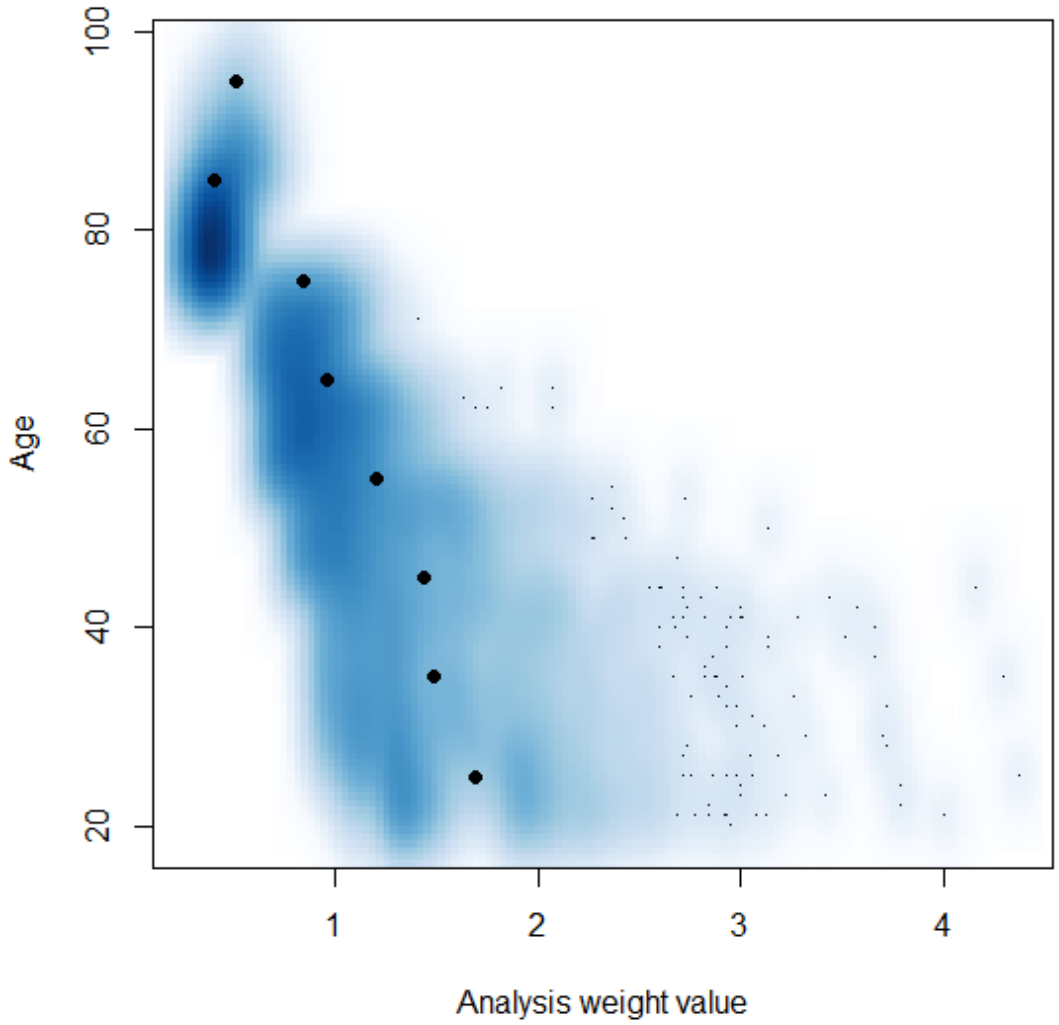


Figure 6.2: Density of analysis weights plotted against continuous age. Large points indicate the group mean of the agegroup in 10-year intervals (see table 6.3).

No critical sign of unstable weights is seen since there are no zero value weights and the largest weight is 4.4. The denisities show that older respondents tend to

receive smaller weights and in the younger agegroups there are more variation in weight values (figure 6.2). If we take a closer look at the weights plotted against age, gender and education (figure 6.3) we find that most of the weight values are between 0.5 and 2.0. The highest weights are received by males and respondents with low level education. Examining the weights more we see in table 6.3 that the highest weight is received by 30-39-year –old male with low education.

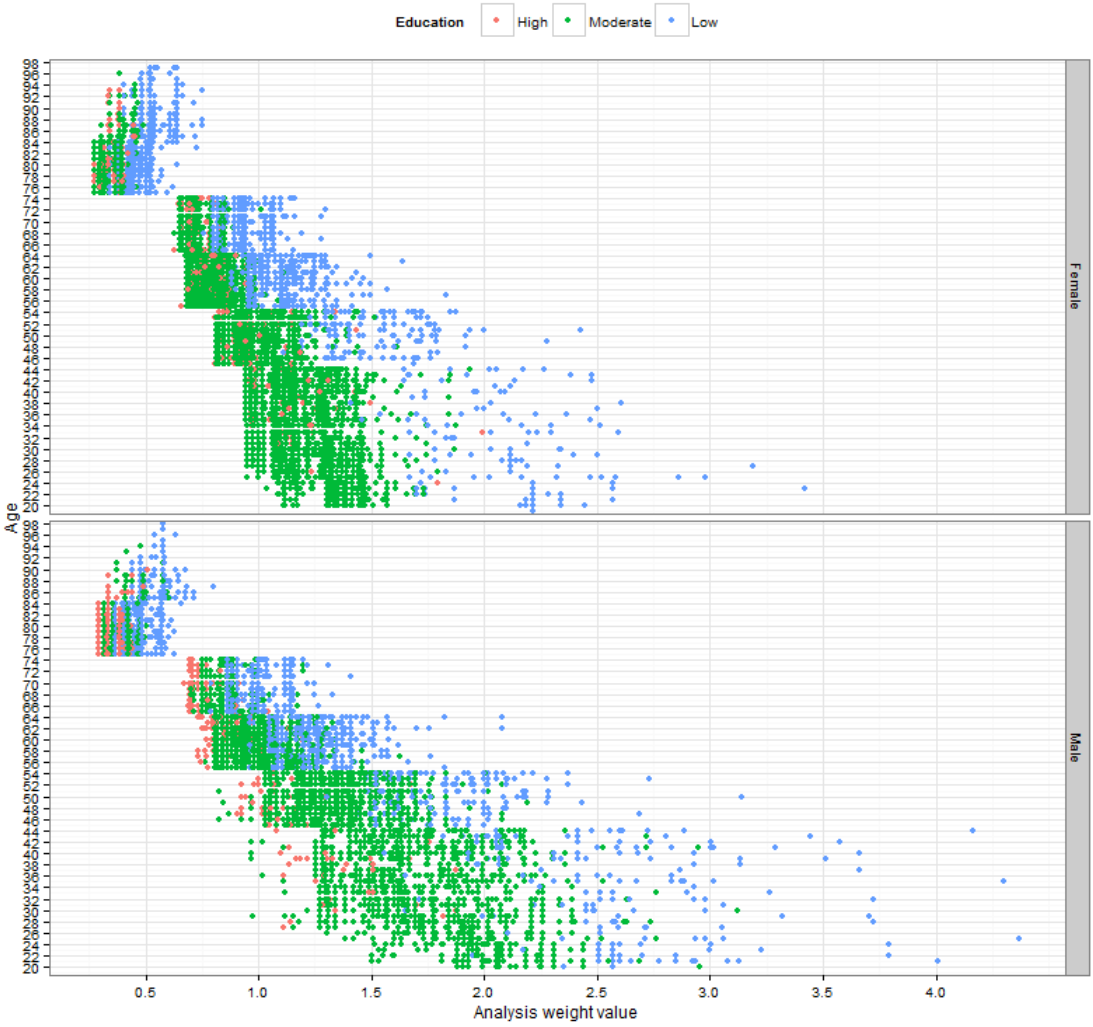


Figure 6.3: Weight values for both genders plotted against continuous age. Color indicates the education group.

Results

Let's consider CC estimates and IPW estimates of self-reported depression prevalences in table 6.4. There is also register rates for people who received reimburse-

Table 6.3: Basic statistics of analysis weights by gender, age and education and number of complete cases and sample and response rates by group.

Group	**min**	**mean**	**max**	**st. Dev.**	**Respondents**	**Sample**	**Response rate (%)**
Female	0.27	0.88	3.42	0.37	8705	15884	54.8
Male	0.29	1.16	4.37	0.57	6588	15116	43.6
20-29	1.04	1.69	4.01	0.46	703	2221	31.7
30-39	0.95	1.49	4.37	0.43	1684	4694	35.9
40-49	0.94	1.44	4.30	0.44	1624	4229	38.4
50-59	0.81	1.21	3.14	0.31	2416	5155	46.9
60-69	0.66	0.97	2.08	0.19	3099	5293	58.5
70-79	0.62	0.85	1.41	0.13	2080	3157	65.9
80-89	0.27	0.41	0.72	0.06	2995	4847	61.8
90+	0.50	0.79	1.37	0.21	671	1360	49.3
High	0.27	0.74	1.99	0.32	878	1390	63.2
Moderate	0.27	1.09	3.12	0.42	5031	10460	48.1
Low	0.34	0.87	4.37	0.57	9384	19150	49.0
Overall	0.27	1.00	4.37	0.49	15272	30956	49.3

Table 6.4: Self-reported depression prevalences of the four study areas in two age-groups for CC analysis and IPW

Area		**age 25 to 64, % (CI)**	Std error	**age over 65, % (CI)**	Std error
Kainuu	CC	10.5 (9.2-11.7)	0.63	13.2 (11.4-15.0)	0.94
	IPW	10.8 (9.5-12.2)	0.68	13.7 (11.8-15.6)	0.95
	Register*	9.9		10.9	
Northern Ostrobothnia	CC	8.3 (7.1-9.4)	0.59	12.4 (10.2-14.6)	1.12
	IPW	8.8 (7.5-10.0)	0.66	13.0 (10.7-15.3)	1.16
	Register*	9.9		12.6	
Turku	CC	12.3 (11.0-13.6)	0.67	12.7 (10.8-14.7)	0.98
	IPW	13.2 (11.7-14.7)	0.76	13.4 (11.4-15.4)	1.02
	Register*	11.3		11.4	
Finland	CC	10.7 (9.0-11.3)	0.85	10.0 (7.6-12.3)	1.20
	IPW	11.3 (9.4-13.1)	0.94	10.5 (8.9-13.9)	1.25
	Register*	10		11.9	

** Register from Social Insurance Institution; Rate of people that received reimbursement of depression medication in 2010 (SOTKAnet Statistics and Indicator Bank, 2005 - 2013)*

ments of depression medication in the research year 2010 (SOTKAnet Statistics and Indicator Bank, 2005 - 2013) but only to examine the differences distribution-wise since medical reimbursement and self-reported depression are two slightly different aspects. Compared to CC rates IPW estimates tend to be a little higher. Biggest difference weighting causes in estimates in city of Turku, where the register based rate is also highest. Compared to register distributions the difference between the two age groups is more significant in IPW estimates. Especially in Kainuu and Northern Ostrobothnia the difference in IPW estimated rates are significantly greater between agegroups than the register based distribution suggests.

Here we see clearly how weighting causes increase in variance and affects confidence intervals and standard errors because respondents with low estimated response propensity receive large weights but their influence on the estimate may be small. Since depression is somewhat associated with noresponse, increase in variance can be justified bias reducing effect of weighting (Little and Rubin, 2002).

If we consider the effect of weighting on odds ratio point estimates for depression (table 6.5), weighting changes the estimate for marital status and age. Significant change can also be seen in education class odds ratio estimates, where low education has a more significant affect and moderate education has less significant affect in weighted estimates. Also confidence intervals are increased in width but not substantially.

6.2 Correcting item nonresponse with WSHDI

Application

Using the IPW weights created in the previous section we now proceed to correct the nonresponse bias in self-reported depression with Weighted Sequential Hot Deck Imputation (WSHDI). This WSHDI algorithm corrects the potential bias of simple hot deck imputation by using weighting technique. WSHDI preserves the methodology of the unweighted procedure, but allows all respondents the chance to be a donor and uses sampling and/or analysis weights to restrict the number of times a respondent value can be used for imputation (Cox, 1980). Weights of the non-respondents are rescaled to sum to the total of the respondent weights. The algorithm is designed

Table 6.5: Odds ratio estimates of complete data model variables for CC analysis and IPW.

	CC			IPW		
	OR	**95 % CI**		**OR**	**95 % CI**	
Area Finland	1.00			1.00		
Area Kainuu	1.08	0.90	1.29	1.04	0.86	1.26
Area Northern Ostrobothnia	0.86	0.71	1.05	0.84	0.68	1.03
Area Turku	1.21	1.01	1.44	1.22	1.01	1.48
Education high	1.00			1.00		
Education low	1.27	0.66	2.44	1.51	1.15	1.98
Education moderate	1.46	0.78	2.71	0.92	0.71	1.20
Single	1.00			1.00		
Unknown	1.64	0.20	13.24	1.46	0.48	4.43
Married	1.17	0.77	1.77	0.66	0.57	0.78
Divorced	1.36	1.13	1.63	1.43	1.18	1.74
Widowed	1.22	0.76	1.95	1.38	1.12	1.69
Gender Male	1.00			1.00		
Gender Female	1.19	0.91	1.57	1.22	1.07	1.38
20s	1.00			1.00		
30s	0.76	0.40	1.47	0.99	0.72	1.35
40s	1.21	0.67	2.22	1.12	0.84	1.51
50s	1.10	0.61	1.98	1.32	1.00	1.74
60s	0.72	0.39	1.35	0.96	0.72	1.27
70s	0.95	0.51	1.77	1.29	0.97	1.72
80s	1.32	0.70	2.50	2.04	1.52	2.74
90s	0.87	0.16	4.86	1.98	1.10	3.56

so that, over repeated imputations, the weighted mean obtained from the imputed values is equal in expectation to the weighted mean of the respondents alone within imputation strata (Andridge and Little, 2010). The procedure is in this case carried out with SUDAAN PROC HOTDECK.

"Similarity" of donor to recipient is still controlled by the selection of model variables. In this case the variable set for the model is the same as in unit nonresponse added with information of the response wave (early or late) since it seems to be associated with the answer. The other determining factor in variable choices is that PROC HOTDECK does not handle well missing values in donor determining variables and missingness in other than sample based predictors of depression associates highly with the missingness in depression variable. Over-specification of the model also increases the risk of too small donor pools.

Results

Table 6.6: Self-reported depression rates of four study areas in two agegroups for WSHDI and CC analysis.

Area		**age 25-64**	Std error	**age over 65**	Std error	**N missing**
Kainuu	CC	10.5 (9.2-11.7)	0.63	13.2 (11.4-15.0)	0.94	649
	WSHDI	10.5 (9.1-11.8)	0.69	13.6 (11.9-15.3)	0.87	5
	Register*	9.9		10.9		
Northern Ostrobothnia	CC	8.3 (7.1-9.4)	0.59	12.4 (10.2-14.6)	1.12	313
	WSHDI	8.8 (7.4-10.2)	0.7	13.6 (11.4-15.8)	1.12	23
	Register*	9.9		12.6		
Turku	CC	12.3 (11.0-13.6)	0.67	12.7 (10.8-14.7)	0.98	276
	WSHDI	13.7 (11.8-15.5)	0.94	13.9 (11.7-16.2)	1.16	17
	Register*	11.3		11.4		
Finland	CC	10.7 (9.0-11.3)	0.85	10.0 (7.6-12.3)	1.20	179
	WSHDI	11.0 (9.2-12.8)	0.91	10.1 (7.9-12.2)	1.09	13
	Register*	10		11.9		

** Register from Social Insurance Institution; Rate of people that received reimbursement of depression medication in 2010 (SOTKAnet Statistics and Indicator Bank, 2005 - 2013)*

The final model variables are answer wave, agegroup (25-64 and 65+), area, education, profession, marital status and special medical reimbursement of psyche and other medication (diabetes medication excluded). With this model PROC HOTDECK produces 1362 imputed values for depression. 64 values remain missing maybe due to lack of appropriate donor. The maximum use of single donor is

4 times and altogether 1354 different donors are used. Results show very similar results as CC analysis as far as proportions show but standard errors increase with WSHDI in the younger agegroup. Compared to register based distribution WSHDI seems to respect the relation of age and depression. The imputation efficiency is the best in Turku case in terms of the rate of missing values after imputation proportioned to the number of missing values in the beginning (only 0,8 % is left without imputation).

Table 6.7: Odds ratio estimates of complete data model variables for WSHDI and CC analysis.

	CC			**WSHDI**		
	OR	**95 % CI**		**OR**	**95 % CI**	
Area Finland	1.00			1.00		
Area Kainuu	1.08	0.90	1.29	1.05	0.87	1.27
Area Northern Ostrobothnia	0.86	0.71	1.05	0.89	0.73	1.09
Area Turku	1.21	1.01	1.44	1.32	1.08	1.61
Education high	1.00			1.00		
Education low	1.27	0.66	2.44	1.72	1.26	2.36
Education moderate	1.46	0.78	2.71	1.11	0.82	1.50
Single	1.00			1.00		
Unknown	1.64	0.20	13.24	1.23	0.37	4.10
Married	1.17	0.77	1.77	0.55	0.46	0.66
Divorced	1.36	1.13	1.63	1.18	0.94	1.48
Widowed	1.22	0.76	1.95	1.33	1.06	1.66
Gender Male	1.00			1.00		
Gender Female	1.19	0.91	1.57	1.01	1.00	1.01
20s	1.00			1.00		
30s	0.76	0.40	1.47	0.91	0.64	1.29
40s	1.21	0.67	2.22	0.94	0.68	1.31
50s	1.10	0.61	1.98	1.07	0.79	1.46
60s	0.72	0.39	1.35	0.81	0.59	1.10
70s	0.95	0.51	1.77	1.12	0.82	1.53
80s	1.32	0.70	2.50	1.74	1.27	2.40
90s	0.87	0.16	4.86	1.96	1.05	3.66

In gender and marital status WSHDI seems to weaken the affect and decrease the significance of the differences between classes. In the agegroups of 70 and upwards WSHDI seems to result in more significant differences compared to the agegroup of 20s. In the imputed results age seems to be a little less determining in self-reported depression and in marital status marriage seems to affect more negatively than in CC analysis. In education groups the differences are more significant with WSHDI

than with CC especially in the low education class. This may indicate that the low educated people that have not responded to the question may suffer from depression more than the other non-respondents. Areal differences compared to Finland are increased in Turku but otherwise stay the same way as in CC analysis.

6.3 Correcting item nonresponse with MI

Application

Here item nonresponse of self-reported depression is corrected by multiple imputation by chained equations (MICE) described in the section 5.4. Predictors are selected according to van Buuren and Groothuis-Oudshoorn (2011) as follows:

1. all the variables of the "complete data model" are included as predictors. This means that all the variables used in the analysis or modelling the imputed data will be included which are in this case the same predictors as in unit nonresponse analysis (see section 6.1). This preserves the possible predictive relations in the complete data.

2. all the variables associated with the nonresponse, both unit and item nonresponse are included. This is done by using the results of section 6.1 and studying correlations with response indicator of self-reported depression.

3. all the variables that explain the target variables variance are included. This is determined by examining the correlations with the target variables and choosing those exceeding certain correlation.

4. variables with too many missing values in different subgroups of incomplete cases are then removed from the model due to preserving the efficiency.

After these steps the final model is constructed from age, gender, area, profession, education, language, marital status, response indicators of previous and the following alternatives to depression (see figure 6.1). By modelling the item nonresponse with variables that have no missing values (sample variables and response indicators) we see that the best model for item nonresponse is main effects plus interaction of psyche medicine and response indicator of all alternatives and interaction of response

indicators of previous and following alternatives to depression (the best 20 models by BIC in table 6.8). This also indicates that structural qualities of the questionnaire may explain some of the missingness.

In figure 6.4 there are listed all variables that are in the complete data model or their correlation with the target variable self-reported depression or its response indicator is stronger than 0.3. Here the strongest correlations with self-reported depression can be found in mental health related question as nervousness or low mood within the last 4 weeks and psychic overload (based on MHI-meter) but also in self-reported medication use (sedative or depression medication).

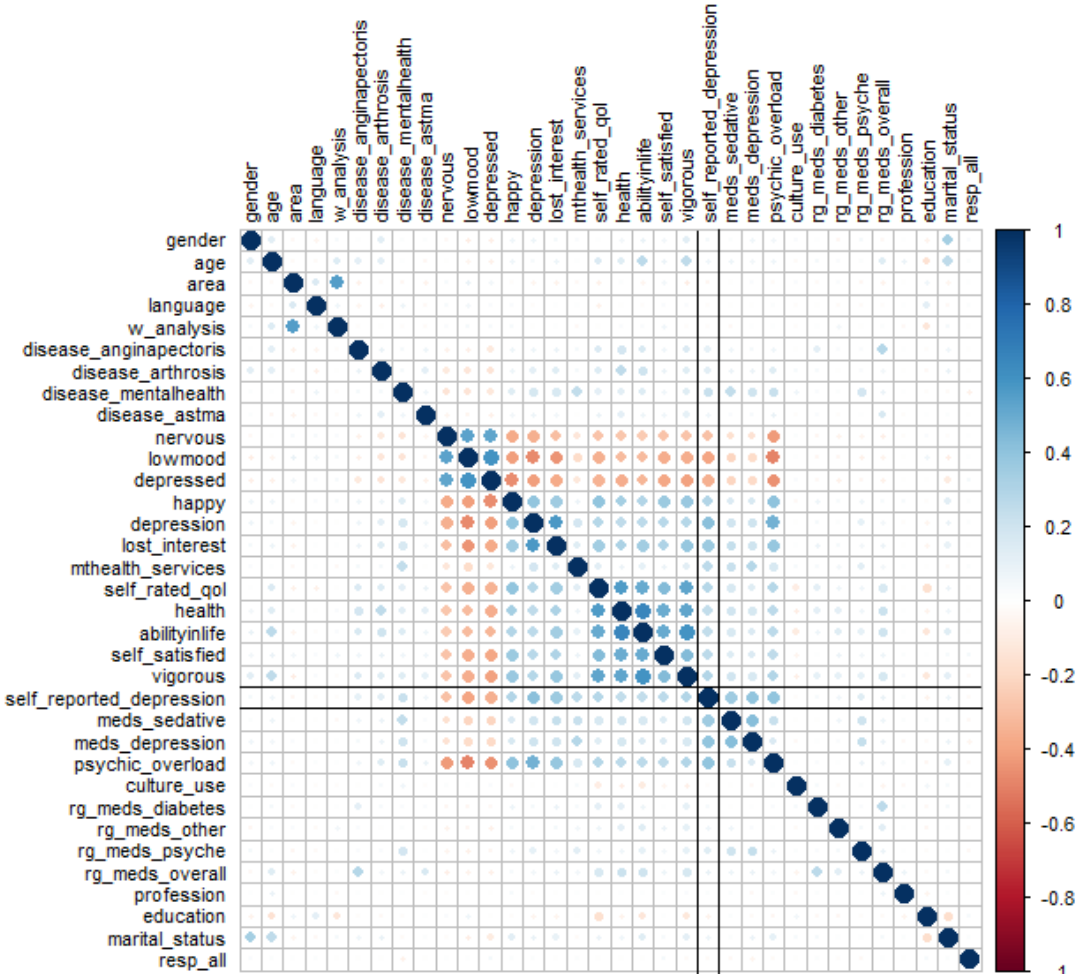


Figure 6.4: Correlations of possible predictor set. Larger circle size and darkness of the circle indicate higher correlation. Question texts of these variables can be found in appendix table 1.

Since also the predictors contain missing values it is important to check the proportion of usable cases in imputation. The proportion of usable cases is a fraction of predictor variable observations that are observed for the missing target variable(van

Table 6.8: Models for item nonresponse of self-reported depression including only variables available for all cases

AIC	BIC	deviance	df	model.name
1632	1944	1550	41	Main effects + Meds:psyche * resp_all + resp_previous * resp_following
1634	1945	1552	41	Main effects + Meds:psyche * resp_following + resp_previous * resp_following
1644	1948	1564	40	Main effects + resp_previous * resp_following
1637	1949	1555	41	Main effects + resp_previous * resp_following + resp_following * resp_all
1630	1949	1546	42	Main effects + Education * resp_all + resp_previous * resp_following
1641	1952	1559	41	Main effects + resp_previous * resp_following + resp_previous * resp_all
1643	1955	1561	41	Main effects + Meds:psyche * resp_previous + resp_previous * resp_following
1644	1956	1562	41	Main effects + Gender * resp_all + resp_previous * resp_following
1644	1956	1562	41	Main effects + Gender * Meds:other + resp_previous * resp_following
1645	1956	1563	41	Main effects + Gender * resp_following + resp_previous * resp_following
1645	1957	1563	41	Main effects + Meds:other * Meds:psyche + resp_previous * resp_following
1646	1957	1564	41	Main effects + Meds:other * resp_all + resp_previous * resp_following
1646	1957	1564	41	Main effects + Gender * Meds:psyche + resp_previous * resp_following
1646	1957	1564	41	Main effects + Gender * resp_previous + resp_previous * resp_following
1646	1957	1564	41	Main effects + Meds:other * resp_previous + resp_previous * resp_following
1646	1957	1564	41	Main effects + Meds:other * resp_following + resp_previous * resp_following
1640	1959	1556	42	Main effects + Education * resp_previous + resp_previous * resp_following
1644	1963	1560	42	Main effects + Meds:other * Education + resp_previous * resp_following
1644	1963	1560	42	Main effects + Meds:other * Answer_wave + resp_previous * resp_following
1645	1964	1561	42	Main effects + Gender * Education + resp_previous * resp_following

Main effects include age, study area, gender, education, profession, maritals status, language, answer wave (early/late), response indicator of previous (resp_previous) and following alternative (resp_following), response indicator of all the alternatives (resp_all, see figure 6.1), right to receive special reimbursement for psyche medication (Meds:psyche) and right to receive special reimbursement for medication other than psyche or diabetes medication (Meds:other).

Buuren and Groothuis-Oudshoorn, 2011). In table 6.9 there are listed the possible predictors that contain missing values. Here we see that alternatives on the same question as the target variable receive a very low proportion of usable cases (between 12 and 15 %) and leaving them out of the predictor set should have no great affect but improve efficiency. Out of scientific interest self-reported other mental health disease might be vital to keep in the predictor set but since its correlation with the target variable is only 0.3 we will leave it out in this case.

Table 6.9: Proportion of usable cases in the considered predictors for MI. Question texts of these variables can be found in appendix table 1.

predictor variable	**proportion of usable cases**
disease_mentalhealth	12.1
disease_astma	13.1
disease_anginapectoris	14.8
meds_depression	31.4
depression	33.6
lost_interest	34.4
nervous	34.4
meds_sedative	34.7
depressed	35.3
lowmood	35.8
happy	35.9
vigorous	39.3
culture_use	41.8
self_satisfied	42.2
health	43.5
abilityinlife	43.8
psychic_overload	44.1
mthealth_services	44.7
self_rated_qol	57.8
disease_arthrosis	91.2

The final predictor set contains 32 variables and predictors with missing values will not be imputed. Iterations was set to 20 and convergence of the mean and standard error seemed to be valid (figure 6.5). Imputation method used is logistic regression with 5 multiple imputations.

Results

Analysis of the imputed data set was carried out by calculating the mean of self-reported depression (0,1 values) for each study area and agegroup separatively in

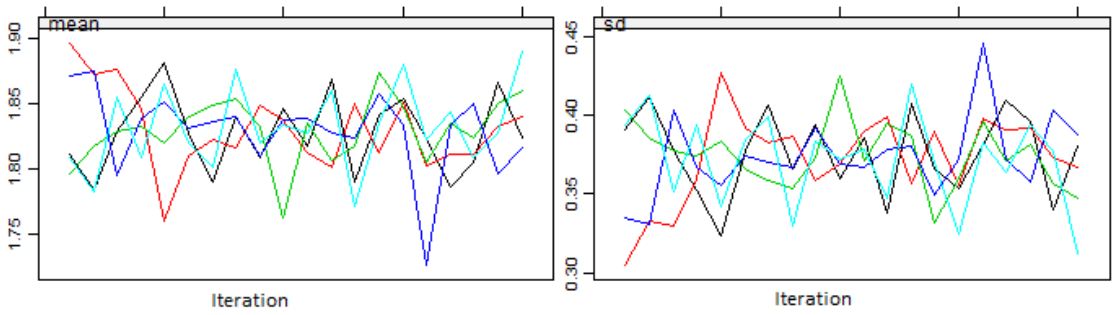


Figure 6.5: Convergence of imputed mean and standard error of self-reported depression.

each imputed data set. The final prevalence of self-reported depression is a weighted mean (with previously calculated IPW weights) of the prevalences in the imputed data sets.

Examining densities of the observed and imputed values shows that the densities of imputed values are higher on value 1 than the observed ones. This might indicate some problems in the imputation model but it also might indicate that nonresponse is associated with self-reported depression and that is why densities differ. In closer examination it shows in the agegroup 25 to 64- year-olds 89 % of the missing values was imputed to value 1 (to have self-reported depression) and in the agegroup of over 65s 93 %. This can also be seen in the result rates of areal self-reported depression in these two agegroups.

Table 6.10: The ratio of original values and the imputed values in the two age groups of interest.

	agegroup 25-64			**agegroup 65 and over**		
original values	imputed values		Total	imputed values		Total
	0	1		0	1	
0	7435	0	7435	3729	0	3729
	100%	0	87%	100%	0	69%
1	0	869	869	0	579	579
	0	100 %	10%	0	100%	11%
missing	32	252	284	78	1046	1124
	11%	89%	3%	7%	93%	21%

In rates of self-reported depression significant differences can be found in every area and agegroup between CC analysis and multiple imputed values. In the

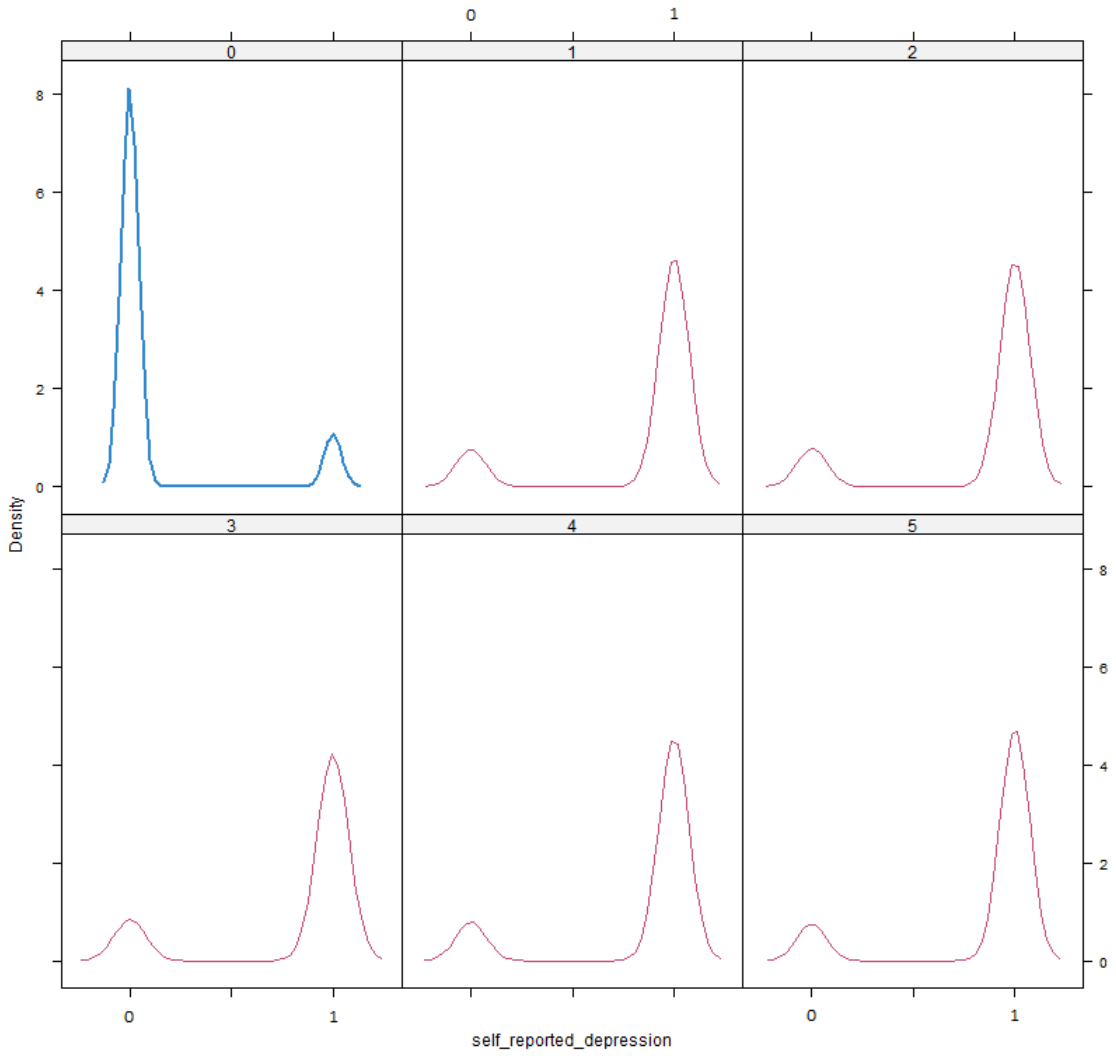


Figure 6.6: Density of observed and imputed values of self-reported depression in every imputation. Blue indicates the observed values and red indicates the imputed values.

agegroup of 25 to 64-year olds only in Turku the rates are not clearly significantly different. In over 65-year-old age group the differences are notable in every area, especially in Kainuu where the multiple imputed rate is twice as high as CC analysis based rate. Standard errors are constantly only slightly higher than in CC analysis. To conclude, multiple imputation based analysis implies that rates of self-reported depression are higher than CC analysis would suggest in every area studied especially in the age group of over 65-year-olds.

Closer examination of the β-estimates in the imputation model showed that self-reported depression medication use and the need of mental health services affect most the imputed values of self-reported depression, which is a reasonable result. The reason for MI resulting in significantly higher prevalences of self-reported depression might be that also respodents answers on nervousness, low or depressed mood, anxiety and vigorousness were significant predictors of imputed self-reported depression. Those feelings might not indicate depression by themselves although they were correlated with self-reported depression and using them as predictors may result in unreliable results in self-reported depression. This is something that should be studied more and also substance expertise on depression should be consulted when building the predictor set.

Table 6.11: Self-reported depression prevalences in the four study areas for MI and CC analysis

Area		**25-64-year-olds**	Std error	**over 65-year-olds**	Std error
Kainuu	CC	10.5 (9.2-11.7)	0.63	13.2 (11.4-15.0)	0.94
	MI	14.2 (12.8-15.7)	0.74	22.9 (20.9-24.9)	1.01
	Register*	9.9		10.9	
Northern Ostrobothnia	CC	8.3 (7.1-9.4)	0.59	12.4 (10.2-14.6)	1.12
	MI	11.2 (9.8-12.6)	0.72	20.1 (17.7-22.6)	1.25
	Register*	9.9		12.6	
Turku	CC	12.3 (11.0-13.6)	0.67	12.7 (10.8-14.7)	0.98
	MI	14.7 (13.2-16.3)	0.78	20.2 (18.0-22.3)	1.09
	Register*	11.3		11.4	
Finland	CC	10.7 (9.0-11.3)	0.85	10.0 (7.6-12.3)	1.20
	MI	13.3 (11.3-15.2)	0.99	16.8 (14.0-19.5)	1.38
	Register*	10		11.9	

** Register from Social Insurance Institution; Rate of people that received reimbursement of depression medication in 2010 (SOTKAnet Statistics and Indicator Bank, 2005 - 2013)*

MI changes the odds ratio area wise in Kainuu and overall in the agegroups of 60 and over. The same effect shows also in the marital status class widowed where there are supposedly older people. Also the low education group's OR is significantly higher than in CC analysis. This shows the same difference as the areal rates that age seems to be more significant aspect in self-reported depression than CC analysis suggests. Overall it seems that item nonresponse for self-rated depression seems to be related to the phenomenon of depression.

Table 6.12: Odds ratio estimates of self-reported depression with complete data model variables for MI and CC analysis.

	CC			**MI**		
	OR	**95 % CI**		**OR**	**95 % CI**	
Area Finland	1.00			1.00		
Area Kainuu	1.08	0.90	1.29	1.44	1.27	1.64
Area Northern Ostrobothnia	0.86	0.71	1.05	1.02	0.88	1.17
Area Turku	1.21	1.01	1.44	1.06	0.92	1.22
Education high	1.00			1.00		
Education low	1.27	0.66	2.44	2.27	1.87	2.75
Education moderate	1.46	0.78	2.71	0.83	0.69	1.01
Single	1.00			1.00		
Unknown	1.64	0.20	13.24	1.76	0.69	4.48
Married	1.17	0.77	1.77	0.94	0.84	1.06
Divorced	1.36	1.13	1.63	1.44	1.23	1.68
Widowed	1.22	0.76	1.95	2.96	2.57	3.41
Gender Male	1.00			1.00		
Gender Female	1.19	0.91	1.57	1.34	1.23	1.46
20s	1.00			1.00		
30s	0.76	0.40	1.47	0.93	0.71	1.23
40s	1.21	0.67	2.22	1.13	0.88	1.46
50s	1.10	0.61	1.98	1.46	1.14	1.85
60s	0.72	0.39	1.35	1.60	1.26	2.04
70s	0.95	0.51	1.77	3.30	2.61	4.17
80s	1.32	0.70	2.50	5.26	4.14	6.69
90s	0.87	0.16	4.86	4.71	3.10	7.17

6.4 Combined results and discussion

Overall the nonresponse bias corrected results of self-reported depression prevalences are quite similar in CC analysis, IPW and WSHDI. MI rates stand out significantly

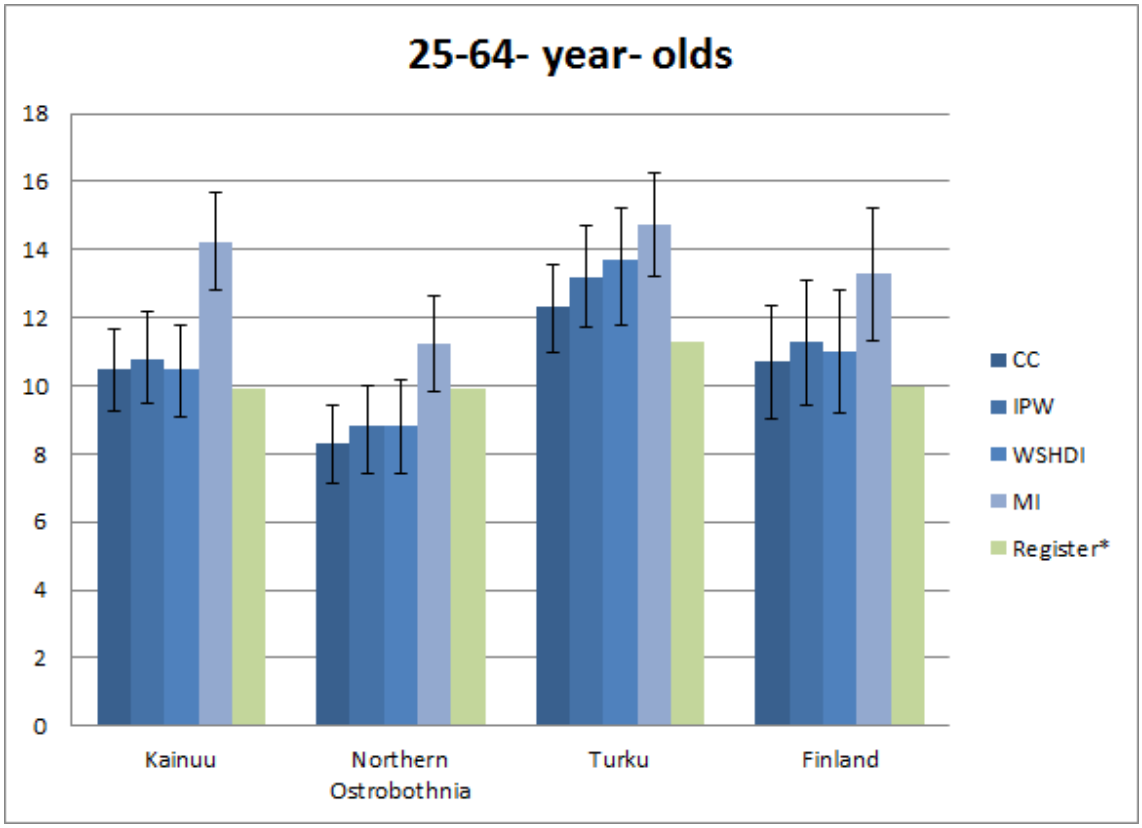


Figure 6.7: Rate estimates for self-reported depression in the agegroup of 25 to 64

especially in over 65-year-olds. All the methods seem to respect the relation between areas and although the rates differ the order stays the same. The highest rate of self-reported depression is in Turku in the agegroup of 25 to 64-year-olds and the lowest in Northern Ostrobothnia in the agegroup of 25 to 64-year-olds according to all the methods. Also the register based distribution of rate of people receiving reimbursement of depression medication supports the results distribution-wise. Some deviation from the general pattern is observable in the older agegroup in figure 6.8 since the register suggests no big differences between areas but the method distributions seem to suggest that the lowest rate is in Finland overall and the highest in Kainuu.

IPW is only used here to correct the unit nonresponse in order to get the background study variable distributions to match the original sample distributions. Although the weighted distributions are close to the original sample distributions IPW is not enough to cover the missing values in one questionnaire item and that is why rate estimates stay close to CC analysis based estimates. More weight was given to those more rare cases who participated the survey but their representation may

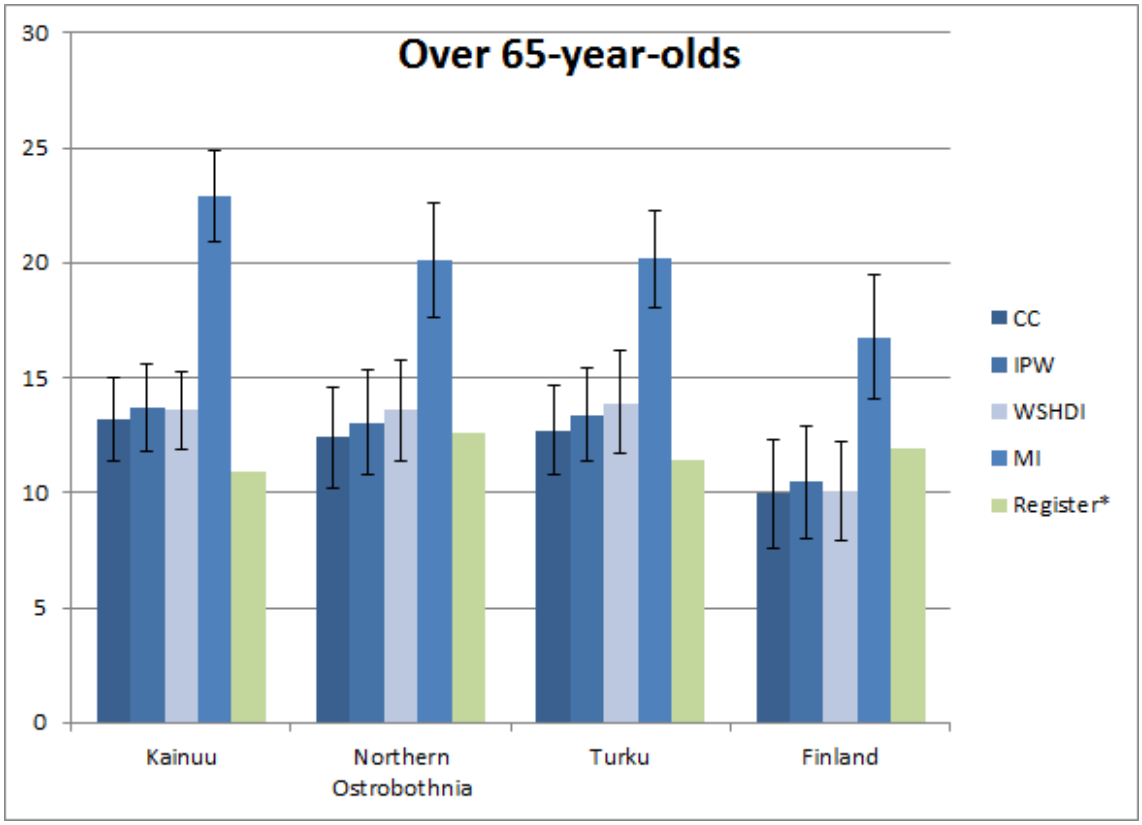


Figure 6.8: Rate estimates for self-reported depression in the agegroup of over 65

have been ineffectual in the item nonresponse case if they had not answered the question of interest. Weight was then given to the observed responses of the item and no consideration was given to the other predictors in the data set since they were not in the unit nonresponse model. More effective IPW would require model specification to the self-reported depression itself including predictors used in MI.

WSHDI was also a bit problematic like IPW since it can only use completely observed predictors. Here we used IPW weights to determine how many times a single donor could be used in the donor pool. Also information on the question structure (see figure 6.1) was applied here as a predictor since response indicators for the previous and following question of depression were available for all cases. Those response indicators were significant predictors for response of the depression so they provided vital information and also pointed out the need for statistical editing in nonresponse correcting. Still WSHDI did not bring out any great differences in the rate estimates compared to IPW and CC analysis and it was not able to impute all the missing values so some information for depression was left missing.

In MI also non-complete predictors were used to correct the nonresponse so that

information was gained for example from mood, self-rated health, quality of life and other mental health related questions. This resulted in significantly higher rate estimates especially in the older agegroup. These results may be biased and overestimate the depression rates in the older group since predictors included information about the mood in the last month, self-rated health and quality of life which are known to differ from the younger age group without signs of diagnosis-based depression (Saarela and Stenberg, 2011). However, the depression rates among the elderly are also known to be underdiagnosed since many factors like memory function and function ability overall may complicate the diagnosis so these multiple imputed rate estimates may reflect also that contradiction. In this case the imputation model may not preserve the connection of age and depression and for further consideration it would be useful to construct predictor sets separatively for the age groups and take into account the interactions between age and the predictors.

When studying these results it is important to keep in mind the original question setting which was that the self-reported depression was to be diagnosed or treated by a doctor. Many people may suffer from depression but have not been diagnosed at the time of answering the question and similarly many people may have momentarily felt low mood or depression which would here indicate depression but was never diagnosed. In some areas health stations might be crowded and that is why there are people just waiting to be diagnosed with depression which may result in higher rates especially in MI where the variables indicating depressive mood are used as predictors. Also the register based rates describe the situation from another point of view which is the reimbursement of depression medication. People may have been diagnosed with depression at the time of the survey but not yet received reimbursements or were denied to receive them and that is why the rates differ.

Overall the results indicate that CC analysis underestimates the rate of depression in every area studied. This indicates that respondents with depression may have been left outside the survey or not have responded the question. Correcting this nonresponse bias is vital especially if there is to be further analysis done with depression in the analysis model. For example analysis of the unmet need of mental health services would be biased if it only accounted for the complete cases. Overall the methods' results also suggest that the willingness to answer the depression

question depends on depression itself which means that the data is MNAR and the nonresponse corrected estimates may still be biased. This would be good to take into account in further applications. Results presented here are also supported by previous studies, for example Haapea (2010) has shown that respondents with a psychiatric disorder participated less actively on epidemiological studies than those without one. In the same study it was also shown that using non-response correcting methods such as IPW resulted in higher prevalences of psychiatric disorders than using only complete cases, which also supports the results in this study.

Table 6.13: Prevalences of self-reported depression in the four study areas according to different nonresponse methods and register prevalence of people receiving reimbursements of depression medication in 2010.

Area		**25-64-year-olds**	Std error	**over 65-year-olds**	Std error
Kainuu	CC	10.5 (9.2-11.7)	0.63	13.2 (11.4-15.0)	0.94
	IPW	10.8 (9.5-12.2)	0.68	13.7 (11.8-15.6)	0.95
	WSHDI	10.5 (9.1-11.8)	0.69	13.6 (11.9-15.3)	0.87
	MI	14.2 (12.8-15.7)	0.74	22.9 (20.9-24.9)	1.01
	Register*	9.9		10.9	
Northern Ostrobothnia	CC	8.3 (7.1-9.4)	0.59	12.4 (10.2-14.6)	1.12
	IPW	8.8 (7.5-10.0)	0.66	13.0 (10.7-15.3)	1.16
	WSHDI	8.8 (7.4-10.2)	0.70	13.6 (11.4-15.8)	1.12
	MI	11.2 (9.8-12.6)	0.72	20.1 (17.7-22.6)	1.25
	Register*	9.9		12.6	
Turku	CC	12.3 (11.0-13.6)	0.67	12.7 (10.8-14.7)	0.98
	IPW	13.2 (11.7-14.7)	0.76	13.4 (11.4-15.4)	1.02
	WSHDI	13.7 (11.8-15.5)	0.94	13.9 (11.7-16.2)	1.16
	MI	14.7 (13.2-16.3)	0.78	20.2 (18.0-22.3)	1.09
	Register*	11.3		11.4	
Finland	CC	10.7 (9.0-11.3)	0.85	10.0 (7.6-12.3)	1.20
	IPW	11.3 (9.4-13.1)	0.94	10.5 (8.9-13.9)	1.25
	WSHDI	11.0 (9.2-12.8)	0.91	10.1 (7.9-12.2)	1.09
	MI	13.3 (11.3-15.2)	0.99	16.8 (14.0-19.5)	1.38
	Register*	10		11.9	

** Register from Social Insurance Institution; Rate of people that received reimbursement of depression medication in 2010 (SOTKAnet Statistics and Indicator Bank, 2005 - 2013)*

Table 6.14: Odds ratio estimates of complete data model variables for different nonresponse methods used.

	CC			**IPW**			**WSHDI**			**MI**		
	OR	**95 % CI**		**OR**	**95 % CI**		**OR**	**95 % CI**		**OR**	**95 % CI**	
Area Finland	1.00			1.00			1.00			1.00		
Area Kainuu	1.08	0.90	1.29	1.04	0.86	1.26	1.05	0.87	1.27	1.44	1.27	1.64
Area Ostrobothnia	0.86	0.71	1.05	0.84	0.68	1.03	0.89	0.73	1.09	1.02	0.88	1.17
Area Turku	1.21	1.01	1.44	1.22	1.01	1.48	1.32	1.08	1.61	1.06	0.92	1.22
Education high	1.00			1.00			1.00			1.00		
Education low	1.27	0.66	2.44	1.51	1.15	1.98	1.72	1.26	2.36	2.27	1.87	2.75
Education moderate	1.46	0.78	2.71	0.92	0.71	1.20	1.11	0.82	1.50	0.83	0.69	1.01
Single	1.00			1.00			1.00			1.00		
Unknown	1.64	0.20	13.24	1.46	0.48	4.43	1.23	0.37	4.10	1.76	0.69	4.48
Married	1.17	0.77	1.77	0.66	0.57	0.78	0.55	0.46	0.66	0.94	0.84	1.06
Divorced	1.36	1.13	1.63	1.43	1.18	1.74	1.18	0.94	1.48	1.44	1.23	1.68
Widowed	1.22	0.76	1.95	1.38	1.12	1.69	1.33	1.06	1.66	2.96	2.57	3.41
Gender Male	1.00			1.00			1.00			1.00		
Gender Female	1.19	0.91	1.57	1.22	1.07	1.38	1.01	1.00	1.01	1.34	1.23	1.46
20s	1.00			1.00			1.00			1.00		
30s	0.76	0.40	1.47	0.99	0.72	1.35	0.91	0.64	1.29	0.93	0.71	1.23
40s	1.21	0.67	2.22	1.12	0.84	1.51	0.94	0.68	1.31	1.13	0.88	1.46
50s	1.10	0.61	1.98	1.32	1.00	1.74	1.07	0.79	1.46	1.46	1.14	1.85
60s	0.72	0.39	1.35	0.96	0.72	1.27	0.81	0.59	1.10	1.60	1.26	2.04
70s	0.95	0.51	1.77	1.29	0.97	1.72	1.12	0.82	1.53	3.30	2.61	4.17
80s	1.32	0.70	2.50	2.04	1.52	2.74	1.74	1.27	2.40	5.26	4.14	6.69
90s	0.87	0.16	4.86	1.98	1.10	3.56	1.96	1.05	3.66	4.71	3.10	7.17

Luku 7

Päätelmät

Tässä työssä kartoitettiin kyselynä kerätyn väestötutkimusaineiston kadon syitä, keinoja sen vähentämiseksi ja sen tilastollisia hallintamenetelmiä jo kerätyssä Alueellisen terveys- ja hyvinvointitutkimuksen vuoden 2010 aineistossa. Vastauskato oli tässä aineistossa selkeästi valikoitunutta, kuten useissa Suomessa lähiaikoina kerätyissä väestötutkimusaineistoissa, ja keskittyi selkeästi esimerkiksi nuorempiin ikäluokkiin ja matalamman koulutuksen saaneiden ryhmään. Tämä osoittaa, että vastauskato saattaa aiheuttaa harhaa aineistosta saataviin tuloksiin. Yksittäisenä esimerkkinä kadonhallintamenetelmiä käytettiin itse raportoituun masennukseen. Vertailukohdaksi saaduille tuloksille käytettävänä oli samankaltaisesta ilmiöstä kertovaa rekisteriperäistä tietoa depressiolääkkeistä korvauksia saaneista, jolloin voitiin vertailla lähinnä aineistosta saatujen osuuksien suhteita alueittain ikäryhmissä 25–64-vuotiaat ja yli 65-vuotiaat.

Tilastolliset kadonhallintamenetelmät lisäävät vastauskatoa sisältävän aineiston luotettavuutta. Tässä työssä pystyttiin osoittamaan, että kadonhallintamenetelmiä käyttämällä saatiin eroavia tuloksia, kuin pelkästään kokonaan havaittua aineistoa käyttämällä. Vastauskatoa pyrittiin korjaamaan painottamalla ja imputoimalla puuttuvaa tietoa sisältävää aineistoa. Saadut tulokset vaihtelivat menetelmittäin, suurin ero kokonaan havaitusta aineistosta saatuihin tuloksiin saatiin moniimputoinnilla, kun toisaalta esimerkiksi *Hot Deck*-imputoinnilla saadut tulokset olivat melko lähellä kokonaan havaitun aineiston tuloksia. Kaikissa käytetyissä menetelmissä keskivirheet kasvoivat verrattuna kokonaan havaitun aineiston keskivirheisiin. Tämä kertoo puuttuviin havaintoihin liittyvästä epävarmuudesta, mutta myös

siitä, että puuttuvien havaintojen huomioimatta jättäminen tuottaa virheellisiä vaihtelua kuvaavia tunnuslukuja.

Inverse probability weighting- painotusmenetelmällä (IPW) pystyttiin melko hyvin korjaamaan yksikkökadon aiheuttamia eroja taustamuuttujien jakaumissa verrattuna alkuperäisen otoksen jakaumiin. Painotetut tulokset eivät kuitenkaan merkittävästi eronneet esimerkiksi nuoremmassa 25–64-vuotiaiden ikäryhmässä korjaamattoman aineiston tuottamista tuloksista itse raportoidun masennuksen tapauksessa, josta voisi päätellä että tarkemmat tulokset vaatisivat tarkempaa mallimäärittelyä juuri itse raportoidun masennuksen suhteen. Muiden muuttujien suhteen, joissa puuttuvuus riippuu vain otosmuuttujista, painottamalla saadaan varmasti luotettavampia estimaatteja kuin tässä tapauksessa käytetystä esimerkkimuuttujasta.

Hot Deck-imputoinnissa saatiin itse raportoidulle masennukselle hyvin samankaltaisia tuloksia kummassakin ikäryhmässä kuin kokonaan havaitusta aineistosta ja painotetusta aineistosta. Tähän yksi syy on se, että imputointiin käytettiin vain kokonaan havaittuja muuttujia, eli samoja muuttujia kuin painottaessa ja jo luotuja analyysipainoja käytettiin rajoittamaan yksittäisen luovuttajan mahdollisia luovutuskertoja imputoidessa. *Hot Deck*- imputoinnista huolimatta aineistoon jäi myös jonkin verran puuttuvaa tietoa, eli menetelmässä olisi vielä tarpeen tarkentaa mallia ja kokeilla erilaisia malleja kokonaisen imputoidun aineiston aikaansaamiseksi.

Moni-imputoinnissa (MI) sen sijaan pystyttiin käyttämään myös osittain havaittuja muuttujia imputointimallissa, joka näkyi myös tuloksissa. Erityisesti yli 65-vuotiaiden estimoitu osuus itse raportoidun masennuksen osalta oli huomattavasti korkeampi kuin muilla menetelmillä. Koska erot muihin menetelmiin olivat vanhemmassa ikäryhmässä suuret, mallin muuttujia tulisi vielä tarkastella lähemmin erityisesti vanhemman ikäryhmän osalta. Tässä yhteydessä tuli esille, että itse raportoidun masennuksen puuttuneisuus saattaa täysin ehdollista eli toisin sanoen valikoitunutta tiettyjen vastaajaryhmien suhteen ja sen tutkiminen tarkemmin olisi tarpeen. Verrattuna rekisteriperusteiseen depressiolääkkeistä vuonna 2010 korvauksia saaneiden jakaumaan, mikään menetelmistä ei juuri muuttanut tutkimusalueiden tai ikäryhmien tuloksien suhteita toisiinsa vaan ne pysyivät hyvin samankaltaisina. Tästä voisi päätellä, että jokainen menetelmä toimi kadon aiheuttaman harhan korjaamisessa mahdollisuuksien mukaan. Tulee kuitenkin huomioida, että itse raportoidun

masennuksen tuli olla kysymysasettelunkin mukaan lääkärin hoitama tai diagnosoima, ja kaikki masennusdiagnoosin saaneet eivät välttämättä saa depressiolääkkeistä korvauksia, joka näkyy rekisteristä ja aineistosta saatujen osuuksien eroina.

Kyselyaineiston lisäksi vastauskadon hallinnassa käytettiin apuna Väestörekisterikeskusta, Tilastokeskuksesta ja Kansaneläkelaitokselta saatavia rekisteritietoja esimerkiksi siviilisäädystä, koulutuksesta, ammatista ja lääkkeiden erityiskorvausoikeudesta tutkimusvuonna. Nämä rekisteritiedot olivat saatavilla koko alkuperäiselle otokselle, jolloin saatiin tärkeää tietoa myös kokonaan tutkimuksesta pois jääneistä vastaajista. Rekisteritietojen käyttö apuna kadon hallintamenetelmissä paransi vastauskatomallien BIC-kerrointa, eli rekisteritiedot paransivat mallien sopivuutta vastauskadon selittämiseen. Tämä merkitsee sitä, että koko otokselle saatavissa olevien rekisteritietojen käyttö on hyödyllistä kyselyaineiston kadon hallintamenetelmissä ja lisää hallintamenetelmillä saatavien tulosten luotettavuutta.

Jatkotutkimuksen kohteena voisi olla moni-imputointimallin tarkempi tutkiminen ja sen muodostaminen ikäryhmille erikseen. Tässä aineistossa voisi toimia myös *doubly robust* moni-imputointi, jossa määritetään kaksi eri mallia, toinen puuttuville arvoille ja toinen puuttuvien arvojen todennäköisyydelle (mm. Long et al. (2012)). Myös muiden Bayes-perusteisten menetelmien käyttö voisi tulla tässä tapauksessa kyseeseen. Tässä työssä jäivät käsittelemättä myös malliperusteiset kadon hallintamenetelmät, kuten suurimpaan uskottavuuteen ja EM-algoritmiin perustuvat menetelmät, joten nekin olisivat hyviä jatkotutkimuksen kohteita. Työn alkuosassa esiteltyjen aikaisten ja myöhäisten vastaajien erot voisivat olla myös yksi mielenkiintoinen jatkotutkimuksen kohde, sillä myöhäisten vastaajien on todettu edustavan jollain tapaa myös katoon jääneitä vastaajia (Peress, 2010).

Vastauskadon huomioiminen ja sen hallintamenetelmät ovat tärkeä osa väestötutkimusaineistojen käyttöä, jolloin kadon huomioimisen tärkeys ja sen hallintamenetelmien käyttökelpoisuus olisi hyvä olla tiedossa myös muilla tutkimusryhmän jäsenillä kuin tilastotieteen osaajilla. Selkeästi dokumentoidut otantamenetelmät, mahdollisen kadon vähentämiseen pyrkiminen jo tiedonkeruuvaiheessa ja hyvät koko otokselle saatavissa olevat rekisteritiedot omalta osaltaan edesauttavat tehokkaiden kadonhallintamenetelmien käyttöä, joilla saadaan luotettavia väestöön yleistettäviss olevia tuloksia vastauskadon yhä mahdollisesti kasvaessa väestötutkimuksissa.

Kirjallisuutta - References

P. Allison. *Missing data*. Sage, Thousand Oaks, 2001. ISBN 0-7619-1672-5.

P. Allison. Handling missing data by maximum likelihood. *SAS Global Forum, Paper 312-2012*, 2012.

R. Andridge and R. Little. A review of hot deck imputation for survey nonresponse. *International Statistical Review*, 78(1):40–64, 2010. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2010.00103.x. URL http://dx.doi.org/10.1111/j.1751-5823.2010.00103.x.

J. Carpenter, J. Bartlett, and M. Kenward. www.missingdata.org.uk. http://missingdata.lshtm.ac.uk/, 2014. ESRC Research Methods Programme.

B. Cox. The weighted sequential hot deck imputation procedure. *ASA Proc Section on Survey Res Methods*, pages 721–726, 1980. URL http://www.amstat.org/sections/srms/proceedings/papers/1980_152.pdf.

D. Dillman. *Mail and Internet surveys : the tailored desing method*. Wiley & Sons, New York, NY, 2000. ISBN 0-471-32354-3 (sid.).

W. Dunkelberg and G. Day. Nonresponse bias and callbacks in sample surveys. *Journal of Marketing Research*, 10, 1973.

G. Durrant. Imputation methods for handling item-nonresponse in the social sciences: A methodological review. *National Centre for Research Methods Working Paper Series*, 2005.

C. Enders. Missing data mechanisms. http://msass.case.edu/downloads/research/missing_data_mechanisms.pdf, 2013. Mandel School of Applied Social Sciences.

Eurostat. Survey sampling reference guidelines: Introduction to sample design and estimation techniques, 2008.

R. Groves, D. Dillman, J. Eltinge, and R. Little. *Survey nonresponse*. John Wiley & Sons,, New York, 2002. ISBN 0-471-39627-3 (sid.).

M. Haapea. *Non-response and information bias in population-based psychiatric research. The Northern Finland 1966 Birth Cohort Study*. PhD thesis, University of Oulu, 2010.

D. Hawkins. Estimation of nonresponse bias. *Sociological Methods and Research*, 3, 1975.

D. Heitjan. Ignorability in general incomplete-data models. *Biometrika*, 81 (4):701–708, 1994.

J. Higgins and S. Green. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, 2011.

D. Howell. Treatment of missing data. http://www.uvm.edu/~dhowell/StatPages/More_Stuff/Missing_Data/Missing.html, 2012.

R. Kaikkonen, J. Murto, T. Koskela, E. Virtala, T. Härkänen, T. Koskenniemi, E. Vartiainen, and S. Koskinen. Alueellinen terveys- ja hyvinvointitutkimus, regional health and wellbeing study. www.thl.fi/ath, 2010a.

R. Kaikkonen, J. Murto, T. Koskela, E. Virtala, T. Härkänen, T. Koskenniemi, E. Vartiainen, and S. Koskinen. Alueellisen terveys- ja hyvinvointitutkimuksen tutkimuslomakeluettelo - question forms 2010. http://www.terveytemme.fi/ath/lomakkeet/2010/A1004-3_ENGLANTI_20-54_EN.pdf, 2010b.

R. Kaikkonen, J. Murto, T. Koskela, E. Virtala, T. Härkänen, T. Koskenniemi, E. Vartiainen, and S. Koskinen. Alueellisen terveys- ja hyvinvointitutkimuksen tutkimusalueet ja vastausaktiivisuus 2010-2011. `http://terveytemme.fi/ath/tulokset/notes/ath_alueet_2010-2011.htm`, 2010c.

G. Kalton and I. Flores-Cervantes. Weighting methods. *Journal of Official Statistics*, 19(2):81–97, 2003.

G. Kalton and D. Kasprzyk. The treatment of missing survey data. In *Survey Methodology*. 1986.

A. Koyama, R. Fukunaga, Y. Abe, Y. Nishi, N. Fujise, and Ikeda M. Item non-response on self-reported depression screening questionnaire among community-dwelling elderly. *Journal of affective disorders*, 162(1):30–33, 2014. doi: 10.1016/j.jad.2014.03.022.

V. Kuusela. Otantamenetelmä on surveytutkimuksen kulmakivi. *Tilastokeskus: Hyvinvointikatsaus*, 4, 2009.

S. Laaksonen. *Painotusmenetelmät*. Luentokurssi, 2009. Helsingin Yliopisto.

S. Laaksonen. *Surveymetodiikka*. Ventus Publishing ApS, 2010. ISBN 978-87-7681-724-4.

S. Laaksonen, R. Lehtonen, and K. Vehkalahti. Topics in survey methodology and survey analysis. Lecture course, 2013. Helsingin Yliopisto.

R. Lehtonen and E. Pahkinen. *Practical methods for design and analysis of complex surveys*. Wiley, Chichester, 2004. ISBN 0470091630 (e-book).

S. Lipsitz, J. Ibrahim, and L. Zhao. A weigthed estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160, 1999.

R. Little and D. Rubin. *Statistical analysis with missing data*. John Wiley, Hoboken, NJ, 2002. ISBN 0-471-18386-5 (sid.).

Q. Long, C. Hsu, and Li Y. Doubly robust nonparametric multiple imputation for ignorable missing data. *Statistica Sinica*, 22:149–172, 2012.

D. Marker, D. Judkins, and M. Winglee. Large-scale imputation for complex surveys. In *Survey Nonresponse*. John Wiley and Sons, 1999.

G. Molenberghs and M. Kenward. *Missing Data in Clinical Studies*. Wiley, 2007.

P. Patrician. Multiple imputation for missing data. *Research in Nursing & Health*, 25(1):76–84, 2002. doi: 10.1002/nur.10015.

M. Peress. Correctiong for survey nonresponse using variable response prospensity. *Journal of the American Statistical Association*, 2010.

K. Pohjanpää. Tilastoinnin merkityksen ymmärtäminen lisää tutkimuksiin osallistumista. *Tilastokeskus: Hyvinvointikatsaus*, 4, 2010.

D. Posada and T. Buckley. Model selection and model averaging in phylogenetics: Advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808, 2004. doi: 10.1080/10635150490522304. URL `http://sysbio.oxfordjournals.org/content/53/5/793.abstract`.

R. Prättälä and H. Tolonen. Ketkä todellisuudessa vastaavat terveyskyselyihin? `http://demo.seco.tkk.fi/tervesuomi/item/ktl:12394`, 2007.

A. Rotnitzky and J. Robins. Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82(4):805–820, 1995.

D. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

D. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley-Interscience, Hoboken, N.J., 2004. ISBN 0-471-65574-0.

T. Saarela and J. Stenberg. Kun mikään ei kelpaa vanhukselle - taustalla persoonallisuushäiriö? *Lääketieteellinen Aikakauskirja Duodecim*, 127(4):397–405, 2011.

J. Schafer and M. Olsen. Multiple imputation for multivariate missing-data problems: a data analyst's perspective. *Multivariate Behavioral Research*, 33 (4):545–571, 1998.

S. Seaman and I. White. Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, 22(3):278–295, 2013. doi: 10.1177/0962280210395740.

S. Seaman, J. Galati, D. Jackson, and J. Carlin. What is meant by "missing at random"? *Statistical Science*, 28(2): 257–268, 2013. doi: 10.1214/13-STS415.

Social Insurance Institute. Special reimbursements for medicines. `http://www.kela.fi/web/en/reimbursements-for-medicine-expences-special-reimbursement`, 2014.

SOTKAnet Statistics and Indicator Bank. *Depressiolääkkeistä korvauksia saaneet, % vastaavanikäisestä väestöstä*. `http://uusi.sotkanet.fi/taulukko/zW2/111/7/3A/0/`, 2005 - 2013.

J. Suvisaari, T. Aalto-Setälä, A. Tuulio-Henriksson, T. Härkänen, SI. Saarni, M. Perälä, J. Schreck, A. Castaneda, J. Hintikka, S.and Latvala A. Kestilä, L. Lähteenmäki, S. Koskinen, M. Marttunen, H. Aro, and J. Lönnqvist. Mental disorders in young adulthood. *Psychological Medicine*, 39:287–299, 2009. doi: doi:10.1017/S0033291708003632.

T. Therneau, B. Atkinson, and B. Ripley. *rpart: Recursive Partitioning and Regression Trees*, 2014. URL `http://CRAN.R-project.org/package=rpart`. R package version 4.1-8.

Tilastokeskus. Tulonjakotilaston menetelmäseloste 2009. `http://www.stat.fi/til/tjt/2009/tjt_2009_2011-05-20_men_001.html`, 2009.

H. Tolonen. *Towards the High Quality of Population Health Surveys*. Publications of National Public Health Institute, 2005. ISBN 951-740-547-2.

A. Tsiatis. *Semiparametric theory and missing data*. Wiley-Interscience, Hoboken, N.J., 2004. ISBN 978-0-387-37345-4.

S. Van Buuren. Software for multiple imputation. `http://www.stefvanbuuren.nl/mi/Software.html`, 2012.

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011. URL `http://www.jstatsoft.org/v45/i03/`.

L. Voigt, T. Koepsell, and J. Daling. Characteristics of telephone survey respondents according to willingness to participate. *American Journal of Epidemiology*, 157, 2003.

Table 1: Question texts of the possible predictors used in multiple imputation in chapter 5.4

predictor variable	**Question text/ other information**
disease_mentalhealth	Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months? other mental health problem
disease_astma	Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months? Asthma
disease_anginapectoris	Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months? coronary disease, angina pectoris (=chest pain under physical strain)
meds_depression	Have you used any of the following types of medicines over the past 7 days? anti-depressants
depression	Over the past 12 months, have you ever had a period of two weeks or more when for most of the time you have felt: down, melancholic or depressed
lost_interest	Over the past 12 months, have you ever had a period of two weeks or more when for most of the time you have felt: have you ever had a period of two weeks or more when for most of the time you have felt: that you have lost your interest in most things that usually give you pleasure
nervous	Over the past 4 weeks, for how much of the time have you felt: very nervous
meds_sedative	Have you used any of the following types of medicines over the past 7 days? sedatives
depressed	Over the past 4 weeks, for how much of the time have you felt: downhearted and sad
lowmood	Over the past 4 weeks, for how much of the time have you felt: so down in the dumps that nothing could cheer you up
happy	Over the past 4 weeks, for how much of the time have you felt: happy
vigorous	Do you have: enough energy for everyday life
culture_use	Average of culture use
self_satisfied	How satisfied are you with: Yourself
health	How satisfied are you with: your health
abilityinlife	How satisfied are you with: your ability to perform your daily living activities
psychic_overload	Psychic overload according to MHI-5-meter
mthealth_services	Do you feel you have been adequately provided with the following social and health care services over the past 12 months? mental health services
self_rated_qol	How would you rate your quality of life?
disease_arthrosis	Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months? Arthrosis
rg_meds_all	Right to receive special reimbursements of any medication
rg_meds_other	Right to receive special reimbursements of any medication other than diabetes and psyche medication
rg_meds_diabetes	Right to receive special reimbursements of diabetes medication
rg_meds_psyche	Right to receive special reimbursements of psyche medication
resp_previous	Have the person responded to the alternative before depression (chronic bronchitis, emphysema)
resp_following	Have the person responded to the alternative after depression (other mental health problem)
resp_all	Have the person responded to all the alternatives in this question
self_reported_depression	Have you had any of the following conditions diagnosed or treated by a doctor over the past 12 months? depression