TAMPEREEN
YLIOPISTO

PRO GRADU -TUTKIELMA

Matematiikan, tilastotieteen ja filosofian laitos
Tilastotiede
Syyskuu 2005

ANTTI LISKI

Lonkkamurtumapotilaiden hoitokustannusten vertailu vastaavuuspistemäärään perustuvalla menetelmällä

Tampereen yliopisto

Matematiikan, tilastotieteen ja filosofian laitos

LISKI, ANTTI: Lonkkamurtumapotilaiden hoitokustannusten vertailu vastaavuuspistemäärään perustuvalla menetelmällä

Pro gradu -tutkielma, 59 s., 8 liites.

Tilastotiede

Syyskuu 2005

Tiivistelmä

Lonkkamurtumat ovat yksi yleisimmistä vanhusten elämänlaadun huononemiseen johtavista syistä ja niiden hoidon kansantaloudelliset vaikutukset ovat merkittävät. Hoidon kustannukset ovatkin eräs tärkeä mittari hoitokäytäntöjä vertailtaessa. Tässä tutkielmassa kustannuksia vertaillaan sairaanhoitopiirien tasolla, koska suurin osa lonkkamurtumien hoidoista muodostuu erilaisista jatkohoidoista, jotka eivät ole sairaalakohtaisia.

Tutkimusaineisto, joka on saatu yhdistelemällä ja muokkaamalla useiden eri rekisterien tiedoista, koostuu kaikista tietyt kriteerit täyttävistä ensimmäisen kerran lonkkansa murtaneista potilaista (16 881 tapausta) vuosilta 1998 - 2001. Työssä kehitellään vastaavuuspistemäärän (propensity score) käyttöön perustuvaa tutkimustekniikkaa terveydenhuollon yksiköiden kustannusten vertailuun sopivaksi menetelmäksi, mikä on uusi lähestymistapa tällä sovellusalueella. Teoreettisen kehittelyn rinnalla menetelmää sovelletaan Helsingin ja Uudenmaan, Varsinais-Suomen ja Satakunnan sairaanhoitopiirien vertailuun.

Havaitaan, että erityisesti Helsingin ja Uudenmaan sairaanhoitopiirin ja Varsinais-Suomen sairaanhoitopiirin kustannukset poikkeavat merkittävästi toisistaan. Menetelmä tarjoaa hyvän lähtökohdan myös kustannuksiltaan erilaisten ryhmien potilasrakenteen tarkasteluun. Työssä tutkitaan myös usean sairaanhoitopiirin samanaikaiseen vertailuun soveltuvaa yleistettyä vastaavuuspistemäärää. Lisäksi tarkastellaan potilaiden ryhmittelyä kustannusjakauman perusteella.

Tässä esitetyillä menetelmillä voidaan toteuttaa suoraviivaisesti Suomen kaikkien sairaanhoitopiirien väliset kustannusvertailut. Saatujen tulosten tulkitseminen hoitokäytäntöjen ja hoidon tulosten kannalta on mielenkiintoinen jatkoaihe. Näkyvätkö kustannuserot esimerkiksi hoitotuloksissa? Monista kiinnostavista jatkotutkimusten aiheista mainittakoon esimerkiksi yleistetty vastaavuuspistemäärä, ydinestimaattorin harha ja vastaavuuspistemäärän käyttö ryhmittelyssä.

Hakusanat hoitokäytäntö, kustannusjakauma, rekisteriaineisto, ydinestimointi, yleistetty vastaavuuspistemäärä

Sisältö

1 Johdanto

Väestön ikääntyminen on vaativa haaste terveydenhuollolle. Kun väestö ikääntyy, hoitotarpeet kasvavat. Lääketieteen kehittyessä myös hoitomahdollisuudet paranevat, mutta taloudelliset voimavarat asettavat käytännön hoitotoiminnalle omat rajoitteensa. Tässä tilanteessa korostuu tarve kehittää hoitokäytäntöjä entistä tehokkaammiksi ja vaikuttavammiksi myös resurssien käytön kannalta.

Hoidon kustannukset ovat eräs tärkeä mittari hoitokäytäntöjä vertailtaessa. Jos oletetaan hoidon tavoitteet kaikkialla samoiksi, niin hoidon tuloksia ja käytettyjä resursseja voidaan vertailla hyvien hoitokäytäntöjen löytämiseksi. Lonkkamurtumien hoidossa tavoitteena on palauttaa potilaan toimintakyky murtumaa edeltäneelle tasolle. Jos jokin terveydenhuollon yksikkö saavuttaa tavoitteensa muita pienemmillä kustannuksilla, voidaan sen hoitokäytännöistä ehkä ottaa oppia.

Kustannusten käyttö hoitokäytäntöjen arvioinnissa on kuitenkin ongelmallista, koska jo pelkkä kustannusten vertailu on osoittautunut käytännössä hankalaksi. Tässä työssä kustannusvertailut perustuvat laajojen rekisteriaineistojen hyväksikäyttöön, mikä asettaa tutkimusotteelle omat erityisvaatimuksensa. Kun sairaanhoitopiirien hoitokäytäntöjä vertaillaan kustannusten avulla, pitäisi kustannuserojen luonnollisesti johtua ensisijaisesti hoitokäytäntöjen välisistä eroista. Kustannusten suora vertailu voi kuitenkin olla harhaanjohtavaa, sillä esimerkiksi piirien potilasaineksen vaihtelu saattaa peittää hoitokäytännöistä johtuvat kustannuserot.

Tutkielmassani esitän vastaavuuspistemäärän (propensity score) käyttöön perustuvan menetelmän, joka soveltuu terveydenhuollon yksiköiden kustannusten vertailuun. Menetelmää, joka on tällä sovellusalueella uusi lähestymistapa, käytetään eri sairaanhoitopiirien lonkkamurtumapotilaiden kustannusten vertailuun. Tällä menetelmällä voidaan yksinkertaisella tavalla kontrolloida muun muassa potilasaineksen systemaattisista eroista johtuvaa sekoittavaa vaikutusta kustannuseroihin. Näin piirien keskimääräiset kustannuserot ovat tulkittavissa entistä luotettavammin. Tässä työssä vertaillaan Helsingin- ja Uudenmaan sairaanhoitopiirin (HUS), Varsinais-Suomen sairaanhoitopiirin ja Satakunnan sairaanhoitopiirin hoitokustannuksia keskenään. Samalla menetelmällä voidaan vertailla myös muiden piirien kustannuksia, mutta laajuutensa vuoksi muiden vertailujen tuloksia ei esitetä tässä työssä.

1.1 Lonkkamurtumien hoito ja kustannukset

1.1.1 Yleistä

Lonkkamurtumat ovat yksi yleisimmistä vanhusten elämänlaadun huononemiseen johtavista syistä. Koska lonkkamurtumien hoito on kallista ja vanhusten osuus väestöstä kasvaa, tulevat sekä lonkkamurtumapotiladen määrä, että hoidon kansantaloudelliset vaikutukset kasvamaan entisestään. Vanhuksilla on yleensä myös muita vaivoja, jotka lisäävät hoitokustannuksia.

Lonkkamurtuma itsessään ei yleensä aiheuta pysyvää elämänlaadun alenemista tai kuolemaa, mutta lonkkamurtumasta saattaa alkaa hoitokierre, joka johtaa laitostumiseen tai jopa kuolemaan. Kun lonkka murtuu, iäkkäillä potilailla yleensä esiintyvät muut terveysongelmat heikentävät potilaan kuntoa ja hidastavat toipumista. Jos toipuminen pitkittyy, kunto jää pitkäksi aikaa alhaiselle tasolle ja potilas altistuu uusille vaivoille. Ennen murtumaa täysin omatoimisista vanhuksista puolet jää osittain ja kolmannes täysin riippuvaiseksi ulkopuolisesta avusta (Narinen ym. 2001).

Alle 50 -vuotiailla lonkkamurtumien vammamekanismi on yleensä suurienergiainen: murtuma on syntynyt esimerkiksi liikennetapaturmassa tai putoamisen yhteydessä. Yli 50 -vuotiaiden lonkkamurtumat syntyvät yleensä pienienergiaisesti liukastumisen, kompastumisen tai vuoteesta putoamisen yhteydessä (Lüthje 1984.) Yleisimmät lonkkamurtuman riskiä kasvattavat syyt ovat alentunut toimintakyky ja heikentynyt lihaskunto (Marks ym. 2003).

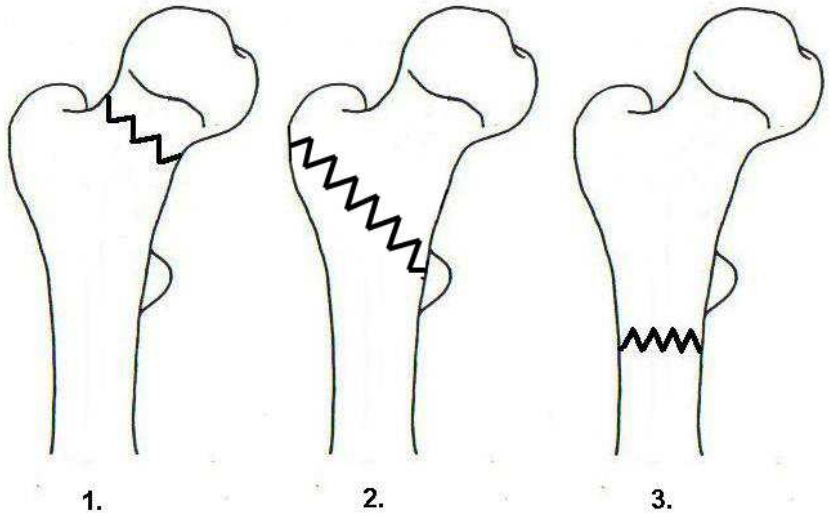
Lonkkamurtuman hoito tapahtuu pitkän ajan kuluessa useissa terveydenhuollon ja sosiaalihuollon toimipisteissä. Hoito alkaa yleensä siitä, kun potilas tulee päivystyspoliklinikan kautta sairaalaan. Käytännössä lähes kaikki lonkkamurtumapotilaat leikataan. Leikkauksen jälkeen potilas viettää noin viikon sairaalassa, minkä jälkeen kuntoutus jatkuu terveyskeskuksessa tai mahdollisesti vanhainkodissa. Hoidon tavoitteena on palauttaa potilas ennen murtumaa olleelle omatoimisuuden tasolle.

Yleinen suositus on, että potilas tulisi leikata mahdollisimman pian, mikäli kunto sen sallii. Nopea leikkaukseen pääsy on tärkeää erityisesti huonokuntoisille potilaille. Sund ja Liski (2005) osoittivat, että hyvin huonokuntoisten nopeasti leikattujen (alle kolme yötä murtumasta) potilaiden kuolleisuus oli alhaisempi kuin leikkausta kauemmin (vähintään kolme yötä murtumasta) odottaneiden hyvin huonokuntoisten potilaiden kuolleisuus.

Lonkkamurtuma tarkoittaa reisiluun yläosan murtumaa. Reisiluu voi murtua useasta eri kohdasta ja murtuman tyyppi määritellään sijainnin perusteella. Tässä työssä erotellaan toisistaan kolme murtumatyyppiä: reisiluun kaulan murtumat, trokanteeriset murtumat ja subtrokanteeriset murtumat. Nämä murtumatyypit on esitetty Kuviossa 1.1.

1.1.2 Lonkkamurtumapotilaat hoitoilmoitusrekisterissä

Sosiaali- ja terveydenhuollon hoitoilmoitusrekisterit (ks. esimerkiksi Sund 2000) ovat Stakesin ylläpitämiä valtakunnallisia rekistereitä, joissa on yksilötason tie-



Kuva 1.1. Murtumatyypit. 1. Reisiluun kaulan murtuma. 2. Trokanteerinen murtuma. 3. Subtrokanteerinen murtuma.

toja potilaista. Rekisterit sisältävät tietoa sosiaali- ja terveydenhuollon laitoshoidosta, asumispalveluista, säännöllisestä kotihoidosta sekä päiväkirurgiasta. Rekisteröitäviä muuttujia ovat muun muassa asiakkaan tunnistetiedot (henkilötunnus), ikä, sukupuoli, kotikunta, hoitoaika, diagnoosi- ja toimenpidetiedot sekä tiedot hoitoon lähettävästä tahosta, hoitopaikasta sekä jatkohoitopaikasta (Sund 2000.)

Tilastokeskus pitää yllä kuolinsyyrekisteriä, johon kirjataan kaikki Suomessa kuolleet. Kuolinsyyrekisteristä saatavia muuttujia ovat muun muassa kuolinpäivä, kuolinsyy ja kuolinpaikka. Myös kuolinsyyrekisterissä henkilöillä on henkilökohtaisena tunnisteenaan henkilötunnus.

Käytännössä kaikki lonkkamurtumat löytyvät terveydenhuollon hoitoilmoitusrekistereistä. Lonkkamurtumat ovat usein hyvin kivuliaita ja vaativat aina sairaalahoitoa. Siksi lähes jokainen murtuma tulee merkityksi rekisteriin. Vaikka kaikista murtumista on rekistereissä merkintä, tutkimusaineistoon kelpaavien murtumien identifiointi ei ole täysin yksiselitteistä.

Työssä tutkitaan uusia lonkkamurtumatapauksia. Siksi aineistosta on jätetty pois potilaat, joilla on ollut lonkkamurtuma tarkasteluvuotta edeltäneen kymmenen vuoden aikana. Aineiston tuottaminen rekistereistä on ollut varsin suuri ja tarkkuutta vaativa prosessi. Sund (2005) on kertonut myös tässä työssä käytettävän aineiston tuottamisesta ja tehdyistä valinnoista.

1.1.3 Tutkimusaineisto

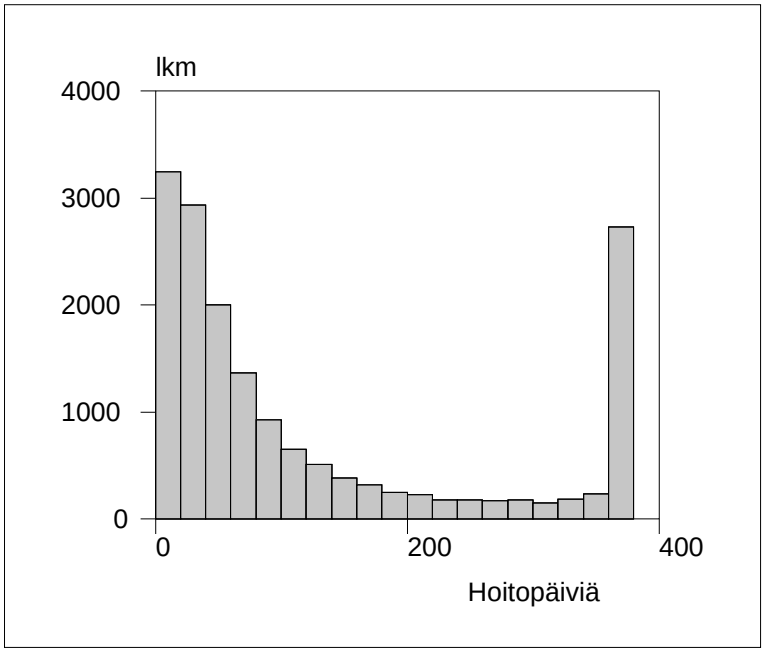
Aineisto on kerätty poimimalla hoitoilmoitusrekisteristä kaikki vuosina 1998-2001 lonkkamurtumadiagnoosilla hoidossa olleet potilaat. Tämän jälkeen näille potilaille on haettu kaikki hoitoilmoitukset poistoilmoitusrekisteristä (ks. esimerkiksi Sund 2000), hoitoilmoitusrekisteristä ja Tilastokeskuksen kuolinsyyrekisteristä vuosilta 1987-2002. Avokäynnit on haettu sairaaloiden Benchmarking projektin tietokannasta. Näin on saatu kaikkien lonkkamurtumapotilaiden tiedot hoitojaksoista sekä kuolemista vähintään murtumaa edeltäneen kymmenen vuoden ajalta sekä murtumaa seuranneen vuoden ajalta. Tiedot eri rekistereistä on yhdistetty henkilötunnusten avulla.

Koska tutkitaan ikääntyneitä lonkkamurtumapotilaita, aineistoon otettiin vain 65 vuotta täyttäneet leikatut potilaat. Tutkimusaineiston potilaille on tehty lonkan korjaus naulaamalla, ruuvaamalla, levyttämällä tai asentamalla lonkkaan osa- tai kokoproteesi. Osaproteeseja on asennettu 50.7%:lle aineston potilaista ja naulauksia, ruuvauksia tai levytyksiä 47.2%:lle. Kokoproteeseja on ainoastaan 2.1%. Ne potilaat, joilla on puuttuvia havaintoja (esimerkiksi leikkauksen päivämäärä puuttuu), on jätetty pois tutkimuksesta. Lopulliseen tutkimusaineistoon kuuluu 16881 potilasta, mikä on 83.3% kaikista ensimmäisen kerran lonkkansa murtaneista potilaista vuosilta $1998 - 2001$.

Rekistereistä on tuotettu kaikille potilaille muuttujat, jotka toisaalta luonnehtivat potilaan saamia hoitoja ja toisaalta potilasta ja hänen terveydentilaansa (ks. muuttujaluettelo, Liite A). Jokaiselle potilaalle on laskettu hoitojen rahamääräiset kokonaiskustannukset (euroina) murtumaa seuranneen vuoden ajalta. Analyyseihin käytetty aineisto sisälsi seuraavat muuttujat: ikä hoitojakson alussa, sukupuoli, murtumatyyppi, suoritettu leikkaus, asumistapa ennen hoitoa (koti, vanhainkoti, terveyskeskus tai sairaala), hoitopäivien lukumäärä murtumaa edeltäneen vuoden sekä murtumaa edeltäneen kahden kuukauden aikana. Aineistossa on myös tiedot tietyistä sairauksista (komorbiditeeteista), joita potilaalla on ollut hoidon alkaessa. Nämä komorbiditeetit ovat: syöpä, diabetes, dementia, verenpainetauti, sydän- ja verisuonitaudit, aivoinfarkti ja aivoverenkierron häiriöt, krooniset hengityselinten sairaudet, anemia, hermoston sairaudet (sisältää Parkinsonin taudin, molemminpuolisen- tai toispuoleisen raajojen halvauksen ja epilepsian), silmäsairaudet (sisältää viherkaihin ja harmaakaihin), ruoansulatusjärjestelmän taudit (sisältää mahahaavan, nivustyrän, maksan sairaudet, sappirakon ja sappiteiden sairaudet ja ärtyneen paksusuolen) ja muita sairauksia (kilpirauhasen liikatoiminta, kihti, munuaisen toiminnanvajaus, nivelrikko, nivelreuma ja muut niveltulehdukset). Komorbiditeettien luokittelussa on käytetty Charlsonin (1987) luokituksesta edelleen kehitettyä luokitusta.

Potilaille on laskettu aineistosta myös elinpäivien lukumäärä murtumaa seuranneen vuoden aikana, hoitopäivien lukumäärä murtumaa seuranneen vuoden aikana ja kotona vietettyjen päivien lukumäärä murtuman ja murtumaa edeltäneen hoitojakson välissä. Kuviossa 1.2 on potilaiden hoitopäivien lukumäärän histogrammi. Siitä nähdään, että aineistossa on eniten potilaita, joilla

on hyvin paljon tai vähän hoitopäiviä.



Kuva 1.2. Hoitopäivien lukumäärän histogrammi

Aineistosta 75.5% on naisia. Tämä johtuu osaksi siitä, että miesten keskimääräinen elinikä on alhaisempi kuin naisilla. Miesten keskimääräinen elinikä aineistossa on 79.2 vuotta ja naisten 81.9. Taulukossa 1.1 on esitetty potilaiden iän luokitus, sekä potilaiden suhtellinen määrä kussakin luokassa. Miehistä 37.6% ja naisista 25.9% on kuollut lonkkamurtumaleikkausta seuranneen vuoden aikana.

Taulukko 1.1. Lonkkamurtuma-aineiston potilaiden ikäjakauma.

Ikä	% -osuus
65-74	19.1
75-79	20.7
80-84	24.3
85-89	23.3
90-104	12.7

Sairaanhoitopiirit ovat useiden kuntien muodostamia hallinnollisia yksiköitä. Kaikki sairaalat, yksityisiä lukuunottamatta, ovat sairaanhoitopiirien alaisuudessa ja omistuksessa. Aineistossa on potilaita kaikista Suomen sairaanhoitopiireistä, Ahvenanmaata lukuunottamatta. Potilaita on siis yhteensä kahdestakymmenestä sairaanhoitopiiristä. Hoitokustannuksia vertailaan sairaanhoitopiiritasolla siksi, että suurin osa lonkkamurtumapotilaiden kustannuksista muodostuu jatkohoidosta, mikä ei ole sairaalaspesifistä. Tässä työssä esitetään vain

Helsingin ja Uudenmaan (HUS), Varsinais-Suomen ja Satakunnan sairaanhoitopiirien hoitokustannusten vertailutuloksia. Kaikkien muiden sairaanhoitopiirien kustannusten vertailu voidaan tehdä tässä työssä esitetyillä tavoilla.

1.1.4 Hoitokustannukset

Potilaiden kustannukset on laskettu Diagnosis Related Groups (DRG) luokituksen (ks. esimerkiksi Mikkola 1998) perusteella. DRG kehitettiin alunperin Yhdysvalloissa sairaalahoitojaksojen ryhmittelyyn ja sitä käytetään erityisesti sairaalalaskutuksessa. Sittemmin DRG:n käyttö on levinnyt useisiin maihin ympäri maailman.

Suomalaiset sairaalat saavat tulonsa laskuttamalla kuntia, joiden omistuksessa sairaalat myös ovat. Näin ollen sairaaloiden on hinnoiteltava toimenpiteensä siten, että menot saadaan mahdollisimman tarkasti katettua. Tämän vuoksi toimenpiteiden hintojen täytyy kuvastaa todellisia kuluja mahdollisimman tarkasti. Epätarkat kuluarviot ovat ongelma sekä kunnille että sairaaloille. Useat kansainväliset tutkimukset osoittavat, että DRG kuvastaa yksittäisen potilaan hoitoon käytettyjä resursseja paremmin kuin vaihtoehtoiset menetelmät (Mikkola 2002.)

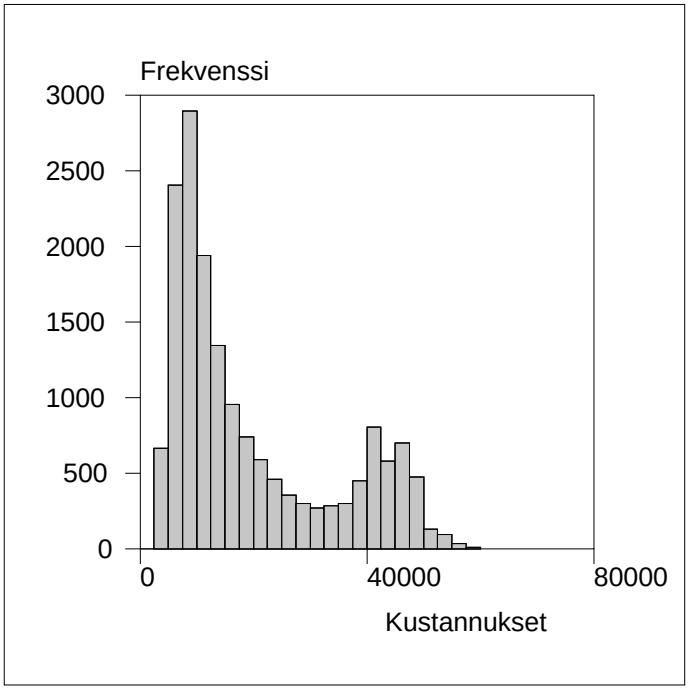
Suomessa DRG:n toimenpideluokitukset perustuvat NOMESKOn leikkausluokitukseen (NOMESCO, Nordic Medico-Statistical Committee). Jokainen toimenpide on luokiteltu ja luokalle on laskettu eräänlainen keskimääräinen "pakettihinta". Jos potilas tulee sairaalaan lonkkamurtuman vuoksi ja lonkka päätetään leikata, hoitojakso hinnoitellaan leikkauksen tyypin perusteella. Hintaan vaikuttaa lisäksi usein myös se, onko leikkaus komplisoitunut. DRG hinta on siis koko hoitojaksolle määritelty keskimääräinen hinta (sisältäen hoitopäivät, lääkityksen ym.) ja se on määritelty kaikille ei-psykiatrisille sairaalakäynneille.

Jokaisen potilaan kustannukset sisältävät ainakin yhden lonkkamurtumasta johtuvan hoitojakson hinnan. Potilaalla saattaa olla erilaisista syistä myös avokäyntejä sairaalassa, jolloin kustannuksiin on lisättu keskimääräinen avokäynnin hinta. Lonkkamurtumaleikkausta seuraavan vuoden aikana potilaalla voi olla myös muista vaivoista johtuvia hoitojaksoja, jolloin myös näiden hoitojaksojen DRG hinnat lisätään potilaan kustannuksiin. Lonkkamurtumaleikkauksen jälkeen potilas pyritään siirtämään mahdollisimman nopeasti terveyskeskukseen kuntoutumaan. Silloin potilaan kustannuksiin lisätään terveyskeskuksessa vietetyn hoitopäivän keskimääräinen hinta kerrottuna terveyskeskuksessa vietettyjen päivien lukumäärällä. Lonkkamurtumaleikkausta seuraavan vuoden aikana annetun psykiatrisen erikoissairaanhoidon, vanhainkotihoidon ja kuntoutushoidon hinnat on myös laskettu keskimääräisten hoitopäivien hintojen avulla. Potilaan *kokonaiskustannus* saadaan tällä tavalla laskemalla yhteen kaikki lonkkamurtumaa seuraavan vuoden aikana potilaan hoidosta kertyneet kustannukset. Tämä kokonaiskustannus on tässä työssä selitettävä muuttuja eli vastemuuttuja.

Kaikkien potilaiden kustannukset on laskettu vuoden 2001 koko maan keski-

määräisiä DRG hintoja käyttäen (ks. Hujanen 2003). Vuosittain kustannusten laskenta hieman muuttuu muun muassa sairaaloiden tuottavuuden vaihtelusta johtuen. Koska kustannuserojen halutaan kertovan ensisijaisesti hoitokäytäntöjen eroista, saadaan kustannukset vertailukelpoisiksi poistamalla laskentaperusteista johtuva vaihtelu.

Kuviossa 1.3 on esitetty potilaiden kustannusten jakauman histogrammi. Siitä nähdään, että kustannusten jakauma on kaksihuippuinen.



Kuva 1.3. Kustannusten jakauma

1.2 Tutkielman tavoitteet

Työni liittyy Sosiaali- ja terveysalan tutkimus- ja kehittämiskeskuksessa (Stakes) toimivan hankkeen "Rekisteritiedon hyödyntäminen erikoissairaanhoidon vaikuttavuustutkimuksessa" lonkkamurtumia käsittelevään osaprojektiin. Projektin yhtenä osatavoitteena on lonkkamurtumahoitojen kustannustehokkuuden vertailu. Tutkielmassani esitän menetelmän, joka soveltuu lonkkamurtumapotilaiden hoitokustannusten vertailuun sairaanhoitopiireittäin. Koska työn käytännön tavoitteena on selkeä hoitokäytäntöjen kustannustehokkuuden vertailu, tulee menetelmän antamien tulosten olla helposti tulkittavissa. Siksi luovuin ensiksi kokeilemistani kustannusten regressiopohjaisista selitysmalleista ja päädyin alaluvussa 1.3 esitettävään lähestymistapaan.

Esittelen tutkielmassani myös kolmen sairaanhoitopiirin (HUS, Varsinais-Suomi ja Satakunta) kustannusvertailujen tuloksia. Tavoitteena on sekä havainnollistaa käyttämääni tutkimusmenetelmää että esittää myös todellisiin rekiste-

ripohjaisiin aineistoihin perustuvia uusia tuloksia. Myöhemmin on mahdollista julkaista myös muiden piirien vertailuja koskevat vastaavat tulokset.

1.3 Tutkimusmenetelmät

Kustannusvertailut perustuvat rekisteriaineistoihin, joita ei ole suunnitelmallisesti kerätty erityisesti näitä vertailuja varten. On mahdollista, että esimerkiksi potilasaineksen eroavuudet sairaanhoitopiirien välillä aiheuttavat systemaattisia eroja hoitokustannuksiin. Tällöin kustannuserot eivät johdu välttämättä hoitokäytäntöjen eroista. On siis pyrittävä kontrolloimaan taustamuuttujien vaikutusta havaittuihin kustannuseroihin. Tämä sekoittavien muuttujien aiheuttama harha on eräs havainnoivan tutkimuksen perusongelma (esim. Rosenbaum 2002, luku 1).

Sovellan työssäni Rosenbaumin ja Rubinin (1983) esittämän vastaavuuspistemäärän käyttöön perustuvaa tutkimustekniikkaa. Menetelmän keskeinen idea on korvata sekoittavien taustamuuttujien (kovariaattien) joukko yhdellä muuttujalla, joka on näiden kovariaattien funktio. Sen jälkeen päättely ehdollistetaan tähän yhteen muuttujaan (vastaavuuspistemäärä). Tällä tavoin ongelma palautuu ikään kuin yhden sekoittavan taustamuuttujan ongelmaksi. Vastaavuuspistemäärä on yleensä estimoitava aineistosta, joten menetelmän käyttöön liittyy myös estimointiongelma.

Vastaavuuspistemäärään perustuva vertailu tehdään aina parittain. Jos vertailtavia ryhmiä on esimerkiksi 4, niin kaikkien mahdollisten parittaisten vertailujen lukumäärä on 6. Kun vertailtavia ryhmiä on useita, saattaa tarvittavien vertailujen määrä tuntua menetelmän heikkoudelta verrattuna mallipohjaiseen analyysiin (esimerkiksi kustannusten selittämiseksi laadittu regressiomalli). Itse asiassa parittaista vertailua voidaan pitää tämän tekniikan vahvuutena (Rubin 1997). On ainakin periaatteessa mahdollista, että kahden vertailtavan sairaanhoitopiirin potilasaines poikkeaa niin täysin toisistaan, että taustamuuttujien arvoalueet eivät lainkaan kohtaa tai niillä on vain vähän yhtymäkohtia. Vastaavuuspistemäärään perustuvalla menetelmällä tämä arvoalueiden erilaisuus havaitaan helposti. Kun vertailtavien taustamuuttujienmuuttujien arvoalueiden päällekkäisyys on riittävä, voidaan taustamuuttujien vaikutus poistaa ja kustannusero estimoida.

Tutkielmassa käytetään myös niin sanottua yleistettyä vastaavuuspistemäärää (Imbens 2000), jonka avulla voidaan vertailla useita ryhmiä samanaikaisesti. Yleistetty vastaavuuspistemäärä ei kuitenkaan säilytä kaikkia vastaavuuspistemäärän ominaisuuksia ja siksi sen käyttömahdollisuudet eivät ole kaikilta osin yhtä laajat kuin vastaavuuspistemäärän.

Vastaavuuspistemäärän rajoitteena voidaan pitää sitä, että sen avulla on mahdollista kontrolloida vain havaittujen muuttujien vaikutusta. Tämä on kuitenkin kaiken havainnoivan tutkimuksen rajoitus verrattuna kokeelliseen tutkimukseen. Vastaavuuspistemäärään perustuvalla menetelmällä pyritään siihen, että taustamuuttujien jakaumat olisivat samanlaiset kahdessa vertailtavassa

ryhmässä. Havaituissa otoksissa empiiriset jakaumat poikkeavat kuitenkin yleensä toisistaan ja satunnaisvaihtelun vaikutus voi olla häiritsevä pienissä otoksissa. Tässä tutkimuksessa havaintomäärät ovat kuitenkin suuria, joten pienten otosten ongelmaa ei ole.

Nykyään vastaavuuspistemäärää soveltavia tutkimuksia on löydettävissä useiden eri alojen lehdistä, joista mainittakoon esimerkiksi biometria, lääketiede ja ekonometria (ks. D'Agostino 1998 ja Rubin 1997). Esimerkiksi terveystieteellisissä tutkimuksissa on usein tavoitteena selvittää jonkin toimenpiteen tai hoidon vaikutus, vaikka havainnot eivät ole satunnaistettujen kokeiden tuloksia, vaan perustuvat johonkin havainnoivaan tutkimusasetelmaan. Joissain soveltavissa ekonometrisissa tutkimuksissa on tarkasteltu esimerkiksi työhön perehdyttävien koulutusohjelmien vaikutusta (Heckman, Ichimura ja Todd 1998) menetelmillä, jotka hyödyntävät Rosenbaumin ja Rubinin (1983) esittämää tekniikkaa.

Tutkielmassani sovellan vastaavuuspistemäärään perustuvaa tekniikkaa usealla eri tavalla. Menetelmä, jossa hyödynnetään epäparametrista regressiota, on verrattain uusi lähestymistapa. Heckman ym. (1998) ovat soveltaneet tätä lähestymistapaa, joskin melko erilaisessa tilanteessa. Yleistetyn vastaavuuspistemäärän käyttö (6. luku) on myös uusi tekniikka (Imbens 2000), joka vaatii lisätutkimusta. Vastaavuuspistemäärän käyttö piirien sisäisten kustannusryhmien erottelussa on niin ikään uusi tapa soveltaa vastaavuuspistemäärää.

1.4 Tilastollinen laskenta

Sovellettava menetelmä on rakentunut lopulliseen muotoonsa vasta tutkimuksen edetessä. Valitsemaani lähestymistapaa ei ole tietääkseni aiemmin käytetty juuri tällä sovellusalueella. Siksi ei ole olemassa myöskään valmista ohjelmistoa tai funktiota, joilla esitettävät sairaanhoitopiirien kustannusvertailut saataisiin suoraan tehtyä. Tutkimuksen edetessä on käytetty useita tilastollisia ohjelmistoja ja yhdistelty niiden ominaisuuksia, jotta tutkimusongelmat on saatu ratkaistua tehokkaasti ja järkevällä tavalla.

Perustyökaluna analyyseissa ja datan hallinnassa on ollut Survo-ohjelmisto. Osa estimoinneista on tehty R- ja SAS-ohjelmistoilla. Näitä kolmea ohjelmistoa yhdistelemällä on löydetty tarvittavat laskennalliset ratkaisut. Stata on ainoa tuntemani ohjelmisto, johon on implementoitu vastaavuuspistemäärää käyttäviä menetelmiä ja jossa on suoraan komentoja niiden toteuttamiseksi. Olen kuitenkin saanut Statan käyttööni vasta sen jälkeen, kun tämän työn empiiriset analyysit oli jo tehty.

Rekisteripohjaisten tutkimusten luonne asettaa omat rajoitteensa käytettäville menetelmille ja ohjelmistoille. Suuret havainto- ja muuttujamäärät vaativat ohjelmistolta hyvää laskentatehokkuutta ja hyviä suurten aineistojen käsittelyominaisuuksia. Rekisteriaineiston muokkaaminen tutkimusaineistoksi vaatii tarkkuutta ja asiantuntemusta. Tämän työn empiiristen tarkastelujen perustana olevan tutkimusaineiston tuottaminen rekistereistä on ollut erittäin suu-

ritöinen ja vaativa prosessi, jonka on toteuttanut Reijo Sund Stakesista.

1.5 Tutkielman rakenne

Tutkielmassa kuljetetaan teoriaa ja sovellusta rinnakkain. Teoriaa havainnollistetaan käytännön esimerkein, jotka on johdettu työssä käsiteltävästä sovelluksesta. Ensimmäisessä luvussa kerrotaan lonkkamurtumista, hoidon kustannuksista ja tutkimusaineistosta. Tässä yhteydessä esitellään myös tutkimusongelma ja kerrotaan hieman kustannusten vertailun ongelmista. Lisäksi kuvaillaan käytettyjä menetelmiä sekä tutkimuksessa tarvittua tilastollista laskentaa.

Toisessa luvussa esitellään vastaavuuspistemäärän käyttöön liittyvät keskeiset oletukset sekä perusteoria. Tässä luvussa keskitytään kahden sairaanhoitopiirin vertailuun, mutta jatkossa tarkastellaan myös usean piirin samanaikaista vertailua. Menetelmän ymmärtämisen kannalta ovat keskeisiä toisessa luvussa sekä liitteissä B ja C esitettävät satunnaismuuttujien riippumattomuuden, ehdollisen riippumattomuuden ja ehdollisen odotusarvon käsitteet. Kolmannessa luvussa sovelletaan toisessa luvussa käsiteltyä teoriaa ja esitetään sairaanhoitopiirien kustannukset vastaavuuspistemäärän funktiona. Kolmannessa luvussa käsitellään myös regressiofunktioiden estimointia sekä niiden tulkintaan liittyviä käsitteitä.

Kun kolmessa ensimmäisessä luvussa on kuvattu tutkimuksen taustaa ja menetelmien teoriaa, sovelletaan neljännessä teoriaa käytäntöön. Ensin estimoidaan vastaavuuspistemäärät ja sen jälkeen kustannukset vastaavuuspistemäärän funktiona. Niiden avulla tutkitaan sairaanhoitopiirien kustannuseroja. Lopuksi esitetään vielä vastaavuuspistemäärän estimointi multinomisella logitmallilla ja tutustutaan esimerkin avulla useiden piirien samanaikaisen vertailun ongelmaan.

Viidennessä luvussa esitetään yleistetyn vastaavuuspistemäärän käsite ja tarkastellaan sen avulla usean piirin samanaikaista vertailua. Kuudennessa luvussa käsitellään myös piirien sisäisten ryhmien vertailuun soveltuvaa menetelmää. Koska piirien sisällä näyttää olevan kaksi kustannuksiltaan selkeästi erilaista potilasryhmää, estimoidaan näiden ryhmien kustannukset erikseen. Kahteen eri piiriin kuuluvan sisäisen ryhmän vertailemiseksi on nämä ryhmät ensin muodostettava kustannusjakauman avulla. Sen jälkeen niitä voidaan vertailla vastaavuuspistemäärään perustuvin menetelmin.

2 Vastaavuuspistemäärään perustuva vertailu

2.1 Tutkimusasetelma

Tutkimusaineistoon on valikoitunut tietyllä aikavälillä sattuneiden lonkkamurtumien seurauksena n potilasta (n=16881). Aineiston synty voidaan ajatella tietynlaisena laskuriprosessina. Murtumatapahtumalle asetetuista rajoitteista on kerrottu alaluvussa 1.1.3. Hoitoa tietyssä sairaanhoitopiirissä nimitetään käsittelyksi (treatment) koesuunnitteluun vakiintuneen tilastollisen terminologian mukaan.

2.1.1 Kahden käsittelyn tilanne

Tarkastellaan nyt kahden käsittelyn tulosten vertailua. Käsittelymuuttuja Z on indikaattori, jonka arvo on 0 tai 1. Jokaista tilastoyksikköä (potilasta) i kohti indikaattori Z_i osoittaa, onko yksikkö saanut tarkastelun kohteena olevan käsittelyn ($Z_i = 1$) vai onko se saanut vertailukäsittelyn ($Z_i = 0$), jota kutsutaan myös kontrollikäsittelyksi.

Olkoon n otoskoko. Silloin n-ulotteinen vektori $(Z_1, Z_2, \ldots, Z_n)$ kertoo, miten käsittelyt ovat jakautuneet tilastoyksiköiden yli. Kun vertaillaan kahden sairaanhoitopiirin kustannuksia, nimitetään toista piiriä vertailuryhmäksi (kontrolli) ja toista käsittelyryhmäksi. Kumpi sairaanhoitopiiri valitaan vertailuryhmäksi, on tutkijan päätettävissä. Tämän työn vertailuissa käytetään aina samaa vertailuryhmää (HUS).

Merkitään yksikön i tulosta (kustannus) Y_{1i}, $1 \leq i \leq n$, kun yksikkö on saanut käsittelyn ja Y_{0i}, kun yksikkö on saanut vertailukäsittelyn (kontrolli). Todellisuudessa havaitaan indikaattorin Z_i arvo ja tulos Y_i, missä

$$\begin{aligned} Y_i &= Z_i Y_{1i} + (1 - Z_i) Y_{0i} \\ &= Y_{0i} + Z_i (Y_{1i} - Y_{0i}), \end{aligned}$$

eli

$$Y_i = \begin{cases} Y_{1i}, & \text{kun } Z_i = 1, \\ Y_{0i}, & \text{kun } Z_i = 0. \end{cases}$$

Lisäksi havaitaan p-ulotteinen kovariaattivektori $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$. Teoreettisesti käsittelyn vaikutus yksikön i tulokseen on $Y_{1i} - Y_{0i}$, mutta tätä vaiku-

tusta ei voi havaita, koska jokaista yksikköä kohti vain joko muuttujan Y_{1i} tai Y_{0i} arvo havaitaan. Vaikka käytännössä vain toinen muuttujista Y_{1i} ja Y_{0i} havaitaan, niin periaatteessa yksikköön olisi voitu kohdistaa kumpi tahansa vaihtoehtoisista käsittelyistä. Potilas olisi siis periaatteessa voitu hoitaa joko piirissä 1 tai piirissä 0.

Tässä työssä käytetään niin sanottua potentiaalisten vasteiden merkintätapaa, jonka Rubin (1974, 1977) on esitellyt ensimmäisenä. Hyvä johdanto tähän ajattelutapaan on Holland (1986), joka puhuu Rubinin kausaalisen päättelyn mallista. Myös Davison (2003, s. 424) käsittelee lyhyesti tätä lähestymistapaa kausaalisen päättelyn yhteydessä. Ajatellaan, että potilaan i hoitokustannukset olisivat Y_{1i}, jos hänet hoidettaisiin piirissä 1 ja Y_{0i}, jos hänet hoidettaisiin piirissä 0. Vastemuuttujan Y_i tarkoituksena on mitata käsittelyn (piirin) tulosta. Tässä mielessä käsittely voi olla vaikutuksen aiheuttaja tai syy. Sen sijaan esimerkiksi potilaan sukupuolta ei voida ajatella vaikutuksen aiheuttajana, koska samaa potilasta ei voida ajatella hoidettavaksi joko naisena tai miehenä.

Tilastollisessa analyysissa käytettävien menetelmien kannalta on oleellista korostaa, että periaatteessa kumpi tahansa kahdesta mahdollisista vasteesta Y_{0i}, Y_{1i} on mahdollinen, mutta käytännössä potilas i on hoidettu vain joko piirissä 0 tai piirissä 1 ja toisen muuttujan havainnot puuttuvat. Aineiston havaintoprosessi on siis esimerkiksi muotoa:

$$\begin{array}{cc} \text{Satunnaismuuttujat} & \qquad\qquad \text{Havainnot} \\ \begin{pmatrix} Z_1 & Y_{01} & Y_{11} & \mathbf{X}_1 \\ Z_2 & Y_{02} & Y_{12} & \mathbf{X}_2 \\ \dots & \dots & \dots & \dots \\ Z_i & Y_{0i} & Y_{1i} & \mathbf{X}_i \\ \dots & \dots & \dots & \dots \\ Z_n & Y_{0n} & Y_{1n} & \mathbf{X}_n \end{pmatrix} & \longrightarrow \begin{pmatrix} 1 & y_{11} & \mathbf{x}_1 \\ 0 & y_{02} & \mathbf{x}_2 \\ \dots & \dots & \dots \\ 0 & y_{0i} & \mathbf{x}_i \\ \dots & \dots & \dots \\ 1 & y_{1n} & \mathbf{x}_n \end{pmatrix}. \end{array}$$

Satunnaismuuttujia merkitään isoilla ja niiden arvoja pienillä kirjaimilla. Esimerkiksi potilas 2 on hoidettu piirissä 0, joten on saatu havainto y_{02} ja satunnaismuuttujan Y_{12} arvoa ei havaita.

Esimerkki 2.1. Yksinkertaisimmassa satunnaistetussa kokeessa yksiköt valitaan käsittely- ja vertailuryhmään heittämällä harhatonta lanttia. Silloin kaikilla yksiköillä i, $1 \leq i \leq n$, on todennäköisyys $\frac{1}{2}$ tulla valituksi käsittelyryhmään eli $P(Z_i = 1) = \frac{1}{2}$. Olkoon $P(Z_i = 1) = \pi_i$, $0 < \pi_i < 1$, yksikön i todennäköisyys tulla valituksi käsittelyryhmään ja olkoot käsittelymuuttujat $Z_1, \dots, Z_n$ keskenään riippumattomat. Silloin saadaan

$$P(Z_1 = z_1, \dots, Z_n = z_n) = \prod_{i=1}^{n} \pi_i^{z_i} (1 - \pi_i)^{1 - z_i}, \tag{2.1}$$

missä $z_i \in \{0, 1\}$, $1 \leq i \leq n$. Mallin (2.1) mukaan yksiköt valitaan käsittely ja vertailuryhmään harhaisten lanttien heitolla, missä yleisessä tilanteessa jo-

kaista yksikköä kohti on oma lantti ja oma valintatodennäköisyys π_i. Oleellista on, että satunnaistetuissa kokeissa tutkija tuntee satunnaistamismekanismin ja valintatodennäköisyydet.

Valintatodennäköisyys voi myös riippua valittavan yksikön ominaisuuksista. Jos tarkasteltava yksikkö on potilas, voimme ehdollistaa valintatodennäköisyyden esimerkiksi potilaan sukupuoleen S ($S = 1$, kun potilas on mies; $S = 0$, kun potilas on nainen) siten, että $P(Z_i = 1|S = 1) = \frac{1}{2}$ ja $P(Z_i = 1|S = 0) = \frac{1}{4}$. Silloin miehet valitaan käsittelyryhmään todennäköisyydellä $\frac{1}{2}$ ja naiset todennäköisyydellä $\frac{1}{4}$. Yleisesti siis yksikön i valintatodennäköisyydet voivat riippua havaituista kovariaattivektorien $\mathbf{X}_i$ arvoista $\mathbf{x}_i$, eli

$$\pi_i = P(Z_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \pi(\mathbf{x}_i),\ 1 \leq i \leq n.$$

Silloin valintatodennäköisyyden malli (2.1) voidaan kirjoittaa muodossa

(2.2)
$$P(Z_1 = z_1, \ldots, Z_n = z_n|\mathbf{X}_1 = \mathbf{x}_1, \ldots, \mathbf{X}_n = \mathbf{x}_n) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{z_i}[1 - \pi(\mathbf{x}_i)]^{1-z_i}.$$

Satunnaistetuissa kokeissa todennäköisyydet $\pi(\mathbf{x}_i)$ tunnetaan, koska tutkija itse tavallisesti konstruoi satunnaistamismekanismin. □

2.1.2 Odotettu käsittelyvaikutus

Merkintöjen yksinkertaistamiseksi jätetään jatkossa alaindeksi i pois. Nyt siis käsittelyn ($Z = 1$) aiheuttama vaste on Y_1 ja kontrollin ($Z = 0$) aiheuttama vaste on Y_0. Muuttujat Y_1 ja Y_0 esittävät siis kahta potentiaalista vastetta. Koska erotuksesta $Y_1 - Y_0$ ei saada havaintoja, tarkastelemme *odotettua käsittelyvaikutusta* (kontrollin suhteen)

(2.3)
$$\begin{aligned} \tau &= E(Y_1 - Y_0) \\ &= E(Y_1) - E(Y_0), \end{aligned}$$

joka voidaan tietyin edellytyksin (ks. alaluku 2.1.3) estimoida. Huomattakoon, että käsittelyvaikutus riippuu aina valitusta kontrolliryhmästä. Se, miten tilastoyksiköt valitaan koeryhmään (käsittely) ja kontrolliryhmään, on ratkaisevan tärkeää. Nämä valintaperiaatteet ovat tilastollisen koesuunnittelun perusasioita.

Havainnoivassa tutkimuksessa (observational study, ks. Rosenbaum 2002, luku 1), kuten esimerkiksi käytettäessä rekisteriaineistoja, tutkija ei voi vaikuttaa havaintojen valintamekanismiin. Silloin sellaiset kokeellisen tutkimuksen periaatteet, kuten esimerkiksi satunnaistaminen, eivät ole käytettävissä. Kuitenkin havainnoivassa tutkimuksessakin ollaan kiinnostuneita joidenkin käsittelyjen, interventioiden tai toimenpiteiden vaikutuksesta vasteeseen. Siinä mielessä se muistuttaa kokeellista tutkimusta. Kokeissa tilastoyksiköt voidaan liit-

tää käsittelyihin kontrolloidusti käsittelyvaikutuksen vertailtavuuden varmistamiseksi. Havainnoivassa tutkimuksessa tämä kontrolli puuttuu.

Jokaista yksikköä kohti havaitaan muuttujapari (Z, Y), missä $Y = ZY_1 + (1-Z)Y_0$ ja Z on indikaattorimuuttuja, joka saa arvon 0 tai 1. Havaintoaineiston perusteella saadaan siis havaintoja vain satunnaismuuttujasta Y_1, kun $Z = 1$ ja satunnaismuuttujasta Y_0, kun $Z = 0$. Näiden havaintojen perusteella voidaan estimoida odotusarvot

$$E(Y|Z = 1) = E(Y_1|Z = 1)$$

ja

$$E(Y|Z = 0) = E(Y_0|Z = 0).$$

On tärkeää huomata, että $E(Y_1)$ ja $E(Y_1|Z = 1)$ eivät tietenkään ole sama asia eivätkä ne yleensä saa samoja arvoja. Vastaava huomautus koskee myös odotusarvoja $E(Y_0)$ ja $E(Y_0|Z = 0)$.

Odotusarvo $E(Y_1)$ on otettu yli koko populaation, kun taas $E(Y_1|Z = 1)$ on odotusarvo yli sellaisten yksiköiden, jotka valikoituvat (tai valitaan) käsittelyyn. Aineistosta pystytään siis laskemaan vain erotus

$$E(Y_1|Z = 1) - E(Y_0|Z = 0),$$

mikä ei yleisesti ole odotettu käsittelyvaikutus (2.3).

Esimerkki 2.2. Vertaillaan sairaanhoitopiirien SHP1 ja SHP0 hoitokustannuksia. Ajatellaan esimerkkinä sellaista hypoteettista käytäntöä, että todennäköisesti vaativaa ja kallista hoitoa tarvitsevat potilaat (lääketieteellisen ennusteen perusteella) ohjataan piiriin SHP1 ($Z = 1$) ja helpommat ja halvemmalla hoidolla selviävät lähetetään piiriin SHP0 ($Z = 0$). Silloin $E(Y_1|Z = 1)$ on odotusarvo sellaisessa osapopulaatiossa, johon potilaiden valintaprosessi ohjaa keskimääräistä kalliimpaa hoitoa tarvitsevat potilaat. Sen sijaan $E(Y_1)$ otetaan yli koko potilaspopulaation, jossa ovat siis mukana myös keskimääräistä halvempaa hoitoa saavat potilaat. Tässä tilanteessa on selvää, että $E(Y_1|Z = 1) > E(Y_1)$. Havaintoja saadaan tietysti vain potilaista, jotka todella on hoidettu piirissä SHP1 (tai piirissä SHP0). Voimme siis havaintojen perusteella estimoida odotusarvon $E(Y_1|Z = 1)$, mutta sitä ei ole suositeltava käyttää $E(Y_1)$:n estimaattina. Odotusarvon $E(Y_1)$ harhatonta estimointia varten tarvittaisiin havaintoja myös niistä hypoteettisista hoitokustannuksista, joita saataisiin, kun piirissä SHP0 hoidettuja potilaita hoidettaisiinkin piirissä SHP1. □

2.1.3 Yksiköiden liittäminen käsittelyihin ja satunnaistaminen

Milloin sitten identiteetit

$$E(Y_0) = E(Y_0|Z = 0) \text{ ja } E(Y_1) = E(Y_1|Z = 1) \tag{2.4}$$

pitävät paikkansa? Jos on perusteltua ajatella, että Z ja Y_1 ovat toisistaan riippumattomat ja samoin Z ja Y_0, niin silloin (2.4) pitää paikkansa. Muuttujien Z ja Y_i riippumattomuutta merkitään $Z \perp\!\!\!\perp Y_i$. Jos siis oletus

$$Y_0 \perp\!\!\!\perp Z \text{ ja } Y_1 \perp\!\!\!\perp Z \tag{2.5}$$

pitää paikkansa, niin identiteetit (2.4) ovat voimassa. Tämä tulos osoitetaan liitteessä C (identiteetit (C.4)).

Esimerkki 2.3. Vertaillaan sairaanhoitopiirin SHP1 ja SHP0 hoitokustannuksia kuten Esimerkissä 2.2. Silloin riippumattomuusoletus $Y_1 \perp\!\!\!\perp Z$ ei voi pitää paikkaansa, sillä riippumattomuudesta seuraa identiteetin (C.3) nojalla (Liite C)

$$E(Y_1|Z = 1) = E(Y_1).$$

Tämä on ristiriidassa sen kanssa, että Esimerkissä 2.2

$$E(Y_1|Z = 1) > E(Y_1).$$

Esimerkissä 2.1 kuvattu satunnaistaminen sen sijaan takaisi riippumattomuusoletuksen toteutumisen. Tämän tutkielman aineiston todellisten sairaanhoitopiirien potilaita ei tietystikään valita satunnaistamalla, vaan potilaat valikoituvat tavalla, jota ei voi kontrolloida. Tätä tilannetta tarkastellaan alaluvussa 2.2. □

Jos tilastoyksiköt jaetaan koeryhmään ja vertailuryhmään täysin satunnaisesti jonkin satunnaistamismekanismin (esimerkiksi lantin heitto) avulla, niin silloin Z on riippumaton muista populaatiossa määritellyistä satunnaismuuttujista ja siis erityisesti myös satunnaismuuttujista Y_0 ja Y_1 sekä tilastoyksikköjä luonnehtivista kovariaateista $X_1, X_2, \ldots, X_p$. Satunnaistamisesta seuraa riippumattomuus

$$(Y_0, Y_1) \perp\!\!\!\perp Z, \tag{2.6}$$

joka on oletusta (2.5) voimakkaampi (ks. Liite B). Relaation (2.6) mukaan siis käsittelymuuttuja Z ja satunnaisvektori (Y_0, Y_1) ovat riippumattomat. Identiteetit (2.4) seuraavat tietysti myös oletusta (2.5) tiukemmasta oletuksesta (2.6).

Jos oletus (2.5) (tai tiukempi oletus (2.6)) pitää paikkansa, niin identiteetit (2.4) ovat voimassa ja odotettu käsittelyvaikutus τ voidaan lausua muodossa

$$\begin{aligned}\tau &= E(Y_1) - E(Y_0) \\ &= E(Y_1|Z = 1) - E(Y_0|Z = 0) \\ &= E(Y|Z = 1) - E(Y|Z = 0),\end{aligned}$$

missä $Y = ZY_1 + (1 - Z)Y_0$. Tämä tarkoittaa sitä, että satunnaistetussa koeasetelmassa havainnoista (Z, Y) voidaan harhattomasti estimoida τ. Lasketaan ensin vastemuuttujan keskiarvot käsittelyryhmässä ja vertailuryhmässä ja muodostetaan keskiarvojen erotus.

2.2 Käsittelyryhmän valinta kovariaattien perusteella - kokeellinen vs. havainnoiva tutkimus

Kokeellisissa tutkimusasetelmissa voidaan soveltaa jotain satunnaistamistekniikkaa ja estimoida käsittelyvaikutus alaluvussa 2.1.3 esitetyllä tavalla. Tilastollisesti satunnaistaminen takaa sen, että Z ja (Y_0, Y_1) ovat riippumattomat ja myös Z ja kovariaatit $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ ovat riippumattomat. Tästä seuraa, että kovariaateilla on sama jakauma sekä koeryhmässä että kontrolliryhmässä. Koska näiden ryhmien yksiköiden voidaan olettaa olevan keskimäärin samanlaisia, ryhmien suora vertailu on mielekästä.

Havainnoivissa asetelmissa tutkija ei voi kontrolloida sitä, miten tilastoyksiköt valikoituvat käsittely- ja vertailuryhmään. Mikään ei silloin takaa sitä, että kovariaattien $\mathbf{X}$ jakauma olisi sama sekä käsittelyryhmässä että vertailuryhmässä. Käsittelymuuttuja Z ja samoin vaste Y voivat riippua kovariaateista. Siksi ryhmäkeskiarvojen vertailu ei tässä tilanteessa anna oikeaa kuvaa käsittelyvaikutuksesta, vaan kovariaattien jakaumien erilaisuus käsittely- ja vertailuryhmissä sotkee päättelyä.

Palataan vielä esimerkkiin 2.1 ja valintatodennäköisyyksien malliin (2.2). Kokeellisessa tutkimuksessa tunnetaan valintatodennäköisyydet $\pi(\mathbf{x}_i)$, $1 \leq i \leq n$, mutta havainnoivassa tutkimuksessa ne ovat tuntemattomia. Sanotaan, että havainnoivassa asetelmassa ei ole piiloharhaa, jos nämä tuntemattomat valintatodennäköisyydet riippuvat vain havaituista kovariaateista. Rubin (1977) kutsuikin mallia (2.2) satunnaistamiseksi kovariaatin perusteella ("randomization on the basis of a covariate"). Jos taas tuntemattomat valintatodennäköisyydet riippuvat myös muuttujista, joita ei havaita, asetelmassa on piiloharhaa.

2.2.1 Ehdollisen riippumattomuuden hypoteesi

Kaikista tilastoyksiköistä havaitaan kovariaattien $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ arvot. Jos voidaan olettaa, että (Y_0, Y_1) ja Z ovat ehdollisesti riippumattomat ehdolla $\mathbf{X}$, niin silloin mahdollisten vasteiden (Y_0, Y_1) vertailu voidaan tehdä vastaavalla tavalla kuin alaluvussa 2.1.3 esitetyn riippumattomuusoletuksen (2.6) vallitessa.

Mahdollisten vasteiden (Y_0, Y_1) ja käsittelymuuttujan Z ehdollista riippumattomuutta ehdolla $\mathbf{X}$ (ks. määritelmä liitteestä B) merkitään

$$(Y_0, Y_1) \perp\!\!\!\perp Z | \mathbf{X}. \tag{2.7}$$

Rubin (1978) ja Rosenbaum ja Rubin (1983) ovat ensimmäisinä täsmällisesti

muotoilleet oletuksen (2.7) ja käsitelleet sen merkitystä havainnoivassa tutkimuksessa.

Ehdollisen riippumattomuuden oletuksesta (2.7) seuraavat tuloksen (B.10) perusteella (Liite B) riippumattomuusrelaatiot

$$Y_0 \perp\!\!\!\perp Z|\mathbf{X} \text{ ja } Y_1 \perp\!\!\!\perp Z|\mathbf{X}. \tag{2.8}$$

Vastaavalla tavalla riippumattomuusoletuksesta (2.6) saatiin relaatiot $Y_0 \perp\!\!\!\perp Z$ ja $Y_1 \perp\!\!\!\perp Z$.

Esimerkki 2.4. Jatketaan sairaanhoitopiirien SHP1 ja SHP0 hoitokustannusten vertailua (vrt. esimerkit 2.2 ja 2.3). Oletetaan, että taudin X esiintyvyys SHP1:ssä on suurempi kuin SHP0:ssa. Silloin siis $P(X = 1|Z = 1) > P(X = 1|Z = 0)$, missä X on taudin esiintymistä osoittava indikaattorimuuttuja ($X = 1$, kun henkilöllä on tauti X, muutoin $X = 0$). Lisäksi tiedetään, että tauti X vaikeuttaa leikkauksesta toipumista ja johtaa tavallisesti keskimääräistä paljon korkeampiin hoitokustannuksiin. Nyt siis pätee epäyhtälö

$$E(Y_j|X = 1) > E(Y_j|X = 0),\ j = 0, 1.$$

Koska SHP1:ssä hoidetaan keskimääräistä enemmän kalliita X-potilaita, niin

$$E(Y_1|Z = 1) > E(Y_1).$$

Odotusarvo $E(Y_1|Z = 1)$ otetaan yli niiden potilaiden, jotka valikoituvat hoidettavaksi SHP1:een ja jotka hoidetaan siellä. Käytännössä potilaat asuvat piirin SHP1 alueella ja saavat siksi hoidon SHP1:ssä. Oletammekin tässä yksinkertaisuuden vuoksi, että $Z = 1$, kun potilas asuu SHP1:n alueella. Vastaavasti $Z = 0$, kun potilas asuu SHP0:n alueella. Odotusarvo $E(Y_1)$ taas otetaan yli kaikkien potilaiden ilman erottelua. Siinä siis ajatellaan, että myös piirin SHP0 potilaat hoidettaisiin piirissä SHP1. Koska potilaat samaistetaan kovariaattien arvoihin ja kovariaattien suhteen identtisiä potilaita hoidetaan kummassakin piirissä, ei potilaiden "siirto" piiristä toiseen ole laskennallisesti ongelmallista.

Jos teemme riippumattomuusoletuksen (2.8), niin silloin $Y_j \perp\!\!\!\perp Z|X$, $j = 0, 1$. Tämä tarkoittaa sitä, että esimerkiksi $Y_1 \perp\!\!\!\perp Z|X = 1$ ja $Y_1 \perp\!\!\!\perp Z|X = 0$. Tarkastellaan vielä lähemmin vaikkapa oletuksen $Y_1 \perp\!\!\!\perp Z|X = 1$ sisältöä. Ehdollisen riippumattomuuden määritelmästä (B.6) (Liite B) seuraa, että tautia X sairastavien hoitokustannusten jakauma ei riipu siitä, minkä piirin alueella potilas on asunut. Kustannukset voivat riippua kovariaateista ja piirien hoitokäytännöistä, mutta oletamme kustannukset riippumattomiksi potilaan asuinpaikasta.

Tässä yksinkertaistetussa esimerkissä on oletettu, että on vain yksi kustannuksiin vaikuttava kovariaatti X. Silloin ehdollisen riippumattomuuden (2.7) nojalla

$$E(Y_j|Z,X) = E(Y_j|X),\ j = 0,1,$$

josta seuraa esimerkiksi

$$E(Y_1|Z=1,X=1) = E(Y_1|X=1).$$

Potilaan asuinpaikka ($Z = 1$ tai $Z = 0$) ei siis vaikuta kustannuksiin. □

Vastaavuuspistemäärä $e(\mathbf{x})$ määritellään todennäköisyytenä

$$e(\mathbf{x}) = P(Z = 1|\mathbf{X} = \mathbf{x}), \tag{2.9}$$

missä $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ sisältää kovariaattien havaitut arvot. Teemme jatkossa sen $\mathbf{x}$:n arvoaluetta koskevan oletuksen, että

$$0 < e(\mathbf{x}) < 1 \tag{2.10}$$

kaikilla $\mathbf{x}$:n arvoilla. Jos esimerkiksi $e(\mathbf{x}) = 0$ jollain kovariaattivektorin $\mathbf{x}$ arvolla, niin sellaisten yksiköiden todennäköisyys saada käsittely on 0. Näillä $\mathbf{x}$:n arvoilla ei voida tarkastella käsittelyn vaikutusta, koska näitä yksiköitä on vain vertailuryhmässä. Vertailu siis rajataan niihin $\mathbf{x}$:n arvoihin, joita vastaavia yksiköitä on molemmissa ryhmissä.

Oletukset (2.7) ja (2.10) yhdessä, eli

$$(Y_0, Y_1) \perp\!\!\!\perp Z|\mathbf{X} \text{ ja } 0 < e(\mathbf{x}) < 1 \tag{2.11}$$

ovat jatkotarkastelujen kannalta keskeisessä asemassa. Jos oletus (2.11) pitää paikkansa, niin Rosenbaumin ja Rubinin (1983) käyttämän sanonnan mukaan käsittelyn liittyminen yksikköön on "strongly ignorable", kun kovariaattien $\mathbf{X}$ arvo on annettu.

On syytä korostaa, että käytännön sovelluksissa ei koskaan voi olla täysin varma siitä, pitääkö oletus (2.11) paikkansa. Käytännössä voidaan vain pyrkiä ottamaan mukaan kovariaateiksi kaikki sellaiset muuttujat, jotka mahdollisesti vaikuttavat vasteeseen tai yksiköiden valikoitumiseen käsittely- tai vertailuryhmään. Oletus, että kaikki relevantit kovariaatit on otettu mukaan, on järkevä likiarvo käytännön tilanteessa. Joidenkin relevanttien muuttujien puuttuminen aiheuttaa käsittelyvaikutuksen estimaattiin piiloharhaa, joten tarpeellisten kovariaattien valitseminen on tärkeä osa tutkimusprosessia.

Vaikka oletus (2.11) pitäisikin paikkansa, eivät kovariaattien havaitut jakaumat koe- ja vertailuryhmissä ole tavallisesti samat satunnaisvaihtelun vuoksi. Tämä koskee sekä kokeellista, että havainnoivaa tilannetta. Satunnnaisvaihtelusta johtuva empiiristen jakaumien erilaisuus tasoittuu otoskoon kasvaessa, joten se on pienten otosten ongelma.

2.2.2 Käsittelyvaikutus havainnoivassa tutkimuksessa

Käsittelyyn liittyvä vaste Y_1 havaitaan vain silloin, kun yksikkö saa käsittelyn eli $Z = 1$. Jos satunnaisesti valittua käsittelyryhmän yksikköä ($Z = 1$) verrataan satunnaisesti valittuun kontrolliryhmän yksikköön ($Z = 0$), vasteiden odotettu erotus on

$$E(Y_1|Z = 1) - E(Y_0|Z = 0). \tag{2.12}$$

Kuten alaluvussa 2.1.2 todettiin, erotus (2.12) ei ole yleensä odotettu käsittelyvaikutus (2.3).

Oletetaan, että kovariaattivektorin arvo $\mathbf{x}$ valitaan satunnaisesti $\mathbf{X}$:n jakaumasta. Tämän jälkeen valitaan sellainen käsittely- ja vertailuyksikkö, että niillä on tämä sama kovariaattivektorin arvo $\mathbf{x}$ (vrt. oletus (2.10)). Käsittelyvaikutuksen odotusarvo tässä 2-vaiheisessa otannassa voidaan lausua muodossa

$$E_{\mathbf{X}}[E(Y_1|\mathbf{X}, Z = 1) - E(Y_0|\mathbf{X}, Z = 0)], \tag{2.13}$$

missä $E_{\mathbf{X}}$ on odotusarvo $\mathbf{X}$:n jakauman suhteen. Jos oletus (2.7) pitää paikkansa, niin ehdollisesta riippumattomuudesta seuraa tuloksen (C.7) mukaan (ks. liite C)

$$E(Y_1|\mathbf{X}, Z = 1) = E(Y_1|\mathbf{X}) \text{ ja } E(Y_0|\mathbf{X}, Z = 0) = E(Y_0|\mathbf{X}). \tag{2.14}$$

Nyt odotusarvo (2.13) voidaan (2.14):n nojalla lausua muodossa

$$E_{\mathbf{X}}[E(Y_1|\mathbf{X}) - E(Y_0|\mathbf{X})].$$

Koska odotusarvo on lineaarinen operaattori ja $E_{\mathbf{X}}[E(Y_i|\mathbf{X})] = E(Y_i)$, $i = 0, 1$ (Casella & Berger 2001, s.164), niin saadaan tulos (Rosenbaum & Rubin 1983)

$$\begin{aligned} &E_{\mathbf{X}}[E(Y_1|\mathbf{X}, Z = 1)] - E_{\mathbf{X}}[E(Y_0|\mathbf{X}, Z = 0)] \\ &= E(Y_1) - E(Y_0) \\ &= \tau. \end{aligned} \tag{2.15}$$

Jos oletetaan (2.11), niin odotettu käsittelyvaikutus voidaan estimoida harhattomasti ehdollistamalla $\mathbf{X}$:n arvoihin.

2.3 Luokittelu vastaavuuspistemäärän avulla

Koska havainnoivassa tutkimusasetelmassa käsittely- ja vertailuryhmät saattavat poiketa systemaattisesti oleellisten taustamuuttujien (kovariaattien) suhteen, on nämä erot otettava ryhmien vertailussa huomioon. Eräs tavallinen

menettely näiden systemaattisten erojen hallitsemiseksi on aineiston luokittelu kovariaattivektorin $\mathbf{x}$ havaittujen arvojen perusteella siten, että $\mathbf{x}$:n vaihtelu kunkin luokan sisällä olisi mahdollisimman pieni. Käsittelyvaikutusta tutkitaan sitten erikseen näissä $\mathbf{x}$:n luokissa. Keskimääräinen käsittelyvaikutus voidaan estimoida harhattomasti jokaisessa luokassa, jos $\mathbf{x}$:n arvo luokan sisällä ei vaihtele. Luokittelulla voidaan yleensä pienentää harhaa erittäin oleellisesti, vaikka $\mathbf{x}$ ei olekaan luokan sisällä vakio (Cochran 1968; Rosenbaum & Rubin 1984). Koko aineiston keskimääräisen käsittelyvaikutuksen estimaatti saadaan luokkaestimaattien painotettuna keskiarvona, missä painoina ovat luokkien suhteelliset frekvenssit.

Edellä kuvattu luokittelumenetelmä saattaa olla käyttökelpoinen, jos kovariaatteja on vähän ja $\mathbf{x}$:n mahdollisten arvojen joukko ei ole kovin suuri. Lonkkamurtumapotilaiden hoitokustannuksia käsittelevässä tutkimuksessani on 31 kovariaattia (ks. liite A), joten valitsemalla esimerkiksi vain 2 luokkaa jokaista kovariaattia kohti, saadaan yhteensä $2^{31} = 2147483648$ luokkaa. Luokkien lukumäärä olisi siis paljon suurempi kuin havaintojen lukumäärä. Silloin tyhjiä luokkia on runsaasti, samoin luokkia, joissa on vain joko käsittelyryhmän tai vertailuryhmän yksiköitä. Kovariaattien vaikutuksen kontrollointi on siis jokseenkin mahdotonta suoraan $\mathbf{x}$:n arvoihin perustuvalla luokittelulla, jos kovariaattien lukumäärä p on suuri (suhteessa havaintojen lukumäärään).

Vastaavuuspistemäärä (ks. alaluku 2.2.1) $e(\mathbf{x}) = P(Z = 1|\mathbf{X} = \mathbf{x})$ on todennäköisyys, että tilastoyksikkö kuuluu käsittelyryhmään, kun kovariaattien arvovektori $\mathbf{x}$ on annettu. Rosenbaum & Rubin (1983) osoittivat, että käsittely- ja vertailuryhmien erot kovariaattien jakauman suhteen voidaan tasapainottaa vastaavuuspistemäärän $e(\mathbf{x})$ arvojen perusteella.

Esimerkki 2.5. Olkoon kahdella potilaalla, joista toinen kuuluu käsittelyryhmään ja toinen vertailuryhmään, sama vastaavuuspistemäärä e. Silloin näillä potilailla on sama todennäköisyys e kuulua käsittelyryhmään ja vastaavasti sama todennäköisyys $1 - e$ kuulua vertailuryhmään. Oletetaan, että emme tiedä, kumpi potilaista kuuluu käsittelyryhmään, mutta meidän pitäisi arvata. Koska kummankin potilaan vastaavuuspistemäärä on e, heillä on samat mahdollisuudet kuulua käsittelyryhmään. Tällaisessa tilanteessa "paras" arvaus saadaan heittämällä harhatonta lanttia. Huomattakoon, että näiden kahden potilaan (sanokaamme k ja v) kovariaattivektorit voivat olla varsin erilaiset ($\mathbf{x}_k \neq \mathbf{x}_v$). Kovariaattivektorien $\mathbf{x}_k$ ja $\mathbf{x}_v$ tunteminen ei siis auta parantamaan arvaustamme, jos $e(\mathbf{x}_k) = e(\mathbf{x}_v)$. □

Koska $e(\mathbf{x})$ on kovariaattien $\mathbf{x}$ skalaariarvoinen funktio, niin edellä kuvattu moniulotteinen luokitteluongelma voidaan palauttaa yksiulotteiseksi. Oletetaan, että jokaista $e(\mathbf{x})$:n havaittua arvoa kohti on olemassa ainakin yksi käsittelyryhmän ja yksi vertailuryhmän yksikkö. Nyt odotettu käsittelyvaikutus voidaan estimoida harhattomasti jokaista kiinnitettyä $e(\mathbf{x})$:n arvoa kohti. Huomaa, että tämän menetelmän teoreettinen perustelu esitetään seuraavassa alaluvussa (alaluku 2.4).

Rosenbaum ja Rubin (1983) ehdottivat vastaavuuspistemäärään perustuvaa luokitteluestimaattoria. Seuraavassa esitetään eräs tapa toteuttaa estimaattori käytännössä. Ositetaan aineisto vastaavuuspistemäärän arvojen perusteella ositteisiin (osa-aineistoihin). Olkoon ositteiden lukumäärä L. Määritellään indikaattorimuuttuja I_{il} siten, että

$$I_{il} = \begin{cases} 1, & \text{kun } (l-1)/L < e(\mathbf{X}_i) \leq l/L \text{ ja} \\ 0 & \text{muutoin,} \end{cases} \tag{2.16}$$

missä l ja i ovat sellaiset luonnolliset luvut, että $1 \leq l \leq L-1$ ja $1 \leq i \leq n$. Huomattakoon, että oletuksesta (2.10) johtuen (2.16):ssa viimeisen luokan (l=L) rajat ovat $(l-1)/L < e(\mathbf{X}_i) < l/L$. Vastaavuuspistemäärän arvoalue $(0,1)$ jaetaan siis tasavälisesti L:ään luokkaan. Jos $I_{il} = 1$, niin yksikön i saama vastaavuuspistemäärän arvo kuuluu välille $((l-1)/L, l/L]$ ja sanomme, että yksikkö kuuluu ositteeseen l. Ositteeseen l, $1 \leq l \leq L$, kuuluu $n_{1l} = \sum_{i=1}^{n} Z_i I_{il}$ käsittelyryhmän yksikköä, $n_{0l} = \sum_{i=1}^{n} (1-Z_i) I_{il}$ vertailuryhmän yksikköä ja yhteensä $n_l = n_{0l} + n_{1l}$ yksikköä, missä Z_i:t ovat käsittelymuuttujia (vrt. alaluku 2.1.1). Jokaista yksikköä i kohti Z_i osoittaa, kuuluuko i. yksikkö käsittelyryhmään ($Z_i = 1$) vai vertailuryhmään ($Z_i = 0$).

Nyt käsittelyn ja kontrollin odotettu vaste voidaan estimoida jokaisessa ositteessa l jokseenkin harhattomasti alaluvussa 2.4 esitettävän teorian nojalla (ks. (2.19)) seuraavasti:

$$\hat{\mu}_{1l} = \frac{1}{n_{1l}} \sum_{i=1}^{n} I_{il} Z_i Y_i \quad \text{ja} \quad \hat{\mu}_{0l} = \frac{1}{n_{0l}} \sum_{i=1}^{n} I_{il} (1 - Z_i) Y_i.$$

Silloin käsittelyvaikutuksen harhaton estimaatti ositteessa l on

$$\hat{\tau}_l = \hat{\mu}_{1l} - \hat{\mu}_{0l}.$$

Vastaavat koko populaation keskimääräisten vasteiden (käsittely- ja vertailu) estimaatit ovat

$$\hat{\mu}_1 = \sum_{l=1}^{L} \frac{n_l}{n} \hat{\mu}_{1l} \text{ ja } \hat{\mu}_0 = \sum_{l=1}^{L} \frac{n_l}{n} \hat{\mu}_{0l}.$$

Keskimääräisen käsittelyvaikutuksen estimaatti on siis

$$\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0 = \sum_{l=1}^{L} \frac{n_l}{n} (\hat{\mu}_{1l} - \hat{\mu}_{0l}).$$

Tämä luokittelumenetelmä voidaan ymmärtää epäparametrisen regression (ks. luku 3) alkeelliseksi versioksi, missä regressiofunktion likiarvona käytetään porrasfunktiota. Luokittelumenetelmässä joudutaan tietysti ottamaan kantaa siihen, mikä on riittävä luokkien lukumäärä. Rosenbaum ja Rubin (1984) ovat esittäneet tutkimustuloksia, joissa viittä luokkaa käyttämällä voitiin poistaa yli 90% harhasta.

2.4 Vastaavuuspistemäärään perustuvan päättelyn teoriaa

Tämän alaluvun tulokset perustuvat Rosenbaumin ja Rubinin (1983) esittämään teoriaan. Kuten alaluvussa 2.2.1 todettiin [oletus (2.11)], tämän työn keskeinen oletus on, että

$$(Y_0, Y_1) \perp\!\!\!\perp Z|\mathbf{X} \text{ ja } 0 < e(\mathbf{X}) < 1.$$

Rosenbaum ja Rubin (1983, Lause 1) osoittivat, että $\mathbf{X}$ ja Z ovat ehdollisesti riippumattomat, kun ehdollistetaan $e(\mathbf{X})$:n arvoihin:

$$\mathbf{X} \perp\!\!\!\perp Z|e(\mathbf{X}). \tag{2.17}$$

Jos oletus (2.11) pitää paikkansa, niin silloin (Rosenbaum & Rubin 1983, Lause 3)

$$(Y_0, Y_1) \perp\!\!\!\perp Z|e(\mathbf{X}) \text{ ja } 0 < P(Z = 1|e(\mathbf{X})) < 1. \tag{2.18}$$

Esimerkki 2.6. Oletus (2.17) tarkoittaa siis ehdollisen riippumattomuuden määritelmän nojalla sitä, että $\mathbf{X}$:llä on sama jakauma käsittelyryhmässä ($Z = 1$) ja vertailuryhmässä ($Z = 0$) jokaista kiinnitettyä vastaavuuspistemäärän arvoa $e(\mathbf{x}) = e$ kohti. Vertaillaan nyt sellaisia SHP1:n ja SHP0:n potilaita, joilla on sama vastaavuuspistemäärän arvo $e(\mathbf{X}) = e$. Silloin tuloksen (2.17) mukaan $\mathbf{X}$:n jakauma on molempien piirien potilailla sama. Tässä mielessä siis piirien potilaat ovat samankaltaisia, kun $e(\mathbf{X}) = e$. Jo Esimerkissä 2.5 tarkasteltiin vastaavuuspistemäärään ehdollistamista ja siis tuloksen (2.17) tulkintaa.

Tuloksen (2.18) ja sen seurauksen (2.21) sisältöä voidaan havainnollistaa samalla tavalla kuin Esimerkissä 2.4 havainnollistettiin ehdollistamista yhteen kovariaattiin X. Jos teemme oletuksen (2.11), niin (2.18):n mukaan

$$(Y_0, Y_1) \perp\!\!\!\perp Z|e(\mathbf{x}) = e \text{ ja } 0 < P(Z = 1|e(\mathbf{X}) = e) < 1.$$

Tästä seuraa esimerkiksi relaatio

$$Y_1 \perp\!\!\!\perp Z|e(\mathbf{x}) = e \text{ ja } 0 < P(Z = 1|e(\mathbf{X}) = e) < 1,$$

jonka mukaan sairaanhoitopiirissä SHP1 hoidettavien potilaiden kustannusjakauma on sama piirin SHP1 ja SHP0 potilailla, mikäli potilailla on sama vastaavuuspistemäärä $e(\mathbf{x}) = e$.

Tuloksen $0 < P(Z = 1|e(\mathbf{X}) = e) < 1$ mukaan on positiivinen todennäköisyys, että ehdon $e(\mathbf{X}) = e$ täyttäviä potilaita löytyy kummastakin piiristä. Se on luonnollinen edellytys piirien hoitokustannusten vertailtavuudelle. Havaituissa otoksissa saattaa tietysti käydä niin, että ehdon $e(\mathbf{X}) = e$ täyttäviä on vain toisesta piiristä. □

Tuloksesta (2.18) seuraa alaluvussa 2.2.2 esitettyä tulosta (2.15) vastaava tulos:

$$
\begin{aligned}
(2.19) \qquad & E_{e(\mathbf{X})}[E(Y_1|e(\mathbf{X}), Z = 1)] - E_{e(\mathbf{X})}[E(Y_0|e(\mathbf{X}), Z = 0)] \\
& = E(Y_1) - E(Y_0) \\
& = \tau.
\end{aligned}
$$

Nyt vain satunnaisvektori $\mathbf{X}$ lausekkeessa (2.15) on korvattu satunnaismuuttujalla $e(\mathbf{X})$. Tulos (2.19) voidaan todistaa odotusarvon ominaisuuksien nojalla. Jos (2.18) pätee, niin ehdollisesta riippumattomuudesta seuraa

$$E(Y_i|e(\mathbf{X}), Z = i) = E(Y_i|e(\mathbf{X})), \tag{2.20}$$

missä $i = 0, 1$. Ottamalla odotusarvoista (2.20) puolittain odotusarvot $e(\mathbf{X})$:n jakauman suhteen saadaan kaksinkertaisen odotusarvon ominaisuuksien nojalla (Casella & Berger 2001, s.164)

$$E_{e(\mathbf{X})}[E(Y_i|e(\mathbf{X}), Z = i)] = E(Y_i),$$

missä $i = 0, 1$. Tästä seuraa tulos (2.19) välittömästi.

Ehdollisen riippumattomuuden oletus (2.17) tarkoittaa siis sitä, että $\mathbf{X}$:n jakaumat ovat samat käsittelyryhmässä ja vertailuryhmässä jokaista annettua $e(\mathbf{X})$:n arvoa kohti. Odotettu käsittelyvaikutus voidaan siis estimoida harhattomasti jokaista $e(\mathbf{X})$:n arvoa kohti, kuten teoreettinen tulos (2.19) osoittaa.

Huomattakoon, että tulokset (2.20) saadaan olettamalla

$$Y_0 \perp\!\!\!\perp Z|e(\mathbf{X}) \text{ ja } Y_1 \perp\!\!\!\perp Z|e(\mathbf{X}). \tag{2.21}$$

Relaatiot (2.21) seuraavat Rosenbaumin ja Rubinin (1983) käyttämästä oletuksesta $(Y_0, Y_1) \perp\!\!\!\perp Z|\mathbf{X}$ (ks. alaluku 2.2.1 ja liite B), joka on siis tiukempi oletus kuin (2.21).

Rosenbaum ja Rubin (1983) tarkastelivat vastaavuuspistemäärää ns. *tasapainottavien pistemäärien* teorian yhteydessä ja osoittivat, että vastaavuuspistemäärällä on erityisrooli tässä teoriassa. He todistivat monet vastaavuuspistemäärän ominaisuudet tässä yleisessä teoriassa. Kovariaattien $\mathbf{X}$ funktiota $\mathbf{b}(\mathbf{X})$ sanotaan tasapainottavaksi pistemääräksi, jos

$$\mathbf{X} \perp\!\!\!\perp Z|\mathbf{b}(\mathbf{X}). \tag{2.22}$$

Huomattakoon, että $\mathbf{b}(\mathbf{X})$ voi olla myös vektoriarvoinen. Rosenbaum ja Rubin (1983, Lause 2) osoittivat, että $\mathbf{b}(\mathbf{X})$ on tasapainottava pistemäärä jos ja vain jos $\mathbf{b}(\mathbf{X})$ on hienojakoisempi kuin $e(\mathbf{X})$ siinä mielessä, että $e(\mathbf{X})$ on lausuttavissa $\mathbf{b}(\mathbf{X})$:n funktiona. Toisin sanoen, on olemassa sellainen funktio f, että

$e(\mathbf{X}) = f[\mathbf{b}(\mathbf{X})]$. Koska $\mathbf{X}$ ja $e(\mathbf{X})$ toteuttavat ehdon (2.22), niin $\mathbf{X}$ ja $e(\mathbf{X})$ ovat tasapainottavia pistemääriä. Esimerkiksi tulos (2.19) pysyy voimassa, kun $e(\mathbf{X})$ korvataan jollain tasapainottavalla pistemäärällä $\mathbf{b}(\mathbf{X})$. Tässä työssä käytetään kuitenkin vain vastaavuuspistemäärää.

2.5 Useiden käsittelytulosten vertailu

Jos tarkastellaan useampaa kuin kahta käsittelyä samanaikaisesti, niin käsittelymuuttujaa merkitään K:lla erotuksena binäärisestä käsittelymuuttujasta Z. Käsittely K on siis satunnaismuuttuja, jonka arvojoukko olkoon $\mathcal{K}$. Jos vertaillaan Suomen kaikkia sairaanhoitopiirejä, niin $\mathcal{K} = \{0, 1, 2, \ldots, 19\}$, missä $k \in \mathcal{K}$ on sairaanhoitopiirin numero. Käsittely k tarkoittaa, että potilas on hoidettu k. sairaanhoitopiirissä. Tässä sairaanhoitopiireihin viitataan yksinkertaisuuden vuoksi numeroin, mutta myöhemmin kustannusten vertailuissa käytetään myös piirien oikeita nimiä.

Jokaiseen tilastoyksikköön (potilaaseen) i, $1 \leq i \leq n$, ja jokaiseen käsittelyyn k liittyy mahdollinen tulos (kustannus), jota merkitään Y_{ki}. Olemme kiinnostuneita kaikkien käsittelyiden $k \in \mathcal{K}$ keskimääräisistä tuloksista (vasteista) $E(Y_{ki})$ ja erityisesti tarkastellaan muotoa

$$E(Y_{ki} - Y_{si}) = E(Y_{ki}) - E(Y_{si}) \tag{2.23}$$

olevia erotuksia, missä $s \in \mathcal{K}$ ja $s \neq k$. Erotus (2.23) on keskimääräinen vaikutus, kun yksiköt saavat käsittelyn k sijasta käsittelyn s. Odotusarvo otetaan tutkimuksen kohteena olevan populaation tai jonkin sen osapopulaation yli. Jokaista otokseen valikoitunutta potilasta i kohti havaitaan käsittely K_i, käsittelyyn K_i liittyvä tulos $Y_{K_i i}$ sekä potilaiden kovariaattien ja hoitoa kuvaavien muuttujien arvot $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. Jos potilas i on hoidettu esimerkiksi sairaanhoitopiirissä 2 (saanut käsittelyn 2), niin $K_i = 2$. Silloin havaitaan satunnaismuuttujan Y_{2i} arvo, eli potilaan hoitokustannukset sairaanhoitopiirissä 2.

Vastaavuuspistemäärää käytettäessä piirejä vertaillaan aina pareittain. Silloin 20:stä piiristä saadaan $\binom{20}{2} = 190$ parittaista vertailua. Jos käytetään aina samaa vertailuryhmää, riittää 19 vertailua. Aidossa usean piirin vertailussa pyritään tarkastelemaan kustannuseroja samanaikaisesti. Tähän yleiseen tapaukseen palataan 5. luvussa.

3 Kustannuserot vastaavuuspistemäärän funktiona

Teemme nyt oletuksen (2.11). Silloin odotettu käsittelyvaikutus $\tau = E(Y_1) - E(Y_0)$ voidaan esittää tuloksen (2.15) perusteella muodossa

$$\tau = E_{\mathbf{X}}[g_1(\mathbf{X}) - g_0(\mathbf{X})],$$

missä $g_0(\mathbf{X}) = E(Y|\mathbf{X}, Z = 0)$ ja $g_1(\mathbf{X}) = E(Y|\mathbf{X}, Z = 1)$ ovat regressiofunktioita. Odotettu käsittelyvaikutus on siis keskimääräinen käsittelyvaikutus koko populaatiossa. Käytännössä tarkastellaan kovariaattien funktioita $g_j(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x}, Z = j)$, $j = 0, 1$. Heckman, Ichimura ja Todd (1997, 1998) tarkastelivat ydinestimoinnin käyttöä regressiofunktioiden $g_0(\mathbf{x})$ ja $g_1(\mathbf{x})$ estimoinnissa. He tutkivat tietyn sosiaalisen koulutusohjelman (työhön koulutus) vaikutusta. Traditionaalinen tapa lähestyä regressiofunktioiden $g_0(\mathbf{x})$ ja $g_1(\mathbf{x})$ estimointia on spesifioida niille esimerkiksi lineaariset mallit ja estimoida ne. Ensisijainen tavoite on kuitenkin käsittelyvaikutuksen τ estimointi. Siihen soveltuu epäparametrinen regressio.

Regressiofunktioiden $g_0(\mathbf{x})$:n ja $g_1(\mathbf{x})$:n epäparametrinen estimointi olisi kuitenkin hankalaa, koska kovariaatteja on paljon (vrt. alaluku 2.4). Rosenbaum ja Rubin (1983, Lause 3) osoittivat, että oletuksesta (2.11) seuraa tulos (2.18). Tämän tuloksen nojalla odotettu käsittelyvaikutus on [ks. (2.19)]

$$\tau = E_{e(\mathbf{X})}\{m_1[e(\mathbf{X})] - m_0[e(\mathbf{X})]\}, \tag{3.1}$$

missä $m_1(e) = E(Y|e(\mathbf{X}) = e, Z = 1)$ ja $m_0(e) = E(Y|e(\mathbf{X}) = e, Z = 0)$ ovat odotetut käsittelyvaikutukset koe- ja vertailuryhmässä vastaavuuspistemäärän e funktiona. Koska regressiofunktiot $m_0(e)$ ja $m_1(e)$ ovat yhden muuttujan funktioita, vältetään vastaavuuspistemäärän käytöllä usean muuttujan regressiofunktion estimointi.

3.1 Kustannusten regressiofunktiot

Nyt siis ehdolliset odotusarvot

$$E(Y|e(\mathbf{x}) = e, Z = 1) = m_1(e)$$

ja

$$E(Y|e(\mathbf{x}) = e, Z = 0) = m_0(e)$$

riippuvat kovariaateista vain vastaavuuspistemäärän e kautta. Lausutaan vastaavuuspistemäärään perustuvat regressioyhtälöt nyt muodossa

$$\begin{aligned} Y_1 &= m_1(e) + \epsilon_1 \\ Y_0 &= m_0(e) + \epsilon_0, \end{aligned} \tag{3.2}$$

missä $E(\epsilon_1|e) = E(\epsilon_0|e) = 0$ kaikilla $e = e(\mathbf{x})$. Virhetermit ϵ_1 ja ϵ_0 ovat keskenään riippumattomat ja

$$E(\epsilon_j) = 0 \text{ ja } var(\epsilon_j) = \sigma^2,\ j = 0, 1.$$

3.2 Keskimääräinen käsittelyvaikutus

Jos $f(e)$ on vastaavuuspistemäärän e tiheysfunktio, niin keskimääräinen käsittelyvaikutus (3.1) on

$$\tau = \int [m_1(e) - m_0(e)] f(e) de, \tag{3.3}$$

missä integraali otetaan yli e:n arvoalueen. Keskimääräinen käsittelyvaikutus τ saadaan siis regressiofunktioiden $m_1(e)$ ja $m_0(e)$ avulla.

Vastaavasti keskimääräiset vasteet koe- ja vertailuryhmässä ovat

$$\mu_1 = E(Y_1) = \int m_1(e) f(e) de \tag{3.4}$$

ja

$$\mu_0 = E(Y_0) = \int m_0(e) f(e) de.$$

Integraalit otetaan tavallisesti yli koko e:n arvoalueen, mutta joskus saatetaan olla kiinnostuneita keskimääräisestä vaikutuksesta jollain osavälillä. Jos esimerkiksi välillä $[a, b]$ kustannusero $m_1(e) - m_0(e)$ on erityisen suuri, voidaan tutkia, millaiset kovariaattien yhdistelmät kuvautuvat kyseiselle välille.

3.3 Regressiofunktioiden estimointi

3.3.1 Paikallisesti lineaarinen ydinestimaattori

Regressiofunktiot (3.2) estimoidaan epäparametrisella *paikallisesti lineaarisella ydinestimoinnilla*, joka kuuluu tasoitusmenetelmiin. Tasoituksessa ei määritetä regressiofunktiolle parametrista mallia, vaan funktion muoto haetaan aineiston perusteella. Regressiofunktiosta oletetaan vain, että se on riittävän tasainen. Matemaattisesti tasaisuusehdot ilmaistaan funktion $m(e)$ derivaattoja koskevina oletuksina (esim. Wand & Jones 1995, luku 5).

Tässä regressiofunktion estimointia käsittelevässä alaluvussa vastemuuttujaa merkitään Y:llä ja regressiofunktiota m:llä. Aineistoon sovitetaan paikallisesti ensimmäisen asteen polynomi

$$E(Y_i) = \beta_0 + \beta_1(e_i - e), \tag{3.5}$$

missä Y_i on i. yksikön vaste ja e_i on i. yksikön havaittu vastaavuuspistemäärä (ks. Wand & Jones 1995, s.116). Nyt käytetään jälleen alaluvussa 2.1.1 esiteltyä merkintätapaa, jossa i viittaa tilastoyksikköön. Tämän tutkielman sovelluksessa regressiofunktiot estimoidaan kuitenkin erikseen käsittelyryhmän ($Z_i = 1$ ja $Y_i = Y_{i1}$) ja vertailuryhmän ($Z_i = 0$ ja $Y_i = Y_{i0}$) yksiköille, kuten edellä yhtälöissä (3.2) on esitetty. Tässä alaluvussa tarkastellaan kuitenkin tavallisen yhden selittäjän lineaarisen regressiomallin estimointia yleisesti. Siinä yhteydessä (3.5):n merkintätapa on tavanomainen.

Jokaisessa pisteessä e saadaan tasoitus sovittamalla aineistoon suora (3.5) painotetulla pienimmän neliösumman menetelmällä, missä havaintojen painoina käytetään *ydinpainoja* $K(\frac{e_i - e}{h})$. Ydinfunktio $K(e)$ on tavallisesti yksihuippuinen, symmetrinen ja positiivinen funktio, joka vähenee, kun $|e|$ kasvaa. Tässä työssä käytetään ydinfunktiona normaalijakauman tiheysfunktiota. Parametri $h > 0$ on tasoitusparametri, jota kutsutaan myös ikkunan leveydeksi. Suoran (3.5) parametrien β_0 ja β_1 pienimmän neliösumman estimaatit $\hat{\beta}_0$ ja $\hat{\beta}_1$ saadaan minimoimalla lauseke

$$\sum_{i=1}^{n} [y_i - \beta_0 - \beta_1(e_i - e)]^2 K(\frac{e_i - e}{h}) \tag{3.6}$$

parametrien β_0 ja β_1 suhteen, kun h on kiinnitetty. Paikallisesti lineaarinen regressioestimaatti on $\hat{m}(e) = \hat{\beta}_0$. Jokainen paikallinen regressiotehtävä määrittää estimaatin pisteessä e. Kun e muuttuu, muuttuvat ydinpainot ja estimaatit $\hat{\beta}_0$, $\hat{\beta}_1$ ja $\hat{m}(e)$. Estimaatteja $\hat{\beta}_0$ ja $\hat{\beta}_1$ pitäisi oikeastaan merkitä $\hat{\beta}_0(e)$ ja $\hat{\beta}_1(e)$.

Tavallisen painotetun pienimmän neliösumman estimointiteorian mukaan estimaattivektorin $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1)^T$ lauseke on muotoa

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_e^T \mathbf{W}_e \mathbf{X}_e)^{-1} \mathbf{X}_e^T \mathbf{W}_e \mathbf{y}, \tag{3.7}$$

missä matriisin $\mathbf{X}_e$ transpoosi

$$\mathbf{X}_e^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ e_1 - e & e_2 - e & \dots & e_n - e \end{pmatrix}$$

on $2 \times n$ matriisi,

$$\mathbf{W}_e = diag \left\{ K(\frac{e_1 - e}{h}), \dots, K(\frac{e_n - e}{h}) \right\}$$

on painojen $n \times n$-diagonaalimatriisi ja $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ on $n \times 1$-havaintovektori. Lausekkeessa (3.7) on tietysti oletettava, että matriisi $\mathbf{X}_e^T \mathbf{W}_e \mathbf{X}_e$ on kääntyvä.

Koska regressiofunktion $E(Y|e) = m(e)$ estimaattori on $\hat{\beta}_0$, niin

$$\hat{m}(e) = \mathbf{a}^T (\mathbf{X}_e^T \mathbf{W}_e \mathbf{X}_e)^{-1} \mathbf{X}_e^T \mathbf{W}_e \mathbf{Y}, \tag{3.8}$$

missä $\mathbf{a}^T = (1, 0)$ ja $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$. Nähdään, että paikallinen regressioestimaattori (3.8) on lineaarinen. Kun lausekkeessa (3.8) merkitään

$$(l_1(e), \ldots, l_n(e)) = \mathbf{a}^T (\mathbf{X}_e^T \mathbf{W}_e \mathbf{X}_e)^{-1} \mathbf{X}_e^T \mathbf{W}_e,$$

niin

$$\hat{m}(e) = \sum_{i=1}^{n} l_i(e) Y_i.$$

Tästä saadaan

$$var[\hat{m}(e)] = [\sum_{i=1}^{n} l_i(e)^2] \sigma^2.$$

Kun $var[\hat{m}(e)]$ estimoidaan, tarvitaan σ^2:n estimaatti. Koska estimaattori $\hat{m}(e)$ ei ole harhaton (Wand & Jones 1995, luku 5), paitsi lineaarisen funktion m tapauksessa, on virhevarianssin estimoinnissa otettava mahdollinen harha huomioon. Siksi σ^2:n estimointiin on käytetty Gasserin ym. (1986) ehdottamaa pseudo-residuaalien menetelmää (Bowman & Azzalini 1997, s. 74), joka pyrkii poistamaan mahdollisen harhan vaikutuksen.

Kun tasoitusparametrin arvo lähenee nollaa, estimaattori painottaa voimakkaasti kaikkein lähimpänä e:tä olevia havaintoja ja regressioestimaattori $\hat{m}(e)$ lähenee käyrää, joka pyrkii kulkemaan kaikkien havaintopisteiden kautta. Toisaalta h:n kasvaessa, estimaattori painottaa kaikkia havaintoja jokseenkin yhtä paljon ja estimaatti $\hat{m}(e)$ lähenee pienimmän neliösumman suoraa. Koska tässä työssä on kaikkien käyrien estimoinnissa käytetty samaa h:n arvoa, se on pyritty valitsemaan niin, että se toimii kohtuullisen hyvin kaikissa tarkasteltavissa sairaanhoitopiireissä.

Lokaalia polynomiestimaattoria käytettäessä on valittava polynomin aste, ydinfunktio K ja ikkunan leveys h. Kuten edellä on mainittu, ydinfunktioksi on valittu normaalijakauman tiheysfunktio. Tämä on käytännössä varsin tavanomainen valinta. Polynomin aste sekä ikkunan leveys vaikuttavat estimaattorin $\hat{m}(e)$ varianssiin ja harhaan (Fan & Gijbels 1996, luku 3; Wand & Jones 1995, luku 5). Käytännössä useimmiten päädytään valitsemaan joko 1. tai 2. asteen polynomi, joita pidetään tilastollisesti parempina kuin esimerkiksi lokaalia vakioestimaattoria (ks. esim. Davison 2003, s.522 tai Ruppert, Wand & Carroll 2003, s 86). Tässä työssä on siis valittu polynomin asteeksi 1, vaikka myös 2. asteen polynomi on harkinnan arvoinen vaihtoehto.

3.3.2 Tasoitusparametrin valinta

Tasoitusparametrin h valintaan liittyvää teoriaa on käsitelty alan kirjallisuudessa runsaasti (esim. Fan & Gijbels 1996, luku 3). Parametrin h valinta muistuttaa mallinvalintaongelmaa. Liian pienen h:n valinta johtaa mallin yliestimointiin, estimaattorin $\hat{m}(e)$ varianssi kasvaa ja harha pienenee. Liian suurella h:n arvolla malli aliestimoituu, $\hat{m}(e)$:n varianssi pienenee ja harha kasvaa.

Tässä työssä parametrin h valintaa on tarkasteltu ristiinvalidoimisen (cross-validation) menetelmällä. Valitaan h siten, että neliösumma

$$CV(h) = \sum_{i=1}^{n} [y_i - \hat{m}_{-i}(e_i)]^2 \tag{3.9}$$

minimoituu (Bowman & Azzalini 1997, alaluku 4.5). Estimaatti $\hat{m}_{-i}(e_i)$ on laskettu aineistosta, josta i. yksikkö on jätetty pois. Ideana on siis ennustaa jokaista vastemuuttujan arvoa y_i jäljellä olevalla aineistolla. Tässä työssä tarkastellaan kuitenkin useita osa-aineistoja, joissa kriteerin (3.9) perusteella päädyttäisiin erilaisiin h:n valintoihin. Siksi on valittu vain yksi h:n arvo h_0 siten, että kaikissa tarkasteltavissa osa-aineistoissa arvo $CV(h_0)$ on kohtuullisen lähellä minimiarvoa.

3.4 Vaihteluvälit

Tiettyjen ehtojen vallitessa estimaattori $\hat{m}(e)$ noudattaa normaalijakaumaa (Fan & Gijbels 1996, luku 3). Silloin

$$\frac{\hat{m}(e) - E[\hat{m}(e)]}{\sqrt{\hat{v}(e)}} \tag{3.10}$$

noudattaa likimain standardoitua normaalijakaumaa, missä $\hat{v}(e)$ on varianssin $var[\hat{m}(e)]$ estimaatti. Kun lasketaan estimaatit $\hat{m}(e)$ ja $\hat{v}(e)$, voidaan jokaisessa pisteessä e muodostaa odotusarvon $E[\hat{m}(e)]$ $100(1-2\alpha)\%$ luottamusvälin estimaatti

$$\hat{m}(e) \pm z_\alpha \sqrt{\hat{v}(e)}, \tag{3.11}$$

missä z_α on standardoidun normaalijakauman α-yläfraktiili. Koska $\hat{m}(e)$ ei ole harhaton, (3.11) ei ole $m(e)$:n luottamusväli. Eräs tapa ratkaista ongelma on pyrkiä korjaamaan väliä (3.11) harhan

$$B_h(e) = E[\hat{m}_h(e)] - m(e)$$

estimaatilla, missä merkinnät $B_h(e)$ ja $\hat{m}_h(e)$ korostavat estimaattoreiden riippuvuutta h:sta. Tämä lähestymistapa osoittautuu kuitenkin hankalaksi. Tässä

työssä on omaksuttu vaihtoehtoinen yksinkertaisempi tapa, jossa ei yritetä korjata väliä (3.11) harhan estimaatilla. Välit (3.11) kuvaavat kyllä regressioestimaattorin vaihtelua jokaisessa pisteessä e, mutta tulkinnassa on otettava huomioon mahdollinen harha. Regressioestimaattorin $\hat{m}_h(e)$ teoriasta tiedetään, että harha $B_h(e)$ pienenee, kun h pienenee.

Vertailemalla eri h:n arvoilla saatuja $m(e)$:n estimaatteja, voitaisiin saada käsitys harhan suuruudesta. Tämä olisi eräs tapa arvioida, kuinka hyvä $m(e)$:n luottamusvälin likiarvo väli (3.11) on. Tässä työssä ei ole kuitenkaan pyritty systemaattiseen harhan arviointiin, vaan se tulee olemaan eräs jatkotutkimuksen aihe. Käytännössä $\hat{m}(e)$, $\hat{var}[\hat{m}(e)]$ ja välit (3.11) on estimoitu Bowmanin ja Azzalinin kirjassa esitetyllä tekniikalla (Bowman & Azzalini 1997, luku 4).

Jos jonkin pisteen e_0 ympäristössä on havaintopisteitä harvassa, $\hat{m}(e_0)$:n estimoinnin tarkkuus pienenee (ks. esim. Davison 2003, s. 522). Erityisesti e:n vaihtelualueen reunoilla saattaa estimointi olla tästä syystä epätarkkaa. Kokonaisuutena lonkkamurtuma-aineistossa on runsaasti havaintoja, joten pienten aineistojen ongelmaa ei tässä tapauksessa ole.

3.5 Keskimääräisten vasteiden ja käsittelyvaikutuksen estimointi

Kun estimaattorit $\hat{m}_1(e)$ ja $\hat{m}_0(e)$ ovat käytettävissä, voidaan keskimääräinen käsittelyvaikutus τ estimoida niiden avulla:

$$\hat{\tau}_e = \frac{1}{n}\sum_{i=1}^{n}\{\hat{m}_1[e(\mathbf{x}_i)] - \hat{m}_0[e(\mathbf{x}_i)]\}, \tag{3.12}$$

missä n on havaintojen lukumäärä. Vaihtoehtoinen τ:n estimaattori saadaan kaavan (3.3) avulla. Estimoidaan ensin e:n tiheysfunktio esimerkiksi ydinestimaattorilla (ks. Wand & Jones 1995, luku 2). Lasketaan sitten integraali (3.3) numeerisesti, kun funktiot $m_1(e)$, $m_0(e)$ ja $f(e)$ korvataan estimaateillaan.

Keskimääräisten vasteiden (3.4) estimaatit ovat vastaavasti

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_1[e(\mathbf{x}_i)] \text{ ja } \hat{\mu}_0 = \frac{1}{n}\sum_{i=1}^{n}\hat{m}_0[e(\mathbf{x}_i)].$$

Tässä yhteydessä voidaan tarkastella myös ehdollisten (ehdolla $e(\mathbf{x}) = e$) vasteiden keskimääräistä vaihtelua. Voidaan ajatella, että keskimääräinen hajonta ryhmässä j on

$$\sigma_j = \int \sqrt{v_j(e)} f(e) de,$$

missä $j = 0, 1$. Sen estimaattina voidaan käyttää lauseketta

$$\hat{\sigma}_j = \frac{1}{n}\sum_{i=1}^{n}\sqrt{\hat{v}[e(\mathbf{x}_i)]},\ j = 0, 1.$$

Keskimääräisen hajonnan estimaatin avulla voidaan vertailla ehdollisten vasteiden (ehdolla e) keskimääräistä vaihtelua.

4 Kustannusten vertailu

Alaluvussa 2.3 tarkasteltiin $\mathbf{x}$:n arvoihin perustuvan ehdollistamisen ongelmallisuutta, kun kovariaatteja on paljon. Lonkkamurtumapotilaiden kustannuksia käsittelevässä tutkimuksessani on 31 kovariaattia ja 16881 havaintoa. Kovariaattien suuri määrä aiheuttaa sen, että suoraan $\mathbf{x}$:n arvoihin perustuva luokittelumenetelmä ei toimi. Sairaanhoitopiirien välinen käsittelyvaikutus voidaan kuitenkin estimoida käyttämällä vastaavuuspistemäärään perustuvaa yksiulotteista luokittelumenetelmää.

Havainnoivassa tutkimuksessa vastaavuuspistemäärä $e(\mathbf{x})$ on kovariaattivektorin $\mathbf{x}$ tuntematon funktio. Tässä kokeellinen ja havainnoiva tutkimus poikkeavat toisistaan. Esimerkiksi täysin satunnaistetussa koeasetelmassa (käsittely vs. kontrolli) kaikkien potilaiden i, $1 < i < n$, vastaavuuspistemäärä olisi $e(\mathbf{x}) = 1/2$. Tuntematon vastaavuuspistemäärä $e(\mathbf{x})$ voidaan kuitenkin estimoida aineistosta. Vaikka yksiköiden vastaavuuspistemäärät tunnettaisiinkin, saatetaan saada parempia käsittelyvaikutuksen estimaatteja käyttämällä tunnetun $e(\mathbf{x})$:n sijasta sen estimaattia. Tätä hieman yllättävää ilmiötä ovat tutkineet esimerkiksi Rosenbaum (1987) ja Rubin ja Thomas (1996). Yleinen johtopäätös on, että estimoidun vastaavuuspistemäään käyttö toimii varsin hyvin.

Tässä työssä vastaavuuspistemäärät estimoidaan käyttäen parametrista lineaarista logistista regressiota. Se on Rosenbaumin ja Rubinin (1984) alunperin ehdottama lähestymistapa, jota käytännössä tavallisimmin sovelletaan (D'Agostino 1998). Vastaavuuspistemäärän käyttöön perustuva vasteiden ja käsittelyvaikutusten estimointi tehdään käytännössä kolmivaiheisesti:

1. Estimoidaan vastaavuuspistemäärä $e(\mathbf{x})$.

2. Estimoidaan ehdolliset odotusarvot $m_i(e) = E(Y|e(\mathbf{x}) = e, Z = i)$, $i = 0, 1$. Estimoinnissa voidaan käyttää joko luokittelumenetelmää (alaluku 2.3) tai epäparametrista paikallisesti lineaarista ydinestimaattoria (alaluku 3.3).

3. Estimoidaan keskimääräiset vasteet laskemalla luokkaestimaattien painotetut keskiarvot (luokittelumenetelmä) tai estimoitujen regressiokäyrien integraalit vastaavuuspistemäärän yli (paikallisesti lineaarinen ydinestimaattori).

4.1 Vastaavuuspistemäärän estimointi

Lonkkamurtuma-aineistossa sairaanhoitopiiri vastaa käsittelyä. Nyt tarkastellaan esimerkkinä Helsingin ja Uudenmaan (HUS) ja Varsinais-Suomen sairaanhoitopiirien muodostamaa osa-aineistoa, jossa on 5152 potilasta. Tämä on siis alaluvussa 2.1.1 kuvattu kahden käsittelyn tilanne, jossa käsittelymuuttuja Z on binäärinen. Verailuryhmäksi ($Z = 0$) valitaan HUS, joten Varsinais-Suomen sairaanhoitopiiri on käsittelyryhmä ($Z = 1$). Kuten jo aiemmin on todettu, tässä työssä valitaan aina HUS vertailuryhmäksi.

Merkitään potilaan i todennäköisyyttä kuulua Varsinais-Suomen sairaanhoitopiiriin $P(Z = 1|\mathbf{X}_i = \mathbf{x}_i) = e(\mathbf{x}_i)$, missä $\mathbf{x}_i$ sisältää potilaan i kovariaattien $(X_{i1}, \ldots, X_{i,31})$ arvot. Vastaavuuspistemäärä estimoidaan käyttäen binomista logit-mallia (esim. Agresti 1990; Dobson 2001)

$$E(Y_i|\mathbf{X}_i = \mathbf{x}_i) = log\left(\frac{e(\mathbf{x}_i)}{1 - e(\mathbf{x}_i)}\right) = \alpha + \boldsymbol{\beta}'\mathbf{x}_i, \tag{4.1}$$

$1 \leq i \leq n$, missä α ja $\boldsymbol{\beta}$ ovat estimoitavia parametreja ja $e(\mathbf{x}_i)$ on potilaan i vastaavuuspistemäärä.

Malliin (4.1) voidaan tietysti ottaa mukaan myös kovariaattien yhdysvaikutustermejä. Esimerkiksi kahden kovariaatin yhdysvaikutustermejä on yhteensä $\binom{31}{2} = 465$. Jokaista parittaista vertailua (171 kappaletta) kohti on estimoitava oma logit-mallinsa. Yhdenkin mallin estimointi olisi mittava mallinvalintatehtävä ja mukaan tulevat muuttujat vaihtelevat mallista toiseen. Siksi työn tässä vaiheessa on vastaavuuspistemääriä estimoitaessa käytetty mallien vertailtavuuden helpottamiseksi aina samoja kovariaatteja selittäjinä. Vastaavuuspistemäärän perustarkoitus on korvata suuri määrä kovariaatteja yhdellä funktiolla, jonka arvoihin päättely voidaan ehdollistaa. Silloin voidaan menetellä ikään kuin olisi vain yksi kovariaatti.

Malli (4.1) estimoitiin suurimman uskottavuuden menetelmällä. HUS:n ja Varsinais-Suomen sairaanhoitopiirin potilaista muodostetusta osa-aineistosta lasketut estimaatit $\hat{e}(\mathbf{x}_i)$, $1 \leq i \leq n$, ovat siis potilaiden i estimoituja todennäköisyyksiä kuulua Varsinais-Suomen piiriin. Myöhemmin analyyseissä käytetään logit-muunnettuja vastaavuuspistemääriä

$$logit[\hat{e}(\mathbf{x}_i)] = log\left(\frac{\hat{e}(\mathbf{x}_i)}{1 - \hat{e}(\mathbf{x}_i)}\right), \tag{4.2}$$

missä log on luonnollinen logaritmi. Logit-muunnettuja vastaavuuspistemääriä kutsutaan seuraavassa lyhyesti logit-pistemääriksi. Logit-pistemäärä voi saada arvoja väliltä $(-\infty, \infty)$, kun taas vastaavuuspistemäärä vaihtelee välillä $(0, 1)$. Logit-pistemäärät sisältävät saman informaation kuin vastaavuuspistemäärät ja ne voidaan tarvittaessa helposti muuntaa vastaavuuspistemääriksi. Jatkossa selviää yhteydestä, kummasta pistemäärästä on kyse.

Mallissa (4.1) käytetyt selittäjät on esitetty alaluvussa 1.1.3 ja liitteessä A. Kaikkia muita selittäjiä on käsitelty dummy muuttujina, paitsi hoitopäivien lukumäärää murtumaa edeltäneiden 60 päivän aikana, hoitopäivien lukumäärää murtumaa edeltäneen vuoden aikana, kotona vietettyjen päivien lukumäärää murtuman ja murtumaa edeltäneen hoitojakson välissä, hoitopäivien lukumäärää murtumaa seuranneen vuoden aikana sekä elinpäiviä murtumaa seuranneen vuoden aikana. Mallissa on yhteensä 31 selittäjää.

4.2 Käsittelyvaikutuksen estimointi: luokittelu vs. regressio

Jokaiselle potilaalle saadaan nyt oma, logit-mallin avulla estimoitu vastaavuuspistemääränsä. Alaluvussa 2.3 esitetyn menetelmän mukaan HUS:n ja Varsinais-Suomen potilaat sisältävä aineisto jaetaan vastaavuuspistemäärän mukaan luokkiin ja odotetut vasteet sekä käsittelyvaikutus (sairaanhoitopiirin vaikutus kustannuksiin) voidaan estimoida jokaisessa luokassa.

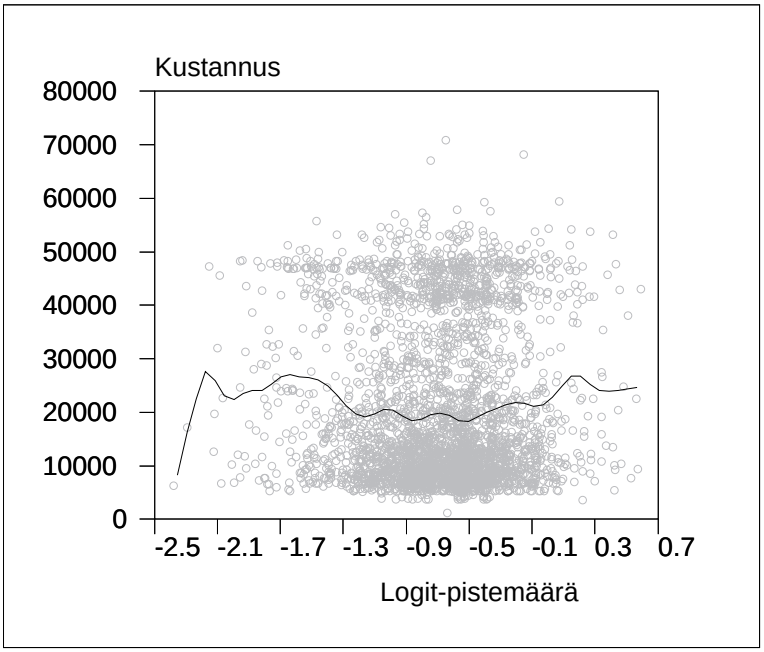
Käsittelyvaikutus voidaan estimoida laskemalla ensin jokaisessa luokassa käsittelyryhmän (Varsinais-Suomen sairaanhoitopiiri) kustannusten ja vertailuryhmän (HUS) kustannusten painotetut keskiarvot ja muodostamalla keskiarvojen erotus. Painoina käytetään luokkien suhteellisia frekvenssejä. Käsittelyvaikutuksen estimointi on harhatonta, jos oletukset (2.11) ovat voimassa. Nyt $\mathbf{X}$ sisältää kaikki kovariaatit, jotka aineiston perusteella ovat käytettävissä ja joiden voidaan ajatella vaikuttavan kustannuksiin tai potilaiden valikoitumiseen sairaanhoitopiireihin. Näin ollen alaluvussa 2.2.1 selvitetyin perustein on tehty riippumattomuusoletus (2.11). Oletuksen (2.11) ehto $0 < e(\mathbf{x}) < 1$ toteutuu nyt estimoitujen vastaavuuspistemäärien tapauksessa kaikilla havaituilla kovariaattien arvoilla $\mathbf{x}$.

Vastaavuuspistemäärät voidaan estimoida kaikista kaksi sairaanhoitopiiriä sisältävistä osa-aineistoista. Tällöin saadaan jokaiselle potilaalle yksi vastaavuuspistemäärän estimaatti jokaisesta estimoidusta mallista. On kuitenkin huomattava, että kutakin vastaavuuspistemäärää voidaan käyttää ainoastaan näiden kyseisten sairaanhoitopiirien välisen käsittelyvaikutuksen estimointiin.

Alaluvussa 2.3 palautettiin moniulotteinen luokitteluongelma vastaavuuspistemäärän avulla yksiulotteiseksi. Luokitellun aineiston jakauma voidaan esittää histogrammina, mutta luokkien leveydellä ja luokkarajoilla on suuri vaikutus histogrammiin. Jakauman muoto saattaa muuttua radikaalisti, kun luokkarajoja siirrellään. Histogrammi esittää todennäköisyystiheyden porrasfunktiona, joka on useimmiten vain tiheysfunktion karkea likiarvo. Kustannuksia estimoitaessa käytetäänkin sen vuoksi luokittelumenetelmän sijasta alaluvussa 3.3 esiteltyä regressiotekniikkaa. Voidaan osoittaa, että histogrammiin perustuva tiheysfunktion estimaattori ei ole yhtä tehokas kuin regressiofunktion estimoinnissa käytetty ydinestimaattori (Wand & Jones, 1995 s.23).

4.3 Kustannusten tasoitus paikallisesti lineaarisella ydinestimaattorilla

Regressiofunktioiden $E(Y_1|e)$ ja $E(Y_0|e)$ estimaatit $\hat{m}_1(e)$ ja $\hat{m}_0(e)$ on laskettu aineistosta alaluvussa 3.3 esitellyllä tavalla. Kuvioon 4.1 on piirretty HUS:n potilaiden pisteparvi, kun x-akselilla on logit-pistemäärä (4.2) ja y-akselilla kustannukset euroissa. Kuviossa 4.1 on HUS:n estimoitu kustannuskäyrä $\hat{m}_0(e)$, joka on saatu käyttäen lokaalia lineaarista ydinestimattoria alaluvussa 3.3 kuvatulla tavalla.



Kuva 4.1. HUS:in potilaiden kustannusten pisteparvi ja tasoitetut kustannukset logit-pistemäärän funktiona.

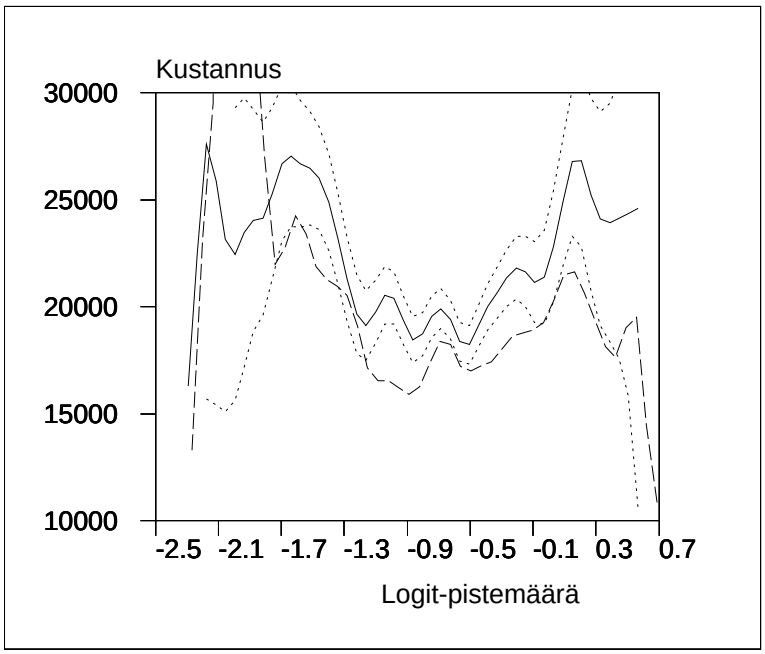
Regressiofunktioille lasketaan myös alaluvussa 3.4 esitellyt vaihteluvälit. Kuvioon 4.1 vaihteluvälejä ei ole piirretty, mutta Kuviossa 4.2 on HUS:n kustannuksille piirretty kahden hajonnan levyiset vaihteluvälit.

4.4 Sairaanhoitopiirien kustannusten vertailu vastaavuuspistemäärän yli tasoitettujen kustannusfunktioiden avulla

Tarkastellaan nyt HUS:in ja Varsinais-Suomen sairaanhoitopiirin estimoituja kustannusfunktioita $\hat{m}_0(e)$ ja $\hat{m}_1(e)$. Luvuissa 2 ja 3 esitettyjen tulosten perusteella funktiot ovat vertailukelpoisia keskenään. Alaluvussa 3.2 esitettiin, miten sairaanhoitopiirien välinen keskimääräinen käsittelyvaikutus τ voidaan estimoida. Nyt ei kuitenkaan tarkastella piirien välisiä keskimääräisiä käsittelyvaikutuksia, jotka ovat skalaariarvoisia. Sen sijaan vertaillaan ehdollisia käsittely-

vaikutuksia eli piirien välistä kustannuseroa vastaavuuspistemäärän funktiona. Näin voidaan tutkia kustannusten käyttäytymistä vastaavuuspistemäärän eri arvoilla.

Sairaanhoitopiirien kustannuskäyrien vertailut on nyt siis tehtävä pareittain, koska vastaavuuspistemäärät on estimoitu pareittaisista osa-aineistoista. Luvun 2 tulosten perusteella on selvää, että vastaavuuspistemäärät on estimoitava uudelleen, kun halutaan vertailla uutta sairaanhoitopiiriä HUS:iin. Kuvioon 4.2 on piirretty HUS:in sekä Varsinais-Suomen sairaanhoitopiirin kustannuskäyrät. HUS:in käyrälle on piirretty kahden hajonnan vaihteluväli, jotta kustannusfunktioiden eroja on helpompi vertailla.

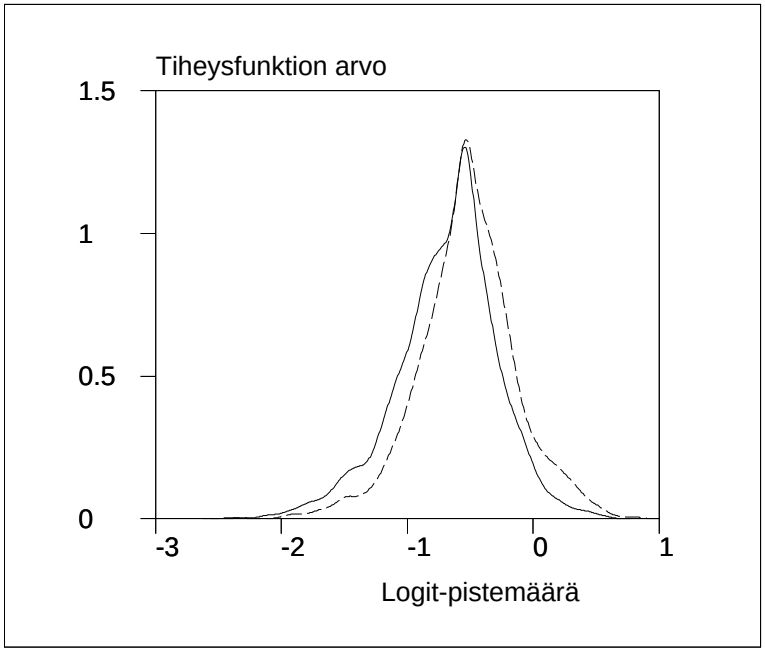


Kuva 4.2. HUS:n kustannukset (——-), HUS:n kustannusten kahden hajonnan levyinen vaihteluväli (.....) ja Varsinais-Suomen sairaanhoitopiirin kustannukset (– – – –). Logit-pistemäärä on x-akselilla ja kustannukset y-akselilla.

Kun tarkastellaan käyriä suurimmilla ja pienimmillä logit-pistemäärien arvoilla, nähdään, että kuvion reunoilla vaihteluvälit kasvavat nopeasti hyvin leveiksi. Verrattaessa Kuviota 4.2 Kuvioon 4.1 havaitaan, että pisteparven reunoilla havainnot käyvät vähiin ja tämä aiheuttaa vaihteluvälin leviämisen. On tarkoituksenmukaista vertailla käyriä ainoastaan siellä, missä havaintoja on riittävästi. Näin vältetään reuna-alueilla esiintyvät estimointiongelmat. Jos havaintoja on hyvin vähän, käyrien estimaatit eivät ole luotettavia. Luotettavan tarkastelualueen määrittämiseksi on tutkittava myös vastaavuuspistemäärän tiheysfunktioita vertailtavissa piireissä. Kuvion 4.3 estimoitujen tiheysfunktioiden perusteella voidaan karkeasti arvioida, että Kuvion 4.2 käyrät ovat vertailukelpoisia esimerkiksi logit-pistemäärän arvosta -1.3 arvoon 0.2.

Kuvioon 4.3 on piirretty HUS:n ja Varsinais-Suomen sairaanhoitopiirin logit-

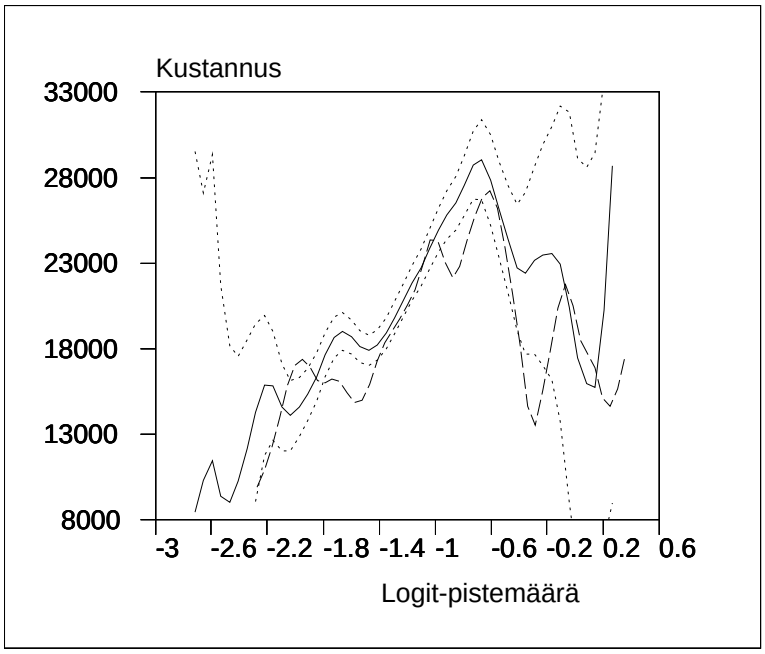
pistemäärien estimoidut tiheysfunktiot. Tiheysfunktiot on estimoitu käyttämällä ydinestimaattoria (ks. esim. Wand & Jones 1995, luku 2). HUS:n ja Varsinais-Suomen sairaanhoitopiirien logit-pistemäärien jakaumien tiheysfunktiot ovat Kuvion 4.3 perusteella melkein identtiset. Piirien kustannukset ovat vertailtavissa vain silloin, kun vastaavuuspistemäärien jakaumat menevät riittävästi päällekkäin. Silloin kutakin annettua vastaavuuspistemäärän arvoa kohti on havaintoja molemmissa vertailtavissa piireissä. Kaikissa tämän työn vertailuissa tilanne on Kuvion 4.3 kaltainen. Suurilla ja pienillä logit-pistemäärän arvoilla ei kuitenkaan ole riittävää määrää havaintoja molemmista piireistä. Näillä arvoilla ei kustannusten vertailu ole tietenkään mielekästä. Kun ehdollistetaan vastaavuuspistemääriin, on aina tarkistettava, millä vastaavuuspistemäärien arvoilla ryhmät ovat vertailtavissa.



Kuva 4.3. Logit-pistemäärien tasoitetut tiheysfunktiot HUS:issa (——) ja Varsinais-Suomen sairaanhoitopiirissä (– – – –).

Kuten Kuviosta 4.2 nähdään, Varsinais-Suomen sairaanhoitopiirin kustannuskäyrä kulkee miltei koko ajan HUS:n vaihteluvälin alarajan alapuolella siellä, missä käyrät ovat vertailukelpoisia. Tämän tuloksen perusteella Varsinais-Suomen sairaanhoitopiirin kustannukset ovat siis HUS:n kustannuksia alhaisemmat, kun potilaiden eroavaisuuksista johtuva vaikutus kustannuksiin on poistettu.

Kuvioon 4.4 on piirretty HUS:n ja Satakunnan sairaanhoitopiirin kustannuskäyrät. Käyrät on estimoitu samalla tavalla kuin HUS:n ja Varsinais-Suomen sairaanhoitopiirin käyrät. Myös kuvioon 4.4 on piirretty HUS:n kustannuskäyrän vaihteluväli. Käyrät ovat nyt vertailukelpoisia logit-pistemäärän arvosta -2.0 arvoon -0.6, mikä on päätelty logit-pistemäärän estimoitujen tiheysfunktioiden avulla (niitä ei ole tässä esitetty).



Kuva 4.4. HUS:n kustannukset (——-), HUS:n kustannusten kahden hajonnan levyinen vaihteluväli (.....) ja Satakunnan sairaanhoitopiirin kustannukset (- - - -).

Satakunnan sairaanhoitopiirin käyrä käy välillä HUS:n vaihteluvälin alarajan alapuolella, mutta pääasiallisesti se kulkee vaihteluvälin sisällä. Tulosten perusteella ei voida sanoa, että Satakunnan sairaanhoitopiirin kustannukset poikkeaisivat merkittävästi HUS:n kustannuksista.

Kun piirien kustannuksia vertaillaan käyrinä, havaitaan, että molemmissa kuvioissa käsittelyryhmän (Kuviossa 4.2 Varsinais-Suomen sairaanhoitopiiri ja Kuviossa 4.4 Satakunnan sairaanhoitopiiri) kustannuskäyrä leikkaa vertailuryhmän (HUS) kustannuskäyrän vaihteluvälin alarajan. Havaitaan myös, että kuviossa 4.2 kustannukset ovat alhasempia kuvion keskellä kuin kuvion reunoilla. Koska x-akselilla on logit-pistemäärä, ei käyristä pystytä suoraan päättelemään, minkälaisilla potilailla on esimerkiksi korkeat kustannukset. Käyrien avulla voidaan kuitenkin jatkossa tehdä lisätarkasteluja, joissa tutkitaan kustannusten ja potilaita sekä hoitoja luonnehtivien taustamuuttujien välisiä yhteyksiä. Tässä työssä rajoitutaan kuitenkin tutkimaan vain sairaanhoitopiirien kustannusten tasoeroja.

4.5 Vastaavuuspistemäärien estimointi multinomisella logit-mallilla

Kun piirejä on 20 ja niitä vertaillaan pareittain, on kaikkien mahdollisten vertailujen tekemiseksi estimoitava $\binom{20}{2} = 190$ logit-mallia. Olisikin havainnollisempaa, jos kaikkia piirejä voitaisiin verrata samanaikaisesti toisiinsa. Toden-

näköisyydet kuulua eri sairaanhoitopiireihin voidaan estimoida samanaikaisesti multinomisella logit-mallilla.

Kun piirejä tarkastellaan samanaikaisesti, niin käsittelymuuttujan K arvojoukko on $\{0, 1, \ldots, 19\}$. Olkoon $P(K = k|\mathbf{x}_i) = \pi_k(\mathbf{x}_i)$ potilaan i todennäköisyys kuulua piiriin $k \in \mathcal{K}$, kun potilaan kovariaattien arvovektori $\mathbf{x}_i$ tunnetaan. Multinomisessa logit-mallissa (Agresti 1990, 9. luku) verrataan jokaisen piirin todennäköisyyttä peruspiirin (tässä piiri 0) todennäköisyyteen. Malli

$$log\left(\frac{\pi_k(\mathbf{x}_i)}{\pi_0(\mathbf{x}_i)}\right) = \alpha + \boldsymbol{\beta}_k'\mathbf{x}_i \tag{4.3}$$

esittää samanaikaisesti kovariaattien $\mathbf{x}_i$, $1 \leq i \leq n$, vaikutuksen kaikkiin 19:ään logittiin ($k = 1, \ldots, 19$).

Multinominen logit-malli estimoidaan suurimman uskottavuuden menetelmällä sovittamalla malli (4.3) aineistoon samanaikaisesti kaikilla arvoilla $k = 1, \ldots, 19$. Mallit voitaisiin sovittaa myös erikseen, mutta samanaikaisesti sovitetuissa malleissa parametrien estimaattien keskivirheet ovat pienemmät kuin erikseen sovitetuissa malleissa (ks. esim. Agresti 1990, alaluku 9.3.1). Estimoinnin tuloksena mallista saadaan estimaatit $\hat{\pi}_k(\mathbf{x}_i)$, $0 \leq k \leq 19$. Jokaista havaintoa kohti saadaan estimoidut todennäköisyydet kuulua eri sairaanhoitopiireihin (20 kappaletta).

Seuraavassa luvussa esitetään, miten ns. yleistetyt vastaavuuspistemäärät voidaan estimoida multinomisen logit-mallin avulla. Multinomitodennäköisyyksiä $\pi_0(\mathbf{x}), \ldots, \pi_{19}(\mathbf{x})$ voidaan kuitenkin käyttää myös tavallisen vastaavuuspistemäärän tapaan. Kun vertaillaan esimerkiksi piirin k (käsittelyryhmä) kustannuksia muiden piirien kustannuksiin (kontrolli). Silloin vastaavuuspistemäärä $e(\mathbf{x}) = \pi_k(\mathbf{x})$ ja $1 - e(\mathbf{x}) = \sum_{j \neq k}^{19} \pi_j(\mathbf{x})$. Tässä asetelmassa vertaillaan siis piirin k kustannuksia kaikkien muiden kustannuksiin. Voidaan myös vertailla usean piirin ryhmiä. Vertaillaan esimerkiksi kolmen eteläisimmän piirin (sanokaamme piirit 0, 1 ja 2) kustannuksia muiden kustannuksiin. Silloin $e(\mathbf{x}) = \pi_0(\mathbf{x}) + \pi_1(\mathbf{x}) + \pi_2(\mathbf{x})$ ja $1 - e(\mathbf{x}) = \sum_{j=3}^{19} \pi_j(\mathbf{x})$.

Tarkastellaan nyt esimerkiksi kolmen (0, 1 ja 2) sairaanhoitopiirin muodostamaa ryhmää. Olkoot piirit HUS, Varsinais-Suomi ja Satakunta. Olisi kätevää vertailla samanaikaisesti näiden kolmen piirin kustannuksia. Kuvioon 4.5 on piirretty kustannuskäyrien

$$m_k(e) = E(Y|\pi_0(\mathbf{x}) = e, K = k) \tag{4.4}$$

estimaatit todennäköisyyden $\pi_0(\mathbf{x})$ estimaatin funktiona, missä K on käsittelymuuttuja ja $P(K = k|\mathbf{X} = \mathbf{x}) = \pi_k(\mathbf{x})$, $0 \leq k \leq 2$. Nyt $\pi_0(\mathbf{x})$ on valittu funktioksi, jonka suhteen ehdollistetaan. Havaitaan vaste

$$Y = D_0Y_0 + D_1Y_1 + D_2Y_2,$$

missä D_k on sellainen indikaattorimuuttuja, että

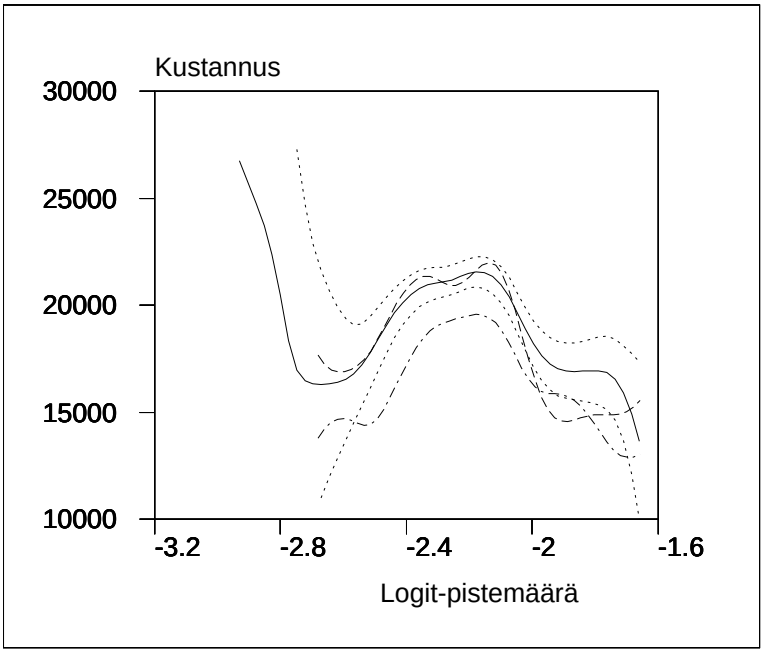
$$D_k = \begin{cases} 1, & \text{kun } K = k \text{ ja} \\ 0 & \text{muutoin.} \end{cases}$$

Kun esimerkiksi $K = 2$, havaitaan piiriin 2 kuuluvan potilaan kustannukset (vrt. alaluku 2.1.1).

Huomattakoon, että nyt $\pi_0(\mathbf{x})$ ei ole vastaavuuspistemäärä eivätkä identiteetit

$$m_k(e) = E(Y_k|\pi_0(\mathbf{x}) = e),\ k = 1, 2 \tag{4.5}$$

seuraa suoraan vastaavuuspistemäärälle esitetystä teoriasta. Siis Kuvan (4.5) käyrät eivät ole vertailukelpoisia keskenään samalla tavalla kuin vastaavuuspistemäärän tapauksessa. Useiden käsittelyiden tilannetta tarkastellaan lähemmin seuraavassa luvussa.



Kuva 4.5. HUS:n kustannukset (——-), HUS:n kustannusten kahden hajonnan levyinen vaihteluväli (.....), Varsinais-Suomen sairaanhoitopiirin kustannukset (.- .- .-) ja Satakunnan sairaanhoitopiirin kustannukset (– – –). Kustannuskäyrät on tasoitettu multinomisesta logit-mallista lasketun logit-muunnetun HUS:in todennäköisyyden yli.

5 Yleistetty vastaavuuspistemäärä

Usein sovelluksissa käsittelyjä on enemmän kuin kaksi, kuten esimerkiksi vertailtaessa lonkkamurtumapotilaiden hoitokustannuksia Suomen sairaanhoitopiirien välillä. Kun vertailussa käytetään vastaavuuspistemäärään perustuvaa menetelmää, sairaanhoitopiirejä tarkastellaan pareittain. Jokaista parittaista vertailua kohti on määritettävä ja estimoitava eri vastaavuuspistemäärät. Alaluvussa 2.5 tarkasteltiin käsittelyiden vertailua, missä käsittelymuuttuja K voi saada kokonaislukuarvot väliltä 0 ja J, joten K:n arvojoukko on $\mathcal{K} = \{0, 1, \ldots, J\}$. Imbens (2000) on esittänyt vastaavuuspistemäärän yleistyksen, jonka avulla on mahdollista estimoida kaikki keskimääräiset käsittelyvaikutukset, kun käsittelyjä on enemmän kuin kaksi. Tämän luvun teoria perustuu Imbensin artikkeliin. Yleistetty vastaavuuspistemäärä ei kuitenkaan jaa populaatiota osapopulaatioihin siten, että käsittelyvaikutuksia voitaisiin vertailla yleistetyn vastaavuuspistemäärän funktiona samaan tapaan kuin vastaavuuspistemäärän avulla.

5.1 Yleistetyn vastaavuuspistemäärän määritelmä

Yleistetty vastaavuuspistemäärä on ehdollinen todennäköisyys saada tietty käsittely $k \in \mathcal{K}$, kun kovariaattien $\mathbf{X}$ arvovektori on annettu:

$$r_k(\mathbf{x}) \equiv P(K = k|\mathbf{X} = \mathbf{x}).$$

Määrittelemme nyt jokaista käsittelyn K tasoa k kohti indikaattorimuuttujan

$$D_k = \begin{cases} 1, & \text{kun } K = k \text{ ja} \\ 0 & \text{muutoin.} \end{cases}$$

Indikaattorin avulla lausuttuna yleistetty vastaavuuspistemäärä on muotoa

$$r_k(\mathbf{x}) = E(D_k|\mathbf{X} = \mathbf{x}),\ k \in \mathcal{K}.$$

Jokaista tilastoyksikköä kohti havaitsemme satunnaismuuttujan $Y = D_0Y_0 + \cdots + D_kY_k + \cdots + D_JY_J$. Koska jokainen yksikkö saa täsmälleen yhden käsittelyn k, niin satunnaismuuttuja Y_k havaitaan vain silloin, kun $D_k = 1$. Näiden havaintojen perusteella voidaan siis estimoida vain ehdollinen odotusarvo $E(Y_k|D_k = 1, \mathbf{X})$, mutta nyt haluamme estimoida odotusarvon $E(Y_k|\mathbf{X})$. Palaamme siis kysymykseen, millä ehdolla identiteetti

$$E(Y_k|\mathbf{X}) = E(Y_k|D_k = 1, \mathbf{X}) \tag{5.1}$$

on voimassa. Vastaavaa kysymystä kahden käsittelyn tilanteessa käsiteltiin alaluvuissa 2.1.3 ja 2.2.2. Jos oletamme, että D_k ja Y_k ovat ehdollisesti riippumattomat ehdolla $\mathbf{X}$, eli

$$D_k \perp\!\!\!\perp Y_k|\mathbf{X} \tag{5.2}$$

kaikilla $k \in \mathcal{K}$, niin seuraa tulos (5.1)(vrt. liite C). Oletuksella (5.2) on tässä samanlainen keskeinen rooli kuin oletuksella (2.7) on vastaavuuspistemäärän yhteydessä (alaluku 2.2.1).

Jokaista käsittelyä $k \in \mathcal{K}$ kohti voidaan nyt muodostaa keskimääräiset vasteet laskemalla kaksinkertainen odotusarvo:

$$E(Y_k) = E_{\mathbf{X}}[E(Y_k|\mathbf{X})], \tag{5.3}$$

joka (5.1):n nojalla voidaan estimoida havainnoista. Keskimääräiset käsittelyvaikutukset $E(Y_k - Y_s)$, $k \neq s$, saadaan odotusarvojen (5.3) avulla. Odotusarvon laskeminen ehdollistamalla suureen määrään kovariaatteja on käytännössä hankalaa syistä, joita käsiteltiin alaluvussa 2.3. Siksi Rosenbaum & Rubin (1983) kehittivät vastaavuuspistemäärän käyttöön perustuvan ehdollistamismenetelmän. Tarkastelemme seuraavassa yleistetyn vastaavuuspistemäärän käyttöä ehdollisten odotusarvojen (5.3) laskemisessa.

5.2 Keskimääräisten vasteiden laskeminen

Riippumattomuusoletuksesta (5.2) seuraa, että keskimääräiset vasteet (5.3) voidaan laskea ehdollistamalla yleistettyyn vastaavuuspistemäärään $r_k(\mathbf{X})$. Jos siis oletus (5.2) pitää paikkansa, niin kaikilla $k \in K$ (Imbens 2000, Lause 1)

$$E(Y_k) = E_{\mathbf{X}}[E(Y_k|r_k(\mathbf{X}))]. \tag{5.4}$$

Keskimääräistä vastetta voidaan tarkastella yleistetyn vastaavuuspistemäärän funktiona

$$m_k(r) = E[Y_k|r_k(\mathbf{x}) = r] \tag{5.5}$$

kaikilla $k \in \mathcal{K}$. Nyt siis jokaista käsittelyä (ryhmää) kohti saadaan regressiofunktio $m_k(r)$. Odotusarvo (5.4) voidaan siis lausua regressiofunktion (5.5) avulla seuraavasti:

$$E(Y_k) = E_{\mathbf{X}}\{m_k[r_k(\mathbf{X})]\} = E_{R_k}[m_k(R_k)], \tag{5.6}$$

missä yleistetty vastaavuuspistemäärä $R_k = r_k(\mathbf{X})$ on satunnaismuuttuja. Odotusarvo voidaan nyt laskea R_k:n jakauman suhteen (vrt. Casella & Berger 2002, s.58).

Tarkastellaan nyt regressiofunktioiden m_k ja m_s erotusta pisteessä $r = r_k(\mathbf{X}) = r_s(\mathbf{X})$, missä $m_k(r)$ on esitetty yleistetyn vastaavuuspistemäärän $r_k = r_k(\mathbf{x})$ funktiona ja $m_s(r)$ vastaavasti r_s:n funktiona. Nyt siis identiteetin (5.5) perusteella saadaan

$$m_k(r) - m_s(r) = E(Y_k|r_k(\mathbf{x}) = r) - E(Y_s|r_s(\mathbf{x}) = r). \tag{5.7}$$

Odotusarvo $E(Y_k|r_k(\mathbf{x}) = r)$ lasketaan yli osapopulaation $S_k(\mathbf{x}) = \{\mathbf{x}|r_k(\mathbf{x}) = r\}$ ja odotusarvo $E(Y_s|r_s(\mathbf{X})) = r$ yli osapopulaation $S_s(\mathbf{x}) = \{\mathbf{x}|r_s(\mathbf{x}) = r\}$. Koska nämä ehdollistamisjoukot poikkeavat yleensä toisistaan, eli $S_r(\mathbf{x}) \neq S_s(\mathbf{x})$, niin erotusta (5.7) ei voida tulkita odotettuna käsittelyvaikutuksena. Käsittelyvasteiden ehdollista vertailua varten odotusarvot pitää ehdollistaa leikkausjoukkoon $S_k(\mathbf{x}) \cap S_s(\mathbf{x})$. Jos (5.2) on voimassa, niin silloin

$$\begin{aligned} &E(Y_k|K = k, r_k(\mathbf{x}), r_s(\mathbf{x})) - E(Y_s|K = s, r_k(\mathbf{x}), r_s(\mathbf{x})). \\ =&E(Y_k - Y_s|r_k(\mathbf{x}), r_s(\mathbf{x})) \end{aligned} \tag{5.8}$$

Kahden käsittelyn ehdollisessa vertailussa jouduttaisiin siis ehdollistamaan kahteen yleistettyyn vastaavuuspistemäärään. Tässä kadotetaan se vastaavuuspistemäärän e yksinkertainen ominaisuus, että odotettua käsittelyvaikutusta τ voidaan tarkastella vain e:n funktiona $\tau(e)$.

5.3 Kahden käsittelyn tilanne

Kahden käsittelyn tilanteessa $\mathcal{K} = \{0, 1\}$ ja $r_1(\mathbf{x}) = 1 - r_0(\mathbf{x})$. Rosenbaum ja Rubin (1983) osoittivat (ks. alaluku 3.2) vastaavuuspistemäärälle $e(\mathbf{x})$ tuloksen

$$\begin{aligned} &E(Y_1|K = 1, e(\mathbf{x}) = e) - E(Y_0|K = 0, e(\mathbf{x}) = e) \\ &= E(Y_1|e(\mathbf{x}) = e) - E(Y_0|e(\mathbf{x}) = e) \\ &= m_1(e) - m_0(e). \end{aligned}$$

Jos käytetään yleistetyn vastaavuuspistemäärän merkintöjä, niin $r_1(\mathbf{x}) = e(\mathbf{x})$ ja $r_0(\mathbf{x}) = 1 - e(\mathbf{x})$. Näillä merkinnöillä edellä esitetty tulos voidaan kirjoittaa muodossa

$$
\begin{aligned}
(5.9) \qquad & m_1(e) - m_0(e) \\
& = E(Y_1|r_1(\mathbf{x}) = e) - E(Y_0|r_0(\mathbf{x}) = 1 - e) \\
& = E(Y_1|r_1(\mathbf{x}) = e) - E(Y_0|r_1(\mathbf{x}) = e) \\
& = E(Y_1 - Y_0|r_1(\mathbf{x}) = e).
\end{aligned}
$$

Odotusarvon määrittämiseksi riittää siis ehdollistaminen pelkästään vastaavuuspistemäärään $r_1(\mathbf{x})$, koska ehdollistamisjoukot $\{\mathbf{x}|r_1(\mathbf{x}) = e\}$ ja $\{\mathbf{x}|r_0(\mathbf{x}) = 1 - e\}$ ovat identtiset. Näin kahden käsittelyn tapauksessa

$$
\begin{aligned}
& E(Y_1 - Y_0|r_0(\mathbf{x}), r_1(\mathbf{x})) \\
& = E(Y_1 - Y_0|1 - r_1(\mathbf{x}), r_1(\mathbf{x})) \\
& = E(Y_1 - Y_0|r_1(\mathbf{x})).
\end{aligned}
$$

Vaikka erotusta (5.7) ei voida tulkita odotetuksi käsittelyvaikutukseksi annetussa osapopulaatiossa, voidaan odotusarvoja (5.6) yli koko yleistetyn vastaavuuspistemäärän jakauman käyttää odotettujen vasteiden muodostamiseen.

5.4 Yleistetyn vastaavuuspistemäärän soveltaminen käytäntöön

Alaluvussa 4.5 estimoitiin jokaiselle potilaalle todennäköisyys $P(K = k|\mathbf{x})$ kuulua sairaanhoitopiiriin k, $0 \leq k \leq 19$ multinomisella logit-mallilla. Näitä todennäköisyyksiä voidaan käyttää yleistettyjen vastaavuuspistemäärien estimaatteina. Yleistettyyn vastaavuuspistemäärään ehdollistettujen kustannuskäyrien erotusta ei voida tulkita samalla tavalla kuin tavallisen vastaavuuspistemäärän tapauksessa. Alaluvussa 5.2 näytettiin, että yleistetyn vastaavuuspistemäärän käyttäminen kahden sairaanhoitopiirin käsittelyvaikutuksen vertailuun edellyttää sitä, että ehdollistetaan kahteen yleistettyyn vastaavuuspistemäärään.

Identiteetin (5.6) perusteella piirien k ja s välinen odotettu käsittelyvaikutus voidaan lausua regressiofunktioiden m_k ja m_s avulla

$$
\tau_{ks} = E(Y_k) - E(Y_s) = E_{R_k}[m_k(R_k)] - E_{R_s}[m_s(R_s)],
$$

missä $k \neq s$. Merkitään yleistettyjen vastaavuuspistemäärien $R_k = r_k(\mathbf{X})$, $0 \leq k \leq J$ ja tiheysfunktioita $f_k(r_k)$. Yleistettyjen vastaavuuspistemäärien avulla lausuttuna odotetut vasteet ovat muotoa

$$
(5.10) \qquad \mu_k = E[m_k(R_k)] = \int m_k(r_k) f(r_k) dr_k, \ 0 \leq k \leq J,
$$

missä integraali otetaan yli välin $(0, 1)$. Näin voidaan vertailla piirien odotettuja keskimääräisiä kustannuksia (vasteita) toisiinsa, vaikka kustannusfunktioiden

(regressiofunktioiden) erotusta $m_k(r) - m_s(r)$ ei voida tulkita ehdolliseksi käsittelyvaikutukseksi (vrt. alaluku 5.2).

Käytännössä regressiofunktiot (kustannusfunktiot) m_k estimoidaan aineistosta alaluvussa 3.3 esitetyllä tavalla. Funktiot m_k estimoidaan aina kyseisen piirin k, $0 \leq k \leq J$ potilaista. Myös $f(r_k)$ ja $f(r_s)$ ovat tuntemattomia funktioita, mutta ne voidaan estimoida R_k:n ja R_s:n histogrammeilla tai ydinestimaattorilla (vrt. alaluku 4.4). Näin odotusarvot $E[m_k(r_k)]$ ja $E[m_s(r_s)]$ voidaan estimoida esimerkiksi alaluvussa 2.3 esitetyllä luokittelumenetelmällä.

Kun funktioiden $m_k(\cdot)$, $0 \leq k \leq J$ estimaatit $\hat{m}_k(\cdot)$ ovat käytettävissä, saadaan odotusarvojen (5.10) estimaatit alaluvussa 3.5 esitetyllä menetelmällä:

$$\hat{\mu}_k = \sum_{i=1}^{n} \hat{m}_k[\hat{r}_k(\mathbf{x}_i)],\ 0 \leq k \leq J. \tag{5.11}$$

Huomattakoon, että estimaatissa (5.11) summa voidaan laskea yli kaikkien piirien potilaiden, jolloin $n = 16881$. Silloin estimaatti $\hat{\mu}_k$ on keskimääräinen kustannus, jos kaikki potilaat olisi hoidettu piirissä k. Myös regressiofunktioiden estimaattien $\hat{m}_k(r_k)$ keskimääräiset hajonnat voidaan laskea (vrt. alaluku 3.5).

Keskimääräisen käsittelyvaikutuksen estimaattoria ei nyt kuitenkaan voi muodostaa kaavan (3.12) tapaan, koska kaikki odotusarvot (5.5) lasketaan eri vastaavuuspistemäärän (ja eri jakauman) suhteen. Kun sairaanhoitopiirejä vertaillaan yleistettyjen vastaavuuspistemäärien avulla, on samanaikaisesti kustannuskäyrien keskimääräisten kustannusten estimaattien kanssa otettava huomioon yleistettyjen vastaavuuspistemäärien R_k erilaiset jakaumat.

Taulukko 5.1. HUS:in, Varsinais-Suomen sairaanhoitopiirin ja Satakunnan sairaanhoitopiirin potilaiden keskimääräiset kustannukset ja keskimääräiset hajonnat.

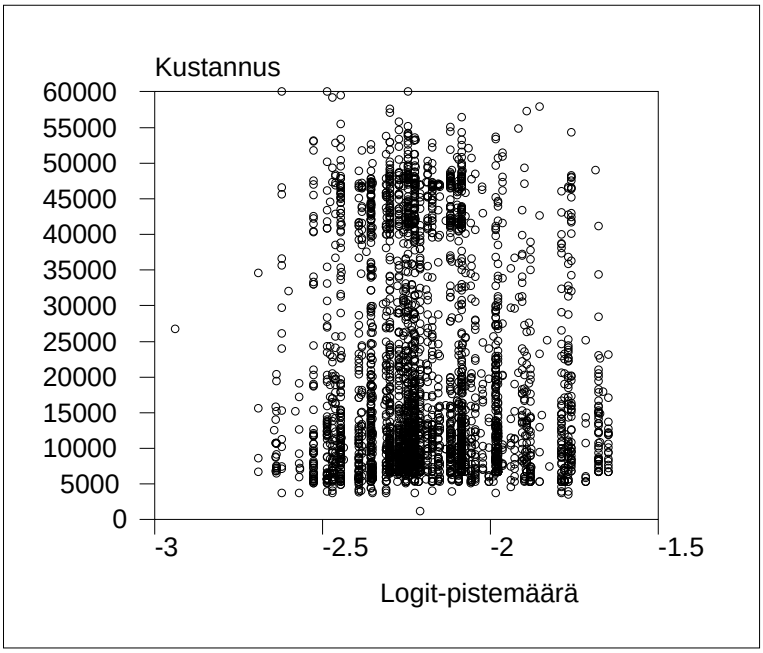
Sairaanhoitopiiri	Keskimääräinen kustannus	Keskimääräinen hajonta
HUS	19980.262	506.786
V-S shp	17726.545	659.066
Stk shp	20022.932	946.321

Taulukossa 5.1 ovat Satakunnan sairaanhoitopiirin potilaiden keskimääräiset kustannukset kaikkein korkeimmat. HUS:n potilaiden kustannukset ovat käytännössä samalla tasolla Satakunnan sairaanhoitopiirin potilaiden kustannusten kanssa, koska eroa on vain noin 40 euroa. Varsinais-Suomen sairaanhoitopiirin potilaiden kustannukset ovat kuitenkin selvästi kahden muun piirin potilaiden kustannuksia alhaisemmat. Osoitimme jo 4. luvussa vastaavuuspistemäärän avulla, että Varsinais-Suomen sairaanhoitopiirin potilaiden kustannukset ovat HUS:n kustannuksia alhaisemmat.

6 Potilasryhmät sairaanhoitopiirien sisällä

Aiemmissa luvuissa on tarkasteltu sairaanhoitopiirin kaikista potilaista laskettuja keskimääräisiä kustannuksia. Alaluvun 1.1.4 Kuviossa 1.3 esitettiin kustannusten histogrammi, josta havaitaan kustannusten jakauman olevan kaksihuippuinen. Koska histogrammista erottuu näin selvästi kaksi eri huippua, voidaan ajatella ilmiön taustalla ajatella olevan kaksi kustannuksiltaan toisistaan poikkeavaa potilasryhmää.

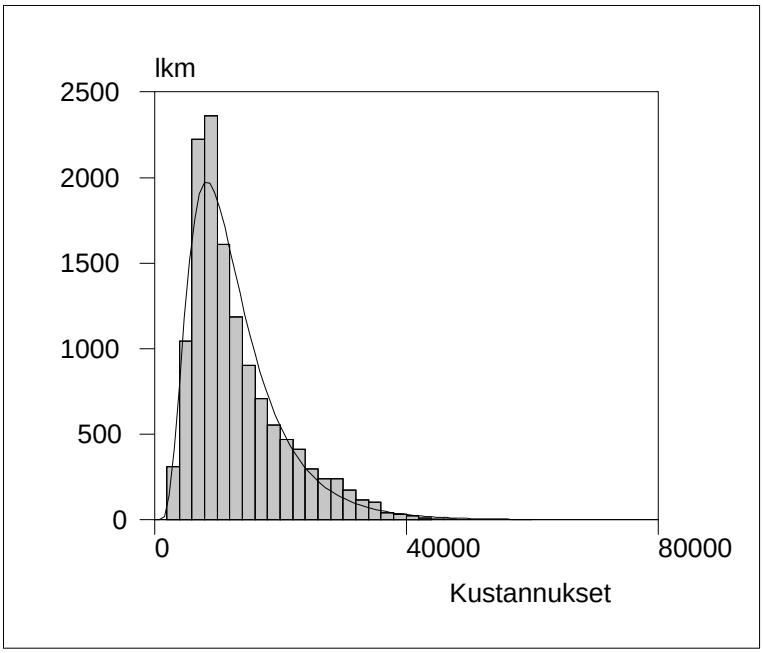
Kuvioon 6.1 on piirretty HUS:n potilaiden kustannusten riippuvuutta logit-pistemääristä kuvaava pisteparvi. Myös Kuvion 6.1 pisteparvesta erottuu kaksi kustannuksiltaan erilaista ryhmää. Nämä ryhmät erottuvat myös muiden piirien vastaavista pisteparvista.



Kuva 6.1. HUS:n potilaiden logit-pistemäärien ja kustannusten välistä riippuvuutta kuvaava pisteparvi.

Jos poistetaan aineistosta potilaat, joilla on ollut murtumaa seuraavan vuoden aikana yli 250 hoitopäivää ja piirretään jäljelle jääneiden potilaiden kustannusjakauma, saadaan Kuvion 6.2 histogrammi. Kalleimmat potilaat pystytään siis erottelemaan melko hyvin hoitopäivien perusteella. Tarvitaan kuitenkin

jatkotutkimuksia, jotta voidaan erotella ja luonnehtia riittävän hyvin kaikkein kalleinta hoitoa saava ryhmä.



Kuva 6.2. Korkeintaan 250 päivää hoidettujen potilaiden kustannusten histogrammi ja siihen sovitetun lognormaalijakauman tiheysfunktio koko maan aineistosta.

Koska aineistosta erottuu näin selkeästi kaksi kustannuksiltaan erilaista potilasryhmää, on kiinnostavaa estimoida molemmille ryhmille omat keskimääräiset kustannuksensa. Näitä ryhmiä voidaan sitten vertailla myös eri piirien välillä ja selvittää miten piirien kustannuserot ovat rakentuneet.

6.1 Ryhmät piirien sisällä ja niiden vertailu

Oletetaan, että aineistosta erottuvien kahden ryhmän potilasrakennetta ei tunneta. Tiedetään vain, että ryhmät poikkeavat kustannuksiltaan toisistaan. Näitä ryhmiä kutsutaan jatkossa kalliin hoidon ryhmäksi (kh-ryhmä) ja normaalihintaisen hoidon ryhmäksi (nh-ryhmä). Kustannusten vaihtelu ryhmien sisällä on paljon pienempi kuin koko aineistossa. Suurin osa kustannusvaihtelusta selittyykin ryhmien välisellä vaihtelulla. Näille sisäisille ryhmille estimoidaan omat kustannuskäyränsä piirin sisäisen kustannusrakenteen havainnollistamiseksi. Luonnollisesti meitä kiinnostaa vertailla myös kahden eri piirin kh-ryhmien kustannuksia keskenään (vastaavasti nh-ryhmien). Tämän luvun tarkastelut tarjoavat myös lähtökohdan piirien sisäisten kustannusryhmien potilasrakenteen selvittämiseksi.

Tämän luvun tarkastelut voidaan jakaa kolmeen pääkohtaan

1. Estimoidaan kh-ryhmän ja nh-ryhmän kustannuskäyrät vastaavuuspistemäärän funktiona sairaanhoitopiirin sisällä (alaluku 6.2).

2. Luokitellaan potilaat koko aineiston kustannusjakauman perusteella kalliin hoidon ryhmään (kh-ryhmä) ja normaalihintaiseen (nh-ryhmä) ryhmään (alaluku 6.3.1).

3. Vertaillaan kahden piirin kh-ryhmiä (vastaavasti nh-ryhmiä) keskenään. Tätä varten vertailtavien piirien kh-ryhmät yhdistetään ja yhdistetystä aineistosta estimoidaan vastaavuuspistemäärä. Sen jälkeen ryhmiä vertaillaan kuten sairaanhoitopiirejä 4. luvussa (alaluku 6.3.2).

Sairaanhoitopiirien kustannusryhmien sisäinen vertailu on sinällään mielenkiintoinen ja tärkeä tehtävä. Varsin suuri osa piirien kokonaiskustannuksista muodostuu kh-ryhmän kustannuksista, vaikka se on potilasmäärältään suhteellisen pieni verrattuna nh-ryhmään. Jatkotutkimuksissa voidaan kovariaattien avulla identifioida kustannusryhmien potilasrakenne.

Kun vertaillaan esimerkiksi HUS:in ja Varsinais-Suomen sairaanhoitopiirin kustannuksia, lasketaan vastaavuuspistemäärä yhdistetystä HUS:in ja Varsinais-Suomen sairaanhoitopiirin aineistosta (vrt. luku 4). Näin laskettuja vastaavuuspistemääriä käyttäen voidaan vertailla myös HUS:in (ja vastaavasti Varsinais-Suomen sairaanhoitopiirin) kh-ryhmän ja HUS:in nh-ryhmän kustannuksia keskenään alaluvussa 6.2 esitettävällä painotusmenetelmällä. Jos halutaan tutkia kh-ryhmän ja nh-ryhmän potilasrakennetta, on potilaat luokiteltava eksplisiittisesti näihin ryhmiin.

Luokittelu voidaan tehdä käyttäen ryhmien kustannusjakaumia, jotka on estimoitava myös alaluvussa 6.2 esitettävää painotusmenetelmää varten. Tämän jälkeen mitä tahansa kahta luokittelulla muodostettua ryhmää voidaan vertailla keskenään 4. luvussa esitetyllä tekniikalla. Vertailua varten ryhmät yhdistetään omaksi osa-aineistokseen, toinen ryhmistä valitaan käsittelyryhmäksi ja toinen vertailuryhmäksi ja sitten osa-aineistosta estimoidaan vastaavuuspistemäärä. Ryhmien vertailussa voidaan sen jälkeen käyttää kaikkia vastaavuuspistemäärään perustuvia menetelmiä. Huomattakoon, että tätä luokitteluun perustuvaa menetelmää voidaan aina käyttää alaluvun 6.2 painotusmenetelmän sijasta.

6.2 Piirien sisäisten ryhmien kustannuskäyrät

6.2.1 Estimointi painotusmenetelmällä

Piirien sisäisten ryhmien kustannusten regressiokäyrät estimoidaan alaluvussa 3.2 esitetyllä ydinestimaattorilla. Käyrien estimoimiseksi ryhmiä ei tarvitse muodostaa eksplisiittisesti, vaan estimaatit saadaan painottamalla havaintoihin liittyvää ydinfunktiota kyseisen ryhmän kustannusjakauman tiheysfunktion estimaatilla. Painotusta varten lasketaan ensin kh-ryhmän ja nh-ryhmän kustannusjakaumien tiheysfunktioiden estimaatit $f(y;\hat{\boldsymbol{\theta}}_1)$ ja $f(y;\hat{\boldsymbol{\theta}}_2)$, missä $\hat{\boldsymbol{\theta}}_1$ on kh-ryhmän kustannusjakauman parametrin $\boldsymbol{\theta}_1$ estimaatti ja $\hat{\boldsymbol{\theta}}_2$ on nh-ryhmän vastaavan parametrin $\boldsymbol{\theta}_2$ estimaatti (ks. alaluku 6.2.2).

Ryhmien kustannuskäyrät estimoidaan ydinestimaattorilla (3.8), mutta nyt painotetussa pienimmän neliösumman lausekkeessa (3.6) käytetään ydinpainojen $K(\frac{e_i-e}{h})$ sijasta tiheysfunktion arvoilla painotettuja ydinpainoja. Kh-ryhmän estimoinnissa painot ovat

$$K_1(\frac{e_i-e}{h}) = K(\frac{e_i-e}{h})f(y_i;\hat{\boldsymbol{\theta}}_1) \tag{6.1}$$

ja nh-ryhmän estimoinnissa

$$K_2(\frac{e_i-e}{h}) = K(\frac{e_i-e}{h})f(y_i;\hat{\boldsymbol{\theta}}_2), \tag{6.2}$$

missä $1 \leq i \leq n$. Ryhmien kustannuskäyrät on estimoitu R-ohjelmiston sm-kirjaston regressiofunktiolla (Bowman & Azzalini 1996, sm.regression).

6.2.2 Kustannusjakaumien tiheysfunktioiden estimointi

Ryhmien kustannusjakaumien tiheysfunktiot $f(y;\boldsymbol{\theta}_1)$ ja $f(y;\boldsymbol{\theta}_2)$ estimoidaan sovittamalla kahden lognormaalijakauman sekoitus koko aineiston kustannuksiin. Merkitään kalliin hoidon (kh) ryhmään kuuluvien potilaiden kustannuksia Y_1 ja normaalihintaisen hoidon ryhmään (nh) kuuluvien kustannuksia Y_2. Koska oletamme kustannusten kummassakin ryhmässä noudattavan lognormaalijakaumaa, niin

$$log(Y_1) \sim N(\mu_1, \sigma_1^2)$$

ja

$$log(Y_2) \sim N(\mu_2, \sigma_2^2),$$

missä log on luonnollinen logaritmi ja $N(\cdot,\cdot)$ normaalijakauma. Lognormaalijakaumaa noudattavan satunnaismuuttujan Y_1 tiheysfunktio on

$$f(y_1;\mu_1,\sigma_1^2) = \frac{1}{\sigma_1\sqrt{2\pi}y_1}e^{-(log(y_1)-\mu_1)^2/2\sigma_1^2}$$

(Casella & Berger 2002, s.109).

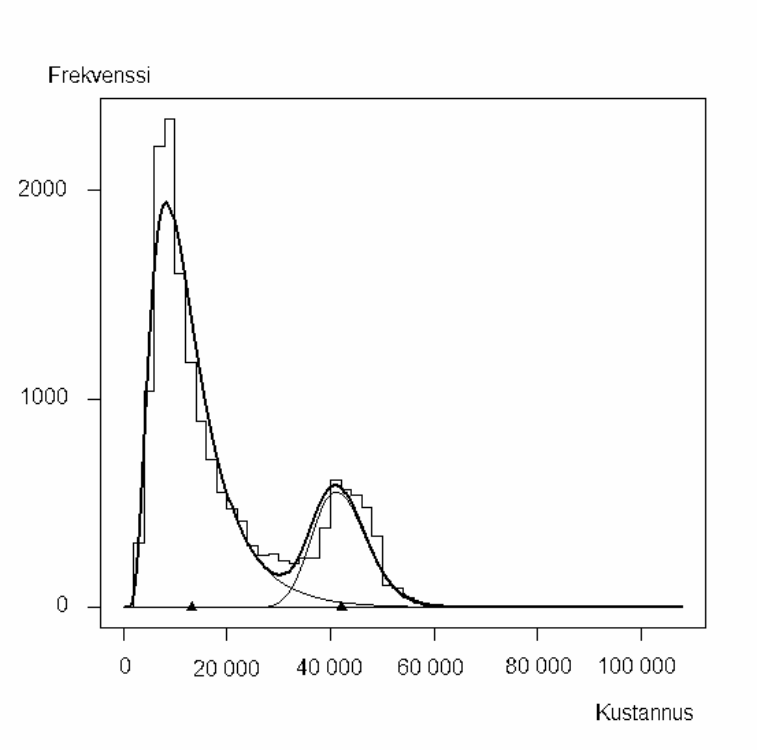
Koska kustannukset Y mallinnetaan ryhmien lognormaalijakaumien sekoituksena, niin

$$Y = RY_1 + (1-R)Y_2, \tag{6.3}$$

missä R on sellainen indikaattori (satunnaismuuttuja), että $P(R=1)=\pi$ ja $P(R=0)=1-\pi$. Nyt siis todennäköisyydellä π saadaan $(R=1)$ havainto kh-ryhmästä (Y_1). Vastaavasti todennäköisyydellä $1-\pi$ havaitaan Y_2. Silloin Y:n tiheysfunktio on

$$g(y;\pi,\boldsymbol{\theta}_1,\boldsymbol{\theta}_2) = \pi f(y;\boldsymbol{\theta}_1) + (1-\pi)f(y;\boldsymbol{\theta}_2), \tag{6.4}$$

mistä parametrit $\pi, \boldsymbol{\theta}_1 = (\mu_1, \sigma_1^2)$ ja $\boldsymbol{\theta}_2 = (\mu_2, \sigma_2^2)$ voidaan estimoida suurimman uskottavuuden menetelmällä. Kuvioon 6.3 on piirretty koko maan lonkkamurtumapotilaiden kustannusten histogrammi ja kahden lognormaalijakauman sekoituksen estimoitu tiheysfunktio.



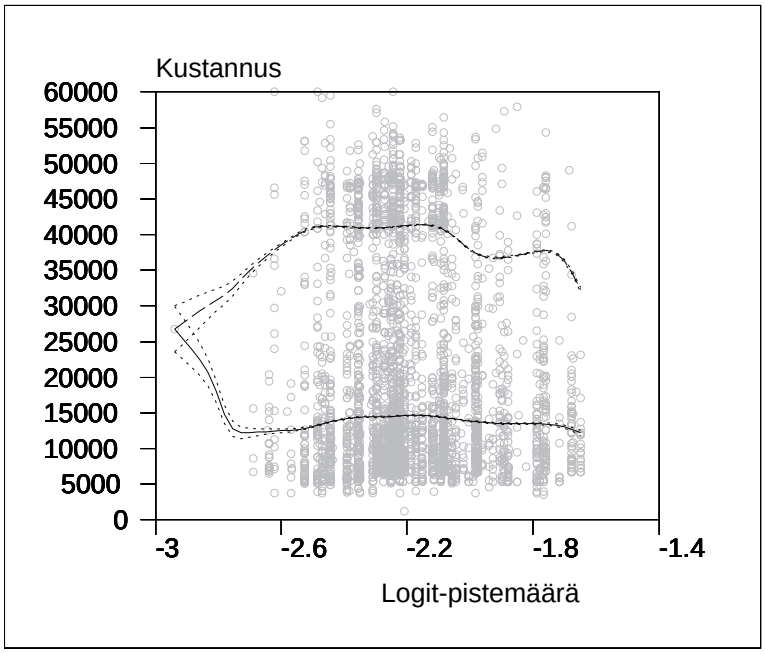
Kuva 6.3. Koko maan lonkkamurtumapotilaiden kustannusten histogrammi, sekä kustannuksiin sovitettu kahden lognormaalijakauman sekoituksen tiheysfunktio.

Sekoitetun jakauman uskottavuusfunktion suoraviivainen maksimointi on numeerisesti hankalaa, mutta onnistuu esimerkiksi EM-algoritmilla (Hastie, Tibshirani & Friedman 2001, s.238). Estimoin parametrit käyttäen R-ohjelmiston Rmix pakettia (ks. esim. Du 2002, Rmix ei ole saatavilla R-projektin virallisilta kotisivuilta, koska se ei ole ainakaan vielä virallinen R-paketti. Rmix on saatavilla professori Peter Macdonaldin kotisivulta osoitteesta http://icarus.math.mcmaster.ca/peter/). Näin on saatu tiheysfunktioiden estimaatit $f(y; \hat{\boldsymbol{\theta}}_1)$ ja $f(y; \hat{\boldsymbol{\theta}}_2)$, joiden avulla määritellään painotetut ydinpainot (6.1) ja (6.2). Painotettujen havaintojen perusteella estimoidaan sitten regressiofunktiot, kuten alaluvussa 6.2.1 on esitetty.

6.2.3 Kahden ryhmän kustannusten vertailu sairaanhoitopiirin sisällä

Alaluvussa 4.4 vertailtiin HUS:in ja Varsinais-Suomen sairaanhoitopiirin kustannuksia keskenään (Kuvio 4.2). Samoja vastaavuuspistemääriä käyttäen voi-

daan piirtää myös HUS:in ja Varsinais-Suomen sairaanhoitopiirin kh-ryhmän ja nh-ryhmän tasoitetut kustannuskäyrät vastaavuuspistemäärän funktiona. Kuviossa 6.4 on HUS:in potilaille piirretty kustannusten ja logit-pistemäärän riippuvuutta kuvaavaan pisteparveen sekä kh-ryhmän, että nh-ryhmän potilaiden estimoidut kustannuskäyrät ja niille kahden hajonnan vaihteluvälit. Havaitaan, että ryhmien kustannuskäyrien vaihteluvälit kapenevat rajusti sairaanhoitopiirin kaikkien potilaiden kustannuskäyrän vaihteluväliin verrattuna (vrt. Kuva 4.2).



Kuva 6.4. Kalliiden (– – –) ja normaalihintaisten potilaiden (——-) kustannukset Helsingin ja Uudenmaan sairaanhoitopiirissä logit-pistemäärän funktiona. Taustalla on himmeänä potilaiden kustannusten pisteparvi. Molempien ryhmien käyrille on piirretty kahden hajonnan levyinen vaihteluväli (.....).

6.3 Kahden eri piirin potilasryhmien vertailu

6.3.1 Potilaiden luokittelu

Jos halutaan vertailla esimerkiksi kahden eri piirin kh-ryhmien kustannuksia vastaavuuspistemäärään perustuvilla menetelmillä, niin silloin piirin potilaat on ensin luokiteltava kustannusryhmiin, kuten alaluvussa 6.1 todettiin. Tämä siksi, että vastaavuuspistemäärät on estimoitava vertailtavista aineistoista. Jos taas kh-ryhmää ja nh-ryhmää vertaillaan vain piirin sisällä, ei luokittelua tarvita, vaan voidaan käyttää alaluvun 6.2.1 painotusmenetelmää.

Alaluvussa 6.2.2 esitetyn mallin (6.3) mukaan havainto Y on peräisin joko kh-ryhmästä (Y_1) tai nh-ryhmästä (Y_2). Kun saadaan havainto $Y = y$, on

päätettävä, kummasta ryhmästä se on peräisin. Tämä on klassinen luokitteluongelma (ks. esim. Anderson 1984, luku 6.). Nyt π on havainnon ennakkotodennäköisyys (prior probability) kuulua kh-ryhmään ja vastaavasti $(1-\pi)$ on nh-ryhmään kuulumisen ennakkotodennäköisyys. Kustannusten jakaumien tiheysfunktiot ryhmissä ovat (6.4):n mukaan $f(y;\boldsymbol{\theta}_1)$ ja $f(y;\boldsymbol{\theta}_2)$. Optimaalinen luokittelusääntö, joka minimoi virheluokittelun todennäköisyyden, on seuraava:
Sijoita havainto kh-ryhmään, jos

$$\pi f(y;\boldsymbol{\theta}_1) > (1-\pi) f(y;\boldsymbol{\theta}_2), \tag{6.5}$$

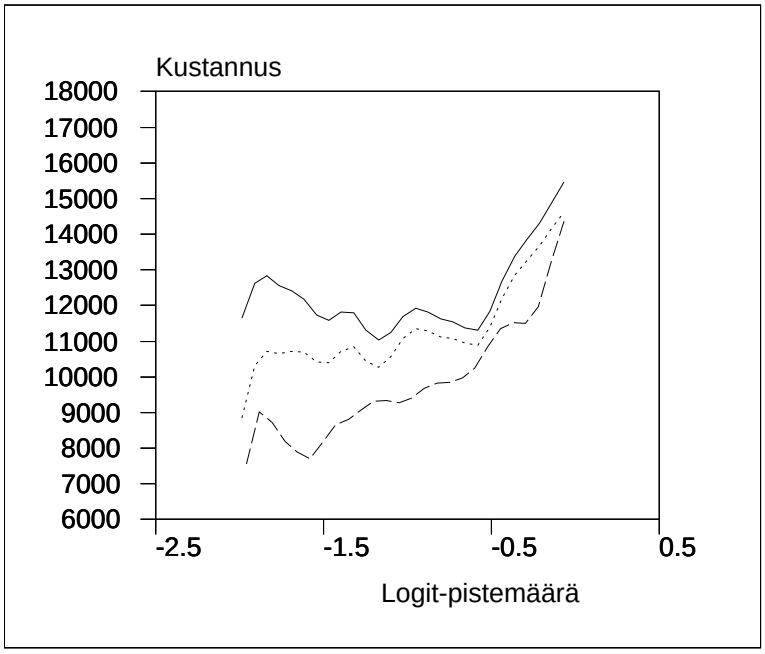
ja muutoin sijoita havainto nh-ryhmään (ks. Anderson 1984, s.200). Jotta sääntöä (6.5) voidaan soveltaa käytännössä, korvataan tuntemattomat parametrit π, $\boldsymbol{\theta}_1$ ja $\boldsymbol{\theta}_2$ suurimman uskottavuuden estimaateillaan $\hat{\pi}, \hat{\boldsymbol{\theta}}_1$ ja $\hat{\boldsymbol{\theta}}_2$ (ks. alaluku 6.2.2).

6.3.2 Eri piireihin kuuluvien ryhmien kustannuskäyrät

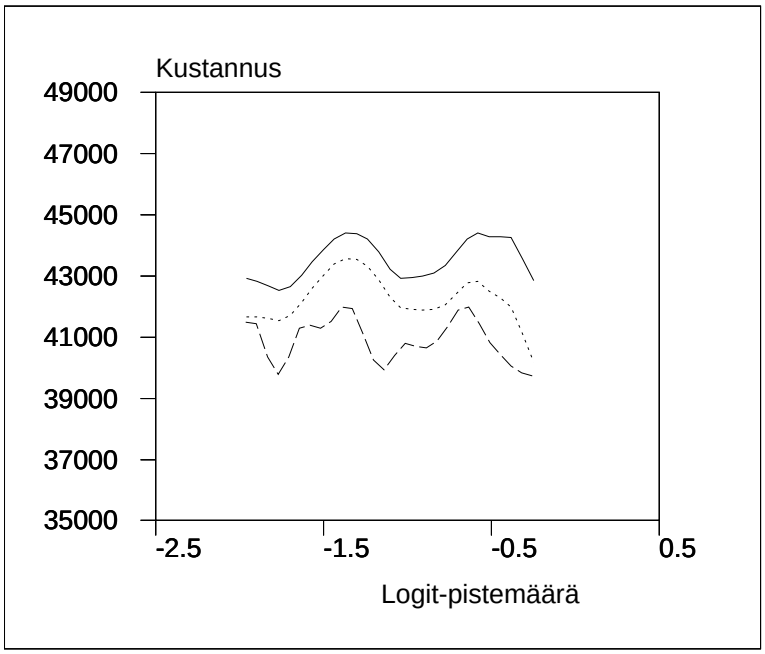
Kun määritetään ensin estimoidut funktiot $\hat{\pi} f(y;\hat{\boldsymbol{\theta}}_1)$ ja $(1-\hat{\pi}) f(y;\hat{\boldsymbol{\theta}}_2)$, voidaan piirin aineisto jakaa nh- ja kh-ryhmään luokittelusäännöllä (6.5). HUS:n potilas sijoitetaan kh-ryhmään, jos $y > 33800$; muutoin potilas sijoitetaan nh-ryhmään. Esimerkiksi HUS:in ja Varsinais-Suomen sairaanhoitopiirien kh-ryhmien vertailemiseksi muodostetaan HUS:in kh-ryhmästä ja Varsinais-Suomen kh-ryhmästä uusi aineisto, josta estimoidaan vastaavuuspistemäärä. Sitten aineistosta estimoidaan tämän vastaavuuspistemäärän avulla piirien kh-ryhmien kustannuskäyrät vastaavuuspistemäärän funktiona. Samoin menetellään vertailtaessa nh-ryhmien kustannuksia kh-ryhmien kustannusten sijasta.

Kahden eri piirin (esim. HUS ja Varsinais-Suomi) sisäisten ryhmien (esim. kh-ryhmät) vertailemiseksi muodostetaan osa-aineisto edellä kuvatulla tavalla. Sen jälkeen näitä sisäisiä ryhmiä voidaan vertailla tässä osa-aineistossa samalla tavalla kuin alaluvussa 4.4. Kuviossa 6.5 on esitetty HUS:in ja Varsinais-Suomen nh-ryhmien kustannuskäyrät. Nähdään, että Varsinais-Suomen piirin nh-ryhmän kustannukset ovat kaikilla vastaavuuspistemäärillä HUS:n nh-ryhmän kustannuksia alhaisemmat.

Kuvassa 6.6 ovat HUS:in ja Satakunnan sairaanhoitopiirin kh-ryhmien kustannuskäyrät, joista havaitaan, että Satakunnan sairaanhoitopiirin kh-ryhmän kustannusten olevan HUS:n kh-ryhmän kustannuksia alhaisemmat.



Kuva 6.5. HUS:in (——-) ja Varsinais-Suomen sairaanhoitopiirin (- - - -) nh-ryhmien kustannuskäyrät. HUS:in käyrälle on piirretty kahden hajonnan levyisen vaihteluvälin alaraja(.....).



Kuva 6.6. HUS:in (——-) ja Satakunnan sairaanhoitopiirin (- - - -) kh-ryhmien kustannuskäyrät. HUS:in käyrälle on piirretty kahden hajonnan levyisen vaihteluvälin alaraja (.....).

7 Lopuksi

Tutkielmassa esitetään vastaavuuspistemäärän käyttöön perustuva menetelmä, jolla voidaan estimoida sairaanhoitopiirien kustannukset ja piirien väliset kustannuserot. Menetelmää sovelletaan laajoihin rekisteripohjaisiin aineistoihin ja saadaan uusia ja käyttökelpoisia empiirisiä tuloksia. Tietääkseni rekisteriaineistoihin perustuvia kustannusvertailuja ei ole aikaisemmin tehty näillä menetelmillä. Kustannusten tarkastelu vastaavuuspistemäärän funktiona ei ole aivan tavanomaista, vaikka esimerkiksi vastaavuuspistemäärään perustuva luokitteluestimaattori saattaa olla joillain sovellusalueilla suosittu menetelmä.

On perusteltua olettaa, että lonkkamurtumien hoidon tavoitteet ovat kaikissa sairaanhoitopiireissä samat. Hoitokäytäntöjä voidaan vertailla esimerkiksi hoitotulosten ja hoitoon käytettyjen resurssien perusteella. Tässä työssä on keskitytty hoitokustannusten vertailuun. Jos jokin sairaanhoitopiiri saavuttaa yhtä hyvät hoitotulokset kuin vertailupiiri, mutta alhaisemmilla kustannuksilla, on syytä tutkia tarkemmin käytettyjen hoitokäytäntöjen eroja. Silloin hoidon tuloksia on luonnehdittava muilla indikaattoreilla kuin kustannuksilla.

Työssä esitetyn menetelmän perusidea on korvata suuri joukko kustannuksiin vaikuttavia taustamuuttujia näiden muuttujien funktiolla, jota kutsutaan vastaavuuspistemääräksi. Pistemäärää käytetään sitten ikään kuin olisi vain yksi selittävä muuttuja. Ehdollistamalla vastaavuuspistemäärään voidaan eliminoida taustamuuttujien sekoittava vaikutus kustannuseroihin. Kustannuksille voitaisiin tietysti rakentaa erilaisia usean selittäjän regressiomalleja, kuten esimerkiksi monitasomalleja. Tällaisten mallien spesifionnissa ja tulkinnassa on omat ongelmansa. Silloin tavallisesti tehdään oletuksia esimerkiksi kustannusten ja selittäjien välisen riippuvuuden funktionaalisesta muodosta. Tällaisia oletuksia ei vastaavuuspistemäärään perustuvassa lähestymistavassa tarvita. Toisaalta käytännössä vastaavuuspistemäärän estimointi perustuu usean selittäjän logit-malliin.

Olen vertaillut tällä menetelmällä kolmen sairaanhoitopiirin kustannuksia. Vastaavat vertailut ovat suoraviivaisesti tehtävissä Suomen kaikille sairaanhoitopiireille. Tulosten perusteella havaitaan, että erityisesti HUS:in ja Varsinais-Suomen sairaanhoitopiirin kustannukset poikkeavat merkittävästi. Piirien kustannuskäyrien tarkempi analysointi tulee olemaan kiintoisa jatkotarkastelujen aihe. Jatkotutkimusten avulla voitaneen saada hyödyllistä lisätietoa myös eri sairaanhoitopiirien kustannusrakenteesta ja mahdollisesti löytää potilasryhmiä, joihin on kiinnitettävä erityistä huomiota.

Työn kestäessä nousi esiin runsaasti sekä menetelmää että sovellusta koske-

via jatkokysymyksiä, joista sovelluksen kannalta kiinnostavimmat liittyvät sairaanhoitopiirien potilasrakenteen selvittämiseen. Miten kustannuksiltaan erilaisten potilasryhmien potilaat poikkeavat toisistaan? Näkyvätkö kustannuserot myös hoidon tuloksissa? Menetelmällisesti kiinnostavia kysymyksiä ovat muun muassa kustannuskäyrien harhan tutkiminen ja kustannusten estimointi myös vaihtoehtoisilla menetelmillä sekä piirien sisäisten kustannusryhmien erottelu. Yleistetty vastaavuuspistemäärä on potentiaalisesti hyvin käyttökelpoinen lähestymistapa, jonka sovellusmahdollisuuksien kehittely on mielenkiintoinen tutkimushaaste. Sitä koskevaa teoreettista tutkimustakin on olemassa vasta niukasti.

Lopuksi haluan kiittää ohjaajaani Reijo Sundia, joka johdatti minut tutkimaan tätä aihetta, sekä antoi valmiin tutkimusaineiston käyttööni. Hän tarjosi myös tuoreita ideoita useissa työn vaiheissa ja antoi apua työn aikana ilmenneiden ongelmien ratkaisussa.

Kirjallisuutta

Agresti, A. (1990), *Categorical Data Analysis*, John Wiley & Sons, Inc.

Anderson, T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc.

Bowman, A.W. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations.* Oxford University Press, Oxford.

Casella, G. & Berger, R.L. (2001), *S*tatistical Inference, 2nd edition, Duxbury Press.

Charlson, M.E., Pompei, P., Ales, K.L. ym. (1987), "A New Method for Classifying Prognostic Comorbidity in Longitudinal Studies: Development and Validation" *Journal of Chronic Disease*, 40, 373-83.

Cochran, W.G., (1968), "The Effectiveness of Adjustment by Subclassification in Removing Bias in observational studies", *Biometrics*, 24, 295-313.

D'Agostino, R. B. Jr., (1998), "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group", *Statistics in Medicine*, 17, 2265-2281.

Davison, A.C., (2003), *Statistical Models*, Cambridge Unversity Press.

Dawid, A.P., (1979), "Conditional Independence in Statistical Theory", *Journal of the Royal Statistical Society. Series B (Methodological)*, 41, 1-31.

Dobson, A.J., (2001), *An Introduction to Generalized Linear Models, 2nd edition*, Chapman & Hall/CRC.

Du, J., (2002), "Combined Algorithms for Constrained Estimation of Finite Mixture Distributions with Grouped Data and Conditional Data", McMaster University, Hamilton, Ontario, Saatavilla Internetistä: `http://icarus.math.mcmaster.ca/peter/mix/Rmix.pdf`, Luettu 11.7.2005.

Fan, J. & Gijbels, I. (1996), *Local Polynomial Modelling and its Applications*, Chapman & Hall, London.

Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986), "Residual variance and residual pattern in nonlinear regression", *Biometrika*, 73, 625-33.

Hastie, T., Tibshirani R. & Friedman, J. (2001), *Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag New York Inc.

Heckman, J.J., Ichimura, H. & Todd, P. (1997), "Matching As An Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Programme", *Review of Economic Studies*, 64, 605-654.

Heckman, J.J., Ichimura, H. & Todd, P. (1998), "Matching As An Econometric Evaluation Estimator", *Review of Economic Studies*, 65, 261-294.

Holland, P.W. (1986), "Statistics and Causal Inference", *Journal of the American Statistical Association*, 81, 945-60.

Hujanen, T. (2003), "Terveydenhuollon yksikkökustannukset Suomessa vuonna 2001", Aiheita 1/2003, Stakes, Helsinki.

Imbens, G. (2000), "The Role of the Propensity Score in Estimating Dose-Response Functions", *Biometrika*, 87, 706-710.

Lüthje (1984), "Reisiluunkaulan ja trokantterin murtumapotilaiden hoito ja ennuste sekä hoidon kustannukset", väitöskirja, Tampereen yliopisto, Lääketieteellinen tiedekunta.

Marks, R., Allegrante, J. P., MacKenzie, C. R. & Lane, J. M. (2003), "Hip fractures among the elderly: causes, consequences and control", *Ageing Research Reviews*, 2, 57-93.

Mikkola, H-M., Keskimäki, I. & Häkkinen, U. (1998), "Tietoa DRG:stä", Aiheita 39/1998. Stakes. Helsinki

Mikkola, H-M., Keskimäki, I. & Häkkinen, U. (2002), "DRG-Related Prices Applied in a Public Health Care System - Can Finland Learn from Norway and Sweden?", *Health Policy*, 59, 37-51.

Narinen, A., Nurmi, I., Tanninen, S. & Lüthje, P. (2001), *Reisiluun yläosan murtumapotilaiden hoitotulokset, selviytyminen ja kustannukset vuoden aikana Pohjois-Kymenlaaksossa*, Kymenlaakson sairaanhoitopiirin julkaisuja A1/2001

Rosenbaum, P. R. (1987), "Model-Based Direct Adjustment", *Journal of the American Statistical Association*, 82, 387-94.

Rosenbaum, P.R. (2002), *Observational Studies*, New York: Springer.

Rosenbaum, P. R. & Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70,41-55.

Rosenbaum, P. R. & Rubin, D. B. (1984), "Reducing Bias in Observational Studies Using Subclassification on the Propensity Score", *Journal of the American Statistical Association*, 79, 387, 516-524.

Rubin, D.B. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies", *Journal of Educational Psychology*, 66, 688-70.

Rubin, D.B. (1977), "Assignment to Treatment Group on the Basis of a Covariate", *Journal of Educational Statistics*, 2, 1-26.

Rubin, D.B. (1978), "Bayesian Inference for Causal Effects: The Role of Randomization", *The Annals of Statistics*, 6, 34-58.

Rubin, D.B. (1997), "Estimating Causal Effects from Large Data Sets Using Propensity Scores", *Annals of Internal Medicine*, 127, 8(Part 2), 757-763.

Rubin, D.B. & Thomas N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice", *Biometrics*, 52, 249-64.

Ruppert, D., Wand, M.P., Carroll, R.J. (2003), *Semiparametric Regression*, Cambridge University Press.

Sund, R. (2000) "Tilastollisia menetelmiä dynaamisten potilaspopulaatioiden mallintamiseen. Tapahtumahistoria-analyysia hoitoilmoitusrekisterin skitsofreenikoille" Aiheita 26/2000, Stakes, Helsinki.

Sund, R. (2005) "Utilisation of Routinely Collected Register Data in Health System Performance Assessment - the Case of Hip Fracture" Käsikirjoitus (2005).

Sund, R. & Liski, A. (2005), "Quality Effects of Operative Delay on Mortality in the Case of Hip Fracture Treatment", *Quality and safety in health care*, 14, 371-377.

Wand, M.P. & Jones, M.C. (1995), *Kernel Smoothing*, London: Chapman & Hall.

Williams, D. (2001), *Weighing the Odds: A Course in Probability and Statistics*, Cambridge University Press.

Liite A

Muuttujaluettelo

Taulukossa A.1 on esitetty yhteenveto aineiston muuttujista. Lyhennettyjen muuttujanimien selitteet ovat taulukossa A.2. Taulukossa A.2 on selitetty myös luokitteluasteikollisten muuttujien koodaustapa.

Ikä on analyyseja varten luokiteltu alaluvussa 1.1.3 esitetyllä tavalla. Jokaista luokkaa kohti on tehty dummy-muuttuja (saa arvon 1, kun potilas kuuluu luokkaan, muuten 0), joita on käytetty selittäjinä. Murtumatyyppi, leikkaustyyppi ja asuintapa on myös esitetty dummy-muuttujina, joten selittäjiä on yhteensä 31.

Taulukko A.1. Lonkkamurtuma -aineiston muuttujat.

muuttuja	mitta-asteikko	pienin arvo	suurin arvo	mediaani	moodi	keskiarvo
ikä	suhde	65	104	81.2	-	81.3
sukup	luokittelu	0	1	-	1	-
mtyyppi	luokittelu	0	2	-	2	-
ltyyppi	luokittelu	0	2	-	2	-
atapa	luokittelu	0	3	-	0	-
hp365	suhde	0	365	12.7	-	76.8
hp60	suhde	0	60	0	-	16.3
syöpä	luokittelu	0	1	-	0	-
diabetes	luokittelu	0	1	-	0	-
dementia	luokittelu	0	1	-	0	-
vptauti	luokittelu	0	1	-	0	-
svtaudit	luokittelu	0	1	-	0	-
avh	luokittelu	0	1	-	0	-
avkhäiriöt	luokittelu	0	1	-	0	-
khesair	luokittelu	0	1	-	0	-
anemia	luokittelu	0	1	-	0	-
hsairaudet	luokittelu	0	1	-	0	-
ssairaudet	luokittelu	0	1	-	0	-
rjsairaudet	luokittelu	0	1	-	0	-
muut	luokittelu	0	1	-	0	-
ku365	luokittelu	0	1	-	0	-
eplkm	suhde	1	366	366	-	289.9
hplkm	suhde	1	365	63.5	-	126.1
väli	suhde	0	366	55.0	-	139.3
kust	suhde	345.9	105148.0	14463.9	-	19765.2

Taulukko A.2. Muuttujien nimien selitteet.

muuttuja	selite
ikä	ikä (luokiteltu ja käytetty dummy -muuttujia)
sukup	sukupuoli
mtyyppi	murtumatyyppi (0=trokanteerinen , 1=subtrokanteerinen 2=reisiluun kaulan murtuma)
ltyyppi	leikkaustyyppi (0=osaproteesi, 1=kokoproteesi 2=naulaus, ruuvaus ja levytys)
atapa	asumistapa (0=koti, 1=vanhainkoti, 2=terveyskeskus, 3=sairaala)
hp365	hoitopäivien lukumäärä 365 päivän aikana ennen murtumaa
hp60	hoitopäivien lkm. 60 päivän aikana ennen murtumaa
syöpä	syöpä
diabetes	diabetes
dementia	dementia
vptauti	verenpainetauti
svtaudit	sydän- ja verisuonitaudit
avh	aivoinfarkti
avkhäiriöt	aivoverenkierron häiriöt
khesair	krooniset hengityselinten sairaudet
anemia	anemia
hsairaudet	hermoston sairaudet
ssairaudet	silmäsairaudet
rjsairaudet	ruoansulatusjärjestelmän sairaudet
muut	muut sairaudet
ku365	indikaattorimuuttuja, joka kertoo, onko potilas kuollut murtumaa seuranneen vuoden aikana
eplkm	elinpäivien lukumäärä murtumaa seuranneen vuoden aikana (arvo 366 tarkoittaa, että potilas on elänyt yli 365 päivää)
hplkm	hoitopäivien lukumäärä murtumaa seuranneen vuoden aikana
väli	kotona vietettyjen päivien lukumäärä murtuman ja murtumaa edeltäneen hoitojakson välissä (arvo 366 tarkoittaa, että potilas ollut yli 365 päivää kotona)
kust	potilaan kokonaiskustannus murtumaa seuranneen vuoden ajalta

Liite B

Riippumattomuus ja ehdollinen riippumattomuus

Olkoot X ja Y satunnaismuuttujia. Merkitään satunnaisvektorin (X, Y) yhteisjakauman tiheysfunktiota $f(x, y)$, X:n reunajakauman tiheysfunktiota $f_X(x)$ ja X:n ehdollisen jakauman tiheysfunktiota $f(x|y)$, kun $Y = y$. Jos satunnaismuuttuja on diskreetti, niin tiheysfunktio voidaan korvata vastaavalla todennäköisyysfunktiolla. Satunnaismuuttujien X ja Y riippumattomuutta merkitään $X \perp\!\!\!\perp Y$. Jos $X \perp\!\!\!\perp Y$, niin (Casella & Berger 2002, s. 152)

$$f(x, y) = f_X(x) f_Y(y), \tag{B.1}$$

kaikilla X:n ja Y:n arvoilla x ja y. Hajotelman (B.1) kanssa matemaattisesti yhtäpitävä riippumattomuuden määrittelevä ehto on, että yhtäsuuruus

$$f(x|y) = f_X(x), \tag{B.2}$$

on voimassa kaikilla X:n ja Y:n arvoilla x ja y (ks. esim. Casella & Berger 2002, 4. luku tai Dawid 1979). Määritelmät voidaan suoraviivaisesti yleistää satunnaisvektoreille korvaamalla satunnaismuuttujat satunnaisvektoreilla.

Alaluvussa 2.1.3 todettiin, että satunnaistamisesta seuraa tulos (relaatio (2.6))

$$(Y_0, Y_1) \perp\!\!\!\perp Z. \tag{B.3}$$

Jos (B.3), niin ehdon (B.2) mukaan

$$f(y_0, y_1|Z = z) = f(y_0, y_1), \tag{B.4}$$

kun $z = 0, 1$. Integroimalla (B.4) puolittain yli y_1:n arvojen saadaan y_0:n reunajakauman tiheysfunktio

$$f(y_0|Z = z) = f_{Y_0}(y_0),$$

missä $z = 0, 1$. Integroimalla vastaavasti yli y_0:n arvojen, saadaan

$$f(y_1|Z = z) = f_{Y_1}(y_1),$$

missä $z = 0, 1$. Ehdon (B.2) nojalla siis nähdään, että oletuksesta (B.3) seuraa

$$Y_0 \perp\!\!\!\perp Z \text{ ja } Y_1 \perp\!\!\!\perp Z. \tag{B.5}$$

Mutta ehdosta (B.5) ei seuraa (B.3), koska satunnaismuuttujien reunajakaumista ei yleisesti voida johtaa niiden yhteisjakaumaa. Ehto (B.3) on siis vahvempi kuin (B.5).

Tarkastellaan X:n ja Y:n lisäksi myös kolmatta satunnaismuuttujaa Z. Merkitään $Y \perp\!\!\!\perp Z|X = x$, kun Y ja Z ovat riippumattomat ($Y \perp\!\!\!\perp Z$) ehdolla $X = x$. Merkintä $Y \perp\!\!\!\perp Z|X$ tarkoittaa, että $Y \perp\!\!\!\perp Z|X = x$ kaikilla X:n arvoilla $X = x$. Analogisesti riippumattomuuden määritelmien (B.1) ja (B.2) kanssa voimme määritellä ehdollisen riippumattomuuden tiheysfunktioiden (tai todennäköisyysfunktioiden) avulla.

Satunnaismuuttujat Y ja Z ovat ehdollisesti riippumattomat ehdolla X ($Y \perp\!\!\!\perp Z|X$), kun

$$f(y, z|x) = f(y|x)f(z|x) \tag{B.6}$$

kaikilla X:n, Y:n ja Z:n arvoilla x, y ja z. Ehdon (B.6) kanssa yhtäpitävää on, että

$$f(y|z, x) = f(y|x), \tag{B.7}$$

kaikilla x,y ja z. Määritelmät yleistyvät jälleen suoraviivaisesti satunnaisvektoreille. Dawidin (1979) artikkeli on tärkeä ehdollista riippumattomuutta käsittelevä perustutkimus, josta löytyvät keskeisimmät ehdollista riippumattomuutta koskevat tulokset. Myös merkintä $Y \perp\!\!\!\perp Z|X$ on peräisin kyseisestä Dawidin artikkelista.

Ehdollinen riippumattomuus

$$(Y_0, Y_1) \perp\!\!\!\perp Z|\mathbf{X} \tag{B.8}$$

oli Rosenbaumin ja Rubinin (1983) tutkimuksessa eräs keskeinen oletus, missä (Y_0, Y_1) ja $\mathbf{X} = (X_1, \ldots, X_p)$ ovat satunnaisvektoreita. Oletuksesta (B.8) seuraa tulos

$$Y_0 \perp\!\!\!\perp Z|\mathbf{X} \text{ ja } Y_1 \perp\!\!\!\perp Z|\mathbf{X}. \tag{B.9}$$

Tulos (B.9) voidaan osoittaa samalla tekniikalla kuin edellä tulos (B.5). Jos oletetaan (B.8), niin ehdollisen riippumattomuuden määritelmän ehdon (B.7) nojalla

(B.10) $$f(y_0, y_1|z, \mathbf{x}) = f(y_0, y_1|\mathbf{x}).$$

Integroimalla saadaan Y_0:n ja Y_1:n ehdollisia reunajakaumia koskevat identiteetit

$$f(y_0|z, \mathbf{x}) = f(y_0|\mathbf{x}) \text{ ja } f(y_1|z, \mathbf{x}) = f(y_1|\mathbf{x}),$$

jotka ehdon (B.7) nojalla implikoivat tuloksen (B.9). Samoin perustein kuin ehtojen (B.3) ja (B.5) yhteydessä voidaan tässäkin todeta, että ehto (B.8) on voimakkaampi kuin (B.9).

Liite C

Ehdollinen odotusarvo

Riippumattomat satunnaismuuttujat

Olkoon Z indikaattorimuuttuja, joka voi saada arvon 0 tai 1. Silloin pätee identiteetti $P(Z = 1) = 1 - P(Z = 0)$. Lisäksi oletetaan, että $P(Z = 1)P(Z = 0) > 0$. Satunnaismuuttujan Y ehdollinen odotusarvo ehdolla Z voidaan määritellä lausekkeena (Williams 2001, s.385)

$$E(Y|Z) = \frac{E(ZY)}{P(Z = 1)}. \tag{C.1}$$

Jos Y ja Z ovat riippumattomat ($Y \perp\!\!\!\perp Z$), niin

$$\begin{aligned} E(ZY) &= E(Z)E(Y) \\ &= P(Z = 1)E(Y), \end{aligned} \tag{C.2}$$

joten (C.1):stä seuraa identiteetti

$$E(Y|Z) = E(Y). \tag{C.3}$$

Tuloksen (C.2) ensimmäinen yhtäsuuruus seuraa siitä, että riippumattomien satunnaismuuttujien tulon odotusarvo on niiden odotusarvojen tulo (Casella & Berger 2001 s. 154).

Alaluvussa 2.1.3 oletettiin, että $Y_0 \perp\!\!\!\perp Z$ ja $Y_1 \perp\!\!\!\perp Z$ [oletus (2.5)]. Tällöin tuloksen (C.3) mukaan

$$E(Y_0|Z) = E(Y_0) \text{ ja } E(Y_1|Z) = E(Y_1). \tag{C.4}$$

Näin siis oletuksesta (2.5) seuraavat identiteetit (2.4).

Ehdollisesti riippumattomat satunnaismuuttujat

Oletetaan nyt, että $P(Z = 1|\mathbf{X})P(Z = 0|\mathbf{X}) > 0$ ja

$$Y \perp\!\!\!\perp Z|\mathbf{X}, \tag{C.5}$$

missä $\mathbf{X}$ on satunnaisvektori. Satunnaismuuttujan Y ehdollinen odotusarvo ehdolla Z ja $\mathbf{X}$ voidaan määritellä (C.1):n tapaan siten, että

$$E(Y|Z,\mathbf{X}) = \frac{E(ZY|\mathbf{X})}{P(Z=1|\mathbf{X})}. \tag{C.6}$$

Jos $Y \perp\!\!\!\perp Z|\mathbf{X}$, niin

$$\begin{aligned} E(ZY|\mathbf{X}) &= E(Z|\mathbf{X})E(Y|\mathbf{X}) \\ &= P(Z=1|\mathbf{X})E(Y|\mathbf{X}), \end{aligned}$$

joten (C.6):sta seuraa

$$E(Y|Z,\mathbf{X}) = E(Y|\mathbf{X}). \tag{C.7}$$

Liitteessä B osoitettiin, että oletuksesta $(Y_0, Y_1) \perp\!\!\!\perp Z|\mathbf{X}$ seuraavat riippumattomuusrelaatiot $Y_0 \perp\!\!\!\perp Z|\mathbf{X}$ ja $Y_1 \perp\!\!\!\perp Z|\mathbf{X}$. Tuloksen (C.7) perusteella voidaan siis päätellä, että

$$E(Y_0|Z,\mathbf{X}) = E(Y_0|\mathbf{X}) \text{ ja } E(Y_1|Z,\mathbf{X}) = E(Y_1|\mathbf{X}). \tag{C.8}$$

Identiteetit (C.8) seuraavat siis oletusta (B.8) heikommasta oletuksesta (B.9).