Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kaarel Lusmägi

Ülegenoomsete uuringute metaanalüüs ja metaregressioon

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendajad: PhD Märt Möls
PhD Reedik Mägi

Tartu 2026

ÜLEGENOOMSETE UURINGUTE METAANALÜÜS JA METAREGRESSIOON

Bakalaureusetöö

Kaarel Lusmägi

Lühikokkuvõte

Metaanalüüsiks nimetatatakse analüüsimeetodit, kus sünteesitakse mitme erineva uuringu tulemused, mis käsitlevad ühesugust uurimisteemat. Käesoleva bakalaureusetöö eesmärk oli kasutada metaanalüüsi mudelit, et leida viie ülegenoomse assotsiatsiooniuuringu põhjal geenivariandid, millel on seos rasedusdiabeeti haigestumisega ning kirjeldada, kuidas nende geenivariantide mõju muutub populatsiooniti, kasutades neid uuringupopulatsioone iseloomustavaid tunnuseid. Töö esimeses osas antakse ülevaade vajalikest geneetika põhimõistetest ning tutvustatakse erinevaid statistilisi meetodeid ja teste, mida kasutatakse metaanalüüsi läbiviimisel. Teises osas viiakse läbi näiteandmestikul metaanalüüs ja metaregressioon ning võrreldakse erinevate metaanalüüsi mudelite headust.

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

Märksõnad: Metaanalüüs, metaregressioon, geneetika.

METAANALYSIS AND METAREGRESSION OF GENOME-WIDE ASSOCIATION STUDIES

Bachelor thesis

Kaarel Lusmägi

Abstract

Meta-analysis is an analytical method in which the results of several different studies addressing the same research topic are synthesized. The aim of this bachelor's thesis was to use a meta-analysis model to identify gene variants associated with gestational diabetes based on five genome-wide association studies, and to describe how the effects of these gene variants vary across populations using characteristics that define those study populations. The first part of the thesis provides an overview of essential genetic concepts and introduces various statistical methods and tests used in conducting a meta-analysis. The second part performs a meta-analysis and meta-regression on a sample dataset and compares the performance of different meta-analysis models.

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

Key Words: Meta-analysis, meta-regression, genetics.

Sisukord

Sissejuhatus **4**

1 Geneetika alusmõisted **6**

2 Fikseeritud mõjudega metaanalüüs **7**

3 Juhuslike mõjudega metaanalüüs **10**

4 Suurima tõepära meetod **12**

4.1 Mudeli parameetrite hindamine 13

5 Tõepärasuhte test **17**

6 Peakomponentanalüüs **18**

7 Metaanalüüsi näide **19**

Kokkuvõte **31**

Sissejuhatus

Rasedusdiabeet (ing. *Gestational diabetes mellitus*) on süsivesikute ainevahetushäire, mille tõttu tõuseb rasedatel naistel veresuhkru tase märgatavalt ja mis on seotud mitmete raseduskomplikatsioonidega (Ida-Tallinna Keskhaigla, 2026). Käesoleva töö eesmärgiks on kasutada metaregressiooni rasedusdiabeeti haigestumist mõjutavate geenivariantide otsimiseks. Samuti üritatakse metaregressiooni abil kirjeldada nende geenivariantide mõju muutumist erinevates populatsioonides.

Metaanalüüsiks nimetatatakse analüüsimeetodit, kus sünteesitakse mitme erineva uuringu tulemused, mis käsitlevad ühesugust uurimisteemat, nt kindla ravimi mõju vererõhu langusele. Metaanalüüsi kasutades on võimalik leida täpsemad hinnangud faktori mõju suurusele ning saavutada suurem võimsus võrreldes individuaalsete uuringutega. (Guerra ja Goldstein, 2010)

Töö teoreetiline osa jaguneb kolmeks osaks. Esimeses osas antakse ülevaade peamistest geneetika mõistetest, mida töös vaja läheb. Teises osas tutvustatakse põhjalikumalt kahte peamist metaanalüüsi meetodit: fikseeritud mõjudega metaanalüüs ja juhuslike mõjudega metaanalüüs. Kolmandas osas käsitletakse erinevaid teste ja meetodeid, mille abil saab — kasutades metaanalüüsi mudelit — testida, kas geenivariandil on statistiliselt oluline mõju rasedusdiabeedi tekkele, leida hinnanguid geenivariandi mõju suurusele ning uurida, kas geenivariandi mõju suurus erineb populatsiooniti.

Analüüsi osa kätkeb endas metaanalüüsi näite peatükk, mis jaguneb kaheks. Metaanalüüsi näite esimeses osas viiakse näiteandmestikul läbi metaanalüüs viie ülegenoomse assotsiatsiooniuuringu põhjal, mis viidi läbi erinevates populatsioonides, et leida geenivariandid, millel on statistiliselt oluline mõju rasedusdiabeedi tekkele, kusjuures eeldatakse, et geenivariantide tegelik mõju

on kõikides populatsioonides ühesugune. Teises osas proovitakse leida geenivariandid, mille tegeliku mõju suurus erineb populatsiooniti ning leida hinnangud mõju suurustele ning teha kindlaks hinnangute täpsus.

Töö kirjutamiseks kasutati tekstiküljendussüsteemi LaTeX. Andmete analüüsiks kasutati rakendustarkvara R-i. Koodi kirjutamisel ning jooniste puhul kasutati abiks tehisintellekti (OpenAI, 2026).

1 Geneetika alusmõisted

Rakk on elu väikseim ehitusühik, mis suudab kas üksi või suurema organismi osana kasvada, areneda ja paljuneda. Iga rakk sisaldab kogu liigile omast geneetilist materjali, mis enamjaolt paikneb raku tuumas kromosoomides. Kromosoom on biheeliksi vormis keerdunud DNA molekul. DNA koosneb omakorda orgaaniliste molekulide ahelast, mida nimetatakse nukleotiididiks, mis on moodustunud järgmise kolme ühendi liitumisel: lämmastikalus, suhkur (desoksüriboos) ja fosforhappe jääk. DNA struktuuris esineb neli erinevat lämmastikalust: adeniin (A), guaniin (G), tümiin (T) ja tsütosiin (C). DNA ahel tekib nukleotiidide omavahelise liitumise tulemusel, mis toimib komplementaarsusprintsiibi alusel, kus ühe ahela adeniini vastas on alati teise ahela tümiin ja guaniini vastas tsütosiin. (Kaart ja Möls, 2009)

Geen on DNA lõik, mis sisaldab infot ühe valgu sünteesiks. Geen võib koosneda vaid 1000 aluspaarist, aga ka miljonitest, sealjuures ei ole geen katkematu, vaid võib asuda tükeldatult mitmes DNA piirkonnas. (Mändul, 2016)

Alleel on üks variant kahest või enamast DNA järjestusest kindlas genoomi lokatsioonis (National Human Genome Research Institute, 2026a). SNP ehk üksiku nukleotiidi polümorfism on DNA ahela teisend, mis kujutab endas mingit nukleotiidi muutust kindlas DNA positsioonis. Ehk üks nukleotiid võib olla asendunud ülejäänud kolmega. (Mändul, 2016)

Kõige levinumad on kahe alleeliga SNPid, st kus on ainult kaks võimalikku nukleotiidi varianti. Antud töös vaatleme ka ainult selliseid SNPe.

Et teha kindlaks SNPi seost mingi haiguse või tunnustega, kasutatakse ülegenoomseid assotsiatsiooniuuringuid. Selle uuringu käigus vaadatakse läbi paljude inimeste genoom — ehk nende kogu geneetiline informatsioon — ning leitakse need geenivariandid, mis esinevad suurema sagedusega inimes-

tel, kellel on uuritav haigus või tunnus. (National Human Genome Research Institute, 2026b)

2 Fikseeritud mõjudega metaanalüüs

Fikseeritud mõjudega mudeli puhul eeldame, et kõikides uuringutes on vaadeldava faktori mõju uuritavale tunnusele ühesugune. Ehk kõik tegurid, mis võivad mõjutada faktori mõju suurust on kõikides uuringutes ühesugused, seega tõelise mõju suurus on võrdne kõikides uuringutes. (Borenstein *et al.*, 2009)

Faktoriks võib olla nt mingi ravim, kus uuritav tunnus on vererõhk; dieet, kus uuritavaks tunnuseks on kaalulangus jne. Antud töös on vaadeldavaks faktoriks SNP ning uuritavaks tunnuseks rasedusdiabeedi olemasolu.

Tähistagu θ faktori tõelise mõju suurust ning Y_i i-nda uuringu hinnangut sellele mõjule. Kuna kõikides uuringutes on faktori tegelik mõju eelduse kohaselt ühesugune, siis sellest järeldub, et erinevused hinnangutes tulenevad vaid valimiveast ehk lõpmatu valimimahu korral oleksid erinevate uuringute hinnangud samasugused. Praktikas pole muidugi kõik hinnangud täpsed, mistõttu saame oma metaanalüüsi mudeli avaldada kujul $Y_i = \theta + \varepsilon_i$, kusjuures on loomulik eeldus, et hinnangud on nihketa, st $EY_i = \theta$. (Kool, 2010)

Kui tahame fikseeritud mõjudega metaanalüüsi mudeliga leida erinevate uuringute pealt hinnang tegelikule mõjule, siis oleks mõttekas erinevatele uuringutele ka kaal anda, nt et uuringutel, mis hindavad täpsemalt tegelikku mõju - st nendel uuringutel, mille standardviga on väikseim - on suurem kaal ja suurema standardveaga uuringutel väiksem kaal.

Et meie metaanalüüsihinnang tegelikule mõjule oleks nihketa, võtame i-nda uuringu kaaluks $w_i = \frac{1}{\sigma_i^2}$, kus σ_i^2 on $i-$nda uuringu hinnangu dispersioon, st

$\sigma_i^2 = DY_i$. Hinnangute kaalutud keskmine ehk metaanalüüsi hinnang meid huvitavale mõjule avaldub siis kujul:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i}. \tag{1}$$

Ning saadud fikseeritud mõjudega metaanalüüsihinnangu dispersioon avaldub kujul:

$$D\hat{\theta} = \frac{1}{\sum_{i=1}^{n} w_i}.$$

Kaalutud keskmine on meie hinnang faktori tegelikule mõjule. (Borenstein *et al.*, 2009)

Näide

Kolmes populatsioonis on uuritud SNPi rs12565286 mõju diabeeti haigestumisele (hinnati logaritmi šansside suhtest). Saadud hinanngud ja nende hinnangute standardvead on toodud tabelis 1, kusjuures referentsalleeliks on guaniin (G) ja alternatiivalleeliks tsütosiin (C). Eeldame, et tegelikud mõ-

Tabel 1: Tegeliku mõju hinnangud ja nende hinnangute standardvead.

Populatsioon	y_i	SE_i
1	−0,0134	0,4075
2	−0,0223	0,4031
3	0,0112	0,4075

jud on igas uuringus ühesugused. Kaalutud keskmise valemi põhjal tuleb

ligikaudne hinnang tegelikule mõjule

$$\hat{\theta} \approx \frac{\frac{1}{0{,}41^2} \cdot (-0,01) + \frac{1}{0{,}4^2} \cdot (-0,02) + \frac{1}{0{,}41^2} \cdot 0,01}{\frac{1}{0{,}41^2} + \frac{1}{0{,}4^2} + \frac{1}{0{,}41^2}} \approx -0{,}007,$$

ehk inimesel, kellel on selle SNPi puhul alleeliks guaniin on ligikaudu $e^{0{,}007} \approx 1{,}007$ korda suuremad šansid nakatuda diabeeti võrreldes inimesega, kellel on alleeliks tsütosiin. Tuleb ka meeles pidada, et fikseeritud mõjudega metaanalüüsi meetodit kasutades saadud hinnang on kõigest hinnang ja võib olla ebatäpne. Ligikaudne 95% usaldusintervall tegelikule mõjule tuleb kujul $\hat{\theta} \pm 1{,}96\sqrt{D\hat{\theta}}$, mis antud andmete põhjal tuleb $(-0{,}47; 0{,}45)$. See sisaldab nulli, seega me ei saa ka metaanalüüsi kasutades öelda, et sellel SNPil on statistiliselt oluline mõju diabeeti haigestumisele.

On täheldatud, et konkreetse mutatsiooni mõju suurus võib erinevates uuritavates populatsioonides olla erinev, näiteks geenimutatsiooni mõju võib olla erinev naiste ja meeste jaoks. Seega puhas fikseeritud mõjudega mudel meile täpselt ei sobi, kuid saame kasutada selle veidi moonutatud varianti. Eeldame, et faktori tegelik mõju i-ndas uuringus θ_i sõltub p-st erinevast uuringupopulatsiooni kirjeldavast tunnusest ja see seos on ühesugune iga uuringu puhul. Siis saab regressioonimudeli panna kirja kujul:

$$\begin{aligned} Y_i &= \theta_i + \varepsilon_i \\ &= \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \varepsilon_i, \end{aligned}$$

kus $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ on tundmatu parameetervektor ja ε_i on juhuslik hindamisviga, $\varepsilon_i \sim N(0, \sigma_i^2)$. Antud mudelit nimetatakse metaregressiooni mudeliks. Peamine erinevus lineaarse regressiooniga seisneb selles, et analüüsi-

objektiks on uuringud, koos hinnangute ja standardvigadega, mitte üksikud vaatlused (Schmid, Stijnen ja White, 2021). Selles töös kasutatakse mudeli argumentidena SNPide efektialleelide sageduste põhjal arvutatud esimest ja teist peakomponenti, mis kirjeldavad populatsiooni päritolu. Mudeli täpsem kirjeldus esitatakse peatükis 7.

3 Juhuslike mõjudega metaanalüüs

Juhuslike mõjudega mudeli korral eeldatakse, et faktori tegelik mõju erineb uuringute puhul. Paljudel juhtudel on väheusutav, et kõikides uuringutes faktori tegelik mõju on sama — mida eeldatakse fikseeritud mõjudega mudelis —, kuna uuringud viiakse läbi erinevates populatsioonides, kus võivad olla erinevused geenides, haridustasemes, vanuses, kehamassiindeksis, mis kõik võivad olla teguriteks faktori mõju suuruses, seega ka tegelik mõju neis uuringutes on erinev. (Kool, 2010)

Eelnevalt aga nägime, et kui faktori tegelik mõju i-ndas uuringupopulatsioonis sõltub mingitest seda populatsiooni kirjeldavatest tunnustest ja see seos on ühesugune iga uuringu puhul, siis on võimalik kasutada ka fikseeritud mõjudega mudelit. Kui faktori tegelik mõju ei ole aga sellise lihtsa seosega kirja pandav, siis kasutatakse uuringutulemuste sünteesimiseks juhuslike mõjudega metaanalüüsi mudelit (Kool, 2010).

Tähistame erinevate uuringute tegelike mõjude keskmist tähega θ ja i-nda uuringu tegelikku mõju θ_i, siis i-nda uuringu tegelik mõju avaldub kujul: $\theta_i = \theta + \xi_i$, kus ξ_i tähistab i-nda uuringupopulatsiooni eripära. (Borenstein *et al.*, 2009)

Seega i-nda uuringu hinnang tegelikule mõjule tuleb:

$$\begin{aligned} Y_i &= \theta_i + \varepsilon_i \\ &= \theta + \xi_i + \varepsilon_i. \end{aligned}$$

Tavaliselt eeldame, et juhuslike mõjudega mudeli korral faktori tegelikud mõjud on normaaljaotusega, st $\theta_i \sim N(\theta, \tau^2)$ (Kool, 2010).

Prognoosimaks, kui palju erineb i-nda uuringu hinnang tegelikust keskväärtusest θ, oleks ka vaja teada juhuslike suuruste ξ_i ja ε_i hajuvust. Eeldame, et uuringutest on võimalik leida informatsiooni hinnanguvigade ε_i dispersiooni kohta $D(\varepsilon_i)$, siis jääb meile endale ülesandeks hinnata $D(\xi_i) = D(\theta_i) =: \tau^2$ (Möls, 2026).

Parameeter τ^2 tähistab uuringutevahelist hajuvust. Ehk kui me teaksime iga uuringu puhul tegelikku faktori mõju θ_i ja arvutaksime nende kaudu välja dispersiooni lõpmata paljude uuringute põhjal, siis see dispersioon oleks τ^2. Üheks võimaluseks τ^2 hinnata oleks kasutades DerSimoniani ja Lairdi hinnangut:

$$\hat{\tau}^2 = \frac{Q - (n-1)}{C},$$

kus

$$Q = \sum_{i=1}^{n} w_i Y_i - \frac{\left(\sum_{i=1}^{n} w_i Y_i\right)^2}{\sum_{i=1}^{n} w_i},$$

ja

$$C = \sum_{i=1}^{n} w_i - \frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i}$$

ning n tähistab uuringute arvu. (Borenstein *et al.*, 2009)
Lisaks võib τ^2 hindamiseks kasutada ka suurima tõepära meetodit. Suurima tõepära meetodi kasutamisel maksimeeritakse valimi tõepära numbrilisi meetodeid kasutades.

4 Suurima tõepära meetod

Antud peatükk põhineb aine "Tõenäosusteooria ja statistika II" materjalidel. Olgu $f(x|\theta)$ juhusliku suuruse X_i tihedusfunktsioon, kui X_i on pidev juhuslik suurus, ja tõenäosusfunktsioon, kui X_i on diskreetne, siis valimi $\mathbf{x} = (x_1, \ldots, x_n)$ tõepärafunktsioon on defineeritud kui

$$L_n(\theta) = L_n(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Suurima tõepära hinnanguks parameetrile θ nimetatakse hinnangut

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L_n(\theta),$$

seega STH leiab sellise $\hat{\theta}$, mis maksimiseerib valimi tõepärafunktsiooni ehk mis parameetri all on kõige tõenäolisem näha antud valimit. Kuna logaritm on rangelt kasvav funktsioon, siis samaväärselt kasutatakse ka logaritmilist tõepärafunktsiooni

$$\ell_n(\theta) = \ell_n(\theta|\mathbf{x}) = \log \prod_{i=1}^{n} f(x_i|\theta) = \sum_{i=1}^{n} \log f(x_i|\theta).$$

Suurima tõepära hinnangu leidmiseks võrdsustatakse tõepärafunktsiooni tuletis nulliga

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = 0$$

ning määratakse kindlaks, mis argumendi väärtusel saavutab ta oma globaalse maksimumi.

4.1 Mudeli parameetrite hindamine

Leiame fikseeritud mõjudega metaanalüüsi mudeli puhul suurima tõepära hinnangu tegelikule mõjule. Olgu meil valim $\mathbf{y} = (Y_1, \ldots, Y_n)$, kus Y_i tähistab i-nda uuringu hinnangut tegelikule mõjule. Eeldame, et $Y_i \sim N(\theta, \sigma_i^2)$, kus uuringute dispersioonid on meile teada ja on vaja hinnata parameetrit θ. Siin y_i tähistab juhusliku hinnangu Y_i realisatsiooni. Siis logaritmiline tõepärafunktsioon avaldub kujul:

$$\begin{aligned}\ell_n(\theta) &= \sum_{i=1}^{n} \log f(y_i|\theta) = \sum_{i=1}^{n} \log\left(\frac{1}{\sigma_i\sqrt{2\pi}} \exp(-\frac{(y_i-\theta)^2}{2\sigma_i^2})\right) \\ &= \sum_{i=1}^{n} \left(\log(\frac{1}{\sigma_i\sqrt{2\pi}}) - \frac{(y_i-\theta)^2}{2\sigma_i^2}\right).\end{aligned}$$

Ning tuletis sellest funktsioonist tuleb:

$$\frac{\partial \ell_n(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{y_i - \theta}{\sigma_i^2} = \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} - \theta \sum_{i=1}^{n} \frac{1}{\sigma_i^2}.$$

Võrdsustades tuletise nulliga avaldub hinnang kujul:

$$\hat{\theta}_{\text{STH}} = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \; w_i = \frac{1}{\sigma_i^2},$$

kus STH tähistab suurima tõepära hinnangut ning

$$\frac{\partial^2 \ell_n(\theta)}{\partial \theta^2} = -\sum_{i=1}^{n} \frac{1}{\sigma_i^2} < 0.$$

Seega tegemist on maksimumkohaga. Saadud hinnang on esialgsete hinnangute kaalutud keskmine.

Tundmatut parameetervektorit $\boldsymbol{\beta}$ on võimalik hinnata üldistatud vähimruutude meetodiga, mis saadakse minimiseerides avaldist $(\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta})^T\mathbf{W}^{-1}(\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta})$. Siin $\mathbf{X}$ on disaini- ehk mudelimaatriks, $\mathbf{y}$ on uuringuhinnangute vektor ja $\mathbf{W}$ on juhusliku vektori $\mathbf{y}$ kovariatsioonimaatriks, kusjuures eeldame, et uuringud on teineteisest sõltumatud, seega tegemist on diagonaalmaatriksiga. Hinnang avaldub kujul:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-1}\mathbf{y}.$$

Tundmatut parameetervektorit on võimalik hinnata ka suurima tõepära meetodiga. Olgu meil valim $\mathbf{y} = (Y_1, \ldots, Y_n)$ ja eeldame, et $Y_i \sim N(\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i}, \sigma_i^2)$. Maatrikskujul, kus $\mathbf{X}_{n\times(p+1)}$ on disainimaatriks, $\mathbf{y}_{n\times 1}$ on uuringuhinnangute vektor, $\mathbf{V}_{n\times n} = \mathrm{diag}(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2})$ ning $\boldsymbol{\beta}_{(p+1)\times 1}$ on parameetrite vektor, avaldub logaritmiline tõepärafunktsioon kujul:

$$\ell_n(\boldsymbol{\beta}) = c - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = c - \frac{1}{2}(\mathbf{y}^T\mathbf{V}\mathbf{y} - \mathbf{y}^T\mathbf{V}\mathbf{X}\boldsymbol{\beta} -$$
$$\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}\mathbf{X}\boldsymbol{\beta}) = c - \frac{1}{2}(\mathbf{y}^T\mathbf{V}\mathbf{y} - 2\mathbf{y}^T\mathbf{V}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}\mathbf{X}\boldsymbol{\beta}).$$

Siin $c = -\frac{1}{2}\log(|2\pi\boldsymbol{V}^{-1}|)$ tähistab konstanti. Keskmine liige saadakse võrdusest $\mathbf{y}^T\mathbf{V}\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}\mathbf{y}$, kuna nad on mõlemad skalaarid ja esimene maatriks saadakse teist transponeerides. Olgu $\boldsymbol{y}$ m-elemendiline vektor ja $\boldsymbol{x}$

n-elemendiline vektor, edaspidi kasutame tähistust:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \frac{\partial y_m}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_n} \end{pmatrix}.$$

Arvestades, et $\frac{\partial \boldsymbol{x}^T \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^T(\boldsymbol{A} + \boldsymbol{A}^T)$ ning $\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}$ (Barnes, 2006), saame et

$$-\frac{1}{2}\frac{\partial \boldsymbol{\beta}^T \mathbf{X}^T \mathbf{V} \mathbf{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = -\frac{1}{2} 2 \boldsymbol{\beta}^T \mathbf{X}^T \boldsymbol{V} \boldsymbol{X} = -\boldsymbol{\beta}^T \mathbf{X}^T \boldsymbol{V} \boldsymbol{X}$$

ning

$$\frac{\partial \mathbf{y}^T \mathbf{V} \mathbf{X} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X}.$$

Võrdsustades tuletise nulliga ning eeldades, et pöördmaatriks $(\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1}$ eksisteerib, tuleb hinnang kujul:

$$\begin{aligned} \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X} - \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} &= 0 \\ \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X} &= \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X} \\ \boldsymbol{\beta}^T &= \boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X} (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \\ \boldsymbol{\beta} &= ((\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1})^T (\boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X})^T \\ \boldsymbol{\beta} &= ((\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^T)^{-1} (\boldsymbol{y}^T \boldsymbol{V} \boldsymbol{X})^T \\ \hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{y}. \end{aligned}$$

Näeme, et antud hinnang kattub täpselt vähimruutude hinannguga. Veendume, et tegemist on maksimumpunktiga, selleks leiame logaritmilise tõepärafunktsiooni hessiaani, mis avaldub kujul:

$$\boldsymbol{H} = \frac{\partial^2 \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = -\frac{\partial \boldsymbol{\beta}^T \mathbf{X}^T \boldsymbol{V} \boldsymbol{X}}{\partial \boldsymbol{\beta}} = -\mathbf{X}^T \boldsymbol{V} \boldsymbol{X}.$$

Tõestus on järgmine. Võtame $\boldsymbol{A} = \mathbf{X}^T \boldsymbol{V} \boldsymbol{X} = (a_{ij})$ ning olgu $\ell_n(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{A} = (f_1(\boldsymbol{\beta}), \ldots, f_{p+1}(\boldsymbol{\beta}))$. Maatriksite korrutamise definitsiooni põhjal saame, et $f_j(\boldsymbol{\beta}) = \sum_{i=1}^{p+1} \beta_i a_{ij}$ ning ilmselt $\frac{\partial f_j(\boldsymbol{\beta})}{\partial \beta_k} = a_{kj}$, mistõttu

$$-\frac{\partial \boldsymbol{\beta}^T \boldsymbol{A}}{\partial \boldsymbol{\beta}} = -\boldsymbol{A}^T = -\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}.$$

Tõestamaks, et punktis $\boldsymbol{\beta}$ on range lokaalne maksimum, peame näitama, et hessiaan $\boldsymbol{H}$ on negatiivselt määratud. Tõestame enne ära, et $\boldsymbol{H}$ on negatiivselt poolmääratud. Veendumaks, et hessiaan on negatiivselt poolmääratud, piisab näidata, et hessiaan avaldub kujul $\boldsymbol{H} = -\boldsymbol{A}^T \boldsymbol{A}$, kus $\boldsymbol{A}$ on mingi maatriks, sest siis $-\boldsymbol{v}^T \boldsymbol{A}^T \boldsymbol{A} \boldsymbol{v} \leq 0$ iga vektori $\boldsymbol{v}$ korral, sest vektor $\boldsymbol{A}\boldsymbol{v}$ korrutatud enda tranponeeritud vektoriga on selle vektori skalaarkorrutis iseendaga, mis on alati mittenegatiivne ning lisades miinusmärgi muudab selle summa mittepositiivseks. Paneme tähele, et $\boldsymbol{H} = -\boldsymbol{X}^T \boldsymbol{V}^{\frac{1}{2}} \boldsymbol{V}^{\frac{1}{2}} \boldsymbol{X}$, kus $\boldsymbol{V}^{\frac{1}{2}}$ on defineeritud kui maatriks, mille korral $\boldsymbol{V}^{\frac{1}{2}} \boldsymbol{V}^{\frac{1}{2}} = \boldsymbol{V}$. Võtame $\boldsymbol{A} = \boldsymbol{V}^{\frac{1}{2}} \boldsymbol{X}$, siis $\boldsymbol{H} = -\boldsymbol{A}^T \boldsymbol{A}$, mistõttu hessiaan on negatiivselt poolmääratud eelneva arutluse tõttu. Näitame nüüd, et hessiaan $\boldsymbol{H}$ on negatiivselt määratud, selleks piisab näidata, et $-\boldsymbol{H} = \boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X}$ on positiivselt määratud, sest suvalise positiivselt määratud maatriksi $\boldsymbol{A}$ korral on maatriks $-\boldsymbol{A}$ negatiivselt määratud. Eelduse kohaselt on sümmeetriline maatriks $-\boldsymbol{H}$ pööratav. Samuti on teada, et pööratava maatriksi kõik omaväärtused erinevad nullist ning positiivselt poolmääratud maatriksi kõik omaväärtused on mittenegatiivsed (Horn ja Johnson, 2013). Kuna $-\boldsymbol{H}$ on meil pööratav ja positiivselt poolmääratud, siis sellest järeldub, et kõik tema omaväärtused on positiivsed. Kuna maatriks on positiivselt määratud parajasti siis, kui tema omaväärtu-

sed on kõik positiivsed (Kollo, 2026), siis on $-\boldsymbol{H}$ positiivselt määratud ehk $\boldsymbol{H}$ on negatiivselt määratud, millest järeldub, et punktis $\boldsymbol{\beta}$ on logaritmiline tõepärafunktsioon maksimiseeritud.

5 Tõepärasuhte test

Olgu tõepärafunktsioon defineeritud nii nagu eelnevalt:

$$L(\theta) = L_n(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta).$$

Tähistagu Θ tervet parameeterruumi, siis tõepärasuhte teststatistik kontrollimaks hüpoteese $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta$ on defineeritud kujul:

$$\lambda(\boldsymbol{x}) = -2 \log \frac{\sup\limits_{\theta \in \Theta_0} L(\theta)}{\sup\limits_{\theta \in \Theta} L(\theta)} = -2 \log \frac{L_0}{L},$$

kus θ võib olla skalaar või vektor. Wilksi teoreemi kohaselt on antud teststatistik nullhüpoteesi kehtides asümptootiliselt hii-ruut jaotusega vabadusastmete arvuga $\dim(\Theta) - \dim(\Theta_0)$, kui on täidetud nn regulaarsuse tingimused. Üks tingimustest on näiteks, et eksisteerib $\frac{\partial}{\partial\theta} \log f(x|\theta)$ iga $x \in X$ ja $\theta \in \Theta$ korral, kus X on tihedusfunktsioonide määramispiirkond. (Rossi, 2018)
Kõiki tingimusi me ei hakka siin loetlema, kuid nad on leitavad eelnevalt viidatud raamatust.
Tõepärasuhte testi kasutatakse tavaliselt, et leida kahest erinevast mudelist sobivam:

H_0 : võime kasutada lihtsamat mudelit — vähem hinnatavaid parameetreid

H_1 : peame kasutama keerukamat mudelit — rohkem hinnatavaid parameetreid.

6 Peakomponentanalüüs

Antud peatükk põhineb aine "Mitmemõõtmelised statistilised meetodid" materjalidel ja Getter Põru bakalaureusetööl "Veekogude klassifitseerimine satelliidiandmetelt multinomiaalse logistilise mudeliga".

Peakomponentanalüüs on olemuselt andmemahu vähendamise meetod. Kui analüüsitavate tunnuste arv on liiga suur, siis peakomponentanalüüsi abil saab konstrueerida uued tunnused, mida saab algsete tunnuste asemel statistilises analüüsis kasutada, tingimusel, et informatsioonikadu ei ole väga suur. Olgu meil p sõltuvat juhuslikku suurust $\boldsymbol{X} = (X_1, \ldots, X_p)^T$, keskväärtusvektoriga

$$\boldsymbol{\mu} = E\boldsymbol{X} = (EX_1, \ldots, EX_p)^T$$

ja kovariatsioonimaatriksiga

$$\boldsymbol{\Sigma} = Cov(\boldsymbol{X}) = E[(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})^T].$$

Peakomponentanalüüsi eesmärgiks on suurused $X_1, \ldots, X_p$ asendada väiksema arvu suurustega $\xi_1, \ldots, \xi_t$ $(t \leq p)$, mis on lineaarkombinatsioonid esialgsetest suurused:

$$\xi_j = \boldsymbol{b}_j^T \boldsymbol{X} = b_{j1}X_1 + \cdots + b_{jp}X_p, \;\; j = 1, \ldots, t.$$

Kordajad $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_t$ tahame valida nii, et suurusel ξ_1 oleks maksimaalne võimalik hajuvus, suurusel ξ_2 suuruselt järgmine hajuvus jne, ning et $\xi_1, \ldots, \xi_t$ oleksid omavahel mittekorreleeritud. Et lahend oleks ühene, eeldame, et maatriksi $\boldsymbol{\Sigma}$ omaväärtused on kõik erinevad (Hyvärinen, Hurri ja Hoyer, 2009) ning seame vektoritele normeerituse tingimuse: $\boldsymbol{b}_j^T \boldsymbol{b}_j = 1, \forall j \in \{1, \ldots, t\}$.

Olgu $\lambda_1 > \lambda_2 > \ldots > \lambda_j > \ldots > \lambda_p > 0$ maatriksi $\boldsymbol{\Sigma}$ omaväärtu-

sed. On võimalik näidata, et kui valida $\boldsymbol{b}_1 = \boldsymbol{v}_1, \ldots, \boldsymbol{b}_t = \boldsymbol{v}_t$, kus $\boldsymbol{v}_j$ on omaväärtusele λ_j vastav normeeritud omavektor, siis on eelnevalt loetletud tingimused täidetud. Suurust ξ_1 nimetatakse esimeseks peakomponendiks, suurust ξ_2 teiseks peakomponendiks jne, kusjuures nende dispersioonid on $D\xi_j = \boldsymbol{v}_j^T \boldsymbol{\Sigma} \boldsymbol{v}_j = \lambda_j,\ j \in \{1, \ldots, t\}$. Osakaalu esialgsete tunnuste hajuvusest, mida esimesed q peakomponenti ära kirjeldavad, on leitav jagatisest

$$\psi_q = \frac{\sum_{i=1}^{q} \lambda_i}{\sum_{i=1}^{p} \lambda_i}.$$

7 Metaanalüüsi näide

Antud peatükis viiakse näiteandmestikul läbi metaanalüüs viie ülegenoomse assotsiatsiooniuuringu põhjal, mis uurisid erinevate SNPide mõju rasedusdiabeedile ning mis viidi läbi erinevates populatsioonides, et leida SNPid, millel on mõju selle haiguse avaldumisele.

Vaatame praegu mudelit $Y_i = \theta + \varepsilon_i$, kus kontrollime hüpoteese $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$. Siin Y_i-d $(i = 1, \ldots, 5)$ tähistavad tähistavad erinevatest uuringutest saadud hinnanguid uuritava SNPi mõjule ning hindavad logaritmilist šansside suhet, kus lugejas on rasedusdiabeedi esinemise šanss alternatiivalleeliga inimese puhul ja nimetajas šanss referentsalleeli puhul. Analüüsi kaasati vaid need SNPid, mille mõju uuriti kõigis viies populatsioonis ning tegeliku mõju hinnang $\hat{\theta}$ saadi kasutades valemiga (1) kirjeldatud kaalutud keskmist. Fikseeritud mõjudega mudeli puhul eeldame, et uuringuhinnangud $Y_i \sim N(\theta, \sigma_i^2)$, kus θ on tegeliku mõju suurus. Maatriks $\boldsymbol{V}$ tähistab juhuslikku vektori $\boldsymbol{y} = (Y_1, \ldots, Y_n)^T$ kovariatsioonimaatriksi pöördmaatriksit. Kuna

antud näite puhul on tegemist erinevates populatsioonides tehtud uuringutega, mis kasutavad erinevaid valimeid, siis võime eeldada, et leitud hinnangud on üksteisest sõltumatud ehk $\boldsymbol{V} = \text{diag}(\frac{1}{\sigma_1^2}, \ldots, \frac{1}{\sigma_n^2})$. Maatriks $\boldsymbol{X}$ tähistab disainimaatriksit, antud vabaliikmega mudeli puhul $\boldsymbol{X} = 1_{n\times 1}$. Arvestades, et $D\hat{\theta} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1} = (\sum_{i=1}^n w_i)^{-1}$, on võimalik leida ligikaudne usaldusintervall tegelikule mõjule. Testimaks, kas $\theta = 0$ üldkogumis, kasutatakse tõepärasuhte testi. Siin $\Theta_0 = \{0\}$ ja $\Theta = \mathbb{R}$. Toome välja selle teststatistiku arvutamise ja p-väärtuse leidmise SNPi rs36179555 puhul. Testitavad hüpoteesid on kujul $H_0 : \theta = 0$ vs. $H_1 : \theta \in \mathbb{R}$. Teststatistik avaldub järgmiselt:

$$
\begin{aligned}
\lambda(\boldsymbol{y}) &= -2\left(\log\sup_{\theta=0}\prod_{i=1}^n f(y_i|\theta) - \log\sup_{\theta\in\mathbb{R}}\prod_{i=1}^n f(y_i|\theta)\right)\\
&= -2\left(\log\prod_{i=1}^n f(y_i|\theta=0) - \log\prod_{i=1}^n f(y_i|\hat{\theta})\right)\\
&= -2\left(\sum_{i=1}^n\left(\log\frac{1}{\sigma_i\sqrt{2\pi}} - \frac{y_i^2}{2\sigma_i^2}\right) - \sum_{i=1}^n\left(\log\frac{1}{\sigma_i\sqrt{2\pi}} - \frac{(y_i-\hat{\theta})^2}{2\sigma_i^2}\right)\right)\\
&= \sum_{i=1}^n\frac{y_i^2-(y_i-\hat{\theta})^2}{\sigma_i^2} = 2\hat{\theta}\sum_{i=1}^n w_i y_i - \hat{\theta}^2\sum_{i=1}^n w_i \qquad \left(w_i := \frac{1}{\sigma_i^2}\right)\\
&= 2\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}\sum_{i=1}^n w_i y_i - \left(\frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}\right)^2\sum_{i=1}^n w_i\\
&= 2\frac{(\sum_{i=1}^n w_i y_i)^2}{\sum_{i=1}^n w_i} - \frac{(\sum_{i=1}^n w_i y_i)^2}{\sum_{i=1}^n w_i} = \frac{(\sum_{i=1}^n w_i y_i)^2}{\sum_{i=1}^n w_i}.
\end{aligned}
$$

SNPi rs36179555 puhul tuli uuringuhinnangute vektor ja standardvigade vektor vastavalt:

$$\boldsymbol{y} = (-0{,}8176; 0{,}5320; 0{,}0968; -0{,}0459; 0{,}0625)^T$$
$$\boldsymbol{s} = (0{,}7368; 0{,}2390; 0{,}1109; 0{,}0290; 0{,}0816)^T.$$

Antud andmete põhjal tuleb

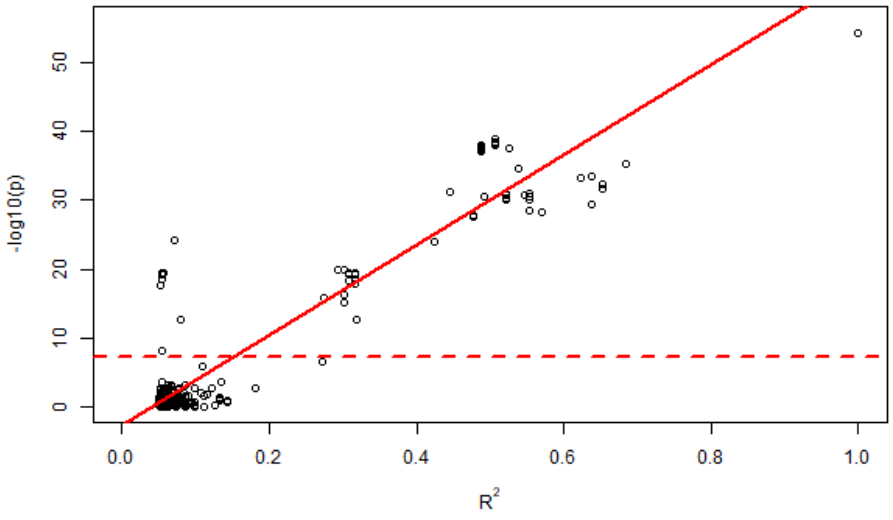
$$\lambda(\boldsymbol{y}) = \frac{(\frac{1}{0{,}7368^2} \cdot (-0{,}8176) + \ldots + \frac{1}{0{,}0816^2} \cdot 0{,}0625)^2}{\frac{1}{0{,}7368^2} + \ldots + \frac{1}{0{,}0816^2}} \approx 0{,}6$$

ning arvestades, et teststatistik on asümptootiliselt jaotusega χ^2_1, saame p-väärtuse arvutada kujul $p = 1 - F(0{,}6) \approx 0{,}44$, kus F on vabadusastmega üks hii-ruut jaotuse jaotusfunktsioon.

Tabelis 2 on toodud välja metaanalüüsi käigus saadud kolm kõige väiksema p-väärtusega sõltumatut SNPi.

Kuna SNPid, mis on üksteisele lähedal, päranduvad väga sageli koos edasi, siis on nad omavahel tugevalt korreleeritud (vt *Linkage Disequilibrium*). See tähendab, et kui ühel SNPil on tuvastatud statistiliselt oluline seos mingi haigusega, siis ka paljudel selle SNPi vahetus ümbruses olevatel SNPidel tuvastatakse seos samuti. Seega me tahame leida sellist otsustusmeetodit, mille põhjal määrata, kas SNPi olulisus võib olla tingitud tema lähedal paikneva teise, antud piirkonnas kõige väiksema p-väärtusega SNPi (nn peaSNP) mõjust või mitte. Soovime lisada tulemustesse vaid neid SNPe, mis võiksid kirjeldada sõltumatut mõju haigestumisele. Otsuse tegemiseks võime uurida, kui kaugel peaSNPist võime veel kohata palju statistiliselt olulisi (p-väärtus väiksem kui 5×10^{-8}) SNPe. Kaugust võime sealjuures mõõta nii aluspaarides kui ka SNPide omavahelist korrelatsiooni kasutades. Et määrata, mis nivoost

alates loeme peaSNPi mõju kadunuks, kasutame joonisel 1 toodud hajuvusgraafikut. Korrelatsiooni ruut (R^2) on siin arvutatud iga SNPi ja nn peaSNPi (kõige madalam p-väärtus) vahel. Kuna SNPe on palju, siis peame arvestama mitmese testimise probleemiga. Geneetikas on leitud, et katseviisilise vea kontrollimiseks ülegenoomsetes uuringutes tuleks võtta võrdlusviisilise vea tõenäosuseks 5×10^{-8} (Xu *et al.*, 2014), millele vastab antud joonisel katkendlik joon. Otsustusnivoo määrame regressioonisirge ja katkendliku joone kokkupuutepunkti x-koordinaadi põhjal, mis antud juhul tuleb 0,15. Kui SNPi $R^2 \in [0; 0{,}15]$ ja SNP on oluline, st $-\log_{10}(p) > -\log_{10}(5 \times 10^{-8})$, siis loeme, et see SNP on sõltumatu peaSNPist ja tema olulisus ei ole tingitud peaSNPi mõjust, sest lõigus $[0; 0{,}15]$ on peaSNPi mõju suuresti kadunud regressioonisirge põhjal.



Joonis 1: SNP ja peaSNPi vaheline R^2 ja SNPi statistiline olulisus.
PeaSNPi rollis on rs10830963, graafikul on kujutatud peaSNPi ümbruses asuvate teiste SNPide statistiline olulisus logaritmilisel skaalal.

Antud juhul on fikseeritud SNPiks rs10830963, millel tuvastati kõige väik-

sem p-väärtus, SNPidevahelised korrelatsioonid siin on arvutatud Bangladeshi bengalide populatsiooni põhjal. Joonise 1 põhjal saame öelda, et selles kromosoomis on mitu peaSNPist sõltumatut SNPi, mis on ka statistiliselt olulised. Tabelisse 2 kaasasime nendest vähima p-väärtusega SNPi, milleks oli rs4753426.

Tabel 2: Kolm kõige väiksema p-väärtusega sõltumatut SNPi vabaliikmega mudeli korral (eeldame samasugust SNPi mõju kõigis uuringutes).

SNP	Kr	Referentsalleel	Alternatiiv	P-väärtus	OR (95% UI)
rs10830963	11	G	C	$4{,}4 \times 10^{-55}$	1,41 (1,34-1,47)
rs4753426	11	C	T	$7{,}7 \times 10^{-25}$	1,24 (1,18-1,29)
rs7903146	10	T	C	$1{,}7 \times 10^{-17}$	1,21 (1,16-1,27)

Kr - kromosoom.

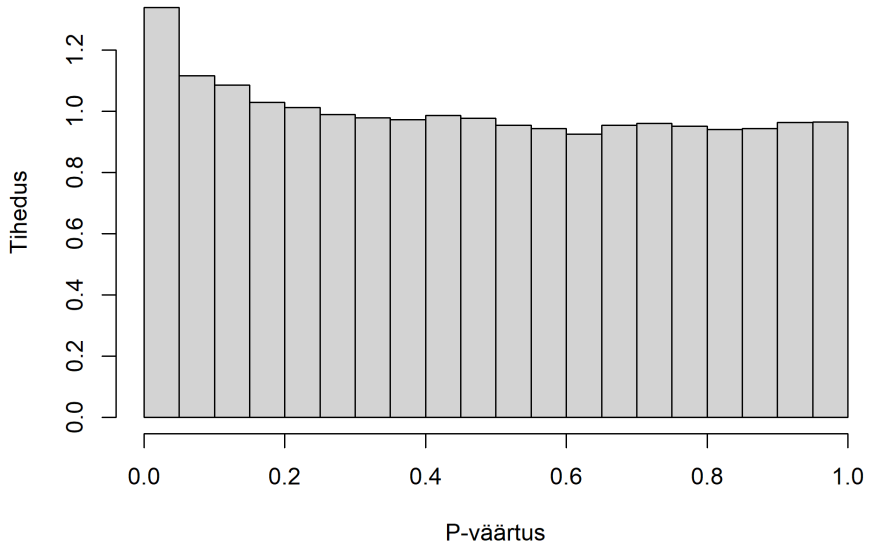
Näiteks SNPi rs10830963 puhul on naisel, kes kannab alleeli C ligikaudu 1,41 korda suuremad šansid haigestuda rasedusdiabeeti võrreldes naisega, kellel on alleeliks G. Tabelis esitatud ligikaudne 95% usaldusintervall šansside suhtele on arvutatud valemiga:

$$\text{UI} = \left(\exp\left(\hat{\theta} - 1{,}96\sqrt{D\hat{\theta}}\right), \exp\left(\hat{\theta} + 1{,}96\sqrt{D\hat{\theta}}\right)\right).$$

Kui SNPidel ei ole mõju, siis p-väärtuste jaotus on ühtlase jaotusega. Põhjendus on järgmine. Kuna meie teststatistik on asümptootiliselt hii-ruut jaotusega, siis p-väärtus avaldub kujul $p = 1 - F(T)$, kus F on hii-ruut jaotuse jaotusfunktsioon ning T on teststatistik. On teada, et kui juhusliku suuruse X jaotusfunktsioon on G ning G on pidev, siis $G(X) \sim U(0,1)$ (Kangro, 2024), seega kui antud juhul nullhüpotees kehtib: $T \sim F$, siis $F(T) \sim U(0,1)$, mistõttu ka $p \sim U(0,1)$.

Enamus kromosoomide puhul selle andmestiku juures ei ole p-väärtuste jao-

tus ühtlane — vaata näiteks joonist 2 — aga pole ka ühtegi SNPi, mille puhul $p < 5 \times 10^{-8}$. Seega on SNPe, millel on mõju, aga meil pole piisavalt andmeid, et need tuvastada.



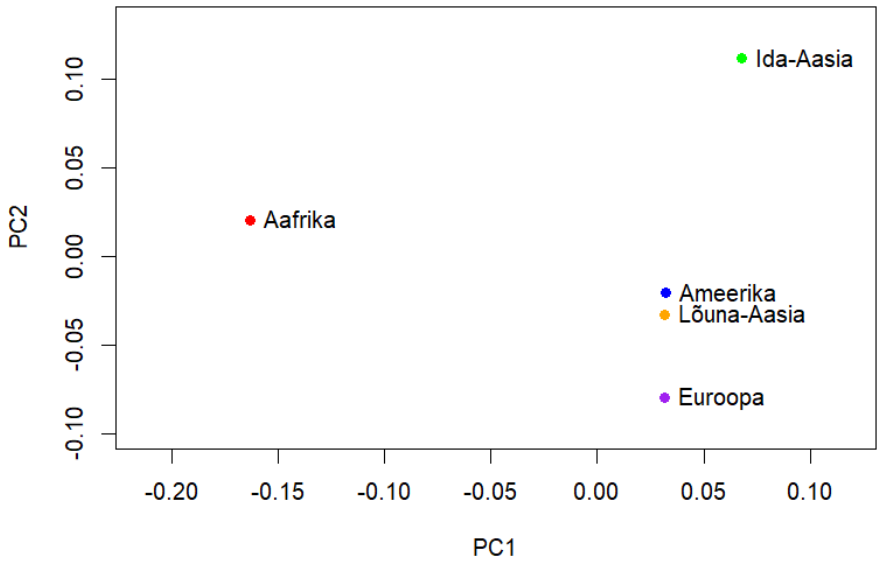
Joonis 2: SNPide p-väärtuste histogramm esimese kromosoomi puhul.

Vaatasime ka iga SNPi puhul metaregressiooni mudelit, mis lubab SNPi tegelikul mõjul erinevates populatsioonides olla erinev:

$$Y_i = \overbrace{\beta_0 + \beta_1 PC1_i + \beta_2 PC2_i}^{\theta_i} + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma_i^2), \tag{2}$$

kus $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^T$ hinnang leiti valemiga $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{y}$. Argumentideks kasutasime peakomponente, mis on arvutatud referentsalleelide sageduste põhjal. Siin $PC1 = (PC1_1, \ldots, PC1_5)$, kus $PC1_1$ tähistab esimest peakomponenti, mis on arvutatud välja Aafrika populatsiooni kohta, $PC1_2$ tähistab esimest peakomponenti, mis on arvutatud välja Ameerika po-

pulatsiooni kohta jne, teise peakomponendiga on tähistus analoogiline. Siin $\boldsymbol{X} = (\mathbf{1}, PC1^T, PC2^T)$. Tavaliselt eristab esimene peakomponent Aafrika populatsioone Euroopa ja Aasia omadest ning teine peakomponent Aasia populatsioone Euroopa omadest nagu on näha jooniselt 3.



Joonis 3: Populatsioonide PC1 ja PC2 hajuvusgraafik.

Arvutame alljärgnevalt parameetervektori $\boldsymbol{\beta}$ hinnangu $\hat{\boldsymbol{\beta}}$ SNPi rs3766582 jaoks. Erinevates uuringutes leitud hinnangud antud SNPi mõjule koos hinnangute standardvigadega on ära toodud tabelis 3.

Tabel 3: Tegelike mõjude (θ_i) hinnangud ja nende hinnangute standardvead.

Populatsioon	y_i	SE_i
Aafrika	0,1418	0,3263
Ameerika	−0,3461	0,1645
Ida-Aasia	0,0386	0,0780
Euroopa	−0,0122	0,0483
Lõuna-Aasia	0,2582	0,1272

Disainimaatriks tuleb siis

$$X = (\mathbf{1}, PC1^T, PC2^T) = \begin{pmatrix} 1 & -0{,}1632 & 0{,}0206 \\ 1 & 0{,}0322 & -0{,}0205 \\ 1 & 0{,}0677 & 0{,}1117 \\ 1 & 0{,}0315 & -0{,}0792 \\ 1 & 0{,}0317 & -0{,}0327 \end{pmatrix}.$$

Kus maatriksi esimese rea viimased kaks elementi on Aafrika populatsiooni peakomponendid, teise rea viimased kaks elementi Ameerika populatsiooni peakomponendid jne. Kuna uuringud on viidud läbi erinevate populatsioonide põhjal, siis uuringuhinnangud on üksteisest sõltumatud, st $Cov(Y_i, Y_j) = 0$ $(i \neq j)$, seega $\boldsymbol{V} = \text{diag}(\frac{1}{0{,}33^2}; \frac{1}{0{,}16^2}; \ldots; \frac{1}{0{,}13^2})$. Uuringuhinnangute vektorit tähistame $\boldsymbol{y}$-ga. Parameetervektori suurima tõepära hinnang tuleb siis:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}\mathbf{y}.$$

Antud juhul tuleb parameetrite hinnang kujul:

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}04 \\ -0{,}57 \\ 0{,}33 \end{pmatrix}.$$

On samuti teada, et $D\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}\boldsymbol{X})^{-1}$, seega kovariatsioonimaatriks tuleb:

$$D\hat{\boldsymbol{\beta}} = \begin{pmatrix} 0{,}006 & -0{,}099 & 0{,}023 \\ -0{,}099 & 2{,}363 & -0{,}398 \\ 0{,}023 & -0{,}398 & 0{,}297 \end{pmatrix}.$$

Regressioonikordaja β_i usaldusintervall usaldusnivool $1-\alpha$ tuleb

$$\hat{\beta}_i \pm z_{1-\frac{\alpha}{2}} \sqrt{(D\hat{\boldsymbol{\beta}})_{ii}}.$$

Leidmaks SNPid, mille mõju erineb populatsiooniti, sobitasime igale SNPile andmestikus mudeli (2). Iga SNPi puhul kontrollisime hüpoteeside paari $H_0 : \beta_1 = 0 \ \& \ \beta_2 = 0$ vs. $H_1 : \beta_1 \neq 0 \ \vee \beta_2 \neq 0$, mille testimiseks kasutasime samuti tõepärasuhte testi. Kolme kõige väiksema p-väärtusega SNPid antud mudeli puhul on nähtaval tabelis 4. Kui mudel ei ole statistiliselt oluline, siis jääme nullhüpoteesi juurde, mis ütleb, et ei ole erinevust tegelikes mõjudes erinevate populatsioonide vahel.

Tabel 4: Kolm kõige väiksema p-väärtusega SNPi mudeli (2) korral. Kontrollitakse nullhüpoteesi, mis väidab, et SNPi mõju on kõigis uuringupopulatsioonides ühesugune.

SNP	Referents	Alternatiiv	Mudeli olulisus	$\hat{\beta}_1$	$\hat{\beta}_2$
rs9348441	A	T	$9{,}8 \times 10^{-9}$	2,93	1,37**
rs3804141	T	C	8×10^{-7}	−7,72**	−0,7
rs11020102	G	A	$1{,}6 \times 10^{-6}$	0,8	−1,78***

*p<0,05; **p<0,01; ***p<0,001.

Vaatleme praegu SNPi rs9348441, mille puhul $\hat{\beta}_0 = 0{,}08$. Kuna $\hat{\beta}_2 = 1{,}37$, siis see tähendab, et teise peakomponendi kasvades nt 0,06 ühiku võrra, suureneb mõju suurus keskmiselt $1{,}37 \times 0{,}06 = 0{,}082$ võrra. Kui võrdleme Ameerika ja Euroopa populatsioonide jaoks leitud teise peakomponendi väärtuste erinevust, siis nende vahe on ligikaudu 0,06. Kui võrdleme kahte Euroopa naist, ühel neist on SNPi rs9348441 genotüübiks TT — st emalt ja isalt päritud homoloogilistes kromosoomides on mõlemad alleelid T — ja teisel AA, siis TT genotüübiga naisel on ligikaudu $\exp(2 \cdot (0{,}08 + 2{,}93 \cdot 0{,}03 + 1{,}37 \cdot (-0{,}08)))$ korda suurem šanss haigestuda rasedusdiabeeti. Kui võrdleme kahte Ameerika

naist, siis on TT genotüübiga naisel $\exp(2\cdot(0{,}08+2{,}93\cdot 0{,}03+1{,}37\cdot(-0{,}02))) = \exp(2\cdot(0{,}08+2{,}93\cdot 0{,}03+1{,}37\cdot(-0{,}08)))\cdot\exp(2\cdot 1{,}37\cdot 0{,}06)$ korda suurem šanss haigestuda rasedusdiabeeti. Seega on genotüübi TT mõju (TT vs. AA) $\exp(2\cdot 1{,}37\cdot 0{,}06)\approx 1{,}17$ korda suurem Ameerika populatsioonis.
Tabelis 5 on toodud välja SNPi rs9348441 esialgne mõju hinnang ja metaregressiooni mudeli (2) hinnang koos 95% usaldusintervallidega.

Tabel 5: Esialgne mõju hinnang ja metaregressiooni mudeli hinnang erinevates populatsioonides. Hinnangud on logaritmilisele šansside suhtele.

Populatsioon	Esialgne (95% UI)	Metaregressioon (95% UI)
Aafrika	-0,39 (-1,03; 0,25)	-0,37 (-1,00; 0,26)
Ladina-Ameerika	0,23 (-0,03; 0,50)	0,15 (0,09; 0,20)
Ida-Aasia	0,43 (0,32; 0,54)	0,43 (0,32; 0,54)
Euroopa	0,06 (0,01; 0,11)	0,06 (0,01; 0,11)
Lõuna-Aasia	0,13 (-0,02; 0,27)	0,13 (0,08; 0,18)

Metaregressiooni mudeli puhul tuleb 95% usaldusintervall tegelikule mõjule i-ndas populatsioonis kujul: $Y_{prog} \pm 1{,}96\sqrt{D(Y_{prog}|PC1_i, PC2_i)}$, kus

$$\begin{aligned} D(Y_{prog}|PC1_i, PC2_i) &= D(\overbrace{\hat{\beta}_0 + \hat{\beta}_1 PC1_i + \hat{\beta}_2 PC2_i}^{\hat{\theta}_i} - \theta_i) \\ &= D\hat{\beta}_0 + PC1_i^2 D\hat{\beta}_1 + PC2_i^2 D\hat{\beta}_2 \\ &+ 2\,PC1_i\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) + 2\,PC2_i\,\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_2) \\ &+ 2\,PC1_i PC2_i\,\mathrm{Cov}(\hat{\beta}_1, \hat{\beta}_2). \end{aligned}$$

Kovariatsioonid saime kätte maatriksist: $D\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V} \boldsymbol{X})^{-1}$. Antud regressioonimudeli peamiseks mõtteks on siluda hinnangute juhuslikkusest tulenevat viga, seega tegeliku mõju hinnanguks konkreetses populatsioonis võtame

metaregressiooni hinnangu, mitte esialgse hinnangu. Kui vaadata tabeli 5 usaldusintervalle, siis on näha, et enamasti usaldusintervallid on sama laiusega, välja arvatud Ladina-Ameerika ja Lõuna-Aasia, kus metaregressiooni usaldusintervall on märgatavalt kitsam. Seega saavutasime metaregressiooni mudeliga täpsemad hinnangud SNPi rs9348441 mõjule vähemalt mõnede uuritavate populatsioonide jaoks. Usaldusintervallidest on näha ka, et Ida-Aasia populatsiooni tegelik mõju erineb teistest populatsioonidest, kuna tegeliku mõju usaldusintervall Ida-Aasia populatsioonis ei kattu ühegi teise populatsiooni usaldusintervalliga.

Tuletame meelde, et kui faktori mõju suurus muutub erinevates uuringupopulatsioonides raskesti seletataval viisil, st pole adekvaatselt regressiooniseosega hinnatav, siis tuleks kasutada juhuslike mõjudega metaanalüüsi mudelit. Seega kerkib üles küsimus, kas äkki juhuslike mõjudega mudel pole praeguses kontekstis parem viis, et SNPi mõju hinnata. Selle kontrollimiseks saame kasutada Akaike informatsioonikriteeriumit (AIC), mida kasutatakse erinevate mudelite omavaheliseks võrdlemiseks. Kui ühel mudelil on väiksem AIC kui teisel, siis tuleks seda eelistada. Tabelis 6 on välja toodud metaregressiooni mudelite ja juhuslike mõjudega mudelite AIC SNPide rs9348441, rs3804141 ja rs68113313 puhul, mis metaregressiooni puhul tulid kõige väiksemate p-väärtustega. Hindamaks τ^2, kasutati suurima tõepära hinnangut.

Tabel 6: Metaregressiooni AIC ja juhuslike mõjudega AIC võrdlus.

SNP	Metaregressioon (AIC)	Juhuslike mõjudega (AIC)
rs9348441	-8,9	2,6
rs3804141	-1,4	10
rs11020102	-10,1	-1,4

Nagu tabelist on näha, siis on iga SNPi puhul AIC metaregressiooni mude-

liga väiksem, seega eelistame metaregressiooni mudelit juhuslike mõjudega mudelile nende SNPide puhul.

Kokkuvõte

Bakalaureusetöö eesmärgiks oli tutvuda metaregressiooni mudeliga ning kasutada seda kirjeldamaks SNPide mõju muutust erinevates uuringupopulatsioonides. Töös kasutati viie ülegenoomse assotsiatsiooniuuringu andmeid, mis viidi läbi erinevates populatsioonides.

Vabaliikmega mudeli korral, st kus eeldasime, et igas populatsioonis on tegeliku mõju suurus ühesugune ja puudusid täiendavad argumenttunnused, olid kõige väiksema p-väärtusega SNPid rs10830963, rs7903146 ja rs10811661, kuid seose tugevuse iseloomustamiseks kasutatavad šansside suhted viitasid pigem nõrgale seosele.

Metaregressiooni mudeli puhul oli ainsaks oluliseks SNPiks rs9348441, mille puhul tuvastati statistiliselt oluline erinevus tegelikes mõjudes Ida-Aasia populatsiooni ja teiste populatsioonide vahel ning mille puhul andis metaregressiooni mudel täpsemad hinnangud tegelikele mõjudele võrreldes esialgsete ülegenoomsete uuringutega. Tuleb rõhutada, et nullhüpoteesi jäämine enamus SNPide puhul metaregressiooni mudeliga ei tähenda, et pole erinevust tegelikes mõjudes erinevate populatsioonide vahel. Võib nii olla, et tõepärasuhte testi võimsus oli väike antud andmete põhjal, mudelisse valitud tunnused polnud parimad, et mingit erinevust mõjudes tuvastada erinevate populatsioonide vahel või hoopiski ei sobinud lineaarne mudel kirjeldamaks seost peakomponentide ja SNPi mõju vahel.

Võrreldes juhuslike mõjude mudelit metaregressiooni mudeliga SNPide rs9348441, rs3804141 ja rs11020102 põhjal, tuli välja, et metaregressiooni mudeli hinnanguviga on väiksem kui juhuslike mõjudega mudeli oma ning seega metaregressiooni mudel oli parem variant prognoosimaks tegelikke mõjusid erinevates populatsioonides.

Kasutatud allikad

Barnes, Randal (2006). *Matrix differentiation*. URL: https://atmos.washington.edu/~dennis/MatrixCalculus.pdf (vaadatud 15.04.2026).

Borenstein, Michael, Larry Hedges, Julian Higgins ja Hannah Rothstein (2009). *Introduction to Meta-Analysis*. Wiley, lk. 63–67, 71–73.

Guerra, Rudy ja Darlene Goldstein (2010). *Meta-analysis and Combining Information in Genetics and Genomics*. Chapman & Hall/CRC, lk. 3.

Horn, Roger ja Charles Johnson (2013). *Matrix Analysis*. Cambridge University Press, lk. 47, 430.

Hyvärinen, Aapo, Jarmo Hurri ja Patrik Hoyer (2009). *Natural Image Statistics*. Springer, lk. 125.

Ida-Tallinna Keskhaigla (2026). *Gestatsioonidiabeet*. URL: https://www.itk.ee/patsiendile/patsiendi-infomaterjalid/haigused/gestatsioonidiabeet (vaadatud 15.04.2026).

Kaart, Tanel ja Tõnu Möls (2009). "Populatsioonigeneetika genotüüpide tasemel". Loengukonspekt, Tartu ülikool. URL: http://ph.emu.ee/~ktanel/MTMS_02_007/2009/loeng_01_2009.pdf (vaadatud 15.04.2026).

Kangro, Raul (2024). "Monte-Carlo meetodid". Loengukonspekt, Tartu ülikool. URL: https://moodle.ut.ee/pluginfile.php/2938402/mod_resource/content/3/MC_2024.pdf (vaadatud 15.04.2026).

Kollo, Tõnu (2026). "Maatriksid statistikas". Loengukonspekt, Tartu ülikool. URL: https://moodle.ut.ee/pluginfile.php/2952408/mod_resource/content/17/RL2026k.pdf (vaadatud 15.04.2026).

Kool, Pille (2010). "Sõltuvate uuringute meta-analüüs". Magistritöö. Tartu ülikool. URL: https://dspace.ut.ee/server/api/core/bitstreams/7a3e7964-a4f7-49cb-ad0a-7399059a0060/content (vaadatud 15.04.2026).

Kuljus, Kristi (2026). "Mitmemõõtmelised statistilised meetodid". Loengukonspekt, Tartu ülikool. URL: https://moodle.ut.ee/mod/folder/view.php?id=1419503 (vaadatud 15.04.2026).

Mändul, Merli (2016). "Geneetilise tagasiside mõju ravitulemusele". Tartu ülikool. URL: https://dspace.ut.ee/server/api/core/bitstreams/06715ba5-2613-423d-8624-032a564c103b/content (vaadatud 15.04.2026).

Möls, Märt (2026). "Biostatistika". Loengukonspekt, Tartu ülikool. URL: https://www-1.ms.ut.ee/mart/biostat2026/loeng4.pdf (vaadatud 15.04.2026).

National Human Genome Research Institute (2026a). *Allele*. URL: https://www.genome.gov/genetics-glossary/Allele (vaadatud 15.04.2026).

National Human Genome Research Institute (2026b). *Genome-Wide Association Studies (GWAS)*. URL: https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies-GWAS (vaadatud 15.04.2026).

OpenAI (2026). *ChatGPT*. Vestluspõhine abiline koodi kirjutamisel. URL: https://chatgpt.com/ (vaadatud 11.05.2026).

Põru, Getter (2018). "Veekogude klassifitseerimine satelliidiandmetelt multinomiaalse logistilise mudeliga". Tartu ülikool. URL: https://dspace.ut.ee/server/api/core/bitstreams/0a0138f7-c160-42b6-afca-cb79c4e9673a/content (vaadatud 15.04.2026).

Rossi, Richard (2018). *Mathematical Statistics: An Introduction to Likelihood Based Inference*. Wiley, lk. 205, 310.

Schmid, Christopher, Theo Stijnen ja Ian White (2021). *Handbook of Meta-Analysis*. Chapman & Hall/CRC, lk. 130.

Sova, Joonas (2023). "Tõenäosusteooria ja statistika II". Loenguslaidid, Tartu ülikool. URL: https://moodle.ut.ee/pluginfile.php/2784220/mod_resource/content/20/loeng.pdf (vaadatud 15.04.2026).

Xu, ChangJiang, Joanna Tachmazidou, Klaudia Walter, Antonio Ciampi, Eleftheria Zeggini ja Celia Greenwood (2014). "Estimating Genome-Wide Significance for Whole-Genome Sequencing Studies". *National Library of Medicine*. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC4489336/ (vaadatud 15.04.2026).

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kaarel Lusmägi,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Ülegenoomsete uuringute metaanalüüs ja metaregressioon", mille juhendajad on Märt Möls ja Reedik Mägi, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kaarel Lusmägi
12.05.2026