TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

Matemaatilise statistika instituut

Matemaatilise statistika eriala

Põhjuslikud mudelid Tartu Ülikooli Eesti Geenivaramu metaboloomika ja toitumise andmetel

Magistritöö (30 EAP)

Koostaja: Kristi Helekivi

Juhendaja: Krista Fischer, PhD

Tartu 2015

Põhjuslikud mudelid Tartu Ülikooli Eesti Geenivaramu metaboloomika ja toitumise andmetel

Käesolevas magistritöös uuritakse põhjuslikke seoseid kohvi tarbimise, metaboliitide kontsentratsiooni taseme ja vererõhu vahel, kasutades Mendeli randomiseerimise meetodit. Riskitegurite põhjusliku mõju hindamisel huvipakkuvale haigusele või tervisenäitajale, on nimetatud meetodi korral instrumenttunnusena kasutusel geneetilised markerid, mille mõju eksponenttunnusele on teada. Töös on otsitavatele parameetritele hinnangud leitud nii Mendeli randomiseerimist kui ka lineaarset regressioonanalüüsi kasutades. Ilmnes, et lineaarne regressioonanalüüs annab küll statistiliselt olulised seosed, ent Mendeli randomiseerimisel saadud tulemuste põhjal ei ole võimalik kinnitada seose põhjuslikkust. Lisaks pakutakse töös välja Mendeli randomiseerimise meetodi edasiarendus juhule, kus põhjuslik seoseahel on keerukam. Simulatsioonieksperiment kinnitab, et meetodi edasiarendus annab eelduste kehtimise korral soovitud tulemused. Reaalsete andmete korral osutusid instrumenttunnused aga liiga nõrgaks, et soovitud täpsusega tulemusi saada.

Märksõnad: *randomiseerimine*, *regressioonanalüüs*, *geneetiline muutlikkus*, *põhjuslikkus*, *geenid*, *ühenukleotiidsed polümorfismid*

Causal Models based on the Estonian Genome Center metabolomics and nutrition data

In this master's thesis causal relationships between coffee consumption, the level of concentration of metabolites and blood pressure are examined, using Mendelian randomization method. In order to assess the causal influence of risk factors to disease of interest or health indicator, genetics variants as proxies for exposures are used. In this thesis, estimates for parameters are found using both Mendelian randomization and linear regression analysis. It appeared that linear regression analysis gives statistically significant linkages but based on the results obtained with Mendelian randomization it is not possible to say that the relationships are causal. In addition, a further development of the Mendelian randomization method is suggested in which causal relationships between the chains are more complex. Simulation experiments confirmed that proposed method gives expected results when assumptions are fulfilled. With real data, however, used instrumental variables proved to be too weak to provide accurate results.

Keywords: *randomization*, *regression analysis*, *genetic variation*, *causality*, *genes*, *Single Nucleotide Polymorphisms*

Sisukord

Sissejuhatus

Haiguste teket, levikut ja tõrjet uurivas arstiteaduse harus – epidemioloogias – on pikka aega püütud teha kindlaks kas inimeste eluviisidest või keskkonnast tulenevad riskitegurid on põhjuslikeks faktoriteks rasketele haigustele. Peamine raskus seisneb selles, et kahe tunnuse vahel olev seos on harva üheselt tõlgendatav – statistiline seos ei pruugi tähendada põhjuslikkust.

Olukorras, kus riskiteguriga on seotud mõni tunnus, mille otsene mõju uuritavale haigusele või tervisenäitajale on välistatud, saab kasutada nn instrumenttunnustel põhinevaid hinnanguid põhjuslikele mõjudele. Viimasel ajal on populaarseks muutunud Mendeli randomiseerimine – geneetiliste muutujate kasutamine instrumentidena, et hinnata riskitegurite põhjuslikku mõju huvipakkuvale haigusele või tervisenäitajale.

Käesolevas töös on meetodit lähemalt uuritud ning kasutades Mendeli randomiseerimist on Tartu Ülikooli Eesti Geenivaramu andmete põhjal uuritud kuidas mõjutab kohvi joomine metaboliitide taset inimese kehas, kuidas mõjutavad metaboliidid vererõhku ning kas esineb põhjuslik mõju kohvi joomise, metaboliitide taseme ja vererõhu vahel.

Töö esimeses ja teises peatükis kirjeldatakse teoreetilist osa: mis on Mendeli randomiseerimine ning milles seisneb statistilise seose ja põhjusliku mõju erinevus. Teises peatükis näidatakse ka meetodi laiendamise võimalusi ning katsetatakse metoodikat simuleeritud andmestikul. Töö kolmandas peatükis antakse ülevaade kasutatavatest andmetest ning kirjeldatakse, miks on töösse valitud just sellised tunnused. Neljandas peatükis rakendatakse metoodikat Tartu Ülikooli Eesti Geenivaramu andmetele ning tõlgendatakse tulemusi.

Magistritöö kirjutamiseks on kasutatud tekstitöötlusprogrammi Microsoft Word 2010. Analüüsid on läbi viidud statistikapaketiga R. Joonised on tehtud programmidega Microsoft Excel 2010 ja Adobe InDesign CS5. Viited kasutatud allikatele on nurksulgudes.

Autor tänab käesoleva magistritöö juhendajat, Tartu Ülikooli Eesti Geenivaramu vanemteadurit Krista Fischerit huvitava probleemipüstituse ning rohkete nõuannete eest. Magistritöö ajal välismaal tudeerimist toetas riiklik Kristjan Jaagu stipendiumiprogramm, mida viib ellu Sihtasutus Archimedes koostöös Haridus- ja Teadusministeeriumiga.

1. Põhjuslik mõju

1.1. Statistiline seos *vs* põhjuslik mõju (epidemioloogias)

Uurides korraga mitut näitajat, on kaks enim huvipakkuvat küsimust tavaliselt kas tunnused on omavahel seotud (näiteks, kas ilmneb statistiliselt oluline seos kõrge kehamassiindeksi ja teist tüüpi diabeedi vahel) ning kas või kuidas üks tunnus mõjutab teist (näiteks, kas kõrge kehamassiindeksiga inimestel on suurem tõenäosus haigestuda teist tüüpi diabeeti). Oluline on teadvustada, et need kaks küsimust ei ole samaväärsed.

Olgu meil kaks huvipakkuvat näitajat – ekspositsioon X (eluviisidest või keskkonnast tulenev riskitegur) ning väljundnäitaja Y (näiteks haigestumine, suremus või mõni tervisenäitaja nagu vererõhk). Küsimused „Kas X ja Y vahel on seos?“ ning „Kas tunnus X mõjutab tunnust Y?“ ei ole samaväärsed.

1.1.1. Kas seos X ja Y vahel on deterministlik, statistiline või põhjuslik?

Deterministlikuks seoseks nimetatakse olukorda, kus ühtede muutujate, katsetingimuste või tunnuste väärtuste muutmisel muutub ka meid huvitava näitaja väärtus. Meid huvitava tunnuse väärtuse saab üheselt leida arvutusvalemi abil, juhuslikkust pole. [1]

Näiteks soovides teisendada tolle sentimeetriteks, peame teadma mõõtühikute omavahelist vahekorda. Teades, et 1 toll vastab 2,54 sentimeetrile, saame arvutusvalemi

$$y = 2{,}54 \cdot x,$$

kus y on otsitav pikkus sentimeetrites ja x pikkus tollides.

Statistiliseks seoseks (ingl *association*) nimetatakse olukorda, kus ühtede muutujate, katsetingimuste või tunnuste väärtuste muutumisel muutub ka meid huvitava näitaja jaotus. Katsetingimuste teadmine ei pruugi meile veel täpselt öelda, milline tuleb katsetulemus, aga teatud katsetingimuste juures on mõned katsetulemused

tõenäolisemad kui teiste katsetingimuste korral. Statistiline seos on sümmeetriline – kui eksisteerib seos tunnuste X ja Y vahel, siis eksisteerib ka seos tunnuste Y ja X vahel. [1]

Näiteks koheselt pärast munemist inkubaatorisse paigutatud linnumunast koorub linnupoeg tõenäosusega 0,9. Kui aga paigutada muna inkubaatorisse 8 päeva pärast munemist, koorub sealt välja tibu vaid tõenäosusega 0,75 – seega on tunnuste „ooteperioodi pikkus“ ja „koorumisedukus“ (koorub/ ei kooru) vahel statistiline seos. [1]

Põhjuslikkuseks (ingl *causation*) nimetatakse nähtustevahelist seost, kus üks nähtustest (nimetatakse põhjuseks) tingib teise nähtuse toimumise (nimetatakse tagajärjeks). [2]

Põhjuslikke mõjusid hinnates peame esmalt tegema selgeks millise mõju hindamine on tegelikult huvi pakkuv. Näiteks, kas meid huvitab kõrge kehamassiindeksiga inimeste diabeeti haigestumise sagedus või diabeedi kui haiguse mõju inimeste kehamassiindeksile. Samuti tuleb otsustada kas ja milliseid eeldusi oleme nõus tegema konkreetse uuringu jooksul, et huvipakkuvad mõjud oleks hinnatavad. Hinnatud põhjuslike mõjude tõlgendus ja hinnangute valiidsus ehk kehtivus sõltuvad tehtud eeldustest ning nende paikapidavusest. [3]

1.2. Statistiline seos *vs* põhjuslik mõju ja randomiseerimine

Pöördume tagasi eelmises alapeatükis mainitud näitajate – ekspositsiooni X ja väljundnäitaja Y – juurde.

Olgu mõlemad vaadeldavad tunnused binaarsed, st väärtused 0 ja 1 vastavad ekspositsiooni/ tervisenäitaja puudumisele või olemasolule. Seos tunnuste X ja Y vahel on olemas, kui $P(Y = 1|X = 1) \neq P(Y = 1|X = 0)$. Seda on lihtne testida, kuid kahjuks ei näita kirjutatud tinglike tõenäosuste mittevõrratus põhjusliku seose suunda. [3]

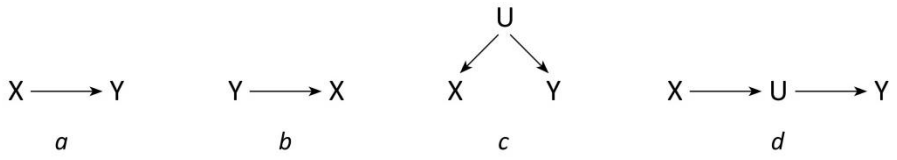
Seosele tunnuste X ja Y vahel on mitmeid võimalikke selgitusi:

a) tunnus X mõjutab põhjuslikult tunnust Y (Joonis 1.a);

b) tunnus Y mõjutab põhjuslikult tunnust X (Joonis 1.b);

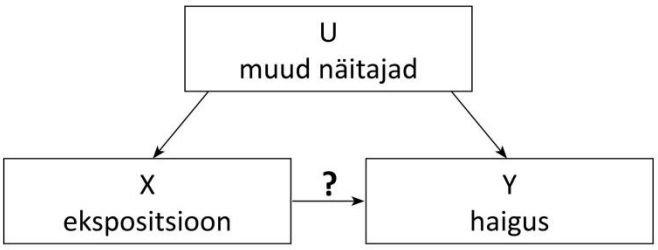
c) eksisteerib kolmas faktor U, mis mõjutab nii tunnuseid X kui ka Y (ehk tunnustel X ja Y on vähemalt üks ühine põhjuslik tegur – ingl *common cause*) (Joonis 1.c);

d) seos, kus ekspositsioon mõjutab väljundnäitajat läbi kolmandate faktorite. Selline on näiteks juht, kus X on geen, U inimese pikkus ja Y kopsude võimekus (Joonis 1.d). [1]



Joonis 1. Neli võimalikku seost tunnuste X ja Y vahel

Tõenäosused $P(Y = 1|X = 1)$ ja $P(Y = 1|X = 0)$ iseloomustavad haiguse (näitaja Y) esinemissagedust kahes rahvastikurühmas. Neid rühmi eristab ekspositsioon X (näiteks suitsetamine: jah/ ei), kuid nad võivad erineda veel paljude muude näitajate poolest. Segavad faktorid mõjutavad tavaliselt korraga nii ekspositsiooni kui ka väljundnäitajat, ent sageli ei ole nende olemasolu võimalik mõõta ega kontrollida, sest me ei mõista seda täielikult. Lihtne lineaarne regressioonanalüüs annaks meile, segajaid arvestamata, nihkega hinnangu otsitavale parameetrile või hoopis väära seose suuna (Joonis 2). [3]



Joonis 2. Nihkega hinnang ekspositsioontunnuse ja väljundnäitaja vahel, segajaid teadmata

Kujutame ette hüpoteetilist olukorda, kus on võimalik muuta tunnuse X väärtust vastavalt soovile kas 1-ks või 0-ks kõigil indiviididel.

$P(Y = 1|do(X) = 1)$ kirjeldab tõenäosust, et $Y = 1$ kui tunnuse X väärtus on kogu üldkogumis seatud olema 1. Samamoodi defineerime $P(Y = 1|do(X) = 0)$.

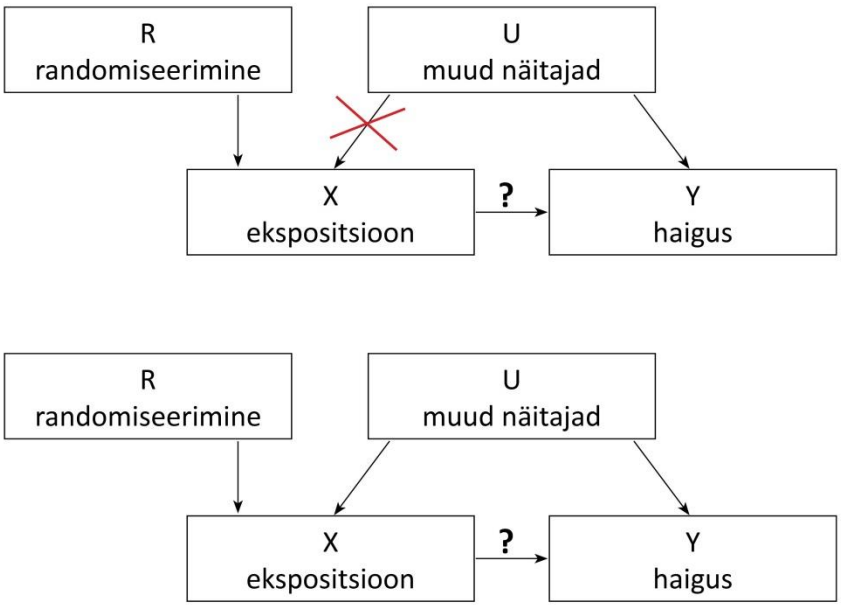
Üldkogumi keskmine põhjuslik mõju (ingl ACE – *Average Causal Effect*) on nüüd defineeritav kui $ACE = P(Y = 1|do(X) = 1) - P(Y = 1|do(X) = 0)$. Alternatiivselt saame defineerida nn potentsiaalsed tunnused $Y(1) = Y(do(X) = 1)$ ja $Y(0) = Y(do(X) = 0)$. Siis $ACE = E[Y(1) - Y(0)]$.

Vahel on mõttekam rääkida ekspositsiooni mõjust neile, kes tegelikult eksponeeritud on: meid ei huvita niivõrd see, kui palju suureneks praeguse mittesuitsetaja haiguserisk, kui ta suitsetaks, vaid huvitab, kui palju väheneks suitsetaja haiguserisk, kui ta ei suitsetaks.

Keskmine ekspositsiooni mõju eksponeeritutel (ingl ATT – *Average Treatment effect in the Treated* või AEE – *Average Exposure effect in the Exposed*) on, kasutades eelnevat tähistust, välja kirjutatav kujul $ATT = P(Y = 1|X = 1) - P(Y = 1|X = 1, do(X) = 0)$ ehk $ATT = E(Y - Y(0)|X = 1)$.

Üldkogumi keskmine põhjuslik mõju on teoreetiliselt hinnatav, kui X on randomiseeritav – juhuvalim üldkogumist jagatakse juhuslikkuse alusel kaheks osaks, ühele osale määratakse $X = 1$, teisele $X = 0$. Sellises uuringus $P(Y = 1|do(X) = x) = P(Y = 1|X = x)$. Kahjuks ei ole selline uuring enamasti praktiliselt teostatav. [3]

Ideaalse uuringu korral on ekspositsioon X ja väljundnäitaja Y seotud vaid läbi ühe tee – läbi otsitava seose (Joonis 3 ülal). Hinnangu otsitavale parameetrile saame sel juhul ka lihtsa lineaarse regressioonanalüüsi abil. Tegelikkuses esineb aga peaaegu alati hulk mittemõõdetavaid või raskestimõõdetavaid tunnuseid, mis mõjutavad nii ekspositsiooni kui ka väljundnäitajat ning mida arvestamata saame väära või nihkega hinnangu otsitavale parameetrile (seosele tunnuste X ja Y vahel). (Joonis 3 all).



Joonis 3. Põhjusliku mõju uurimine ideaalses randomiseeritud uuringus (ülal) ning reaalses randomiseeritud uuringus (all)

Kasutades randomiseerimist, saame uuritavate rühma juhuslikkuse alusel jagada kaheks. Näiteks ravimkatsetes saame eristada kahte gruppi, andes pooltele ravimit (katsegrupp) ja pooltele mitte (kontrollgrupp). Nii saame olla veendunud, et ekspositsioontunnus on kontrolli all ehk ravimi saamine ei ole seotud segavate tunnustega, mis võetava ravimi kogust ja haigust mõjutada võivad (näiteks kui inimene tunneb ennast haigena, võtab ta tõenäoliselt korralikumalt rohtu, kui inimene, kes ennast haigena ei tunne).

Ideaalse randomiseeritud uuringu korral, kus katsealused võtaksid ravimit nii nagu määratud, toimiks korrektse tulemuse saamiseks lihtne lineaarne regressioonanalüüs. Kui mõned kontrollgrupi isikud ikkagi võtavad rohtu ja mõned, kellele ravim on määratud, ei võta ravimit korralikult, saame kahte gruppi võrdlevat t-testi kasutades siiski korrektselt testida ravimi mõju olemasolu, sest ravi määramine saab tulemust mõjutada vaid ravimi efekti kaudu.

Samas lineaarne regressioonanalüüs, mis kasutab tegelikult võetud ravimikogust argumenttunnusena, võib anda nihkega hinnangu ravimi põhjusliku mõju parameetrile. Randomiseerimine mõjutab väga tugevalt seda, kas inimene võtab rohtu või mitte ja kui palju ta rohu võtab. Tänu sellele lähenemine instrumenttunnuste kaudu toimib.

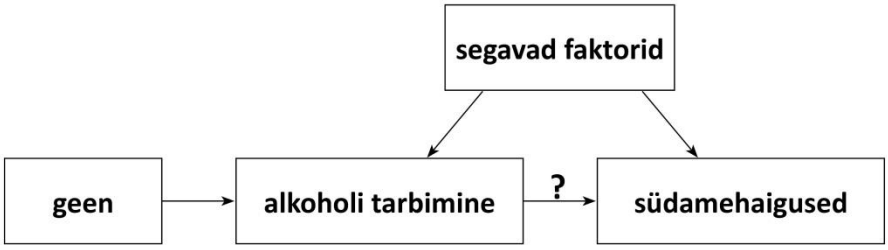
1.3. Randomiseerimine looduse poolt – Mendeli randomiseerimine

Kõik epidemioloogias huvipakkuvad ekspositsioonid, mis põhjustavad kroonilisi haigusi, ei saa olla randomiseeritud. Mõnikord teeb aga randomiseerimise meie eest ära loodus.

Kui randomiseeritud uuringut korraldada ei saa, võib abi olla geenidest. Eriti, kui on teada geneetilised markerid, mille kohta on teada, et nad mõjutavad otseselt vaid ekspositsiooni X, aga mitte väljundnäitajat Y. Väljundnäitajat tohivad need geenimarkerid mõjutada ainult läbi ekspositsiooni, mitte ühtegi teist teed pidi – see on oluline eeldus, mida kahjuks ei ole võimalik statistiliselt testida. [3]

Ühed enamlevinumad geneetilised markerid, mida uuringutes randomiseerijatena kasutatakse, on üksiku nukleotiidi polümorfismid ehk üksiknukleotiidsed polümorfismid (edaspidi SNP – S*ingle Nucleotide Polymorphisms*). SNP-d on DNA järjestuse variatsioonid, mis on toimunud ühe genoomi nukleotiidi (A, T, C või G) muutumisel. Variatsioonid DNA järjestuses võivad määrata, kuidas arenevad inimestel haigused, kuidas toimub reageerimine patogeenidele ja kemikaalidele aga ka ravimitele või vaktsiinidele, millised võivad olla kõrvaltoimed ning mõjutavad toiduainete tarbimist. [4]

Näiteks on varasematest uuringutest teada, et teatud geneetiline mutatsioon takistab alkoholi lagundamist seedesüsteemi poolt. Teadlased on üsna veendunud, et seesama mutatsioon ei mõjuta otseselt südamehaigustesse haigestumist. Sisuliselt käitub see geen just samamoodi nagu randomiseerimine – muud süstemaatilist erinevust selle geeniga ja ilma selleta inimeste hulgas, peale nende erineva alkoholitarbimise taseme, ei ole (Joonis 4). [3]



Joonis 4. Randomiseerimine looduse poolt

Uuringut, kus on kogutud sobivaid geeniandmeid, samuti andmeid ekspositsiooni ja väljundtunnuse kohta, saab analüüsida sarnaselt tavalisele randomiseeritud uuringule. Selle erinevusega, et enamasti randomiseeritakse indiviidid kahte gruppi, kuid ühel geenimarkeril on kolm võimalikku väärtust ning mitme geneetilise markeri põhjal koostatud riskiskoor on vaadeldav pideva tunnusena.

2. Ülevaade Mendeli randomiseerimisest

2.1. Mendeli randomiseerimise põhimõte

Epidemioloogias on pikka aega püütud teha kindlaks kas ekspositsioontegurid on põhjuslikeks faktoriteks rasketele haigustele. Peamine raskus seisneb selles, et kahe tunnuse vahel olev seos on harva ühesuunaline. Uuritavaid tunnuseid mõjutavad lisaks ka teised segavad tunnused, vastupidine põhjuslikkus ning mõõtmisest või küsitlusest tekkinud nihe. Käesolevas töös selgitame, kas ja kuidas on võimalik hinnata põhjuslikke mõjusid, kasutades geneetilisi muutujaid instrumentidena – IV-muutujatena (ingl IV – *Instrumental Variable*). [5]

Mendeli randomiseerimine on põhjusliku analüüsi meetod, et selgitada riskitegurite põhjuslikku mõju huvipakkuvatele haigustele või tervisenäitajatele, kasutades instrumendina geenimarkereid. [5]

Meetodi idee seisneb geneetiliste muutujate (näiteks SNP-markerite) kasutamises, mille kohta on eelnevalt teada, et nad mõjutavad ekspositsiooninäitajat X. Geneetiliste muutujate kasutamise põhjenduseks on nende alleelide juhuslik pärandumine järglastele, mistõttu saab neid pidada sõltumatuteks teguriteks. Meetodi nimetus – Mendeli randomiseerimine – tulenebki sellest, et geenialleelid jagunevad juhuslikult juba meioosi ehk taandjagunemise käigus (ühe genotüübi ühest alleelist pärandub vaid üks juhuslikult järglasele, seda nii ema kui ka isa poolt). [6]

Mendeli randomiseerimise ja IV-muutujate kasutamist soodustab teadmine, et lisaks ekspositsioonile X mõjutab nii väljundnäitajat Y kui ka ekspositsiooni ennast peaaegu alati hulk mõõtmatuid või raskestimõõdetavaid segajaid. Seega annaks lineaarne regressioonanalüüs nihkega hinnangu meid huvitavale otsitavale parameetrile. Nende võimalike mõõtmatute segajate olemasolu tavaliselt motiveeribki kasutama Mendeli randomiseerimist. [6]

Mendeli analüüs ongi eriti vajalik olukordades, kus eeldame segajate olemasolu, ent nende olemasolu ei ole võimalik mõõta või kontrollida (sest me ei mõista seda täielikult). Kui saaksime olla kindlad, et segajaid ei eksisteeri, oleks IV-muutujate analüüs ebavajalik ning korrektse tulemuse saaks ka tavalise regressioon-analüüsiga. [7]

Meetodil on omad piiranguid, ent edusammud geneetikas aitavad neid ületada ning tõenäoliselt suurendavad meetodi kasulikkust, et avastada haiguste riskifaktoreid. [5]

2.2. Eeldused

Nagu teisedki IV-muutujate analüüsid, tuginevad ka Mendeli randomiseerimist kasutavad uuringud eeldustele. [6]

Et saada nihketa hinnang, kuidas ekspositsioon X mõjutab väljundnäitajat Y, kasutades IV-muutujana geenimarkereid (näiteks SNP-sid), peavad kehtima järgmised eeldused.

1. Geneetiline muutuja on seotud ekspositsiooniga, st joonisel ühendab ekspositsiooni X ja SNP-d nool, mille seose suund on võimalik täpselt kindlaks teha.

2. Geneetiline muutuja on sõltumatu segavatest faktoritest, st joonisel ei ole ühtegi noolt (kummaski suunas), mis ühendaks SNP-sid segavate tunnustega.

3. Geneetilisel muutujal puudub otsene mõju väljundnäitajale. Intuitiivselt lähtudes tähendab see, et kõik otsesed teed ehk nooled graafikul SNP-de juurest väljundnäitajasse Y läbivad ekspositsiooni X. [6]

Valiidsete hinnangute saamiseks peavad nimetatud eeldused olema põhjendatud, arvestades bioloogiast tulenevaid taustateadmisi. Statistiliselt ei ole teist ega kolmandat eeldust võimalik testida, kuna nad sõltuvad segavatest faktoritest, mis on definitsiooni kohaselt mittemõõdetavad.

2.3. Piirangud

Kuigi uuringus võib ilmneda seos kahe muutuja vahel, tuleb tähele panna juba varasemalt mainitut: seos ei tähenda põhjuslikkust. Ainult randomiseeritud kontrolluuringuid (ingl RCTs – *Randomized Controlled Trials*) kasutades on võimalik kontrollida põhjusliku seose olemasolu. Paljude ekspositsioonide puhul, mille kohta on vaatlusuuringutes leitud, et nad on seotud väljundnäitajaga (haigusega), on uuringut RCT-ga testides ilmnenud, et ekspositsioon siiski ei ole põhjuslikuks faktoriks. [5]

Peamised põhjused, miks ilmnevad erinevused vaatlusuuringute ja RCT-uuringute vahel, on segajad (ingl *confounding*), vastandlik ehk mitmepidine põhjuslikkus (ingl *reverse causation*), nihe (ingl *bias*) ja mõõtmisviga (ingl *measurement error*). [5]

2.3.1. Segajad

Segavateks teguriteks loetakse epidemioloogias faktoreid, mis on seotud nii huvipakkuva riskifaktoriga kui ka väljundnäitajaga. Segajaid mitte arvestades saadakse nihkega parameetri hinnang seosele ekspositsiooni ja väljundnäitaja vahel. Epidemioloogilised uuringud on segajatest kergesti mõjutatavad, kuna ekspositsiooni näitajad (eluviisid ja keskkonna riskitegurid) on sageli üksteisega tihedasti korreleeritud. Enamus ekspositsioonitegureid ei avaldu üksinda. Näiteks indiviididel, kellel on E-vitamiini vaegus on tavaliselt ka kõrgem kehamassiindeks, sageli tarbivad nad rohkem alkoholi, suitsetavad rohkem ning on madalamast sotsiaalsest klassist (kui need inimesed, kellel E-vitamiini vaegust ei ole). Lisaks võib neil olla mitmeid sotsiaalmajanduslikke ja käitumuslikke riskitegureid, mis muudavad nad vastuvõtlikumaks südamehaigustele (ingl CHD – *Coronary heart disease*) ja teistele rasketele haigustele. [5]

On olemas meetodeid, mis lubavad arvestada võimalikke segavaid faktoreid, ent kohandades mudelit segajatele, teeme eelduse, et segajad on kõik õigesti mõõdetud ning kõik võimalikud segajad on mudelisse lisatud. See eeldus on tõenäoliselt ebareaalne. [5]

2.3.2. Põhjusliku seose suund

Täiendavaks probleemiks vaatlusuuringute juures on sageli võimetus määrata täheldatud seoste suunda või sündmuste ajalist järjestust. Nähtust, kus haigus mõjutab varemoletatud riskifaktorit ja mitte vastupidi, nimetatakse epidemioloogias vastupidiseks põhjuslikkuseks. [5]

Näiteks on CHD-ga patsientidel leitud suurenenud C-reaktiivse valgu (ingl CRP – *C-reactive protein*) taset, võrreldes kontrollrühmaga. See on tekitanud huvi vaadelda CRP-d kui võimalikku haigustekitajat, kuid seos ei ole selline nagu esmapilgul tundub. CRP põhjusliku rolli ümberhindamisel Mendeli randomiseerimise raamistikus on ilmnenud, et mitte suurenenud CRP ei põhjusta südamehaigusi, vaid CRP tase on tõenäoliselt suurenenud põletikuliste protsesside tulemusel, mis kaasnevad CHD-ga. [5]

2.3.3. Nihe

Subjektiivne aruandlus, küsitleja eelarvamused ja vastaja kallutatus on järgmised selgitused, miks vaatlusandmete juures on leitud seos, ent seda ei toetatud randomiseeritud kontrolluuring. Haigestunud inimesed võivad sageli vastata küsimustele oma eluviiside (ekspositsiooninäitaja X) kohta teisiti kui üldine populatsioon, sest nad võivad olla eriti tundlikud kõige suhtes, mis võis nende haigust põhjustada ning seetõttu võivad nad üle tähtsustada oma ekspositsiooninäitajaid (aruandluse erapoolikus). [5]

Üks näide selle kohta on suurem teadlikkus kodulähedaste elektriliinide mõjust nende lapsevanemate hulgas, kelle laps on haigestunud leukeemiasse, võrreldes lapsevanematega, kelle laps ei ole haigestunud (kontrollrühm). Sarnaselt võivad ka intervjueerijad küsida küsimusi juhtumgrupilt ja kontrollgrupilt erinevalt (intervjueerija kallutatus). Ning inimesed võivad suurema tõenäosusega osaleda uuringus, kui nad usuvad, et teatud kindel tegur on põhjustanud nende haiguse (vastaja kallutatus). [5]

2.3.4. Mõõtmisviga

Vaatlusuuringud ei suuda sageli mõõta ekspositsiooninäitajaid täpselt ning selline mõõtmisviga võib viia valede seosteni ekspositsioontunnuste ja haiguste vahel. „Müra“, mis tekib, mõõtes tunnuseid juhtumgruppides ja kontrollgruppides, võib tõenäoliselt viia nõrgema seoseni ekspositsioontunnuse ja haigustunnuse vahel. See tähendab, et kõik riskifaktorid ei pruugi olla kindlaks tehtud. [5]

Näiteks paljud uuringud erinevate toiduainete tarbimise kohta kasutavad küsimustikku, milles uuritakse toiduaine tarbimise sagedust. On ilmnenud, et sellised küsimustikud põhjustavad mõõtmisviga. See on tingitud kombinatsioonist, kus uuringualused raporteerivad valesti toidu tarbimise kohta, ebatäpsetest küsimustest toidu tarbimise (sageduse) kohta ja mõõtmisveast, kuidas teisendada tarbitud toitu toitainete tasemele. See võib selgitada, miks ka pärast tuhandeid uuringuid ei ole veel selge, millised toiduained on riskiteguriteks ja millised kaitsevad sagedamini esinevate haiguste eest. [5]

2.4. Näide: üks esimesi Mendeli randomiseerimisel põhinevaid uuringuid

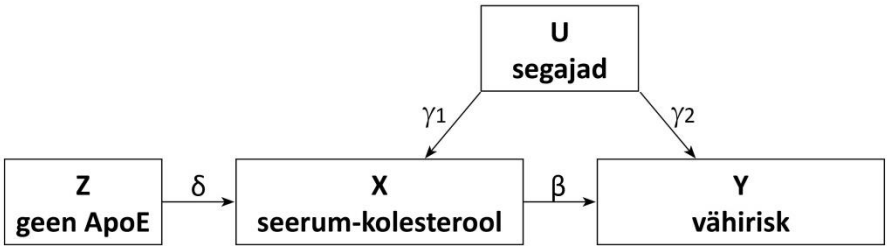
1980. aastatel (1986, Katan) arutleti selle üle, kas madal seerum-kolesterooli tase saab otseselt suurendada vähiriski. Alternatiivsed variandid täheldatud seosele olid näiteks, et kolesterooli tase on alanenud, kuna vähihaigetel patsientidel on juba olemas latentsed ehk mittemõõdetavad kasvajad (tagurpidi põhjuslikkus) või et nii vähirisk kui ka kolesteroolitase on mõjutatud sellistest segavatest teguritest nagu toitumine ja/ või suitsetamine. [7]

Täheldamine, et üksikisikutel, kellel oli *abetalipoproteinaemia* (ja seeläbi väga väike, mittearvestatav seerum kolesterooli tase kehas), ei tundunud olevad eelsoodumust vähile, viis Katani ideele, et tuleks uurida suuremat gruppi inimesi, kellel juba on geneetiliselt suurem kalduvus madalala kolesterooli tasemele. Oli teada, et geen ApoE (*apolipoprotein E*) mõjutab seerum-kolesterooli taset üldiselt ning geenivarianti ApoE2 seostati seerum-kolesterooli madalama tasemega. Katani idee seisnes

arvamises, et paljud inimesed kannavad geneetiliselt ApoE2-varianti ning seetõttu on juba sünnist saati madalama kolesteroolitasemega. [7]

Seega, kuna geenid määratakse juhuvalikuga juba meioosi käigus, ei ole ApoE2 kandjad millegi muu poolest, kui ApoE ja ApoE2 geenide erinevus, süstemaatiliselt erinevad inimestest, kes kannavad teisi ApoE alleele. Ainult siis, kui madal seerum-kolesterooli tase on põhjuslikuks teguriks haigusele, peaks vähihaigetel olema rohkem ApoE2 alleele kui kontrollrühmal. Kui põhjuslikkus puudub, peaksid ApoE alleelide jaotused olema mõlemas grupis ühesugused. Seda saab kergesti kontrollida, vaadates jaotusi. [7]

Arutlust iseloomustab Joonis 5, kus Z tähistab geeni ApoE, X seerum-kolesterooli taset kehas ning Y vähki haigestumist. Segavad tunnused on joonisel tähistatud tähega U.



Joonis 5. Mendeli randomiseerimise skeem Katani näite põhjal

2.4.1. Mendeli randomiseerimise matemaatiline põhjendus

Eeldame, et kõik Joonisel 5 kujutatud seosed on lineaarsed ehk kehtivad järgmised regressioonivõrrandid:

$$X = \alpha_x + \delta Z + \gamma_1 U + \varepsilon_x,$$

$$Y = \alpha_y + \beta X + \gamma_2 U + \varepsilon_y,$$

kus juhuslikud vead ε_x ja ε_y on sõltumatud ning $E(\varepsilon_x|Z,U) = E\left(\varepsilon_y\middle|X,U\right) = 0$.

Samuti eeldame, et $U \perp Z$, kus $\perp$ tähistab statistilist sõltumatust.

Jooniselt 5 näeme, et lihtne regressioonanalüüs, mis uurib seerum-kolesterooli mõju vähiriskile, annab meile nihkega hinnangu, sest

$$E(Y|X) = E\left(\alpha_y + \beta X + \gamma_2 U + \varepsilon_y \middle| X\right) = \alpha_y + \beta X + \gamma_2 E(U|X) + \varepsilon_y$$

ehk tunnuse X otsitav kordaja β sõltub ka parameetrist γ_2 ning X ja U vahelisest seosest.

Võttes abiks geneetilise muutuja Z, saame kirjutada

$$E(X|Z) = E(\alpha_x + \delta Z + \gamma_1 U + \varepsilon_x|Z) = \alpha_x + \delta Z, \quad (1)$$

sest eelduse kohaselt $E(U|Z) = 0$.

Kuna nii X kui ka Z on vaadeldud tunnused, on kordaja δ hinnatav lineaarsest regressioonimudelist, kus funktsioontunnuseks on X ja argumenttunnuseks Z.

Kasutades võrdust (1), saame

$$E(Y|Z) = E\left(\alpha_y + \beta X + \gamma_2 U + \varepsilon_y \middle| Z\right) = \alpha_y + \beta E(X|Z) + \gamma_2 E(U|Z) =$$
$$\alpha_y + \beta(\alpha_x + \delta Z) = \alpha_y^* + \beta\delta Z. \quad (2)$$

Seega on βδ hinnatav ning saame leida otsitava hinnangu parameetrile β:

$$\hat{\beta} = \frac{\widehat{\beta\delta}}{\hat{\delta}}$$

kus $\hat{\delta}$ on parameetri hinnang mudelist (1) ja $\widehat{\beta\delta}$ hinnang mudelist (2).

Standardvea hinnang hinnangule $\hat{\beta}$ saadakse kas nn võileivameetodil (ingl *sandwich method*) $\widehat{\beta\delta}$ ja $\hat{\delta}$ jagatise standardvea asümptootilisel lähendamisel või kasutades nn *bootstrap*-meetodit.

2.4.2. R-i funktsiooni *tsls* tööpõhimõte

Statistikatarkvara R funktsioon *tsls* (ingl *Two-Stage Least Squares*) paketist *sem* (ingl *General Structural Equation Models*) abistab põhjusliku mõju hindamise juures. Pakett *sem* on mõeldud struktuurivõrrandite mudelite hindamiseks. Mendeli randomiseerimine on struktuurivõrrandite mudelite erijuht, kus me eeldame kindlat seosestruktuuri.

R-i funktsiooni *tsls* sobib väga hästi eespool kirjeldatud seosestruktuuri (peatükis 2.4 kirjeldatud Katani näide) hindamiseks, kus soovime leida eksposits iooninäitaja põhjuslikku mõju väljundtunnusele, kasutades korrektse hinnangu saamiseks instrumenttunnust.

Funktsioon *tsls* hindab Mendeli randomiseerimise põhimõttel, leides otsitava kordaja, kasutades hinnangu saamiseks mitut etappi. Nagu meetodi nimigi ütleb on parameetrile hinnangu leidmine 2-astmeline: esmalt koostatakse regressioonivõrrandid hinnangute $\hat{\delta}$ ja $\widehat{\beta\delta}$ saamiseks seejärel leitakse saadud kahe hinnangu põhjal otsitava kordaja hinnang $\hat{\beta}$. Funktsioon on kasulik, kuna annab üheaegselt otsitava hinnanguga kordajale välja ka vastava standardvea hinnangu.

Funktsioon kirjutatakse kujul *tsls(mudel, instrument, andmestik)*. Lisada saab ka näiteks kaalud vaatlustele, et leida kaalutud hinnanguid; alamvaatluste vektori ja kriteeriumi, mida teha puuduvate väärtustega. [8]

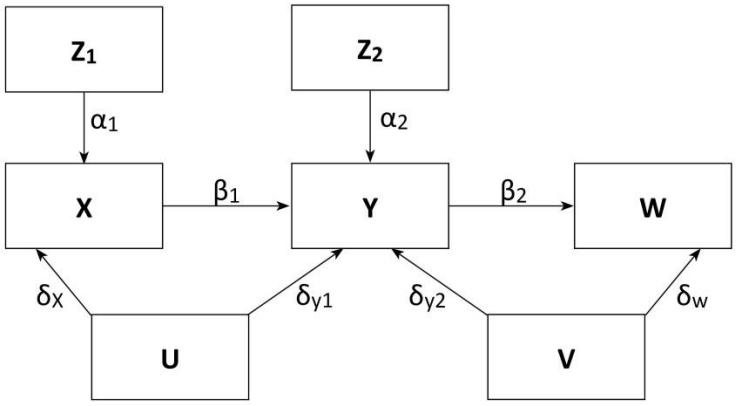
Näiteks eespool kirjeldatud Katani näite puhul, kus selgitati, milline on seerum-kolesterooli põhjuslik mõju vähiriskile, kasutades instrumendina geeni ApoE, tuleks andmed funktsiooni sisse kirjutada kujul

$$> tsls(formula = vähirisk{\sim}seerumkolesterool, instruments = ApoE, data = andmed)$$

2.5. Meetodi edasiarendus keerulisemale seosestruktuurile

Mendeli randomiseerimise laiendamisena saab tavalist meetodit kasutada ka keerulisemal seosestruktuuril. Laiendatud meetodit on graafiliselt kujutatud Joonisel 6, kus otsitakse hinnanguid parameetritele β_1 ja β_2.

Z_1 ja Z_2 tähistavad geneetilisi markereid, mida kasutatakse instrumentidena, otsides hinnanguid parameetritele β_1 ja β_2. Tunnused X, Y ja W on mõõdetud näitajad, mille omavahelisi põhjuslikke seoseid soovime uurida. Nii ekspositsiooninäitajaid kui ka väljundtunnuseid mõjutavad lisaks erinevad segavad faktorid U ja V, mis Mendeli randomiseerimise eelduste kohaselt ei ole seotud instrumenttunnustega.



Joonis 6. Mendeli randomiseerimise metoodika laiendamine

Joonisel 6 eeldame geenimarkerite Z_1 ja Z_2 mõju vastavalt näitajatele X ja Y. Tunnus X mõjutab otseselt tunnust Y ning läbi tunnuse Y ka tunnust W.

Näeme, et kordaja β_1 leidmine toimub analoogiliselt eelmises alapeatükis kirjeldatud viisile (ehk sarnaselt Katani näitele): kasutades instrumendina geneetilist markerit Z_1, saame leida hinnangu parameetrile α_1 ning kui α_1 ja $\alpha_1 \cdot \beta_1$ on hinnatavad, on võimalik leida ka kordaja β_1 otsitav hinnang. Samamoodi toimub hinnangu leidmine parameetrile β_2, kasutades vahendina geenimarkerit Z_2.

Mendeli randomiseerimise edasiarendusena saame leida hinnangu parameetrile β_2 ka teisel viisil – lähtudes geenimarkerist Z_1. Meetodi laiendamine on kasulik olukordades, kus geneetiline marker Z_2 kas puudub või on vaadeldava funktsioontunnusega Y nõrgalt seotud. Sel juhul saame hinnangu leidmiseks kordajale β_2 kasutada instrumenttunnusena geenimarkerit Z_1 ning otsitav kordaja avaldub kujul $\widehat{\beta_2} = \frac{\widehat{\alpha_1 \cdot \beta_1 \cdot \beta_2}}{\widehat{\alpha_1 \cdot \beta_1}}$.

2.5.1. Mudelite koostamine matemaatiliselt

Vastavalt Joonisele 6 saame välja kirjutada regressioonivõrrandid

$$X = \mu_x + \alpha_1 \cdot Z_1 + \delta_x \cdot U + \varepsilon_x,$$

$$Y = \mu_y + \beta_1 \cdot X + \delta_{y1} \cdot U + \delta_{y2} \cdot V + \varepsilon_y,$$

$$W = \mu_w + \beta_2 \cdot Y + \delta_w \cdot V + \varepsilon_w,$$

kus U ja V tähistavad segavaid faktoreid ning ε_x, ε_y ja ε_w on juhuslikud vead.

Eeldame, et $E(\varepsilon_x|Z,U) = E\left(\varepsilon_y\middle|X,U,V\right) = E(\varepsilon_w|Y,V) = 0$. Lisaks eeldame, et segavad faktorid on teineteisest sõltumatud ($U \perp V$, kus $\perp$ tähistab statistilist sõltumatust) ning $E(U) = E(V) = 0$.

Jooniselt 6 järelduvad ka eeldused $Z_1 \perp U$, $Z_1 \perp V$, $Z_2 \perp U$ ning $Z_2 \perp V$.

Näitame matemaatiliselt, kuidas leida hinnang parameetrile β_2, kasutades instrumendina geneetilist markerit Z_1. Kasutame tinglikku keskväärtust, tinglikustades üle kasutatava geenimarkeri Z_1.

Hinnangute leidmine parameetritele β_1 ja β_2, lähtudes vastavalt geenimarkeritest Z_1 ja Z_2, on analoogiline eelmises alapeatükis kirjeldatud näitele ja on välja toodud alapeatükis 2.4.1.

$$E(X|Z_1) = E(\mu_x + \alpha_1 \cdot Z_1 + \delta_x \cdot U + \varepsilon_x|Z_1)$$
$$= E(\mu_x|Z_1) + E(\alpha_1 \cdot Z_1|Z_1) + E(\delta_x \cdot U|Z_1) + E(\varepsilon_x|Z_1)$$
$$= \mu_x + \alpha_1 \cdot Z_1 + \delta_x \cdot E(U|Z_1) + 0$$
$$= \mu_x + \alpha_1 \cdot Z_1 + \delta_x \cdot 0 = \mu_x + \alpha_1 \cdot Z_1 \quad (3)$$

$$E(Y|Z_1) = E\left(\mu_y + \beta_1 \cdot X + \delta_{y1} \cdot U + \delta_{y2} \cdot V + \varepsilon_y\middle|Z_1\right)$$
$$= \mu_y + \beta_1 \cdot E(X|Z_1) + \delta_{y1} \cdot E(U|Z_1) + \delta_{y2} \cdot E(V|Z_1) + E\left(\varepsilon_y\middle|Z_1\right)$$
$$= \mu_y + \beta_1 \cdot (\mu_x + \alpha_1 \cdot Z_1) = \mu_y^* + \alpha_1 \cdot \beta_1 \cdot Z_1 \quad (4)$$

$$E(W|Z_1) = E(\mu_w + \beta_2 \cdot Y + \delta_w \cdot V + \varepsilon_w|Z_1)$$
$$= \mu_w + \beta_2 \cdot E(Y|Z_1) + \delta_w \cdot E(V|Z_1) + E(\varepsilon_w|Z_1)$$
$$= \mu_w + \beta_2 \cdot \left(\mu_y^* + \alpha_1 \cdot \beta_1 \cdot Z_1\right)$$
$$= \mu_W^* + \alpha_1 \cdot \beta_1 \cdot \beta_2 \cdot Z_1 \quad (5)$$

Võrranditest (3), (4) ja (5) järeldub, et hinnatavateks parameetriteks on kordaja α_1 ning korrutised $\alpha_1 \cdot \beta_1$ ja $\alpha_1 \cdot \beta_1 \cdot \beta_2$. Seega saame β_2 hinnanguks

$$\widehat{\beta_2} = \frac{\widehat{\alpha_1 \cdot \beta_1 \cdot \beta_2}}{\widehat{\alpha_1 \cdot \beta_1}}$$

Et analoogselt peatükis 2.4.1 kirjeldatule on β_2 hinnanguks ka $\widehat{\beta_2} = \frac{\widehat{\alpha_2 \cdot \beta_2}}{\widehat{\alpha_2}}$, on sama parameetri hindamiseks kaks võimalust. Näeme, et kordaja β_2 on hinnatav ka siis, kui üks instrumentidest (Z_1 või Z_2) on kas puudu või väga nõrgalt seotud vastava funktsioontunnusega (X või Y).

2.6. Bootstrapi põhimõte parameetri hinnangu standardvea leidmiseks

Soovides hinnata valimi põhjal parameetritele β_1 ja β_2 saadud hinnangute hajuvust ehk standardviga (populatsiooni muutlikkuse keskmine mõõde), võtame lisaks kasutusele *bootstrap*-meetodi. Peatükis 2.4.2 kirjeldatud funktsioon *tsls*, annab koos parameetri hinnanguga välja ka hinnangu standardvea, ent funktsiooni ei saa kasutada Mendeli randomiseerimise metoodika laiendatud variandi korral.

Bootstrap-meetod ehk „saapapaela meetod" on simuleerimismeetod, mis põhineb eeldusel, et empiiriline (valimi) jaotusfunktsioon on ligikaudu tegelik uuritava tunnuse jaotus populatsioonis. Seega eeldatakse *bootstrap*-meetodit kasutades, et olemasolev andmestik (valim) kirjeldab üldkogumit (populatsiooni). Kogu informatsioon saadakse ning hinnangud otsitava parameetri θ kohta tehakse algse valimi X^0 kaudu. [9]

Kasutades lihtsat juhuslikku tagasipanekuga valimit (see tähendab, et esialgse andmestiku element võib uues valimis korduda ning kõigi saadud elementide valimise tõenäosus on võrdvõimalik), võetakse olemasolevast – üldkogumit kirjeldavast valimist – uus valim, mida nimetatakse *bootstrap*-valimiks või pseudoandmestikuks. Niimoodi saadud pseudoandmestik on samast jaotusest nagu esialgne andmestik, kusjuures saadud andmestikus on sama palju elemente, kui oli esialgses andmestikus. Protsessi korratakse m korda, tavaliselt vähemalt 1000 või 10 000 korda. [9]

Igas simuleeritud *bootstrap*-valimis arvutatakse meid huvitava statistiku (näiteks regressiooniparameetri) väärtus θ_i^* $(i = 1, \dots, m)$. Tulemuseks saadakse m väärtust uuritavale parameetrile: $\theta_1^*, \dots, \theta_m^*$. Saadud väärtuste pealt saame rekonstrueerida teststatistiku jaotuse või hinnata seda jaotust iseloomustava arvkarakteristiku väärtuse (hindame hinnangute standardhälbe). [9]

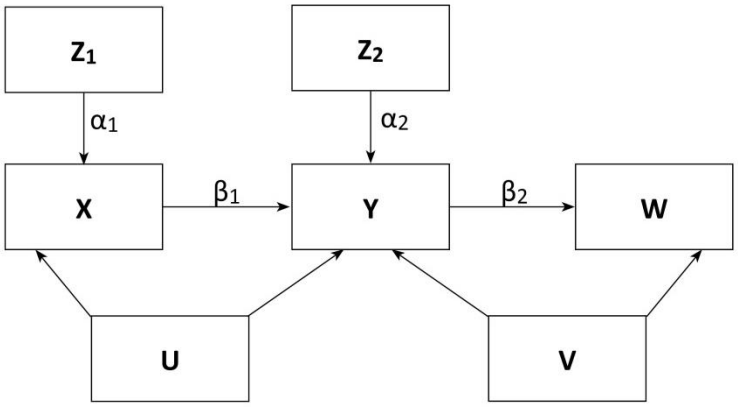
Nendele väärtustele vastav empiiriline jaotus hindab $\hat{\theta}$ („algse valimi" pealt saadud hinnang parameetrile θ) jaotust üldkogumi jaotuse suhtes. Saadud empiirilist jaotust kutsutakse ka *bootstrap*-jaotuseks. *Bootstrap*-jaotuse põhjal saab arvutada uuritava

parameetri keskmise ning standardhälbe üldkogumi jaoks. *Bootstrap*-meetod põhineb tsentraalsel piirteoreemil: juhuslikult moodustatud pseudoandmestikud on kõik teineteisest sõltumatud ning sama jaotusega ja seetõttu on tsentraalse piirteoreemi eeldused täidetud. [9]

Simulatsiooni kood statistikatarkvaras R on välja toodud Lisas 1.

2.7. Simulatsioonieksperiment Mendeli randomiseerimise meetodi edasiarenduse testimiseks

Simulatsiooni abil uurime, kui hästi on eespool kirjeldatud Mendeli randomiseerimise metoodika kasutatav hüpoteetilise andmestiku peal, et leida hinnangud parameetritele β_1 ja β_2. (Joonis 7)



Joonis 7. Mendeli randomiseerimise metoodika laiendatud skeem. Uuritavad seosed ja seosesuunad koos kordajate ning segavate faktoritega

Joonisel 7 tähistavad Z_1 ja Z_2 instrumenttunnustena kasutusel olevaid geneetilisi markereid. Tunnused X, Y ja W on mõõdetud näitajad, mille omavahelisi põhjuslikke seoseid soovime uurida. Nii ekspositsiooninäitajaid kui ka väljundtunnuseid mõjutavad lisaks erinevad segavad faktorid U ja V, mis Mendeli randomiseerimise eelduste kohaselt ei ole seotud instrumenttunnustega.

Simulatsiooni rakendamiseks koostatakse hüpoteetiline andmestik, mis sisaldab eespool nimetatud tunnuseid. Kordajad seoste moodustamisel (parameetrite ees olevad kordajad ning võrrandite vabaliikmed) on valitud juhuslikult. Käesolevas simulatsioonis on genereeritud 1000 andmestikku.

Andmed, valimimahuga $n = 2000$, on genereeritud järgnevalt.

1. Nii tunnuseid X kui ka Y mõjutavate markerite skoor on binoomjaotusega vastavalt $Z_1 \sim Bin(2; 0.4)$ ja $Z_2 \sim Bin(2;\ 0.3)$.
2. „Tundmatute" kordajate β_1 ja β_2 väärtused on vastavalt 1 ja 2 (valitud juhuslikult) või 0 ja 0 (kontrollimaks olukorda, kus tegelikku põhjuslikku mõju ei esine).
3. Vaadeldavad tunnused X, Y ja W on omavahel sõltuvad järgnevalt:

$$X = -2 + \alpha_1 \cdot Z_1 - 4 \cdot U + \varepsilon_x$$

$$Y = -10 + \alpha_2 \cdot Z_2 + \beta_1 \cdot X + U + 2 \cdot V + \varepsilon_y$$

$$W = 5 + \beta_2 \cdot Y + 3 \cdot V + \varepsilon_w$$

kus nii segajad U ja V kui ka juhuslikud vead ε_x, ε_y ning ε_w on normaaljaotusega $N(0, 1)$. Kõik parameetrite ees olevad kordajad ning võrrandite vabaliikmed on valitud juhuslikult.

Kordajate α_1 ja α_2 väärtused näitavad instrumenttunnuste mõju tugevust. Simulatsiooni käigus vaadeldakse instrumenttunnuste kahte olukorda: esiteks kui kasutusel on uuritavat tunnust nõrgalt mõjutavad instrumenttunnused ($\alpha_1 = 0.2$, $\alpha_2 = 0.3$) ning teiseks kui on valitud prognoositava tunnusega tugevamalt seotud instrumenttunnused ($\alpha_1 = 2$, $\alpha_2 = 3$).

Simulatsiooni käigus kontrollime esmalt ekspositsiooni mõju väljundnäitajale tavalise lineaarse regressioonanalüüsi abil (tunnuse Y sõltuvus tunnusest X parameetri β_1 saamiseks ning tunnuse W sõltuvus tunnusest Y parameetri β_2 saamiseks, kasutamata geneetiliste markerite skoore) ning hiljem vaatleme, kas

saame parameetri õige väärtuse kätte ka Mendeli randomiseerimise kaudu, kasutades instrumenttunnustena geneetilisi markereid Z_1 ja Z_2.

Simuleeritud andmete korral on meil küll teada uuritavaid tunnuseid mõjutavad segajad U ja V, ent kuna segavaid faktoreid ei ole võimalik mõõta reaalseid andmeid kasutades, siis kasutame neid vaid andmete genereerimise etapis, jäljendades tegelike segavate faktorite olemasolu reaalsete andmete korral. On selge, et mida tugevamalt mõjutavad segavad faktorid prognoositavat tunnust, seda ebatäpsemaks muutub lineaarne regressioonhinnang, kus saame segavate faktorite olemasolu küll eeldada, ent mitte otseselt mõõta. Ka Mendeli randomiseerimist kasutades eeldame segavate tunnuste olemasolu, mida otseselt mõõta ei saa, ent saame uuritavat seost n-ö ringiga hinnata.

Lineaarsel regressioonanalüüsil leiame otsitava parameetri väärtuse, vaadates lineaarset seost ekspositsiooninäitaja ja väljundnäitaja vahel (näitaja Y sõltuvus näitajast X ning näitaja W sõltuvus näitajast Y). Mendeli randomiseerimisel leiame parameetri väärtuse instrumenttunnusega lähenemise abil, võttes instrumentideks geneetilised markerid Z_1 ja Z_2.

Meetodi edasiarendamisena kasutame lisaks Mendeli randomiseerimise laiendamist, jagades lineaarsete regressioonmudelite parameetrid nii, nagu kirjeldatud peatükis 2.5.1. (mudeli koostamine matemaatiliselt).

Parameetrite hinnangud koos hinnangute standardvigadega on välja toodud Tabelis 1.

Tabel 1. Parameetrite β_1 ja β_2 õiged väärtused koos lineaarsel regressioonanalüüsil ning Mendeli randomiseerimisel (MR) saadud tulemustega. Mendeli randomiseerimise korral sulgudes esmalt standardvigade hinnangute keskmine, teisena standardviga üle kõigi simuleeritud andmete

Parameeter	Parameetri õige väärtus	Hinnang lineaarsel regressioonil (standardveaga)	Keskmine hinnang Mendeli randomiseerimisel ($\alpha_1 = 0.2, \alpha_2 = 0.3$)	Keskmine hinnang Mendeli randomiseerimisel ($\alpha_1 = 2, \alpha_2 = 3$)
β_1	1	0.3055 (0.0127)	1.0943 (23.2845; 47.6524)	0.9995 (0.0508; 0.0039)
β_2 (MR korral lähtudes Z_2-st)	2	1.1437 (0.0186)	2.0215 (22.1011; 41.0181)	2.0018 (0.0513; 0.0053)
β_2 (MR korral lähtudes Z_1-st; meetodi edasiarendus)	2		2.0721 (19.0744; 32.6149)	1.9985 (0.0367; 0.0028)
β_1	0	–0.7271 (0.0129)	0.1801 (37.3119; 110.2869)	0.0063 (0.0515; 0.0045)
β_2 (MR korral lähtudes Z_2-st)	0	–0.4351 (0.0159)	0.2952 (22.7027; 22.1788)	0.0059 (0.0367; 0.0025)
β_2 (MR korral lähtudes Z_1-st; meetodi edasiarendus)	0		0.2985 (11.4605; 25.3077)	–0.0006 (0.0363; 0.0021)

Tabelist 1 näeme, et Mendeli randomiseerimist kasutades saame parameetrite õigetele väärtustele tunduvalt lähedasemad hinnangud kui regressioonanalüüsi abil. Seejuures on meetodi edasiarendust kasutades saadud sama täpsusega või isegi täpsemad hinnangud parameetrile β_2, võrreldes tavalise Mendeli randomiseerimise meetodil lähenemisega.

Kuid olukorras, kus instrumenttunnuse mõju uuritavale tunnusele on nõrk ($\alpha_1 = 0.2$, $\alpha_2 = 0.3$) ei saa me Mendeli randomiseerimist kasutades olla kindlad, et saadud hinnang on korrektne, kuna hinnangute standardvead (nii standardvigade hinnangute keskmine kui ka standardvead üle kõigi simuleeritud β-de) on kordades suuremad parameetrile saadud hinnangust. Kuna Mendeli randomiseerimise meetod kasutab otseselt instrumenttunnust põhjusliku mõju prognoosimiseks, on nimetatud meetodi korral parameetrite standardvead seda väiksemad, mida tugevam on seos instrumenttunnuse ja väljundnäitaja vahel.

Lineaarsel regressioonanalüüsil ilmneb ekslik seos olukorras, kus tegelikku põhjuslikku mõju ei esine ($\beta_1 = 0,\ \beta_2 = 0$). Selles olukorras on Mendeli randomiseerimist kasutades variandi puhul, kus instrumenttunnused on eksponsitsiooninäitajaga tugevalt seotud, selgelt näha, et tegelikku põhjuslikku mõju vaadeldavate tunnuste vahel ei esine. Instrumenttunnuste vahelise nõrgema seose korral näitab ka Mendeli randomiseerimine parameetrite hinnanguteks nullist erinevaid kordajaid, ent tulemus on suurte standardvigade tõttu nii või teisiti kaheldav.

Simulatsiooni kood statistikatarkvaras R on välja toodud Lisas 2.

3. Tartu Ülikooli Eesti Geenivaramu andmete analüüs Mendeli randomiseerimise põhimõttel

3.1. Ülevaade andmetest

Tartu Ülikooli Eesti Geenivaramu (edaspidi TÜ geenivaramu) on teadus- ja arendusasutus, mille eesmärk on edendada geeniuuringute arengut, koguda teavet Eesti rahvastiku terviseandmete ja päriliku informatsiooni kohta ning rakendada uusimaid uuringutulemusi rahva tervise parandamiseks. [10]

TÜ geenivaramu andmebaasis on ligikaudu 52 000 vabatahtliku geenidoonori andmed, millest varaseimad pärinevad 2002. aasta oktoobrist. Et oleks võimalik uurida seoseid geenide ja haiguste vahel, täidavad kõik geenidoonorid küsimustiku, mis sisaldab isikuandmeid, terviseseisundi kirjeldust ja sugupuuandmeid. Lisaks võetakse igalt geenidoonorilt vereproov, millest eraldatakse DNA, vereplasma ja valged verelibled, mida säilitatakse vedelas lämmastikus. [11, lk 5]

Käesolevas töös analüüsitakse 4497 geenidoonori andmeid. Antud valim (juhuvalim kõigist TÜ geenivaramu geenidoonoritest) koosneb indiviididest, kelle puhul olid mõõdetud teatud geneetilised markerid – üksiku nukleotiidi polümorfismid (edaspidi SNP-d) ning kelle puhul oli samaaegselt teada nelja lipoproteiini kontsentratsioon. Töös on kasutusel lipoproteiinid tsitaat (Cit), α-1 glükoproteiin (Gp), LDL-kolesterool (LDL-C) ning üldkolesterool (Serum-C).

Käesolevas töös on kasutusel viis geenimarkerit, mille kohta on teada, et nad mõjutavad indiviidi kohvi tarbimist (rs1260326, rs1481012, rs6968554, rs6265 ja rs2472297) ning viis metaboliite mõjutavat geenimarkerit (rs7412, rs17112596, rs16848079, rs217181 ja rs712959).

Lisaks kokku kümnele geenimarkerile ja neljale metaboliidile on analüüsitavasse andmetabelisse valitud veel seitse tunnust geenidoonorite poolt täidetud küsimustikust: sugu, vanus, kehamassiindeks (KMI), info selle kohta, kas

geenidoonor on suitsetaja, kohvi tarbimise sagedus päevas ning ülemine (süstoolne) ja alumine (diastoolne) vererõhk.

3.1.1. Taustatunnused

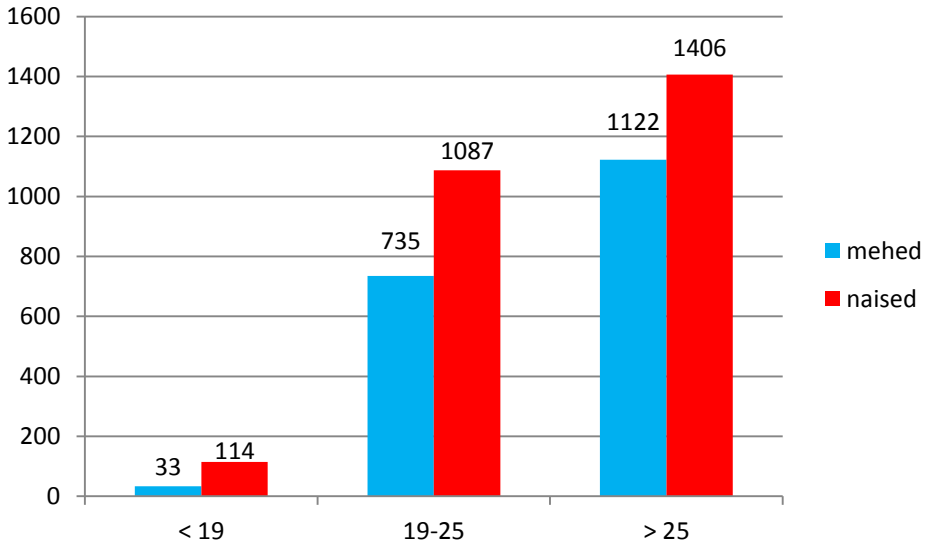
Analüüsitavatest 4497-st geenidoonorist on 2607 naised, mis moodustab 58% kõigist analüüsitavas andmetabelis olevatest geenidoonoritest, ning 1890 mehed, mis teeb 42% koguarvust.

Noorimad geenidoonorid olid andmete kogumise hetkel 18-aastased (140 inimest), vanim 103-aastane (üks inimene). Analüüsitavate geenidoonorite keskmine vanus on 46,6 eluaastat. (Tabel 2)

Tabel 2. Sagedustabel soo ja vanuse järgi

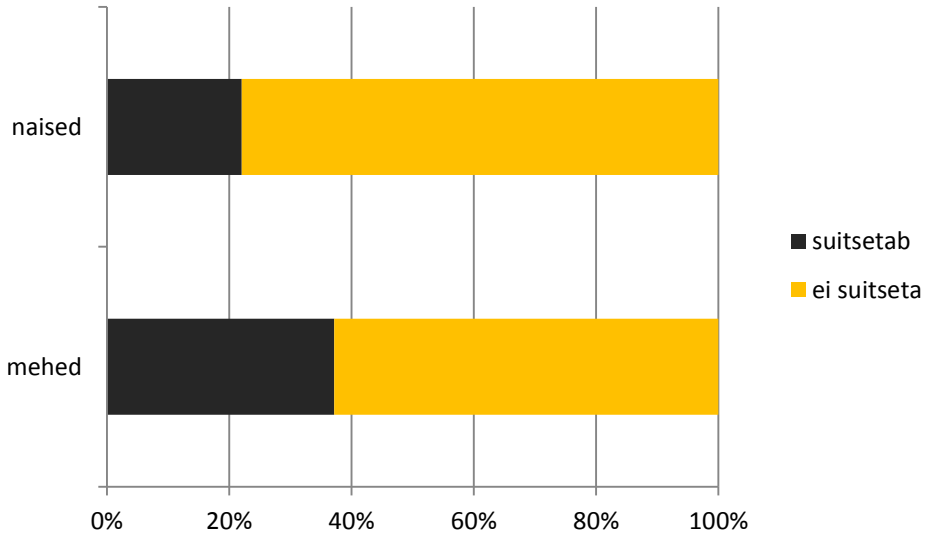
	Sugu		Vanus liitumisel (aastates)							
	Mees	Naine	18–27	28–37	38–47	48–57	58–67	68–77	78–87	88–103
Indiviidide arv	1890	2607	907	748	747	693	643	529	207	23
Osakaal (%)	42,0	58,0	20,2	16,6	16,6	15,4	14,3	11,8	4,6	0,5
Kokku	4497		4497							

Geenidoonorite keskmine kehamassiindeks (KMI) oli 26,6 ühikut, mis kaldub pigem ülekaalu kui normaalkaalu poole [4, vana]. Üle poolte (2528 indiviidi ehk 56,2%) andmestikus olevate geenidoonorite KMI ületab soovitusliku normaalkaalu piiri (25 ühikut). Alakaalulisi (KMI alla 19 ühiku) geenidoonoreid oli andmestikus 147 indiviidi ehk 3,3% kõigist geenidoonoritest. Normaalkaalu piiridesse jäävaid isikuid oli 1822 ehk 40,5% geenidoonorite KMI jääb soovitusliku 19-25 ühiku piiresse. Nii meeste kui ka naiste keskmine KMI on sarnane ning sarnaneb ka geenidoonorite keskmise KMI-ga, olles meeste puhul 26,8 ühikut ning naiste korral 26,4 ühikut. (Joonis 8)



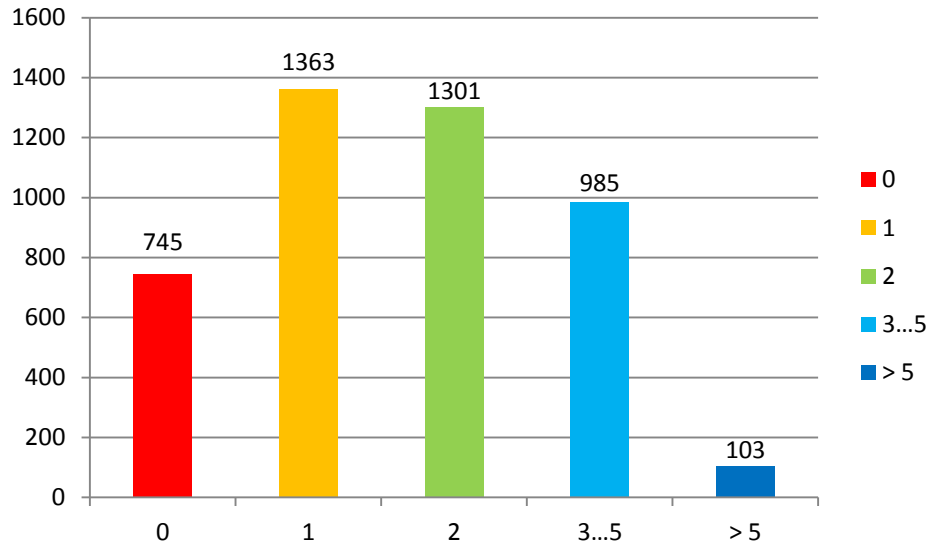
Joonis 8. Geenidoonorite kehamassiindeksi jaotus soo järgi

Geenidoonoritest 1278 inimest (28,4%) olid vereproovi andmise ajal suitsetajad, 3219 (71,6%) mittesuitsetajad. Seejuures meestest oli suitsetajaid 703 inimest (37,2% kõigist andmetabelis olevatest meestest), naistest 575 inimest (22,1%). (Joonis 9)



Joonis 9. Geenidoonorite suitsetamisharjumused sugude lõikes

Ainsa toitumisharjumust kirjeldava tunnusena on käesolevas töös kasutusel tunnus „kohvi joomine“. Geenidoonoritelt küsiti, kui palju joovad nad päevas kohvi. Ilmneb, et keskmiselt tarbivad geenidoonorid 2 tassi kohvi päevas. Rohkem kui viis tassi päevas joovad kohvi 103 (2,3%) geenidoonorit, mitte üldse ei joo kohvi 745 geenidoonorit (16,6%). (Joonis 10)



Joonis 10. Geenidoonorite kohvi tarbimise harjumused (tasside arv päevas)

Tervisenäitajatena on analüüsi valitud süstoolne (ülemine, ingl sbp – *systolic blood pressure*) ja diastoolne (alumine, ingl dbp – *diastolic blood pressure*) vererõhk. Vererõhk on arteriaalne rõhk, mis tekib südame kontraktsioonil ja vere pumpamisel suure vereringe arteritesse. Nii nimetatakse rõhku, mida avaldab veresoontes voolav veri veresoonte seintele. Vererõhu määravad südame löögimaht ja perifeerne veresoonte vastupanu. Terve inimese vererõhk on üsna konstantne. Vananedes veresoonte elastsus siiski väheneb, perifeerne vastupanu tõuseb ja vererõhul on kalduvus tõusta. Lisaks vanusele mõjutavad vererõhku ka kehaline aktiivsus, hirm, kaalutõus, liigne soola tarbimine, alkoholi igapäevane tarvitamine, stress ning uni. Vererõhku väljendatakse kahe arvuga. [12]

Ülemine ehk süstoolne rõhk näitab rõhku südame kontraktsiooni ajal. Süstoolne rõhk tekib vasaku vatsakese kontraktsioonil ja näitab südame, arterite ja arterioolide terviklikkust. Alumine ehk diastoolne rõhk tekib vasaku vatsakese lõõgastumisel ja näitab veresoonte resistentsust (vastupanu). Süstoolse vererõhu normiks loetakse 110-140 mmHg ning diastoolse vererõhu normiks 70–90 mmHg. [12]

Geenidoonorite keskmine süstoolne vererõhk on 127,6 mmHg (meestel 130,8; naistel 125,3), diastoolne 78,5 mmHg (meestel 80,2; naistel 77,2). Mõlemad näitajad jäävad nii naistel kui ka meestel soovitusliku normi piiresse.

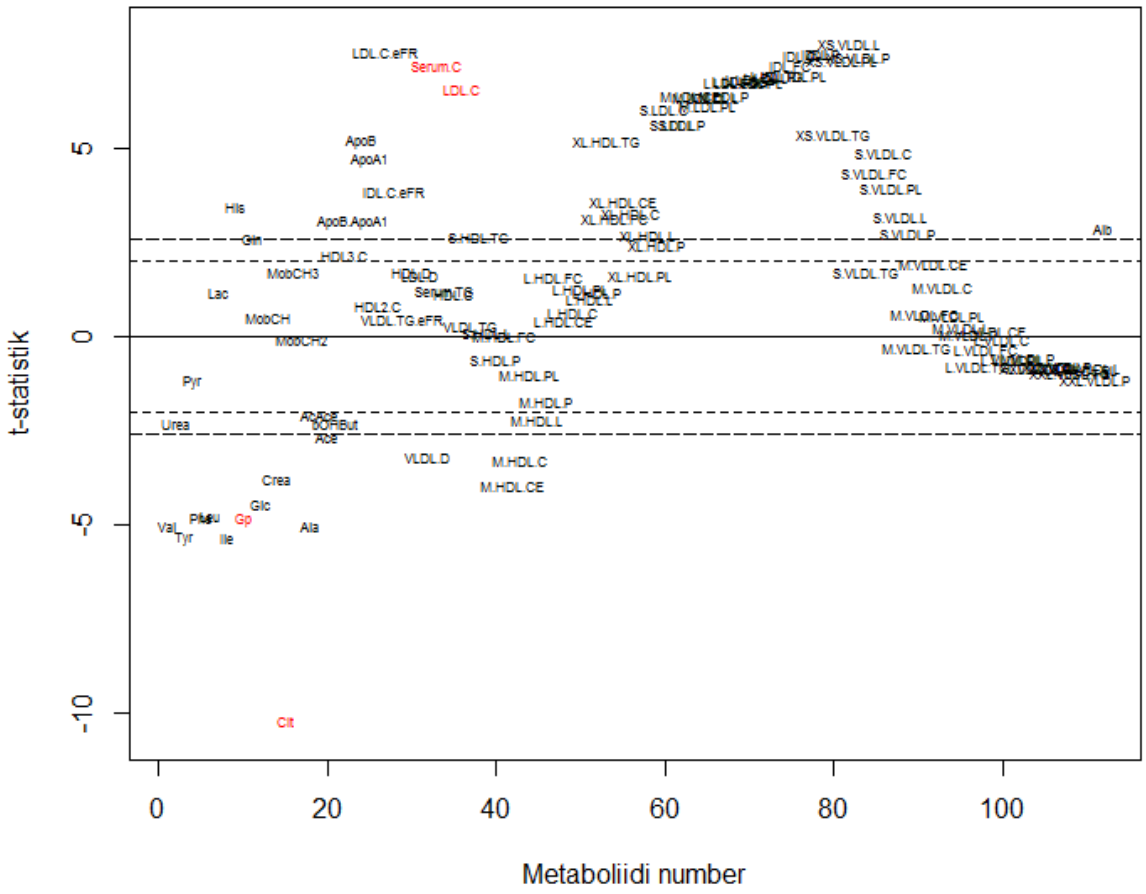
3.1.2. Metaboliidid

TÜ geenivaramu metaboloomika-andmete hulgas on üle saja tunnuse, mis iseloomustavad eri tüüpi lipoproteiinide kontsentratsiooni. Lipoproteiinid on valgu ehk proteiini ja lipiidi kompleksid, mis tagavad mitmete erinevate rasvamolekulide (sh kolesterooli) liikumise vereringesse [13]

Uurides kohvi joomise mõju igale üksikule metaboliidile, kasutades lineaarset regressioonanalüüsi, kus arvestatud on ka soo, vanuse, kehamassiindeksi ja suitsetamise mõju, ilmnes, et kohvi tarbimine mõjutab tugevalt selliseid metaboliite nagu Cit, LDL-C-eFR, Serum-C ja LDL-C (Joonis 11). Tuginedes saadud seostele ning eelnevatele teadmistele kohvi joomise mõjust metaboliitide, on analüüsimiseks välja valitud neli metaboliiti: tsitaat (Cit), α-1 glükoproteiin (Gp), LDL-kolesterool (LDL-C) ning üldkolesterool (Serum-C). Väljavalitud metaboliidid on Joonisel 11 kujutatud punase värviga.

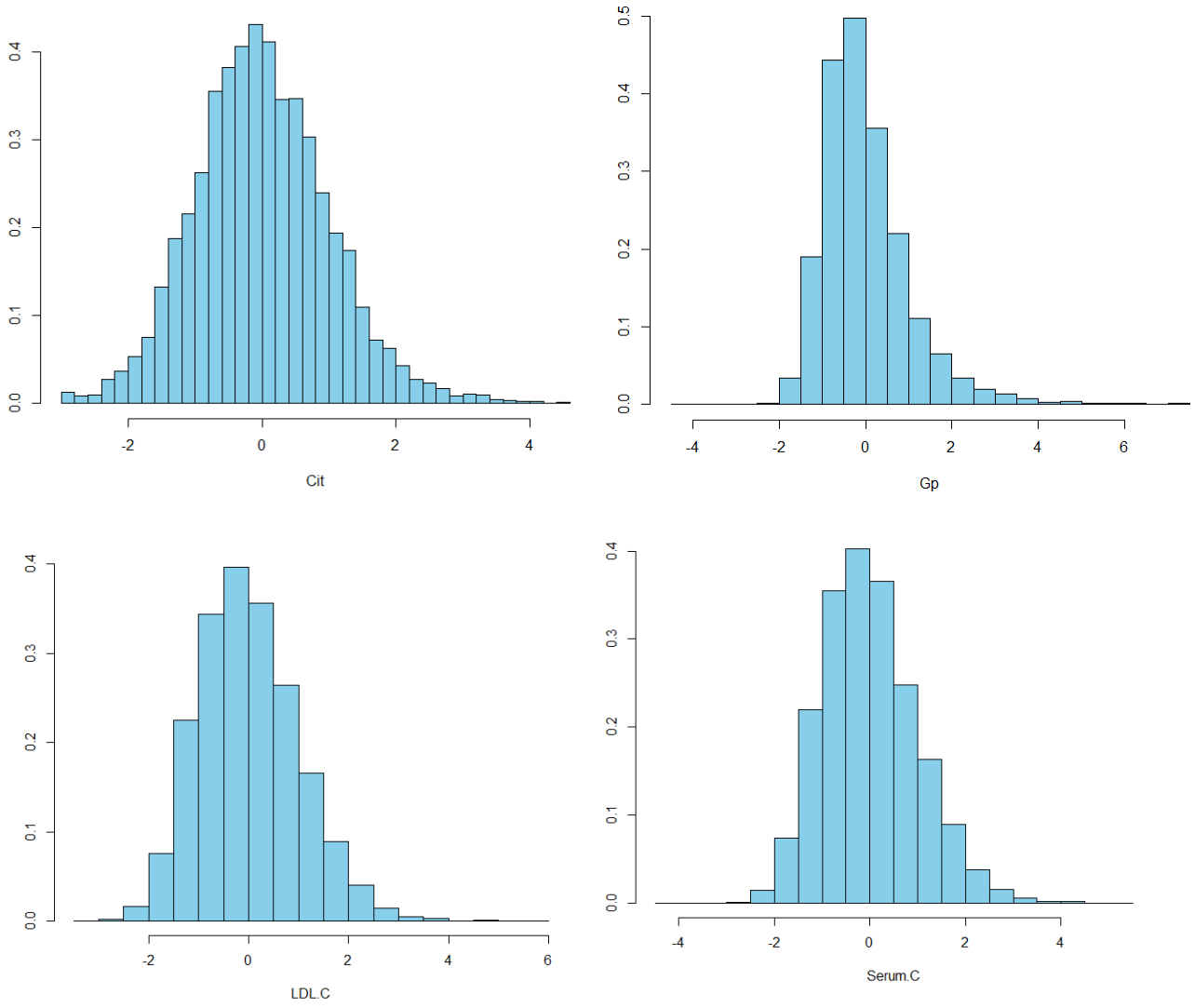
Igale metaboliidile TÜ geenivaramu andmestikust hinnati lineaarne regressioonimudel, kus argumentideks olid päevane tarbitav kohvitasside arv, sugu, vanus, kehamassiindeks ning suitsetamisstaatus vereproovi võtmise ajal. Mudeli hindamisel on salvestatud tunnusele „kohvi tarbimine päevas“ vastava t-statistiku väärtus iga lipoproteiini korral. Nii on saadud t-statistiku väärtused, mis on kantud joonisele. Vertikaalsel teljel on kujutatud t-statistiku väärtusi, horisontaalsel lipoproteiine.

Horisontaalse musta punktiirjoonega on tähistatud piirid -2 ja $+2$ ning $-2{,}6$ ja $+2{,}6$, mis vastavad normaaljaotuse vastavalt 97,5 ja 99,5 kvantiilidele. Seega olulisusenivool $\alpha = 0{,}05$ loetakse statistiliselt mitteoluliseks seosed lipoproteiinidega, mis jäävad sisemiste piirjoonte vahele. Olulisusenivool $\alpha = 0{,}01$ on statistiliselt mitteolulised välimiste piirjoonte vahele jäävad seosed. Välja on toodud ka nullpunkt.



Joonis 11. Kohvi joomise mõju lipoproteiinidele lineaarsel regressioonanalüüsil

Joonisel 12 on välja toodud lipoproteiinide jaotusi kirjeldavad histogrammid koos vastava tihedusfunktsiooniga. Vertikaalsel teljel on tihedus ning horisontaalsel lipoproteiini suurusjärk.



Joonis 12. Lipoproteiinide jaotusi kirjeldavad histogrammid

3.1.3. Geneetilised markerid

Geenimarkeritena, mida kasutada instrumentidena, et hinnata valitud tunnuste omavahelist põhjuslikku mõju, on käesolevas töös kasutusel SNP-markerid.

Geneetilistest markeritest, mis mõjutavad indiviidi kohvi tarbimist, on käesolevas töös kasutusel viis geenimarkerit (rs1260326, rs1481012, rs6968554, rs6265 ja rs2472297). Liites markerid, on arvutatud kohvimarkerite skoor, kasutades teadaolevaid kordajaid. [14]

Tabelis 3 on näha kasutatavad kohvi tarbimist mõjutavad geneetilised markerid koos taustainfoga. Välja on toodud võrdlus töös kasutatud eelnevalt teadaolevate kordajate ning TÜ geenivaramu andmete põhjal saadud hinnangute vahel. Lisaks on välja kirjutatud geenimarkerite pealt moodustatud riskiskoor.

Metaboliite mõjutatavate geenimarkeritena on antud töös vaadeldud viit erinevat markerit (rs7412, rs17112596, rs16848079, rs217181, rs712959). Neist kolm (rs7412, rs17112596, rs16848079) mõjutavad teadaolevalt metaboliiti Serum-C, kolm (rs7412, rs17112596, rs16848079) metaboliiti LDL-C, kaks tükki metaboliiti Gp (rs217181, rs16848079) ning üks (rs712959) metaboliiti Cit.

Tabelis 4 on nimetatud markerid koos taustainfoga välja toodud. Ent kuna TÜ geenivaramu andmete põhjal ei osutunud kõik geneetilised markerid statistiliselt olulisteks, on käesolevas töös iga metaboliidi kohta kasutusel üks geneetiline marker, millele vastav p-väärtus oli kasutatavate andmete põhjal väikseim (Cit – rs712959, Gp – rs217181, LDL-C – rs7412, Serum-C – rs7412). Metaboliidi skoor on taas arvutatud, kasutades vastavat kordajat Tabelist 4. Samas tabelis on välja toodud ka saadud riskiskoor.

Tabel 3. Kasutatavad kohvi tarbimist mõjutavad geenimarkerid ja markerite skoor koos taustainfoga (Chr – kromosoom, Pos – positsioon, A1 – Alleel1, A2 – Alleel2, EAF – efektialleeli sagedus, mis leitakse $\frac{\bar{x}_i}{2}$, kui x_i kodeeritud kui $0, 1, 2$) ning parameetrite hinnangu ja hinnangu standardhälbega

Geneetiline marker	Chr	Pos	A1	A2	EAF	Mõju kohvi joomisele			
						GWAS [14]		TÜ geenivaramu andmed	
						$\hat{\beta}$	Standard-viga	$\hat{\beta}$	Standard-viga
rs1260326	2	27730940	T	C	0,41	–0,04	0,01	–0,04	0,02
rs1481012	4	89039082	A	G	0,89	0,06	0,01	0,06	0,03
rs6968554	7	17287106	A	G	0,39	–0,10	0,01	–0,05	0,02
rs6265	7	27679916	T	C	0,19	–0,04	0,01	–0,03	0,03
rs2472297	15	75027880	T	C	0,24	0,14	0,01	0,15	0,02
Skoor SNP_1	–	–	–	–	–	–	–	1,03	0,1711

Tabel 4. Kasutatavad metaboliite mõjutavad geenimarkerid ja markerite skoor koos taustainfoga (Chr – kromosoom, Pos – positsioon, A1 – Alleel1, A2 – Alleel2, EAF – efektialleeli sagedus, mis leitakse $\frac{\bar{x}_i}{2}$, kui x_i kodeeritud kui $0, 1, 2$) ning parameetrite hinnangu, hinnangu standardhälbe ja vastava p-väärtusega

Metaboliit	Geneetiline marker	Chr	Pos	A1	A2	EAF	Mõju metaboliidi kontsentratsioonile					
							GWAS [15]			TÜ geenivaramu andmed		
							$\hat{\beta}$	Standard-viga	p-väärtus	$\hat{\beta}$	Standard-viga	p-väärtus
Cit	rs712959	22	17539015	C	T	0,59	-0,14	0,02	$1{,}52 \cdot 10^{-15}$	-0,03	0,02	0,132
Gp	rs217181	16	70671503	C	T	0,17	0,17	0,02	$1{,}43 \cdot 10^{-15}$	0,13	0,02	$2{,}57 \cdot 10^{-8}$
	rs16848079	4	73630887	C	T	0,04	0,35	0,05	$1{,}51 \cdot 10^{-11}$	0,07	0,04	0,0694
LDL-C	rs7412	19	50103919	C	T	0,04	-0,69	0,05	$2{,}94 \cdot 10^{-51}$	-0,43	0,03	$<2 \cdot 10^{-16}$
	rs17112596	1	55892884	C	T	0,02	-0,57	0,06	$3{,}77 \cdot 10^{-19}$	0,01	0,07	0,9455
	rs16848079	4	73630887	C	T	0,04	0,38	0,05	$2{,}64 \cdot 10^{-12}$	0,06	0,04	0,1419
Serum-C	rs7412	19	50103919	C	T	0,04	-0,48	0,05	$2{,}51 \cdot 10^{-26}$	-0,32	0,03	$<2 \cdot 10^{-16}$
	rs17112596	1	55892884	C	T	0,02	-0,53	0,06	$1{,}25 \cdot 10^{-16}$	-0,01	0,08	0,930
	rs16848079	4	73630887	C	T	0,04	0,38	0,05	$7{,}86 \cdot 10^{-13}$	0,06	0,04	0,101
Cit	Skoor SNP_2	–	–	–	–	–	–	–	–	0,23	0,15	0,127
Gp	Skoor SNP_2	–	–	–	–	–	–	–	–	0,72	0,13	$7{,}78 \cdot 10^{-8}$
LDL-C	Skoor SNP_2	–	–	–	–	–	–	–	–	0,61	0,04	$<2 \cdot 10^{-16}$
Serum-C	Skoor SNP_2	–	–	–	–	–	–	–	–	0,66	0,06	$<2 \cdot 10^{-16}$

4. Mudelid Mendeli randomiseerimise põhimõttel

4.1. Ülevaade koostatavatest mudelitest

TÜ geenivaramu andmetele tuginedes on meil teada järgmised andmed:

- geenimarkerid, mis mõjutavad kohvi tarbimist (SNP_1 ehk kohvimarkerid);
- geenimarkerid (geenimarkerite skoor), mis mõjutavad metaboliite Cit, Gp, LDL-C ja Serum-C (SNP_2);
- 8086 indiviidi kohvitarbimisharjumused;
- nende indiviidide eespool nimetatud metaboliitide kontsentratsioon indiviidi kehas;
- tervisenäitajatena süstoolne ja diastoolne vererõhk.

Järgnevate seoste uurimisel eeldame geenimarkerite SNP_1 ja SNP_2 mõjusid vastavalt kohvi tarbimisele ja metaboliitide kontsentratsioonile kehas. Lisaks on kirjeldatud erinevaid tunnuste omavahelisi seoseid, mida soovime lähemalt uurida (kohvi tarbimise mõju metaboliitide kontsentratsioonile indiviidi kehas, metaboliitide mõju vererõhule ning kohvi tarbimise mõju vererõhule). Peale geenimarkerite mõjutab nii indiviidi kohvi tarbimist kui ka metaboliitide kontsentratsiooni kehas hulk tunnuseid, mis on joonisel tähistatud segajatena.

Kõigi kasutavate mudelite puhul, mille tulemused on järgnevates alapeatükkides välja toodud, on kasutusel olnud mudel, kus eksponenttunnus ja väljundnäitaja muutuvad vastavalt uurimiseesmärgile ning taustatunnustena on arvestatud soo, vanuse, kehamassiindeksi ja suitsetamise mõju.

4.2. Eelduste täidetus

Et saada korrektset hinnangut otsitavatele parameetritele ning tagada, et geneetiline muutuja oleks kasutatav IV-muutujana, tuleb järgida teatud eeldusi (kirjeldatud lähemalt peatükis 2.2).

Esimene eeldus – geneetiline muutuja peab olema seotud ekspositsiooninäitajaga – on kolmest ainuke, mida on võimalik statistiliselt kontrollida, kasutades vaatlusandmete tulemusi. On selge, et Mendeli randomiseerimine töötab seda efektiivsemalt, mida tugevam on seos kasutatava IV-muutuja ja ekspositsiooninäitaja vahel.

TÜ geenivaramu andmete põhjal on seosed IV-muutujatena kasutatavate geenimarkerite ja ekspositsiooninäitajate „kohvi tarbimine“ ning „metaboliitide tase“ vahel matemaatilises mõistes küllaltki nõrgad. Kõige tugevam korrelatsiooniseos on metaboliidi LDL-C taseme ja metaboliidile LDL-C vastava SNP_2-skoori vahel ($\rho = 0{,}1725$).

Kohvimarkerite skoori ja kohvi joomise vaheline seos samuti on üsna nõrk ($\rho = 0{,}0898$). Nõrk korrelatsiooniseos võib viidata sellele, et kasutades instrumentidena kohvimarkerite skoori, ei pruugi me saada tugevat tulemust, hinnates kohvi joomise põhjuslikku mõju metaboliitide tasemele ega tugevat seost, hinnates metaboliitide põhjuslikku mõju vererõhule. Arvestada tuleb aga, et geenimarkerid ei määra, kas keegi joob kohvi või mitte, vaid mõjutavad tarbimist. Seega võime need korrelatsiooniseosed lugeda rahuldavateks, eriti teades, et oluliselt tugevamaid seoseid geenimarkerite ja komplekssete fenotüübitunnuste vahel (nt toitumisharjumused) pole leitud.

Veendumusele, et teine eeldus – geneetiline muutuja ei ole seotud ja on sõltumatu segavatest faktoritest – võiks üldjuhul olla täidetud, aitab kaasa teadmine, et Mendeli randomiseerimisel kasutatavad alleelid määratakse juhuvaliku tulemusena juba meioosi käigus. Seega saab geneetilisi muutujaid pidada sõltumatuteks teguriteks ning on võimalik eeldada, et valitud geneetilist muutujat ei mõjuta segajad, mida võid üldjuhul eeldada sellisest riskitegurite-haiguste vahelisest sõltuvusseosest.

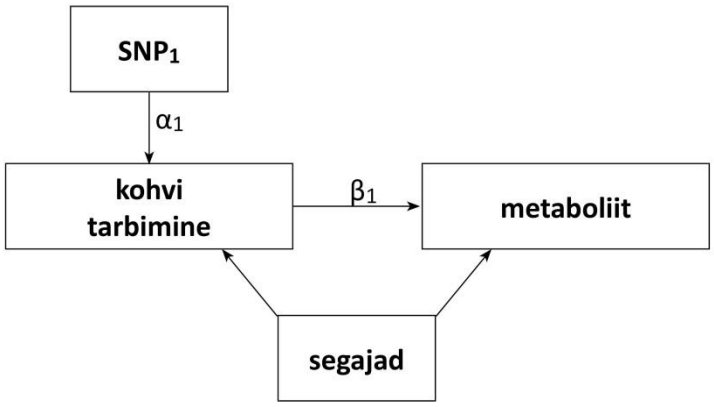
Analoogiliselt teise eeldusega ei saa ka kolmandat eeldust – geneetiline muutuja on sõltumatu väljundnäitajast – kontrollida matemaatiliste meetodite abil. Eelduse kontrollimiseks on vaja teada mõõdetavate tunnuste bioloogilist tausta, et saada eksperthinnanguid sellele, et kui tõenäoline selle eelduse täidetus on.

4.3. Mudelid

4.3.1. Kohvijoomise põhjuslik mõju verelipiididele ja metaboliitidele

Esimese uurimisküsimusena soovime selgitada, milline on kohvijoomise mõju metaboliitidele Cit, Gp, LDL-C ja Serum-C, kasutades instrumendina kohvimarkerite skoori (Joonis 13). Otsime kordajat β_1, millele leiame hinnangu kahel viisil: nii lineaarse regressioonanalüüsi abil kui ka Mendeli randomiseerimise teel. Mendeli randomiseerimise metoodikat kasutades on parameetrit võimalik hinnata, kasutades statistikatarkvara R põhjuslike mõjude hindamiseks välja töötatud paketti *tsls()* (kirjeldatud lähemalt peatükis 2.4.2 ning matemaatiliselt näidatud peatükis 2.4.1).

Lisaks leiame ka kohvimarkerite riskiskoori mõju igale metaboliidile lineaarse regressioonanalüüsi abil. Kasutades kohvimarkerite riskiskoori mõju metaboliidile ning teades kohvimarkerite mõju kohvi tarbimisele (peatükk 3.1.3., Tabel 3; kordaja α_1 Joonisel 13), on meil olemas kaks lineaarsel regressioonanalüüsil saadud hinnangut. Nende jagatis ongi aga parameetri β_1 hinnang Mendeli randomiseerimise kaudu: $\widehat{\beta_1} = \frac{\widehat{\alpha_1 \cdot \beta_1}}{\widehat{\alpha_1}}$.



Joonis 13. Kohvijoomise põhjuslik mõju verelipiididele ja metaboliitidele. Uuritavad seosed koos kordajatega

Uurides kohvijoomise mõju metaboliitide kontsentratsioonile inimese kehas (Tabel 5) annab lineaarne regressioonanalüüs statistiliselt olulised seosed kõigi nelja metaboliidi puhul (p-väärtus < 0.001). Mendeli randomiseerimise korral jäävad parameetrite hinnangute standardvead kordades suuremaks, kui parameetritele endile saadud hinnangud. Tõenäoliselt tuleneb see kasutatavate geneetiliste markerite valikust – juba korrelatsiooniseos kohvimarkerite ning tunnuse „kohvi tarbimine“ vahel jäi suurusjärku 0,1 ning nagu nägime simuleeritud andmeid kasutades (peatükk 2.7) on parameetrite standardvead seda suuremad, mida nõrgem on seos instrumenttunnuse ja väljundnäitaja vahel.

Tabelist 5 näeme ka, et kohvi tarbimise riskiskoor ei ole metaboliitidega seotud. See kinnitab arvamust, et kasutatavate andmete puhul ei anna Mendeli randomiseerimise metoodika statistiliselt olulist tulemust nõrkade instrumenttunnuste tõttu.

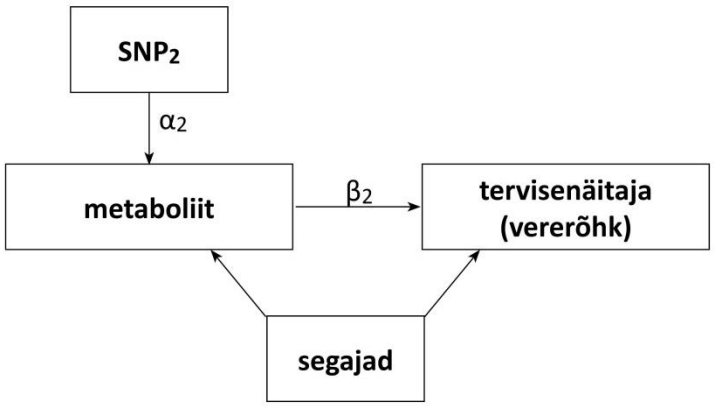
Tulemused lubavad aga oletada, et lineaarsel regressioonanalüüsil nähtav tugev seos ei pruugi olla põhjuslik – seos kohvi joomise ja metaboliidi kontsentratsiooni vahel on tõenäoliselt mõjutatud mingist muust ühisest segavast faktorist.

Tabel 5. Kohvi tarbimise mõju metaboliidile lineaarsel regressioonanalüüsil ning Mendeli randomiseerimise (MR) meetodil (vasakul) koos kohvimarkerite riskiskoori mõjuga metaboliitidele lineaarsel regressioonanalüüsil (paremal)

Metaboliit	Kohvi joomise mõju metaboliidile lineaarsel regressioonanalüüsil (standardveaga)	Kohvi joomise mõju metaboliidile MR meetodil (standardveaga)		Kohvijoomise riskiskoori mõju metaboliidile lineaarsel regressioonanalüüsil (standardveaga)
Cit	–0,11 (0,01)	0,02 (0,12)		0,02 (0,13)
Gp	–0,04 (0,01)	–0,01 (0,11)		–0,01 (0,12)
LDL-C	0,07 (0,01)	–0,01 (0,11)		–0,01 (0,12)
Serum-C	0,07 (0,01)	–0,02 (0,11)		–0,02 (0,11)

4.3.2. Metaboliitide mõju vererõhule

Teise skeemi (Joonis 14) juures soovime selgitada metaboliitide mõju tervisenäitajale vererõhk, kasutades instrumendina metaboliitide markerite skoori. Otsime kordajat β_2, mille hinnangu leiame samuti nii lineaarse regressioonanalüüsi abil kui ka Mendeli randomiseerimise kaudu. Lisaks leiame riskimarkerite skoori SNP_2 otsese mõju tervisenäitajale (nii süstoolsele kui ka diastoolsele vererõhule).



Joonis 14. Metaboliitide mõju vererõhule, kasutades riskiskoori SNP_2. Uuritavad seosed koos kordajatega

Vaadates metaboliitide taseme mõju vererõhule, on taas statistiliselt olulised vaid lineaarsel regressioonanalüüsil saadud seosed (Tabel 6; Tabel 7). Vaadates lineaarse regressioonanalüüsi tulemusi on vaid metaboliidil Cit negatiivne mõju nii süstoolsele kui ka diastoolsele vererõhule. Taaskord ei kinnita aga seose põhjuslikkust Mendeli randomiseerimine – nii metaboliitide mõju vererõhule kui ka metaboliitide riskiskoori mõju vererõhule ei anna meetodi põhjal statistiliselt olulisi hinnanguid.

Sarnaselt eelmise uurimistulemusega näeme Tabelitest 6 ja 7, et metaboliite mõjutavate markerite geneetilised skoorid ei ole tervisenäitajaga „vererõhk“ statistiliselt oluliselt seotud.

Tabel 6. Metaboliitide mõju süstoolsele vererõhule lineaarsel regressioonanalüüsil ning Mendeli randomiseerimise (MR) meetodil (vasakul) koos metaboliite mõjutavate markerite riskiskoori mõjuga süstoolsele vererõhule lineaarsel regressioonanalüüsil (paremal)

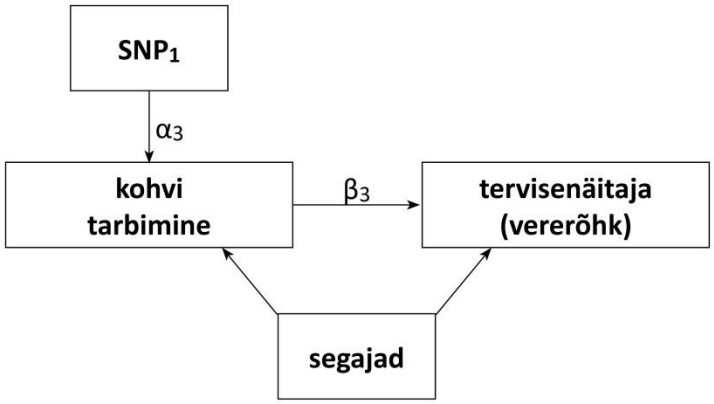
Metaboliit	Metaboliidi mõju süstoolsele vererõhule lineaarsel regressioonanalüüsil (standardveaga)	Metaboliidi mõju süstoolsele vererõhule MR meetodil (standardveaga)		Metaboliidi riskiskoori mõju süstoolsele vererõhule lineaarsel regressioonanalüüsil (standardveaga)
Cit	–0,88 (0,23)	4,02 (10,04)		0,92 (2,20)
Gp	1,25 (0,23)	–1,18 (2,94)		–0,85 (2,10)
LDL-C	1,01 (0,24)	0,86 (1,16)		0,52 (0,71)
Serum-C	1,34 (0,24)	1,15 (1,55)		0,75 (1,02)

Tabel 7. Metaboliitide mõju diastoolsele vererõhule lineaarsel regressioonanalüüsil ning Mendeli randomiseerimise (MR) meetodil (vasakul) koos metaboliite mõjutavate markerite riskiskoori mõjuga diastoolsele vererõhule lineaarsel regressioonanalüüsil (paremal)

Metaboliit	Metaboliidi mõju diastoolsele vererõhule lineaarsel regressioonanalüüsil (standardveaga)	Metaboliidi mõju diastoolsele vererõhule MR meetodil (standardveaga)		Metaboliidi riskiskoori mõju diastoolsele vererõhule lineaarsel regressioonanalüüsil (standardveaga)
Cit	–0,67 (0,15)	–1,67 (6,49)		–0,39 (1,49)
Gp	0,47 (0,16)	–2,92 (2,08)		–2,10 (1,43)
LDL-C	0,89 (0,16)	–0,50 (0,79)		–0,30 (0,48)
Serum-C	1,09 (0,16)	–0,67 (1,06)		–0,44 (0,69)

4.3.3. Kohvi mõju vererõhule

Kolmandaks vaatleme kohvi tarbimise mõju tervisenäitajale „vererõhk“, kasutades instrumendina taas kohvimarkerite skoori SNP_1 (Joonis 15). Otsime kordajat β_3. Sarnaselt eelmisele kahele uurimisküsimusele leiame otsitava kordaja kahel viisil (lineaarse regressioonanalüüsi kaudu ning Mendeli randomiseerimise teel) ning toome välja kohvimarkerite skoori otsese mõju vererõhule.



Joonis 15. Kohvi tarbimise mõju vererõhule, kasutades instrumendina kohvimarkerite riskiskoori. Uuritavad seosed koos kordajatega

Vaadates kohvi joomise mõju vererõhule (Tabel 8), ilmnevad taas statistiliselt olulised seosed vaid lineaarsel regressioonanalüüsil. Seosesuunad, võrreldes lineaarset regressioonanalüüsi ja Mendeli randomiseerimist, on omavahel vastassuundades: lineaarse regressioonanalüüsi põhjal saaksime öelda, et kohvi tarbimine langetab nii süstoolset kui ka diastoolset vererõhku. Mendeli randomiseerimise tulemusi vaadates saame öelda, et mõju ei ole põhjuslik. Seega, kuigi lineaarsel regressioonanalüüsil on näha tugev statistiliselt negatiivne seos, ei saa me järeldada, et kohvi joomine langetaks vererõhku. Ilmnenud tugev seos ei ole põhjuslik. Vaadeldav seos kohvi joomise ja vererõhu vahel tuleneb tõenäoliselt mõnest muust ühisest segavast faktorist tunnuste „kohvi tarbimine“ ja „vererõhk“ vahel.

Tabelist 8 näeme ka, et kohvi joomist mõjutav geenimarkerite skoor ei ole seotud ei süstoolse ega ka diastoolse vererõhuga.

Tabel 8. Kohvi tarbimise mõju süstoolsele ja diastoolsele vererõhule lineaarsel regressioonanalüüsil ning Mendeli randomiseerimise (MR) meetodil (vasakul) koos kohvimarkerite riskiskoori mõjuga süstoolsele ja diastoolsele vererõhule (paremal)

	Kohvi tarbimise mõju vererõhule lineaarsel regressioonanalüüsil (standardveaga)	Kohvi tarbimise mõju vererõhule MR meetodil (standardveaga)		Kohvimarkerite skoori mõju vererõhule lineaarsel regressioonanalüüsil (standardveaga)
Süstoolne vererõhk	–0,61 (0,16)	0,78 (1,75)		0,81 (1,80)
Diastoolne vererõhk	–0,25 (0,11)	1,09 (1,20)		1,13 (1,22)

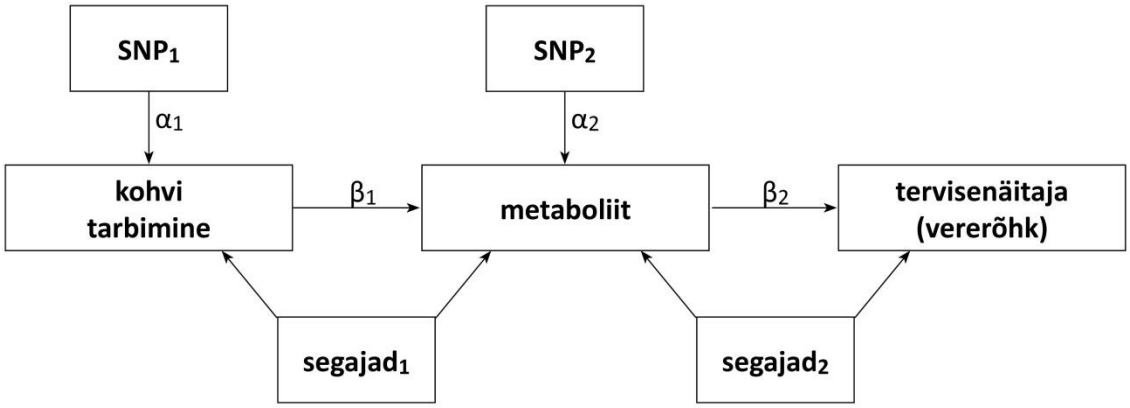
4.3.4. Ühendatud mudel: Mendeli randomiseerimise metoodika edasiarendus

Järgmise uurimisküsimusena soovime teada saada kuidas mõjutab kohvi tarbimine metaboliitide taset ning kuidas mõjutab metaboliitide kontsentratsioon tervisenäitajat. Seega otsime kordajaid β_1 ja β_2. Paneme tähele, et kordajat β_2 on võimalik kätte saada kahel viisil: lähtudes nii kohvimarkeritest kui ka metaboliite mõjutatavatest markeritest.

Seega ühendame eelmistes alapeatükkides (peatükid 4.1.1 ja 4.1.2) kirjeldatud meetodid, laiendades Mendeli randomiseerimise metoodikat ning püüdes üheaegselt uurida kohvi tarbimise mõju metaboliitidele, metaboliitide mõju tervisenäitajatele ning kohvi tarbimise mõju vererõhule läbi metaboliitide taseme.

Kõik kolm uurimisküsimust on korraga näha joonisel 16. Paneme tähele, et neist kaks on läbi tehtud juba eespool: peatükis 4.1.1 vaatlesime kohvi tarbimise mõju metaboliitidele, peatükis 4.1.2 metaboliitide mõju vererõhule läbi geenimarkerite SNP_2. Mendeli randomiseerimise metoodika laiendusena saame vaadelda kohvi tarbimise mõju vererõhule ka läbi metaboliitide, kasutades geenimarkerit SNP_1 ning leida hinnangu metaboliitide mõjust vererõhule läbi geneetiliste markerite SNP_1 skoori. See on kasulik olukorras, kui meil ei oleks infot geneetiliste markerite SNP_2 kohta või kui instrumenttunnuse SNP_2 seos metaboliitidega on väga nõrk.

Lisaks geenimarkerite SNP_1 ja SNP_2 mõjudele vastavalt kohvi tarbimisele ja metaboliitide kontsentratsioonile kehas, eeldame ka kohvi joomise mõju metaboliitide tasemele ning metaboliitide kontsentratsiooni mõju omakorda tervisenäitajatele. Vastavad seosesuunad on välja toodud Joonisel 16.



Joonis 16. Mendeli randomiseerimise metoodika laiendatud skeem. Uuritavad seosed koos kordajatega

Mudeli edasiarenduses, soovides leida hinnangut metaboliitide mõjust tervisenäitajale „vererõhk“, kasutades geneetilise markerina kohvi joomist mõjutavate geenimarkerite skoori, statistiliselt olulisi efekte ei ilmnenud. (Tabel 9)

Tulemus on oodatav, arvestades, et TÜ geenivaramu andmeid kasutades ei leidnud Mendeli randomiseerimist kasutades juba kohvi joomise mõju metaboliitide

kontsentratsioonile kinnitust. Seega võis eeldada, et ka mudeli edasiarendus ei anna soovitud täpsusega tulemusi, mis ka juhtuski: arvestades parameetritele saadud hinnanguid, on kõik standardvead küllaltki suured ning hinnangute usaldusvahemikud sisaldavad kõigi nelja metaboliidi korral nullpunkti.

Seoste suunad Mendeli randomiseerimise edasiarenduse juures, võrreldes metaboliitide mõju süstoolsele ja diastoolsele vererõhule, on samapidised.

Parameetrite hinnangud, vaadates metaboliitide kontsentratsiooni mõju süstoolsele ja diastoolsele vererõhule on tunduvalt erinevad tulemustest, mis saime kordaja hinnanguks, kasutades geneetilist skoori SNP_2 (peatükk 4.3.2.), olles kordades suuremad laiendatud meetodit kasutades.

Tabel 9. Laiendatud Mendeli randomiseerimise metoodika tulemused: metaboliitide kontsentratsiooni mõju süstoolsele ja diastoolsele vererõhule, kasutades kohvimarkerite riskiskoori

Metaboliit	Hinnang süstoolsele vererõhule	Hinnang diastoolsele vererõhule
Cit	33,70 (134,79)	47,07 (335,93)
Gp	–70,92 (101,04)	–99,05 (95,15)
LDL-C	–281,62 (155,33)	–393,35 (308,64)
Serum-C	–45,31 (302,64)	–63,29 (329,37)

4.4. Kokkuvõte tulemustest

Kokkuvõtvalt saame öelda, et kuigi lineaarne regressioonanalüüs näitas statistiliselt olulisi seoseid nii kohvi joomise ja metaboliitide kontsentratsiooni, metaboliitide taseme ja vererõhu kui ka kohvi joomise ja vererõhu vahel, ei ole Mendeli randomiseerimise tulemuste põhjal võimalik väita, et seosed oleksid põhjuslikud. Seega ei ole võimalik väita ei seda, et kohvi joomine vererõhku tõstaks, ega ka mitte seda, et kohvi joomine vererõhku langetaks.

Tõenäoliselt tuleneb see kasutatavate geneetiliste markerite valikust – seosed antud töös kasutatavate instrumenttunnuste ning ekspositsiooninäitajate vahel ei ole väga tugevad. Kuigi geneetikas loetakse selliseid seoseid tugevateks, nägime simuleeritud andmete juures (peatükk 2.7), et parameetrite standardvead on seda suuremad, mida nõrgem on seos instrumenttunnuse ja väljundnäitaja vahel.

Kokkuvõte

Magistritöö eesmärgiks oli tutvuda Mendeli randomiseerimise kui ühe põhjusliku analüüsi meetodiga, uurida meetodi edasiarendamise võimalikkust ning rakendada meetodit reaalsetele andmetele.

Lõputöö raames uuriti põhjuslikke seoseid tunnuste „kohvi tarbimine päevas“, „metaboliitide kontsentratsioon inimese kehas“ ja „vererõhk“ vahel, kasutades Tartu Ülikooli Eesti Geenivaramu andmeid. Valim koosnes 4497-st geenidoonorist. Metaboliitidest oli analüüsi kaasatud neli lipoproteiini (Cit, Gp, LDL-C ja Serum-C), vererõhu puhul vaadeldi eraldi ülemist ja alumist vererõhku.

Koos Mendeli randomiseerimisega vaadeldi uuritavaid seoseid ka tavalise lineaarse regressioonanalüüsi abil. Simuleeritud andmete puhul andis parameetri õigele väärtusele lähedasema hinnangu Mendeli randomiseerimine. Saadud hinnang oli seda täpsem, mida tugevam oli genereeritud seos instrumendina kasutuseloleva geneetilise markeri ja väljundnäitaja vahel. Lineaarne regressioonanalüüs jäi liiga palju sõltuma segavatest mittemõõdetavatest faktoritest ning andis otsitavale parameetrile väära hinnangu.

Reaalseid Tartu Ülikooli Eesti Geenivaramu andmeid kasutades olid kasutatavate instrumenttunnuste seosed eksposits iooninäitajatega küllaltki nõrgad ning võib eeldada, et seetõttu ei näidanud Mendeli randomiseerimine statistiliselt olulisi mõjusid kohvi tarbimise, metaboliitide kontsentratsiooni ja vererõhu vahel. Statistiliselt olulised seosed ilmnesid küll lineaarsel regressioonanalüüsil, kuid ei saa väita, et saadud mõjud oleksid põhjuslikud. Seega ei ole tehtud analüüsi põhjal võimalik väita ei seda, et kohvi joomine vererõhku tõstaks, ega ka mitte seda, et kohvi joomine vererõhku langetaks.

Causal Models based on the Estionan Genome Center metabolomics and nutrition data

Master thesis

Kristi Helekivi

Summary

The purpose of this thesis was to study Mendelian randomization as one causal analysis method, study the feasibility of further development of the method and apply the method to real data.

The thesis aims to examine the causal relationships between traits "daily consumption of coffee", "concentration of metabolites in the human body" and "blood pressure", using data from the Estonian Genome Center. The analysis was carried out using data of 4497 individuals. In the analysis there were included four lipoproteins (Cit, Gp, LDL-C, Serum-C); systolic and diastolic blood pressure were observed separately.

Along with Mendelian randomization studied relationships were also looked in linear regression analysis. For the simulated data the closest estimation to the correct value of parameter was given by Mendelian randomization. The resulting estimate was more accurate the stronger the link between the instrumental variable and outcome variable was generated. Linear regression analysis was too dependent on non-measurable confounding factors and gave an incorrect assessment of the searched parameter.

Using real data from the Estonian Genome Center the relationship between instrumental variables and exposures were relatively weak, and therefore it can be expected why Mendelian randomization did not show statistically significant effects between coffee consumption, concentration of metabolites and blood pressure. Statistically significant relationships occurred during linear regression analysis, but we cannot state that the resulting impacts would be causal. Thus, based on the thesis one cannot say that drinking coffee raises blood pressure or lowers blood pressure.

Kasutatud kirjandus

[1] Möls, M. (2012). *Statistiline seos ja Hii-ruut test.* Loenguslaidid. Tartu: Tartu ülikool, matemaatilise statistika instituut

[2] Põhjuslikkus. Wikipedia, URL (vaadatud 13.03.2015) http://en.wikipedia.org/wiki/Causality

[3] Fischer, K. (2010). *Põhjuslikkus ja statistika.* Loenguslaidid. Tartu: Tartu ülikool, matemaatilise statistika instituut

[4] Üksiku nukleotiidi polümorfism. Wikipedia, URL (vaadatud 11.02.2015) http://et.wikipedia.org/wiki/Üksiku_nukleotiidi_polümorfism

[5] Lewis, SJ. (2010). *Mendelian Randomization as Applied to Coronary Heart Disease, Including Recent Advances Incorporating New Technology.* Department of Social Medicine, University of Bristol, United Kingdom. http://www.ncbi.nlm.nih.gov/pubmed/20160203 (vaadatud 11.02.2015)

[6] Glymour, MM., Tchetgen Tchetgen, EJ., & Robins, JM. (2011) *Credible Mendelian Randomization Studies: Approaches for Evaluating the Instrumental Variable Assumptions.* Harvard School of Public Health, Boston, Massachusetts

[7] Sheehan, AN., Didelez, V., Burton, PR., & Tobin, MD. (2008) *Mendelian Randomisation and Causal Inference in Observational Epidemiology.* Department of Health Sciences, University of Leicester, Leicester, United Kingdom; Department of Mathematics, University of Bristol, Bristol, United Kingdom

[8] Package „SEM". R-project.org, URL (vaadatud 12.03.2015) http://cran.r-project.org/web/packages/sem/sem.pdf

[9] Möls, M. (2012). Monte-Carlo meetodid. Käsikirjaline loengukonspekt. Tartu: Tartu ülikool, matemaatilise statistika instituut

[10] Geenivaramu põhikiri. Tartu Ülikooli Eesti Geenivaramu kodulehekülg, URL (vaadatud 11.02.2013) http://www.geenivaramu.ee/info/ametlik-info/geenivaramu-pohikiri.html

[11] *Estonian Genome Center*. Aastaraamat 2001–2011. Tartu: Estonian Genome Center, University of Tartu 2011, lk 5

[12] Oolo, T., Uiga, B., Usberg, G. (2009). Õendustoimingute töövihik. Tartu

[13] Lipoproteiinid. Wikipedia, URL (vaadatud 11.02.2013) http://en.wikipedia.org/wiki/Lipoprotein

[14] Cornelis, MC., Byrne, EM., Esko, T., …, Chasman DI. (2014) *Genome-wide meta-analysis identifies six novel loci associated with habitual coffee consumption.*

[15] Kettunen, J., Tukiainen, T., Sarin A-P., …, Ripatti, S. (2011) *Genome-wide association study identifies multiple loci influencing human serum metabolite levels*

Lisad

Lisa 1. *Bootstrap*-meetodi simulatsioon

```
library("boot")

a <- 1   # instrumendi (geneetilise markeri) mõju X-tunnusele
b <- 1   # väljundnäitaja; parameeter, mida tahame hinnata
z <- rbinom(10000,2,0.2)  # geneetilise markeri andmed
u <- rnorm(10000)   # segav faktor U
x <- 10 + a*z + 5*u + rnorm(10000,0,5)   # X sõltub instrumendist ja U-st

summary(lm(x~z))

y <- 15 + b*x - 4*u + rnorm(10000,0,5)  # Y sõltub X-st ja U-st

summary(lm(y~x))   # mudel mis annab "vale" tulemuse
summary(lm(y~x+u)) # lineaarne mudel, mis annaks täpse tulemuse, kui meil
oleks reaalses elus andmed segavate faktorite kohta

summary(tsls(y~x,~z))      #      väljundnäitaja      hindamine      Mendeli
randomiseerimisega, kasutades geneetilist markerit instrumendina


## Bootstrap meetodi jaoks andmestiku ja kasutatava mudeli koostamine
andm<-data.frame(x,y,z)

mudel <- function(data,s)
{
  d <- data[s,]
  m1<-lm(x~z,data=d)
  m2<-lm(y~z,data=d)
  m2$coef[2]/m1$coef[2]
}

## test 1 valimiga
i <- sample(1:10000,size=1000,replace=T) #1000-elemeendiline valim
mudel(andm,i)
tsls(y~x,~z,data=andm[i,])

## Bootstrap-meetodi rakendamine
bt <- boot(andm,mudel,500) # arvutab Bootstrap-meetodil standardhälbe
bt
```

Lisa 2. Mendeli randomiseerimise simulatsioon

```
library("boot")

# Geenimarkerite tugevam seos uuritava tunnusega;
# Parameetrite beta1 ja beta2 kordajad erinevad nullist

n <- 2000
beta1 <- 1
beta2 <- 2

andmed <- data.frame(z1=rbinom(n, 2, 0.4), z2=rbinom(n, 2, 0.3),
u=rnorm(n), v=rnorm(n))

andmed$x <- with(andmed, -2 + 2*z1 - 4*u + rnorm(n))
andmed$y <- with(andmed, -10 + 3*z2 + beta1*x + u + 2*v + rnorm(n))
andmed$w <- with(andmed, 5 + beta2*y + 3*v + rnorm(n))

nsim <- 1000
out <- matrix(NA, nrow=nsim, ncol=6)


# Kontrollid:
summary(lm(y~x,data=andmed))  # parameetri kordaja pole ligikaudu 1 (beta1)
summary(lm(w~y,data=andmed)) # parameetri kordaja pole ligikaudu 2 (beta2)


mudel1 <- function(data, s)
{
  d  <- data[s,]
  gamma1 <- lm(y ~ z1, data=d)$coef[2]
  alfa1 <- lm(x ~ z1, data=d)$coef[2]
  beeta1 <- gamma1 / alfa1
 beeta1
}

mudel1(andmed,1:n) #kontroll


mudel2 <- function(data, s)
{
  d  <- data[s,]
  gamma1 <- lm(y ~ z1, data=d)$coef[2]
  alfa1 <- lm(x ~ z1, data=d)$coef[2]
  beeta1 <- gamma1 / alfa1
  gamma2 <- lm(w ~ z1, data=d)$coef[2]
  beeta2 <- gamma2 / (alfa1 * beeta1)
beeta2
}

mudel2(andmed,1:n) #kontroll
```

```
mudel3 <- function(data, s)
{
  d  <- data[s,]
  gamma3 <- lm(w ~ z2, data=d)$coef[2]
  alfa2 <- lm(y ~ z2, data=d)$coef[2]
  beeta2 <- gamma3 / alfa2
beeta2
}

mudel3(andmed,1:n) #kontroll


## tsükkel
for (i in 1:nsim) {
  andmed <- data.frame(z1=rbinom(n, 2, 0.4), z2=rbinom(n, 2, 0.3),
u=rnorm(n), v=rnorm(n))

  andmed <- within(andmed,
{
  x <- -2 + 2*z1 - 4*u + rnorm(n)
  y <- -10 + 3*z2 + beta1*x + u + 2*v + rnorm(n)
  w <- 5 + beta2*y + 3*v + rnorm(n)
})

bt <- boot(andmed, mudel1, 1000)
beeta1 <- bt[1]$t0
sd1 <- sd(bt$t)

bt <- boot(andmed, mudel2, 1000)
beeta2 <- bt[1]$t0
sd2 <- sd(bt$t)

bt <- boot(andmed, mudel3, 1000)
beeta2.2 <- bt[1]$t0
sd2.2 <- sd(bt$t)

out[i, 1] <- beeta1
out[i, 2] <- beeta2
out[i, 3] <- beeta2.2
out[i, 4] <- sd1
out[i, 5] <- sd2
out[i, 6] <- sd2.2
}

#out
colMeans(out) # Standardvigade hinnangute keskmine

sd(out[,4]) # Standardvead üle kõigi simuleeritud beetade
sd(out[,5])
sd(out[,6])


# Võrdluseks funktsiooni tsls tulemused:
library(„sem“)
summary(tsls(y ~ x, ~z1, data=andmed))
summary(tsls(w ~ y, ~z2, data=andmed))
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Kristi Helekivi** (sünnikuupäev: 29.07.1991)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Põhjuslikud mudelid Tartu Ülikooli Eesti Geenivaramu metaboloomika ja toitumise andmetel“, mille juhendaja on Tartu Ülikooli Eesti Geenivaramu vanemteadur Krista Fischer (PhD),

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 13.05.2015