CHRNA3-CHRNA5-CHRNB4 -geenialueen assosiaatio painoindeksiin ja verenpaineeseen eri muuttujanvalintamenetelmillä

Tilastotieteen pro gradu -tutkielma

Heli Liuski

Matemaattisten tieteiden laitos

Oulun yliopisto

Kevät 2013

Sisältö

1 Johdanto

Monet nykyään havaittavista sairauksista johtuvat muuttuneista elämäntavoista. Liikalihavuus ja verenpaineen nousu vaikuttavat moneen nykypäivän sairauteen ja aiheuttavat ennenaikaisia kuolemia. Suomalaisesta 15-64 -vuotiaasta väestöstä ylipainoisia vuonna 2011 oli 51% ja lihavia 17%. Vuodesta 1990 ylipainoisten osuus on lisääntynyt noin 13 prosenttiyksikköä ja lihavien noin 16 prosenttiyksikköä. (Terveyden ja hyvinvoinnin laitos 2012.) Lihavuus ei kuitenkaan ole vain suomalaisten ongelma, vaan ylipainosta ja lihavuudesta on tullut yksi maailman yleisimmistä terveysongelmista. Sekä korkea että matala verenpaine voivat olla pitkällä aikavälillä kohtalokkaita. Ihmisten painoindeksiin ja verenpaineeseen vaikuttavat huomattavasti elämäntavat, mutta geeneillä on myös oma osuutensa asiaan. Esimerkiksi vuonna 2008 löydettiin FTO -geeni, joka vaikuttaa ylipainoon ihmisillä (Loos & Bouchard 2008).

Tutkielmassa tarkastellaan CHRNA3-CHRNA5-CHRNB4 -geenialuetta ja sen yhteyttä painoindeksiin ja verenpaineeseen eri muuttujanvalintamenetelmien avulla. Tutkielman tarkoituksena on selvittää, onko kyseisellä geenialueella vaikutusta tarkasteltaviin vasteisiin, vai selittyykö assosiaatio ainakin osittain tupakoinnin kautta. Tutkielman avulla halutaan myös saada selville, ovatko eri muuttujanvalintamenetelmät tarpeeksi luotettavia havaitsemaan assosiaatioita CHRNA3-CHRNA5-CHRNB4 -geenialueen ja vasteiden välillä.

2 Populaatiopohjainen assosiaatiotutkimus

Geneettiset assosiaatiotutkimukset pyrkivät havaitsemaan assosiaatiota yhden tai useamman geenikohdan ja fenotyyppipiirteen tai ominaisuuden välillä. Tämä ominaisuus voi olla joko jokin tauti, muu diskreetti ominaisuus tai kvantitatiivinen arvo. Assosiaatiotutkimuksissa sama alleeli (tai alleelit) on assosioitunut ominaisuuden arvoihin samalla lailla koko populaatiossa. Geneettistä assosiaatiota syntyy vain, jos eri populaatiot jakavat yhteisen historian. (Balding 2006; Cordell & Clayton 2005.)

Toinen assosiaatiotutkimuksen tapa on käyttää perheitä tutkimuksen havaintoyksikkönä. Populaatiopohjaista assosiaatiotutkimusta voidaankin verrata perhepohjaiseen assosiaatiotutkimukseen. Perhepohjaisessa assosiaatiotutkimuksessa käytetään perheitä kontrolloimaan mahdollisia tutkimuspopulaatiossa havaittavia alleelifrekvenssien eroja tietyssä lokuksessa. Eroja voi esiintyä esimerkiksi eri etnisten ryhmien välillä tai syntyä ympäristötekijöiden vuoksi. (Chen & Abecasis 2007.) Yleisimmin perhepohjaisessa assosiaatiotutkimuksessa käytetään TDT-testiä (eng. transmission disequilibrium test), joka tarkastelee alleelien siirtymistä heterotsygoottisilta vanhemmilta heidän jälkeläisilleen. TDT-testi on robusti osittuneelle populaatiolle, vaikka tarkasteltaisiin vain yhtä markkeria kerrallaan. Tämä kuitenkin johtaa siihen, että TDT-testi menettää tehoaan. (Chen & Abecasis 2007.) Populaatiopohjaisissa assosiaatiotutkimuksissa sen sijaan vertaillaan yksilöitä, jotka eivät oletettavasti ole sukua keskenään. Tätä ei kuitenkaan pystytä tutkimuksessa tarkalleen todentamaan, joten sukulaisuuden oletetaan olevan tuntematonta tai riittävän kaukaista. (Balding 2006.)

Assosiaatio geneettisen polymorfismin ja fenotyyppipiirteen tai ominaisuuden välillä voi populaatiossa esiintyä kolmella eri tavalla. Polymorfismilla voi olla kausaalinen rooli (suora assosiaatio) tai polymorfismilla ei ole kausaalista

roolia, mutta se on assosioitunut läheisen kausaalisen geenikohdan kanssa (epäsuora assosiaatio). Assosiaatio voi myös olla seurausta osittuneesta tai sekoittuneesta populaatiosta (sekoittava assosiaatio). (Balding 2006; Cordell & Clayton 2005.) Jos suoraa assosiaatiota ei pystytä havaitsemaan, on mahdollista, että pystytään havaitsemaan epäsuora assosiaatio markkerilokuksen ja fenotyypin välillä tilastollisten tunnuslukujen avulla (Esim. r^2 -arvo, joka kuvastaa tilastollista voimaa havaita kytkentäepätasapaino kahden lokuksen välillä). Kytkentäepätasapainolla (linkage disequilibrium, LD) eli alleeliassosiaatiolla tarkoitetaan tilannetta, jossa eri lokuksissa sijaitsevat alleelit esiintyvät yhdessä useammin tai harvemmin kuin niiden odotetaan esiintyvän. Tämä johtuu siitä, että tarpeeksi läheisten lokusten välille ei ehdi syntyä geenienvaihduntaa sukupolvien aikana niin paljon, että alussa vallinnut assosiaatio purkautuisi täydellisesti. (Zhao ym. 2003.)

Suoran assosiaation tutkimukset kohdentuvat polymorfismeihin, jotka ovat itsessään kausaalisia variantteja. Tämäntyyppinen tutkimus on helpointa analysoida, ja se on kaikkein voimakkain assosiaation tyyppi. Suoran assosiaatiotutkimuksen vaikeaksi tekee kandidaattipolymorfismin tunnistaminen ja mahdolliset kausaaliset muunnokset. Esimerkiksi mutaatio kodonissa, joka johtaa aminohapon muuttumiseen, on kausaalinen muunnos. Kuitenkin on todennäköistä, että monet kausaaliset muunnokset, jotka vaikuttavat monitekijäisen mutaation heritabiliteettiin (geneettisen muuntelun periytyvyysaste), ovat ei-koodaavia. Suoralla assosiaatiotutkimuksella on mahdollista löytää vain joitain geneettisiä syitä tautiin ja tautiin liittyviin ominaisuuksiin. Epäsuorassa assosiaatiossa polymorfismi on korvike kausaaliselle lokukselle. Tällainen assosiaatio antaa mahdollisuuden etsiä kausaaliset geenit. Koska epäsuorat assosiaatiot ovat heikompia kuin suorat assosiaatiot, on yleensä tarpeellista määritellä useita ympäröiviä markkereita epäsuoran assosiaation löytämiseksi. (Balding 2006; Cordell & Clayton 2005.)

Sekoittunut assosiaatio on seurausta osittuneesta populaatiosta, joka on jakautunut kahteen tai useampaan ryhmään jonkin geneettisen ominaisuuden vuoksi. Tällaisessa ongelmallisessa populaatiorakenteessa tarkasteltavia fenotyyppipiirteitä, tautia, muuta diskreettiä ominaisuutta tai kvantitatiivista arvoa esiintyy eri suhteissa itsenäisissä populaatioissa tai geneettisissä alaryhmissä. Osittunut populaatiorakenne voi aiheuttaa vääriä löydöksiä (positiivinen sekoitus) ja todellista kausaalista assosiaatiota (negatiivinen sekoitus). (Balding 2006.) Jos tutkijat eivät ole havainneet ongelmallista populaatiorakennetta, antaa se todennäköisesti virheellisempiä tuloksia kuin se, että yliedustettuja alleeleita ei huomata testauksessa. Ongelmallista populaatiorakennetta lisää se, että kausaalisia genotyyppejä esiintyy alaryhmässä enemmän erilaisten ympäristöolojen takia tai että, jotain alaryhmää suositaan toista enemmän. Esimerkiksi tiettyä erityistä alaryhmää voidaan tutkia tarkemmin terveyspalveluissa kuin tavallista populaatiota, joten kohdetapahtumat alaryhmästä tulevat todennäköisemmin tutkimukseen mukaan kuin kohdetapahtumat tavallisesta populaatiosta. Myös ilmeisesti homogeeninen, eristynyt populaatio kuten Islanti on herkkä sekoitukselle, sillä sinne on tapahtunut muuttoliikettä erilaisista lähtöpopulaatioista. Osittunutta tai sekoittunutta populaatiorakennetta voidaan parantaa genomisen kontrollin avulla. Genomisessa kontrollissa hävitetään osittuneen populaatiorakenteen vaikutus havainnoitaviin SNP-kohtiin. (Balding 2006; Cordell & Clayton 2005; Ziv & Gonzalez Burchard 2003.)

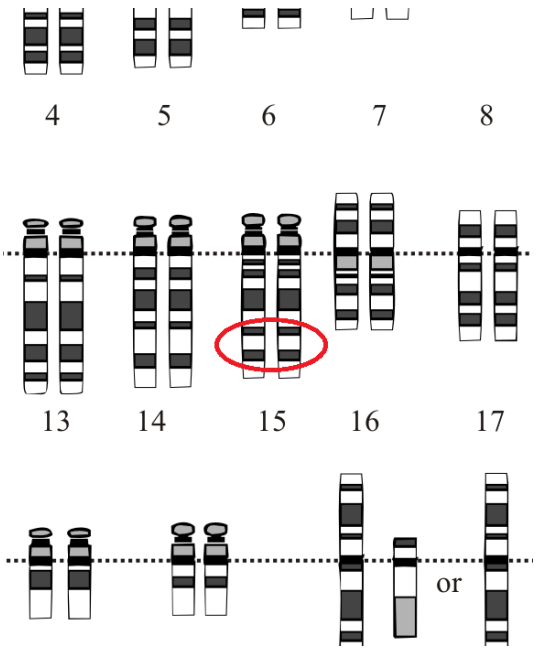
Assosiaatiotutkimuksissa voidaan käyttää seuraavia neljää eri tutkimustyyppiä, jotta pystytään havaitsemaan assosiaatio halutun ominaisuuden kanssa. **Ehdokaspolymorfismitutkimuksissa** keskitytään yksittäisiin polymorfismeihin, joiden epäillään olevan merkitseviä tarkasteltavan taudin vaikutukseen, kun taas **ehdokasgeenitutkimukset** sisältävät 5-50 SNP-kohtaa (yhden nukleotidin polymorfismi, single nucleotidi polymorphism) geenin sisältä. Ehdokasgeeni voi olla paikkaan sidonnainen tai sillä voi olla toiminnalli-

nen asema. **Hienojakoinen kartoitus** yleensä viittaa tutkimukseen, joka on suoritettu ehdokasalueella ja voi sisältää useita satoja yhden nukleotidin polymorfismeja. Ehdokasalue voi olla tässä tapauksessa tunnistettu kytkentätutkimuksella (linkage study), ja se sisältää 5-50 geeniä. **Genominlaajuisessa assosiaatiotutkimuksessa** pyritään tunnistamaan yhteiset kausaaliset muunnokset läpi koko genomin, jolloin tarvitaan yli 300 000 hyvin valittua SNP:tä. Näin monen markkerin tarkastelu on tullut mahdolliseksi kansainvälisen HapMap -projektin ansiosta. Nämä luokittelut eivät kuitenkaan ole tarkkoja. Esimerkiksi jotkin kandidaattigeenitutkimukset voivat sisältää satoja geenejä ja ovat tämän myötä samanlaisia kuin genominlaajuiset assosiaatiotutkimukset. (Balding 2006.)

3 Tutkimusongelma, -aineisto ja mittausmenetelmät

3.1 Tutkimusongelma

Tutkimuksen kohteena on CHRNA5-CHRNA3-CHRNB4 -geenialue kromosomissa 15 ja sen mahdolliset assosiaatiot painoindeksin ja verenpaineen kanssa. Tarkaksi geenialueen kohdaksi määritellään 15q25, josta geenialue CHRNA5-CHRNA3-CHRNB4 voidaan löytää (Kuva 1). Tutkimukseen valittiin 19 eri SNP-kohtaa geenialueelta CHRNA5-CHRNA-CHRNB4. Nämä 19 eri SNP-kohtaa tarkistettiin ja huomattiin, että yhden SNP-kohdan, rs8192475, harvinaisemman alleelin frekvenssi (MAF, minor allele frequency) oli pienempi kuin 0.05. Tämä SNP-kohta jätettiin pois tutkimuksista. Tarkasteltavat kahdeksantoista SNP-kohtaa geenialueelta CHRNA5-CHRNA3-CHRNB4 on esitelty taulukossa 1.



Kuva 1: Alueen 15q25 sijainti kromosomissa 15.

Taulukko 1: Tutkimuksessa käytettävät SNP-kohdat (*Lihavoituna harvinaisempi alleeli).

Kromosomi	rs numero	Esiintyvät alleelit*	Harvinaisemman alleelin frekvenssi
15	rs8034191	**G**/A	0.33
15	rs3885951	**G**/A	0.06
15	rs2036534	A/**G**	0.28
15	rs6495306	A/**G**	0.38
15	rs680244	G/**A**	0.38
15	rs621849	A/**G**	0.38
15	rs1051730	**A**/G	0.32
15	rs6495309	G/**A**	0.27
15	rs1948	G/**A**	0.34
15	rs950776	A/**G**	0.33
15	rs12594247	**A**/G	0.21
15	rs12900519	A/**G**	0.14
15	rs1996371	**G**/A	0.35
15	rs6495314	**C**/A	0.35
15	rs8032156	G/**A**	0.30
15	rs8038920	G/**A**	0.27
15	rs4887077	**A**/G	0.33
15	rs11638372	**A**/G	0.33

Tutkimuksessa analysoidaan kaikkien kahdeksantoista SNP-kohdan vaikutusta tarkasteltaviin vasteisiin. Eri muuttujanvalintamenetelmien avulla pyritään selvittämään, vaikuttaako vasteeseen yksi vai useampi SNP-kohta. Assosiaatioita käsitellään sekä koko aineistossa että tupakoivilla ja tupakoimattomilla henkilöillä erikseen. Halutaan nähdä, onko kyseisellä geenialueella assosiaatioita pelkästään käsiteltäviin vasteisiin, vai selittyvätkö assosiaatiot tupakoinnin kautta. Tutkimus pohjautuu Kaakisen ym. (2012) käyttämään aineistoon, jossa selvitettiin painoindeksin ja verenpaineen keskinäistä yhteyttä tupakoivilla ja ei-tupakoivilla kyseisellä geenialueella CHRNA5-CHRNA3-CHRNB4 yksittäisissä SNP-kohdissa. Aineistona Kaakisen ym. (2012) tutkimuksessa oli Pohjois-Suomen syntymäkohorttitutkimus 1966. Tässä tutkimuksessa myös selvitetään, ovatko tutkimuksessa mahdollisesti havaittavat SNP-

kohdat samoja, kuin mitä Kaakinen ym. (2012) tutkimuksessaan havaitsivat.

Analyysiä varten Kaakinen ym. (2012) muodostivat 10 pääkomponenttia (pc, principal components) korjaamaan populaatiorakennetta ja mahdollisia sukulaissuhteita tutkimuksessa olleilla henkilöillä. He käyttivät näistä kolmea ensimmäistä tällaisen ongelmallisen assosiaation korjaamiseksi. Tässä tutkimuksessa on myös tutkittu näiden pääkomponenttien vaikutusta ja tarpeellisuutta muuttujien valintamenetelmissä. Kaakinen ym. (2012) laskivat pääkomponentit 22:sta suoraan genotyypatusta autosomista, joissa kaikilla MAF oli pienempi kuin 1 prosentti, Hardy-Weinbergin tasapainon mukaan lasketu p-arvo oli pienempi kuin 0.005, call rate oli suurempi kuin 99,5 prosenttia, ja millään kahdella SNP-kohdalla ei ollut LD-arvo r^2 suurempi kuin 0,2 minkään muun kanssa. Aineistoa supistettiin niin, että yksi viidestätoista SNP:stä valittiin mukaan pääkomponenttianalyysiin. Aikaisemmin on havaittu, että tutkimuksessa käytettävästä datasta lasketut pääkomponentit vastaavat hyvin tähän kohorttiin kuuluvien maantieteellistä taustaa. (Sabatti ym. 2009.)

Kyseistä geenialuetta CHRNA5-CHRNA3-CHRNB4 on tutkittu monissa tutkimuksissa. On huomattu, että CHRNA5-CHRNA3-CHRNB4 -geenialue on assosioitunut tupakoinnin kanssa. Ducci ym. (2011) selvittivät, onko kahden geenialueen TTC12-ANKK1-DRD2 ja CHRNA5-CHRNA3-CHRNB4 assosiaatiossa tupakoinnin kanssa eroja iän mukaan Pohjois-Suomen syntymäkohortissa 1966. He havaitsivat, että TTC12-ANKK1-DRD2 -geenialue vaikutti tupakointiin murrosiässä epäsuorasti. CHRNA5-CHRNA3-CHRNB4 -geenialue ja sieltä erityisesti lokus rs1051730 vaikuttivat tupakoinnin runsauteen ja jatkuvuuteen aikuisiällä. Tutkimuksia on tehty monessa eri maassa, sillä tämä geenialue koodaa nikotiiniasetyylikoliinireseptorin osia $\alpha 3$, $\alpha 5$ ja $\beta 4$ (Kaakinen ym. 2012).

Ihmisten lisäksi samaa geenialuetta on tutkittu eläimillä, ja on huomattu, että se osallistuu myös sydän- ja verisuonitoimintoihin kuten verenpaineen ja sydämen sykkeen säätelyyn (Moore ym. 2011; Li ym. 2009). ihmisillä CHRNA5-CHRNA3-CHRNB4 -geenialueen assosiaatiota sydän- ja verisuonitoimintoihin ei kuitenkaan ole paljoakaan tutkittu. Rana ym. (2009) kaksostutkimuksessaan löysivät assosiaation CHRNA5-CHRNA3-CHRNB4 -geenialueen ja systolisen verenpaineen välillä. Myös Freathy ym. (2011) 24000 henkilön meta-analyysissä löysivät interaktion tupakointistatuksen ja painoindeksin välillä. Huomattiin, että tupakoinnin määrä oli assosioitunut matalamman painoindeksin kanssa tupakoivilla. Kaakinen ym. (2012) tutkivat, löytyykö CHRNA5-CHRNA3-CHRNB4 -geenialueelta näyttöä geneettisestä variaatiosta painoindeksin ja verenpaineen suhteen. Heidän tutkimuksessaan analysoitiin myös kyseisten variaatioiden vaikutusta tupakointistatukseen, jos näyttöä variaatiosta löytyi. Tutkimuksessa huomattiin, että variaatio tarkasteltavalla geenialueella on yhteyksissä tupakointistatuksen kanssa vaikuttaen painoindeksiin ja systoliseen verenpaineeseen.

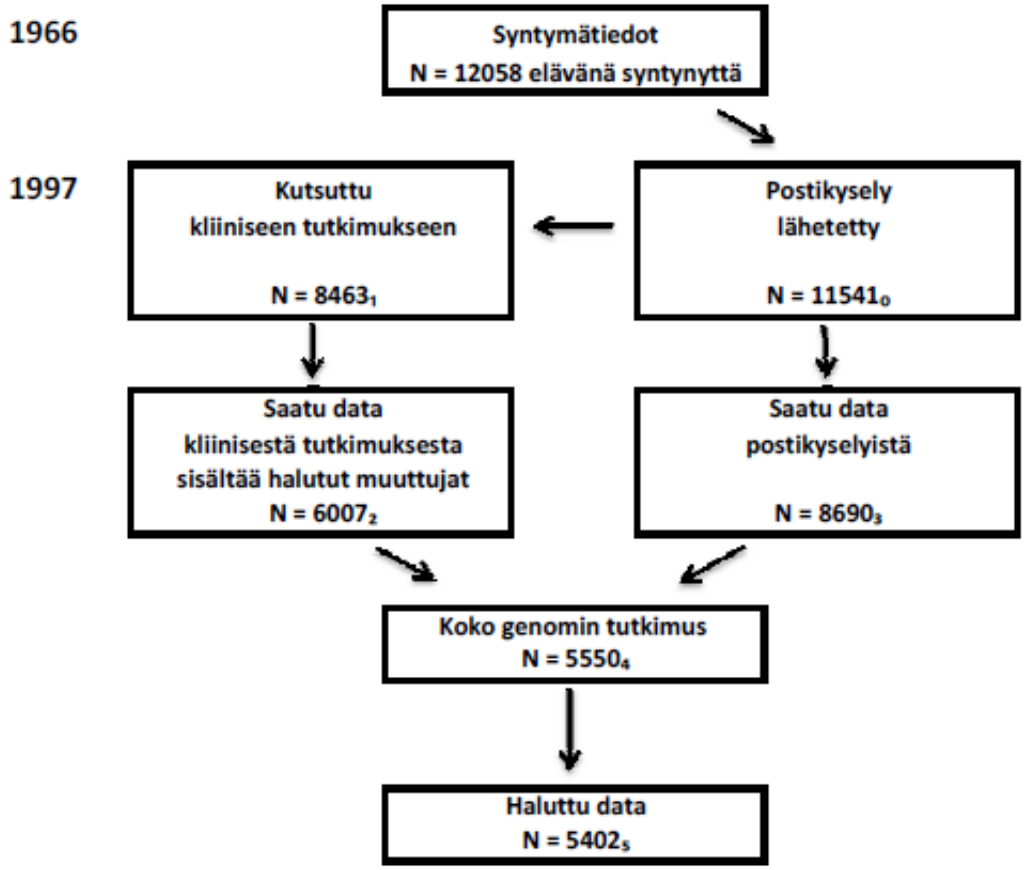
Tutkijat eivät ole varmoja siitä, johtuuko matalampi verenpaine tupakoivilla heidän huonommasta ruokahalustaan vai siitä, että niin tupakoinnilla, painoindeksillä kuin verenpaineella on yhteinen geneettinen tausta. (Rana ym. 2009; Mineur ym. 2011; Thorgeirsson ym. 2008.) Tutkimuksen tavoitteena onkin löytää mahdollisia yhteyksiä CHRNA5-CHRNA3-CHRNB4 -geenialueen välillä joko suorasti tai epäsuorasti samoista lokuksista, mitä Kaakinen ym. (2012) ovat tutkimuksessaan käyttäneet. Koska yhteys vasteiden ja tarkasteltavan CHRNA5-CHRNA3-CHRNB4 -geenialueen välillä ei ole varmaa, tuloksia on tärkeä analysoida oikein. Vaikka saataisiinkin tuloksia, joissa jokin lokus näyttäisi olevan assosioitunut tarkasteltavien vasteiden painoindeksin ja verenpaineen kanssa, tulee miettiä, johtuvatko havainnot epäsuorasta assosiaatiosta.

3.2 Tutkimusaineisto

Tutkimusaineistona on Pohjois-Suomen syntymäkohorttitutkimus 1966, joka on epidemiologinen pitkittäistutkimusohjelma. Sen tarkoituksena on edistää väestön terveyttä ja hyvinvointia. Tutkimukseen valittiin äidit, jotka asuivat Suomen kahdessa pohjoisimmassa läänissä, Oulun ja Lapin lääneissä, ja joiden lasketut ajat osuivat aikavälille 1.1.-31.12.1966. Tutkimuksen aloitti professori Paula Rantakallio jo vuonna 1965, jolloin jotkut äidit olivat raskaana. Tutkimuskohortti sisältää kyseisten äitien synnyttämät lapset. Tutkimusväestön äitejä ja lapsia on seurattu siitä lähtien, kun äidit ilmoittautuivat ensimmäiseen äitiysneuvolakäyntiinsä. Kohorttiin kuuluu 12055 äitiä, jotka synnyttivät 12068 kertaa (13 naista synnytti kahdesti). Tutkimus sisälsi kaikki elossa ja kuolleena syntyneet vauvat, joiden syntymäpaino oli 600 grammaa tai enemmän. Yhteensä 12231 lasta syntyi kohorttiin, joista 12058 oli elävänä syntynyttä (Rantakallio 1969). Tutkimusväestö sisälsi 96.3% kaikista tutkimusalueella vuonna 1966 syntyneistä lapsista.

Vuonna 1992 kohortin jäsenten ollessa 31-vuotiaita toteutettiin poikittaistutkimus, joka koostui postikyselystä, terveystarkastuksesta ja laboratoriomittauksista. Postikysely sisälsi kysymyksiä mm. terveydentilasta, elämäntavoista ja ammatista. (Kaakinen ym. 2012.) Postikysely lähetettiin kaikille kohortin jäsenille, jotka olivat elossa, ja joiden osoite oli tiedossa ($N = 11541$). Kyselylomakkeen palautti 8767 kohortin jäsentä (73.3%). Kaikki ne henkilöt, jotka asuivat Oulun ja Lapin lääneissä tai Helsingin alueella kyseisenä aikana, kutsuttiin terveystutkimukseen ($N = 8463$). Siihen osallistui 6033 henkilöä (71.2%), joista 10 oli kohdealueiden ulkopuolelta tulleita kohortin jäseniä. (Kaakinen ym. 2012; Järvelin ym. 2004.) Verinäytteet otettiin terveystutkimukseen osallistuvilta henkilöiltä. Saaduista verinäytteistä DNA saatiin sekvensoitua onnistuneesti 5733 henkilöltä. Tutkimuksessa käytettävä otos sisältää vain ne henkilöt, jotka antoivat luvan käyttää DNA-näytettään

myöhemmissä tutkimuksissa ($N = 5402$) (Kuva 2).



₀ Kysely lähetetty elossa oleville, joiden osoite on tiedossa
₁ Kliiniseen tutkimukseen kutsuttu ne, jotka asuivat Oulun tai Lapin läänin alueilla tai pääkaupunkiseudulla
₂ Tutkimukseen osallistui 6033, joista 26 ei antanut suostumusta datan käyttöön.
₃ Tutkimukseen osallistui 8768, joista 77 ei antanut käyttää tietojaan.
₄ DNA-näyte vain 5753 henkilöltä.
₅ Pois jätettyjen genomidatassa datan pilaantumista, sukupuolen yhteensopimattomuutta, genotyyppaus kahteen kertaa, sukulaisuussuhteita, suostumuksen perumista ja pientä heterotsygotiaa.

Kuva 2: Kaavio datan keräämisestä kohortti 1966:ssa.

3.3 Mittausmenetelmät

Painoindeksi (engl. body mass index, BMI) on yleisesti käytetty aikuisten ali- tai ylipainoisuuden mittari perustuen yksilön painoon ja pituuteen.

$$\text{Painoindeksi} = \frac{\text{massa(kg)}}{\text{pituus}^2(\text{m}^2)}$$

Painoindeksin arvot ovat iästä riippumattomia ja samat molemmille sukupuolille. Painoindeksi luokitellaan katkaisukohtien mukaan ali-, normaali- ja ylipainoisiin (Maailman terveysjärjestö 2012). (Taulukko 2)

Taulukko 2: Painoindeksin luokittelu.

Alipaino	< 18.50
Normaalipaino	$18.50 - 24.99$
Lievä ylipaino	$25.00 - 29.99$
Lihavuus	≥ 30.00

Systolinen ja diastolinen verenpaine mitattiin kahdesti, jolloin käytettiin mittausten keskiarvoa. Verenpaineen mittaus suoritettiin verenpainemittarilla henkilön oikeasta kädestä tämän ollessa istuma-asennossa. Mittaus suoritettiin vasta, kun henkilö oli istunut 15 minuuttia. Mittauksen laatua tarkkailtiin jatkuvasti, ja mittauksen suorittivat koulutetut hoitajat yhtenäisin järjestelyin. (Vartiainen ym. 2000.) Systolisen ja diastolisen verenpaineen arvot luettiin 2 mmHg:n tarkkuudella. Niiden henkilöiden ($N = 95$), jotka ilmoittivat postikyselyssä käyttävänsä verenpainelääkitystä, systolisen verenpaineen arvoon lisättiin 15 mmHg ja diastolisen verenpaineen arvoon 10 mmHg. (Kaakinen ym. 2012.) Verenpaine tarkoittaa ihmisen suurimmissa valtimoissa olevaa painetta. Suurimmillaan verenpaine on sydämen pumpatessa verta valtimoihin. Tätä kutsutaan systoliseksi verenpaineeksi. Alimmillaan veren-

paine on pumppauksien välissä, jolloin sydän lepää. Tätä kutsutaan diastoliseksi verenpaineeksi. Systolisen verenpaineen ihannearvo on $90 - 129$ ja diastolisen $60 - 84$. Hyvin alhainen verenpaine voi aiheuttaa solujen kuolemaa. Kohonnut verenpaine altistaa muiden tautien (sepelvaltimotauti, aivohalvaus, ym.) kehittymiselle. (Suomen sydänliitto ry. 2012) Verikokeet kohortin henkilöiltä otettiin aamuisin kahdeksan ja yhdentoista välillä. (Sabatti ym. 2009). Kuten Kaakinen ym. (2012) ilmoittavat artikkelissaan, koko genomin genotyyppaus suoritettiin Illumina HumanCNV370DUO Analysis Beadchip -alustalla Broad -instituutissa, Yhdysvalloissa, ensin vain 4936 yksilölle ja jälkeenpäin samoilla menetelmillä lopuille, jolloin saatiin 5550 havaintoa. Näistä 148 poistettiin, koska havainnoissa esiintyi pilaantumista, sukupuolen erilaisuutta genotyyppi- ja fenotyyppihavainnoissa, kaksinkertaista genotyyppausta, sukulaisuutta, suostumuksen perumista ja matalaa heterotsygotiaa. Laaduntarkkailun jälkeen jäi 5402 yksilöä käytettäväksi analyyseihin. (Kaakinen ym. 2012.) (Kuva 2)

3.4 Tilastollinen analyysi

Aineiston muokkaus toteutettiin IBM SPSS Statistics-ohjelmiston versiolla 21.0.0.0. Esitetyt tulokset tehtiin R-ohjelmiston versiolla 2.12.0.

4 Tilastolliset menetelmät

4.1 Lineaarinen regressio

Alaluku perustuu seuraaviin lähteisiin: Yan (2009), Christensen (2011), Seber (1977), Dunn & Clark (1974), Neter ym. (1990), Rawlings (1998), Draper & Smith (1981), Läärä (2009), Rahiala (2005), Casella & Berger (1990) ja Rao (1973).

Ensimmäisen kerran regressio -termiä ja kahden muuttujan suhteiden tutkimisen metodeja esitti brittiläinen Francis Galton 1800-luvun lopulla tutkiessaan perinnöllisyyttä. Hän käytti yksinkertaista lineaarista regressiota, joka tyypillisesti esitetään muodossa

$$y = \beta_0 + \beta_1 x + \varepsilon,$$

jossa y on vastemuuttuja, x on selittävä muuttuja, β_0 on vakiotermi, β_1 on regressiokerroin ja ε on mallin virhetermi. Yksinkertaisessa lineaarisessa regressiossa tutkitaan yhden vastemuuttujan y ja yhden selittävän muuttujan x suhdetta.

Yleisessä lineaarisessa regressiomallissa, jossa epäillään useamman selittävän muuttujan vaikuttavan vastemuuttujan y käyttäytymiseen samanaikaisesti, aikaisempi yhtälö yleistyy muotoon

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \varepsilon_i \\ &= \beta_0 + \sum_{j=1}^{m} \beta_j x_{ij} + \varepsilon_i, \end{aligned}$$

jossa Y_i on vastemuuttuja ($i = 1, \ldots, n$), $x_1, \ldots, x_m$ ovat selittäviä muuttujia, joille x_{ij} on x_j:n i. arvo, β_0 on vakiotermi, β_j ($j = 1, \ldots, m$) ovat selittäjien vaikutusta kuvaavia regressiokertoimia ja ε_i on virhetermi. Oletuksena on,

että $\varepsilon_i \sim N(0, \sigma_i^2)$ ja $\varepsilon \perp\!\!\!\perp \{x_1, \ldots, x_m\}$. Samalla oletetaan, että $\varepsilon_i, \ldots, \varepsilon_n$ ovat täydellisesti riippumattomia toisistaan ja $\sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$.

Määritellään seuraavat matriisit, jotta yleinen lineaarinen regressiomalli voidaan esittää matriisimuodossa:

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

jossa $\mathbf{Y}$ on $n \times 1$ -vektori, joka sisältää havainnot vastemuuttujasta Y_i, $\mathbf{X}$ on $n \times p$ -matriisi, joka sisältää kaikkien selittävien muuttujien arvot sekä vakiotermiä vastaavan yksikkövektorin ($p = m + 1$), $\boldsymbol{\beta}$ on $p \times 1$ -vektori regressiokertoimista ja $\boldsymbol{\varepsilon}$ on $n \times 1$ -vektori virhetermeistä, missä $\mathbf{E}(\boldsymbol{\varepsilon}) = 0$ ja $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$.

Näillä määritelmillä malli voidaan kirjoittaa muodossa

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Tästä seuraa, että satunnaisen vektorin $\mathbf{Y}$ odotusarvo on $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ ja varianssi-kovarianssimatriisi on $\mathrm{Cov}(\mathbf{Y}) = \mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$, jossa $\mathbf{I}$ on $n \times n$ -yksikkömatriisi. Kun $\boldsymbol{\varepsilon}$ on normaalijakautunut, myös $\mathbf{Y}$ on moniulotteisesti normaalijakautunut $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

4.1.1 Lineaarisen regressiomallin parametrien estimoiminen

Suurimman uskottavuuden (Method of maximum likelihood) -menetelmässä parametrit estimoidaan maksimoimalla uskottavuusfunktio. Uskottavuus-

funktio yleisen lineaarisen regressiomallin tapauksessa voidaan esittää vastemuuttujien tiheysfunktioiden avulla:

$$L(\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} f(Y_i; \boldsymbol{\beta}, \sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}.$$

Tästä yhtälöstä saadaan muodostettua logaritmoitu uskottavuusfunktio, joka on muotoa

$$\log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}(\log(2\pi) + \log(\sigma^2)) - \frac{1}{2\sigma^2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2.$$

Uskottavuusfunktiosta voidaan huomata, että se saavuttaa maksiminsa parametrien $\boldsymbol{\beta}$ suhteen silloin ja vain silloin, kun eksponenttitermissä esiintyvä neliösumma

$$\begin{aligned} Q(\boldsymbol{\beta}) &= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_m x_{im})^2 = \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \mathbf{Y}^T\mathbf{Y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \end{aligned}$$

minimoituu. Tässä on käytetty hyväksi sitä, että $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{Y} = (\mathbf{Y}^T\mathbf{X}\boldsymbol{\beta})^T$.

Yhtälön minimipiste voidaan havaita derivoimalla neliösumma $\mathrm{Q}(\boldsymbol{\beta})$ vektorin $\boldsymbol{\beta}$ kaikkien koordinaattien suhteen erikseen, jolloin saadaan

$$\frac{\partial Q}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial Q}{\partial \beta_0} \\ \frac{\partial Q}{\partial \beta_1} \\ \vdots \\ \frac{\partial Q}{\partial \beta_m} \end{bmatrix} = 2(\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{X}^T\mathbf{Y}).$$

Kun saatu lauseke asetetaan nollaksi ja jaetaan yhtälö kahdella, saadaan pienimmän neliösumman yhtälöt yleiselle lineaariselle regressiomallille. Tämä voidaan esittää muodossa

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{Y}.$$

Näitä yhtälöitä kutsutaan normaaliyhtälöiksi, jonka ratkaisu on parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Tämä pätee, kun $\mathbf{X}^T\mathbf{X}$:n käänteismatriisi on olemassa eli $\text{rank}(\mathbf{X}^T\mathbf{X}) = p$ ja $\det(\mathbf{X}^T\mathbf{X}) \neq 0$. Jos käänteimatriisi ei ole olemassa, voidaan käyttää harjanne regressiota (eng. ridge regression), jolloin matriisista $\mathbf{X}^T\mathbf{X}$ muodostetaan kääntyvä mielivaltaisen parametrin α avulla. Tällöin saatava estimaatti on muotoa $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y}$, jossa $0 \leq \alpha \leq \infty$.

Koska ohessa suoritettiin neliösummalausekkeen minimointi, kutsutaan saatua estimaattoria myös pienimmän neliösumman estimaattoriksi.

Kun käytetään aikaisempaa esitystapaa $\widehat{\boldsymbol{\beta}}$:lle, $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$, nähdään että regressiokertoimet ovat vasteen $\mathbf{Y}$ lineaarisia funktioita. Jos oletetaan, että virhetermit ovat harhattomia eli $\mathbf{E}(\boldsymbol{\varepsilon}) = 0$, niin

$$\begin{aligned}\mathbf{E}(\widehat{\boldsymbol{\beta}}) &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}(\mathbf{Y}) \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}.\end{aligned}$$

Tällöin $\widehat{\boldsymbol{\beta}}$ on parametrivektorin $\boldsymbol{\beta}$ harhaton estimaattori. Tämä pätee sillä edellytyksellä, että $\mathbf{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ vasteen odotusarvolle. Vektorin $\widehat{\boldsymbol{\beta}}$ varianssi-kovarianssimatriisin saadaan muotoon

$$\begin{aligned}\text{Cov}(\widehat{\boldsymbol{\beta}}) &= \text{Cov}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y})] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{Cov}[\mathbf{Y}]\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}.\end{aligned}$$

Kun $\boldsymbol{\varepsilon}$ on normaalijakautunut, noudattaa $\widehat{\boldsymbol{\beta}}$ moniulotteista normaalijakaumaa $\widehat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1})$. Yksittäisen regressiokertoimen β_j luottamustason $100(1-\alpha)\%$ luottamusväli on muotoa $\widehat{\beta}_j \pm t_{1-\alpha/2}(n-m) \times \text{SE}(\widehat{\beta}_j)$.

Parametrin σ^2 suurimman uskottavuuden estimaatti saadaan derivoimalla logaritmoitu uskottavuusfunktio kyseisen parametrin suhteen. Derivoituun lausekkeeseen voidaan sijoittaa suurimman uskottavuuden estimaattori $\widehat{\boldsymbol{\beta}}$ parametrin $\boldsymbol{\beta}$ arvoksi ja asettaa tämän jälkeen derivoitu lauseke nollaksi. Kun yhtälö $\frac{\partial \log L(\widehat{\boldsymbol{\beta}},\sigma^2)}{\partial \sigma^2} = 0$ ratkaistaan parametrin σ^2 suhteen, saadaan σ^2:n suurimman uskottavuuden estimaatiksi

$$\widehat{\sigma^2} = \frac{1}{n}||\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}||^2 = \frac{1}{n}\text{SSE}.$$

Saatu estimaatti on hieman harhainen ja poikkeaa harhattomasta estimaatista, joka on tavanomainen jäännöskeskineliösumma MSE $= \frac{\text{SSE}}{n-m}$ (ks. alaluku 4.1.2).

Vastemuuttujan arvojen sovitteet $\widehat{\mathbf{Y}}$ määräytyvät seuraavasti

$$\widehat{\mathbf{Y}} = \begin{bmatrix} \widehat{Y}_1 \\ \widehat{Y}_2 \\ \vdots \\ \widehat{Y}_n \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

ja residuaalit $\mathbf{e}$ vastaavasti

$$\mathbf{e} = \begin{bmatrix} \widehat{e}_1 \\ \widehat{e}_2 \\ \vdots \\ \widehat{e}_n \end{bmatrix} = \mathbf{Y} - \widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

Kun käytetään yhtälössä hattumatriisia $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, sovitteet $\widehat{\mathbf{Y}}$ ja jäännöstermit $\mathbf{e}$ voidaan esittää muodossa $\widehat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} = \mathbf{H}\mathbf{Y}$ ja $\mathbf{e} = (\mathbf{I}-\mathbf{H})\mathbf{Y}$. Hattumatriisi on $n \times n$ -neliömatriisi, joka on sekä symmetrinen eli $\mathbf{H} = \mathbf{H}^T$ että idempotentti eli $\mathbf{HH} = \mathbf{H}$.

4.1.2 Varianssianalyysi

Neliösummien erittely voidaan aloittaa vastemuuttujan Y arvojen kokonaisvaihtelua kuvaavasta kokonaisneliösummasta SSY. Vastaavasti regressioneliösumman SSR ja jäännösneliösumman SSE lausekkeet voidaan esittää annetussa muodossa:

$$\text{SSY} = \|\mathbf{Y} - \bar{\mathbf{Y}}\|^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$\text{SSE} = \mathbf{e}^T\mathbf{e} = \|\mathbf{Y} - \widehat{\mathbf{Y}}\|^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

$$\text{SSR} = \|\widehat{\mathbf{Y}} - \bar{\mathbf{Y}}\|^2 = \sum_{i=1}^{n} (\widehat{Y}_i - \bar{Y})^2.$$

Determinaatiokerroin R^2 on regressioneliösumman ja kokonaisneliösumman osamäärä

$$R^2 = \frac{\text{SSR}}{\text{SSY}} = 1 - \frac{\text{SSE}}{\text{SSY}}.$$

Determinaatiokerroin kertoo mallin selittämän varianssin suhteellisen osuuden vastemuuttujan Y kokonaisvarianssista arvoalueenaan $0 \leq R^2 \leq 1$. Uusien muuttujien lisääminen malliin voi ainoastaan kasvattaa R^2 -arvoa eikä koskaan pienentää sitä. Tämä johtuu siitä, että jäännösneliösumma SSE ei voi koskaan tulla suuremmaksi muuttujia lisättäessä, ja kokonaisneliösumma SSY on aina sama annetuille vasteille. Tämän vuoksi toisinaan suositellaan käytettäväksi determinaatiokertoimesta muokattua suuretta, jossa otetaan huomioon mukana olevien muuttujien määrä. Muokattu determinaatiokerroin on siis muotoa

$$R_a^2 = 1 - \left(\frac{n-1}{n-m}\right)\frac{\text{SSE}}{\text{SSY}}.$$

Jokaiselle neliösummalle saadaan laskettua vapausasteet, jotka kertovat vapaiden muuttujien määrän. Kokonaisneliösumman tapauksessa vapausasteita

on $n-1$ kappaletta, regressioneliösumman tapauksessa $m-1$ kappaletta ja jäännösneliösumman tapauksessa $n-m$ kappaletta. Vapausasteita käyttämällä saadaan laskettua regressiokeskineliösumma MSR ja jäännöskeskineliösumma MSE:

$$\mathrm{MSR} = \frac{\mathrm{SSR}}{m-1} \qquad\qquad \mathrm{MSE} = \frac{\mathrm{SSE}}{n-m}.$$

Keskineliösummien suhteesta saadaan laskettua F-suure

$$F = \frac{\mathrm{MSR}}{\mathrm{MSE}}.$$

Sitä voidaan käyttää nollahypoteesin $H_0 : \beta_1 = \beta_2 = \cdots = \beta_m = 0$ testisuureena, jolloin testataan, onko vastemuuttujan Y ja selittävien muuttujien $X_1, \ldots, X_m$ välillä yhteyttä. Jos nollahypoteesi pätee, $F \sim F_{m-1,n-m}$. Tämä pätee sen takia, koska $\frac{\mathrm{SSR}}{\sigma^2} \sim \chi^2_{m-1}$ ja $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2_{n-m}$.

Uskottavuusosamäärämenetelmä

F-suure on yksi muoto uskottavuusosamäärätestisuureesta. Uskottavuusosamäärämenetelmä hypoteesin testauksessa liittyy suurimman uskottavuuden estimaattoreihin, ja sillä vertaillaan kahta eri mallia H_0 ja H_1, joiden tulee olla sisäkkäisiä. Uskottavuusosamäärätestillä (Likelihood ratio test) halutaan yleensä testata tilanteita, joissa tarkastellaan useita tuntemattomia parametreja.

Olkoon B parametriavaruus eli parametrin $\boldsymbol{\beta}$ mahdollisten arvojen joukko. Havaintojen ja hypoteesin $H_0 : \boldsymbol{\beta} \in \mathrm{B}_0$ mukaisen mallin välistä yhteensopivuutta voidaan mitata maksimiarvolla $\max_{\boldsymbol{\beta} \in \mathrm{B}_0} L(\boldsymbol{\beta}, \sigma^2)$, jossa B_0 on jokin parametriavaruuden B osajoukko. Vastahypoteesina tällöin on $H_1 : \boldsymbol{\beta} \in \mathrm{B}_1$. Parametriavaruuksien B_0 ja B_1 leikkaus on tyhjä joukko ja unioni parametriavaruus B. ($\mathrm{B}_0 \cap \mathrm{B}_1 = \emptyset$, $\mathrm{B}_0 \cup \mathrm{B}_1 = \mathrm{B}$) Jotta voidaan verrata hypoteeseja H_0 ja H_1 keskenään, vertausarvoksi otetaan koko parametriavaruudesta mitatun

uskottavuusfunktion maksimiarvo $\max_{\boldsymbol{\beta}\in \mathrm{B}} L(\boldsymbol{\beta}, \sigma^2)$. Uskottavuusosamäärätestisuure voidaan siis esittää muodossa

$$\lambda(\mathbf{Y}) = \frac{\max_{\boldsymbol{\beta}\in \mathrm{B}_0} L(\boldsymbol{\beta}, \sigma^2)}{\max_{\boldsymbol{\beta}\in \mathrm{B}} L(\boldsymbol{\beta}, \sigma^2)},$$

kun $\mathbf{Y} \in \mathbb{R}^n$. Uskottavuusosamäärätestisuure hypoteesille $H_0 : \boldsymbol{\beta} \in \mathrm{B}_0$ saa arvoja välillä $0 \leq \lambda(\mathbf{Y}) \leq 1$.

Uskottavuusosamäärätestisuure $\lambda(Y)$ voidaan muuntaa devianssiksi

$$D = -2\log[\lambda(\mathbf{Y})] = 2\log[L(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2)] - 2\log[L(\widehat{\boldsymbol{\beta}}_0, \widehat{\sigma}_0^2)]$$

jossa $\widehat{\boldsymbol{\beta}}$ on parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori ja $\widehat{\boldsymbol{\beta}}_0$ on parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori, joka on määrätty olettaen, että testin nollahypoteesi pätee. Devianssille pätee tiettyjen ehtojen vallitessa ja nollahypoteesin pätiessä seuraava asymptoottinen jakaumatulos

$$D = -2\log[\lambda(\mathbf{Y})] \sim \chi^2(n-p).$$

Normaalin lineaarisen regression tapauksessa devianssi saadaan johdettua muotoon $D = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$. Tällä on yhteys jäännösneliösummaan, sillä tässä tapauksessa $D = \frac{1}{\sigma^2}\mathrm{SSE}$.

4.2 Muuttujien valinta

Tärkein osa mallin rakentamisessa on löytää oikeat ja tarpeelliset selittävät muuttujat tutkimuksessa tarkasteltavalle vasteelle. Osa muuttujista voidaan tuntea aikaisempien tutkimusten pohjalta. Muuttujien valinnassa tulee ottaa huomioon myös sellaiset muuttujat, jotka eivät ole kiinnostuksen kohteena, mutta jotka vähentävät sekoittuneisuutta tai pienentävät jäännöskeskihajontaa. (Andersen & Skovgaard 2010.)

Andersen & Skovgaard (2010) listaavat kirjassaan neljä näkökohtaa, jotka kannattaa huomioida muuttujien sisällyttämisessä tutkimukseen. Kun tehdään uutta tutkimusta, tulee ensiksi huomioida muuttujat, jotka ovat kiinnostuksen kohteena. Tällaisia voivat olla tutkimuksesta riippuen esimerkiksi tutkittavat SNP-kohdat, tarkasteltava hoito tai sairauden kehittymiseen liittyvä uusi riskitekijä. Koska nämä muuttujat ovat mielenkiinnon kohteena, on niiden sisällyttäminen tutkimukseen ymmärrettävää. Uusien muuttujien arvioinnissa tarvitaan samalla huolellista harkintaa, koska muuttujien vaikutusta tarkasteltaviin vasteisiin ei tiedetä. Toiseksi kannattaa huomioida aikaisempien tutkimusten muuttujat, joiden on jo havaittu vaikuttavan tarkasteltavaan vasteeseen tietyllä tavalla. Periaatteena on yleensä sisällyttää tällaiset muuttujat malliin ja säilyttää ne mallissa koko tutkimuksen ajan. Näin vältytään siltä, etteivät uudet muuttujat korvaa vanhoja muuttujia, jolloin aikaisemmin havaitut vaikutukset eivät muutu. Kolmanneksi on syytä huomioida perusmuuttujat kuten ikä ja sukupuoli. Näiden oletetaan melkein aina olevan selittävinä tekijöinä tarkasteltavaan vasteeseen. Neljänneksi kannattaa listata kaikki loput mahdolliset selittävät muuttujat. Listasta voi tulla pitkä, eikä usein ole mahdollista sisällyttää kaikkia yhtäaikaisesti malliin. Näitä muuttujia voidaan tarkastella yksitellen tai paria muuttujaa yhdessä, jotta saadaan selville näiden muuttujien mahdollinen tärkeys tarkasteltavalle mallille.

Päätös siitä, kuinka monta muuttujaa ja mitkä mahdolliset muuttujat sisällytetään malliin, voi olla vaikea. Yhden muuttujan vaikutus mallissa voi riippua siitä, mitkä muut muuttujat on sisällytetty malliin. Muuttujien valinta merkitsee samalla valintaa sen osalta, mihin tieteelliseen kysymykseen tutkimuksessa vastataan. Tämä kannattaa huomioida tutkimuksen teon aikana, vaikka järjestyksen tulisi olla käänteinen. Ensin valitaan ne muuttujat, jotka ovat välttämättömiä kysymykseen vastattaessa. Tämän takia tutkimuksen alussa on kannattavaa määritellä tarkasti kysymys, johon halutaan vastaus, eikä vain päämäärättömästi suorittaa mahdollisia analyysejä ja tulkita niiden antamia tuloksia. (Andersen & Skovgaard 2010.) Liian monen kovariaatin valitseminen malliin voi tuottaa epätarkkoja tuloksia. Toisaalta mahdollisten sekoittavien tekijöiden jättäminen pois mallista aiheuttaa harhaa. Kompromissin tekeminen näiden kahden välillä on hyvin hankalaa varsinkin, jos mallissa on vain vähän havaintoja. Ohjeena voidaan käyttää, että mallissa pitää olla ainakin kymmenen kertaa enemmän havaintoja yhtä selittävää muuttujaa kohti, kun tarkastellaan jatkuvaa vastetta. (Andersen & Skovgaard 2010; Vittinghoff ym. 2005.)

Valitessa muuttujia malliin kannattaa olla erityisen varovainen muuttujien välisten suhteiden kanssa. Sekoittavat (confounder), välittävät (mediator) ja vaikutusta muovaavat (effect modifier) tekijät aiheuttavat vääristyneitä ja harhaisia estimaatteja, jos niiden aiheuttamia ongelmia ei huomioda mallin valinnassa. Sekoittavalla tekijällä on yhteys sekä tarkasteltavaan vasteeseen että muuttujaan, niin että molempien arvot riippuvat sekoittavan tekijän arvoista. Välittävä tekijä vaikuttaa tarkasteltavaan vasteeseen, mutta myös tarkasteltava muuttuja vaikuttaa sen arvoihin. Mahdollista on myös, että vaikutus voi vaihdella jonkun tekijän suhteen systemaattisesti, jolloin tätä kyseistä tekijää kutsutaan vaikutusta muovaavaksi tekijäksi. (Andersen & Skovgaard 2010.)

Seuraavassa on esitelty erinäisiä selittävien muuttujien valintamenetelmiä ja -kriteerejä.

4.2.1 Täydellinen haku ja parhaan osajoukon algoritmi

Täydellisen haun (All possible regressions) tekniikan esittelivät Schatzoff, Tsao ja Fienberg vuonna 1968. Täydellinen haku ottaa harkintaan tietyn selittäjäjoukon kontekstissa kaikki mahdolliset regressiomallit ja identifioi muutaman hyvän selittävien muuttujien osajoukon tietyn valintakriteerin mukaan. Selittävien muuttujien osajoukoista saatuja malleja voidaan tutkia sen jälkeen tarkemmin. Tämä johtaa lopullisen regressiomallin valitsemiseen. Jos tarkasteltavana on 10 muuttujaa, joista täydellisen haun periaatteiden mukaan halutaan muodostaa regressiomalli, saadaan $2^{10} = 1024$ erilaista regressiomallivaihtoehtoa. Tämä proseduuri voi olla hyvin aikaavievä, joten malliin harkittavien muuttujien määrä kannattaa pitää pienenä tai käyttää parhaan osajoukon algoritmia. (Seber 1977; Neter ym. 1990; Draper & Smith 1981.)

Täydellisestä hausta on myös muokattu menetelmä nimeltä parhaan osajoukon algoritmit (Best subsets algorithms). Algoritmi käy läpi kaikki regressiomallivaihtoehdot, joissa on maksimissaan h kappaletta muuttujia ($h < n$). Jokaiselle muuttujalukumäärälle algoritmi muodostaa osajoukon, jossa on l kappaletta regressiomalleja. Regressiomallien määrä l yhdessä osajoukossa on yleensä pieni, esimerkiksi $2 - 8$ kappaletta. Osajoukkoon valitaan valintakriteerin arvon mukaan parhaimmat regressiomallit eli yleensä ne, joissa valintakriteerin arvo on pienimmillään. Kaikkien osajoukkojen regressiomallit ($h \times l$ kappaletta) laitetaan tämän jälkeen paremmuusjärjestykseen valintakriteerin mukaan. Osajoukkojen etsiminen vie paljon vähemmän aikaa kuin täydellisessä haussa kaikkien mahdollisten muuttujakombinaatioiden etsintä

(Neter ym. 1990; Draper & Smith 1981.) Parhaan osajoukon valinnassa tällaisella menetelmällä on taipumus tuottaa malleja, jotka näyttävät olevan parempia, kuin ne todellisuudessa ovat. Esimerkiksi käytettäessä korjattua determinaatiokerrointa valintakriteerinä, se valitsee mallin, jossa on pienin jäännöskeskineliösumma. (Christensen 2011.)

4.2.2 Askeltavat menetelmät

Tapauksissa, joissa mahdollisia selittäviä muuttujia on useita kymmeniä, parhaan osajoukon menetelmä ei ole toteuttamiskelpoinen. Automaattinen etsintämenetelmä, joka valitsee yhden jollakin kriteerillä ”parhaan” selittävien muuttujien osajoukon regressiomallista, voi olla tällaisissa tapauksissa hyödyllinen. Askeltavien menetelmien valintakriteereinä voi käyttää joko AIC:tä (Akaiken informaation kriteeri) tai BIC:tä (Bayesiläistä informaatiokriteeriä). Myös F-suuretta on paljon käytetty askeltavissa menetelmissä.

Askeltavista menetelmistä ensimmäinen on taaksepäin eliminoiva menetelmä (The backward elimination procedure). Se on huomattavasti taloudellisempi kuin täydellisen haun menetelmä, koska se yrittää löytää parhaimman regressiomallin, joka sisältää vain menetelmän valitsemat ”merkittävät” muuttujat. Merkitään selittäviä muuttujia x_j, jossa $j = 1, \ldots, m$. Menetelmä käyttää seuraavia vaiheita mallin valitsemiseksi:

1. Valitaan regressioyhtälöön kaikki selittävät muuttujat m kappaletta.
2. Lasketaan valintakriteerin arvo $V_{m-1}(x_j)$ jokaiselle selittävälle muuttujalle x_j erikseen, kuin se olisi ensimmäinen regressiomallista poistettava muuttuja.
3. Alhaisinta kriteerin arvoa $\min V_{m-1}(x_j)$ verrataan regressiomallista, jos-

sa on m-muuttujaa, saatuun valintakriteerin arvoon $V_m(x_j)$.

- Jos $\min V_{m-1}(x_j) < V_m(x_j)$, niin alhaisimman valintakriteerin tuottanut muuttuja x_j poistetaan ja jatketaan jäljelle jäävien muuttujien ($m-1$ kappaletta) kanssa kohdasta 2.
- Jos $\min V_{m-1}(x_j) > V_m(x_j)$, niin hyväksytään regressiomalli, jossa on $m-1$ kappaletta muuttujia.

Askeltamista jatketaan niin kauan, kunnes regressiomalli on hyväksytty tai kaikki muuttujat on poistettu mallista. Taaksepäin askeltavaa meneltelmää käytettäessä tulee huomioida, että kun menetelmä on poistanut jonkin muuttujan pois mallista, sitä ei voida saada enää malliin mukaan, vaikka se olisikin merkitsevä lopullisessa mallissa. (Draper & Smith 1981.)

Toinen askeltavista menetelmistä on eteenpäin valitseva menetelmä (Forward selection algorithm). Sen avulla pyritään samanlaiseen lopputulokseen kuin taaksepäin askeltavassa menetelmässä, mutta lähtien liikkeelle nollamallista, jossa ei ole yhtään selittävää muuttujaa mukana. Malliin yritetään vuorollaan sovittaa aina yksi uusi selittävä muuttuja kerrallaan mallissa jo olevien lisäksi. Valinta jatkuu siihen asti, kunnes joko kaikki selittävät muuttujat ovat mallissa mukana tai lopetuskriteeri on tullut vastaan. Merkitään selittäviä muuttujia x_j, jossa $j = 1, \ldots, m$. Eteenpäin askeltava menetelmä käyttää seuraavia vaiheita mallin valitsemiseksi:

1. Valitaan regressiomalli, jossa ei ole muuttujia mukana.
2. Lasketaan informaatiokriteerin arvo $V_1(x_j)$ jokaiselle tarkasteltavalle selittävälle muuttujalle x_j, kuin se olisi ensimmäinen regressiomalliin lisättävä muuttuja.

3. Alhaisinta kriteerin arvoa $\min V_1(x_j)$ verrataan regressiomallista, jossa ei ole ollenkaan muuttujia, saatuun kriteerin arvoon $V_0(x_j)$.

    - Jos $\min V_1(x_j) < V_0(x_j)$, niin lisätään muuttuja x_1 ja jatketaan jäljelle jäävien muuttujien kanssa kohdasta 2.
    - Jos $\min V_1(x_j) > V_0(x_j)$, niin hyväksytään regressiomalli.

Tätä menettelytapaa jatketaan niin kauan, kunnes regressiomalli on hyväksytty tai kaikki muuttujat on valittu malliin. Selittävä muuttuja, joka on aiemmassa vaiheessa valittu mukaan malliin, voi osoittautua tarpeettomaksi lopullisessa mallissa kyseisen selittävän muuttujan ja muiden mallissa olevien selittävien muuttujien yhteyden vuoksi. (Yan 2009; Draper & Smith 1981.)

On myös mahdollista tarkastella eteenpäin ja taaksepäin askeltavan menetelmän yhdistelmää, jota kutsutaan askeltavaksi valintamenetelmäksi (Stepwise algorithm). Askeltavassa valintamenetelmässä käytetään vuorotellen eteenpäin valitsevaa ja taaksepäin eliminoivaa menetelmää. Valintamenetelmää toistetaan, kunnes yhtään muuttujaa ei voida lisätä malliin tai poistaa mallista. Askeltava menetelmä yhdistää automaattisten valintamenetelmien hyvät puolet, sillä se tarkastelee paremmin mallin kokonaisuutta eikä vain yhden muuttujan poistamisen tai lisäämisen paremmuutta kuten eteenpäin tai taaksepäin askeltavissa menetelmissä.

Askeltavat menetelmät ovat herättäneet keskustelua siitä, valitsevatko ne todella parhaimman mallin, koska askeltavat menetelmät päätyvät vain yhteen regressiomalliin lisäämällä tai poistamalla vain yhden muuttujan kerrallaan. Onkin ehdotettu seuraavaa: Kuten aiemminkin toteutetaan eteenpäin askeltava valinta, jonka jälkeen tarkastellaan saadun mallin selittävien muuttujien lukumäärää k. Päätetäänkin olla hyväksymättä kyseinen malli. Sen sijaan tarkastellaan kaikista selittävistä muuttujista m muodostuvia osa-

joukkoja, jossa on k selittävää muuttujaa. Näistä osajoukoista valitaan parhain, joka hyväksytään lopulliseksi malliksi. Tällainen muokattu menetelmä voi antaa paremman tuloksen, mutta toisaalta se vie kovin paljon aikaa eikä välttämättä tuota parempaa tulosta. (Christensen 2011; Draper & Smith 1981.)

4.2.3 Valintakriteerit

Tässä kappaleessa on esitelty aiemmin mainituille eri mallinvalintamenetelmille sopivia valintakriteerejä.

Determinaatiokerroin ja muokattu determinaatiokerroin

Kuten aikaisemmin on jo esitetty, determinaatiokerroin ja muokattu determinaatiokerroin ovat muotoa

$$R^2 = \frac{\text{SSR}}{\text{SSY}} \qquad ja \qquad R_a^2 = 1 - \left(\frac{n-1}{n-m}\right)(1-R^2).$$

Malleja, joissa on eri määrä selittäviä muuttujia, ei voida verrata determinaatiokertoimella eikä muokatulla determinaatiokertoimella.

Determinaatiokertoimen arvoon R^2 perustuvassa valintakriteerissä ei voida valita mallia, joka maksimoi determinaatiokertoimen arvon. Tämä johtuu siitä, että jäännösneliösumma SSE ei voi koskaan suurentua, jos lisätään uusia selittäviä muuttuja malliin. Niinpä tarkoituksena on löytää kohta, jolloin selittävien muuttujien x_j lisääminen ei ole merkityksellistä. Tällöin muuttujien lisääminen johtaa vain pieneen R^2 -arvon lisäykseen. (Neter ym. 1990.)

Mallowsin $\mathbf{C_p}$ -kriteeri

Mallowsin C_p-suure, jonka kehitti C. L. Mallows vuonna 1973, voidaan esittää

muodossa

$$C_p = \frac{\text{SSE}_p}{\text{MSE}} - (n - 2p),$$

jossa SSE_p on jäännösneliösumma mallista, joka sisältää p regressiokerrointa sisältäen myös β_0-parametrin, ja MSE on jäännöskeskineliösumma koko mallista, jossa on mukana kaikki parametrit. Kun yhtälö, jossa on p parametria, on sopiva, saadaan approksimoimalla $\mathbf{E}(C_p) = p$. Tällöin kuvaajasta, jossa suure C_p on verrattuna parametrien lukumäärään, pisteistä saadaan likimain seuraava yhtälö: $C_p = p$. Paras malli valitaan vasta, kun suureen C_p -kuvaajaa ollaan analysoitu. (Draper & Smith 1981.)

Ennusteneliösumma

Ennusteneliösumma on perustava valintakriteeri eli PRESS-kriteeri, joka pohjautuu ennusteresiduaaleihin $d_i = Y_i - \widehat{Y}_{i(i)}$, jossa $\widehat{Y}_{i(i)}$ on mallin antama ennuste i. havaintoyksikölle, kun regressiomalli on sovitettu ilman i. havaintoa. PRESS-kriteeri on ennusteresiduaalien neliösumma:

$$\text{PRESS}_p = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (Y_i - \widehat{Y}_{i(i)})^2.$$

Malleja, joilla on pieni PRESS-arvo, ajatellaan hyviksi ehdokasmalleiksi. (Draper & Smith 1981.)

Schwarzin bayesiläinen informaatiokriteeri

Bayesiläisen informaatiokriteerin (BIC, Bayesian information criterion) johti G. E. Schwarz vuonna 1978 Bayes-faktorin pohjalta. Bayes-faktorin avulla tarkastellaan, kumpi kahdesta havaitulle datalle $\mathbf{Y}$ muodostetusta mallista M_1 ja M_2 (M_s, $s = 1, 2$) on parempi. Kumpikin malli voi olla lineaarinen regressiomalli, ja mallit voivat sisältää eri määrän selittäviä muuttujia. Bayes-faktori on muotoa

$$\text{BF}_{12} = \frac{p(\mathbf{Y}; M_1)}{p(\mathbf{Y}; M_2)},$$

jossa $p(\mathbf{Y}; M_s) = \int f(\mathbf{Y}; \boldsymbol{\beta}, M_s) p(\boldsymbol{\beta}; M_s)\, d\boldsymbol{\beta}$ on mallin M_s marginaaliuskottavuusfunktio. Siinä integroidaan kaikkien mahdollisten parametrin $\boldsymbol{\beta}$ arvojen yli. Tuntematon parametri $\boldsymbol{\beta}$ sisältää kummankin mallin tapauksessa tarvittavat parametrit $\beta_0 \ldots \beta_m$. Funktio $p(\boldsymbol{\beta}; M_s)$ on parametrin $\boldsymbol{\beta}$ priorijakauman tiheysfunktio. Priorijakauma kuvastaa ennakkokäsityksiä parametrista $\boldsymbol{\beta}$. Kun data $\mathbf{Y}$ on hankittu, parametrin $\boldsymbol{\beta}$ jakauma päivitetään posteriorijakaumaksi käyttämällä Bayesin kaavaa. Funktio $f(\mathbf{Y}; \boldsymbol{\beta}, M_s)$ on uskottavuusfunktio mallille M_s.

Bayes-faktori antaa numeerisen mitan datan antamasta näytöstä mallin M_1 hyväksi malliin M_2 verrattuna. Usein on helpointa tarkastella Bayes-faktorista muunnosta $T_B = -2 \log \mathrm{BF}_{12}$, sillä se antaa tuloksen samalla skaalalla kuin devianssi. Suuretta T_B voidaan approksimoida Laplacen approksimaation avulla kahden mallin vertailuun

$$T_B \approx 2[\log L(\widehat{\boldsymbol{\beta}}_2) - \log L(\widehat{\boldsymbol{\beta}}_1)] - (q_2 - q_1) \log(n),$$

jossa q_s on parametrien lukumäärä mallissa M_s. BIC-kriteeri yhden mallin vertailuun voidaan esittää muodossa

$$\mathrm{BIC} = -2 \log L(\widehat{\boldsymbol{\beta}}) + q \log(n),$$

jossa q on parametrien lukumäärä mallissa. Kun vastemuuttuja oletetaan normaalijakautuneeksi lineaarisen regressiomallin yhteydessä, voidaan BIC-kriteerille johtaa mallin jäännösneliösummasta riippuva esitystapa. Tällöin logaritmoitu uskottavuusfunktio tulee ratkaista parametrin σ^2 suhteen, kun $\boldsymbol{\beta}$ on saatu ratkaistua kuten aikaisemmin (s. 21). Sen jälkeen estimoitu logaritminen uskottavuusfunktio voidaan sijoittaa aiempaan BIC-kriteerin kaavaan, ja saadaan seuraava esitysmuoto BIC-kriteerille:

$$\mathrm{BIC} = n \log\left(\frac{\mathrm{SSE}}{n}\right) + (p+1) \log(n),$$

jossa n on havaintojen lukumäärä, $p+1$ on parametrien lukumäärä mallissa

ja SSE on jäännösneliösumma. (Kass & Raftery 1995; Posada & Buckley 2004; Kuha 2004; Burnham & Anderson 2004.)

Akaiken informaatiokriteeri

Vuonna 1974 H. Akaike johti informaatioteoriaan pohjautuvan kriteerin, joka nykyään tunnetaan Akaiken informaatiokriteerinä (AIC, Akaike information criterion).

Oletetaan, että havaittavan vasteen $\mathbf{Y}$ todellinen malli, jonka jakauman tiheysfunktio olkoon $g(\mathbf{Y})$, on tuntematon. Todellisen mallin ei oleteta liittyvän mihinkään muuhun tarkasteltavaan malliin, joten tiheysfunktion $g(\mathbf{Y})$ ei tarvitse olla samanlainen minkään ajatellun mallin M_s uskottavuusfunktion $f(\mathbf{Y}; \boldsymbol{\beta}, M_s)$ kanssa millään parametrin $\boldsymbol{\beta}$ arvolla. Näin ollen mallin valinnan tulosta ei voida samaistaa mihinkään tiettyyn malliin, vaan täytyy luoda yksinkertaisempia malleja, jotka ovat hyviä approksimaatioita todellisesta mallista. Vertailu todellisen mallin ja ehdokasmallin välillä voidaan esittää tiheyksien $g(\mathbf{Y})$ ja $f(\mathbf{Y}; \boldsymbol{\beta}, M_s)$ jollakin etäisyysmitalla. Kullbackin-Leiblerin etäisyys (KL-etäisyys) on perustana Akaiken informaatiokriteerille ja se voidaan esittää muodossa

$$\begin{aligned} I[g(\mathbf{Y}),\, f(\mathbf{Y}; \boldsymbol{\beta},\, M_s)] &= \int g(\mathbf{Y}) \log [g(\mathbf{Y})] dY - \int g(\mathbf{Y}) \log [f(\mathbf{Y}; \boldsymbol{\beta},\, M_s)] dY \\ &= \mathrm{E}_{\mathbf{Y}}\{\log [g(\mathbf{Y})]\} - \mathrm{E}_{\mathbf{Y}}\{\log [f(\mathbf{Y}; \boldsymbol{\beta},\, M_s)]\} \\ &= C - \mathrm{E}_{\mathbf{Y}} \log [f(\mathbf{Y}; \boldsymbol{\beta},\, M_s)], \end{aligned}$$

jossa $C = \int g[\mathbf{Y}] \log [g(\mathbf{Y})] d\mathbf{Y}$ ja $\mathrm{E}_{\mathbf{Y}}$ kuvaa vasteen $\mathbf{Y}$ odotusarvoa ottaen huomioon todellisen jakauman.

Ilmeinen strategia olisi valita mallien vertailuun sellainen malli, joka on lähimpänä todellista mallia KL-etäisyydellä mitattuna. Tätä ei pystytä kokonaan kuitenkaan toteuttamaan, sillä useat suureet ovat tuntemattomia. Suu-

re C jää tuntemattomaksi, sillä se riippuu vain todellisesta jakaumasta. Koska se toisaalta on sama kaikille ehdokasmalleille M_s, se voidaan poistaa ja huomioida vain etäisyyden $I[g(\mathbf{Y}),\ f(\mathbf{Y};\boldsymbol{\beta},\ M_s)]$ eroja eri mallien välillä.

KL-etäisyys kertoo todellisen mallin ja yhden ehdokasmallin M_s välisen eroavaisuuden parametrin arvolla $\boldsymbol{\beta}$. Sen sijaan olisi mielekästä vertailla tiheysfunktion g ja kaikkien niiden tiheysfunktioiden, jotka sisältävät ehdokasmallin M_s, välistä eroa. Tätä voidaan tarkastella etäisyyden $I[g(\mathbf{Y}),\ f(\mathbf{Y};\boldsymbol{\beta}^M,\ M_s)]$ avulla, jossa $\boldsymbol{\beta}^M$ on arvo, joka minimoi KL-etäisyyden kaikista mahdollisista parametrin $\boldsymbol{\beta}$ arvoista.

Tuntematon parametri $\boldsymbol{\beta}^M$ voidaan estimoida korvaamalla se suurimman uskottavuuden estimaatilla $\widehat{\boldsymbol{\beta}}$. Näin ollen kahta mallia pystytään vertaamaan käyttämällä erotusta $I[g(\mathbf{Y}),\ f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_1,\ M_1)] - I[g(\mathbf{Y}),\ f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_2,\ M_2)]$. Erotuksen tarkastelussa syntyy ongelma, jos mallit M_1 ja M_2 ovat sisäkkäisiä. Tämän takia tulee tarkastella suurimman uskottavuuden estimaattia, joka perustuu erillisiin, itsenäisiin otoksiin datasta $\mathbf{Y}$, jonka oletetaan noudattavan samaa todellista mallia, jonka tiheys on jakauma g. Näin ollen saadaan suure

$$\begin{aligned} T_A &= -2\mathrm{E}_{\mathbf{Y}}\left\{I[g(\mathbf{Y}),\ f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_1, M_1)] - I(g(\mathbf{Y}),\ f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_2, M_2)]\right\} \\ &= -2\mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}}\left\{\log\left[f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_2;\ M_2)\right] - \log\left[f(\mathbf{Y};\widehat{\boldsymbol{\beta}}_1;\ M_1)\right]\right\}. \end{aligned}$$

Tästä muunnoksesta saatavat estimaatit ovat samalla asteikolla devianssin kanssa. Suure T_A on se, jota AIC estimoi. Kun T_A on positiivinen, odotettu KL-etäisyys on pienempi mallin M_2 ja todellisen mallin välillä kuin mallin M_1 ja todellisen mallin. Tässä tapauksessa T_A on todellisen jakauman $\mathbf{Y}$ ominaisuus ja sen vuoksi tuntematon. Tätä silti pystytään estimoimaan havaitun datan avulla ja tiettyjen ehtojen vallitessa, jolloin konsistentti estimaattori suureelle T_A on

$$T_A \approx \mathrm{AIC}_e = 2[\log L(\widehat{\boldsymbol{\beta}}_2) - \log L(\widehat{\boldsymbol{\beta}}_1)] - 2(q_2 - q_1),$$

jossa q_s on parametrien lukumäärä mallissa M_s. Tätä suuretta voidaan käyttää kahden mallin vertailuun. Yhden mallin vertailuun voidaan Akaiken informaatiokriteeri esittää yleisessä muodossaan

$$\mathrm{AIC} = -2 \log L(\widehat{\boldsymbol{\beta}}) + 2q.$$

Kun vastemuuttuja oletetaan normaalijakautuneeksi lineaarisen regressiomallin yhteydessä, voidaan BIC-kriteerille johtaa mallin jäännösneliösummasta riippuva esitystapa. Tällöin toimitaan samoin kuin BIC-kriteerin tapauksessa ja ratkaistaan logaritmoitu uskottavuusfunktio σ^2 suhteen kuten sivulla 21. Tällöin saadaan seuraava esitysmuoto AIC-kriteerille:

$$\mathrm{AIC} = n \log \left(\frac{\mathrm{SSE}}{n} \right) + 2(p+1),$$

jossa n on havaintojen lkm, $p+1$ on parametrien lukumäärä mallissa ja SSE on jäännösneliösumma. (Yan 2009; Posada & Buckley 2004; Kuha 2004; Burnham & Anderson 2004.)

Akaiken informaatiokriteerin arvo antaa tehokkaimman valinnan pienessä otoksessa, kun taas baysiläinen informaatiokriteeri toimii parhaiten suuressa otoksessa. (Yan 2009).

4.2.4 Mallinvalintamenetelmien kritiikki

Jokaisessa edellä mainitussa mallinvalintamenetelmässä on jotain epäluotettavuutta lisääviä tekijöitä, jotka heikentävät menetelmän käytettävyyttä. Jotkut tutkijat uskovatkin, että valintamenetelmillä saadut mallit johtuvat ainoastaan vaikutusvaltaisista havainnoista. Vaikutusvaltaiset havainnot, jotka poikkeavat paljon muusta havaitusta datasta, ovat ongelma regressioanalyyseissä. Muuttujien valintamenetelmissä sovitetaan useita malleja havaitulle datalle, jolloin vaikutusvaltaisten havaintojen ongelma moninkertaistuu.

Useiden tilastotieteilijöiden mielestä vaikutusvaltaisilla havainnoilla on liian suuri rooli mallinvalintamenetelmissä, minkä takia he eivät tutkimuksissaan halua käyttää näitä menetelmiä ollenkaan. (Christensen 2011.)

Muuttujien valinta on kokeilevaa. Jos tiedetään, että tietyt muuttujat ovat tärkeitä, ne tulee silloin sisällyttää malliin. Jos mallin valinta ja mallin sopivuuden tarkastelu tehdään samalla datalla, voidaan saada aikaiseksi harhaisia estimaatteja. Tämän vuoksi myös saadut testisuureiden arvot ja luottamusvälit voivat olla harhaisia, ja niitä kannattaa tarkastella varoen. Mahdollisuutena on jakaa käytettävä data puoliksi. Toisella osalla tehdään mallin valinta ja toista osaa käytetään mallin sopivuuden tarkastelemiseen (Christensen 2011.) Tällainen ristiinvalidointimenetelmä antaa usein tuloksen, että muuttujat, jotka aluksi valittiin malliin, eivät olekaan niin vahvoja muuttujia, kuin mitä mallinvalintamenetelmillä saatujen tuloksien perusteella voitiin luulla. Jos malli antaa puoliksi jaetun datan toisella osalla hyvin erilaisia tuloksia, kuin mitä toisella osalla saatiin mallin valinnassa, ei valittuun malliin voida luottaa. (Andersen & Skovgaard 2010.)

Eteenpäin ja taaksepäin askeltavat menetelmät voivat olla harhaanjohtavia, sillä ne voivat valita ensimmäisenä malliin mukaan muuttujan, joka ei lopulta olekaan tärkeä lopullisen mallin kannalta. Esimerkiksi eteenpäin valitsevan menetelmän tapauksessa ensimmäinen malliin mukaan otettu muuttuja ei välttämättä ole merkittävä lopullisessa mallissa. Voi olla myös mahdollista, että ensimmäinen muuttuja, jonka taaksepäin eliminoiva menetelmä mallista poistaa, on ensimmäinen muuttuja, jonka eteenpäin valitseva menetelmä ottaa malliin mukaan. Askeltavat menetelmät eivät yleensä anna parasta mallia, ja pahimmassa tapauksessa mikään askeltavan menetelmän valitsemista muuttujista ei sisälly parhaaseen malliin. Tämä johtuu siitä, että nämä menetelmät tarkastelevat vain yhtä muuttujaa kerrallaan. (Hocking 1976.)

Tällaiset automaattiset mallinvalintamenetelmät kärsivät myös herkkyydestä, sillä pienikin muutos yhdessä muuttujassa tai muutos havainnoissa voi johtaa täysin erilaiseen malliin. Osaltaan tämä johtuu askeltavissa menetelmissä siitä, että nämä menetelmät käsittelevät kaikkia muuttujia samalla lailla. Jokainen muuttuja on samanarvoinen toisen muuttujan kanssa, eikä yhteisvaikutusta huomioida lainkaan. Askeltaviin menetelmiin on mahdollista ottaa mukaan interaktiotermejä, jos ne on laskettu erikseen, ennen kuin käytetään askeltavia menetelmiä. Interaktiotermien käyttäminen askeltavissa menetelmissä aiheuttaa usein vieläkin harhaisempia malleja kuin askeltavien menetelmien käyttäminen ilman interaktiotermejä. Myöskin tällaisen mallin tulkinnasta tulee äärimmäisen monimutkaista. (Andersen & Skovgaard 2010.)

5 Tulokset

Pohjois-Suomen syntymäkohortista 1966 peräisin olevasta aineistossa naisia oli 2810 (52%) ja miehiä 2592 (48%). Taulukossa 3 on esitelty naisten ja miesten keskiarvot, keskihajonnat sekä vaihteluväli erikseen tupakoiville ja tupakoimattomille. Tupakoivia oli aineistossa 2098 (39%) henkilöä, joista 941 (45%) oli naisia ja 1157 (55%) miehiä. Vasteiden arvoissa tupakoimattomien ja tupakoivien populaatiossa ei ole havaittavissa suuria eroja.

Taulukko 3: Vastemuuttujien keskiarvot, keskihajonnat ja vaihteluväli tupakoivilla ja tupakoimattomilla.

	sukupuoli	tupakointi	Keskiarvo	Keskihajonta	Minimi	Maksimi
Painoindeksi	Naiset	ei	24.1	4.3	16.1	54.4
(kg/m^2)	Naiset	kyllä	24.6	5.2	15.4	53.8
	Miehet	ei	25.2	3.4	16.3	44.5
	Miehet	kyllä	25.2	3.7	15.3	47.6
Systolinen	Naiset	ei	121	12.4	85	195
verenpaine	Naiset	kyllä	119	12.7	82	179
(mmHg)	Miehet	ei	131	12.8	94	204
	Miehet	kyllä	130	12.8	95	188
Diastolinen	Naiset	ei	76	10.6	37	122
verenpaine	Naiset	kyllä	74	11.3	35	123
(mmHg)	Miehet	ei	81	11.5	42	121
	Miehet	kyllä	80	11.3	46	130

Tuloksien seuraamisen helpottamiseksi SNP-kohdista ei käytetä todellisia rs-numeroita, vaan muuttujien niminä ovat taulukossa 4 esitettävät. Näistä yksi SNP-kohta geno8 eli rs8192475 jätettiin tutkimuksista pois, ja tämän vuoksi SNP-kohtien numerointi ei ole jatkuva. Pääkomponenteista käytetään tulostuksissa nimeä *pc* juoksevalla numeroinnilla (*pc1, pc2, pc3, pc4, pc5, pc6, pc7, pc8, pc9, pc10*). Tulostuksissa esiintyy myös ajoittain sukupuoli-

muuttujaa vastaava termi sex. Vasteista käytetään myös lyhenteitä. Painoindeksiä kuvataan tulostuksissa termillä *bmi*. Systolisesta ja diastolisesta verenpaineesta käytetään englanninkielisiä lyhenteitä *sbp* ja *dbp* (systolic blood pressure, diastolic blood pressure). Kuvassa 3 on esitetty SNP-kohtien väliset korrelaatiokertoimet, jotka kertovat kytkentäepätasapainon määrästä SNP-kohtien välillä. Kuvasta on havaittavissa voimakkaita korrelaatioita joidenkin SNP-kohtien välillä.

Taulukko 4: Tutkimuksessa käytettävät SNP-kohdat ja niistä tuloksissa käytettävät nimet.

rs numero	nimi tuloksissa
rs8034191	geno1
rs3885951	geno2
rs2036534	geno3
rs6495306	geno4
rs680244	geno5
rs621849	geno6
rs1051730	geno7
rs6495309	geno9
rs1948	geno10
rs950776	geno11
rs12594247	geno12
rs12900519	geno13
rs1996371	geno14
rs6495314	geno15
rs8032156	geno16
rs8038920	geno17
rs4887077	geno18
rs11638372	geno19

	geno1	geno2	geno3	geno4	geno5	geno6	geno7	geno9	geno10	geno11	geno12	geno13	geno14	geno15	geno16	geno17	geno18	geno19
geno1	1,000																	
geno2	0,369	1,000																
geno3	-0,439	-0,145	1,000															
geno4	-0,524	-0,216	-0,476	1,000														
geno5	0,525	0,216	0,476	-0,998	1,000													
geno6	-0,523	-0,215	-0,476	0,997	-0,998	1,000												
geno7	-0,958	-0,375	0,426	0,530	-0,531	0,529	1,000											
geno9	0,423	0,135	-0,890	0,473	-0,474	0,474	-0,414	1,000										
geno10	0,440	0,203	0,428	-0,862	0,863	-0,862	-0,485	-0,445	1,000									
geno11	-0,429	-0,200	-0,409	0,831	-0,832	0,832	0,474	0,424	-0,910	1,000								
geno12	0,214	0,113	-0,598	0,348	-0,348	0,348	-0,205	0,609	-0,348	0,345	1,000							
geno13	-0,127	-0,113	-0,100	0,175	-0,174	0,173	0,135	0,125	-0,206	0,138	0,222	1,000						
geno14	0,698	0,353	-0,435	-0,267	0,268	-0,267	-0,694	0,417	0,313	-0,306	0,356	-0,303	1,000					
geno15	0,697	0,353	-0,435	-0,265	0,266	-0,265	-0,693	0,417	0,312	-0,305	0,355	-0,302	0,996	1,000				
geno16	0,236	0,164	-0,234	0,022	-0,022	0,023	-0,237	0,212	0,030	0,027	0,387	-0,607	0,469	0,470	1,000			
geno17	0,427	0,166	-0,036	-0,406	0,406	-0,407	-0,426	0,028	0,416	-0,462	-0,315	0,238	0,449	0,448	-0,401	1,000		
geno18	-0,662	-0,356	0,415	0,251	-0,251	0,252	0,656	-0,398	-0,298	0,293	-0,333	0,295	-0,957	-0,957	-0,448	-0,435	1,000	
geno19	-0,662	-0,356	0,415	0,251	-0,251	0,251	0,656	-0,399	-0,298	0,293	-0,333	0,292	-0,957	-0,957	-0,446	-0,437	0,997	1,000

Kuva 3: SNP-kohtien välinen pareittainen korrelaatio.

Tutkimuksessa tarkastellaan eteenpäin valitsevan menetelmän tuloksia sekä painoindeksille että systoliselle ja diastoliselle verenpaineelle koko aineistossa sekä erikseen tupakoivilla että tupakoimattomilla käyttäen Akaiken informaatiokriteeriä. Jokaiselle vasteelle muodostettiin jokaisessa eri populaatiossa kolme erilaista kokonaisuutta (I, II, III), joissa eteenpäin valitsevaa menetelmää käytettiin. Jokaiseen kokonaisuuteen (I, II, III) otettiin sukupuoli-muuttuja suoraan selittäjäksi eli lähtömalliin, sillä vasteissa on eroa sukupuolen mukaan. Ensimmäiseen kokonaisuuteen (I) otettiin tarkasteluun vain SNP-kohdat, joista eteenpäin valitseva menetelmä sai valita tarvittavat selittävät muuttujat malliin. Toisessa kokonaisuudessa (II) tarkasteltiin edelleen vain SNP-kohtia, mutta lähtömalliin lisättiin kolme ensimmäistä pääkomponenttitekijää sukupuolen lisäksi. Näitä kolmea ensimmäistä pääkomponenttitekijää Kaakinen ym. (2012) käyttivät myös omassa tutkimuksessaan. Kolmannessa kokonaisuudessa (III) tarkasteltiin sekä SNP-kohtia että kaikkia kymmentä pääkomponenttitekijää. Tällöin lähtömallissa oli vain sukupuoli kuten ensimmäisessä kokonaisuudessa (I).

Eteenpäin valitsevan menetelmän tuottamista malleista valitaan jokaiselle vasteelle ”paras” malli, jota arvioidaan myös piste-estimaattien, P-arvojen ja luottamusvälien avulla. ”Parhaimmaksi” malliksi valitaan se malli, jolla Akaiken informaatiokriteerin arvo on pienin.

Eteenpäin valitsevan menetelmän lisäksi raportoidaan myös parhaan osajoukon algoritmin antamia tuloksia. Parhaan osajoukon algoritmi suoritetaan niin, että tarkasteltavaan malliin voidaan valita sukupuoli-muuttuja ja SNP-kohdat. Osajoukkoja arvioidaan BIC-informaatiokriteerillä. Parhaan osajoukon algoritmi voi valita yhdestä kahdeksaan muuttujaa malliin, ja jokaiselle muuttujamäärälle tuotetaan neljä osajoukkoa eli neljä toisistaan poikkeavaa regressiomallia.

Painoindeksi

Assosiaatioiden tarkastelu aloitettiin painoindeksistä koko aineistossa. Jokaisessa kolmessa eri kokonaisuudessa (I, II, III) saatiin samantapaisia tuloksia. Eteenpäin askeltava menetelmä valitsi malliin niin ensimmäisessä (I) kuin toisessakin (II) kokonaisuudessa saman SNP-kohdan, vaikka toisessa kokonaisuudessa (II) lähtömallissa oli myös kolme ensimmäistä pääkomponenttia selittäjänä. (Taulukko 5 ja 6) Ainoastaan AIC-kriteerin arvo on ensimmäisen kokonaisuuden (I) muodostamassa mallissa pienempi. Paras AIC-kriteerin arvo saadaan kolmannen kokonaisuuden (III) mallissa, jossa kolme pääkomponenttitermiä ovat pienentämässä AIC-kriteerin arvoa (Taulukko 7). Kaikissa kolmessa kokonaisuudessa (I, II, III) malliin tuli valituksi sama SNP-kohta: rs1996371.

Taulukko 5: Eteenpäin valitseva menetelmä painoindeksille koko aineistossa ensimmäisessä kokonaisuudessa (I). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		15287.38
2	+ geno14	-3.36

Taulukko 6: Eteenpäin valitseva menetelmä painoindeksille koko aineistossa toisessa kokonaisuudessa (II). Lähtömallissa selittävänä muuttujina sukupuoli, *pc1*, *pc2* ja *pc3*.

	Askel	AIC
1		15287.78
2	+ geno14	-3.48

Seuraavaksi analysoitiin tupakoivien osajoukossa havaittavaa assosiaatiota painoindeksin ja SNP-kohtien välillä. Taulukoista 8, 9 ja 10 huomataan, että tulokset poikkeavat koko aineistosta saatavista tuloksista. Kaikissa kolmessa kokonaisuudessa (I, II, III) valittiin toisiinsa nähden samat SNP-kohdat:

rs6495309, rs1996371 ja rs4887077. Näistä SNP-kohta rs1996371 tuli valituksi myös koko aineistossa. Paras AIC-kriteerin arvo saadaan kolmannessa kokonaisuuden (III) mallissa, jossa on myös yksi pääkomponenttitermi mallissa mukana. Pääkomponenttitermi pienentää AIC-kriteerin arvoa kahteen ensimmäiseen kokonaisuuteen (I, II) verrattuna.

Taulukko 7: Eteenpäin valitseva menetelmä painoindeksille koko aineistossa kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		15287.38
2	+ *pc9*	-6.39
3	+ geno14	-3.19
4	+ *pc2*	-0.90
5	+ *pc3*	-0.54

Taulukko 8: Eteenpäin valitseva menetelmä painoindeksille tupakoivilla ensimmäisessä kokonaisuudessa (I). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		6220.117
2	+ geno9	-6.11
3	+ geno14	-1.16
4	+ geno18	-1.34

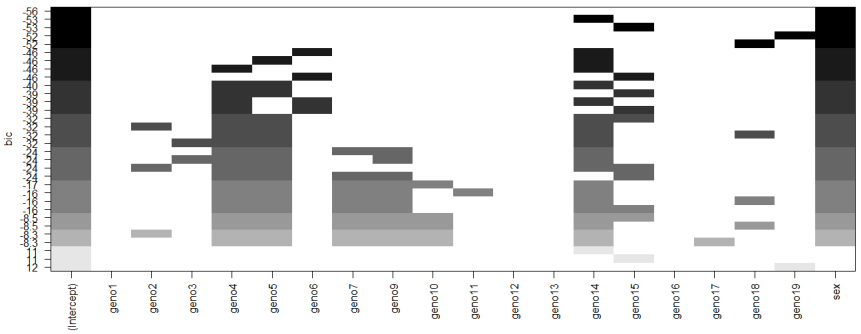
Taulukko 9: Eteenpäin valitseva menetelmä painoindeksille tupakoivilla toisessa kokonaisuudessa (II). Lähtömallissa selittävänä muuttujina sukupuoli, *pc1*, *pc2* ja *pc3*.

	Askel	AIC
1		6224.340
2	+ geno14	-6.36
3	+ geno9	-1.38
4	+ geno18	-1.16

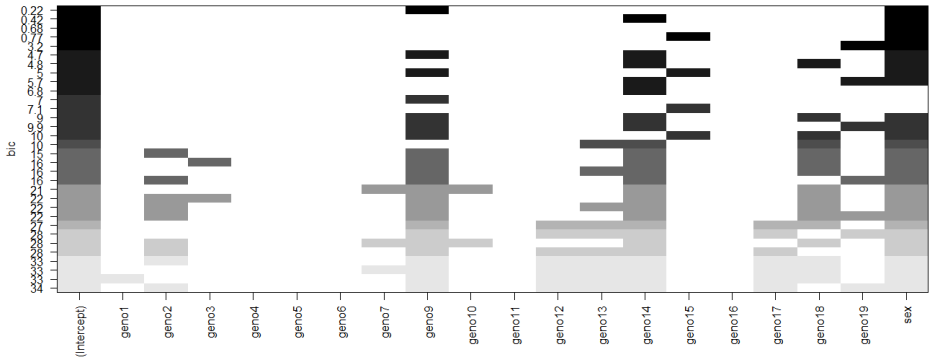
Taulukko 10: Eteenpäin valitseva menetelmä painoindeksille tupakoivilla kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		6220.117
2	+ geno9	-6.11
3	+ $pc7$	-1.99
4	+ geno14	-0.68
5	+ geno18	-1.14

Kuvissa 4 ja 5 on esitetty parhaan osajoukon algoritmin antama tulos BIC-kriteerillä painoindeksille molemmissa aineistoissa. Algoritmi on ajettu pelkästään SNP-kohdille sukupuolien kanssa. Malleissa on yhdestä kahdeksaan muuttujaa, joista jokaisesta on neljä osajoukkoa. Tämä tarkoittaa, että yhden muuttujan malleja on neljä samoin kuin kahden muuttujan malleja ja niin edelleen. Tämän jälkeen mallit on laitettu paremmuusjärjestykseen BIC-kriteerin mukaisesti. Kuvassa ylhäällä on BIC-kriteerin mukaan paras malli ja alimpana huonoin. Koko aineiston tapauksessa parhaaksi malliksi tulee valittua malli, jossa selittävänä muuttujana on sukupuoli. Tupakoivilla parhain malli sisältää sukupuolien lisäksi SNP-kohdan rs6495309.



Kuva 4: Parhaan osajoukon algoritmi koko aineistossa painoindeksille neljällä osajoukolla käyttäen BIC-kriteeriä.



Kuva 5: Parhaan osajoukon algoritmi tupakoivilla painoindeksille neljällä osajoukolla käyttäen BIC-kriteeriä.

Systolinen verenpaine

Seuraavaksi analysoitiin systolisen verenpaineen assosiaatiota SNP-kohtiin koko aineistossa. Systolisen verenpaineen tapauksessa kolmen eri kokonaisuuden (I, II, III) välillä saadut tulokset vaihtelivat paljon, eivätkä olleet yhtä systemaattisia kuin painoindeksille saadut tulokset. Ensimmäisessä kokonaisuudessa (I) eteenpäin valitseva menetelmä valikoi malliin kolme SNP-kohtaa: rs1948, rs6495306 ja rs12900519 (Taulukko 11). Toisessa kokonaisuudessa (II) kolmen ensimmäisen pääkomponenttitermin seuraksi malliin saatiin kaksi SNP-kohtaa: rs12900519 ja rs8038920. (Taulukko 12). Kolmannen kokonaisuuden (III) mallissa saatiin pienin AIC-kriteerin arvo, mutta malliin tuli valituksi kuuden pääkomponenttitermin lisäksi vain yksi SNP-kohta, rs12900519 (Taulukko 13).

Taulukko 11: Eteenpäin valitseva menetelmä systoliselle verenpaineelle koko aineistossa ensimmäisessä kokonaisuudessa (I). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		27444.81
2	+ geno10	-2.13
3	+ geno4	-0.87
4	+ geno13	-0.58

Taulukko 12: Eteenpäin valitseva menetelmä systoliselle verenpaineelle koko aineistossa toisessa kokonaisuudessa (II). Lähtömallissa selittävänä muuttujina sukupuoli, *pc1*, *pc2* ja *pc3*.

	Askel	AIC
1		27428.59
2	+ geno13	-1.31
3	+ geno17	-0.33

Taulukko 13: Eteenpäin valitseva menetelmä systoliselle verenpaineelle koko aineistossa kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		27444.81
2	+ *pc1*	-14.06
3	+ *pc4*	-9.18
4	+ *pc5*	-3.71
5	+ *pc2*	-1.70
6	+ geno13	-1.33
7	+ *pc10*	-0.85
8	+ *pc3*	-0.40

Systolisen verenpaineen assosiaatiota analysoitiin myös vain tupakoivilla. Ensimmäisessä kokonaisuudessa (I) malliin tuli valituksi neljä SNP-kohtaa, jotka ovat rs1948, rs6495306, rs1051730 ja rs8034191 (Taulukko 14). Toisessa kokonaisuudessa (II) kolmen ensimmäisen pääkomponenttitekijän kanssa malliin sisältyy kuusi SNP-kohtaa: rs1948, rs6495306, rs6495309, rs2036534,

rs1051730 ja rs8034191 (Taulukko 15). SNP-kohtien määrä toisessa kokonaisuudessa (II) on muihin verrattuna hyvin suuri. Kolmannessa kokonaisuudessa (III) malliin tuli valituksi neljän pääkomponenttitekijän lisäksi kaksi SNP-kohtaa. Nämä SNP-kohdat ovat rs1948 ja rs6495306 (Taulukko 16). Systemaattisesti jokaiseen kolmeen malliin tulivat valituksi nämä kaksi SNP-kohtaa. Paras AIC-kriteerin arvo saadaan tässäkin kolmannen kokonaisuuden (III) mallissa. Saadut tulokset ovat hyvin erilaiset kuin koko aineistosta saadut tulokset.

Taulukko 14: Eteenpäin valitseva menetelmä systoliselle verenpaineelle tupakoivilla ensimmäisessä kokonaisuudessa (I). Lähtömallissa selittävänä muuttujana vain sukupuoli.

Askel		AIC
1		10691.21
2	+ geno10	-5.79
3	+ geno4	-6.06
4	+ geno7	-0.58
5	+ geno1	-2.52

Taulukko 15: Eteenpäin valitseva menetelmä systoliselle verenpaineelle tupakoivilla toisessa kokonaisuudessa (II). Lähtömallissa selittävänä muuttujana sukupuoli, *pc1*, *pc2* ja *pc3*.

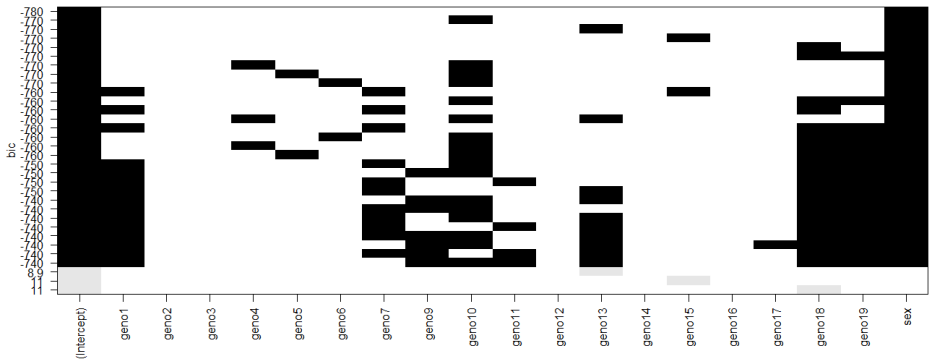
Askel		AIC
1		10686.40
2	+ geno10	-5.15
3	+ geno4	-6.15
4	+ geno9	-0.05
5	+ geno3	-0.08
6	+ geno7	-0.66
7	+ geno1	-0.79

Kuvissa 6 ja 7 on esitetty parhaan osajoukon algoritmin antama tulos sekä koko aineistossa että pelkästään tupakoivien osajoukossa systoliselle verenpaineelle, kun käytetään BIC-informaatiokriteeriä. Tässäkin tapauksessa

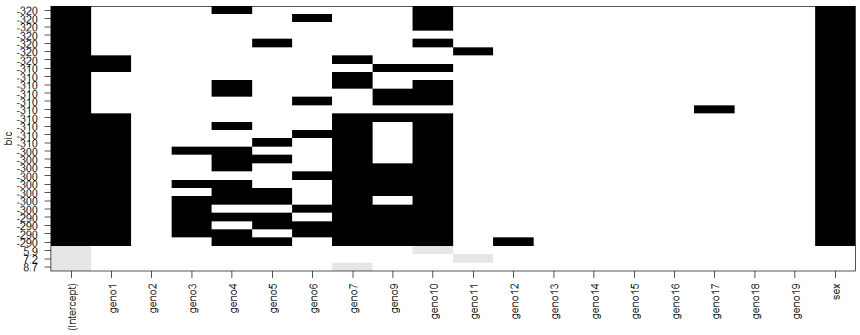
muuttujia voi olla yhdestä kahdeksaan, ja käytössä on neljä osajoukkoa. Koko aineistossa BIC-kriteerin mukaan parhaaksi malliksi tulee vain sukupuolen sisältävä malli. Tupakoivilla parhaaksi malliksi tulee malli, jossa selittäviä muuttujia ovat sukupuolen lisäksi SNP-kohdat rs1948 ja rs6495306. Nämä SNP-kohdat ovat samat kuin mitä eteenpäin valitsevassa menetelmässä esiintyi kaikissa kolmessa kokonaisuudessa (I, II, III), kun tarkasteltiin vain tupakoivia henkilöitä.

Taulukko 16: Eteenpäin valitseva menetelmä systoliselle verenpaineelle tupakoivilla kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana vain sukupuoli.

	Askel	AIC
1		10691.21
2	+ *pc4*	-7.90
3	+ geno10	-5.49
4	+ geno4	-5.88
5	+ *pc2*	-3.61
6	+ *pc10*	-1.91
7	+ *pc3*	-0.56



Kuva 6: Parhaan osajoukon algoritmi koko aineistossa systoliselle verenpaineelle neljällä osajoukolla käyttäen BIC-kriteeriä.



Kuva 7: Parhaan osajoukon algoritmi tupakoivilla systoliselle verenpaineelle neljällä osajoukolla käyttäen BIC-kriteeriä.

Diastolinen verenpaine

Analysoitaessa diastolisen verenpaineen assosiaatiota valitun geenialueen eri SNP-kohtien kanssa koko aineistossa malliin ei kahdessa kokonaisuudessa (I, II) tapauksessa tullut valituksi kuin SNP-kohta rs8038920. Kuvassa 17 on esitetty kolmas kokonaisuus (III) diastoliselle verenpaineelle. Tässäkin kokonaisuudessa malliin tulee valituksi kyseinen SNP-kohta, mutta sen lisäksi myös neljä pääkomponenttitermiä. AIC-kriteerin arvo kyseisessä kolmannessa kokonaisuudessa (III) oli pienin muihin tilanteisiin verrattuna. Ensimmäisessä tapauksessa AIC-kriteerin arvo oli 26064.16.

Kun diastolista verenpainetta käsiteltiin pelkästään tupakoivien aineistossa, saatiin kaikissa kolmessa tapauksessa malliin valituksi samat kaksi SNP-kohtaa. Nämä kohdat ovat rs8038920 ja rs6495309. Kuvassa 18 on esitetty tulos, joka saatiin sekä ensimmäisestä että kolmannesta kokonaisuudessa (I, III).

Taulukko 17: Eteenpäin valitseva menetelmä diastoliselle verenpaineelle koko aineistossa kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana sukupuoli.

	Askel	AIC
1		26068.00
2	+ geno17	-3.84
3	+ *pc5*	-3.12
4	+ *pc7*	-2.78
5	+ *pc9*	-0.75
6	+ *pc6*	-0.23

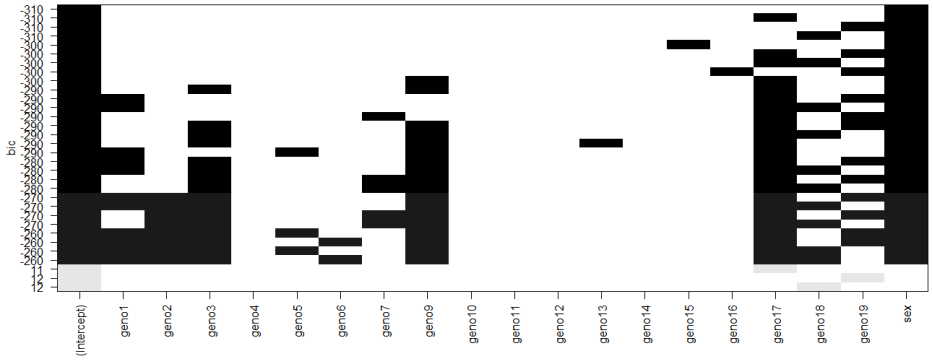
Taulukko 18: Eteenpäin valitseva menetelmä diastoliselle verenpaineelle tupakoivilla ensimmäisessä ja kolmannessa kokonaisuudessa (III). Lähtömallissa selittävänä muuttujana sukupuoli.

	Askel	AIC
1		10178.94
2	+ geno17	-4.90
3	+ geno9	-2.30

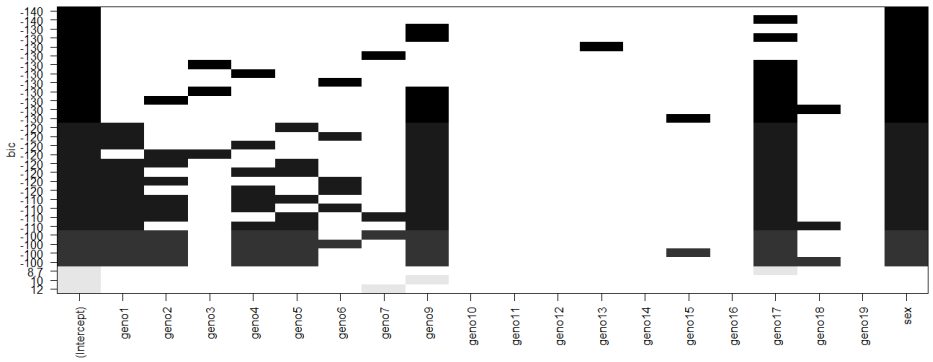
Diastolisen verenpaineen assosiaation tarkastelussa sovellettiin myös parhaan osajoukon algoritmia koko aineistossa ja vain tupakoiville BIC-informaatiokriteerillä. Muuttujia voitiin valita yhdestä kahdeksaan neljällä osajoukolla kuten aikaisemminkin painoindeksille ja systoliselle verenpaineelle. Kuvissa 8 ja 9 on esitetty saadut tulokset. Tulokset eivät ole yhtä johdonmukaisia kuin eteenpäin valitsevassa menetelmässä, sillä parhaan osajoukon algoritmin tapauksessa jokaisen neljästä osajoukosta tulee olla toisistaan poikkeava. Parhaaksi malliksi tulee sekä koko aineiston että vain tupakoivien tapauksessa valittua malli, jossa selittävänä muuttujana on pelkästään sukupuoli.

Kaikille kolmelle vasteelle sovellettiin eteenpäin valitsevaa menetelmää myös silloin, kun populaatio rajattiin vain tupakoimattomiin henkilöihin. Tuloksia saatiin vain systolisen verenpaineen kohdalla, ja silloinkin malliin valittiin pelkkiä pääkomponenttitermejä. Vasteita tarkasteltiin myös taaksepäin

eliminoivalla menetelmällä, mutta tulokset olivat identtisiä eteenpäin valitsevalla menetelmällä saatujen kanssa.



Kuva 8: Parhaan osajoukon algoritmi koko aineistossa diastoliselle verenpaineelle neljällä osajoukolla käyttäen BIC-kriteeriä.



Kuva 9: Parhaan osajoukon algoritmi tupakoivilla diastoliselle verenpaineelle neljällä osajoukolla käyttäen BIC-kriteeriä.

Piste-estimaatit ja luottamusvälit ”parhaimmille” malleille

Jokaiselle vasteelle valittiin sekä koko aineistosta että vain tupakoivilla eteenpäin askeltavalla menetelmällä saatu optimaalinen malli. ”Parhaimmat” mallit on esitetty taulukossa 19.

Taulukko 19: 'Parhaimmiksi' valitut mallit jokaiselle vasteelle eri aineistoissa.

Vaste	Aineisto	Selittävät muuttujat
Painoindeksi	Kaikki	sex *pc2* *pc3* *pc9* geno14
	Tupakoivat	sex *pc7* geno9 geno14 geno18
Syst. verenpaine	Kaikki	sex *pc1* *pc2* *pc3* *pc4* *pc5* *pc10* geno13
	Tupakoivat	sex *pc2* *pc3* *pc4* *pc10* geno4 geno10
Dias. verenpaine	Kaikki	sex *pc5* *pc6* *pc7* *pc9* geno17
	Tupakoivat	sex geno9 geno17

Näille ”parhaimmille” malleille laskettiin regressiokertoimien piste-estimaatit, luottamusvälit ja piste-estimaatin merkitsevyyttä kuvaava p-arvo. Nämä tunnusluvut painoindeksin osalta koko aineistolle on esitetty taulukossa 20 ja vain tupakoiville taulukossa 21. Koko aineistoa koskevista tunnusluvuista voidaan havaita, että mallin ainoa SNP-kohta ei vaikuta painoindeksin arvoon kovinkaan merkittävästi. Tupakoivia kuvaavien tunnuslukujen perusteella näyttäisi siltä, että SNP-kohta rs1996371 vaikuttaa pienentävästi painoindeksiin.

Taulukko 20: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit koko aineistosta taulukon 7 mukaisessa mallissa, jossa vasteena painoindeksi sekä selittävinä muuttujina sex, *pc9*, geno14, *pc2* ja *pc3*.

	Estimaatti (kg/m^2)	Luottamusväli	P-arvo
sex	-1.0	(-1.2, -0.7)	2×10^{-16}
pc9	11.9	(3.8, 19.9)	0.004
geno14	-0.2	(-0.4, -0.03)	0.020
pc2	-7.0	(-15.1, 1.1)	0.089
pc3	-6.5	(-14.6, 1.5)	0.113

Taulukko 21: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit tupakoivilla taulukon 10 mukaisessa mallissa, jossa vasteena painoindeksi sekä selittävinä muuttujina sex, geno9, *pc7*, geno14 ja geno18.

	Estimaatti (kg/m^2)	Luottamusväli	P-arvo
sex	-0.7	(-1.1, -0.3)	0.000164
geno9	-0.3	(-0.6, -0.01)	0.060
pc7	-13.0	(-26.9, 1.0)	0.069
geno14	-1.1	(-2.2, -0.1)	0.030
geno18	-0.9	(-2.0, 0.1)	0.077

Regressiokertoimien piste-estimaatit, luottamusvälit ja piste-estimaattien p-arvot on systoliselle verenpaineelle ”parhaimmaksi” valituissa malleissa esitetty taulukoissa 22 ja 23. Taulukossa 22 ovat tulokset koko aineistolle, ja tulosten perusteella näyttäisi siltä, että ainoa SNP-kohta rs12900519 ei vaikuta systoliseen verenpaineeseen ollenkaan. Kun tarkastellaan pelkkiä tupakoivia taulukossa 23, kahdella SNP-kohdalla, rs1948 ja rs6495306, näyttäisi olevan pienentävä vaikutus systoliseen verenpaineeseen.

Taulukko 22: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit koko aineistossa taulukon 13 mukaisessa mallissa, jossa vasteena systolinen verenpaine ja selittävinä muuttujina sex, *pc1*, *pc4*, *pc5*, *pc2*, geno13, *pc10* ja *pc3*.

	Estimaatti (mmHg)	Luottamusväli	P-arvo
sex	-10.1	(-10.8, -9.4)	2×10^{-16}
pc1	-49.8	(-74.6, -25.0)	0.00008
pc4	-42.2	(-67.0, -17.4)	0.0008
pc5	-30.2	(-55.0, -5.4)	0.017
pc2	24.7	(-0.1, 49.6)	0.051
geno13	0.6	(-0.1, 1.3)	0.080
pc10	21.4	(-3.4, 46.2)	0.091
pc3	-19.6	(-42.4, 3.2)	0.121

Taulukko 23: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit tupakoijilla taulukon 16 mukaisessa mallissa, jossa selittävänä muuttujana systolinen verenpaine ja selittävinä muuttujina sex, *pc4*, geno10, geno4, *pc2*, *pc10* ja *pc3*.

	Estimaatti (mmHg)	Luottamusväli	P-arvo
sex	-10.6	(-11.7, -9.5)	2×10^{-16}
pc4	-66.5	(-109.0, -24.0)	0.002
geno10	-3.1	(-4.7 -1.5)	0.0001
geno4	-2.3	(-3.8, -0.7)	0.004
pc2	48.8	(7.1, 90.5)	0.022
pc10	40.0	(-0.8, 80.8)	0.054
pc3	-32.5	(-72.4, 7.4)	0.110

Taulukossa 24 on esitetty tunnusluvut koko aineistossa diastoliselle verenpaineelle, jolloin SNP-kohta rs8038920 näyttäisi vaikuttavan diastoliseen verenpaineeseen. Taulukossa 25 pelkillä tupakoivilla havaitaan sama vaikutus. Tupakoivilla myös SNP-kohta rs6495309 näyttäisi vaikuttavan diastoliseen verenpaineeseen.

Taulukko 24: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit koko aineistossa taulukon 13 mukaisessa mallissa, jossa vasteena diastolinen verenpaine sekä selittävinä muuttujina sex, geno17, *pc5*, *pc7*, *pc9* ja *pc6*.

	Estimaatti (mmHg)	Luottamusväli	P-arvo
sex	-5.6	(-6.2, -5.0)	2×10^{-16}
geno17	-0.6	(-1.1, -0.1)	0.015
pc5	-25.3	(-47.1, -3.4)	0.024
pc7	-24.5	(-46.3, -2.6)	0.028
pc9	18.5	(-3.4, 40.5)	0.097
pc6	-16.7	(-38.5, 5.2)	0.135

Taulukko 25: Regressiokertoimien piste-estimaatit ja niiden luottamusvälit koko aineistossa taulukon 13 mukaisessa mallissa, jossa vasteena diastolinen verenpaine sekä selittävinä muuttujina sex, geno17 ja geno9.

	Estimaatti (mmHg)	Luottamusväli	P-arvo
sex	-6.2	(-7.2, -5.3)	2×10^{-16}
geno17	-1.0	(-1.8, -0.2)	0.011
geno9	-0.8	(-1.6, -0.1)	0.0368

6 Pohdinta

Jos tarkastellaan tuloksia ilman ennakkokäsitystä eri muuttujanvalintamenetelmistä ja aikaisemmissa tutkimuksissa saaduista tuloksista, voitaisiin tulosten pohjalta tehdä hyvinkin vaihtelevia päätelmiä. Nämä päätelmät voisivat antaa hyvin vääristyneen kuvan vasteiden todellisesta assosiaatiosta tutkittavaan CHRNA3-CHRNA5-CHRNB4 -geenialueeseen. Saatuja tuloksia käsiteltäessä onkin ehdottoman tärkeää huomioida muuttujanvalintamenetelmien vaarat.

Tulosten perusteella näyttäisi, että CHRNA3-CHRNA5-CHRNB4 geenialueella on jonkinlaista vaikutusta tarkasteltaviin vasteisiin. Tulosten perusteella jokainen merkittävä SNP-kohta pienentää vasteen arvoa. Tarvittaisiin lisää havaintoja kertomaan, ovatko tulokset oikeita vai harhaisia. Tulosten luotettavuutta parantaisi myös se, että eteenpäin valitseva menetelmä ja siitä saataville malleille toteutettu piste-estimaattien ja luottamusvälien laskenta olisi toteutettu osapopulaatioissa eli tarkasteltava kohortti olisi jaettu kahteen osaan.

Tulosten avulla ei pystytä täysin sanomaan, vaikuttaako geenialue vasteisiin vain tupakoivilla tai tupakoinnin kautta. Kuitenkaan eri muuttujanvalintamenetelmät eivät löytäneet tupakoimattomien henkilöiden aineistossa mitään todellista assosioivaa tekijää vasteiden kanssa, kun taas tupakoivien aineistossa saatiin malleihin selittäviä muuttujia. Näin ollen voisi olettaa, että koko aineistosta saadut tulokset sovittavat tämän vuoksi tupakoivien ja tupakoimattomien eroja kyseisissä SNP-kohdissa. Kun tupakoimattomat yhdistetään tupakoivien kanssa koko aineistoksi, löytävät eri muuttujanvalintamenetelmät assosiaatiota tarkasteltavien vasteiden ja muuttujien kanssa, vaikka pelkästään tupakoimattomien osajoukossa assosiaatiota ei löytynyt. On mahdollista, että analysoitavat SNP-kohdat ovat assosioituneet tupakoin-

nin kanssa ja tupakointi itsessään vaikuttaa painoindeksiin ja verenpaineeseen. Tämän päätelmän varmistamiseksi tarvittaisiin useampia tutkimuksia myös muussa suomalaisessa populaatiossa.

Kun verrataan tutkimuksessa saatuja tuloksia tupakoivilla Kaakisen ym. (2012) saamiin tuloksiin, voidaan huomata, että askeltavat menetelmät valitsivat painoindeksin tapauksessa samoja SNP-kohtia, joille Kaakinen ym. löysivät assosiaatiota. Kaakisen ym. (2012) tutkimuksessa löydettiin assosiaatiota kuuden SNP-kohdan ja painoindeksin välillä. Näistä kolme rs6495309, rs1996371 ja rs4887077 tuli valituksi myös eteenpäin askeltavalla menetelmällä malliin. SNP-kohdalle rs1996371 havaittiin assosiaatiota myös koko aineistossa. Kaakinen ym. löysivät assosiaation kahden SNP-kohdan, rs1948 ja rs950776, ja systolisen verenpaineen välillä. Näistä SNP-kohta rs1948 tuli valituksi malliin eteenpäin valitsevalla menetelmällä systoliselle verenpaineelle, ja tulosten perusteella tämä SNP-kohta näyttäisikin pienentävän systolista verenpainetta. Kaakisen tutkimuksessa ei löydetty kunnollista assosiaatiota tarkasteltavien SNP-kohtien ja diastolisen verenpaineen välillä, vaikka tässä tutkimuksessa löytyikin kaksi SNP-kohtaa selittäviksi muuttujiksi malliin. Kaakisen ym. (2012) tutkimuksessa tarkasteltiin lähinnä yksittäisten SNP-kohtien vaikutuksia vasteisiin, kun tässä tutkimuksessa tarkasteltiin kaikkien vaikutusta yhdessä vasteisiin. Saadut tulokset eivät siis tämän takia ole täysin vertailukelpoisia. Kuitenkin voidaan varovasti todeta, että SNP-kohdalla rs1948 voisi olla todellista vaikutusta systolisen verenpaineen arvoon tupakoivilla samoin kuin SNP-kohdilla rs6495309, rs1996371 ja rs4887077 painoindeksiin, sillä nämä kyseiset SNP-kohdat havaittiin molemmissa tutkimuksessa.

Malleihin sisältyi paljon muuttujia, jotka ovat mallissa vain parantaakseen informaatiokriteerin arvoa. Tällaisina voidaan pitää malleihin valittuja pääkomponenttitermejä, koska kyseisillä termeillä itsessään ei ole suurta vaikutus-

ta tarkasteltuihin vasteisiin. Käytettyjen pääkomponenttien arvot ovat hyvin pieniä ja ne eivät poikkea toisistaan paljon, jolloin eteenpäin valitsevan menetelmän on helppo valita tällainen muuttuja malliin selittämään jotain vastetta. Mahdollista on, että valitut pääkomponenttitermit ilmaisevat muilta geenialueilta löydettävää assosiaatiota vasteisiin, sillä pääkomponenttitermit sovittavat tarkasteltavan henkilön koko genomia. Tämä on kuitenkin epätodennäköistä askeltavien menetelmien luoteen vuoksi. Askeltavat menetelmät pyrkivät vain mahdollisimman pieneen informaatiokriteerin arvoon. Muuttujanvalintamenetelmät ja erityisesti askeltavat menetelmät eivät ota huomioon muuttujien välisiä keskinäisiä suhteita ja niiden yhteisvaikutuksia. Tämän vuoksi malleista voi tulla todellisuudesta poikkeavia ja mallit sisältävät paljon turhia muuttujia. Se, että eteenpäin askeltava menetelmä valitsi muutaman jo aikaisemmassa tutkimuksessa löydetyn SNP-kohdan malliin, voi olla joko sattumaa tai oikeanlainen löytö. Ei siis voida sanoa, että askeltavat menetelmät löytäisivät merkittävimpiäkään assosiaatioita.

On suositeltavaa, että askeltavia menetelmiä käytettäisiin tutkimuksessa varoen, sillä niiden antamiin tuloksiin sellaisinaan ei ole syytä luottaa. Ennustemallien tarkastelussa askeltavat menetelmät voivat olla hyödyllisiä, vaikka ne seuraavat vain informaatiokriteerin parantumista yhdellä muuttujalla kerrallaan. Askeltavat menetelmät valitsevat malliin vain yhden, joskus hyvinkin sattumanvaraisen, muuttujan, joka pienentää informaatiokriteerin arvoa. Ainoa tapa, millä askeltavista menetelmistä saataisiin käyttökelposia olisikin se, että menetelmä tarkastelisi myös muuttujien yhteisvaikutusta aina kun se on lisäämässä tai poistamassa jotakin muuttujaa. Ei voida olettaa, että ainakaan tällaisessa SNP-kohtia analysoitavassa tutkimuksessa käsiteltävillä muuttujilla ei olisi yhteisvaikutusta. Askeltavien menetelmien kyky havaita todellisuutta vastaavia assosiaatioita on riittämätön. Menetelmät sisältävät liikaa epävarmuutta valinnoissaan ja ne tuijottavat liikaa yksittäisen kriteerin arvoa. Vaikka SNP-kohtien mahdollista assosiaatiota vasteisiin olisikin

tarkasteltu jollain toisella tilastollisella menetelmällä, varmuutta tulosten ja käytettyjen menetelmien hyvyydestä ei voida taata. Sen vuoksi askeltavia menetelmiä kohtaan astettava kritiikki kohdistuu lähinnä muiden tutkijoiden kuten Hocking (1976) päätelmiin askeltavien menetelmien heikkoudesta.

Tässä tutkimuksessa hyödynnettiin myös parhaan osajoukon algoritmiä. Tätä menetelmää käytetään yleensä tarkasteltavien muuttujien määrän vähentämisen työkaluna eikä itsessään tuloksien löytämiseen. Parhaan osajoukon algoritmi valitsi tässä tutkimuksessa samoja muuttujia kuin eteenpäin askeltava menetelmä jokaiselle vasteelle. Parhaan osajoukon algoritmi voi siis toimia muuttujien vähentämisessä, mutta sen käyttö on merkittävää silloin kun käsittelyssä on useita kymmeniä muuttujia. Tällaisessa tutkimuksessa, jossa vasteisiin vaikuttavia muuttujia on todella vähän, on realistisempaa tutkia muuttujia itse kuin parhaan osajoukon algoritmin avulla. Jos haluaa vähentää tarkasteltavien muuttujien määrää, kannattaa se tehdä itse kokeilemalla ja testaamalla tai käyttämällä hyväksi aikaisempaa tietopohjaa.

Nykyisistä tutkimuksista, esimerkiksi geenien assosiaation analysoimisesta, on tullut hyvin vaativia, mikä vaatii tutkijalta sekä oikeanlaisia menetelmiä että riittävää tietoa tarkasteltavien kohteiden ominaisuuksista. Olisi hyvä, jos tutkimus pystyttäisiin suorittamaan kahdella täysin erilaisella menetelmällä. Näin tuloksia pystyttäisiin vertaamaan ja havainnoimaan menetelmien väliset ongelmakohdat. Geenien vaikutukset ovat vielä suurilta osin epävarmaa ja geenien assosioitumisen tarkastelu eri vasteiden kanssa on monimutkainen prosessi. Vääränlainen menetelmä ja hätiköity tulkinta voivat pilata merkittävän tutkimuksen, jolla muutoin saataisiin tietoa geenin vaikutuksista.

Lähteet

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716-723

Andersen, P. K. & Skovgaard, L. T. (2010) *Regression with linear predictors*. Springer, New York, 2010.

Balding, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781-791.

Burnham, K. P. & Anderson, D. R. (2004) Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, 32(2):261-304.

Casella, G & Berger, R. L. (1990) *Statistical inference*. Wadsworth & Brooks/ Cole, Pacific Grove, CA.

Chen, W. M & Abecasis, G. R. (2007) Family-based association tests for genomewide association scans. *The American Journal of Human Genetics*, 81(5):913-916.

Christensen, R. (2011) *Plane answers to complex questions: Theory of linear models*. Springer Texts in Statistics, New York.

Cordell, H. J. & Clayton, D. J. (2005) Genetic epidemiology 3: Genetic association studies. *The Lancet*, 366(9491):1121-1131.

Draper, N. R. & Smith, H. (1981) *Applied regression analysis*. John Wiley & Sons, New York.

Ducci, F., Kaakinen, M., Pouta, A., Hartikainen, A. L. ym. (2011) TTC12-ANKK1-DRD2 and CHRNA5-CHRNA3-CHRNB4 influence different pathways lading to smoking behaviour from adolescence to mid-adulthood. *Biolocical Psychiatry*, 69(7):650-660.

Dunn, O. J. & Clark, V. A. (1974) *Applied statistics: Analysis of variance and regression*. John Wiley & Sons, New York.

Freathy, R. M., Kazeem, G. R., Morris, R. W., Johnson, P. C. ym. (2011) Genetic variation at CHRNA5-CHRNA3-CHRNB4 interacts with smoking status to influence body mass index. *International Journal of Epidemiology*, 40(6):1617-1628.

Hocking, R. R. (1976) The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1-49.

Järvelin, M. R., Sovio, U., King, V., Lauren, L. ym. (2004) Early life factors and blood pressure at age 31 years in the 1966 Northern Finland birth cohort. *Hypertension*, 44(6):838-846.

Kaakinen, M., Ducci, F., Sillanpää, M. J., Läärä, E. & Järvelin, M. R. (2012) Associations between variation in CHRNA5-CHRNA3-CHRNB4, body mass index and blood pressure in the Northern Finland birth cohort 1966. *PLoS ONE*, 7(9):e46557.

Kass, R. E. & Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, 90(430):773-795.

Kuha, J. (2004) AIC and BIC: Comparisons of assumptions and performance. *Sosiological Methods & Research*, 33(2):188-229.

Li, Y. F., LaCroix, C. & Freeling, J. (2009) Specific subtypes of nicotinic cholinergic involved in sympathetic and parasympathetic cardiovascular responses. *Neuroscience Letters*, 462(1):20-23.

Loos, R. J. F. & Bouchard, C. (2008) FTO: The first gene contributing to common forms of human obesity. *Obesity Reviews*, 9(3):246-250.

Läärä, E. (2009) Data-analyysin perusmenetelmät. Oulun yliopisto.

Maailman terveysjärjestö (WHO). Global database on body mass index. *http://apps.who.int/bmi*.

Mallows, C. L. (1973) Some comments on C_p. *Technometrics*, 15(4):661-675

Mineur, Y. S., Abizaid, A., Rao, Y., Salas, R. ym. (2011) Nicotine decreases food intake through activation of pome neurons. *Science*, 332(6035):1330-1332.

Moore, C., Wang, Y. & Ramage, A. G. (2011) Nicotine's central cardiovascular actions: Receptor subtypes involved and their possible physiological role in anaestherizaed rats. *Neuroscience Letters*, 462(1):20-23.

Neter, J., Wasserman, W. & Kutner, M. H. (1990) *Applied linear statistical models*. Irwin, Burr Ridge (IL).

Posada, D. & Buckley, T. R. (2004) Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793-808.

Rahiala, M. (2005) Lineaariset mallit. Oulun yliopisto.

Rana, B. K., Wessel, J., Mahboubi, V., Rao, F. ym. (2009) Natural variation

within the neuronal nicotinic acetylcholine receptor cluster on human chromosome 15q24: Influence on heritable autonomic traits in twin pairs. *The Journal of Pharmacology and Experimental Therapeutics*, 331(2):419-428.

Rantakallio, P. (1969) Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatrica Scandinavica*, 193:1-71.

Rao, C. R. (1973) *Linear statistical inference and its applications*. John Wiley & Sons, New York.

Rawlings, J. O. (1998) *Applied regression analysis: a research tool*. Springer, New York.

Sabatti, C., Service, S. K., Hartikainen, A. L., Pouta, A. ym. (2009) Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, 41(1):35-46.

Schatzoff, M., Tsao, R., & Fienberg, S. (1968) Efficient calculations of all possible regressions. *Technometrics*, 10(4):768-779.

Schwarz, G. E. (1978) Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464.

Seber, G. A. F. (1977) *Linear regression analysis*. John Wiley & Sons, New York.

Suomen Sydänliitto ry. (2012) Verenpaine. *http://www.sydanliitto.fi/verenpaine*.

Terveyden ja hyvinvoinnin laitos. Tilasto- ja indikaattoripankki SOTKANET 2005-2012. *http://uusi.sotkanet.fi/taulukko/o12/91,101,111,112/3/3A/0/*.

Thorgeirsson, T. E., Geller, F., Sulem, P., Rafnar, T. ym. (2008) A vari-

ant associated with nicotine dependence, lung cancer and peripheral aertial disease. *Nature*, 452(7187):638-642.

Vartiainen, E., Jousilahti, P., Alfthan, G., Sundvall, J. ym. (2000) Cardiovascular risk factor changes in Finland, 1972-1997. *Acta Paediatrica Scandinavica*, 29(1):49-56.

Vittinghoff, E., Glidden, D. J., Shiboski, S. C. & McCulloch, C. E. (2005) *Regression methods on biostatistics: Linear, logistic, survival and repeated measures models*. Springer, New York.

Yan, X. (2009) *Linear regression analysis: Theory and computing*. World Scientific, Singapore.

Zhao, H., Pfeiffer, R. & Gail, M. H. (2003). Haplotype analysis in population genetics and association studies. *Pharmagogenomics*, 4(2):171-178.

Ziv, E. & Gonzalez Burchard, E. (2003) Human population structure and genetic association studies. *Pharmacogenomics*, 4(4): 431-441.