Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Lisette Pajula

Geneetilise korrelatsiooni hindamine LD-skoori regressiooni meetodiga

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendajad:
Märt Möls
Reedik Mägi

Tartu 2018

Geneetilise korrelatsiooni hindamine LD-skoori regressiooni meetodiga

Bakalaureusetöö

Lisette Pajula

Lühikokkuvõte. Geneetiline korrelatsioon näitab, kui tugevalt on kahe fenotüübi geenidest mõjutatud osad omavahel seotud. Bakalaureusetöös antakse ülevaade ühest selle hindamise meetodist – LD-skoori regressioon. Töö eesmärk on veenduda selle paikapidavuses nii teoorias kui ka praktikas, et seejärel seda rakendada. Tutvustatakse asjakohaseid geneetika termineid, mille abil selgitatakse meetodi olemust. Lisaks viiakse läbi simulatsioon, et võrrelda meetodi hinnangu käitumist erinevates olukordades. Lõpetuseks hinnatakse ülegenoomsete assotsiatsioonuuringute andmetelt seitsme tunnuse vahelised geneetilised korrelatsioonid.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Märksõnad: andmeanalüüs, simulatsioon, regressioonanalüüs, biomeetria, geneetilised assotsiatsiooniuuringud, korrelatsioonanalüüs

Estimating genetic correlation with LD-score regression method

Bachelor's thesis

Lisette Pajula

Abstract. Genetic correlation is a measure of association between the genetic influences on two traits. This thesis provides an overview of one of the methods of its estimation – LD-score regression. The aim is to assure it works in theory as well as in practice. Relevant genetic terms are introduced to explain the essence of this method. In the practical part of the thesis a simulation is conducted to compare the characteristics of LD-score regression estimate in different situations. Finally, the method is applied to genome-wide association studies' data to estimate genetic correlations between seven traits.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics

Keywords: data analysis, simulation, regression analysis, biometrics, genetic association studies, correlation analysis

Sisukord

Sissejuhatus

Inimeste arengut ning tervist puudutavate küsimuste uurimisel on üha rohkem hakatud väärtustama geneetilist informatsiooni. Viiakse läbi laialdaselt uuringuid, et saada teada, missugused geenid erinevate tunnuste ja haiguste kujunemist mõjutavad. Selliste uuringute tulemused omakorda võimaldavad kõikvõimalikke geneetilisi seoseid veelgi põhjalikumalt uurida. Üheks huvipakkuvaks aspektiks sealjuures on tunnustevaheline geneetiline korrelatsioon.

Bakalaureusetöös seletatakse lahti geneetilise korrelatsiooni mõiste ning antakse ülevaade ühest selle hindamise meetodist: LD-skoori regressioon. Töö eesmärk on veenduda meetodi toimimises nii teoorias kui ka praktikas, et seejärel seda reaalsetel andmetel rakendada.

Töö koosneb kolmest osast. Esimene peatükk keskendub erinevatele geneetikaga seotud terminitele, et töö edasine käik oleks lugejale mõistetav. Teises peatükis selgitatakse, kuidas vastavaid termineid LD-skoori regressiooni meetodi kirjelduses käsitletakse, ning seejärel antakse ülevaade meetodi olemusest. Muuhulgas täiendatakse ja parandatakse varasemalt publitseeritud tõestuseid meetodi paikapidavuse kohta.

Viimane peatükk kujutab endast töö praktilist osa, mille käigus viiakse läbi simulatsioon, et võrrelda LD-skoori meetodi hinnangu käitumist erinevates olukordades. Praktilise osa teises pooles rakendatakse meetodit kahe antropomeetrilise tunnuse ning viie vere keemilise näitaja vaheliste geneetiliste korrelatsioonide hindamiseks. Tuuakse ülevaade nii andmestikest, arvutuskäigust kui ka tulemustest.

Bakalaureusetöö on vormistatud tekstitöötlusprogrammiga LaTeX. LD-skoori regressiooni meetodi rakendamiseks ja jooniste tegemiseks on kasutatud statistikatarkvara R (versioon 3.4.3).

Autor tänab siiralt bakalaureusetöö juhendajaid Märt Mölsi ja Reedik Mäge töösse panustatud aja ning loendamatute selgituste ja näpunäidete eest.

1 Terminoloogia

Järgnev peatükk põhineb allikal Heinaru (2012), kui pole viidatud teisiti.

1.1 Geneetika alusmõisted

Inimese organismi väikseim iseseisvalt kasvav, arenev ja paljunev osa on rakk. Igas keharakus on kokku 23 paari kromosoome ehk DNA molekule, mille peamiseks funktsiooniks on geneetilise informatsiooni säilitamine ning edasikandmine järglasrakkudesse. DNA molekul koosneb kahest järjestatud nukleotiidide (A – adeniin, G – guaniin, T – tümiin ja C – tsütosiin) ahelast, mis püsivad omavahel koos kindla reegli ehk DNA komplementaarsusprintsiibi alusel.

Geen on spetsiifilise bioloogilise funktsiooniga DNA molekuli lõik, mis asub kromosoomi kindlas punktis ehk lookuses. Paarilisi kromosoome, mille korral üks koopia on pärit isalt ja teine emalt, nimetatakse homoloogseteks kromosoomideks. Neis on samad geenid, kuid igas lookuses võib esineda kaks (või enam) geeniteisendit ehk alleeli. Alternatiivsed geeniteisendid mõjutavad küll sama(de) tunnus(t)e kujunemist, kuid põhjustavad selle (nende) erinevaid avaldumisastmeid. Organismi geenilookuste alleelset koosseisu nimetatakse genotüübiks. Alleelide avaldumisel moodustunud tunnuste kogumit nimetatakse aga fenotüübiks, mille kujunemist mõjutavad ka keskkonnategurid.

1.2 Geneetilise varieeruvuse allikad

Inimesed on geneetiliselt suurel määral identsed. DNA replikatsiooni ja järglastele edasikandmise käigus tekib aga erinevatel põhjustel vigu, millest tulenevalt geneetiline materjal muutub. Vastavaid muutusi nimetatakse mutatsioonideks, milleta puuduksid geenide erivormid ehk alleelid. Kõikvõimalikke DNA osi, mis populatsioonis varieeruvad, nimetatakse üldisemalt geneetilisteks markeriteks (Zheng, Yang, Zhu & Elston, 2012).

Kõige sagedasem geneetilise varieeruvuse avaldumisvorm on üksiknukleotiidne polümorfism ehk SNP (ingl *single nucleotide polymorphism*), mis tähendab ühe nukleotiidi vahetusvarieeruvust populatsioonis. Näiteks, kui enamasti on mingis lookuses nukleotiidide järjestuseks „T A **G** C A“, kuid samas lookuses esineb populatsioonis ka järjestust „T A **C** C A“, siis selles kohas on SNP. Keskmiselt esineb seda üks kord iga 200–300 aluspaari kohta. SNP-ide (või muude geneetiliste elementide) kombinatsiooni, mis pärandub ühtse üksusena, nimetatakse haplotüübiks.

Et välja selgitada, millised SNP-id on seotud mingi vaadeldava tunnuse või haigusega, viiakse läbi ülegenoomseid assotsiatsioonuuringuid (ingl *genome-wide association study*) ehk GWAS-e (Zheng *et al.*, 2012).

1.3 Aheldustasakaalutus ehk LD

Nähtust, mille korral mingit geneetiliste markerite alleelide kombinatsiooni esineb populatsioonis oluliselt sagedamini (või harvemini), kui tõenäosuslikult eeldada võiks, nimetatakse aheldustasakaalutuseks (ingl *linkage disequilibrium, LD*). LD viitab markeritevahelisele sõltuvusele (Zheng *et al.*, 2012).

Vaatleme Zheng *et al.* (2012) näite põhjal kahte dialleelsete geenide lookust, kus esimeses on alleelid A ja a ning teises B ja b. Esinegu vastavad alleelid populatsioonis sagedustega p_A, p_a, p_B ja p_b ning nendest moodustunud haplotüübid sagedustega p_{AB}, p_{Ab}, p_{aB} ja p_{ab}. Haplotüübile AB vastav LD koefitsient on defineeritud kui

$$LD_{AB} = p_{AB} - p_A p_B.$$

Et toodud ühikut ka teiste haplotüüpide LD koefitsientidega võrrelda saaks, tuleb seda standardiseerida. Kasutades Cauchy-Schwartzi võrratust, saame kirjutada:

$$|LD_{AB}| = |p_{AB} - p_A p_B| \leqslant \sqrt{p_A p_a p_B p_b}.$$

Sellest tulenevalt on LD kirjeldamiseks kasutusel standardiseeritud koefitsient

$$r_{AB} = \frac{LD_{AB}}{\sqrt{p_A p_a p_B p_b}},$$

mille korral kehtib $-1 \leqslant r_{AB} \leqslant 1$. Olukord $r_{AB} \neq 0$ viitab ahelduse mittetasakaalulisusele – öeldakse, et vastavad alleelid on omavahel LD-s.

1.4 Päritavus

Olgu mingi kvantitatiivse tunnuse oodatavaks fenotüübiliseks väärtuseks y, siis saame selle kirja panna kõrvalekaldena populatsiooni keskväärtusest μ:

$$y = \mu + g + e,$$

kus g tähistab geneetilisi mõjusid ning e keskkonnategureid (ingl *environment*). Kuna μ on konstant, siis vaadeldava tunnuse varieeruvus populatsioonis on tingitud vaid viimasest kahest liidetavast. Seega, eeldades geneetiliste mõjude ja keskkonnategurite sõltumatust:

$$D(y) = D(g) + D(e).$$

Päritavuseks nimetatakse kvantitatiivse tunnuse populatsioonisisese muutlikkuse seda osa, mis on tingitud indiviidide genotüüpide eripäradest. Kuna genotüübiline muutlikkus sõltub nii geenide kui ka alleelide omavahelisest toimest, siis eristatakse selle juures kolme komponenti. Neist tähtsaim on eri geenide alleelide aditiivsest toimest tulenev varieeruvus ($D(a)$), millele võib lisanduda alleelide domineerimisest ($D(d)$) ja erinevate geenide alleelide interaktsioonist ($D(i)$) põhjustatud muutlikkus:

$$D(g) = D(a) + D(d) + D(i).$$

Seetõttu saab päritavust käsitleda kahes tähenduses: laiemas ja kitsamas. Laiatähenduslik päritavus (H^2) on kogu genotüübilise varieeruvuse suhe tunnuse populatsioonisisesesse kogumuutlikusse:

$$H^2 = \frac{D(g)}{D(y)}.$$

Alleelide domineerimisest ja interaktsioonist tulenevat varieeruvust on raskem hinnata kui aditiivset mõju. Kuna nende panus populatsiooni genotüübilisse koguvarieeruvusse on ka väiksem, siis tihtipeale räägitakse päritavusest kitsamas tähenduses (h^2), mis võtab arvesse vaid geenide alleelide aditiivse toime:

$$h^2 = \frac{D(a)}{D(y)}.$$

Nii laia- kui ka kitsatähenduslik päritavuskoefitsient varieerub piirides 0 kuni 1. Päritavus näitab, kui suurt osa kvantitatiivse tunnuse varieeruvusest mõjutab geneetiline komponent.

1.5 Geneetiline korrelatsioon

Olgu nüüd vaatluse all kaks kvantitatiivset tunnust y_1 ja y_2:

$$y_1 = \mu_1 + g_1 + e_1,$$
$$y_2 = \mu_2 + g_2 + e_2.$$

Tunnuste y_1 ja y_2 vaheline geneetiline korrelatsioon r_g näitab, kui tugevalt on nende geenidest mõjutatud osad omavahel korreleeritud (Searle, 1961) ehk

$$r_g(y_1, y_2) = \mathrm{cor}(g_1, g_2) = \frac{\mathrm{cov}(g_1, g_2)}{\sigma(g_1)\sigma(g_2)}, \quad -1 \leqslant r_g \leqslant 1.$$

Ka siin saame geenimõjude komponendi jagada mitmeks osaks. Geneetilisest korrelatsioonist rääkides tehakse seda üldjuhul lihtsustatult: vaadeldav tunnus avaldatakse kahe liidetava kaudu, millest üks tähistab geenide aditiivset mõju ja teine keskkonnamõju koos ülejäänud geeniefektidega (Falconer, 1960). Näiteks:

$$y_1 = a_1 + e_1,$$
$$y_2 = a_2 + e_2.$$

Seetõttu kasutatakse geneetilise korrelatsiooni tähistusena vahel ka r_a, kuna peetakse silmas korrelatsiooni kahe tunnuse aditiivsete geenimõjude vahel:

$$r_a(y_1, y_2) = \mathrm{cor}(a_1, a_2).$$

Geneetilist korrelatsiooni tõlgendatakse sarnaselt tavalise korrelatsiooni mõistega. Kui korrelatsioon on võrdne nulliga, siis järelikult ühe tunnuse geneetilised mõjud on täielikult sõltumatud teise tunnuse geneetilistest mõjudest. Seevastu (absoluutväärtuselt) ühele lähedane korrelatsioon näitab väga tugevat seost tunnuste geneetiliste taustade vahel.

Korrelatsiooni positiivne suund tähendab, et kui üks vaadeldav tunnus on geneetiliselt soodustatud suuri (väikseid) väärtuseid omama, siis ka temaga positiivselt korreleeritud tunnuse geneetilised mõjud soodustavad üldjuhul vastava tunnuse suuri (väikseid) väärtuseid. Negatiivse korrelatsiooni puhul on aga tunnuste väärtuste geneetiliste mõjude vahel vastassuunaline seos.

2 LD-skoori regressiooni meetod

Geneetiliste korrelatsioonide hindamiseks on mitmeid erinevaid meetodeid. Osad neist eeldavad perekonnapõhiseid andmeid, mille kogumine võib osutuda raskeks ja kalliks. Seetõttu on välja töötatud teisigi meetodeid, mis kasutavad vaid ülegenoomsete assotsiatsioonuuringute andmeid. Neist omakorda suur osa analüüsib vaid geneetilisi markereid, mis GWAS-i tulemusel on statistiliselt oluliseks osutunud. Sellised meetodid on tõhusad vaid tunnuste korral, mida mõjutavadki väga paljud statistiliselt olulised SNP-id. Enamik kvantitatiivseid fenotüüpe on aga mõjutatud tuhandete geneetiliste variantide poolt, mille efektisuurused eraldivõetuna on väga väikesed. Lisaks sellele leidub GWAS-ide metaanalüüside seas laialdaselt omavahel kattuvaid valimeid, mille korral enamik meetodeid annab nihkega hinnanguid. Et nimetatud probleemidega toime tulla, on välja töötatud LD-skoori regressiooni meetod. (Bulik-Sullivan *et al.*, 2015)

2.1 Tähistused

Alapeatükk põhineb kahel Bulik-Sullivani *et al.* (2015) artiklil.

2.1.1 Kvantitatiivse tunnuse esitus

Olgu valimis N inimest, kellel on mõõdetud mingi kvantitatiivse tunnuse väärtus. Vastava fenotüüpide vektori $\phi : N \times 1$ saame esitada kujul

$$\phi = X\beta + \epsilon, \tag{1}$$

kus $X : N \times M$ on genotüüpide maatriks (igale indiviidile vastab üks rida M SNP-i väärtusega) ja $\beta : M \times 1$ on genotüüpide efektisuuruste vektor. $X\beta$ kirjeldab seega geenide aditiivset mõju ja $\epsilon : N \times 1$ on keskkonnamõjude (ja mitteaditiivsete geenimõjude) vektor. Suurused X, β ja ϵ on juhuslikud ning omavahel sõltumatud, kusjuures $\beta_j \perp \beta_k$, $j \neq k$.

Olgu X standardiseeritud kujul ehk

$$\forall\, i,j : E(X_{ij}) = 0,\ D(X_{ij}) = 1, \quad i \in \{1,\dots,N\},\ j \in \{1,\dots,M\}.$$

Eeldame, et ka $D(\phi_i) = 1,\ \forall\, i \in \{1,\dots,N\}$. Lisaks kehtigu $\forall\, k,j : E(\beta_k^2) = E(\beta_j^2)$ ning $E(\beta_j) = E(\epsilon_j) = 0,\ k,j \in \{1,\dots,M\}$.

Lause 1. Kitsatähenduslik päritavuskoefitsient avaldub geeniefektide kaudu:

$$h^2 = M \cdot E(\beta_j^2).$$

Tõestus. Avaldugu indiviidi i vaadeldav fenotüüp kui $\phi_i = X_i\beta + \epsilon_i$. Siis (kitsatähendusliku) päritavuse definitsiooni kohaselt

$$h^2 = \frac{D(X_i\beta)}{D(\phi_i)} = D(X_i\beta). \tag{2}$$

Leiame aditiivse geneetilise komponendi $X_i\beta$ dispersiooni:

$$\begin{aligned}
D(X_i\beta) &= D\left(\sum_{j=1}^{M} X_{ij}\beta_j\right) \\
&= \sum_{j=1}^{M} D(X_{ij}\beta_j) + \sum_{k\neq l}^{M} \operatorname{cov}(X_{ik}\beta_k, X_{il}\beta_l) \\
&= \sum_{j=1}^{M} E(X_{ij}\beta_j - E(X_{ij}\beta_j))^2 + \sum_{k\neq l}^{M} [E(X_{ik}\beta_k X_{il}\beta_l) - E(X_{ik}\beta_k)E(X_{il}\beta_l)] \\
&\overset{(X_{ij}\perp\beta_j)}{=} \sum_{j=1}^{M} E[X_{ij}\beta_j - E(X_{ij})E(\beta_j)]^2 + \\
&\quad + \sum_{k\neq l}^{M} [E(X_{ik}X_{il})E(\beta_k)E(\beta_l) - E(X_{ik})E(\beta_k)E(X_{il})E(\beta_l)] \\
&\overset{E(\beta_j)=0}{=} \sum_{j=1}^{M} E(X_{ij}\beta_j)^2 \\
&\overset{X_{ij}\perp\beta_j}{=} \sum_{j=1}^{M} E(X_{ij}^2)E(\beta_j^2) \\
&\overset{E(X_{ij}^2)=1}{=} \sum_{j=1}^{M} E(\beta_j^2) \\
&= M \cdot E(\beta_j^2) \\
&\approx \sum_{j=1}^{M} \hat{\beta}_j^2.
\end{aligned}$$

■

Lausest 1 saame järeldada, et $D(\beta_j) = E[\beta_j - E(\beta_j)]^2 = E(\beta_j^2) = h^2/M$. Peale selle on ilmne, et $D(\epsilon) = 1 - h^2$. Eeldame, et indiviidi genotüüp ei sõltu teiste indiviidide genotüüpidest ehk

$$\forall\, k, l, j : X_{kj} \perp X_{lj}, \quad k, l \in \{1, \ldots, N\},\ j \in \{1, \ldots, M\},\ k \neq l.$$

Ühe indiviidi erinevad SNP-id võivad aga aheldustasakaalutuse tõttu olla omavahel sõltuvad. Seetõttu võtame kasutusele korrelatsiooni tähistuse

$$r_{jk} := E(X_{ij} X_{ik}), \tag{3}$$

mis ei sõltu i-st, ja defineerime j-nda SNP-i LD-skoori:

$$\ell_j := \sum_{k=1}^{M} r_{jk}^2. \tag{4}$$

2.1.2 Geneetilise korrelatsiooni esitus

Olgu nüüd vaatluse all kaks valimit suurustega N_1 ja N_2, kus ühes on mõõdetud tunnus 1 ja teises 2. Esitame fenotüüpide vektorid y_1 ja y_2 kujul

$$y_1 = X\beta + \epsilon$$
$$y_2 = Y\gamma + \delta,$$

kus kõik parameetrid on kirjeldatavad analoogselt avaldisega (1). Juurde lisandub aga asjaolu, et j-nda SNP-i mõjud β_j ja γ_j võivad olla omavahel sõltuvad. Seevastu $j \neq k$ korral kehtib $\beta_j \not\perp \gamma_k$. Juhul, kui indiviid i on kaasatud mõlemasse valimisse, võib ka keskkonnamõjude ϵ_i ja δ_i vahel sõltuvus olla. Olgu selliste indiviidide arvu tähistuseks N_s ja paiknegu nendele vastavad vaatlused vektorite y_1 ja y_2 alguses ehk esimeste N_s elementide seas.

Kõikide indiviidide genotüüpide vektorid (nii maatriksi X kui ka Y read) on sõltumatud sama jaotusega juhuslikud suurused kovariatsioonimaatriksiga R (3). Leidub vaid N_s maatriksi X ja Y rida, mis on omavahel (täielikult) sõltuvad, kuna kirjeldavad samade indiviidide markereid.

Eeldame, et $\forall\, k, j : E(\beta_k \gamma_k) = E(\beta_j \gamma_j)$, $k, j \in \{1, \ldots, M\}$.

Lause 2. Kahe tunnuse vaheline geneetiline kovariatsioon avaldub nende geeniefektide kaudu:

$$\rho_g(y_1, y_2) = M \cdot E(\beta_j \gamma_j).$$

Tõestus. Olgu $X_i : 1 \times M$ i-nda indiviidi standardiseeritud genotüüpide vektorile vastav juhuslik suurus. Antud olukorras avaldub tunnust 1 mõjutav aditiivne geneetiline komponent kui $X_i\beta$. Tunnuse 2 korral on selleks analoogselt $X_i\gamma$.

Geneetilise kovariatsiooni definitsiooni kohaselt

$$\begin{aligned}
\rho_g(y_1, y_2) &= \operatorname{cov}(X_i\beta, X_i\gamma) \\
&= \operatorname{cov}\left(\sum_{j=1}^{M} X_{ij}\beta_j, \sum_{j=1}^{M} X_{ij}\gamma_j\right) \\
&= \sum_{j=1}^{M}\sum_{k=1}^{M} \operatorname{cov}(X_{ij}\beta_j, X_{ik}\gamma_k) \\
&= \sum_{j=1}^{M}\sum_{k=1}^{M} E[X_{ij}\beta_j - E(X_{ij}\beta_j)][X_{ik}\gamma_k - E(X_{ik}\gamma_k)] \\
&\overset{X_{ij} \perp \beta_j}{=} \sum_{j=1}^{M}\sum_{k=1}^{M} E[X_{ij}\beta_j - E(X_{ij})E(\beta_j)][X_{ik}\gamma_k - E(X_{ik})E(\gamma_k)] \\
&\overset{E(X_{ij})=0}{=} \sum_{j=1}^{M}\sum_{k=1}^{M} E(X_{ij}\beta_j X_{ik}\gamma_k) \\
&= \sum_{j\neq k}^{M}\sum^{M} E(X_{ij}X_{ik}\beta_j\gamma_k) + \sum_{j=1}^{M} E(X_{ij}^2\beta_j\gamma_j) \\
&= \sum_{j=1}^{M} E(X_{ij}^2)E(\beta_j\gamma_j) \\
&\overset{E(X_{ij}^2)=1}{=} \sum_{j=1}^{M} E(\beta_j\gamma_j) \\
&= M \cdot E(\beta_j\gamma_j) \\
&\approx \sum_{j=1}^{M} \hat{\beta}_j\hat{\gamma}_j.
\end{aligned}$$

■

Lausest 1 ja 2 saame tuletada, et (β, γ) kovariatsioonimaatriks on

$$\Sigma(\beta, \gamma) = \frac{1}{M} \begin{pmatrix} h_1^2 I_M & \rho_g I_M \\ \rho_g I_M & h_2^2 I_M \end{pmatrix},$$

milles h_1^2 on esimese tunnuse päritavus ja h_2^2 teise. Tähistame $\rho := \rho_g + \rho_e$, kus $\rho_e = \mathrm{cov}(\epsilon_i, \delta_i)$ on keskkondade vaheline korrelatsioon, mis on hinnatav N_s indiviidi põhjal. Saame kirja panna ka (ϵ, δ) kovariatsioonimaatriksi:

$$\Sigma(\epsilon, \delta) = \begin{pmatrix} (1 - h_1^2) I_{N_1} & \rho_e I_{N_s}^{*1} \\ \rho_e I_{N_s}^{*2} & (1 - h_2^2) I_{N_2} \end{pmatrix},$$

kus $I_{N_s}^{*1}$ on $N_1 \times N_2$ maatriks kujul

$$I_{N_s}^{*1} = \begin{pmatrix} I_{N_s} & 0 \\ 0 & 0 \end{pmatrix}.$$

Maatriks $I_{N_s}^{*2}$ on analoogne, kuid mõõtmetega $N_2 \times N_1$.

Lause 3. Kahe tunnuse vaheline geneetiline korrelatsioon on leitav kui

$$r_g(y_1, y_2) = \frac{\rho_g(y_1, y_2)}{\sqrt{h_1^2 h_2^2}}.$$

Tõestus. Eeldame sama olukorda nagu lause 2 tõestuses, siis geneetilise korrelatsiooni definitsiooni kohaselt

$$\begin{aligned} r_g(y_1, y_2) &= \mathrm{cor}(X_i\beta, X_i\gamma) \\ &= \frac{\mathrm{cov}(X_i\beta, X_i\gamma)}{\sigma(X_i\beta)\sigma(X_i\gamma)} \\ &\overset{(2)}{=} \frac{\rho_g(y_1, y_2)}{\sqrt{h_1^2 h_2^2}}. \end{aligned}$$

■

Nii geneetilisest korrelatsioonist kui ka kovariatsioonist rääkides eelistame tähistustes alaindeksit g, kuid sisuliselt käsitleme LD-skoori regressiooni meetodiga siiski seost kahe tunnuse aditiivsete geenimõjude vahel.

2.2 Mudel

LD-skoori meetodiga geneetilise korrelatsiooni tuvastamine põhineb lineaarse mudeli parameetrite hindamisel. Kuna käsitletavad vaatlused võivad olla erineva hajuvusega, siis kasutatakse selleks üldistatud vähimruutude meetodit.

2.2.1 Üldistatud vähimruutude meetod

Alapeatükk põhineb dispersioonanalüüsi segamudelite loengukonspektil (Möls, 2010).

Iga lineaarne mudel on maatrikskujul kirjeldatav kui

$$Y = X\beta + \epsilon,$$

kus $Y = (y_1, \ldots, y_n)^T$ on uuritavate tunnuste vektor, $X : n \times m$ on argumenttunnuste maatriks, $\beta = (\beta_0, \ldots, \beta_{m-1})^T$ on tundmatute parameetrite vektor ja $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T$ tähistab mudeli prognoosivigu. Kui viimaste keskväärtus on null ning hajuvus konstantne, saame β väärtust hinnata vähimruutude meetodiga:

$$\hat{\beta} := \arg\min_{\beta}(Y - X\beta)^T(Y - X\beta) = \arg\min_{\beta} \sum_{i=1}^{n} \hat{\epsilon}_i^2.$$

Saadud hinnang on kujul

$$\hat{\beta} = (X^T X)^{-1}(X^T Y).$$

Üldjuhul võivad aga mudeli jäägid olla sõltuvad ning erineva hajuvusega. Olgu V jääkide vektori kovariatsioonimaatriks. Eeldades, et see on positiivselt määratud, saame selle spektraalse dekompositsiooni abil esitada kujul $V = ODO^T$, kus O on V omavektoritest moodustatud ortogonaalne maatriks ja D on diagonaalmaatriks, mille elementideks V omaväärtused. Moodustades maatriksi $V^{1/2} := OD^{1/2}O^T$, kus $D^{1/2}$ on D elementide ruutjuurtest koosnev diagonaalmaatriks, on ilmne, et kehtib $V^{1/2}V^{1/2} = V$. Olgu $V^{1/2}$ pöördmaatriksiks $V^{-1/2}$.

Loome uue mudeli

$$Y_* = X_*\beta + \epsilon_*,$$

kus $Y_* = V^{-1/2}Y$, $X_* = V^{-1/2}X$ ja $\epsilon_* = V^{-1/2}\epsilon$. Jääkide vektori ϵ_* kovariatsioonimaatriksiks on seega $V^{-1/2}V^{1/2}V^{1/2}V^{-1/2} = I$ ehk saadud mudeli vead on konstantse

hajuvusega. Saame tundmatuid parameetreid hinnata vähimruutude meetodil:

$$\hat{\beta} := \arg\min_{\beta}(Y_* - X_*\beta)^T(Y_* - X_*\beta) = \arg\min_{\beta}(Y - X\beta)^T V^{-1}(Y - X\beta).$$

Seega, üldistatud vähimruutude meetodil saadud hinnang on

$$\hat{\beta} = ({X_*}^T X_*)^{-1} {X_*}^T Y_* = (X^T V^{-1} X)^{-1} X^T V^{-1} Y.$$

2.2.2 LD-skoori regressioon

Käesolev alapeatükk põhineb Bulik-Sullivani *et al.* (2015) artiklitel.

Kehtigu kõik peatükis 2.1.2 mainitud eeldused. Et kirjeldada j-nda SNP-i mõju tunnusele 1, leiame vastava z-statistiku:

$$z_{1j} := \frac{\hat{\beta}_j}{\sigma(\hat{\beta}_j)}. \tag{5}$$

Selleks hindame β_j vähimruutude meetodil, mille tulemusena $\hat{\beta}_j = (X^T X)^{-1} X^T y_1$. Maatriksi $X^T X$ elementideks on $(X^T X)_{jk} = \sum_{i=1}^{N_1} X_{ij} X_{ik} = N_1 E(X_{ij} X_{ik}) \overset{(3)}{=} N_1 r_{jk}$. Seega, peadiagonaalil asetsevad elemendid $(X^T X)_{ii} = N_1$. Ülejäänud elementide puhul eeldavad artiklite autorid siinkohal, et keskeltläbi on need võrdsed nulliga. Eeldus tugineb ilmselt asjaolule, et üksteisest kaugel paiknevad SNP-id ongi enamasti sõltumatud. Tänu sellele saame β_j hinnanguks $\hat{\beta}_j = (N_1 I_M)^{-1} X^T y_1 = X^T y_1 / N_1$ ning $\sigma(\hat{\beta}_j) = 1/\sqrt{N_1}$. Järelikult on z-statistik leitav kui

$$z_{1j} = X^T y_1 / \sqrt{N_1}.$$

Analoogne tuletuskäik kehtib ka $z_{2j} = Y^T y_2 / \sqrt{N_2}$ kohta.

Lause 4. Eeldefineeritud tingimustel avaldub kahe tunnuse z-statistikute korrutiste keskväärtus kui

$$E(z_{1j} z_{2j} | \ell_j) = \frac{\sqrt{N_1 N_2}\rho_g}{M} \ell_j + \frac{N_s \rho}{\sqrt{N_1 N_2}}.$$

Tõestus. Keskväärtuse saame leida tinglikke keskväärtuseid keskmistades:

$$E(z_{1j} z_{2j} | \ell_j) = E[E(z_{1j} z_{2j} \,| X, Y) | \ell_j].$$

Leiame esmalt z-statistikute korrutiste tingliku keskväärtuse fikseeritud X ja Y korral:

$$\begin{aligned}
E(z_{1j}z_{2j}\,|X,Y) &\overset{(5)}{=} \frac{1}{\sqrt{N_1N_2}}E(X_j^T y_1 Y_j^T y_2\,|X,Y)\\
&= \frac{1}{\sqrt{N_1N_2}}E(X_j^T y_1 y_2^T Y_j\,|X,Y)\\
&= \frac{1}{\sqrt{N_1N_2}}X_j^T E[(X\beta+\epsilon)(Y\gamma+\delta)^T|X,Y]Y_j\\
&= \frac{1}{\sqrt{N_1N_2}}X_j^T [XE(\beta\gamma^T)Y^T + XE(\beta\delta^T) + E(\epsilon\gamma^T)Y^T + E(\epsilon\delta^T)]Y_j\\
&= \frac{1}{\sqrt{N_1N_2}}X_j^T [XE(\beta\gamma^T)Y^T + E(\epsilon\delta^T)]Y_j\\
&= \frac{1}{\sqrt{N_1N_2}}\left(\frac{\rho_g}{M}X_j^T XY^T Y_j + \rho_e X_j^T I_{N_s}^{*1} Y_j\right). \qquad (6)
\end{aligned}$$

Viimane võrdus tuleneb eeltoodud kovariatsioonimaatriksite $\Sigma(\beta,\gamma)$ ja $\Sigma(\epsilon,\delta)$ esitustest. Saadud avaldise keskväärtuse leiame osade kaupa:

$$\begin{aligned}
E(X_j^T XY^T Y_j|\ell_j) &= E\left[\sum_{m=1}^{M}\left(\sum_{i=1}^{N_1}X_{ij}X_{im}\right)\left(\sum_{i=1}^{N_2}Y_{ij}Y_{im}\right)\right]\\
&= \sum_{m=1}^{M}E\left(\sum_{i=1}^{N_1}X_{ij}X_{im}\right)\left(\sum_{i=1}^{N_2}Y_{ij}Y_{im}\right)\\
&= \sum_{m=1}^{M}\left[\sum_{i=1}^{N_1}\sum_{k=1}^{N_2}E(X_{ij}X_{im}Y_{kj}Y_{km})\right]\\
&= \sum_{m=1}^{M}\left[\sum_{X_{ij}\perp Y_{kj}}^{N_1}\sum^{N_2}E(X_{ij}X_{im})E(Y_{kj}Y_{km}) + \sum_{X_{ij}\not\perp Y_{kj}}\sum E(X_{ij}X_{im}Y_{kj}Y_{km})\right]\\
&\overset{(3)}{=} \sum_{m=1}^{M}\left[\sum_{i=1}^{N_1N_2-N_s}r_{jm}^2 + \sum_{i=1}^{N_s}E(X_{ij}^2X_{im}^2)\right].
\end{aligned}$$

Valemist $E(X_{ij}^2X_{im}^2) = D(X_{ij}X_{im}) = D(X_{ij})D(X_{im}) + E(X_{ij}X_{im})^2$ saame:

$$\begin{aligned}
E(X_j^T XY^T Y_j|\ell_j) &= \sum_{m=1}^{M}[(N_1N_2-N_s)r_{jm}^2 + N_s(1+r_{jm}^2)]\\
&= N_1N_2\sum_{m=1}^{M}r_{jm}^2 + \sum_{m=1}^{M}N_s\\
&\overset{(4)}{=} N_1N_2\ell_j + MN_s.
\end{aligned}$$

Veel on vaja leida:

$$\begin{aligned} E(X_j^T I_{N_s}^{*1} Y_j) &= E\left(\sum_{i=1}^{N_s} X_{ij} Y_{ij}\right) \\ &= \sum_{i=1}^{N_s} E(X_{ij}^2) \\ &= N_s. \end{aligned}$$

Kokku saame avaldise (6) keskväärtuseks:

$$\begin{aligned} E(z_{1j} z_{2j} | \ell_j) &= \frac{1}{\sqrt{N_1 N_2}} \left[\frac{\rho_g}{M} (N_1 N_2 \ell_j + M N_s) + \rho_e N_s \right] = \\ &= \frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{N_s \rho_g}{\sqrt{N_1 N_2}} + \frac{N_s \rho_e}{\sqrt{N_1 N_2}} = \\ &= \frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{N_s (\rho_g + \rho_e)}{\sqrt{N_1 N_2}} = \\ &= \frac{\sqrt{N_1 N_2} \rho_g}{M} \ell_j + \frac{N_s \rho}{\sqrt{N_1 N_2}}. \end{aligned}$$

■

Lause 4 ütleb, et geneetilise kovariatsiooni tuvastamiseks tuleb hinnata lineaarne regressioonimudel, kus uuritavaks tunnuseks on kahe fenotüübi z-statistikute korrutis ning argumenttunnuseks LD-skoor ℓ_j. Vähimruutude meetodil leitavast LD-skoori kordajast saamegi avaldada geneetilise kovariatsiooni hinnangu. Sellist hindamisviisi nimetatakse LD-skoori regressiooni meetodiks.

Paneme tähele, et kui kahe fenotüübi rolli võtta üks ja sama tunnus, hindab LD-skoori regressiooni meetod vastava fenotüübi päritavust (järeldub lausetest 1 ja 2). Sellisel juhul $N_1 = N_2 = N_s$ ja $\rho = 1$, seega

$$E(z_{ij}^2 | \ell_j) = \frac{N_i h_i^2}{M} \ell_j + 1. \tag{7}$$

Et meetod annaks täpsemaid tulemusi, kasutatakse üldistatud vähimruutude meetodit ehk kaalutakse vaatluseid nende dispersiooni pöördväärtusega. Dispersiooni leidmiseks kasutame autorite toodud valemit

$$\begin{aligned} D(z_{1j} z_{2j} | \ell_j) &= D(z_{1j} | \ell_j) D(z_{2j} | \ell_j) + E(z_{1j} z_{2j} | \ell_j)^2 \\ &= E(z_{1j}^2 | \ell_j) E(z_{2j}^2 | \ell_j) + E(z_{1j} z_{2j} | \ell_j)^2, \end{aligned}$$

millest lause 4 põhjal saame:

$$D(z_{1j}z_{2j}|\ell_j) = \left(\frac{N_1 h_1^2}{M}\ell_j + 1\right)\left(\frac{N_2 h_2^2}{M}\ell_j + 1\right) + \left(\frac{\sqrt{N_1 N_2}\rho_g}{M}\ell_j + \frac{N_s\rho}{\sqrt{N_1 N_2}}\right)^2. \quad (8)$$

Näeme, et kaalufunktsioon $D(z_{1j}z_{2j}|\ell_j)^{-1}$ sisaldab parameetreid h_1^2, h_2^2, ρ ja N_s, mida me ei pruugi alati teada, rääkimata ρ_g väärtusest. Seetõttu leiame kaalud kahe sammuga.

1. Hindame mudeli, kasutades kaalufunktsioonis päritavuse hinnanguid, mille oleme omakorda LD-skoori regressiooni meetodiga leidnud. Korrutise ρN_s võtame võrdseks nulliga ning $\hat{\rho}_g := (\sqrt{N_1 N_2})^{-1} \sum_{j=1}^{M} z_{1j}z_{2j}$.

2. Hindame mudeli, asendades kaalufunktsioonis ρN_s ja ρ_g hinnangud esimesel sammul saadutega.

Et suurtest LD regioonidest pärit markereid liiga suure kaaluga arvesse ei võetaks, korrutame kaalufunktsiooni ka LD-skoori pöördväärtusega, mis on leitud vaid nende SNP-ide pealt, mida regressioonis kasutame. Toome selgituseks lihtsa näite.

Olgu kokku kaks SNP-i X_1 ja Y_1, mis vaadeldavaid fenotüüpe mõjutavad. Kuulugu neist esimene haplotüüpi (ehk LD-plokki), mis koosneb kahest täielikult LD-s olevast markerist X_1 ja X_2, seega $\ell_{X_1} = \ell_{X_2} = 2$. Teine SNP olgu aga pärit kolmesest samuti täielikus LD-s olevast markerite plokist ehk $\ell_{Y_1} = \ell_{Y_2} = \ell_{Y_3} = 3$.

Markeril X_2 puudub tegelik mõju, kuid X_1-ga LD-s olemise tõttu osutuvad nende z-statistikud uuringus täpselt samaks. Analoogselt ka teises plokis olevate markerite korral. Kovariatsiooni hindamisel võetaks seetõttu esimest mõjutajat arvesse kaks korda ning teist kolm korda – vaatluste läbi jagamine LD-skooriga tagab siinkohal realistlikuma tulemuse.

LD-skoori regressiooni meetodi hinnangute standardhälbe leidmine põhineb jackknife'i meetodil, kuid käesolevas bakalaureusetöös seda ei käsitleta.

3 Praktiline osa

3.1 Simulatsioon

Veendumaks, et LD-skoori regressiooni meetod ka praktikas toimib, viidi läbi simulatsioon. Eesmärk oli simuleerida fikseeritud geneetilise korrelatsiooniga fenotüüpide andmeid ning vaadata, kas LD-skoori regressiooni hinnang vastab ootustele. Lisaks sooviti võrrelda hinnangu käitumist erinevates olukordades.

3.1.1 Ülesehitus

Reaalseid geneetilisi mõjusid kirjeldavaid andmeid jäljendada on raske. Eesmärgi saavutamiseks oli aga peamine, et simuleeritud andmed vastaksid peatükis 2.1.2 toodud eeldustele.

Antud juhul sooviti näha hinnangute käitumist kahe konkreetse ja fikseeritud fenotüübi korral, mis üheskoos vastaksid soovitud päritavuste $(h^2_{1_{fix}}, h^2_{2_{fix}})$ ja korrelatsiooni $(r_{g_{fix}})$ väärtustele. Fenotüübi fikseerimisega kaasnevad aga ka fikseeritud geenimarkerite mõjud, millega rikuti eeldust, et β ja γ peaksid olema juhuslikud ning keskväärtusega null. Et lause 1 ja 2 sellegipoolest kehtiks, tuli LD-skoori regressiooni kaasata vaid üksteisest sõltumatuid markereid. Sellisel juhul

$$
\begin{aligned}
\rho_g(y_1, y_2) &= \operatorname{cov}(X_i\beta, X_i\gamma |\, \beta, \gamma) \\
&\overset{\text{(lause 2)}}{=} \sum_{j\neq k}^{M}\sum^{M} E(X_{ij}X_{ik}\beta_j\gamma_k |\, \beta, \gamma) + \sum_{j=1}^{M} E(X_{ij}^2\beta_j\gamma_j |\, \beta, \gamma) \\
&\overset{X_{ij}\perp X_{ik}}{=} \sum_{j\neq k}^{M}\sum^{M} E(X_{ij})E(X_{ik})\beta_j\gamma_k + \sum_{j=1}^{M} E(X_{ij}^2)\beta_j\gamma_j \\
&\overset{E(X_{ij}=0)}{=} \sum_{j=1}^{M} E(X_{ij}^2)\beta_j\gamma_j \\
&\overset{E(X_{ij}^2)=1}{=} \sum_{j=1}^{M} \beta_j\gamma_j.
\end{aligned}
$$

Päritavuse korral on arutelu sarnane (fikseeritud β näitel):

$$
\begin{aligned}
h_1^2 &\overset{\text{(lause 1)}}{=} D\left(\sum_{j=1}^{M} X_{ij}\beta_j \,\middle|\, \beta\right) \\
&\overset{X_{ik}\perp X_{ij}}{=} \sum_{j=1}^{M} D(X_{ij}\beta_j|\,\beta) \\
&\overset{X_{ij}\perp\beta_j}{=} \sum_{j=1}^{M} \beta_j^2 D(X_{ij}) \\
&\overset{D(X_{ij})=1}{=} \sum_{j=1}^{M} \beta_j^2.
\end{aligned}
$$

Kirjeldatud olukorra loomiseks genereeriti esmalt genotüüpide maatriks X, kus markerid olid plokiti sõltuvad. Iga plokk oli juhusliku suurusega vahemikus 5 kuni 20. LD-skooride arvutamisel võeti kõik sõltuvused arvesse, kuid regressioonimudeli jaoks hinnati igast plokist vaid ühe markeri z-statistikud (mõlema fenotüübi jaoks). Juhuslikud keskkonnamõjud ϵ_i ja δ_i genereeriti normaaljaotusest dispersioonidega vastavalt $1 - h^2_{1_{fix}}$ ja $1 - h^2_{2_{fix}}$, et tagada tingimus $D(y_{1_i}) = D(y_{2_i}) = 1$. Mõlema fenotüübi kirjeldamiseks kasutati sama genotüüpide maatriksit, seega $N_1 = N_2 = N_s$.

Iga huvipakkuva aspekti uurimiseks korrati andmete simuleerimise ja nende pealt hinnangu leidmise tsüklit 500 korda. Loodavate LD-plokkide arv igas andmestikus oli 200, mis tegi SNP-ide koguarvuks umbes 2 500. Valimimahuks määrati 10 000. Simulatsiooniprogrammi kood on leitav lisas 1.

3.1.2 Tulemused

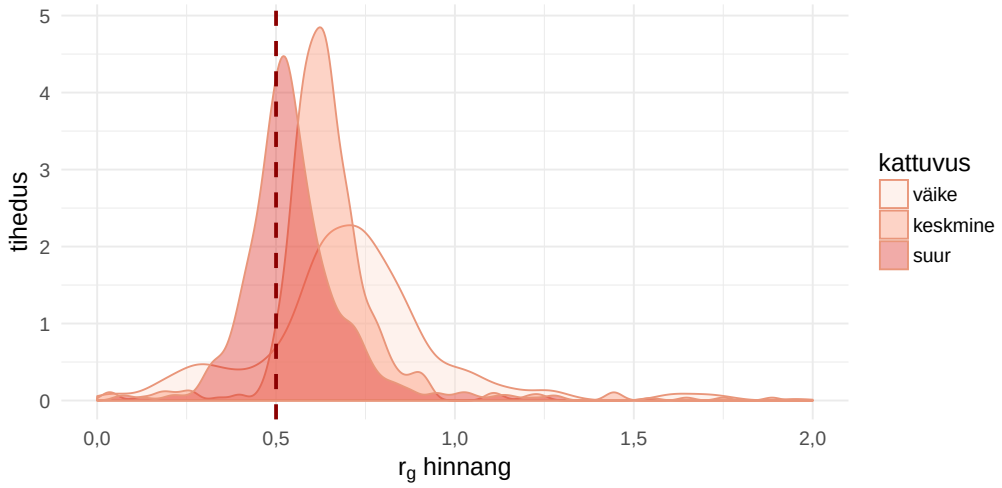
Alustuseks simuleeriti andmestikke, kus mõlema fenotüübi fikseeritud päritavus oli 0,5 ning nendevaheline geneetiline korrelatsioon 0,6. Keskkonnamõjude vastassuunalisuse tõttu oli aga fenotüüpide vaheline korrelatsioon tervikuna negatiivne, keskmiselt ligikaudu -0,2. Eesmärk oli näha, kas LD-skoori regressioon suudab üldisele negatiivsele korrelatsioonile vaatamata tuvastada keskmisest tugevamat geneetilist korrelatsiooni. Saadud 500 geneetilise korrelatsiooni hinnangu keskmiseks osutus 0,64 standardhälbega 0,23. Hinnangu keskväärtuse 95% usaldusintervalliks on seega (0,62...0,66), kust tegelik geneetilise korrelatsiooni väärtus siiski välja jääb.

Katsetati ka vastupidist olukorda: positiivselt korreleeritud fenotüübid, mille vahel geneetiline korrelatsioon puudub. Simuleeritud fenotüüpide korrelatsioon oli keskmiselt 0,51. LD-skoori regressiooni meetod suutis sellegipoolest geneetilise korrelatsiooni puudumise tuvastada. Hinnangute keskmine tuli null ja standardhälve 0,09. Näeme, et LD-skoori meetod suudab fenotüübilist ja geneetilist korrelatsiooni selgelt eristada. Arvestades väga väikest andmemahtu (reaalsete uuringutega võrreldes), annab meetod keskmiselt küllaltki (absoluutväärtuselt) lähedasi tulemusi tegelikule fikseeritud väärtusele.

Edasi sooviti võrrelda, kuidas mõjutab hinnangu täpsust see, kui paljud geneetilised mõjutajad on kahel fenotüübil ühised. Nimelt, kuna $\rho_g(y_1, y_2) = \sum_{j=1}^{M} \beta_j \gamma_j$, siis on võimalik tunnuseid tugevalt (geneetiliselt) korreleeruma saada nii mõne üksiku kui ka väga paljude ühiste mõjutavate SNP-idega. Esimesel juhul tuleb vastavate SNP-ide efektisuurused üsna suureks määrata, teisel juhul aga geenimõjud mitmete markerite peale ära jaotada, seega iga mõju üksikuna on pigem väike.

Vaatluse alla võeti kolm erinevat varianti: fenotüüpi mõjutas kas iga neljakümnes, iga kahekümne viies või iga kahekümnes SNP. Kui kahe fenotüübi kirjeldamisel esimene ja viimane variant kombineeriti, saadi ühiste mõjutavate markerite arvuks 63, teise ja viimase korral analoogselt 26 ning esimese ja teise korral vaid 13. Mida vähem oli ühiseid mõjutajaid, seda rohkem oli aga selliseid markereid, mis panustasid eri fenotüüpide päritavusse, kuid mitte nendevahelisse korrelatsiooni.

Simulatsioonist selgus, et suurem mõjutajate kattuvus annab geneetilise korrelatsiooni hindamisel tegelikule parameetrile absoluutväärtuselt lähedasemaid tulemusi. Kõigi kolme kombinatsiooniga simuleeriti andmeid, kus $h^2_{1_{fix}} = h^2_{2_{fix}} = 0,8$ ja $r_{g_{fix}} = 0,5$. Suurima kattuvusega kombinatsiooni korral saadi hinnangute keskmiseks 0,55 ja standardhälbeks 0,15. Teise kombinatsiooni ehk nii-öelda keskmise kattuvuse korral olid vastavad näitajad 0,65 ja 0,17 ning vähima kattuvuse korral 0,74 ja 0,44. Tegelikku fikseeritud geneetilise korrelatsiooni väärtust ei kata hinnangu keskväärtuse 95% usaldusintervall ühelgi juhul. Nendeks saadi vastavalt (0,54...0,57), (0,63...0,66) ja (0,70...0,78). Saadud hinnangute jaotuseid on võrreldud joonisel 1, kus vertikaalne joon tähistab tegelikku geneetilist korrelatsiooni. Jooniselt on välja jäetud mõned üksikud äärmuslikud väärtused.

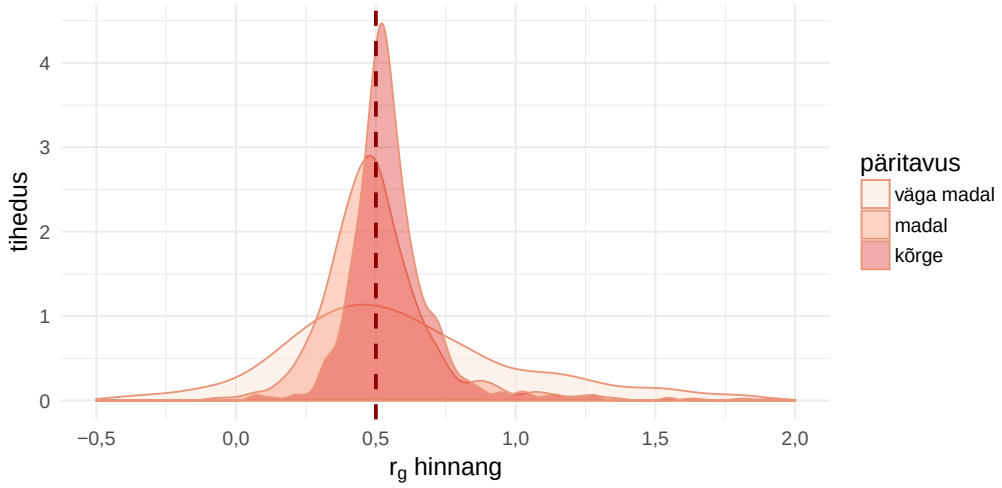


Joonis 1. Hinnangu jaotus erineva ühiste mõjutavate SNP-de arvu korral

Kuna kirjanduses leidub väiteid, et fenotüüpide madala päritavuse korral võib geneetilise korrelatsiooni hinnang üsna ebatäpseks osutuda, sooviti lähemalt uurida ka seda aspekti. Omavahel võrdlemiseks simuleeriti jällegi kolm erinevat varianti: väga madala ($h^2_{1_{fix}} = h^2_{2_{fix}} = 0,05$), madala ($h^2_{1_{fix}} = 0,1$ ja $h^2_{2_{fix}} = 0,2$) ning kõrge ($h^2_{1_{fix}} = h^2_{2_{fix}} = 0,8$) päritavusega fenotüübid. Geneetiline korrelatsioon fikseeriti kõigi variantide korral keskmise tugevusega ($r_{g_{fix}} = 0,5$).

Tulemused olid ootuspärased: madalaim päritavus andis ebatäpseima hinnangu. Päritavusega 0,05 osutus 500 hinnangu keskmiseks 0,69 ning nende standardhälve oli 0,77 ehk senistest näidetest kõrgeim. Hinnangu keskväärtuse 95% usaldusintervalliks saadi (0,63...0,76). Selgus, et isegi vähene päritavuse suurendamine annab juba tunduvalt paremaid tulemusi. Nimelt, teise variandi ehk madala päritavuse korral saadi hinnangud, mille keskmine oli 0,52 ja standardhälve 0,27.

Üllataval kombel erines kõrge päritavusega saadud hinnangute keskmine (0,55) absoluutväärtuselt tegelikust geneetilisest korrelatsioonist rohkem kui madala päritavuse korral, kuid nende varieeruvus oli oluliselt väiksem (standardhälve 0,15). Võrdluseks leiti ka kahel viimasel juhul hinnangu keskväärtuse 95% usaldusintervall, milleks kõrge päritavuse korral saadi (0,54...0,56) ja madala päritavuse korral (0,50...0,54). Selgus, et viimane oli uuritud juhtudest ainus, mille korral hinnangu keskväärtuse 95% usaldusintervall fikseeritud tegelikku väärtust 0,5 sisaldas. Analoogselt eelmise näitega on joonisel 2 toodud võrdluseks hinnangute tihedusgraafik.



Joonis 2. Hinnangu jaotus erinevate päritavuse väärtuste korral

Kõik simulatsioonis saadud hinnangud olid küllaltki suure hajuvusega ning enamik leitud usaldusintervalle ei katnud tegelikku geneetilise korrelatsiooni väärtust. Ilmselt on see tingitud väiksest SNP-ide arvust, kuna reaalsed GWAS-ide andmestikud sisaldavad üldjuhul üle miljoni markeri efektisuuruseid, kusjuures minimaalseks kaasatavate SNP-ide arvuks loetakse 100 000 (Zheng *et al.*, 2012). Et näha, kuidas SNP-ide arvu suurendamine hinnangu omadusi muudab, viidi üks näide läbi ka 300 LD-plokiga, mis andis markerite arvuks ligikaudu 3 800. Parameetrid fikseeriti sarnaselt eeltoodud kõrge päritavuse näitega. Hinnangute keskmine osutus absoluutväärtuselt tõepoolest tegelikule lähemaks (0,54) ja vähenes ka hinnangute standardhälve (0,13). Usaldusintervall (0,53…0,56) siiski oodatud väärtust ei sisaldanud.

3.2 Geneetilise korrelatsiooni hindamine

Üheks peamiseks epidemioloogia eesmärgiks on mõista, kuidas on omavahel seotud erinevad inimestel avaldunud tunnused ja haigused. Enamasti uuritakse, kuidas üks või teine tunnus mingi haiguse kujunemist mõjutab. Geneetilise korrelatsiooni tuvastamine selliste potentsiaalsete mõjutajate vahel võib sealjuures anda meile olulist lisainformatsiooni. (Bulik-Sullivan *et al.*, 2015)

Bakalaureusetöö raames rakendati LD-skoori regressiooni meetodit, et hinnata geneetilisi korrelatsioone seitsme tunnuse vahel. Kõik uuritud tunnused on seotud erinevate metaboolsete haigustega – ennekõike teist tüüpi diabeediga.

3.2.1 Andmestike kirjeldus

Geneetiliste korrelatsioonide hindamiseks kasutati GIANT (ingl *Genetic Investigation of Anthropometric Traits*) ja MAGIC (ingl *the Meta-Analyses of Glucose and Insulin-related traits Consortium*) konsortsiumite andmestikke, kuhu on panustanud ka TÜ Eesti Geenivaramu. Neist esimesse on koondatud antropomeetrilisi tunnuseid uurivate GWAS-ide meta-analüüside andmed. Teises konsortsiumis leidub sarnaselt vere glükoosi- ja insuliinitasemega seonduvate tunnuste andmeid.

Antropomeetria vallast valiti analüüsimiseks kehamassiindeks ehk BMI (Locke *et al.*, 2015) ning BMI suhtes korrigeeritud talje ja puusa ümbermõõdu suhe ehk WHRadjBMI (Shungin *et al.*, 2015). Teisest konsortsiumist kasutati proinsuliini (Strawbridge *et al.*, 2011), glükeeritud hemoglobiini HbA1c (Soranzo *et al.*, 2010), paastu vereglükoosi ehk FG, insuliini resistentsuse näitaja (HOMA-IR) ja beeta-rakkude funktsionaalsuse indeksi (HOMA-β) andmestikke (Dupuis *et al.*, 2010).

Iga andmestik sisaldab informatsiooni üle miljoni SNP-i kohta. Mugavamaks kasutamiseks viidi kõik andmestikud LDSC (Bulik-Sullivan, 2015) programmi abil samasse formaati: esimeses veerus SNP-i identifitseeriv rs-number, teises veerus referentsalleeli (nukleotiidi) tähis, kolmandas riskialleeli tähis, neljandas z-statistik (5) ning viiendas geeniefekti hindamiseks läbi viidud uuringu valimimaht. Viimane jäi eri andmestikes suurusjärku 40 kuni 300 tuhat.

Kuna tegemist on GWAS-ide koondandmetega, pole nende pealt võimalik markerite LD-skoore arvutada. Seetõttu kasutati juba olemasolevaid LD-skooride andmestikke (kokku 22 – iga kromosoomi jaoks üks), mis on hinnatud Euroopa päritolu geeniandmete pealt (Broad Institute, 2016). Ka nendes andmestikes on üheks tunnuseks SNP-i identifitseeriv rs-number. Peale selle ja LD-skoori väärtuse on iga SNP-i kohta välja toodud, mitmendas kromosoomis see asub ning kui kõrge on selle harvemini esineva alleeli sagedus ehk MAF (ingl *minor allele frequency*). Ülejäänud kahte andmestikes olevat tunnust arvutustes ei kasutatud.

3.2.2 LD-skoori meetodi rakendamine

LD-skoori regressiooni meetodit saab korraga rakendada vaid kahe fenotüübi vahelise geneetilise korrelatsiooni hindamiseks. Lõppeesmärk oli hinnata kõigi seitsme tunnuse geneetilisi seoseid, kuid järgnevalt kirjeldatakse arvutuskäiku vaid BMI ja FG näitel. Hinnangute leidmisel lähtuti meetodi teoreetilisest kirjeldusest ning Bulik-Sullivani *et al.* (2015) artiklites toodud lisamärkustest. Lisaks uuriti ja võeti arvesse Pythoni programmis LDSC (Bulik-Sullivan, 2015) kasutatud võtteid, mida artiklites ei mainitud.

Meetodi rakendamiseks tuli alustada andmestike ühendamisest. Kõigepealt koondati eri kromosoomide SNP-ide LD-skoorid ühte andmestikku, mis kokku sisaldas informatsiooni 1 293 150 markeri kohta. Nii BMI kui ka FG andmestik koosnes algselt 1 217 311 reast, kuid pärast puuduvate väärtuste eemaldamist jäi alles vastavalt 1 046 190 ja 1 063 492 rida. Neist omakorda 1 042 509 kirjeldasid samu markereid, mis kaasati ühte ühisesse andmestikku. Viimane pandi kokku algselt loodud LD-skooride andmestikuga – nii jäi kokkuvõttes analüüsi 1 040 941 markerit.

Arvutustes on parameetrina M soovitatud kasutada nende SNP-ide koguarvu, mida arvestati LD-skooride arvutamisel, kuid mille MAF jääb vahemikku 0,05–0,5. Korrelatsiooni hindamisel ei mängi M tegelikult küll rolli, kuna kovariatsiooni ja päritavuste hinnanguid jagades see taandub. Sellegipoolest võeti soovitust arvesse, kuna see tagab realistlikumaid vahehinnanguid (päritavused ja kovariatsioon). Kriteeriumile vastavaid markereid oli kokku 1 176 350. Kuna BMI geeniefektide andmestikku on koondatud erinevate uuringute tulemusi, pole N_1 konstantne, vaid varieerub vahemikus 156 008–322 156. FG tulemused on leitud valimimahuga kuni 46 186, kuid täpsem info iga SNP-i kohta eraldi puudub. Seetõttu kasutati N_2 rollis konstantselt seda väärtust.

LD-skoori regressiooni meetod eeldab, et mõlema vaadeldava fenotüübi analüüsimisel on SNP-ide efektisuurused leitud sama referentsalleeli suhtes. Nende markerite z-statistikud, mis sellele eeldusele ei vastanud, tuli seetõttu muuta vastassuunaliseks (ehk $z_{ij} = -z_{ij}$). Tunnuse BMI z-statistikud varieerusid vahemikus -26,2 kuni 25,6,

teisel tunnusel vahemikus -18,3 kuni 18,1. Leitud kahe fenotüübi z-statistikute korrutiste keskväärtus 0,11 viitas positiivsele (geneetilisele) korreleeritusele. Andmestikus olevate SNP-ide keskmine LD-skoori väärtus oli ligikaudu 23, maksimaalne aga 305.

Hinnangute leidmiseks kirjutati kaks funktsiooni (vt lisa 2), millest üks hindab fenotüübi päritavust ning teine kahe fenotüübi vahelist geneetilist korrelatsiooni. Päritavuse funktsioonis võetakse regressioonimudelis kasutatavate kaalude tuletamisel eeskuju Bulik-Sullivani (2015) kirjutatud programmist, kuna artiklitest sellekohast informatsiooni ei leitud. Parameeter h_i^2 algväärtustatakse kui

$$\hat{h}_i^2 = M\frac{(\overline{z_{ij}^2} - \hat{\alpha}_i)}{\overline{N_i \ell_j}},$$

kus $\hat{\alpha}_i = 1$ tähistab mudeli (7) vabaliiget. Puudusid algandmed, et leida LD-skoorid, milles arvestatakse vaid regressiooni kaasatud SNP-ide sõltuvusi. Seetõttu kasutati kaaludes algupäraseid LD-skoore, mis ei tohiks tulemust oluliselt muuta. Kaalud ise on kokkuvõttes kujul

$$w_{ij} = \frac{1}{\ell_j}\left[\left(\hat{\alpha}_i + \frac{N_i \ell_j \hat{h}_i^2}{M}\right)^2\right]^{-1}.$$

Pärast lineaarse regressioonimudeli (7) hindamist saavad $\hat{h}_i^2$ ja $\hat{\alpha}_i$ uued väärtused, et seejärel nendega protsessi korrata. Jõudmaks sama tulemuseni nagu LDSC (Bulik-Sullivan, 2015) programm, korratakse mudeli ümberhindamist kokku kolm korda, pärast mida toimub praktiliselt koondumine.

Geneetilise korrelatsiooni funktsioonis kasutatakse vaatluste kaalumisel kõiki päritavuse funktsiooniga hinnatud parameetreid. Seega, valemit (8) kasutades saavad kaalud kuju

$$w_j = \frac{1}{\ell_j}\left[\left(\frac{N_1\hat{h}_1^2}{M}\ell_j + \hat{\alpha}_1\right)\left(\frac{N_2\hat{h}_2^2}{M}\ell_j + \hat{\alpha}_2\right) + \left(\frac{\sqrt{N_1 N_2}\hat{\rho}_g}{M}\ell_j + \hat{\alpha}\right)^2\right]^{-1},$$

kus $\hat{\alpha}$ on esialgu võrdne nulliga ja

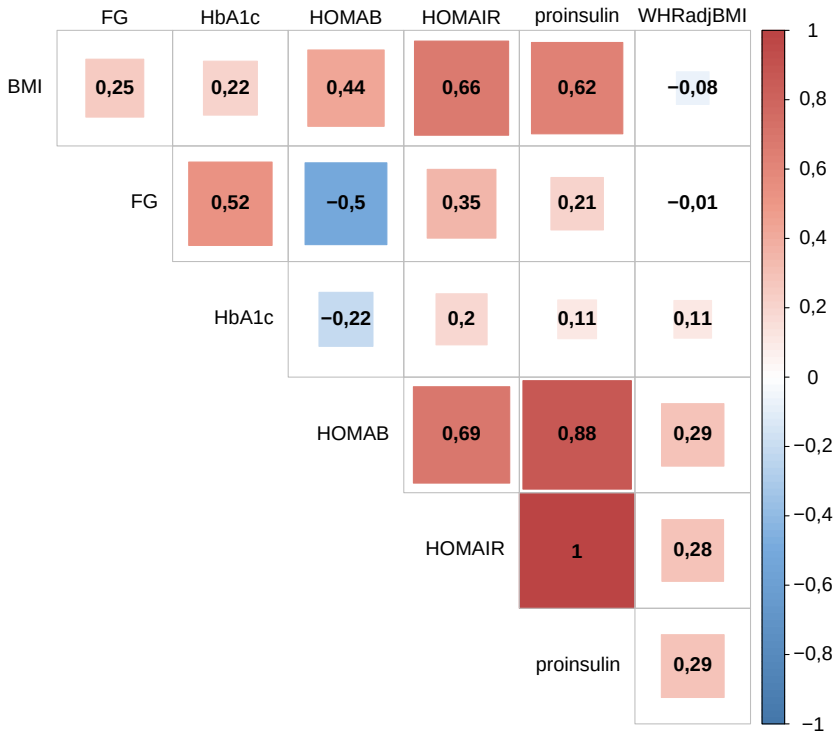
$$\hat{\rho}_g = M\frac{\overline{z_{1j}z_{2j}}}{\ell_j\sqrt{N_1 N_2}}. \tag{9}$$

Analoogselt päritavuse funktsiooniga hinnatakse lõpliku kovariatsiooni hinnangu saamiseks parameetrid $\hat{\rho}_g$ ja $\hat{\alpha}$ kolmel korral ümber. Funktsiooni lõpus kombineeritakse leitud väärtused $\hat{h}_1^2$, $\hat{h}_2^2$ ja $\hat{\rho}_g$ lause 3 põhjal korrelatsiooni hinnanguks.

3.2.3 Tulemused

LD-skoori regressiooni meetodiga leiti, et BMI päritavus on hinnanguliselt 0,14 ning FG oma veidi madalam: 0,10. Vabaliikmeid hindas mudel väärtustega 0,65 ja 0,99. Valemit (9) rakendades saadi geneetilise kovariatsiooni algväärtuseks 0,0542, mis pärast mudeli hindamise esimest kordust kahanes ligikaudu 0,0288 peale. Järgmiste korduste käigus nimetatud väärtus oluliselt ei muutunud: võrreldes järgmisega toimus muutus alates viiendast komakohast ning viimase ümberkaalumisega muutus hinnang alates seitsmendast kümnendkohast.

Saadud (ümardamata) väärtused andsid kokku geneetilise korrelatsiooni hinnanguks ligikaudu 0,25. Tegu on keskmisest nõrgema, kuid siiski märkimisväärse geneetilise seosega. On leitud, et vaadeldavate fenotüüpide vaheline korrelatsioon tervikuna on 0,18 ehk veidi madalam (Amato *et al.*, 2014). Analoogselt uuriti geneetilisi seoseid kõigi seitsme tunnuse vahel, mille tulemused on välja toodud alljärgneval joonisel.



Joonis 3. Geneetiliste korrelatsioonide hinnangud

Selgus, et uuritud tunnustest kõrgeima päritavusega on glükeeritud hemoglobiin ehk HbA1c (0,17) ja madalaimaga insuliini resistentsuse näitaja HOMA-IR (0,05). Tunnused HOMA-IR ja proinsuliin, mis on mõlemad insuliini tootmisega seotud, osutusid täielikult geneetiliselt korreleerituks. Kõrgeid korrelatsioone leiti mitmeid teisigi. Märkimist väärib kehamassiindeksi tugev seotus insuliininäitajatega – pealtnäha täiesti erinevaid omadusi kirjeldavad tunnused, kuid keskmisest tugevam geneetiline korrelatsioon viitab vastupidisele. Esines ka üksikuid negatiivseid korrelatsioone, millest tugevaim ehk -0,5 oli tunnuste FG ja HOMA-β vahel. Tunnuste BMI ja WHRadjBMI vaheline geneetiline korrelatsioon osutus pea olematuks (-0,08), mis oli küllaltki ootuspärane, arvestades WHRadjBMI sisulist olemust.

Geneetiliste korrelatsioonide tuvastamise kasutegureid on erinevaid. Näiteks võimaldavad need hinnata, kas mingil meid huvitaval tunnusel on eraldiseisev geneetiline komponent või on kahe tunnuse väärtuste erinevused tingitud vaid keskkonnast. Vaatame lähemalt täielikult geneetiliselt korreleerituks osutunud tunnuste HOMA-IR ja proinsuliin näidet. On leitud, et proinsuliin on oluline suremusega seotud biomarker, mistõttu võib täheldada sarnast mõju ka HOMA-IR puhul. Kui uurida insuliini resistentsusindeksi ja proinsuliini mõju suremusele korraga, osutub HOMA-IR seevastu ebaoluliseks. Üldiseks nende tunnuste vaheliseks korrelatsiooniks on hinnatud 0,53. Järelikult on nende väärtused mõjutatatud erinevatest keskkonnafaktoritest. Geneetilisest korrelatsioonist teadlik olles saame oletada, et HOMA-IR mõju suremusele tekib vaid tugeva geneetilise korreleerituse tõttu proinsuliiniga ning tema enda seos suremusega pole tõenäoliselt kausaalne. (Alssema *et al.*, 2010)

Kokkuvõte

Bakalaureusetöö eesmärk oli anda ülevaade geneetilist korrelatsiooni hindavast LD-skoori regressiooni meetodist ning veenduda selle paikapidavuses, et seda ülegenoomsete assotsiatsioonuuringute andmetel rakendada.

Töös seletati lahti erinevad geneetika valdkonna mõisted, mille tundmist LD-skoori regressiooni meetodist arusaamine eeldab. Seejärel anti ülevaade meetodi olemusest, mida toetati mitme tõestusega. Varasema publikatsiooniga võrreldes said viimased oluliselt täiendust, kusjuures parandati ka mitmeid vigu.

Et võrrelda, kuidas LD-skoori regressiooni meetod erinevatel juhtudel käitub, viidi läbi simulatsioon. Tulemused näitasid, et meetod teeb selgelt vahet üldisel tunnustevahelisel ning geneetilisel korrelatsioonil. Simulatsioon kinnitas ka väidet, et väga madala päritavusega fenotüüpide korral võib geneetilise korrelatsiooni hinnang olla suure nihkega. Lisaks viitasid tulemused sellele, et suurem kahe fenotüübi ühiste geneetiliste mõjutajate arv annab oodatavale tulemusele absoluutväärtuselt lähedasemaid hinnanguid.

Autori oletusel annaks simulatsiooniga uurida hinnangu käitumist ka muudes aspektides. Täpsemate ja huvitavamate tulemuste saamiseks oleks aga vaja tunduvalt suurem arv markereid simuleerida, kuna meetod on tegelikult loodud töötama GWAS-ide andmetel, mille mahud küündivad töös loodust sadu või lausa tuhandeid kordi kõrgemale. Arvamus tugineb ka töös katsetatud rohkemate markeritega simulatsioonile, mille käigus leitud hinnangud näisid olevat täpsemad.

Viimases peatükis rakendati LD-skoori regressiooni meetodit ka reaalsetel ülegenoomsete assotsiatsioonuuringute andmetel, mille tulemusena leiti seitsme tunnuse vahelised geneetiliste korrelatsioonide hinnangud. Saadud hinnangud varieerusid vahemikus -0,5 kuni 1.

Kasutatud kirjandus

Alssema, M. *et al.* (2010). Proinsulin Concentration Is an Independent Predictor of All-Cause and Cardiovascular Mortality. *Diabetes Care*, 28(4), 860–865.

Amato, M. C. *et al.* (2014). Visceral Adiposity Index (VAI) Is Predictive of an Altered Adipokine Profile in Patients with Type 2 Diabetes. *PLoS ONE*, 9(3): e91969.

Broad Institute (2016). https://data.broadinstitute.org/alkesgroup/LDSCORE/ (vaadatud 28.03.2018).

Bulik-Sullivan, B. *et al.* (2015). An atlas of genetic correlations across human diseases and traits. *Nature Genetics*, 47(11), 1236–1241.

Bulik-Sullivan, B. *et al.* (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, 47(3), 291–295.

Bulik-Sullivan, B. (2015). LDSC [programm]. https://github.com/bulik/ldsc/ (vaadatud 28.03.2018).

Dupuis, J. *et al.* (2010). New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nature Genetics*, 42(2), 105–116.

Falconer, D. S. (1960). *Introduction to quantitative genetics.* New York: The Ronald Press Company.

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

Locke, A. E. *et al.* (2015). Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518, 197–206.

Möls, M. (2010). *Dispersioonanalüüsi segamudelid.* Tartu: Tartu Ülikool.

Searle, S. R. (1961). Phenotypic, Genetic and Environmental Correlations. *Biometrics*, 17(3), 474–480.

Shungin, D. *et al.* (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature*, 518, 187–196.

Soranzo, N. *et al.* (2010). Common variants at 10 genomic loci influence hemoglobin A_1(C) levels via glycemic and nonglycemic pathways. *Diabetes*, 59(12), 3229–3239.

Strawbridge, RJ. *et al.* (2011). Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes*, 60(10), 2624–2634.

Zheng, G., Yang, Y., Zhu, X. & Elston, R. C. (2012). *Analysis of Genetic Association Studies*. New York: Springer-Verlag New York.

Lisad

Lisa 1. Simulatsiooniprogrammi kood

```
#funktsioon genotüüpide andmestiku loomiseks
f = function(n,m, epsilon , delta , beta ,gamma, plokid){
  SNP = data.frame(matrix(nrow = n, ncol = m))
  esimene = 1
  viimane = 0
  #tsükkel erineva suurusega haploplokkide tekitamiseks
  for(i in 1:length(plokid)){
    viimane = viimane + plokid[i]
    p1 = runif(1)
    #genereerime ühe ühise "alleeli" plokisisese korrelatsiooni
        loomiseks
    X1 = rbinom(n,1,p1)
    #tsükkel ühe fikseeritud suurusega haploploki tekitamiseks
    for(j in esimene:viimane){
      p2 = runif(1)
      #genereerime teise "alleeli"
      X2 = rbinom(n,1,p2)
      #genotüüp j on kahe "alleeli" väärtuste summa
      SNP[,j] = X1 + X2
      #standardiseerime loodud genotüübi
      SNP[,j] = (SNP[,j]-mean(SNP[,j]))/sd(SNP[,j])
    }
    esimene = viimane + 1
  }
  X = as.matrix(SNP) #genotüüpide maatriks
  SNP$Y1 = X%*%beta + epsilon #esimene fenotuup
  SNP$Y2 = X%*%gamma + delta #teine fenotuup
  #salvestame ka tegelikult realiseerunud seosekordajad
  cov_tegelik = cov(X%*%beta, X%*%gamma)
  cor_tegelik = cor(X%*%beta, X%*%gamma)
  return(c(SNP, cov_tegelik ,cor_tegelik))
}
```

```
#funktsioon LD-skooride arvutamiseks
l = function(d){
  cor = cor(d)^2
  l = rowSums(cor)
  return(l)
}
#funktsioon z-skooride leidmiseks
z = function(d,m){
  n = nrow(d)
  z = matrix(nrow = 2, ncol = m)
  X = as.matrix(d[,1:m])
  z[1,] = (t(X)%*%d$Y1)/sqrt(n)
  z[2,] = (t(X)%*%d$Y2)/sqrt(n)
  return(z)
}
#funktsioon päritavuse hindamiseks
paritavus = function(Z,LD,N,M,LD2){
  LDN = LD*N
  vabaliige = 1
  h2_hinnang = M * (mean(Z^2) - vabaliige)/mean(LDN)
  kaalud = 1/((vabaliige+LDN*h2_hinnang/M)^2)
  mudel1 = lm(Z^2 ~ LDN, weights = kaalud/LD2)
  h2_hinnang = summary(mudel1)$coefficients[2,1]*M
  vabaliige = summary(mudel1)$coefficients[1,1]
  kaalud = 1/((vabaliige+LDN*h2_hinnang/M)^2)
  mudel2 = lm(Z^2 ~ LDN, weights = kaalud/LD2)
  h2 = summary(mudel2)$coefficients[2,1]*M
  return(h2)
}
#funktsioon geneetilise kovariatsiooni hindamiseks
cov_g = function(d,N,h2_1,h2_2,M,LD2){
  d$LDN = d$LD*N
  cov.g_hinnang = 1/N*sum(d$Z1*d$Z2)
  kaalud = 1/((N*h2_1*d$LD/M + 1)*(N*h2_2*d$LD/M +
      1)+(N*cov.g_hinnang*d$LD/M)^2)
  mudel1 = lm(Z1*Z2 ~ LDN, data=d, weights = kaalud/LD2)
  cov.g_hinnang = summary(mudel1)$coefficients[2,1]*M
  vabaliige = summary(mudel1)$coefficients[1,1]
  kaalud = 1/((N*h2_1*d$LD/M + 1)*(N*h2_2*d$LD/M +
```

```
      1)+(N*cov.g_hinnang*d$LD/M + vabaliige)^2)
  mudel2 = lm(Z1*Z2 ~ LDN, data = d, weights = kaalud/LD2)
  cov.g = summary(mudel2)$coefficients[2,1]*M
  return(cov.g)
}
#funktsioon fikseeritud korrelatsiooniga geeniefektide loomiseks
bg = function(h2_1_fix,h2_2_fix,cor_g_fix,m,c1,c2){
  beta = rep(0,m)
  gamma = rep(0,m)
  c_samad = intersect(c1,c2)
  c_erinevad = setdiff(c2,c_samad)
  cov_g_fix = sqrt(h2_1_fix*h2_2_fix)*cor_g_fix
  liidetav = cov_g_fix/length(c_samad)
  h2_2 = 1
  while(h2_2>h2_2_fix){
    beta[c1] = runif(length(c1))
    h2_1 = sum(beta^2)
    beta = beta/sqrt(h2_1)*sqrt(h2_1_fix)
    gamma[c_samad] = liidetav/beta[c_samad]
    h2_2 = sum(gamma^2)
  }
  gamma[c_erinevad] = sqrt((h2_2_fix - h2_2)/length(c_erinevad))
  return(cbind(beta,gamma))
}

n_pl = 200 #loodavate haploplokkide arv
plokid = sample(5:20,n_pl,T) #plokkide suurused
n = 10000 #valimimaht
m = sum(plokid) #SNP-ide arv kokku
#hindame iga ploki viimase SNP-i z-skoori (kohtadel "indeksid")
indeksid = cumsum(plokid)
#kombinatsioonid SNP-dest, mis mõjutavad esimest tunnust
c11 = seq(1,m,40)
c12 = ...
#kombinatsioonid SNP-dest, mis mõjutavad teist tunnust
c21 = seq(1,m,20)
c22 = ...
kombinatsioonid1 = list(c11,c12,...)
kombinatsioonid2 = list(c21,c22,...)
```

```
h1 = c(0.8,0.1,...)
h2 = c(0.8,0.2,...)
rg = c(0.5,0.6,...)

for(k in 1: ...){
  c1 = unlist(kombinatsioonid1[k])
  c2 = unlist(kombinatsioonid2[k])
  #fikseerime päritavused ja korrelatsiooni
  h2_1_fix = h1[k]
  h2_2_fix = h2[k]
  cor_g_fix = rg[k]
  #leiame fikseeritud väärtustele vastavad geeniefektid
  geeniefektid = bg(h2_1_fix, h2_2_fix, cor_g_fix,m,c1,c2)
  beta = geeniefektid[,1]
  gamma = geeniefektid[,2]
  #genereerime juhuslikud keskkonnamõjud
  epsilon = rnorm(n,sd = sqrt(1-h2_1_fix))
  delta = rnorm(n,sd = sqrt(1-h2_2_fix))
  N = 500
  h2_1 = rep(NA,N)
  h2_2 = rep(NA,N)
  cov.g = rep(NA,N)
  cor.g = rep(NA,N)
  cov_f = rep(NA,N)
  cor_f = rep(NA,N)
  cor.g_tegelik = rep(NA,N)
  cov.g_tegelik = rep(NA,N)

  #simuleerimistsükkel
  for(i in 1:N){
    andmed = f(n,m,epsilon,delta,beta,gamma,plokid)
    cov.g_tegelik[i] = andmed[m+3]
    cor.g_tegelik[i] = andmed[m+4]
    andmed = data.frame(andmed[1:(m+2)])
    cov_f[i] = cov(andmed$Y1,andmed$Y2)
    cor_f[i] = cor(andmed$Y1,andmed$Y2)
    ld = l(andmed[,1:m])
    z_andmed = andmed[,c(indeksid,m+1,m+2)]
    ld2 = l(z_andmed[,1:n_pl])
```

```
    z_skoorid = z(z_andmed,n_pl)
    z1 = z_skoorid[1,]
    z2 = z_skoorid[2,]
    h2_1[i] = paritavus(z1,ld[indeksid],n,m,ld2)
    h2_2[i] = paritavus(z2,ld[indeksid],n,m,ld2)
    cov_andmed = data.frame(Z1 = z1, Z2 = z2,LD = ld[indeksid])
    cov.g[i] = cov_g(cov_andmed,n,h2_1[i],h2_2[i],m,ld2)
    if(h2_1[i]>0 & h2_2[i]>0) cor.g[i] = cov.g[i]/sqrt(h2_1[i]*h2_2[i])
  }
  cor.g_tegelik = unlist(cor.g_tegelik)
  cov.g_tegelik = unlist(cov.g_tegelik)
  tulemus = data.frame(h1 = h2_1, h2 = h2_2, cov_g = cov.g, cor_g =
      cor.g, cov.g_tegelik = cov.g_tegelik, cor.g_tegelik =
      cor.g_tegelik, cor_f=cor_f, m = m)
  save(tulemus,file = paste("tulemus",k,".RData", sep=""))
}
```

Lisa 2. Funktsioonid päritavuse ja geneetilise korrelatsiooni hindamiseks LD-skoori regressiooni meetodil

```
paritavus = function(Z,LD,N,M,n){
  LD = pmax(LD,1)
  LDN = LD*N
  vabaliige = 1
  lugeja = M * (mean(Z^2) - vabaliige)
  nimetaja = mean(LD*N)
  h2 = lugeja / nimetaja #h^2 algväärtustamine
  #tükkel kaalude ümberhindamiseks
  for(i in 1:n){
    h2 = max(h2,0)
    h2 = min(h2,1)
    kaalud = (1/(vabaliige+LDN*h2/M)^2)/LD
    m = lm(Z^2 ~LDN, weights = kaalud)
    vabaliige = summary(m)$coefficients[1,1]
    h2 = summary(m)$coefficients[2,1]*M #saadud päritavuse hinnang
  }
  return(c(vabaliige,h2))
}
```

```
cor_g = function(y1,y2,ld,n){
  M = nrow(ld[ld$MAF>0.05,])
  y1 = y1[!(is.na(y1$Z)),] #esimene fenotüüp
  y2 = y2[!(is.na(y2$Z)),] #teine fenotüüp
  colnames(y1) = c("SNP", "A1_1","A2_1","Z_1","N_1")
  colnames(y2) = c("SNP", "A1_2","A2_2","Z_2","N_2")

  #Ühendame andmestikud
  data = merge(y1,y2, by="SNP")
  data = merge(data, ld, by="SNP")
  colnames(data)[14] = "LD"
  #muudame z-statistikud "samasuunaliseks"
  data$Z_2[data$A1_1!=data$A1_2] = data$Z_2[data$A1_1!=data$A1_2]*(-1)
  data$LDN1N2 = data$LD*sqrt(data$N_1*data$N_2)

  #leiame mõlema fenotüübi päritavuse hinnangud
  h1 = paritavus(data$Z_1, data$LD, data$N_1, M, n)
  h2 = paritavus(data$Z_2, data$LD, data$N_2, M, n)
  h2_int_1 = h1[1] #y1 h^2 mudeli vabaliige
  h2_1 = h1[2]     #y1 h^2 hinnang
  h2_int_2 = h2[1] #y2 h^2 mudeli vabaliige
  h2_2 = h2[2]     #y2 h^2 hinnang

  #parameetrid kaalude jaoks
  ld = pmax(data$LD, 1)
  sqrt_n1n2 = sqrt(data$N_1*data$N_2)
  h2_1 = max(min(h2_1,1),0)
  h2_2 = max(min(h2_2,1),0)
  a = (data$N_1* h2_1 * ld) / M + h2_int_1
  b = (data$N_2* h2_2 * ld) / M + h2_int_2

  #gencovi ja vabaliikme algväärtused enne kaalutsüklit
  lugeja = M*(mean(data$Z_1*data$Z_2))
  nimetaja = mean(data$LD*sqrt_n1n2)
  gencov = lugeja/nimetaja
  vabaliige_gencov = 0
```

```
  #tsükkel kaalude ümberhindamiseks
  for(i in 1:n){
    gencov = min(gencov, 1)
    gencov = max(gencov,-1)
    c = (sqrt_n1n2* gencov * ld) / M + vabaliige_gencov
    kaalud = (1 / (a*b + c^2))/ld
    m = lm(Z_1*Z_2 ~ LDN1N2, data=data, weights = kaalud)
    vabaliige_gencov = summary(m)$coefficients[1,1]
    gencov = summary(m)$coefficients[2,1]*M
  }
  #geneetilise korrelatsiooni hinnang
  gencor = gencov/sqrt(h2_1*h2_2)
  return(gencor)
}
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Lisette Pajula,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Geneetilise korrelatsiooni hindamine LD-skoori regressiooni meetodiga", mille juhendajad on Märt Möls ja Reedik Mägi,

    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, 8.05.2018