TARTU ÜLIKOOL

MATEMAATIKA-INFORMAATIKATEADUSKOND

MATEMAATILISE STATISTIKA INSTITUUT

Oliver Aasmets

Geneetiliste mõjude hindamine kinnitava faktoranalüüsiga

Bakalaureusetöö (9 EAP)

Juhendaja:

Krista Fischer, PhD

TARTU

2015

Geneetiliste mõjude hindamine kinnitava faktoranalüüsiga

Genoomikapõhise personaalse meditsiini väljatöötamiseks soovitakse inimese genotüübiandmete põhjal ennustada haiguste tekkimise riske. Geneetiliste mõjude hindamisel kasutatakse enim ühenukleotiidsete polümorfismide (SNP) markereid, mis on inimese geneetilise varieeruvuse põhilisemaid avaldumisviise. DNA- ahelal lähestikku paiknevad SNP-d on omavahel tugevasti korreleeritud, seetõttu kasutatakse geeni mõju hindamisel enamasti ainult piirkonna kõige olulisemat markerit.

Käesoleva bakalaureusetöö eesmärk on anda ülevaade struktuurivõrrandite mudelitest ning rakendada metoodika ühte erijuhtu- kinnitavat faktoranalüüsi, hindamaks geenipiirkonna mõju, kasutades kõiki piirkonnas mõõdetud geneetilisi markereid.

Märksõnad: *struktuurianalüüs, faktoranalüüs, ühenukleotiidsed polümorfismid*

Evaluation of the genetic effects using confirmatory factor analysis

The aim of the genomics-based personal medicine is to predict the risk of occurrence of the disease using the human genome data. For assessing the genetic effects single nucleotide polymorphism (SNP) markers are most commonly used, which are one of the most basic manifestations of human genetic variation. The SNP-s that are located close to each other on the DNA chains are strongly correlated with each other, therefore only the most important marker of the gene region is used in assessing the effects of the gene.

The purpose of this thesis is to provide an overview of the structural equation modelling (SEM) and to apply special case of SEM- confirmatory factor analysis in order to evaluate the effects of a gene region using all genetic markers measured in that region.

Keywords: *structural analysis*, *factor analysis*, *single nucleotide polymorphisms*

Sisukord

Sissejuhatus

Tartu Ülikooli Eesti Geenivaramu üks eesmärkidest on leida seoseid indiviidi geneetilise materjali ja avaldunud tunnuste vahel. Saadud tulemusi soovitakse rakendada personaalses meditsiinis, mis tähendab seda, et inimese geneetilisi andmeid kasutatakse haiguse tekkimise riski hindamisel ning seejärel on võimalik rakendada ennetusmeetmeid või personaalset ravi. Kõige sagedamini kasutatakse geneetiliste mõjude hindamisel suurte ülegenoomsete assotsiatsiooniuuringute tulemusena raporteeritud sõltumatuid geenimarkereid, mille seos haigusega on osutunud genotüübis olulisemaks.

Käesoleva töö eesmärk on võtta mõjude hindamisel kasutusele lisaks nimetatud olulisematele markeritele ka neile genoomis lähedal paiknevad markerid, millega nad on korreleeritud. Töös hinnatakse viie geenipiirkonna mõju teist tüüpi diabeedile ning teist tüüpi diabeeti inditseerivale veresuhkru tasemele.

Töö jaguneb kaheks suuremaks peatükiks. Esimeses peatükis antakse ülevaade töö teoreetilisest osast: tutvustatakse üksiknukleotiidsete polümorfismide kui geneetilise varieeruvuse põhilise avaldumisviisi olulisust geneetilistes uuringutes ning käsitletakse struktuurivõrrandite mudelite teoreetilist tausta, millest põhjalikumalt keskendutakse meetodi erijuhule- kinnitavale faktoranalüüsile.

Bakalaureusetöö teine osa on praktiline. Praktilises osas hinnatakse Tartu Ülikooli Eesti Geenivaramu genotüübiandmetele kinnitava faktoranalüüsi mudel. Lisaks kasutatakse lineaarset ja logistilist regressioon, hindamaks ülegenoomsete assotsiatsiooniuuringute tulemusena raporteeritud olulisemate sõltumatute markerite mõju diabeedi esinemisele ja veresuhkru tasemele uuritavas valimis. Lõpuks hinnatakse lineaarse ja logistilise regressiooni mudelid ka kinnitava faktoranalüüsi abil saadud faktorite skooridele ning võrreldakse tulemusi vaid olulisemaid sõltumatuid markereid kasutavate mudelitega.

Töö praktilise osa läbiviimiseks on kasutatud statistikatarkvara R, struktuurivõrrandite mudelite jaoks kasutati paketti „lavaan“. Töö on vormistatud kasutades tekstitöötlustarkvara Microsoft Word 2013.

Autor soovib tänada töö juhendajat Krista Fischerit rohkete nõuannete ja suunamiste eest geneetika ning struktuurivõrrandite mudelite osas. Samuti on soov tänada Märt Mölsi abistavate selgituste eest paketiga „lavaan“ töötamiseks.

1 Teoreetiline osa

1.1 Geneetika alusmõisted

Organismid koosnevad rakkudest. Rakutuumas paiknevad kromosoomid, mis koosnevad valdavalt kahte tüüpi keemilistest molekulidest, milleks on valgud ja nukleiinhapped. Nukleiinhappeid on kahte tüüpi: RNA ja DNA, millest DNA-s säilitatakse geneetiline informatsioon. DNA on polümeer, mis koosneb nukleotiididest. Nukleotiidid koosnevad omakorda fosfaatgrupist, viiesüsinikulisest suhkrust ning lämmastikalusest, milleks võib olla: adeniin (A), guaniin (G), tümiin (T) ning tsütosiin (C).

Geeniks nimetatakse DNA segmenti, mis määrab organismis mingi elementaartunnuse tekke. Fenotüübiks nimetatakse indiviidil avaldunud tunnuste kogumit, mis on määratud indiviidi genotüübi ja keskkonnamõjude koostoimes. (Heinaru, 2012)

Inimese geneetilise varieeruvuse põhiliseks avaldumisviisiks on üksiknukleotiidsed polümorfismid ehk SNP-d (Kim & Misra, 2007). SNP- ga on tegu juhul, kui DNA järjestuses on asendunud üks nukleotiid teisega. Näiteks on tegu SNP-ga kui kahe erineva isiku DNA fragmendid on vastavalt C**T**A ja C**C**A. Enamasti on SNP-del kaks erinevat esinemise vormi ehk alleeli. See tähendab, et geenipositsioonil, kus SNP esineb, esineb populatsioonis kaks erinevat nukleotiidi. (Heinaru, 2012)

SNP-de on seotud mitmete haiguste tekkega: näiteks on otseselt SNP-de põhjustatud haigused laktoositalumatus ja hemofiilia ehk veritsustõbi. Lisaks näitavad SNP-d soodumusi teatud haiguste tekkeks ning samuti võivad SNP-d määrata, kuidas reageerib inimese organismi kemikaalidele, ravimitele ja vaktsiinidele ning millised võivad olla nende kõrvaltoimed. (Heinaru, 2012; Carlson, 2008)

1.2 Struktuurivõrrandite mudelitest

Peatükk 1.2. põhineb Kenneth A. Bolleni 1989. aastal kirjutatud raamatul „Structural Equation Modelling with Latent Variables“.

1.2.1 Struktuurvõrrandite mudelite kontseptsioon

Struktuurivõrrandite mudelid (SEM) on statistiliste meetodite kogum, mille kontseptsioon põhineb latentsete ehk mittemõõdetavate tunnuste analüüsil. Näiteks on

struktuurivõrrandite erijuhud nii regressioonanalüüs, dispersioonanalüüs kui ka kovariatsioonianalüüs. Erinevus seisneb asjaolus, et mõõdetud ja prognoositud väärtuste erinevuste funktsioonide minimiseerimise asemel minimiseeritakse valimi kovariatsioonide ja mudeli poolt prognoositud kovariatsioonide vahe.

Struktuurivõrrandite mudelite eesmärk on leida latentsete tunnuste abil struktuur, mis kirjeldaks võimalikult hästi ära valimi kovariatsioonimaatriksi. Meetodi peamine hüpotees on, et mõõdetud tunnuste kovariatsioonimaatriks avaldub teatud stuktuuriparameetrite funktsioonina ning seega on korrektset mudelit ning nimetatud parameetreid teades võimalik täpselt reprodutseerida üldkogumi kovariatsioonimaatriks. Lisaks on eesmärk saada teada, kuidas ja mis ulatuses on mõõdetud tunnused seotud latentsete tunnustega: kui tugev on regressioonseos latentsete tunnuste ja mõõdetud tunnuste vahel.

Käesolevas töös on latentseteks tunnusteks viie geenipiirkonna summaarsed mõjud, mida hindavad omavahel korreleeritud SNP markerid, mis on määratud ülegenoomse geenikiibi „Illumina CardioMetabochip“ abil.

Alljärgnevas töös keskendutakse enam struktuurivõrrandite erijuhule- kinnitavale faktoranalüüsile. Kinnitava faktoranalüüsi puhul keskendutakse latentsete tunnuste ja mõõdetud tunnuste vahelistele seostele, latentsete tunnuste omavahelist seost ei uurita. Edasises nähtub, et kinnitava faktoranalüüsi mudel on struktuurivõrrandite mudelite erijuht.

Nii SEM kui ka kinnitav faktoranalüüs eeldavad, et on teada informatsioon latentsete tunnuste olemasolu kohta. Informatsioon võib põhineda kas teoorial, empiirilisel uuringul või mõlemal. Esmalt postuleeritakse mõõdetud tunnuste ja latentsete tunnuste vahelised seosed ning seejärel kontrollitakse määratud struktuuri sobivust statistiliselt.

1.2.2 Mudeli kuju

Üldine struktuurivõrrandite mudel jaguneb kaheks: mõõtmismudel *(measurement model)* ja struktuuri- ehk latentsete tunnuste mudel. Mõõtmismudelis on struktuurivõrrandid, mis esindavad seost latentsete- ja mõõdetud tunnuste vahel. Struktuurimudel näitab seoseid latentsete tunnuste vahel. Nii struktuurimudelis kui ka mõõtmismudelis on kasutusel tunnuste hälbimused nende keskmisest.

Lisaks jagatakse SEM puhul latentsed ehk mittemõõdetud tunnused kaheks: endogeensed ja eksogeensed. Eksogeensed ehk sõltumatud latentsed tunnused on välistekkelised: nad ei sõltu mudelisiseselt teistest tunnustest. Endogeensed ehk sõltuvad latentsed tunnused on määratud mudelis olevate tunnuste poolt: nad on sõltuvad mõnest teisest mudelis olevast tunnusest.

Struktuurivõrrandide üldkuju on:

1) Struktuurimudel latentsetele tunnustele:

$$\eta = B\eta + \Gamma\xi + \zeta$$

2) Mõõtmismudel eksogeensete latentsete tunnustega seotud indikaatortunnustele:

$$x = \Lambda_x\xi + \delta$$

3) Mõõtmismudel endogeensete latentsete tunnustega seotud indikaatortunnustele:

$$y = \Lambda_y\eta + \varepsilon$$

Kus struktuurimudelis:

- η on $m \times 1$- mõõtmeline endogeensete latentsete tunnuste vektor,
- ξ on $n \times 1$- mõõtmeline eksogeensete latentsete muutujate vektor,
- Γ on $m \times n$- mõõtmeline eksogeensete latentsete tunnuste koefitsientide maatriks,
- B on $m \times m$- mõõtmeline endogeensete latentsete tunnuste koefitsientide maatriks,
- ζ on $m \times 1$- mõõtmeline vektor, mis representeerib vigu võrdustes, mis seovad η ja ξ,
- Φ on $n \times n$- mõõtmeline eksogeensete latentsete tunnuste kovariatsioonimaatriks,
- Ψ on $m \times m$- mõõtmeline vigade korrelatsioonmaatriks.

Muutujad mõõtmismudelis:

- x on $q \times 1$- mõõtmeline vektor eksogeensetest latentsete tunnuste indikaatoritest,
- y on $p \times 1$- mõõtmeline vektor endogeensete latentsete tunnuste indikaatoritest,
- ξ on $n \times 1$- mõõtmeline eksogeensete latentsete muutujate vektor,
- η on $m \times 1$- mõõtmeline endogeensete latentsetest muutujate vektor,
- δ on $q \times 1$- mõõtmeline tunnuse x vigade vektor,
- ε on $p \times 1$- mõõtmeline tunnuse y vigade vektor,
- Λ_x on $q \times n$- mõõtmeline maatriks, mis seob n eksogeenset latentset tunnust q tunnusega, mis latentseid tunnuseid eeldatavasti mõõdavad,

- Λ_y on $p \times m$- mõõtmeline maatriks, mis seob m endogeenset latentset tunnust p tunnusega, mis latentseid tunnuseid eeldatavasti mõõdavad,
- $\Theta_\delta = E(\delta\delta')$ on q × q mõõtmeline y- tunnuste mõõtmisvigade kovariatsioonimaatriks,
- $\Theta_\varepsilon = E(\varepsilon\varepsilon')$ on p × p mõõtmeline x- tunnuste mõõtmisvigade kovariatsioonimaatriks.

Maatriksite Λ_x ja Λ_y elementideks on faktorkaalud ehk regressioonikordajad, mis näitavad, kui palju muutub mõõdetud tunnus ühikulise latentse tunnuse muutuse korral.

1.2.3 Mudeli eeldused

Struktuurimudeli eeldused:

- $E(\eta) = 0$,
- $E(\xi) = 0$,
- $E(\eta) = 0$,
- ζ pole korreleeritud ξ-ga,
- $(I - B)$ pole singulaarne ehk on pööratav.

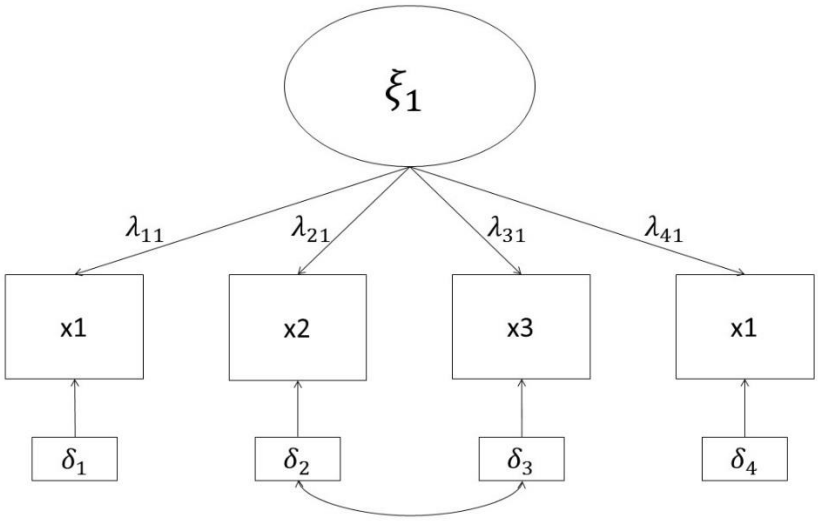
Mõõtmismudeli eeldused:

- $E(\delta) = 0$,
- $E(\varepsilon) = 0$,
- ε ei ole korreleeritud ξ, η ja δ-ga.

1.2.4 Teeanalüüs *(Path analysis)*

Teeanalüüs kujutab endast mudeli graafilist esitlust, milles kujutatakse uurija poolt paika pandud võrrandite süsteemi. Mõõdetud tunnused kujutatakse joonisel ristkülikutena, latentsed tunnused ringide või ovaalidena. Vead või segavad faktorid kujutatakse joonisel ilma ümbriseta. Sirged ühesuunalised jooned kujutavad endast põhjuslikke seoseid tunnuste vahel, kahesuunalised kõverjooned kujutavad tunnustevahelist sõltuvust.

Teeanalüüs aitab kirjeldada kahe tunnuse vahelist kovariatsiooni või korrelatsiooni mudeli parameetrite funktsioonina. Näide mõõtmismudelist, kus ühel faktoril on neli indikaatorit, on kujutatud Joonisel 1.



Joonis 1 Teeanalüüsi näide

Selle mõõtmismudeli esituse põhjal $x_i = \lambda_{ij}\xi_j + \delta_i$ ja seega saab joonisel kujutatud diagrammi abil avaldada $COV(x_1, x_4)$:

$$COV(x_1, x_4) = COV(\lambda_{11}\xi_1 + \delta_1, \lambda_{41}\xi_1 + \delta_4) = \lambda_{11}\lambda_{41}\mathrm{COV}(\xi_1, \xi_1) = \lambda_{11}\lambda_{41}\Phi_{11}$$

Seega $COV(x_1, x_4)$ on latentse tunnuse ξ_1 ja mõõdetud tunnuste x_1 ja x_4 vaheliste seosekordajate ja latentse tunnuse ξ_1 dispersiooni funktsioon.

1.3 Kinnitav faktoranalüüs *(Confirmatory factor analysis)*

Kinnitav faktoranalüüs (CFA) kujutab ainult ühte võrrandit struktuurivõrrandite süsteemist. Tegu on struktuurivõrrandite mudelite erijuhuga, mille puhul uuritakse ainult mõõtmismudelit ehk mudelit, mis seob mõõdetud tunnused latentsete tunnuste ehk faktoritega.

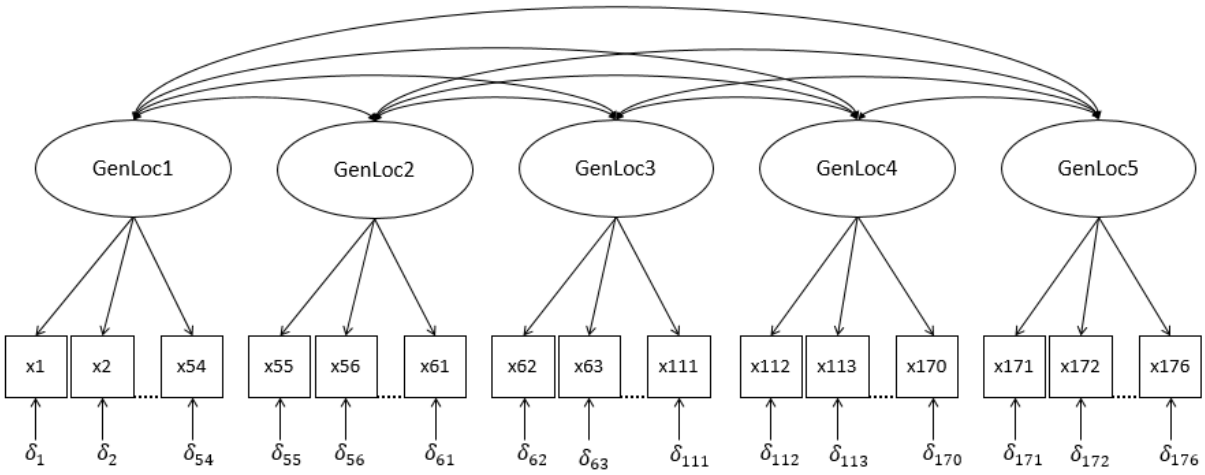
Kinnitava faktoranalüüsi puhul jagatakse analüüsi läbiviimine viieks etapiks:

- Mudeli kirjeldamine teeanalüüsi diagrammi abil,
- mudeli identifitseeritavuse kontrollimine,
- mudeli parameetrite hindamine,
- mudeli headuse hindamine,
- mudeli täpsustamine.

1.3.1 Teeanalüüs

Käesolevas bakalaureusetöös uuritakse viie geenipiirkonna mõjusid. Esialgses mudelis on seega viis latentset tunnust, mille tähistusena kasutatakse „GenLoc1“, „GenLoc2“,... ja

„GenLoc5“. Nimetatud piirkonnad on ebavõrdse suurusega – viiele piirkonnale vastavad vastavalt 54, 7, 49, 59 ja 6 geenimarkerit. Ühes piirkonnas asuvad markerid on eelduse kohaselt seotud ühe faktoriga ja on omavahel korreleeritud. Samuti eeldatakse mudeli üldkujus, et kõik latentsed tunnused võivad olla paarikaupa korreleeritud. Kirjeldatud mudelile vastav teediagramm, kus geenimarkereid tähistatakse $x_1 \dots x_{176}$ ja vigu $\delta_1 \dots \delta_{176}$, on toodud Joonisel 2.



Joonis 2 Teediagramm kinnitavale faktoranalüüsile

1.3.2 Mudeli kuju

Kinnitava faktoranalüüsi puhul on mudeli kuju esitamiseks kaks järgnevat võimalust lähtuvalt struktuurivõrrandite mudeli esitusest:

$$x = \Lambda_x \xi + \delta \quad (1.1)$$

$$y = \Lambda_y \eta + \varepsilon \quad (1.2)$$

kus x ja y on mõõdetud tunnused, ξ ja η on latentsed tunnused ja δ ja ε on mudeli vead. Mudelid (1.1) ja (1.2) on kinnitava faktoranalüüsi jaoks samaväärsed. Mõõdetud tunnused sõltuvad ühest või enamast latentsest tunnusest. Edaspidi lähtutakse mudeli esitusest kujul (1.1).

Seosekordajad, mis kirjeldavad latentsete tunnuste mõju mõõdetud tunnustele, asuvad maatriksites Λ_x. Iga $x_i = \lambda_{ij}\xi_j + \delta_i$ jaoks on λ_{ij} arv, mis näitab, mitu ühikut x_i muutub, kui latentne tunnus ξ_j muutub ühiku võrra. Kui mitu latentset tunnust ξ mõjutavad tunnust x_i, siis λ_{ij} on oodatud muutus latentse tunnuse ühikulise muutuse korral, kui teised latentsete tunnuste väärtused jäävad samaks.

Mudelile kehtivad eeldused, et vigade keskväärtus on 0 ning vead on latentsetest tunnustest sõltumatud:

- $E(\delta) = 0$
- $E(\xi\delta^t) = 0$

1.3.3 Mudeli kovariatsioonimaatriks

Struktuurvõrrandite mudelite, k.a kinnitava faktoranalüüsi lahendamine põhineb kovariatsioonstruktuuride analüüsil. Olgu mõõdetud tunnuste x kovariatsioonimaatriks üldkogumis Σ. Kovariatsiooni struktuuri kohta käiv nullhüpotees on:

$$\Sigma = \Sigma(\theta)$$

kus $\Sigma(\theta)$ on kovariatsioonimaatriks, mis on esitatud mudeli vabade parameetrite θ funktsioonina. Võrdus nõuab, et iga üldkogumi kovariatsioonimaatriksi element on avaldatav ühe või mitme hinnatava mudeli parameetri funktsioonina.

Et tundmatud x on hälbed neile vastavatest keskmistest, siis x kovariatsioonimaatriks on võrdne xx' ooteväärtusega. Tunnuste x üldkogumi kovariatsioonimaatriks $\Sigma(\theta)$ avaldub parameetrite θ funktsioonina kujul:

$$\begin{aligned}\Sigma(\theta) &= E(xx^t)\\ &= E[(\Lambda_x\xi + \delta)(\Lambda^t{}_x\xi^t + \delta^t)]\\ &= \Lambda_x E(\xi\xi^t)\Lambda^t{}_x + \Theta_\delta\\ &= \Lambda_x\Phi\Lambda^t{}_x + \Theta_\delta \qquad (1.3)\end{aligned}$$

Võrdus (1.3) näitab, et x kovariatsioonimaatriksi $\Sigma(\theta)$ saab avaldada latentsete tunnuste ξ kovariatsioonimaatriksi Φ, mõõtmisvigade kovariatsioonimaatriksi Θ_δ ja faktorkaalude maatriksi Λ_x abil. Võrdsustades Σ ja $\Sigma(\theta)$ vastavad elemendid, saadakse, et üldkogumi dispersioonid ja kovariatsioonid maatriksis Σ avalduvad mõõtmismudeli struktuuriparameetrite *(structural parameters)* funktsioonina.

1.3.4 Mudeli identifitseeritavus

Mudeli identifitseeritavuse küsimus tekib parameetrite hindamisel: kas parameetrite hinnang on ühene. Kinnitava faktoranalüüsi jaoks on küsimus, kas struktuuriparameetrite Λ_x, Φ ja Θ_δ hindamiseks leidub ühene lahend.

Vektor θ mõõtmetega $t\ x\ 1$ sisaldab kõiki mitteteadaolevaid mudeli parameetreid. Teadaolevate *(identified)* parameetrite all mõeldakse üldkogumi parameetreid, mille hindamise jaoks on valimisuurused olemas, näiteks tunnuste dispersioonid. Tundmatute parameetrite all mõistetakse parameetreid, mida ei teata olevat identifitseeritud: näiteks kovariatsioonid, mille olemasolu eelnevalt ei ole teada. Identifitseeritavust näidatakse sellega, et demonstreeritakse, et tundmatud parameetrid on teadaolevate parameetrite funktsioonid.

Parameetrid vektoris θ on identifitseeritud, kui kahe vektori θ_1 ja θ_2 korral $\Sigma(\theta_1) = \Sigma(\theta_2)$ siis ja ainult siis kui $\theta_1 = \theta_2$. Kui mitu erinevat parameetrite vektorit viivad samade mudelipõhiste kovariatsioonimaatriksiteni *(implied covariance matrix)*, siis ei ole mudel identifitseeritud.

Hinnata tuleb kõiki mudeli struktuurikordajaid, mis seovad indikaator- ja latentseid tunnuseid, latentsete tunnuste dispersioone ja korrelatsioonikordajaid, mis seovad latentseid tunnuseid. Et mudeli hindamine toimub kovariatsioonimaatriksi abil, siis m indikaatortunnuse korral on lähtesuurusi ehk kõikide tunnuste vahelisi kovariatsioone kokku $m(m+1)/2$.

Selleks, et mudel oleks identifitseeritav, peab olema mudeli vabadusastmete arv positiivne. Vabadusastmete arv avaldub kujul:

$$df = \frac{m(m+1)}{2} - \ (mudeli\ parameetrite\ arv - \ kitsenduste\ arv\ parameetritele)$$

Kus mudeli parameetrite arv, millest on lahutatav kitsenduste arv, on kokku kõikide hinnatavate parameetrite arvuks. Kitsenduste sissetoomine vähendab hinnatavate parameetrite arvu. (Traat, 2014)

1.3.5 Mudeli parameetrite hindamine

Analüüsiks on kasutada valimi kovariatsioonimaatriks S, mille põhjal arvutatakse hinnangud mudeli parameetritele. Eesmärk on leida parameetritele väärtused, mis viivad mudelipõhise kovariatsioonimaatriksi $\hat{\Sigma} = \Sigma(\hat{\theta})$, kus $\hat{\theta}$ on hinnatud parameetrite vektor, valimi kovariatsioonimaatriksile S nii lähedale kui võimalik. Maatriksite $\hat{\Sigma}$ ja S läheduse hindamiseks on defineeritud mitmeid funktsioone, mille minimiseerimine annab hinnangu parameetervektorile θ.

Nimetatud parameetrite hindamise funktsioonidel *(fitting functions)* $F(S, \Sigma(\theta))$ on järgnevad omadused:

- $F(S, \Sigma(\theta))$ on skalaar,
- $F(S, \Sigma(\theta)) \geq 0$,
- $F(S, \Sigma(\theta)) = 0$ ainult siis, kui $\Sigma(\theta) = S$,
- $F(S, \Sigma(\theta))$ on pidev

Peamised parameetrite hindamiseks kasutatavad funktsioonid on:

1) Suurima tõepära meetod:

$$F_{ML} = log|\Sigma(\theta)| + tr[S\Sigma^{-1}(\theta)] - log|S| - q$$

Kus $tr[S\Sigma^{-1}(\theta)]$ on maatriksi $S\Sigma^{-1}(\theta)$ peadiagonaali elementide summa ehk maatriksi jälg. Eeldatakse, et $\Sigma(\theta)$ ja S on positiivselt määratud ehk nad ei ole singulaarsed. Funktsiooni minimiseerimiseks kasutatakse üldiselt numbrilisi meetodeid, täpsed lahendid on võimalik leida vaid teatud juhtudel. Lisaks eeldatakse, et kõik indikaatortunnused on normaaljaotusega.

2) Kaalutumata vähimruutude meetod:

$$F_{ULS} = \frac{1}{2} tr\{[S - \Sigma(\theta)]^2\}$$

Funktsiooniga F_{ULS} minimiseeritakse jääkide maatriksis $(S - \Sigma(\theta))$ iga elemendi ruutude summa. Maatriks $(S - \Sigma(\theta))$ koosneb valimi kovariatsioonide ja vastavate mudeli poolt prognoositud kovariatsioonide vahest.

3) Üldistatud vähimruutude meetod:

$$F_{GLS} = \frac{1}{2} tr\{[I - \Sigma(\theta)S^{-1}]^2\}$$

1.3.6 Mudeli headuse näitajad

Mudeli hindamiseks kasutatavad näitajad jagunevad mudeli headuse näitajateks ning mudeli komponentide headuse mõõdikuteks. Mudeli headuse näitajad jagunevad omakorda kaheks: absoluutsed indeksid ning võrdlevad indeksid.

1) Absoluutsed indeksid

Kovariatsiooni struktuuri kohta käiv nullhüpotees on, et $\Sigma = \Sigma(\theta)$. Üldised headuse näitajad aitavad hüpoteesi kinnitada või aitavad hinnata, kui palju Σ erineb $\Sigma(\theta)$-st.

Põhiline test hindamaks mudeli sobivust, on hii-ruut test, mis kasutab valimi kovariatsioonimaatriksi S ja selle mudelipõhise hinnangu $\hat{\Sigma}$ elementide vahede ruutude

summat. Hii-ruut test sobib aga kasutamiseks struktuurivõrrandite mudelite kohta väikeste valimite jaoks: kui valimimaht on juba üle 400 vaatluse, loetakse testi halvaks. Heaks loetakse testi võimekust hinnata mudeli sobivust, kui vaatlusi on aga alla 200. (Kennedy, 2014)

Enimkasutatavaks indeksiks on lähenduse keskmine ruutviga (RMSEA):

$$RMSEA = \sqrt{\frac{\chi^2 - df}{(n-1)df}}$$

Kus χ^2 on hii-ruut statistik vabadusastmetega df. Heaks loetakse mudelit, mille RMSEA on alla 0,1 (Kennedy, 2014).

Heaks indikaatoriks loetakse veel keskmist ruutviga (RMR) ja standardiseeritud keskmist ruutviga (SRMR). Indeksi RMR väärtuseks on ruutjuur valimi kovariatsioonimaatriksi ja mudelist prognoositud kovariatsioonimaatriksi elementide vahede ruutude keskväärtusest:

$$RMR = \sqrt{2\sum_{i=1}^{m}\sum_{j=1}^{i}\frac{(s_{ij} - \hat{s}_{ij})^2}{m(m+1)}}$$

Hea mudeli korral RMR < 0,06 ja SRMR < 0,08. (Kennedy, 2014)

Teised enamlevinud absoluutsed indeksid põhinevad indeksil *Goodness-of-Fit Index* (GFI), kuid nende puhul tuuakse välja, et nende hinnangud on liialt mõjutatud valimi suurusest. (Kennedy, 2014)

2) Võrdlevad indeksid

Need indeksid võrdlevad mudelit teatud baasmudeliga, milleks on mudel, kus puuduvad igasugused sõltuvused tunnuste vahel. Baasmudeli korral on vabaduastmete arvuks:

$$df_b = \frac{m(m-1)}{2} - m$$

Põhiline võrdlev indeks, mida mudeli headuse hindamisel kasutatakse, on võrdlev headuse näitaja (*comparative fit index*- CFI):

$$CFI = \frac{|({\chi_b}^2 - df_b) - ({\chi_m}^2 - df_m)|}{|{\chi_b}^2 - df_b|}$$

Hea mudeli korral loetakse CFI > 0,95. Samas märgitakse, et CFI hinnang sõltub valimi keskmisest korrelatsioonist, mis võib indeksi väärtust vähendada. (Kennedy, 2014)

Mudelite võrdlemiseks sobib ka Akaike informatsioonikriteeriumit (AIC). AIC omab mõtet ainult siis, kui võrrelda kahte mudelit, üldise mudeli headuse näitajana teda kasutada ei saa. Kriteerium AIC arvutatakse kujul:

$$AIC = \chi^2 + m(m+1) - 2df$$

Kus m on mudeli parameetrite arv ning df on mudeli vabadusastmete arv. Väiksem AIC väärtus inditseerib mudeli paremat sobivust.

3) Komponentide headuse näitajad *(component fit measures)*

Individuaalsete parameetrite hindamisel võivad tekkida ebaloogilised tulemused, mis võivad jääda tähelepanuta, kui hinnatakse vaid mudeli üldist sobivust. Seega on vajalik uurida mudeli komponente eraldi.

Põhiline mõõt komponentide headuse hindamiseks on R-ruut:

$$R_{x_i}{}^2 = 1 - \frac{var(\delta_i)}{\hat{\sigma}_{ii}}$$

Kus $\hat{\sigma}_{ii}$ on tunnuse x_i mudeli poolt hinnatud dispersioon. R-ruut näitab, kui suur osa tunnuse x_i hajuvusest mudeli poolt kirjeldatakse.

1.3.7 Regressioonimudelid

Geneetiliste mõjude hindamiseks kasutatakse lineaarset regressiooni ning logistilist regressiooni. Regressioonanalüüsi eesmärk on seletada ühte pidevat tunnust teiste tunnuste kaudu. Mitme argumendiga lineaarse regressiooni mudeli kuju on:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji} + \varepsilon_i$$

Kus β_0 on vabaliige, β_j $(j = 1,..,5)$ on regressioonikordajad ning ε_i on juhuslikud vead konstantse hajuvusega ning keskväärtusega null. Mudeli parameetrid hinnatakse vähimruutude meetodil selliselt, et uuritava tunnuse erinevused mõõdetud ja prognoositud väärtuste vahel oleksid minimaalsed. (Käärik, 2015)

Logistilise regressiooni puhul on uuritaval tunnusel kaks võimalikku väärtust, enamasti 0 ja 1. Huvi pakub seos uuritava tunnuse väärtuse 1 esinemise tõenäosuse ja indikaatortunnuste vahel.

Mudeli kuju on:

$$\ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_j x_{ji}$$

Kus π on sündmuse esinemise tõenäosus, β_0 on mudeli vabaliige ja β_j $(j = 1,..,5)$ on regressioonikordajad. (Käärik, 2015)

2 Praktiline osa

2.1 Andmestiku kirjeldus

Töös on kasutatud andmeid Tartu Ülikooli Eesti Geenivaramu geenidoonorite kohta, kelle geeniandmed on kaardistatud ehk genotüpiseeritud ülegenoomse kiibiga „Illumina CardioMetabochip“. See konkreetne valim on koostatud juht-kontrolluuringu põhimõttel, kus juhtudeks on 900 teist tüüpi diabeeti põdevat isikut ja kontrollideks 900 sarnase soo-vanusjaotusega isikut, kes on normaalkaalus ja kellel ei ole diabeeti diagnoositud. Käesolevas töös kasutatakse ainult nende isikute andmeid sellest valimist, kellel on määratud ka veresuhkru ehk glükoosi tase vereplasmas (NMR- metoodika abil)- 635 diabeeti põdevat isikut ja 735 kontrolli. Andmed jagunevad kolme andmestiku vahel, millest kaks koondavad nimetatud indiviidide fenotüübi ja genotüübi andmeid ning kolmas sisaldab ülegenoomse assotsiatsiooniuuringu tulemusi analüüsiks kasutatavate SNP-de ja nende mõjude kohta teist tüüpi diabeedi esinemisele (meta-analüüsi põhjal hinnatud logistilise regressiooni parameetrid, nende standardvead ja olulisuse tõenäosused).

Andmestikus „gwas10“ on andmed viie sõltumatu geenimarkeri kohta kümnendalt kromosoomilt, mis on oluliselt seotud teist tüüpi diabeediga. Tegu on suure ülegenoomse assotsiatsiooniuuringu (GWAS) meta-analüüsi põhjal raporteeritud markeritega, millel on antud geenipiirkondades diabeediga kõige tugevam seos. (Morris, 2012)

Andmestikus „c10“ on 1388 indiviidi 176 SNP andmed. Tegu on kümnenda kromosoomi geenimarkeritega, mis jagunevad viie piirkonna vahel, mis asuvad genoomis raporteeritud olulisemate geenimarkerite ümber. Nimetatud piirkonnad on ebavõrdse suurusega – viie piirkonna suurused on vastavalt 54, 7, 49, 59 ja 6 markerit. Vastavate piirkondade siseselt on markerid omavahel korreleeritud, erinevate piirkondade markerite puhul võib eeldada, et nad korreleeritud ei ole.

Andmestikus „fen“ on kodeeritud tunnus *sugu* järgnevalt: 0- naised, 1- mehed, diabeeti inditseeriv tunnus *gr*: 1- diabeet on, 2- diabeeti ei ole. Andmestikus „c10“ olevate SNP-de puhul on kokku loetud minoorsete alleelide esinemiste arv. Minoorseks nimetatakse SNP alleeli, mida esineb populatsioonis vähem. Näiteks kui esinevad alleelid on A ja C ning alleel A on minoorne, siis genotüüp AA = 2, AC = 1 ja CC = 0.

2.2 Fenotüübiandmete kirjeldav analüüs

Andmestikus on inimesi vanuses 32 kuni 93 aastat, keskmine vanus indiviididel on 57 aastat. Geenimarkerid on teada 1388 inimese kohta, kellest 860 on naised ning 528 mehed. Võib arvata, et diabeeti põdevatel isikutel on veresuhkru tase ravimite mõjul kunstlikult langetatud ja seetõttu kasutatakse analüüsis niinimetatud modifitseeritud veresuhkru taset, mis on saadud, liites veresuhkru tasemele juurde 2,3 ühikut juhul kui inimene põeb diabeeti.

Tabel 1 Fenotüübiandmete kirjeldavad statistikud

	Keskmine veresuhkru tase/ modifitseeritud veresuhkru tase		**Keskmine vanus**		**Indiviidide arv**	
	diabeet	**terve**	**diabeet**	**terve**	**diabeet**	**terve**
Mehed	6.95/9.25 (3.06/3.06)	4.22 (1.17)	64.1 (10.1)	53.8 (10.9)	248	280
Naised	6.49 /8.79 (2.72/2.72)	3.87 (1.09)	65.1 (11.3)	50.0 (10.3)	387	473
Mehed ja naised	6.67/8.97 (2.86/2.86)	4.00 (1.13)	64.7 (10.8)	51.4 (10.7)	635	753

2.2.1 Diabeedi ja veresuhkru taseme seos

Teist tüüpi diabeet on haigus, mille korral insuliini tootmine kõhunäärmes järk- järgult väheneb ning seetõttu on veresuhkru tase normist kõrgem. Põhilised riskifaktorid on vanus üle 40 aasta, ülekaalulisus ning samuti eelnevalt haiguse esinemine perekonnas ehk pärilikkus, mistõttu ei pruugi teist tüüpi diabeet olla ennetatav. (Eesti Diabeediliit, 2015)

Kinnitava faktoranalüüsi abil hinnatud latentsete tunnuste väärtuste kaudu on eesmärk prognoosida veresuhkru taset, mille kaudu saab hinnata riski teist tüüpi diabeedi tekkele. Lisa 1 jooniselt 5 nähtud, et valdavalt on modifitseerimata veresuhkru tase diabeedihaigetel

kõrgem kui tervetel indiviididel, kuigi arvestataval osal diabeedihaigetest on veresuhkru tase ka tervete inimestega samal tasemel.

2.3 Mudelid hindamaks diabeedi riski

Ülegenoomsetes assotsiatsiooniuuringutes, mille tulemusena on geenipiirkondade olulisemad markerid leitud, on kasutusel olnud valimid, mille maht on suurem kui 100 000 vaatlust. Et valim 1388 indiviidi kohta on sellega võrreldes väike, kontrolliti, kas viie raporteeritud markeri mõju osutub ka käesolevas bakalaureusetöös kasutatavas andmestikus oluliseks. Lisaks saab võrrelda tulemust kinnitava faktoranalüüsi mudeli abil saaduga, hindamaks, kas kinnitava faktoranalüüsi abil rohkemate markerite kasutamine annab paremaid tulemusi geneetiliste mõjude hindamisel.

Raporteeritud peamarkerite abil geneetiliste mõjude hindamisel kasutati lineaarse regressiooni ning logistilise regressiooni mudeleid, kus uuritavaks tunnuseks on vastavalt modifitseeritud veresuhkru tase *gl1* ning diabeeti inditseeriv tunnus *gr*. Indikaatortunnusteks on raporteeritud markerid andmestikust „gwas10“.

1) Regressioonanalüüs

Tabelis 2 on toodud viiele olulisemale markerile vastavad kordajad modifitseeritud veresuhkru tasemele hinnatud mitmeses lineaarses regressioonimudelis. Regressioonanalüüsi põhjal osutus viiest raporteeritud markerist olulisuse nivool 0,05 statistiliselt oluliseks vaid üks: marker „rs7903146“, mis on piirkonna „GenLoc4“ peamarker.

Tabel 2 Regressioonanalüüs peamarkeritelt

Tunnus	Hinnang	Standardhälve	p-väärtus	GWAS hinnang	GWAS standardhälve	GWAS p-väärtus
rs11257655	0.111	0.159	0.4862	0.073	0.015	9.75e-07
rs12571751	-0.167	0.122	0.1704	-0.076	0.012	2.79e-10
rs1111875	-0.218	0.127	0.0862	-0.113	0.012	5.75e-21
rs7903146	0.387	0.144	0.0072	0.353	0.014	2.60e-148
rs2421016	-0.203	0.126	0.1068	-0.061	0.012	4.33e-07

Mudeli headuse näitajad: $R^2 = 0.01066$.

2) Logistiline regressioon

Tabelis 3 on toodud viiele olulisemale markerile vastavad kordajad diabeeti inditseerivale tunnusele *gr* hinnatud logistilises regressioonimudelis. Analoogselt lineaarse regressiooniga, osutub ka logistilise regressiooni mudelis olulisuse nivool 0,05 oluliseks ainult marker „rs7903146“.

Tabel 3 Logistilise regressiooni mudel peamarkeritelt

Tunnus	Hinnang	Standardhälve	p-väärtus	GWAS hinnang	GWAS standardhälve	GWAS p- väärtus
rs11257655	0.152	0.099	0.1251	0.073	0.015	9.75e-07
rs12571751	-0.126	0.076	0.0985	-0.076	0.012	2.79e-10
rs1111875	-0.143	0.079	0.0709	-0.113	0.012	5.75e-21
rs7903146	0.253	0.090	0.0048	0.353	0.014	2.60e-148
rs2421016	-0.069	0.078	0.3760	-0.061	0.012	4.33e-07

Akaike informatsioonikriteerium: AIC = 1909,6.

Logistilise- ja lineaarse regressiooni parameetrite hinnangud on ülegenoomsete assotsiatsiooniuuringute poolt raporteeritutega samasuunalised: hinnatud efekt teist tüüpi diabeedile avaldub sarnaselt ka käesolevas bakalaureusetöös kasutatavast valimist.

2.4 Kinnitav faktoranalüüs

Kinnitav faktoranalüüs on viidud läbi kasutades statistikatarkvara „R“ struktuurivõrrandite mudelite koostamise paketti „lavaan“. Et lihtsustada mudeli sobitamise ja mudeli täpsustamise protseduuri, võeti igast geenipiirkonnast analüüsi vastava piirkonna peamarkeriga 10 kõige tugevamini korreleeritud markerit. Piirkondades, kus oli alla 11 markeri, võeti andmestikku kõik markerid. Sellest lähtuvalt jagunes andmestik faktorite vahel vastavalt 11, 7, 11, 11 ja 6, kokku 46 markerit.

Analüüsi jaoks kasutati funktsiooni „cfa“, mille puhul esialgse mudeli jaoks tuleb täpsustada mudeli süntaks ning andmestik, mida kasutatakse. Lisaks on võimalus täpsustada lisaparameetreid, näiteks, mida teha puuduvate väärtustega või millist parameetrite hindamise funktsiooni kasutada.

2.4.1 Esialgne mudel

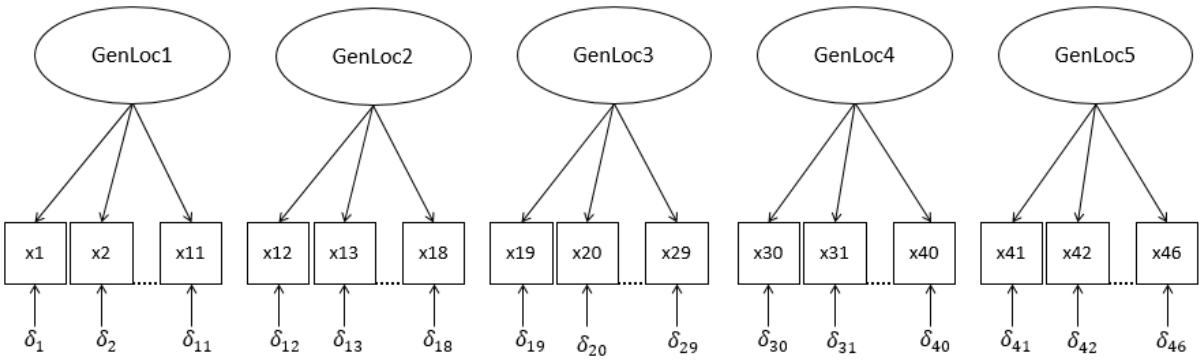
Esialgse mudeli süntaksis täpsustatakse, millised indikaatortunnused mõõdavad millist faktorit. Mudeli kirjapanekuks on markerite nimed asendatud nimedega x1, x2,...,x46 ning latentsed tunnused nimedega f1, f2,...,f6. Latentsete tunnuste siseselt on esimene marker antud piirkonna peamarker ning ülejäänud on järjestatud kahanevalt lähtuvalt markeri ja vastava piirkonna peamarkeri korrelatsiooni absoluutväärtusest.

Paketi „lavaan“ süntaks kasutab põhiliselt operaatoreid =~, mida kasutatakse latentsete tunnuste defineerimiseks, ~ regressioonseose näitamiseks ning ~~ kovariatsioonide ja dispersioonide täpsustamiseks. Programmikood mudeli hindamiseks on toodud lisas 5.

Esialgne mudel defineeritakse seega paketi „lavaan“ süntaksis järgmiselt:

```
Mudel_top10 <-'
#latentsete tunnuste defineerimine
f1=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11
f2=~x12+x13+x14+x15+x16+x17+x18
f3=~x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29
f4=~x30+x31+x32+x33+x34+x35+x36+x37+x38+x39+x40
f5=~x41+x42+x43+x44+x45+x46
'
```

Ülaltoodud mudeli süntaks koos eeldusega, et faktorid on omavahel sõltumatud, vastab teediagrammile, mis on kujutatud joonisel 3.



Joonis 3 Esialgse kinnitava faktoanalüüsi mudeli teediagramm

Faktorite sõltumatuse eeldus lähtub genoomikas teadaolevast: vastavad peamarkerid asuvad genoomis üksteisest liiga kaugel, et teineteist mõjutada. Samuti ei ole teada, et leiduks mõni kaudne mõju, mille kaudu antud geenipiirkonnad on seotud.

2.4.2 Mudeli identifitseeritavuse kontrollimine

Mudeli identifitseeritavuse määrab mudeli vabadusastmete arv:

$$df = \frac{m(m+1)}{2} - \ (mudeli\ parameetrite\ arv - \ kitsenduste\ arv\ parameetritele)$$

kus m on indikaatortunnuste arv, mis antud mudeli korral on kasutatavate geenimarkerite arv 46. Esialgse mudeli parameetrite arvu määravad järgmised komponendid:

- Latentsete tunnuste dispersioonid: kokku 5,
- struktuurikordajad, mis seovad latentseid tunnuseid indikaatortunnustega: kokku 46,
- juhuslike vigade dispersioonid: kokku 46.

Parameetreid lisamata on mudeli vabadusastmete arvuks:

$$df = \frac{46(46+1)}{2} - \ \big((5+46+46) - \ 0\big) = 984$$

Et latentsed tunnused on mittemõõdetud tunnused, siis neil puudub skaala, millel neid hinnata. Seega, et mudel oleks identifitseeritav, tuleb latentsete tunnuste skaala fikseerida. Käesolevas töös kasutatakse selleks faktorite normeerituse kitsendust: faktorite dispersioon fikseeritakse arvuks 1. Faktorite fikseeritud dispersioonide puhul on tegu viie kitsendusega mudelile, mis lisab viis vabadusastet, seega esialgse mudeli vabadusastmete arv on 989. Iga hinnatava parameetri lisamine mudelile vähendab vabadusastmete arvu ühe võrra.

2.4.3 Mudeli hindamine

Mudeli sobitamiseks funktsiooni „cfa“ abil kasutati järgmist süntaksit:

```
fit15 <- lavaan::cfa(mudel_top15,
          data=lavaan_andmed_top15, std.lv=T,
          missing="ml",estimator = "wlsm",orthogonal=T)
```

Parameetrite arvutamiseks kasutati „robustset“ diagonaalselt kaalutud vähimruutude *(diagonally weighed least squares- DWLS*) meetodit. Nimetatud DWLS meetodeid peetakse mitme uuringu põhjal täpseimaks, kui indikaatortunnused on väheste väärtustega järjestustunnused, mis ei ole normaaljaotusega (Gregory R. Hancock, 2006; Mindrila, 2010). Robustsed meetodi variatsioonid parandavad mudeli parameetrite hinnanguid, standardvigu ja mudeli headuse hinnanguid lähtuvalt tunnuste kvalitatiivsest olemusest (Gregory R. Hancock, 2006). Lisaks on kasutusel argumendid `orthogonal=T`, mis määrab latentsete

tunnuste vahelised korrelatsioonid nulliks ning `std.lv=T`, mis fikseerib latentsete tunnuste dispersioonid võrdseks arvuga 1.

Diagonaalselt kaalutud vähimruutude meetod eeldab, et andmestikus ei ole puuduvaid väärtusi. Analüüsist jäeti seetõttu välja indiviidide andmed, kellel olid osade SNP-de väärtused puudu. Kokku eemaldati 13 indiviidi vaatlused: algselt valimis olnud 1388 inimese andmetest kaasati analüüsi 1375.

Mudeli headuse hindamisel kasutati statistikuid RMSEA, CFI, SRMR ja RMR. Statistikute väärtused on toodud Tabelis 4.

Tabel 4 Esialgse mudeli headuse näitajad

RMSEA	CFI	SRMR	RMR
0,049	0,971	0,056	0,025

Esialgse mudeli headuse näitajad inditseerivad mudeli head vastavust andmetega. Lisas 3 joonisel 7 on toodud mudeli jääkide maatriks. Jäägid kujutavad endast mudeli prognoositud kovariatsioonide ja valimi kovariatsioonide vahet. Jooniselt nähtub, et suurimad vead on latentse tunnuse „GenLoc5“ indikaatoritevaheliste kovariatsioonide prognoosides. Samuti on suuremad jäägid „GenLoc3“ markerite x27 ja x28 ning „GenLoc2“ markerite x17 ja x18, x12 ja x18 ning x12 ja x13 vaheliste kovariatsioonide hindamisel. Lisa 2 joonisel 6 on kujutatud valimi kovariatsioonimaatriksit, kus on näha, et väiksemate jääkidega piirkondade „GenLoc1“ ja „GenLoc4“ puhul on kovariatsioonstruktuur palju ühtlasem: indikaatortunnustevahelised kovariatsioonid on kõik positiivsed ning kovariatsioonide suurused kõiguvad piirkonnasiseselt vähem. Siiski ei anna need tähelepanekud põhjust mudelisse parameetreid lisada, sest eeldatud struktuur on bioloogiliselt põhjendatud ning mudeli üldine sobivus on hea.

Hinnatud faktorkaalud on toodud lisa 4 tabelis 8. Kõik indikaatortunnused osutusid olulisteks, see tähendab, et nad sobivad andmete põhjal nimetatud latentseid tunnuseid mõõtma.

Lisaks osutub mudeli jääkide maatriksit analüüsides, et jäägid erinevate geenipiirkondade markerite vahel ei ole nullid. See tuleneb sellest, et valimis on empiirilised kovariatsioonid markerite vahel olemas, kuigi mudelis on nad fikseeritud nullideks. Et hii-ruut statistik arvutatakse jääkide maatriksist ning statistikut mõjutavad kõik nullist erinevad jäägid, siis hii-ruut test ei anna eelkõige suurte valimite tõttu adekvaatseid tulemusi.

2.4.4 Mudeli täpsustamine

Mudeli headuse näitajad inditseerisid algse mudeli puhul väga head kooskõla andmetega, mistõttu mudeli täpsustamine ei ole vajalik ning edaspidi kasutatakse analüüsiks esialgset mudelit, mis on defineeritud peatükis 2.4.1. Käesolevas peatükis tutvustatakse ühte enamlevinud võimalust mudeli täpsustamiseks.

Mudeli täpsustamiseks kasutatakse sageli modifikatsiooniindekseid, mis näitavad, kui palju muutub hii-ruut statistik, kui mõni parameeter „lasta vabaks“. See tähendab, et lisatakse mudelisse mõni hinnatav parameeter, mis eelnevalt eeldatakse olevat null. Näiteks lubatakse tunnuste vigade vaheline kovariatsioon, mis algses mudelis ei ole lubatud ning oli automaatselt fikseeritud nulliks. Vigade *(unique variance)* all mõistetakse seda osa indikaatortunnuse hajuvusest, mida faktorstruktuur ei kirjelda. See tähendab, et vigade korreleerituse põhjuseid on rohkem, kui seda eeldab hinnatav faktormudel.

Modifikatsiooniindeksid lähtuvad aga valimi kovariatsiooni struktuurist ning puhtalt arvutuslikust kujust. Tähtis on jälgida, et mudeli täpsustamisel lähtutakse teoreetilistest kaalutlustest: modifikatsiooniindeksite abil lisatud vaba parameeter peab olema põhjendatud ning kooskõlas teooriaga.

Tabel 5 Modifikatsiooniindeksid algsele mudelile

Lhs	Op	Rhs	Mi	Mi.scaled	Epc
X45	~~	X46	635,37	737,63	0.3933
f2	~~	f3	294,00	342,47	0.0851
X41	~~	X42	268,08	311,22	-0,2547

Tabelis 5 on kirjeldatud parameetrid, mida mudelisse lisada: operaator ~~ näitab tunnuste x45 ja x46 vigade vahelist korrelatsiooni, tunnused Mi *(modification index)* ja Mi.scaled on modifikatsiooniindeksi variatsioonid ning tunnus Epc *(expected parameter change)* näitab rea alguses näidatud parameetri hinnangulist väärtust, kui see parameeter mudelisse lisada.

Tabelist nähtub, et kui eeldada, et tunnuste x45 ja x46 vigade vaheline kovariatsioon ei ole võrdne nulliga ning antud parameeter mudelisse lisada, siis mudeli hii-ruut statistik väheneb eelduslikult 635,37 võrra. Modifikatsiooniindeksite abil mudelit täpsustades tuleb seda teha parameetri kaupa. Pärast parameetri lisamist tuleb mudel uuesti hinnata ning vaadata uuesti modifikatsiooniindekseid.

2.5 Faktoritel põhinevad mudelid glükoositasemele ja teist tüüpi diabeedile

Käesolevas peatükis kasutatakse kinnitava faktoranalüüsi struktuuri geneetiliste mõjude hindamiseks. Eesmärk on rakendada faktorskooridele regressioonanalüüsi ja logistilise regressiooni mudeleid ning võrrelda saadud tulemusi alapeatüki 2.2 omadega.

Faktorskooride arvutamiseks kasutati statistikatarkvara R põhipaketi funktsiooni „predict“, mis arvutab iga indiviidi geenimarkerite komplekti ja kinnitava faktoranalüüsi mudeli poolt hinnatud faktorkaalude abil latentsete tunnuste väärtused.

1) Lineaarne regressioonanalüüs

Tabelis 6 on toodud viiele hinnatud faktorskoorile vastavad kordajad modifitseeritud veresuhkru tasemele hinnatud mitmeses lineaarses regressioonimudelis. Olulisuse nivool 0,05 osutuvad oluliseks faktorite „GenLoc2“, „GenLoc3“ ja „GenLoc4“ mõjud.

Tabel 6 Regressioonanalüüsi parameetrite hinnangud faktorskooridelt

Tunnus	Hinnang	Standardviga	p-väärtus
GenLoc1	0.018	0.088	0.8390
GenLoc2	0.210	0.092	0.0227
GenLoc3	-0.214	0.089	0.0160
GenLoc4	0.238	0.088	0.0072
GenLoc5	-0.054	0.092	0.5546

Mudeli headuse näitajad: $R^2 = 0.0125$.

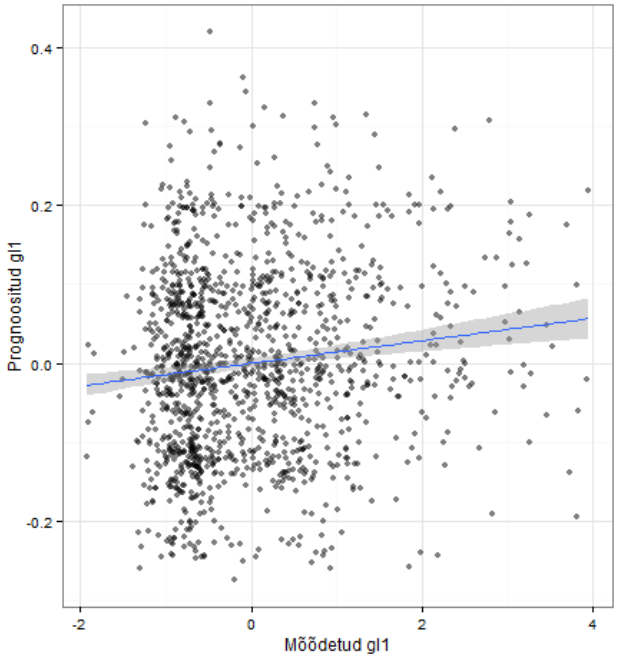
Võrreldes peatükis 2.2 tooduga, on mudeli kirjeldusvõime paranenud: kui ainult peamarkereid kasutav lineaarne regressioonimudel kirjeldas tunnuse *gl1* hajuvusest 1,07%, siis faktorstruktuuri kasutav regressioonimudel kirjeldas juba 1,25%. Lisaks osutusid faktoranalüüsi põhjal oluliseks veel kaks geenipiirkonda: kui enne osutus oluliseks ainult ühe geenipiirkonna peamarker „rs790146“, siis koos toetavate markeritega osutusid oluliseks ka markeritele „rs12571751“ ja „rs1111875“ vastavad geenipiirkonnad.

Faktorskooridel põhineva lineaarse regressiooni efektide hinnangud on samasuunalised peatükis 2.2 esitatud mudeliga, mis võtab arvesse vaid vastavate geenipiirkondade kõige

olulisemad diabeediga seotud markerid. Peatüki 2.2 põhjal on efektid seega samasuunalised ka ülegenoomsetes assotsiatsiooniuuringutes raporteeritutega.

Näiteks oluliseks osutunud geenipiirkonna „GenLoc3“ efekti hinnang tabeli 6 põhjal on -0,214 ning kinnitava faktoranalüüsiga hinnatud faktorkaal piirkonna peamarkeriga „rs1111875“ on 0,604. GWAS põhjal on markeri „rs1111875“ efekti hinnanguks -0,113. Kinnitava faktoranalüüsi mudeli põhjal näitab positiivne faktorkaal, et „Genloc3“ faktorskoori suurenedes kasvab ka peamarkeri prognoositud väärtus. Seega võib nimetada GenLoc3 ja markeri „rs1111875“ mõjusid samasuunalisteks. Asjaolust, et geenipiirkonna efekti hinnang on negatiivne, tuleneb omakorda, et faktoranalüüsi põhjal saadud tulemuses on ka markeri „rs1111875“ hinnatud efekt negatiivne, mis on kooskõlas GWAS-i hinnangutega.

Joonisel 4 on toodud mõõdetud ja prognoositud modifitseeritud veresuhkru taseme *gl1* hajuvusgraafik. Tunnuse *gl1* väärtused on standardiseeritud.



Joonis 4 Mõõdetud ja regressioonanalüüsi abil prognoositud standardiseeritud veresuhkru graafik

2) Logistiline regressioon

Tabelis 7 on toodud viiele hinnatud faktorskoorile vastavad kordajad diabeeti inditseerivale tunnusele *gr* hinnatud logistilise regressiooni mudelis. Olulisuse nivool 0,05 osutuvad samuti oluliseks faktorite „GenLoc2“, „GenLoc3“ ja „GenLoc4“ mõjud.

Tabel 7 Logistilise regressiooni mudel faktorskooridelt

Tunnus	Hinnang	Standardviga	p-väärtus
GenLoc1	0.030	0.055	0.5803
GenLoc2	0.141	0.057	0.0138
GenLoc3	-0.133	0.055	0.0169
GenLoc4	0.160	0.055	0.0036
GenLoc5	-0.028	0.057	0.6261

Akaike informatsioonikriteerium: AIC = 1889,6.

Võrreldes peatükis 2.2 tooduga mudeliga on faktorskooridel põhinev mudel parem: Akaike informatsioonikriteeriumi väärtus on vähenenud.

Sarnaselt lineaarse regressiooniga osutusid logistilise regressiooniga faktoranalüüsi põhjal olulisteks markeritele „rs12571751“ ja „rs1111875“ vastavad geenipiirkonnad. Logistilise regressiooniga hinnatud efektid tabelis 7 on samasuunalised tabelis 6 toodud lineaarse regressiooni efektide hinnangutega. Seetõttu on faktorskooridel põhineva logistilise regressiooni efektide hinnangud samuti samasuunalised peatükis 2.2 esitatud mudeliga, mis võtab arvesse vaid vastavate geenipiirkondade kõige olulisemad diabeediga seotud markerid ning seega ka ülegenoomsetes assotsiatsiooniuuringutes raporteeritutega.

Seega kinnitava faktoranalüüsi faktorskooridele hinnatud lineaarse regressioonanalüüsi ja logistilise regressiooni tulemused on kooskõlas varasemalt raporteeritud efektidega teist tüüpi diabeedile. Lisaks võimaldab faktorstruktuuril põhinev lähenemine võtta kasutusele rohkem informatsiooni geenipiirkonna kohta. Lineaarse ja logistilise regressioonanalüüsi tulemuste ja leitud parameetrite olulisuse tõenäosuste põhjal võib järeldada, et geenipiirkonna 7-11 markerit koondaval faktorskooril on uuritavatele fenotüübitunnustele tihti tugevam mõju kui antud piirkonna kõige olulisemal markeril.

Kokkuvõte

Bakalaureusetöö eesmärk oli uurida, kas struktuurivõrrandite mudelite erijuhu, kinnitava faktoranalüüsi kasutamine geneetiliste mõjude hindamisel aitab rohkem infot kasutades parandada tulemusi geneetiliste mõjude hindamisel.

Töös rakendati kinnitavat faktoranalüüsi 1375 geenidoonori genotüübiandmetele, mis pärinevad Tartu Ülikooli Eesti Geenivaramust. Kinnitava faktoranalüüsi abil võeti analüüsi lisaks ülegenoomsete assotsiatsiooniuuringute GWAS tulemusena raporteeritud sõltumatutele geenimarkeritele ka neile markeritele genoomis lähedal paiknevad markerid. Faktoranalüüsi tulemusena saadi faktorskoorid, mis esindavad geenipiirkondade summaarseid mõjusid.

Kasutati lineaarset ja logistilist regressiooni hindamaks faktorskooride ning sõltumatute markerite mõju teist tüüpi diabeedi esinemisele ja veresuhkru tasemele. Uuritava valimi põhjal osutus, et faktorskooridelt saadud hinnangud oli paremad: igast piirkonnast vaid GWAS tulemusena kõige olulisemaid markereid kasutades osutus valimi põhjal statistiliselt oluliseks vaid üks marker viiest, kuid võttes arvesse geenipiirkonna summaarset mõju, osutusid olulisteks kolm geenipiirkonda. Samuti oli geenipiirkonna, mille olulisem marker oluliseks osutus, mõju tugevam, kui ainult olulisemat markerit kasutades.

Antud töös kasutati vaid viie geenipiirkonna andmeid, kuid tegelikult on praeguse hetke seisuga tuvastatud enam kui 60 geenipiirkonna oluline seos teist tüüpi diabeedi riskiga. Seetõttu oleks edaspidi vaja uurida, kas saadud tulemused peavad üldjoontes paika ka siis, kui kasutatavate geenipiirkondade arvu suurendada. Samuti oleks meetodit vaja testida ka suuremates andmestikes, võttes arvesse ka muude (mittegeneetiliste) riskitegurite mõju.

Kasutatud kirjandus

Bollen, K. A. (1989). *Structural Equations with Latent Variables.*

Carlson, B. (2008). SNPs- A Shortcut to Personalized Medicine. *Genetic Engineering & Biotechnology News*.

Eesti Diabeediliit. (19. Aprill 2015. a.). Allikas: http://www.diabetes.ee/dokumendid/diabeet-2-tyyp.pdf

Gregory R. Hancock, R. O. (2006). *Structural Equation Modeling: A Second Course.*

Heinaru, A. (2012). *Geneetika õpik kõrgkoolile.* Tartu Ülikooli Kirjastus.

Kennedy, D. A. (6. Oktoober 2014. a.). *Measuring Model Fit.* Allikas: http://davidakenny.net/cm/fit.htm

Kim, S., & Misra, A. (2007). SNP Genotyping: Technologies and Biomedical Applications. *Annual Review of Biomedical Engineering*.

Käärik, E. (2015). *Loengukonspekt Andmeanalüüs II.*

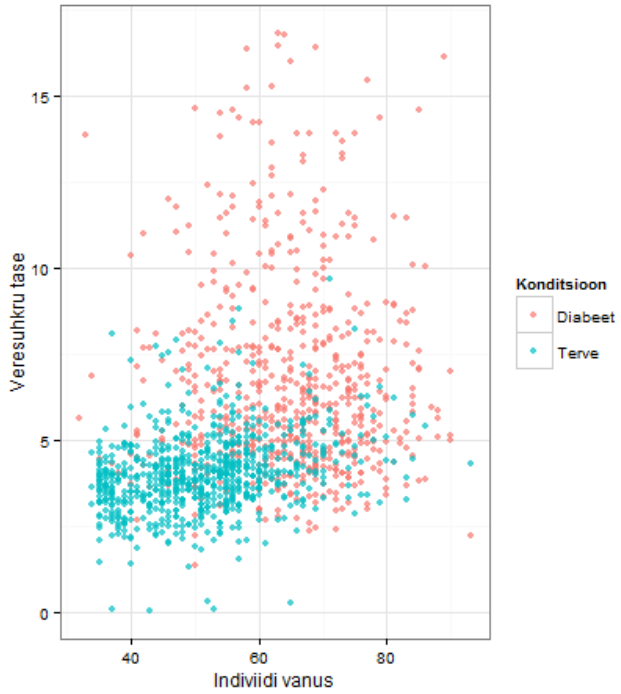
Mindrila, D. (2010). Maximum Likelihood (ML) and Diagonally Weighted Least Squares (DWLS) . *International Journal of Digital Society*.

Morris, A. P. (2012). Large-scale association analysis provides insight into the genetic architecture ant pathopysiology of type 2 diabetes. *Nature Genetics*.

Traat, I. (2014). *Loengukonspekt: Mitmemõõtmeline Analüüs.*

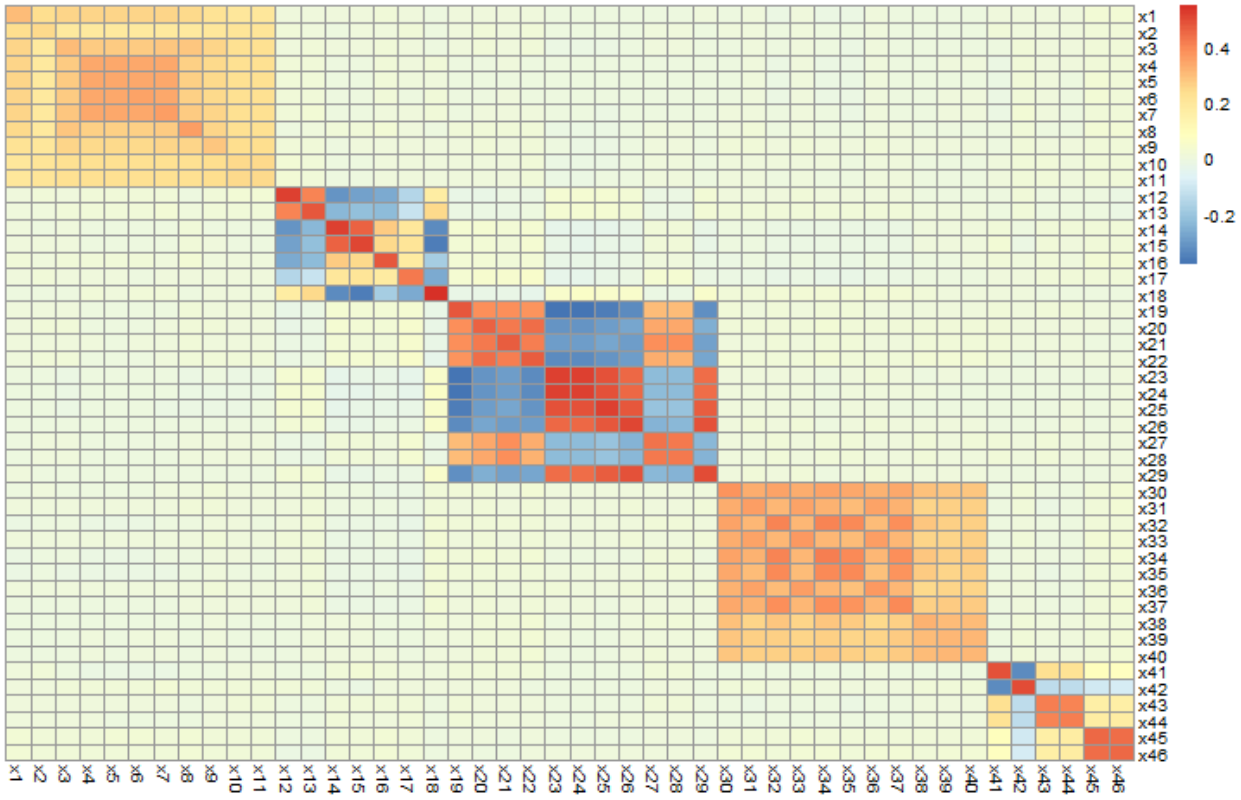
Lisad

Lisa 1. Veresuhkru tase diabeedihaigetel ja tervetel inimestel



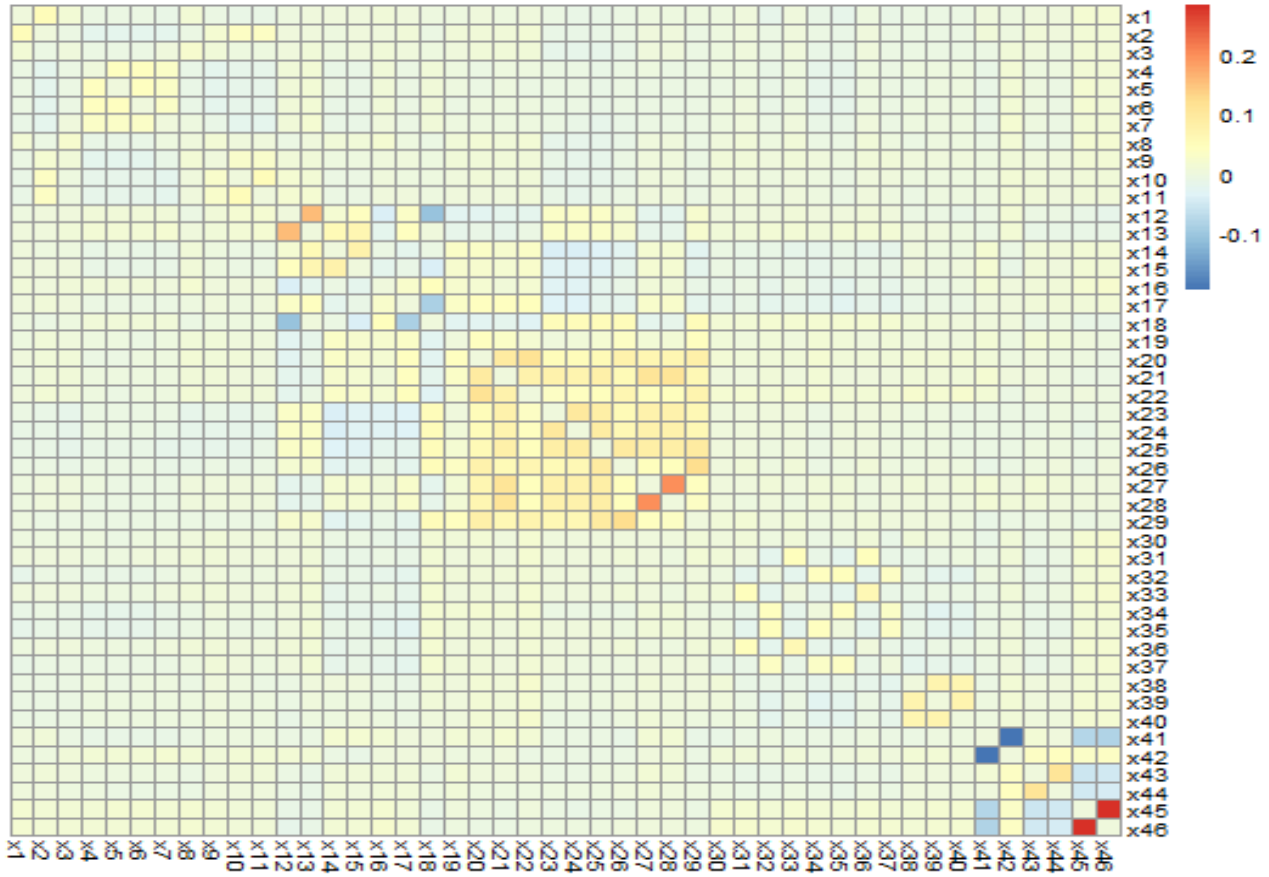
Joonis 5 Modifitseerimata veresuhkru tase diabeedihaigetel ja tervetel inimestel

Lisa 2. Tunnustevahelised kovariatsioonid geenipiirkondades



Joonis 6 Indikaatortunnustevahelised kovariatsioonid

Lisa 3. Kinnitava faktoranalüüsi mudeli jäägid



Joonis 7 Kinnitava faktoranalüüsi mudeli jääkide maatriks

Lisa 4. Mudeli parameetrite hinnangud

Tabel 8 Mudeli poolt hinnatud faktorkaalude koos standardhälvetega

lhs	op	rhs	faktorkaal	standardviga
f1	=~	x1	0.471	0.013
f1	=~	x2	0.384	0.014
f1	=~	x3	0.512	0.012
f1	=~	x4	0.545	0.011
f1	=~	x5	0.543	0.011
f1	=~	x6	0.544	0.011
f1	=~	x7	0.549	0.011
f1	=~	x8	0.498	0.013
f1	=~	x9	0.475	0.013
f1	=~	x10	0.442	0.013
f1	=~	x11	0.441	0.013
f2	=~	x12	-0.527	0.014
f2	=~	x13	-0.462	0.015
f2	=~	x14	0.627	0.012
f2	=~	x15	0.612	0.012
f2	=~	x16	0.439	0.016
f2	=~	x17	0.349	0.016

f2	=~	x18	-0.517	0.015
f3	=~	x19	0.604	0.011
f3	=~	x20	0.571	0.012
f3	=~	x21	0.573	0.012
f3	=~	x22	0.579	0.012
f3	=~	x23	-0.641	0.011
f3	=~	x24	-0.639	0.011
f3	=~	x25	-0.618	0.012
f3	=~	x26	-0.612	0.012
f3	=~	x27	0.475	0.014
f3	=~	x28	0.472	0.014
f3	=~	x29	-0.596	0.012
f4	=~	x30	0.583	0.011
f4	=~	x31	0.552	0.012
f4	=~	x32	0.594	0.011
f4	=~	x33	0.547	0.012
f4	=~	x34	0.599	0.011
f4	=~	x35	0.580	0.011
f4	=~	x36	0.539	0.013
f4	=~	x37	0.591	0.011
f4	=~	x38	0.486	0.014
f4	=~	x39	0.480	0.014
f4	=~	x40	0.483	0.014
f5	=~	x41	0.408	0.016
f5	=~	x42	-0.330	0.018
f5	=~	x43	0.538	0.013
f5	=~	x44	0.536	0.013
f5	=~	x45	0.396	0.016
f5	=~	x46	0.394	0.017

Lisa 5. Kasutatud programmikoodid

```
#Andmete laadimine
load("C:/Users/Samsung/Downloads/chr10_t2d.RData")

#Peamarkerite indeksite leidmine
match(c("rs11257655","rs12571751","rs1111875","rs7903146","rs2
421016"), names(c10_uus))

#Kirjeldav analüüs fenotüübiandmetele
summary(fen)

vanuse_jaotus <- ggplot(fen, aes(x=vanus))+geom_bar()+
          theme_bw()+xlab("Vanus")+ylab("Arv")
```

```
diabeedi_jaotus <- ggplot(fen, aes(x=vanus, y=Glc,
          color=factor(gr)))+ geom_point(alpha = 0.5)+
          labs(x="Indiviidi vanus",y="Modifitseeritud
          glükoositase")+ theme_bw()+
          scale_colour_manual(values = c("#999999","#000000"),
          name="Konditsioon",labels=c("Diabeet", "Terve"))

#Andmestiku jagamine gruppide vahel:
fen_terve=filter(fen,gr==1)
fen_haige=filter(fen,gr==0)

#Meeste ja naiste glükoosisisalduse erinevus
summarise(group_by(fen,sugu,gr), sd(Glc))
summarise(group_by(fen,sugu,gr),mean(gl1-Glc))

#Keskmiste erinevus:
summarise(group_by(fen,gr),mean(gl1))
summarise(group_by(fen,gr),mean(Glc))

#Regressioon ja logistiline regressioon peamarkeritelt
peamised_markerid_data <-
          c10[,c("rs2421016","rs7903146","rs1111875",
          "rs12571751","rs11257655")]
regressioon_peamarkeritelt_data <-data.frame(fen,
          peamised_markerid_data)

#Regressioonmudel esialgsetelt peamarkeritelt
Mudel_peamarkeritelt <-lm(gl1~rs2421016+rs7903146+rs1111875+
          rs12571751+ rs11257655,data=
          regressioon_peamarkeritelt_data)
Mudel_peamarkeritelt_lõplik <- lm(gl1~rs7903146,
          data=regressioon_peamarkeritelt_data)
summary(mudel_peamarkeritelt)

g1_jaotus_mudel2_pohjal <-
          ggplot(regressioon_peamarkeritelt_data,
```

```
          aes(x=gl1, fill= factor(rs7903146)))+
          geom_density(alpha = 0.3)+ theme_bw()+
          xlab("Modifitseeritud glükoositase")+
          ylab("Tihedus")+ scale_fill_manual(values =
          c("#FFFFFF", "#808080", "#000000"),name="SNP")

#Mudeli jaoks andmestiku leidmine: muudab tunnuste nimed,
#järjestab ümber peamarkeriga korrelatsiooni alusel
lavaan_andmed_top10 = loo_andmestik_mudelile(10)

#MUDELI SÜNTAKSI LEIDMINE: funktsioon, mis kasutab saadud
#faktorite suurusi ja leiab mudeli süntaksi
#leia_mudeli_syntax(11,7,11,11,6)

#Mudeli süntaks
Mudel_top10 <-'
#latentsete tunnuste defineerimine
f1=~x1+x2+x3+x4+x5+x6+x7+x8+x9+x10+x11
f2=~x12+x13+x14+x15+x16+x17+x18
f3=~x19+x20+x21+x22+x23+x24+x25+x26+x27+x28+x29
f4=~x30+x31+x32+x33+x34+x35+x36+x37+x38+x39+x40
f5=~x41+x42+x43+x44+x45+x46
'
fit10_algne<- lavaan::cfa(mudel_top10,
          data=lavaan_andmed_top10, std.lv=T,
          estimator = "wlsm",orthogonal=T)

#Kovariatsioonmaatriksid
pheatmap(cov(lavaan_andmed_top10[,2:47], use =
          "pairwise.complete.obs"), cluster_cols=FALSE,
          cluster_rows=FALSE)

jaagid = data.frame(residuals(fit10_algne))
colnames(jaagid)=rownames(jaagid)
pheatmap(jaagid[1:46,1:46], cluster_cols=FALSE,
     cluster_rows=FALSE)
```

```
f1 <- lavaan_andmed_top10[,2:11] #rs11257655
f2 <- lavaan_andmed_top10[,12:19] #rs12571751
f3 <- lavaan_andmed_top10[,20:30] #rs1111875
f4 <- lavaan_andmed_top10[,31:41] #rs7903146
f5 <- lavaan_andmed_top10[,42:47] #rs2421016

#Kovariatsioonmaatriksid faktorite kaupa
cex.before <- par("cex")
par(cex = 0.7)
col <- colorRampPalette(c("red","white", "blue"))
f1_korrelatsioonid = corrplot(cov(f1, use =
          "pairwise.complete.obs"), is.corr=F,
          type="lower",col=col(10), method = "number", cl.cex =
          1/par("cex"))
par(cex = cex.before)

#Mudeli parameetrite hinnangud
fitMeasures(fit10_algne,c("rmsea", cfi","srmr","rmr"))
summary(fit10_algne, fit.measures=TRUE, rsquare=T)
parameterEstimates(fit10_algne)

#fitted on mudeli kovariatsioonmaatriks, residuals on jääkide
maatriks
View(fitted(fit10_algne))
View(residuals(fit10_algne))

#Modifikatsiooniindeksid, mudeli täpsustamiseks
modification = modificationIndices(fit10_algne)
arrange(subset(modification, mi > 10),desc(mi))

#Mudeli diagnostika
#theeta on positiivselt määratud, kui kõik omaväärtused on
positiivsed
eigen(inspect(fit15_tapsustatud, "theta"))$values
inspect(fit15,"theta")  #negativsed peadiagonaalil ei sobi
```

```
#Annab välja prognoosi ja gl1 scatterploti ja jaotused

#Funktsioon ennustab regressioonmudeli abil tunnuse gl1
#väärtust ja lisab selle andmestikku koos mõõdetud tunnustega
prognoosiga_fen = ennusta_mudelist(fit10_algne)

#Loob hajuvusdiagrammi ja prognoositud ja mõõdetud gl1
#tihedusgraafiku
prognoos_gl1_tihedus(prognoosiga_fen,"varv")
prognoos_gl1_scatter(prognoosiga_fen,"varv")

#Mudelid faktoritelt
#juhtmarkerid: x1, x12, x19, x30, x41
faktorite_andmed10 = data.frame(cbind(fen$gr,
          lavaan_andmed_top10, predict(fit10_algne)))

#ainult faktormudelist tuleb f1,f5 ebaolulised
mudel_faktoritelt10 = lm(gl1~f1+f2+f3+f4+f5, data =
          faktorite_andmed15)
mudel_faktoritelt10_lõplik = lm(gl1~f2+f3+f4, data =
          faktorite_andmed15)
summary(mudel_faktoritelt10_lõplik)

#logistiline regressioon diabeedile faktoritelt
log_mudel_faktoritelt10 = glm(fen$gr-1~f1+f2+f3+f4+f5,
          family=binomial(), data = faktorite_andmed15)
log_mudel_faktoritelt10_lõplik = glm(fen$gr-1~f2+f3+f4,
          family=binomial(), data = faktorite_andmed15)
summary(log_mudel_faktoritelt10_lõplik)

#Hindamaks, kas faktorid on olulisemad, st kas omavad rohkem
#informatsiooni peamarkeritest
faktorid_koos_peamarkeritega =
          lm(gl1~f1+f2+f3+f4+f5+x1+x17+x24+x40+x56, data =
          faktorite_andmed10)
faktorid_koos_peamarkeritega_lõplik = lm(gl1~f2+f3+f4, data =
          faktorite_andmed10)
```

```
summary(faktorid_koos_peamarkeritega_lõplik)

#Mudelite võrdlemine
#võrdlus tavalise ja faktoritelt regressiooni vahel
anova(mudel_peamarkeritelt10,faktorid_koos_peamarkeritega)

#võrdlus faktoritelt regressiooni ja koos peamarkeritega
anova(mudel_faktoritelt10_lõplik,mudelt_peamarkeritelt_lõplik)
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Oliver Aasmets (sünnikuupäev 17.09.1993)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Geneetiliste mõjude hindamine kinnitava faktoranalüüsiga“, mille juhendaja on Krista Fischer.
   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, 29.04.2015