Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Hanna Sõnajalg

Statistiliselt ekvivalentsete argumenttunnuste kogumite leidmine

Matemaatilise statistika eriala

Magistritöö (30 EAP)

Juhendajad: Oliver Aasmets, Ph.D
Prof. Krista Fischer, Ph.D

Tartu 2024

STATISTILISELT EKVIVALENTSETE ARGUMENTTUNNUSTE KOGUMITE LEIDMINE

Magistritöö

Hanna Sõnajalg

Lühikokkuvõte

Argumenttunnuste valik on mudeli konstrueerimisel üks olulisemaid ülesandeid. Meetodid nagu samm- ja lassoregressioon tagastavad ühe komplekti tunnustest, millega saavutatakse kõige paremini prognoosiv mudel. Kui andmetes esineb palju tugevalt korreleeritud tunnuseid, võib mitu tunnuste komplekti anda sarnase prognoosimisvõimega mudeleid. Statistiliselt ekvivalentsete argumenttunnuste kogumite leidmise (inglise keeles *statistically equivalent signatures* ehk SES) algoritm rakendab tunnuste valikuks korduvalt tingliku sõltumatuse teste. Lõpuks tagastatakse omavahel ekvivalentsete tunnuste kogumid. Valides igast kogumist täpselt ühe tunnuse, jõutakse erinevate mudeliteni, mis võiksid anda sarnase täpsusega hinnanguid. Magistritöö eesmärk on testida algoritmi Eesti geenivaramu andmetel, kuhu kuuluvad geenidoonorite vere metaboliidi kontsentratsioonid ning metaboliitide kontsentratsioonide suhete väärtused. Lineaarse regressioonimudeli abil prognoositakse kehamassiindeksit ja logistilise regressioonimudeli abil suremust 5 aasta jooksul.

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

Märksõnad: Argumenttunnuste valik, statistiliselt ekvivalentsed mudelid, masinõpe.

SELECTION OF STATISTICALLY EQUIVALENT FEATURE SUBSETS

Master thesis

Hanna Sõnajalg

Abstract

Feature selection is one of the most important tasks in model construction. Methods such as stepwise regression and lasso regression return a single set of variables which have the highest predictive power. When there are many highly correlated variables in the data, multiple sets of features can produce models with similar predictive power. The algorithm for finding sets of statistically equivalent signatures (SES) repeatedly applies conditional independence tests for feature selection. Ultimately, it returns sets of equivalent variables. By selecting exactly one from each set, different models can be created that may provide similarly accurate predictions. The aim of this master's thesis is to test the algorithm on data from the Estonian Biobank, which includes blood metabolite concentrations and the values of metabolite concentration ratios. Using a linear regression model, the body mass index is predicted, and using a logistic regression model, mortality within 5 years is predicted.

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

Key Words: Feature selection, statistically equivalent signatures, machine learning.

Sisukord

Sissejuhatus

Mitmete argumenttunnuste valiku meetodite eesmärk on leida üks alamkogum tunnustest, mille kaasamisel mudelisse saavutatakse kõige parem mudeli sobivus. Ette võib tulla aga olukordi, kus andmetes on tugevalt korreleeritud tunnuste gruppe. Siis võiks olla kasulik rakendada meetodeid, mis annavad võimaluse valida mitme argumenttunnuste komplekti vahel. Üks kindel tunnuste kombinatsioon võib küll anda parima mudeli sobivuse näitaja väärtuse, kuid teiste võimalike kombinatsioonide puhul ei pruugi erinevus olla oluline.

Statistiliselt ekvivalentsete argumenttunnuste kogumite leidmise (inglise keeles *statistically equivalent signatures* ehk SES) algoritm tagastab tunnuste grupid, kust kasutaja saab igast ühest valida täpselt ühe tunnuse selleks, et panna kokku sobiv argumenttunnuste komplekt. Algoritmi töö põhineb tingliku sõltumatuse testide korduval rakendamisel. SES algoritmi rakendamiseks on loodud programmeerimiskeeles R paketti MXM kuuluv funktsioon *SES()*.

Magistritöös tutvustatakse algoritmi tööpõhimõtet ja tuuakse rakendustarkvara R kasutades näiteid funktsiooni *SES()* kasutamisest genereeritud ja ka päriselulistel andmetel. Tartu Ülikooli Eesti geenivaramu andmeid kasutades püütakse leida kehamassiindeksit ja suremise tõenäosust prognoosivate metaboliitide hulka.

1 Regressioonimudelid

Regressioonimudeleid kasutatakse tunnustevaheliste seoste kirjeldamiseks, kus üht sõltuvat tunnust üritatakse kirjeldada teiste sõltumatute tunnuste abil. Sealjuures, kui sõltuv tunnus on pidev (näiteks kehamassiindeks), kasutatakse seoste kirjeldamiseks lineaarset regressiooni. Kui sõltuv tunnus on binaarne (näiteks suremus 5 aasta jooksul), kasutatakse logistilist regressiooni. Järgnevalt kirjeldatakse mudelite kuju, parameetrite hindamise ideed ja üht mudeli headuse näitajat, mida hiljem kasutatakse mudelite valideerimisel ja erinevate argumenttunnuste valiku meetoditega saadud mudelite võrdlemisel. Samuti kirjeldatakse nii lineaarse kui ka logistilise regressioonimudeli korral testi, millega kontrollitakse mingi hulga argumentunnuste statistilist olulisust mudelis korraga. Selgitatakse, kuidas SES algoritm antud testi abil tinglikku sõltumatust testib.

Olgu valimi maht n ning tunnuste arv p. Vaadeldakse indiviidi $i \in \{1, \cdots, n\}$. Tema kohta on teada tunnuste väärtused $x_{1i}, x_{2i}, \cdots, x_{pi}$. Informatsioon indiviidide kohta koondatakse maatriksisse $\mathcal{X}$

$$\mathcal{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Maatriksi $\mathcal{X}$ i. rida tähistatakse sümboliga $\mathbf{x}_i$ ehk $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})$.

1.1 Lineaarse regressioonimudeli kuju ja parameetrite hindamine

Olgu $\mathbf{y} = (y_1, \cdots, y_n)^T$ pidev funktsioontunnus (näiteks kehamassiindeks). Lineaarse regressioonimudeli kohaselt avaldub tunnuse $\mathbf{y}$ keskväärtus argumenttunnuste ja hinnatavate parameetrite lineaarse funktsioonina. Olgu parameetrite vek-

tor $\boldsymbol{\beta} = (\beta_0,\beta_1,\cdots,\beta_p)^T$. Mudeli kuju indiviidi i jaoks on

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i,$$

kus ε_i tähistab juhuslikku viga. Seejuures eeldame, et tunnuse $\mathbf{y}$ mõõtmistulemused on statistiliselt sõltumatud ja juhuslike vigade vektor $\boldsymbol{\varepsilon} = (\varepsilon_1,\cdots,\varepsilon_n)^T$ on n-mõõtmelise normaaljaotusega juhuslik vektor $\boldsymbol{\varepsilon} \sim \mathcal{N}(0;\sigma^2\mathbf{I})$, kus $\mathbf{I}$ tähistab ühikmaatriksit. (Hastie, Tibshirani ja Friedman, 2009)

Parameetrite $\boldsymbol{\beta}$ hinnangute $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p)^T$ leidmiseks kasutatakse vähimruutude meetodit. Eesmärk on minimeerida prognoosivigade $e_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_{ij}$ ruutude summa $\sum_{i=1}^{n} e_i^2$. Selleks võetakse prognoosivigade ruutude summast tuletis parameetervektori järgi ja võrdsustatakse see nulliga. Tähistame mudeli plaanimaatriksi sümboliga $\mathbf{X}$ ehk

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}.$$

Eeldades, et maatriks $\mathbf{X}^T\mathbf{X}$ pole singulaarne, avaldub vähimruutude hinnang parameetervektorile kujul

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

kus $(\mathbf{X}^T\mathbf{X})^{-1}$ tähistab maatriksi $\mathbf{X}^T\mathbf{X}$ pöördmaatriksit. (Hastie, Tibshirani ja Friedman, 2009)

1.2 Logistilise regressioonimudeli kuju ja parameetrite hindamine

Olgu nüüd $\mathbf{y} = (y_1, \cdots, y_n)^T$ binaarne tunnus väärtusega 1, kui huvipakkuv sündmus toimus, ja väärtusega 0, kui ei toimunud. Olgu $\mathbb{P}(y_i = 1) = \pi_i$, kus $i \in \{1, \cdots, n\}$. See tähendab, et y_i on Bernoulli jaotusega parameetriga π_i ja $E\mathbf{y} = \boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)^T$. Logistilise regressioonimudeliga hinnatakse logaritmitud sündmuse toimumise šansse

$$\ln \frac{\pi_i}{1 - \pi_i} = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

ehk

$$\pi_i = \frac{\exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\right)},$$

kus $i \in \{1, \cdots, n\}$.

Funktsiooni, mis seab uuritava tunnuse $\mathbf{y}$ väärtusele y_i vastavusse tema esinemise tõenäosuse $P(\mathbf{y} = y_i) = p(y_i)$, nimetatakse tõenäosusfunktisooniks ja Bernoulli jaotuse tõenäosusfunktsioon on

$$p(y_i) = P(\mathbf{y} = y_i) = (\pi_i)^{y_i}(1 - \pi_i)^{1-y_i}.$$

Parameetervektori $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ hinnang $\hat{\boldsymbol{\beta}}$ leitakse suurima tõepära meetodil. Tõepärafunktsioon avaldub kujul:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} p(y_i) = \prod_{i=1}^{n} (\pi_i)^{y_i}(1 - \pi_i)^{1-y_i}.$$

Logaritmilise tõepärafunktsiooni kuju on

$$\begin{aligned} l(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) &= \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i) = \\ &= \sum_{i=1}^{n} y_i \ln(\pi_i) - y_i \ln(1-\pi_i) + \ln(1-\pi_i) = \\ &= \sum_{i=1}^{n} y_i \ln\frac{\pi_i}{1-\pi_i} + \ln(1-\pi_i). \end{aligned}$$

Funktsioonist $l(\boldsymbol{\beta})$ võetakse tuletis parameetervektori $\boldsymbol{\beta}$ järgi. Tulemus võrdsustatakse nulliga ja jõutakse $\beta_0, \beta_1, \cdots, \beta_p$ suhtes mittelineaarsete võrranditeni, mille lahendamiseks kasutatake Newton–Raphsoni algoritmi. (Hastie, Tibshirani ja Friedman, 2009)

1.3 Mudelite headuse näitajad MSE ja AUC

Lineaarse regressioonimudeli korral võib mudeli täpsuse hindamiseks kasutada ruutjuurt keskmise ruutvea hinnangust (RMSE -*Root Mean Squared Error*)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} e_i^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2}.$$

Erinevalt keskmise ruutvea hinnangust MSE, on RMSE väärtus samal skaalal funktsioontunnuse väärtustega (Hyndman ja Koehler, 2006).

Logistilise regressioonimudeli üks headuse näitaja on ROC-kõvera alune pindala AUC (*area under the curve*). Oletame, et leitud on mudeli parameetrite hinnangud $\hat{\boldsymbol{\beta}}$. Nende abil leiame iga isiku korral $\hat{\pi}_i = P(y_i = 1|\mathbf{x}_i, \hat{\boldsymbol{\beta}})$ ehk prognoosi sündmuse toimumise tõenäosusele. Kui tahaksime tõenäosuste põhjal isikud jagada kahte klassi, peaksime fikseerima mingi piirväärtuse, näiteks 0,5. Kui hinnatud tõenäosus on väiksem kui 0,5, siis liigitame isiku klassi $Y = 0$. Kui hinnatud tõenäosus on 0,5 või suurem, liigitame ta klassi $Y = 1$. Ühe valitud piiri korral saadud tulemused

saame koondada kahemõõtmelisse sagedustabelisse 1.

Tabel 1: Logistilise regressioonimudeli põhjal inimeste klassidesse jagunemine.

	Tegelik olek	
Prognoos	$Y = 0$	$Y = 1$
$Y = 0$	TN	FN
$Y = 1$	FP	TP

Tabeli põhjal saame leida tõeselt positiivsete määra TP/(TP + FN) ja valepositiivsete määra FP/(TN + FP) piirväärtuse 0,5 korral. ROC-kõvera graafikule kujutatakse y-teljele tõeselt positiivsete määr ja x-teljele valepositiivsete määr erinevate piirväärtuste korral. ROC-kõvera kirjeldamiseks kasutatakse AUC väärtust ehk leitakse joonealuse piirkonna pindala. AUC väärtused asuvad 0 ja 1 vahel. Kui AUC = 0,5 tähendab see, et klassifitseerimismudel töötab sama hästi kui juhuslik inimeste klassidesse jagamine, mida suurem on AUC väärtus, seda täpsemini mudel prognoosib. (Fawcett, 2006)

Olgu mudeliga hinnatud tõenäosused $\hat{\pi_1}, \hat{\pi_2}, \cdots, \hat{\pi_n}$ järjestatud kasvavalt ehk moodustatud on variatsioonirida. AUC väärtus on samaväärne tõenäosusega, et juhuslikult valitud isik i, kes (tegelikkuses) kuulub klassi $Y = 1$, asub tõenäosuste järjekorras kaugemal, kui juhuslikult valitud isik, kes (tegelikkuses) kuulub klassi $Y = 0$. See tähendab, et AUC väärtuse leidmine on samaväärne Mann-Whitney U-testi statistiku leidmisega. (Fawcett, 2006)

Mann-Whitney U-testi statistiku väärtus on

$$\mathrm{U} = \sum_{i=1}^{n_1} r_{1i} - \frac{n_1(n_1+1)}{2},$$

kus n_1 tähistab isikute arvu, kes tegelikult kuuluvad klassi $Y = 1$, ja r_{1i} tähistab i. isiku, kes tegelikult kuulub klassi $Y = 1$, tõenäosuse $\hat{\pi}_i$ astakut kasvavas variatsioonireas, mis on moodustatud kõigile isikutele hinnatud tõenäosustest. AUC

väärtuse leidmiseks kasutatakse seost

$$\text{AUC} = \frac{\text{U}}{n_0 n_1},$$

kus $n_0 = n - n_1$ ehk n_0 on inimeste arv, kes tegelikult kuuluvad klassi $Y = 0$. (Mason ja Graham, 2002)

1.4 F-test ja tõepärasuhte test

Lineaarse regressiooni korral testitakse SES algoritmis tinglikut sõltumatust F-testiga. Selleks hinnatakse lineaarne regressioonimudel, kus funktsioontunnus on Y ja argumenttunnused on $X_1, \cdots, X_s$. Seejärel hinnatakse keerukam mudel, kus argumenttunnuste hulka on lisatud X_{s+1}. Kui X_{s+1} on pidev tunnus, siis on keerukamas mudelis hinnatavaid parameetreid ühe võrra rohkem kui lihtsamas mudelis. Kui X_{s+1} on k tasemega faktortunnus, siis on uute parameetrite arv keerukamas mudelis $k-1$. Testitakse, kas tunnus X_{s+1} on keerukamas mudelis statistiliselt oluline ehk kas mõni uutest hinnatud parameetritest on erinev nullist. Kui on, siis see tähendab, et Y sõltub tunnusest X_{s+1} tinglikustatuna tunnuste $X_1, \cdots, X_s$ järgi ehk $(Y \perp\!\!\!\perp X_{s+1}) | X_1, \cdots, X_s$.

Vaatleme kaht lineaarset regressioonimudelit

$$\mathcal{M}_0 : y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-r} x_{i,p-r} + \varepsilon_i \quad \text{ja}$$
$$\mathcal{M}_1 : y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-r} x_{i,p-r} + \beta_{p-r+1} x_{i,p-r+1} + \cdots + \beta_p x_{i,p} + \varepsilon_i,$$

kus $i \in \{1, \cdots, n\}$ ja $r < p$. Paneme tähele, et $\mathcal{M}_0$ on erijuht mudelist $\mathcal{M}_1$, sest võrdsustades mudelis $\mathcal{M}_1$ parameetrid $\beta_{p-r+1}, \cdots, \beta_p$ nulliga, saame mudeli $\mathcal{M}_0$.

Jagame parameetervektori $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_p)^T$ kaheks osaks

$$\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_{\text{vana}} \\ \boldsymbol{\beta}_{\text{uus}} \end{pmatrix},$$

kus $\boldsymbol{\beta}_{\text{vana}} = (\beta_0, \beta_1, \cdots, \beta_{p-r})^T$ ja $\boldsymbol{\beta}_{\text{uus}} = (\beta_{p-r+1}, \cdots, \beta_p)^T$.

Testitakse hüpoteese

$$\begin{aligned} H_0 &: \text{lihtsam mudel on õige ehk } \boldsymbol{\beta}_{\text{uus}} = \mathbf{0}, \\ H_1 &: \text{keerukam mudel on õige ehk } \boldsymbol{\beta}_{\text{uus}} \neq \mathbf{0}, \end{aligned} \tag{1}$$

kus $\mathbf{0}$ on $(p-r)$-mõõtmeline nullvektor. Selleks saab kasutada F-testi. Nullhüopoteesi kehtides kasutatakse tulemust

$$F = \frac{(\text{RSS}_0 - \text{RSS}_1)/r}{\text{RSS}_1/(n-p)} \sim F_{r,n-p}, \tag{2}$$

kus RSS_0 on mudeli $\mathcal{M}_0$ prognoosivigade ruutude summa ja RSS_1 on mudeli $\mathcal{M}_1$ prognoosivigade ruutude summa. Kui $F > F_{r,n-p;\alpha}$, siis kummutatakse H_0. (Hastie, Tibshirani ja Friedman, 2009)

Logistilise regressiooni korral kasutatakse tingliku sõltumatuse testimiseks tõepärasuhte testi. Vaatleme kaht logistilise regressioonimudelit: esimene (lihtsam) mudel on erijuht teisest ehk teisest mudelist saame esimese, kui võrdsustame osa parameetreid nulliga. Olgu esimese mudeli parameetervektor jällegi $\boldsymbol{\beta}_{\text{vana}}$ ja teise mudeli parameetervektor $\boldsymbol{\beta}$.

Selleks, et testida hüpoteese (1), leitakse tõepärasuhe

$$\lambda = \frac{\max_{\boldsymbol{\beta}_{\text{vana}}} L(\boldsymbol{\beta}_{\text{vana}})}{\max_{\boldsymbol{\beta}} L(\boldsymbol{\beta})}$$

ja kasutatakse tulemust, et nullhüpoteesi kehtides

$$-2\ln\lambda \sim \chi^2_r. \tag{3}$$

Kui $-2\ln\lambda > \chi^2_{df=r;\alpha}$, kummutatakse H_0. (McCullagh, 1989)

2 Argumenttunnuste valiku meetodid

Mudeli konstrueerimisel on argumenttunnuste valik üks olulisemaid ülesandeid. Üleliigsete muutujatega mudel võib uutel andmetel anda väga ebatäpseid hinnanguid vaatamata sellele, et treeningandmetele mudel sobitus. Samuti muudab ebavajalike argumenttunnuste eemaldamine mudeli lihtsamini interpreteeritavamaks, sest üksikute tunnuste mõju funktsioontunnusele tuleb selgemini esile. Argumenttunnuste valiku automatiseerimiseks võib kasutada näiteks sammregressiooni või lassoregressiooni meetodeid.

2.1 Sammregressioon

Olgu potentsiaalsete argumenttunnuste koguarv andmestikus p. Siis võimalike kombinatsioonide arv 0 kuni p argumenttunnuse valimiseks on 2^p. Selle asemel, et hinnata kõikvõimalikud mudelid, hinnatakse sammregressiooni meetodi puhul vaid osahulk neist. Sammregressiooni meetodid jagunevad kolmeks: ettepoole ehk kasvav valik, tahapoole ehk kahanev valik ja segavalik. Segavaliku meetodi puhul rakendatakse kasvava ja kahaneva valiku meetodeid vahelduvalt. (James *et al.*, 2013)

Mudelite võrdlemiseks saab kasutada erinevaid hindamiskriteeriume. Näiteks võib parima mudeli otsimist läbi viia Bayesi informatsioonikriteeriumi ehk BIC väärtuse põhjal. Olgu valimi maht n, hinnatud mudel $\mathcal{M}$ ja hinnatud parameetrite arv p.

BIC väärtus avaldub kujul

$$\mathrm{BIC}(\mathcal{M}) = -2 \ln L(\hat{\boldsymbol{\beta}}) + p \ln (n),$$

kus $\hat{\boldsymbol{\beta}}$ tähistab parameetervektorit, mis maksimiseerib tõepära funktsiooni $L(\boldsymbol{\beta})$. BIC väärtuse minimiseerimisel maksimiseeritakse tõepära, kuid samal ajal takistatakse hinnatud parameetrite arvul liiga suureks kasvada. Argumenttunnuste otsimine BIC väärtuse põhjal aitab leida tasakaalu mudeli keerukuse ja selgitusvõime vahel. (Wit, Heuvel ja Romeijn, 2012)

Ettepoole ehk kasvava valikuga sammregressiooni puhul alustatakse mudelist, kus pole ühtegi argumenttunnust. Seejärel hinnatakse iga argumenttunnusega eraldi üks mudel ja leitakse nende BIC väärtused. Lõplikusse mudelisse valitakse tunnus, millega saavutati kõige väiksem BIC väärtus. Igal järgneval sammul lisatakse sarnaselt juurde tunnus, mis vähendab eelmisel sammul hinnatud mudeli BIC väärtust kõige enam. Kui jõutakse sammuni, kus ühegi muutuja lisamine enam BIC väärtust ei paranda, lõpetatakse töö ja väljastatakse mudel, milleni jõuti. Tahapoole ehk kahaneva valikuga sammregressiooni puhul alustatakse täismudelist, kuhu on kaasatud kõik p argumenttunnust. Esimesel sammul leitakse BIC väärtused kõigi mudelite jaoks, kust on üks muutuja eemaldatud. Kõige väiksema BIC väärtusega mudelist välja jäetud tunnus eemaldatakse ka lõplikust mudelist. Igal järgneval sammul jäetakse välja tunnus, mille eemaldamisel saadud mudel on vähima BIC väärtusega. Kui jõutakse sammuni, kus ühegi parameetri eemaldamisel enam BIC väärtus ei kahane, lõpetatakse töö ja väljastatakse mudel, milleni jõuti. (Venables ja Ripley, 2002)

Lisaks BIC väärtusel põhinevale sammregressioonile, viiakse magistritöös läbi kahaneva valiku meetod parameetrite p-väärtuste põhjal. Alustatakse täismudelist ja eemaldatakse kõige suurema p-väärtusega tunnus. Mudel hinnatakse uuesti ja protsess kordub, kuni jõutakse mudelini, kus kõik parameetrid on statistiliselt olulised. Seejuures arvestatakse, et läbi viiakse mitu testi, ehk eelnevalt korrigeeri-

takse olulisuse nivood sõltuvalt argumenttunnustest. Korreleeritud andmete puhul on välja pakutud idee arvesse võtta sõltumatute testide arvu M_{eff}, mille arvutamiseks leitakse andmete korrelatsioonimaatriksi omaväärtused $\lambda_1, \cdots, \lambda_M$, kus M on tunnuste arv. Sõltumatute testide arvu leidmiseks on näiteks välja pakutud valem

$$M_{eff} = \begin{cases} \sum_{i=1}^{M} f(|\lambda_i|) \\ f(x) = I(x \geq 1) + (x - \lfloor x \rfloor), \ x \geq 0, \end{cases} \tag{4}$$

kus $I(x \geq 1) = 1$, kui $x \geq 1$ ja $I(x \geq 1) = 0$, kui $x < 1$, ja $\lfloor x \rfloor$ tähistab maksimaalset täisarvu, mis on väiksem kui x või võrdne x-ga (Li ja Ji, 2005). Teine valem M_{eff} väärtuse leidmiseks on

$$M_{eff} = \left(\frac{\sum_{i=1}^{M} \sqrt{\lambda_i}}{\log \lambda_1} \right)^2 \Bigg/ \left(\frac{\sum_{i=1}^{M} \lambda_i}{\lambda_1} + \sqrt{\lambda_i} \right), \tag{5}$$

kus λ_1 on korrelatsioonimaatriksi suurim omaväärtus. Korrigeeritud olulisuse nivoo on α / M_{eff} (Peluso, Glen ja Ebbels, 2021).

2.2 Lassoregressioon

Lassoregressioon võimaldab teostada argumenttunnuste valikut, penaliseerides parameetrite hinnangute absoluutväärtuste summat. Tulemusena kahanevad ebaoluliste parameetrite hinnangute absoluutväärtused nulliks. (Tibshirani, 1996)

2.2.1 Meetodi idee

Lassoregressiooni idee selgitus põhineb Robert Tibshirani artiklil „Regression Shrinkage and Selection via the Lasso“ (1996). Lineaarse regressioonimudeli parameetrite

lasso meetodi hinnangud leitakse, kui minimiseeritakse avaldist

$$\begin{aligned}\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|=\\ =\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|,\end{aligned} \tag{6}$$

kus $\lambda \geq 0$ on hüperparameeter, mis kontrollib parameetrite absoluutväärtustele rakendatava kahandamise suurust. Seega avaldub lasso hinnang lineaarse regressioonimudeli parameetervektorile $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)^T$ kujul

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Kui $\lambda = 0$, siis $\hat{\boldsymbol{\beta}}^{\text{lasso}}$ on võrdne vähimruutude meetodil saadud hinnaguga $\hat{\boldsymbol{\beta}}$. Erinevatele hüperparameetri $\lambda > 0$ väärtustele vastavad erinevad parameetervektori hinnangud. Parima ennustustäpsusega mudelini viiva λ leidmiseks kasutatakse ristvalideerimsit, mida kirjeldatakse järgmises alapeatükis.

Üldistatud lineaarsete mudelite korral saab samuti argumenttunnuste valikut läbi viia lassoregressiooni meetodil, penaliseerides parameetrite hinnangute absoluutväärtuste summat. Näiteks kui parameetervektori hinnangud leitakse log-tõepära maksimiseerides, saame lasso hinnangud leida, kui asendame avaldises (6) jääkide ruutude summa log-tõepära funktsiooniga. See tähendab, et logistilise regressioonimudeli parameetrite hinnangud leitakse, kui minimiseeritakse avaldist

$$-l(\boldsymbol{\beta})+\lambda\sum_{j=1}^{p}|\beta_j| = -\sum_{i=1}^{n}\left(y_i \ln\frac{\pi_i}{1-\pi_i}+\ln(1-\pi_i)\right)+\lambda\sum_{j=1}^{p}|\beta_j|.$$

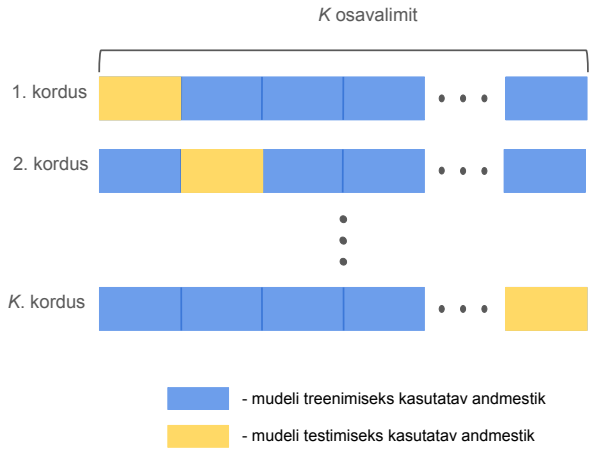
Järelikult avaldub lasso hinnang logistilise regressiooni parameetervektorile kujul

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \text{argmin}_{\boldsymbol{\beta}} \left[-\sum_{i=1}^{n} \left(y_i \ln \frac{\pi_i}{1-\pi_i} + \ln\left(1-\pi_i\right) \right) + \lambda \sum_{j=1}^{p} |\beta_j| \right].$$

2.2.2 Optimaalsete hüperparameetrite valimine

Optimaalsete hüperparameetrite valikuks saab kasutada ristvalideerimise meetodit. Ristvalideerimise eesmärk on leida hüperparameetrie kombinatsioon, millega saavutatakse kõige paremini prognoosiv mudel. Ristvalideerimise ülevaade põhineb teosel „The elements of statistical learning: data mining, inference and prediction“ (2009).

Ristvalideerimise korral jagatakse andmestik võrdsete suurustega osadeks. Andmestiku iga alamhulk täidab mudeli hindamise protsessis testandmestiku rolli, mudeli parameetrie hindamiseks kasutatakse ülejäänud andmeid. K-kordse ristvalideerimise korral on osade arv K. K-kordse ristvalideerimise skeem on toodud joonisel 1.



Joonis 1: K-kordse ristvalideerimise skeem

Vaatleme lineaarse regressioonimudeli hindamise ülesannet. Oletame, et oleme jõudnud $k \in \{1, \cdots, K\}$. sammuni ehk k. andmestiku alamhulka kasutatakse testandmestikuna. Olgu $\kappa : \{1, \cdots, n\} \to \{1, \cdots, K\}$ funktsioon, mis seab iga indiviidi indeksile kogu andmestikus vastavusse andmestiku alamhulga indeksi, kuhu indiviidi juhuslikult määrati. Ülejäänud $K-1$ osal hinnatakse lineaarne regressioonimudel ja jõutakse parameetervektori hinnanguni $\hat{\boldsymbol{\beta}}^{-k}$, kus ülaindeks märgib, et hindamiseks k. osa andmestikust ei kasutatud. Tähistagu $\hat{\boldsymbol{\beta}}^{-\kappa(i)} = (\hat{\beta}_0^{-\kappa(i)}, \cdots, \hat{\beta}_p^{-\kappa(i)})^T$ sellist parameetrite hinnangut, milleni jõuti, kui objekt $i \in \{1, \cdots, n\}$ kuulus testandmestikku (mitte mudeli parameetrite hindamiseks kasutatud andmestikku). Leitakse ristvalideerimise hinnang keskmisele jääkide ruutude summale

$$\mathrm{CV}_{\mathrm{MSE}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - X_i \hat{\beta}^{-\kappa(i)}).$$

Üldisemalt, tähistame hinnatud lineaarse või logalistilise regressioonimudeli funktsioonina $\hat{f}(x)$, mis seab indiviidile $x \in X$ vastavusse kas uuritava tunnuse (näiteks kehamassiindeks või 5 aasta jooksul suremise tõenäosus) prognoosi. Olgu $\mathbf{y} = (y_1, \cdots, y_n)^T$ funktsioontunnuse vektor (ehk näiteks isikute reaalsed kehamassiindeksi väärtused või 5 aasta jooksul suremise indikaatori väärtused). Tähistagu $L(\mathbf{y}, \hat{f}(X))$ kaofunktsiooni, mille abil mõõdetakse viga $\mathbf{y}$ väärtuste ja prognooside vahel. L võib olla lineaarse regressiooni korral näiteks MSE, RMSE või logistilise regressiooni korral näiteks AUC väärtus. Olgu $\hat{f}^{-k}(x, \boldsymbol{\theta})$ mudel fikseeritud hüperparameetrite $\boldsymbol{\theta}$ korral, kus ülaindeks $-k$ tähendab, et k. alamhulka andmestikust pole kasutatud mudeli hindamiseks. Ristvalideerimise hinnang keskmisele kaofunktsiooni väärtusele avaldub kujul

$$\mathrm{CV}(\hat{f}, \lambda) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}^{-\kappa(i)}(x_i, \lambda)).$$

Ristvalideerimise hinnang leitakse mitme $\boldsymbol{\theta}$ korral. Kaofunktsioonist olenevalt (näi-

teks MSE väärtust minimiseeritakse, AUC väärtust maksimiseeritakse), leitakse optimaalne $\boldsymbol{\theta}$ ning seda kasutades hinnatakse mudel kõigil andmetel (ehk kõik ristvalideerimise käigus moodustatud osavalimid ühendatakse). Lassoregressiooni korral koosneb vektor $\boldsymbol{\theta}$ ainult ühest hüperparameetrist λ ja parima lassoregressiooni mudeli saamiseks proovitakse läbi erinevaid $\lambda \geq 0$ väärtuseid.

Magistritöös hinnatakse lassoregressiooni mudelid nii optimaalse λ väärtusega $\hat{\lambda}_{min}$, mis annab parima ristvalideerimise hinnangu keskmisele kaofunktsiooni väärtusele, kui ka λ väärtusega $\hat{\lambda}_{1se}$. Tegu on maksimaalse λ väärtusega, mille korral ristvalideerimise hinnang keskmisele kaofunktsiooni väärtusele ei erine optimaalsest ristvalideerimise hinnangust enam kui ühe standardvea võrra. Kuna $\hat{\lambda}_{min} < \hat{\lambda}_{1se}$, siis võib $\hat{\lambda}_{1se}$ korral olla nulliga võrduvate parameetrite arv suurem. (Friedman, Tibshirani ja Hastie, 2010)

3 SES algoritmi metoodika

Statistiliselt ekvivalentsete argumenttunnuste kogumite leidmise ehk SES algoritm võimaldab leida uuritava tunnusega seotud argumenttunnuste komplekte, mis on oma ennustusvõimelt ekvivalentsed. Algoritm on olemuselt sammregressioon meetod, mis igal sammul kontrollib, kas eelnevalt mudelisse lisatud tunnusele leidub ekvivalentseid tunnuseid argumenttunnuste seast. SES algoritmi on võimalik rakendada erinevat tüüpi tunnuste korral. Funktsioon- ja argumenttunnused võivad olla nominaalsed, pidevad või diskreetsed arvulised tunnused. Funktsioontunnus võib kirjeldada ka elukestust (tsenseerimisega). Tüübist ja jaotusest tulenevalt valib algoritm sobiva tingliku tõenäosuse testi, mida korduvalt rakendama hakatakse.

Peatükk on kirjutatud artikli „Feature Selection with the R package MXM: Discovering Statistically Equivalent Feature Subsets“ (Lagani *et al.*, 2017) põhjal.

3.1 Algoritmi põhitsükkel

Olgu V hulk, kuhu kuuluvad argumenttunnused, ja olgu Y funktsioontunnus. SES algoritmi eesmärk on leida tunnuste hulk, millest pole funktsioontunnust võimalik sõltumatuks muuta ülejäänud argumenttunnuste järgi tinglikustades. Selleks rakendab SES algoritm korduvalt tingliku sõltmutatuse teste. Olgu $R \subset V$ veel läbi vaatamata tunnuste hulk ja $S \subset V$ juba väljavalitud tunnuste hulk. Algoritmi töö alguses $R = V$ ja $S = \emptyset$. Algoritm täidab kordamööda kaht ülesannet:

(a) leiab tunnuse, mis on maksimaalselt seotud tunnusega Y tinglikustatuna mingisuguse alamhulga $Z \subset S$ järgi, ja lisab selle hulka S;

(b) välistab sellise tunnuse hulka S kaasamise, mis ei ole enam seotud tunnusega Y, tinglikustatuna ükskõik millise alamhulga $Z \subset S$ järgi.

Täpsemalt, vaadatakse läbi kõik tunnused $X \in \{R \cup S\}$. Vaadeldava tunnuse X puhul otsitakse seejärel tunnuste hulka $Z \subseteq S \setminus \{X\}, |Z| \leq k$, mille korral kontrollides hüpoteese

$$\begin{aligned} H_0 &: X \perp\!\!\!\perp Y | Z, \\ H_1 &: X \not\perp\!\!\!\perp Y | Z. \end{aligned} \tag{7}$$

jäädakse H_0 juurde. Tunnuste arv hulgas Z on piiratud parameetriga k. Kui selline Z leidub, tähendab see, et teades tunnuste hulga Z väärtusi, ei sõltu funktsioontunnus Y enam argumenttunnusest X. Sel juhul on tunnus X läbi vaadatud ning ta eemaldatakse hulkadest R ja S: $R = R \setminus \{X\}$ ning $S = S \setminus \{X\}$. Paneme tähele, et kui X arvatakse mingis tsüklis välja hulkadest R ja S, siis sinna tagasi see enam ei saa. Enne tunnuse X eemaldamist toimub temaga ekvivalentsete tunnuste otsimine juba väljavalitud tunnuste seast. Seda protsessi kirjeldatakse järgnevas alapeatükis. Paneme tähele, et põhitsükli esimesel iteratisoonil kehtib $Z = \emptyset$. Seega

kontrollitakse siis tavalisi sõltumatuse hüpoteese

$$H_0 : X \perp\!\!\!\perp Y$$
$$H_1 : X \not\perp\!\!\!\perp Y \tag{8}$$

Kui kõik $X \in \{R \cup S\}$ on läbi vaadatud, siis leitakse

$$M = \text{argmin}_{X \in R} \max_{Z \subseteq S}(p_{XY.Z}),$$

kus $p_{XY.Z}$ on hüpoteesipaarile (9) vastav p-väärtus. See tähendab, et funktsioontunnuse Y sõltuvus tunnusest $M \in R$ (erinevate $Z \subseteq S$ järgi tinglikustades) on kõige tugevam. Seejärel kaasatakse M väljavalitud argumenttunnuste hulka $S = S \cup \{M\}$ ning jäetakse välja hulgast R, $R = R \setminus \{M\}$. Siin lõpeb põhitsükli esimene iteratsioon.

Teise iteratsiooni algusega võrreldes on hulgad R ja S muutunud. Pöördutakse tagasi põhitsükli algusesse, et hakata jälle läbi vaatama kõiki tunnuseid $X \in \{R \cup S\}$. Igal iteratsioonil lisatakse hulka S tunnus M, millest on funktsioontunnus Y kõige tõenäolisemalt sõltuv, tinglikustatuna juba väljavalitud tunnuste hulga S mingisuguse alamhulga Z järgi. Seejuures eemaldatakse M veel läbivaatamata tunnuste hulgast R. Kui $R = \emptyset$, korratakse tsüklit veel viimast korda.

3.2 Ekvivalentsete tunnuste kogumid

Oletame, et argumenttunnuste arv on p. Enne algoritmi põhitsükli algust defineeritakse p ekvivalentsete tunnuste hulka Q_i, kus $i \in \{1, \cdots p\}$, sest alguses on kõik tunnused ekvivalentsed vaid iseendaga.

Oletame, et algoritmi põhitsükli alamtsüklis oleme jõudnud tunnuseni $X_1 \in \{R \cup S\}$, mille korral leidub $Z_1 \subseteq S \setminus \{X_1\}$ nii, et Y on X_1-st sõltumatu tinglikustatuna tunnuste hulga Z_1 järgi. See tähendab, et meil on valitud tunnuste seas juba paremaid valikuid funktsioontunnuse Y prognoosimiseks. Enne kui X_1 ära unustatakse,

vaadatakse, kas leidub hulgas Z_1 temaga ekvivalentne tunnus.

Otsitakse tunnust $W \in Z_1$ nii, et testides hüpoteese

$$\begin{aligned} H_0 &: W \perp\!\!\!\perp Y | Z_1', \\ H_1 &: W \not\perp\!\!\!\perp Y | Z_1'. \end{aligned} \tag{9}$$

kus $Z_1' = (Z_1 \cup \{X_1\}) \setminus \{W\}$, jäädakse H_0 juurde ehk Y on W-st sõltumatu tingimusel, et me teame hulga Z_1' tunnuste väärtusi. See tähendab, et tingliku sõltumatuse testides saame X_1 asendada tunnusega W ilma, et testide tulemus erineks, seega SES algoritmi kontekstis on tunnused ekvivalentsed. Siis ühendatakse tunnustega X_1 ja W ekvivalentsete tunnuste hulgad $Q_W = Q_W \cup Q_{X_1}$.

Kui algoritmi põhitsükkel on töötamise lõpetanud, kogutakse kokku kõik väljavalitud tunnuste hulga S elementidele vastavad ekvivalentsete tunnuste hulgad ja väljastatakse need. Kasutaja saab kokku panna sobiva argumenttunnuste komplekti funktsioontunnuse hindamiseks. Näiteks kui väljastatakse $Q_1 = \{X_1, X_4, X_7\}$, $Q_3 = \{X_3, X_9\}$ ja $Q_5 = \{X_5, X_{11}, X_{12}\}$, saab kokku panna $3 \cdot 2 \cdot 3 = 18$ erinevat tunnuste komplekti, mis SES algoritmi põhjal suudavad tunnuse Y varieeruvust kirjeldada ligikaudu sama hästi. Kombinatsiooni moodustamiseks peab igast hulgast valima täpselt ühe tunnuse. Praegu on kolm sobivat kombinatsiooni näiteks $\{X_1, X_3, X_5\}$, $\{X_4, X_9, X_5\}$ ja $\{X_1, X_3, X_{11}\}$.

3.2.1 Lühike näide SES algoritmi tsükli kahest esimesest iteratsioonist

Oletame, et uuritav tunnus on Y ja argumenttunnuste hulka kuuluvad N, O ja P. Väljavalitud tunnuste hulk on $S = \emptyset$ ja veel läbivaatamata tunnuste hulk on $R = \{N,O,P\}$. Fikseerime olulisuse nivoo α.

Esimesel iteratsioonil tehakse kolm sõltumatuse testi (esimene test tunnuste N ja Y vahel, teine O ja Y vahel ning kolmas P ja Y vahel), millele vastavad p-väärtused

olgu $p_{NY} < \alpha$, $p_{OY} < \alpha$ ja $p_{PY} > \alpha$. Kolmanda testi korral jäädakse nullhüpoteesi juurde ehk P ja Y on sõltumatud. Selle tagajärjel $R = R \setminus \{P\} = \{N,O\}$. Tunnus P on nüüd algoritmi töö lõpuni hulgast R (ja ka S) välja arvatud. Oletame, et $p_{NY} < p_{OY}$ ehk seos N ja Y vahel on kõige tugevam. Esimese iteratsiooni lõpus $S = S \cup \{N\} = \{N\}$ ja $R = R \setminus \{N\} = \{O\}$.

Algab algoritmi põhitsükli teine iteratsioon. Teisel iteratsioonil vaadatakse läbi tunnuseid $X \in R \cup S = \{N,O\}$. Läbi viiakse tingliku sõltumatuse teste tinglikustatuna tunnuste $Z = S \setminus \{X\}$ järgi. Ehk kui $X = N$, siis korratakse esimesel iteratsioonil tehtud testi ja saadakse p-väärtus $p_{NY} < \alpha$. Kui $X = O$ ehk $Z = \{N\}$, siis testitakse, kas O ja Y on sõltumatud tinglikustatuna N järgi.

Kui testi p-väärtus $p_{OY.N} > \alpha$, siis saame, et N-i teades tunnus O meile olulist lisainformatsiooni Y kohta ei sisalda. Aga võibolla O ja N on ekvivalentsed. Ehk Y oli N-iga küll tugevamalt seotud, aga seos Y ja O vahel on sarnane. Leitakse $p_{NY.O}$. Kui $p_{NY.O} > \alpha$, siis O ja N loetakse ekvivalentseteks. Kui $p_{NY.O} < \alpha$, tähendab see, et teades tunnust O, sisaldab N veel olulist infot tunnuse Y kohta. Siis tunnuseid O ja N ekvivalentseteks ei loeta.

Kui $p_{OY.N} < \alpha$, siis $S = S \cup \{O\} = \{N,O\}$ ja $R = R \setminus \{O\} = \emptyset$.

3.3 SES mudeli hüperparameetrite valik

Funktsiooni *SES()* kasutades tuleb fikseerida hüperparameetrid α ja k. Olulisusnivoo α tähistab kõigi tingliku sõltumatuse testide maksimaalset lubatud tõenäosust teha esimest liiki viga ehk võtta vastu H_1, kui tegelikult kehtib H_0. Parameeter k tähistab tingliku sõltumatuse testimisel (9) hulga Z maksimaalset suurust ehk tunnuste arvu hulgas Z. Optimaalsete parameetrite väärtuste valikuks on loodud funktsioon *cv.ses()*, mis kasutab K-kordset ristvalideerimise meetodit. Ristvalideerimise täpsem kirjeldus on peatükis 2.2.2.

Fikseeritakse hulk parameetrite $\boldsymbol{\theta} = (\alpha, k)$ väärtuseid, mida testida soovitakse.

Kõigi $\boldsymbol{\theta}$ väärtuste puhul leitakse ristvalideerimise hinnang keskmisele kaofunktsiooni väärtusele (näiteks MSE või AUC). Parameetrite väärtuseid $\boldsymbol{\theta}^* = (\alpha^*, k^*)$, mis annavad optimaalse ristvalideerimise hinnangu, kasutatakse algoritmi rakendamisel kogu andmestikul.

4 Algoritmi töötamine simuleeritud andmetel

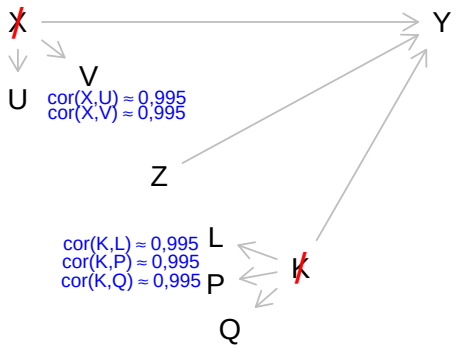
SES algoritmi testiti simuleeritud andmete peal. Standardsest normaaljaotusest genereeriti n-mõõtmelised vektorid X, Z ja K. Seejärel moodustati tunnusega X väga tugevalt korreleeritud tunnused U ja V (lineaarse korrelatsioonikordaja väärtus suurem kui 0,99)

$$U_i = X_i + \varepsilon_i \quad \text{ja}$$
$$V_i = X_i + \varepsilon_i,$$

kus $\varepsilon_i \sim \mathcal{N}(0; 0{,}1)$ ja $i \in \{1, \cdots, n\}$. Samuti moodustati tunnusega K väga tugevalt korreleeritud tunnused

$$L_i = K_i + \varepsilon_i,$$
$$P_i = K_i + \varepsilon_i \quad \text{ja}$$
$$Q_i = K_i + \varepsilon_i,$$

kus $\varepsilon_i \sim \mathcal{N}(0; 0{,}1)$ ja $i \in \{1, \cdots, n\}$. Funktsioontunnuse väärtused moodustati lineaarse kombinatsioonina tunnustest X, K ja Z. Katsetati kaht erinevat lineaarkombinatsiooni: $Y_{1i} = 2 \cdot X_i + 3 \cdot Z_i - 1{,}5 \cdot K_i + \varepsilon_i$ ja $Y_{2i} = 3 \cdot X_i + 2 \cdot Z_i - 1{,}5 \cdot K_i + \varepsilon_i$, kus $\varepsilon_i \sim \mathcal{N}(0; 1)$ ja $i \in \{1, \cdots, n\}$. Esimesel juhul on Y_1 kõige tugevamalt seotud tunnusega Z, millel puuduvad andmestikus tugevalt korreleeritud tunnused, ning teisel juhul on Y_2 kõige tugevamalt seotud tunnusega X, millega tugevalt korreleeritud tunnuseid on andmestikus kaks. SES algoritmile ei antud ette tunnuseid X ja K. Andmete genereerimise skeem on toodud joonisel 2.

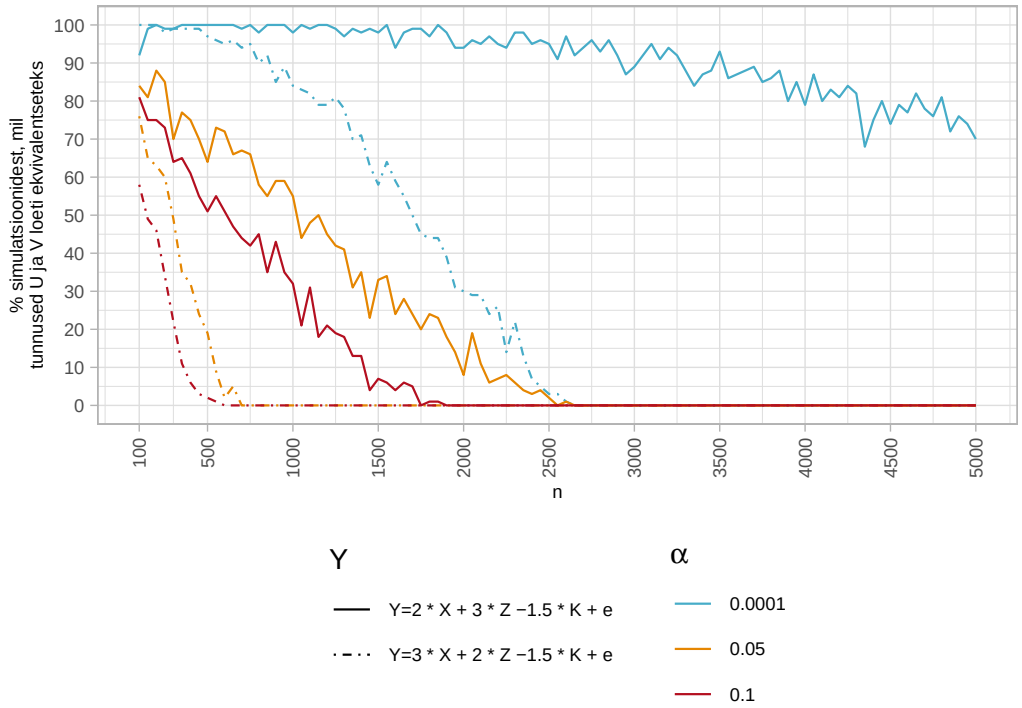


Joonis 2: Andmete genereerimise skeem, kus punase joonega on maha tõmmatud tunnused, mida algoritmile ette ei antud.

Andmeid genereeriti iga valimi mahu $n \in \{100{,}150{,}200, \cdots, 4950{,}5000\}$ puhul 100 korda. Igale andmestikule rakendati SES algoritmi ja vaadati, millised tunnused osutusid valituks, ja kas tunnustega X ja K tugevalt korreleeritud tunnused loeti ekvivalentseteks.

Joonisel 3 on y-teljel kujutatud osakaalu simulatsioonidest, mil tunnused U ja V loeti ekvivalentseteks, ja x-teljel valimi mahtu n. Värviga on tähistatud olulisuse nivoo α väärtus, mida SES algoritm tinglikku sõltumatust testides kasutas. Katkematu joonega on tähistatud tulemused, kui funktsioontunnus Y on kõige tugevamalt seotud tunnusega Z, ja katkendliku joonega on tähistatud tulemused, kui funktsioontunnus Y on kõige tugevamalt seotud tunnusega X.

Mida suurem on valimi maht, seda harvem loetakse tunnused U ja V ekvivalentseteks, ja mida väiksem on olulisuse nivoo α, seda sagedamini loetakse U ja V ekvivalentseteks. On näha ka, et kordajad Y tunnuse lineaarkombinatsioonis mõjutavad tulemusi. Kui Y sõltub kõige tugevamalt tunnusest X, siis loetakse U ja V harvem samaväärseteks. Osakaal langeb valimimahu n kasvades kiiremini, võrreldes teise lineaarkombinatsiooniga. Kõige väiksema $\alpha = 0{,}0001$ ja esimese lineaarkombinatsiooni korral on osakaal simulatsioonidest, mil U ja V loeti ekvivalentseteks, kõige kõrgem, ning valimi mahu kasvades langeb see osakaal kõige aeglasemalt. Valimi mahu $n = 5000$ korral on see ligikaudu 70%.



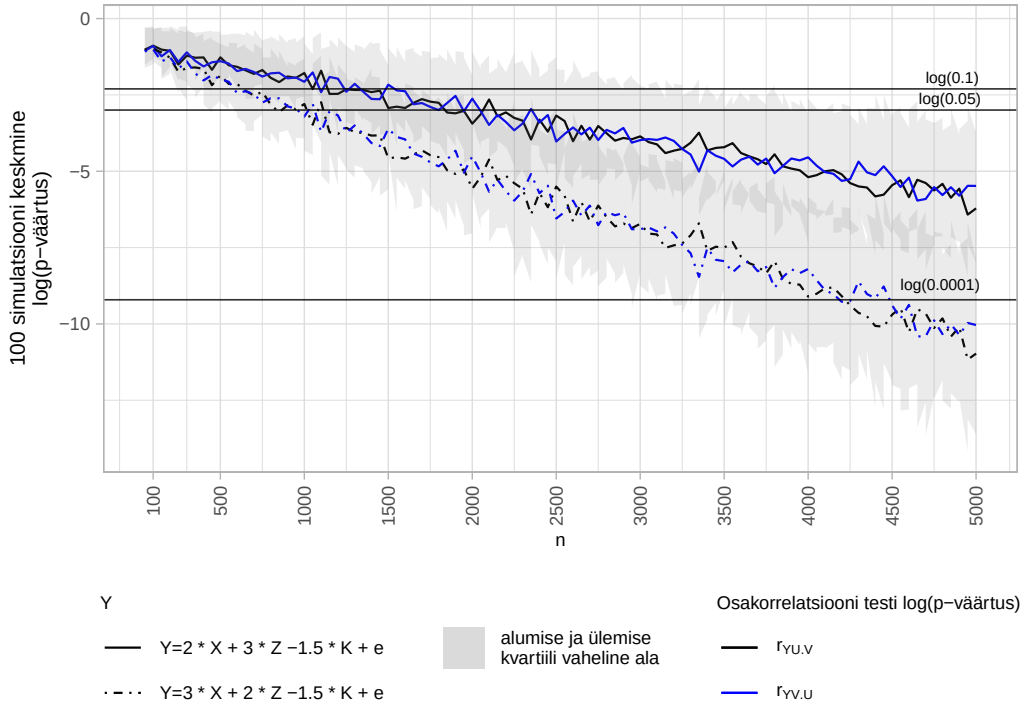
Joonis 3: Osakaal simulatsioonidest, mil tunnused U ja V loeti ekvivalentseteks erinevate olulisuse nivoode korral.

Antud näites pärinevad tunnused mitmemõõtmelisest normaaljaotusest ja sellisel juhul on tunnuste tinglik sõltumatus samaväärne 0-ga võrduva osakorrelatsiooniga (Baba, Shibata ja Sibuya, 2004). Osakorrelatsioon tunnuste Y ja U vahel tinglikustatuna tunnuse V järgi näitab nö puhast sõltuvust tunnuste Y ja U vahel, millest on tunnuse V mõju elimineeritud (Kim, 2015). Tähistame kirjeldatud osakorrelatsiooni valimis $r_{YU.V}$ ja üldkogumis $\rho_{YU.V}$. Tunnuste Y ja U sõltumatust, tinglikustatuna tunnuse V järgi, testitakse hüpoteeside

$$H_0 : \rho_{YU.V} = 0$$
$$H_1 : \rho_{YU.V} \neq 0$$

abil. Joonisel 4 on kujutatud osakorrelatsioonide $\rho_{YU.V}$ ja $\rho_{YV.U}$ testide keskmised logaritmitud p-väärtused 100 simulatsiooni põhjal. Üks neist testidest viiakse läbi

selleks, et kontrollida, kas tunnused U ja V on ekvivalentsed. Testi valik sõltub sellest, kas U või V oli tugevamalt seotud tunnusega Y, ehk kumb neist valiti algoritmi poolt esimesena välja. Kui jäädakse H_0 juurde ehk testi logaritmitud p-väärtus on suurem kui $\log(\alpha)$, loetakse tunnused ekvivalentseteks.

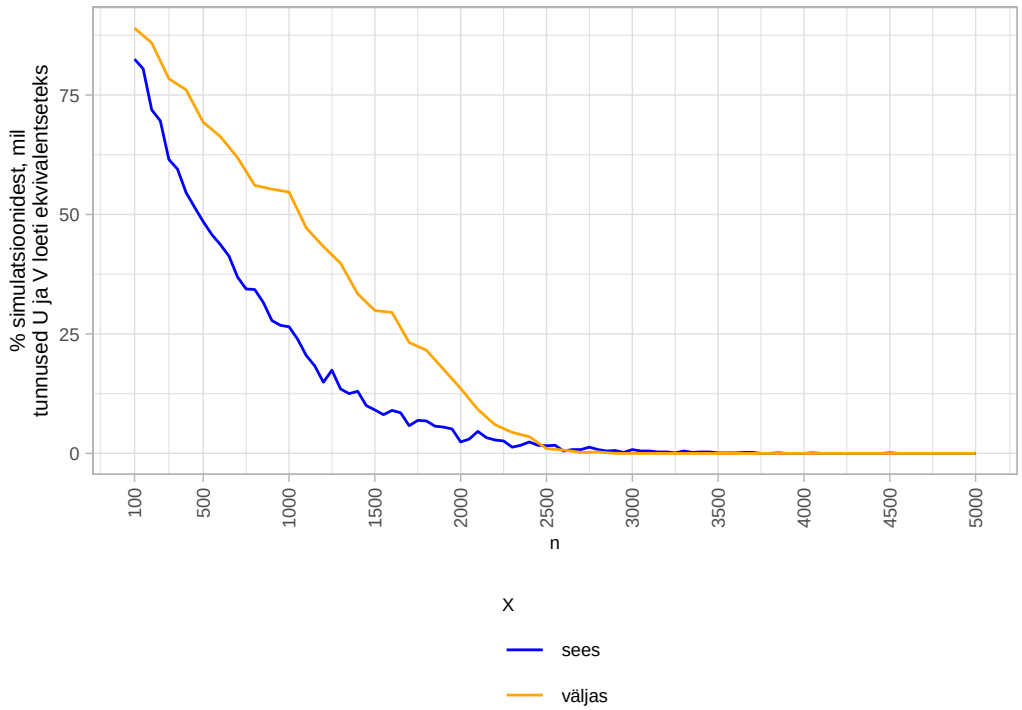


Joonis 4: Tingliku sõltumatuse testide logaritmitud p-väärtused, mille põhjal otsustatakse, kas tunnused U ja V on ekvivalentsed.

Joonisele on märgitud ka logaritmitud olulisuse nivoo väärtused α, et oleks võimalik võrrelda, millise valimi mahu korral loeb algoritm tunnused U ja V ekvivalentseteks erinevate α väärtuste korral. Katkematu joonega on märgitud tulemused esimese lineaarkombinatsiooni korral ja katkendliku joonega on märgitud tulemused teise lineaarkombinatsiooni korral. Helehalliga on tähistatud ala logaritmitud p-väärtuste alumise ja ülemise kvartiili vahel ehk 50 logaritmitud p-väärtust 100-st asus iga $n \in \{100{,}150{,}200, \cdots, 4950{,}5000\}$ korral hallis ribas. Valimi mahu kasvades langevad tingliku sõltumatuse testide p-väärtused. Seetõttu valitakse järjest harvem

X-ga tugevalt korreleeritud tunnuseid ekvivalentseteks.

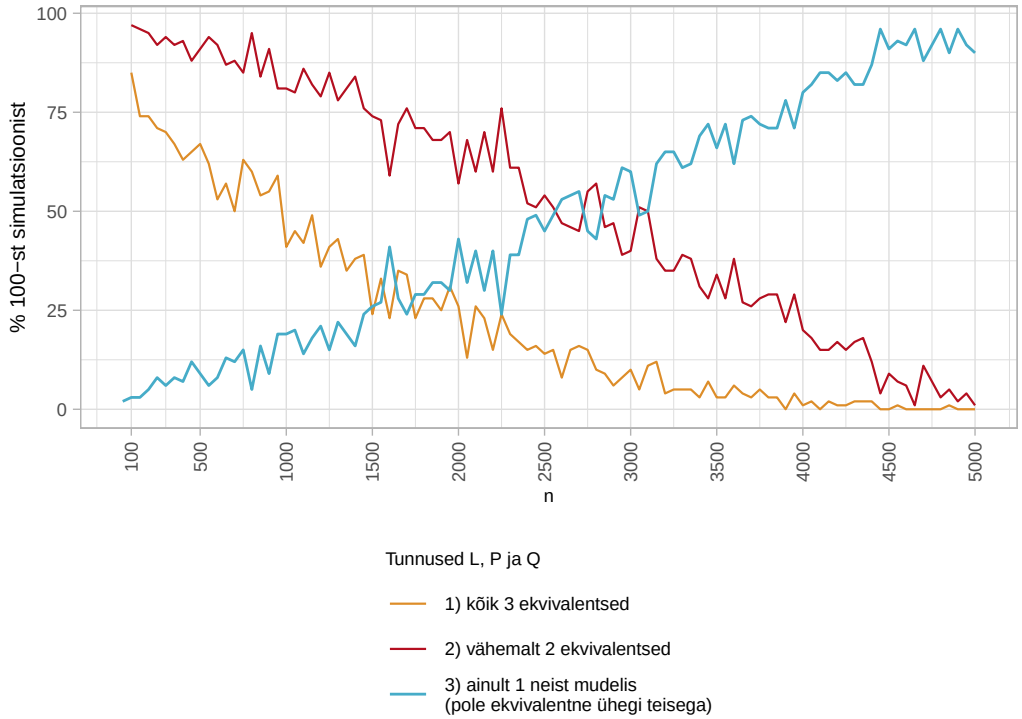
Joonisel 5 on kujutatud, kuidas mõjutab U ja V ekvivalentseteks lugemist asjaolu, kas tunnus X on SES algoritmile teada või mitte. Kujutatud on SES algoritmi tulemused olulisuse nivoo $\alpha = 0.05$ ja funktsioontunnuse $Y = 2 \cdot X + 3 \cdot Z - 1{,}5 \cdot K + \varepsilon$ korral.



Joonis 5: Osakaal simulatsioonidest, mil tunnused U ja V loeti samaväärseteks $\alpha = 0.05$ korral, olenevalt sellest, kas tunnus X oli andmestikus või mitte.

Näha on, et kui tunnus X on andmestikus olemas, langeb simulatsiooonide osakaal, kus U ja V loeti ekvivalentseteks, kiiremini.

Joonisel 6 uuritakse lähemalt tunnuseid L, P ja Q. Kujutatud on SES algoritmi tulemused olulisuse nivoo $\alpha = 0.05$ ja esimese lineaarkombinatsiooni korral.



Joonis 6: Osakaal simulatsioonidest, mil tunnused U ja V loeti ekvivalentseteks; $\alpha = 0{,}05$.

Kollane joon tähistab osakaalu simulatsioonidest, mil kõik kolm tunnust L, P ja Q loeti omavahel samaväärseteks. Punane joon tähistab osakaalu simulatsioonidest, mil vähemalt kaks tunnustest L, P ja Q loeti ekvivalentseteks. Sinise joonega on kujutatud osakaal simulatsioonidest, mil vaid üks tunnustest L, P ja Q valiti välja, kusjuures see tunnus polnud ekvivalentne ühegi teisega.

Näha on, et kuigi vähemalt kaks tunnustest L, P ja Q loetakse ekvivalentseteks sagedamini, kui kõik kolm, siis mõlemad osakaalud langevad valimi mahu kasvades sarnase kiirusega. Seevastu suureneb valimi mahu kasvades simulatsioonide osakaal, mil valiti välja vaid üks vaadeltavatest tunnustest.

5 Näide metaboliitide andmetel

Eesti geenidoonorite metaboloomika andmetel rakendatakse SES algoritmi selleks, et luua lineaarne regressioonimudel kehamassiindeksi ehk KMI (kehakaal (kg)/pikkus $(m)^2$) hindamiseks ja logistiline regressioonimudel suremise tõenäosuse hindamiseks. SES algoritmi poolt välja valitud tunnuseid ning mudelite ennustusvõimet võrreldakse samm- ning lassoregressiooni mudelitega.

5.1 Andmed

Eesti geenivaramu on Tartu Ülikooli genoomika instituudi koosseisus olev teadus- ja arendusasutus, mis on loonud Eestile biopanga. Biopanga eesmärk on koguda ja säilitada Eesti elanikkonna geeni- ja terviseandmeid teaduslikel eesmärkidel. Andmete kogumisega alustati 2002. aastal ja 2023. aastaks on biopangaga liitunud enam kui 210 000 eestimaalase ehk ligikaudu 20% Eesti elanikkonnast. (TÜ genoomika instituut)

Töös kasutati geenidoonorite vere metaboloomika andmeid. Metaboliidid on väikese molekulmassiga ained, mis tekivad organismis ainevahetuse vahe- või lõppsaadusena. Metaboloom on ülevaade kõigist bioloogilises objektis (näiteks veres) leiduvatest metaboliitidest teatud ajahetkel. Veres on enam kui 18000 metaboliidi, mis jagunevad kahte gruppi: rasvlahustuvad ja vesilahustuvad. Üks peamine meetod metaboliitide tuvastamiseks on tuumamagnetresonantsspektroskoopia (inglise keeles *nuclear magnetic resonance spectroscopy*, NMR), mis põhineb vesiniku ja süsiniku isotoopide ^{1}H ja ^{13}C tuumade magneetilistel omadustel. (Kiseleva *et al.*, 2021)

Geenidoonorite kohta on teada sugu, sünniaasta, geenivaramuga liitumise kuupäev, vereproovi andmise kuupäev ja 238 metaboliidi kontsentratsioonide ning metaboliitide kontsentratsioonide suhte väärtused. Juhul, kui geenidoonor on surnud, on teada tema surmakuupäev. Valimisse kaasati doonorid, kes olid vereproovi andmise

ajal 40- kuni 89-aastased. Nende arv on 107 522. Kehamassiindeksi hindamiseks jäeti valimist välja doonorid, kelle kehamassiindeks oli puudu. Lisaks eemaldati isikud, kelle kehamassiindeks oli väiksem kui 15 või suurem kui 60. Valimi maht kehamassiindeksi hindamiseks on 100 292. Teave doonorite vanuselise ja soolise koosseisu kohta mõlemas valimis on toodud tabelis 2.

Tabel 2: Doonorite vanuseline ja sooline jaotus kahes valimis.

	KMI valim			Suremise tõenäosuse valim		
Sugu/ vanusegrupp	Mees (%)	Naine (%)	**Kokku (%)**	Mees (%)	Naine (%)	**Kokku (%)**
40-49	11 892 (11,9)	23 757 (23,7)	**35 649 (35,5)**	12 960 (12,0)	25 193 (23,4)	**38 153 (35,5)**
50-59	9557 (9,5)	20 748 (20,7)	**30 305 (30,2)**	10 329 (9,6)	21 965 (20,4)	**32 294 (30,0)**
60-69	6606 (6,6)	14821 (14,8)	**21 427 (21,4)**	7090 (6,6)	15 844 (14,7)	**22 934 (21,3)**
70-79	3272 (3,3)	7270 (7,2)	**10 542 (10,5)**	3515 (3,3)	7969 (7,4)	**11 484 (10,7)**
80+	723 (0,7)	1646 (1,6)	**2369 (2,4)**	797 (0,7)	1860 (1,7)	**2657 (2,5)**
Kokku	**32050 (32,0)**	**68 242 (68,0)**	**100 292 (100)**	**34 691 (32,3)**	**72 831 (67,7)**	**107 522 (100)**

Geenidoonorid võib ühinemisaasta järgi jagada kahte suuremasse gruppi. Kuni 2011. aastani liituti peamiselt perearstikeskuste kaudu ja aastatel 2018-2019 kampaania „Kingime Eestile juubeliks 100 000 uut geenidoonorit“ raames. Kuna suremus on nendes gruppides erinev, siis on potentsiaalsete argumenttunnuste hulka kaasatud ka binaarne tunnus, mille väärtus on 0, kui doonor ühines geenivaramuga enne 2018. aastat, ja 1, kui liituti 2018. aastal või hiljem. Valimis on enne 2018. aastat liitunute arv 22 788 ja 2018. aastal või hiljem liitunute arv 84 734.

Tabelis 3 on toodud 5 aasta jooksul pärast vereproovi andmist surnud meeste arv igas vanusegrupis vastavalt sellele, kas geenidoonor liitus geenivaramuga enne 2018. aastat või pärast. Tabelis 4 on toodud sama teave naiste kohta. 2018. aastal või

hiljem ühinenute seas on kokku $507 + 452 = 959$ doonorit, kes surid 5 aasta jooksul pärast vereproovi andmist, mis moodustab 40,3% kõigist 5 aasta jooksul surnud isikute arvust $1211 + 1171 = 2382$. 2018. aastal või hiljem liitujate osakaal on valimis 21,2%.

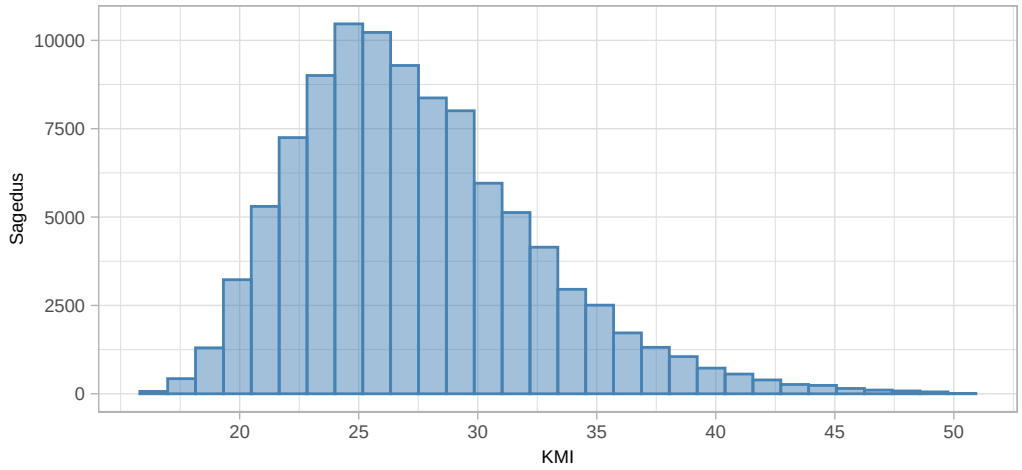
Tabel 3: 5 aasta jooksul pärast vereproovi andmist surnud meeste arv vanusegrupi ja ühinemisaasta järgi.

	Mehed		
Ühinemise aasta/ vanusegrupp	enne 2018. a (%)	2018 või hiljem (%)	**Kokku (%)**
40-49	57 (4,7)	44 (3,6)	**101 (8,3)**
50-59	123 (10,2)	99 (8,2)	**222 (18,3)**
60-69	194 (16,0)	134 (11,1)	**328 (27,1)**
70-79	221 (18,2)	149 (12,3)	**370 (30,6)**
80+	109 (9,0)	81 (6,7)	**190 (15,7)**
Kokku	**704 (58,1)**	**507 (41,9)**	**1211 (100)**

Tabel 4: 5 aasta jooksul pärast vereproovi andmist surnud naiste arv vanusegrupi ja ühinemisaasta järgi.

	Naised		
Ühinemise aasta/ vanusegrupp	enne 2018. a (%)	2018 või hiljem (%)	**Kokku (%)**
40-49	43 (3,7)	32 (2,7)	**75 (6,4)**
50-59	86 (7,3)	71 (6,1)	**157 (13,4)**
60-69	137 (11,7)	112 (9,6)	**249 (21,3)**
70-79	255 (21,8)	145 (12,4)	**400 (34,2)**
80+	198 (16,9)	92 (7,9)	**290 (24,8)**
Kokku	**719 (61,4)**	**452 (38,6)**	**1171 (100)**

Kehamassiindeksi histogramm valimis ($n = 100\ 292$) on toodud joonisel 7.



Joonis 7: Kehamassiindeksi histogramm valimis ($n = 100\ 292$).

Naiste keskmine kehamassiindeks on 27,2 ja standardhälve 5,30 (kg/m^2). Meeste keskmine kehamassiindeks on 27,9 ja standardhälve 4,38 (kg/m^2).

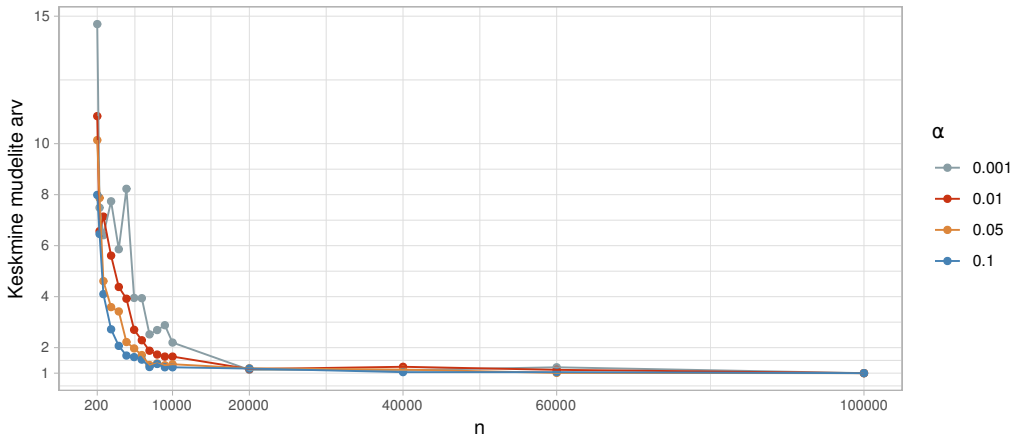
5.2 Kehamassiindeksi mudelid

SES algoritmi rakendati kehamassiindeksi hindamiseks moodustatud valimis erineva suuruse osavalimite peal ja uuriti, kui palju ekvivalentseid argumenttunnuste kogumeid algoritm leiab. Seejärel hinnati mudeleid erinevate argumenttunnuste valiku meetodite abil ning võrreldi tulemusi SES algoritmi poolt väljapakutud mudelitega. Mudelite ennustusvõimet hinnati andmete peal, mis jäid üle pärast mudelite ehitamiseks kasutatud osavalimite eemaldamist.

5.2.1 Ekvivalentsete mudelite arv

Kehamassiindeksi hindamiseks kokkupandud valimist, kuhu kuulub $n = 100\ 292$ isikut, võeti iga valimi mahu $n \in \{200, 500, 1000, 2000, 3000, \cdots, 9000, 10000, 20000, 40000, 60000, 100000\}$ korral tagasipanekuta osavalimeid. Igal osavalimil rakendati SES algoritmi erinevate olulisuse nivoode $\alpha \in \{0{,}1; 0{,}05; 0{,}01; 0{,}001\}$ korral ja osavalimeid võeti iga n puhul 100 korda. Seejärel leiti, mitu ekvivalentset mudelit

SES algoritm leidis. Joonisel 16 on kujutatud 100 osavalimi keskmine samaväärsete mudelite arv iga osavalimi suuruse n korral. Erineva värviga on tähistatud tulemused erinevate α väärtuste korral.



Joonis 8: Ekvivalentsete KMI mudelite keskmine arv sõltuvalt olulisuse nivoost α ja valimi mahust n.

Valimi mahu n kasvades kahaneb ekvivalentsete mudelite arv kiiresti. Alates valimi mahust $n = 20\,000$ on keskmine samaväärsete mudelite arv kõigi α väärtuste korral 1 või lähedal sellele. Kui $n \leq 10\,000$, on näha, et väiksema olulisuse nivoo korral leitakse keskmiselt rohkem samaväärseid mudeleid. Kui $n = 200$, leitakse $\alpha = 0{,}001$ korral keskmiselt ligikaudu 15 samaväärset mudelit. Mudelite arvu aritmeetilist keskmist mõjutavad üksikud suured erindid. Lisades (vt Lisa 1. Ekvivalentsete mudelite arv) on täpsemalt kujutatud samaväärsete mudelite arvude varieeruvust karpdiagrammi abil iga osavalimi suuruse n korral. Kuna sagedamini leiti 1 kuni 2 samaväärset mudelit ja esines ükskuid väga suuri erindeid, on karpdiagrammil ekvivalentsete mudelite arv log-skaalal.

5.2.2 SES algoritmi tulemuste võrdlus teiste meetoditega

Kogu valimist suurusega 100 292 võeti juhuslikult tagasipanekuta valikuga osavalimeid suurusega $n \in \{500; 1000; 5000\}$. Osavalimeid võeti iga n korral 50. Kõigile

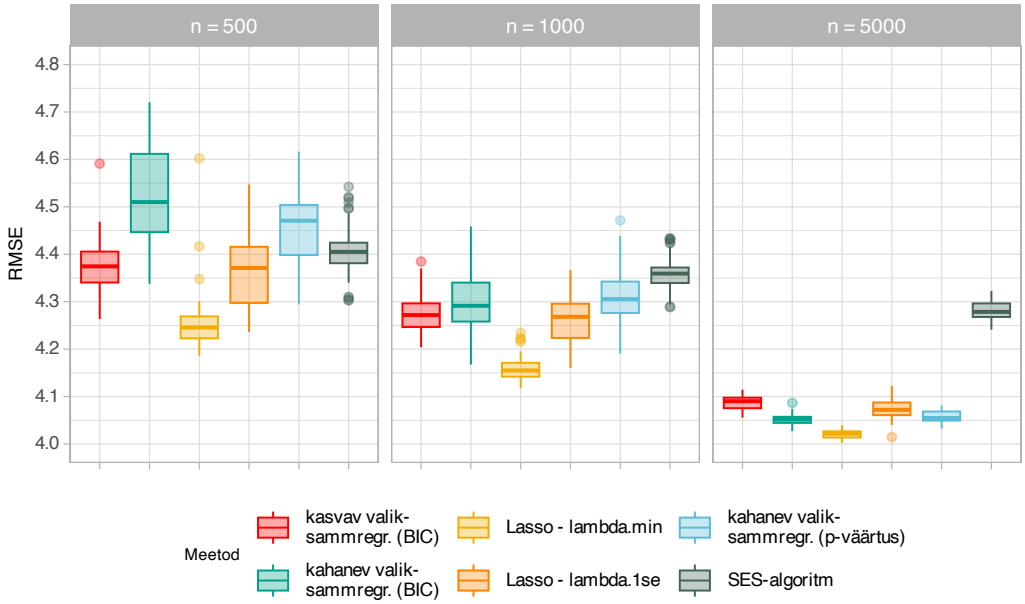
osavalimitele rakendati SES algoritmi, sammregressiooni kasvava ja kahaneva valikuga BIC väärtuse põhjal, lassoregressiooni ($\lambda = \hat{\lambda}_{min}$ ja ka $\lambda = \hat{\lambda}_{1se}$) ja kahaneva valikuga sammregressiooni p-väärtuste põhjal. Mitmese testimisega arvestamiseks korrigeeriti olulisuse nivood. Metaboliitide andmed on tugevalt korreleeritud, näiteks kuuluvad 91 tunnust gruppi, kus paarikaupa korrelatsioonid on kõik kõrgemad kui 0,95. Valemi 4 järgi leiti kogu valimi pealt M_{eff} väärtuseks 36 ja valemi 5 järgi saadi M_{eff} väärtuseks 21. Valiti üks vahepealne väärtus ja korrigeeritud olulisuse nivooks võeti 0,05/25.

Ristvalideerimist kasutati selleks, et valida SES algoritmile välja optimaalne α väärtus hulgast $\{0{,}001; 0{,}01; 0{,}05; 0{,}1\}$ ja optimaalne k väärtus hulgast $\{2; 3; 4; 5\}$. Valituks osutunud hüperparameetrite arvud on toodud tabelis 5. Kõigi osavalimite suuruste korral on kõige enam valitud välja suurim olulisuse nivoo väärtus 0,1. Parameeter k on enamasti olnud kas 2 või 5.

Tabel 5: Ristvalideerimisel leitud optimaalsed parameetrite α ja k väärtused osavalimites.

	Parameetrid							
	α				k			
	0,001	0,01	0,05	0,1	2	3	4	5
$n = 5000$	1	3	9	37	18	7	9	16
$n = 1000$	2	9	9	30	22	11	4	13
$n = 500$	2	10	7	31	24	10	0	16

Pärast sobivaimate hüperparameetrite leidmist ning SES algotritmi poolt ekvivalentsete argumenttunnuste gruppide tuvastamist hinnati osavalimites regressioonimudelid. Hinnatud mudeleid kasutades prognoositi ülejäänud 100 292 $- n$ isiku kehamassiindeksid. Kasutades prognoose ja tegelikke kehamassiindekseid, leiti RMSE väärtus, mille põhjal mudelite prognoosivõimet võrreldi. Joonisel 9 on kujutatud RMSE väärtuste karpdiagramm osavalimite suuruste kaupa, kus erinevate värvidega on tähistatud argumenttunnuste valiku meetodeid.

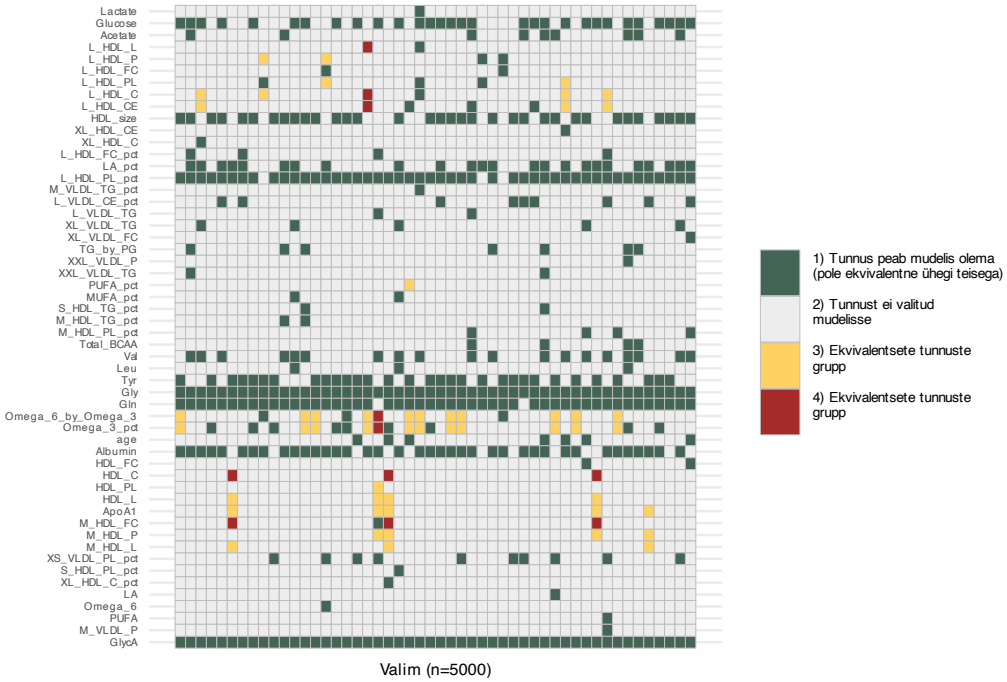


Joonis 9: KMI mudelite võrdlus erinevate tunnuse valiku meetodite vahel RMSE väärtuse põhjal.

Kõige väiksema valimi mahu $n = 500$ korral on SES-mudeli prognoosid olnud veidi täpsemad, kui kahaneva valikuga sammregressiooni meetodite korral. Kui $n = 500$, paikneb SES algoritmi mudelite RMSE väärtuste mediaan ja ka kvartiilid teiste meetodite mediaanide vahel. Kui valimi maht suureneb, siis SES algoritm hakkab küll täpsemaid prognoose tegema enda eelmiste hinnangutega võrreldes, aga teiste algoritmide puhul muutuvad hinnangud kiiremini täpsemaks. Kui $n = 5000$, on SES algoritmi mudelite minimaalne RMSE väärtus märgatavalt suurem, kui teiste meetodite maksimaalne RMSE väärtus. Lisades (vt Lisa 2. Meetodite võrdlus valimite lõikes) on toodud detailsem meetodite võrdlus iga osavalimi korral. (Kui SES algoritm on leidnud vähemalt 2 ekvivalentset mudelit, on nende arv märgitud joonisele.)

5.3 Valitud tunnused

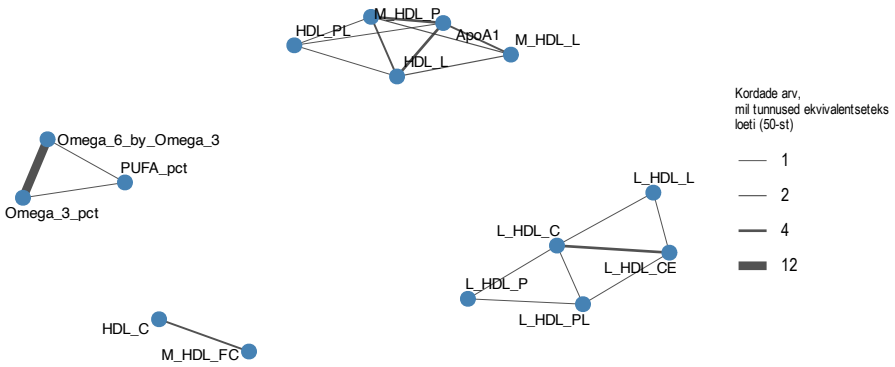
Joonisel 18 on SES algoritmi tunnuste valik kujutatud igas osavalimis suurusega $n = 5000$. Veergudes asuvad 50 osavalimit ja ridades on kõikvõimalikud tunnused, mida algoritm välja valis. Tumerohelisega on tähistatud tunnused, mis on valitud mudelisse, kusjuures neid pole peetud ekvivalentseks ühegi teisega. Helehalliga on märgitud tunnus siis, kui teda välja ei valitud. Kollase ja punasega on kujutatud samaväärseteks loetud tunnused. Näiteks, vasakult esimeses tulbas on kujutatud tunnuste valik esimese osavalimi korral. Lõplikusse mudelisse tuleb kaasata glükoosi ja albumiini tase veres, HDL osakeste suurus ja tunnused L_HDL_PL_pct, Tyr, Gly, Gln, GlycA. Tunnused nimedega Omega_6_by_Omega_3, Omega_3_pct on loetud ekvivalentseteks ehk lõplikusse mudelisse tuleb valida üks neist.



Joonis 10: SES algoritmi tunnuste valik KMI hindamiseks 50 valimi ($n = 5000$) põhjal.

Tumerohelised triibud on joonisele moodustatud tunnuste poolt, mis peaaegu alati lõplikusse mudelisse kaasata tuleb (GlycA, albumiin, Gln, Gly, L_HDL_PL_pct).

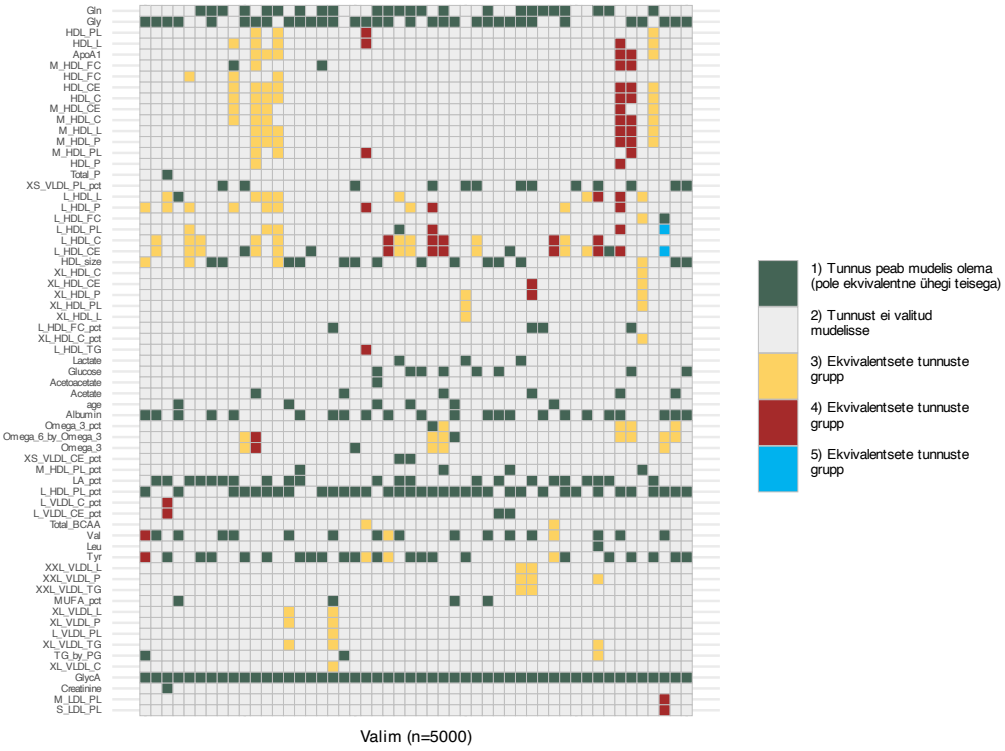
Joonisel 11 on kujutatud vaid neid tunnuseid, mis loeti vähemalt üks kord ekvivalentseks mõne teisega, ehk kujutatud on tunnused, mis joonisel 10 on vähemalt üks kord värvitud kollast või punast värvi. Joonisel samaväärseteks loetud tunnused on ühendatud erineva paksusega joontega vastavalt sellele, mitu korda (50 mudeli hulgast) neid samaväärseteks loeti.



Joonis 11: Samaväärsete tunnuste võrgustik leitud 50 valimi ($n = 5000$) põhjal.

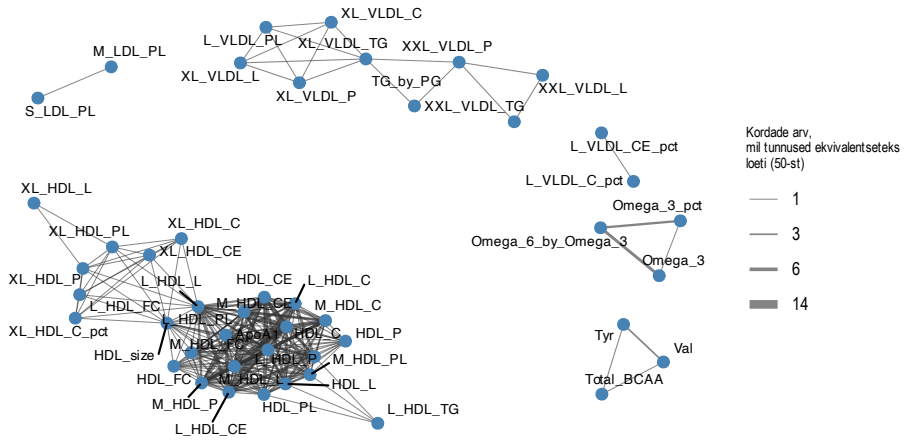
HDL-kolesterooliga seotud tunnuseid on omavahel samaväärseteks peetud. Antud juhul moodustavad need kolm eraldiseisvat gruppi. Kõige enam on ekvivalentseteks peetud tunnuseid Omega_6_by_Omega_3, Omega_3_pct.

Joonisel 12 on kujutatud tunnuste valik valimi mahu $n = 1000$ korral ning on näha, et väiksema n korral peetakse sagedamini tunnuseid omavahel ekvivalentseteks. Kuigi GlycA on jätkuvalt igasse mudelisse valitud, paistab nüüd vähem selliseid tunnuseid, mida järjepidevalt mudelisse kaasatud oleks.



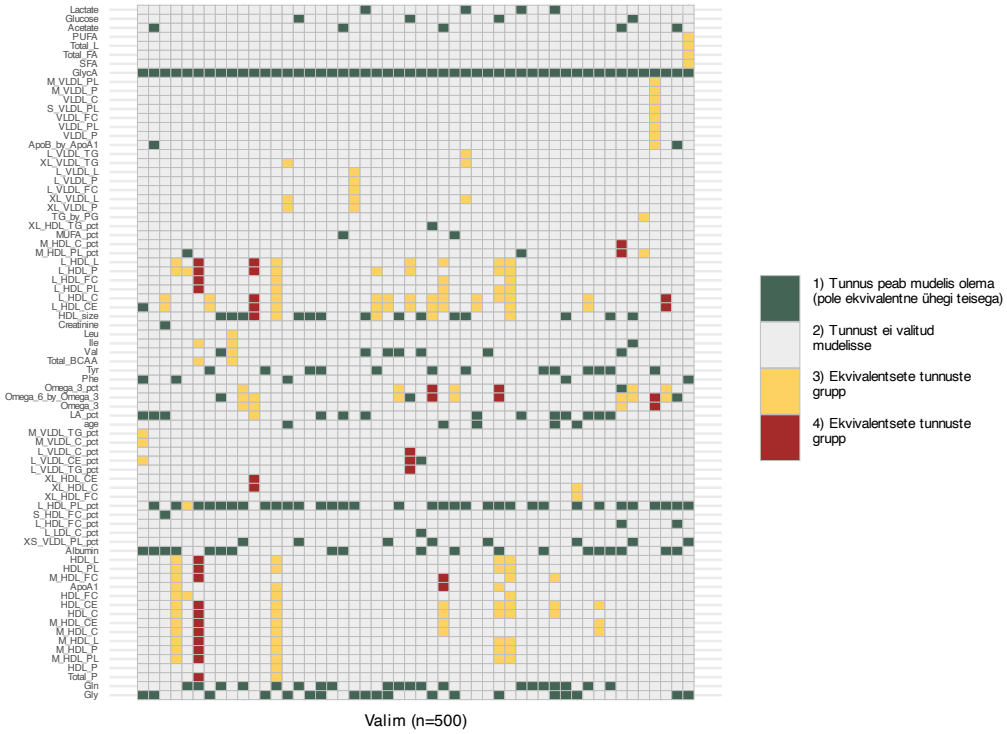
Joonis 12: SES algoritmi tunnuste valik KMI hindamiseks 50 valimi ($n = 1000$) põhjal.

Ekvivalentsete tunnuste võrgustik joonisel 13 on nüüd kirjum kui enne. Tekkinud on kaks peamist gruppi: HDL-kolesterooliga seotud tunnused ja VLDL-kolesterooliga seotud tunnused.



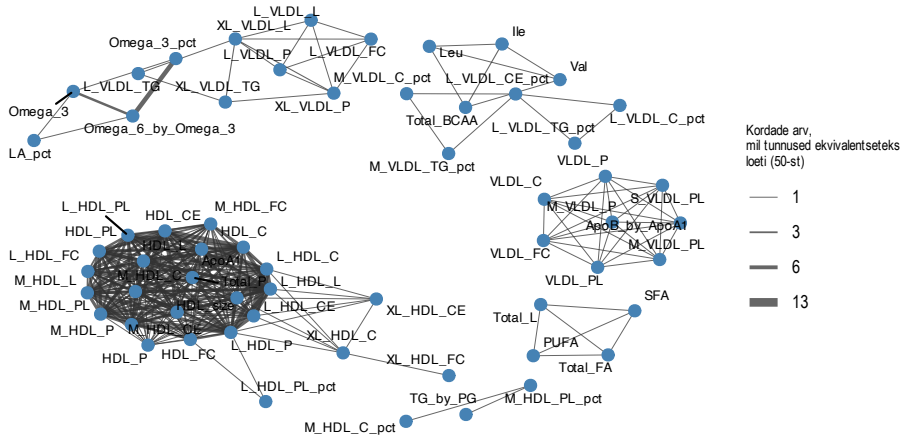
Joonis 13: Samaväärsete tunnuste võrgustik leitud 50 valimi ($n = 1000$) põhjal.

Joonisel 14 on toodud väljavalitud tunnused osavalimites suurusega $n = 500$. Kui $n = 5000$ korral oli tunnuste arv, mis vähemalt ühte mudelisse valiti (ehk tunnuste arv y-teljel), 54, siis valimi mahu kahanedes on tunnuste arv tõusnud, $n = 1000$ korral 65 ja $n = 500$ korral 77.



Joonis 14: SES algoritmi tunnuste valik KMI hindamiseks 50 valimi ($n = 500$) põhjal.

Samaväärsete tunnuste võrgustiku joonisel 15 on tunnuste arv võrreldes $n = 1000$ kasvanud. HDL-kolesterooliga seotud tunnuseid loetakse sagedamini ekvivalentseks kui suuremate valimi mahtude puhul. Ning jällegi on ülejäänud kokku grupeeritud tunnused seotud VLDL-kolesterooliga.



Joonis 15: Samaväärsete tunnuste võrgustik leitud 50 valimi ($n = 500$) põhjal.

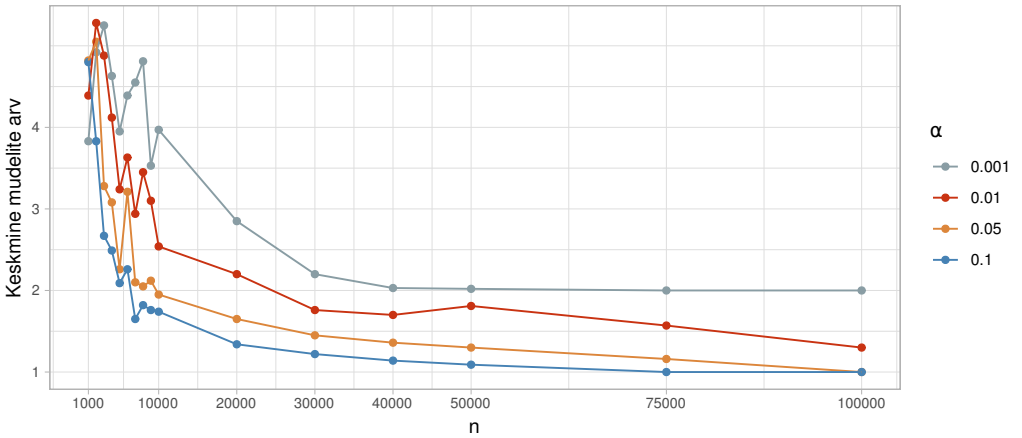
5.4 Surma tõenäosuse mudelid

Sarnaselt kehamassiindeksile hinnati SES algoritmi abil mudelid surma tõenäosusele erineva suurusega valimitel. Uuriti, palju ekvivalentseid mudeleid leitakse, milliseid tunnuseid välja valitakse ja kui täpseid prognoose teevad SES algoritmi poolt väljapakutud mudelid võrreldes samm- ja lassoregressiooni meetoditega.

5.4.1 Ekvivalentsete mudelite arv

Kasutades tervet valimit, kuhu kuulub $n = 107\,522$ isikut, võeti tagasipanekuta valikuga osavalimeid suurustega $n \in \{1000, 2000, \cdots, 9000, 10000, 20000,$
$40000, 40000, 50000, 75000, 100000\}$. Igal osavalimil rakendati SES algoritmi erinevate olulisuse nivoode $\alpha \in \{0{,}1; 0{,}05; 0{,}01; 0{,}001\}$ korral ja osavalimeid võeti iga n puhul 100 korda. Igal osavalimil leiti, mitu ekvivalentset mudelit SES algoritm tagastas. Joonisel 16 on kujutatud 100 osavalimi keskmine samaväärsete mudelite arv iga osavalimi suuruse n korral. Erineva värviga on tähistatud tulemused erinevate α väärtuste korral. Kuna logistilise regressioonimudeli hindamisel on oluline,

et juhtude arv ehk isikute arv klassis $Y = 1$ liiga väike ei oleks, siis algoritmi ei rakendatud osavalimitel väiksema suurusega kui 1000. Arvestades, et kogu valimist moodustab surmade arv ligikaudu 2,2%, siis osavalimisse suurusega 1000 satub keskmiselt 22 isikut, kes kuuluvad $Y = 1$ klassi.



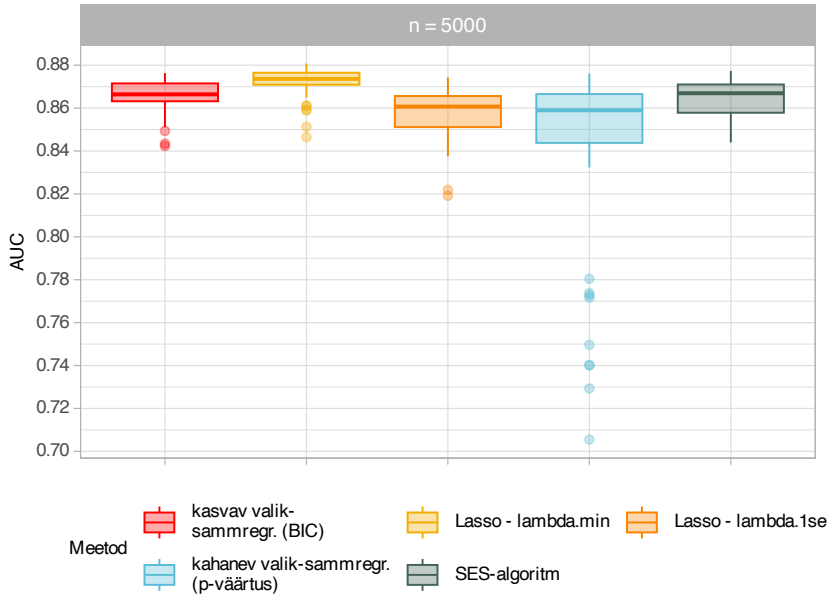
Joonis 16: Keskmine ekvivalentsete logistilise regressiooni mudelite arv sõltuvalt olulisuse nivoost α ja valimi mahust n.

Väiksema olulisuse nivooga on keskmine samaväärsete mudelite arv suurem ja osavalimi n kasvades keskmine samaväärsete mudelite arv kahaneb. Olulisuse nivoo $\alpha = 0{,}001$ korral püsib keskmine ekvivalentsete mudelite arv alates osavalimi mahust $n = 30\ 000$ ligikaudu 2 juures. Lisades (vt Lisa 1. Ekvivalentsete mudelite arv) on kujutatud samaväärsete mudelite arvude varieeruvust karpdiagrammi abil iga osavalimi suuruse n korral.

5.4.2 SES algoritmi tulemuste võrdlus teiste meetoditega

Kogu valimist suurusega 107 522 võeti juhuslikult tagasipanekuta valikuga osavalimeid suurusega $n = 5000$. Osavalimeid võeti kokku 50. Juhtude ehk surmade arv osavalimites oli keskmiselt 110,4 ja minimaalselt 85. Kõigile osavalimitele rakendati SES algoritmi, mille hüperparameetrid leiti ristvalideerimisel. 30 korda valiti α väärtuseks 0,1, väiksemaid väärtuseid (0,001; 0,01 ja 0,05) valiti harvem. 15

korda valiti k väärtuseks 5 ja 23 korda 2, vahepealseid väärtuseid (3 ja 4) valiti kokku 12 korda. Samuti rakendati osavalimitel sammregressiooni kasvava valikuga BIC väärtuse põhjal (kahaneva valikuga sammregressiooni meetodil BIC väärtuse põhjal tekkis mitme osavalimi puhul koondumisprobleeme ning see meetod jäeti võrdlusest välja), lassoregressiooni ($\lambda = \hat{\lambda}_{min}$ ja ka $\lambda = \hat{\lambda}_{1se}$) ja kahaneva valikuga sammregressiooni p-väärtuste põhjal (tunnused loeti statistiliselt oluliseks kui p-väärtus oli madalam kui 0,05/25). Mudelid hinnati osavalimitel ja hinnatud mudelite abil prognoositi ülejäänud $107\ 522 - 5000 = 102\ 522$ isiku 5 aasta jooksul suremise tõenäosused, mille abil arvutati AUC väärtused. Joonisel 17 on kujutatud AUC väärtuste karpdiagramm, kus erinevat värvi karbid tähistavad erinevaid argumenttunnuste valiku meetodeid.



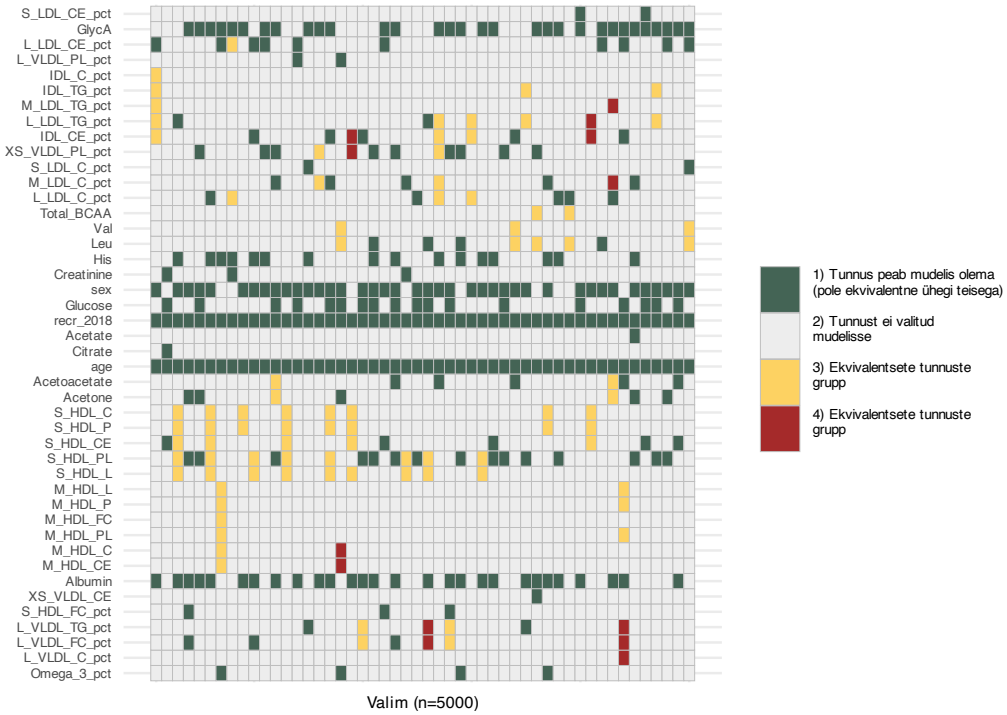
Joonis 17: Logistilise regressiooni mudelite võrdlus erinevate tunnuse valiku meetodite vahel AUC väärtuse põhjal; mudelid hinnatud valimitel suurusega $n = 5000$.

Meetodid prognoosivad AUC väärtuse põhjal 5 aasta suremust sarnaselt, sest karbid asetsevad kohakuti. Üksikud väikesed AUC erindid sammregressioonimeetodite ja ka lassoregressiooni ($\lambda = \hat{\lambda}_{1se}$) korral näitavad, et mõnel osavalimil hinnatud

mudelid on testandmestikul halvemaid prognoose teinud. SES-mudelitel kehvemad erindid puuduvad. Näha on, et lassoregressiooni $\lambda = \hat{\lambda}_{min}$ AUC mediaan on veidi kõrgem SES algoritmi mudelite AUC mediaanist. Detailsem ülevaade meetodite tulemustest igas osavalimis on toodud lisades (vt Lisa 2. Meetodite võrdlus valimite lõikes).

5.5 Valitud tunnused

Erinevaid tunnuseid, mida SES algoritm kõigi 50 osavalimi peale kokku valis, oli 57. Joonisel 18 on kujutatud algoritmi tunnuste valik igas osavalimis. Veergudes asuvad 50 osavalimit ja ridades on kõikvõimalikud tunnused, mida algoritm välja valis.

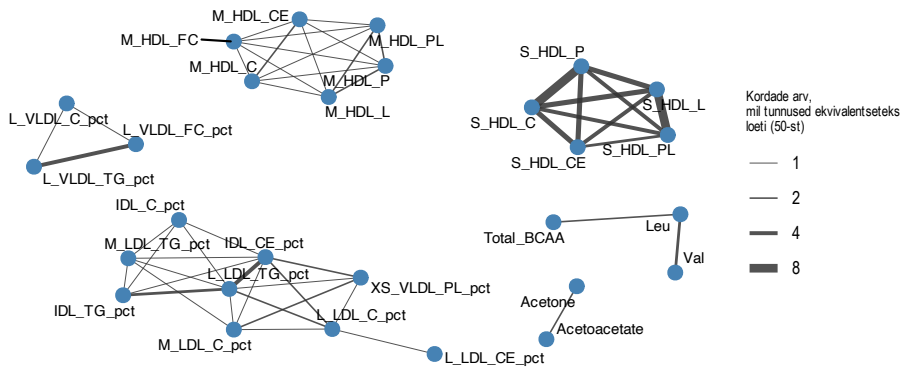


Joonis 18: SES algoritmi tunnuste valik 5 aasta jooksul suremise tõenäosuse hindamiseks 50 valimi ($n = 5000$) põhjal.

Vanus vereproovil ja enne või pärast aastat 2018 ühinemist märkiv tunnus on lõp-

likusse mudelisse vaja kaasata kõigi osavalimite korral. Sugu on mudelisse vaja kaasata 40 valimi korral. Glükoosi tase, albumiin ja tunnus nimega GlycA on mudelitesse kaasatud ligikaudu poolte osavalimite korral.

Joonisel 19 on kujutatud samaväärsete tunnuste võrgustik.



Joonis 19: Samaväärsete tunnuste võrgustik leitud 50 valimi ($n = 5000$) põhjal.

Moodustunud on metaboliitide grupid: esimeses on HDL-kolesterooliga seotud tunnused algusega M_HDL, teises HDL-kolesterooliga seotud tunnused algusega S_HDL, kolmandas on VLDL-kolesterooliga seotud tunnused algusega L_VLDL, kõige suuremasse gruppi on koondatud veel IDL-, LDL- ja VLDL-kolesterooliga seotud tunnuseid. HDL-kolesterooliga seotud tunnuseid algusega S_HDL loeti samaväärseteks kõige enam.

Kokkuvõte

SES algoritm püüab korduvalt tingliku sõltumatus teste rakendades leida, milliseid argumenttunnuseid tuleks mudelisse kaasata. Seejuures väljastatakse ekvivalentsete tunnuste hulgad, kust igast ühest täpselt ühe tunnuse valides on tulemuseks mudelid, mis sobivad andmetele sarnaselt.

Töö eesmärk oli testida SES algoritmi simuleeritud ja reaalsetel andmetel. Algoritmi rakendati genereeritud andmetel, kuhu kuulusid sõltuva muutuja Y moodustanud tunnustega väga tugevalt korreleeritud tunnused (korrelatsioonikordaja suurem kui 0,99). Selgus, et väiksema valimi korral leitakse samaväärsed tunnused sagedamini üles. Valimi mahu kasvades hakkasid ka väga tugevalt korreleeritud tunnused rohkem eristuma ja mudelisse kaasati vaid see tunnus, mis juhuslikult funktsioontunnusega rohkem seotud oli.

Sarnane seaduspärasus tuli esile ka Eesti geenidoonorite andmetel. Kasutati vere metaboliitide andmeid, kus esinevad samuti väga kõrged korrelatsioonid tunnuste vahel. Hinnati lineaarseid regressioonimudeleid kehamassiindeksile ja logistilisi regressioonimudeleid suremise tõenäosusele 5 aasta jooksul pärast vereproovi andmist. Erineva suurusega osavalimeid võttes oli näha, et valimi mahu kasvades väheneb ekvivalentsete mudelite arv.

SES algoritmi poolt väljapakutud kehamassiindeksi mudelid tegid väiksema valimi mahu $n = 500$ korral sarnase täpsusega prognoose teiste argumenttunnuste valiku meetoditega (sammregressioon BIC väärtuse ja p-väärtuste põhjal) võrreldes. Valimi mahu kasvades muutusid aga teiste meetoditega leitud mudelite hinnangud kiiremini täpsemaks ja $n = 5000$ korral oli SES algoritmi mudelite RMSE väärtus märgatavalt kehvem teistest. 5 aasta jooksul suremuse prognoosimisel oli SES algoritmiga leitud mudelid AUC väärtuste poolest sarnased teiste meetoditega. Kusjuures leiti palju ekvivalentseid tunnuseid: LDL-kolesterooliga seotud tunnused ja VLDL-kolesterooliga seotud tunnused.

Kui suurema valimi mahu korral annavad lassoregressiooni meetodil leitud mudelid märgatavalt täpsemaid prognoose, siis väiksema valimi mahu korral pakuvad SES algoritmi mudelid teiste meetoditega sarnast ennustusvõimet. Lisaks annavad SES-mudelid võimaluse valida mitme argumenttunnuste osahulga vahel.

Kasutatud allikad

Kunihiro Baba, Ritei Shibata ja Masaaki Sibuya (2004). "Partial Correlation and Conditional Correlation as Measures of Conditional Independence". *Australian & New Zealand Journal of Statistics* 46.4, lk. 657–664. DOI: https://doi.org/10.1111/j.1467-842X.2004.00360.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-842X.2004.00360.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.2004.00360.x.

Tom Fawcett (2006). "An introduction to ROC analysis". *Pattern Recognition Letters* 27.8. ROC Analysis in Pattern Recognition, lk. 861–874. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2005.10.010. URL: https://www.sciencedirect.com/science/article/pii/S016786550500303X.

Jerome Friedman, Robert Tibshirani ja Trevor Hastie (2010). "Regularization Paths for Generalized Linear Models via Coordinate Descent". *Journal of Statistical Software* 33.1, lk. 1–22. DOI: 10.18637/jss.v033.i01.

Trevor Hastie, Robert Tibshirani ja Jerome Friedman (2009). *The elements of statistical learning: data mining, inference and prediction*. 2. väljaanne. Springer. URL: http://www-stat.stanford.edu/~tibs/ElemStatLearn/.

Rob J. Hyndman ja Anne B. Koehler (2006). "Another look at measures of forecast accuracy". *International Journal of Forecasting* 22.4, lk. 679–688. ISSN: 0169-2070. DOI: https://doi.org/10.1016/j.ijforecast.2006.03.001. URL: https://www.sciencedirect.com/science/article/pii/S0169207006000239.

Gareth James, Daniela Witten, Trevor Hastie ja Robert Tibshirani (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. URL: https://faculty.marshall.usc.edu/gareth-james/ISL/.

Seongho Kim (2015). “ppcor: An R Package for a Fast Calculation to Semi-partial Correlation Coefficients”. *Communications for statistical applications and methods* 22.6, 665–674. DOI: https://doi.org/10.5351/CSAM.2015.22.6.665.

Olga Kiseleva, Ilya Kurbatov, Ekaterina Ilgisonis ja Ekaterina Poverennaya (2021). “Defining Blood Plasma and Serum Metabolome by GC-MS”. *Metabolites* 12(1), lk. 15. DOI: https://doi.org/10.3390/metabo12010015.

Vincenzo Lagani, Giorgos Athineou, Alessio Farcomeni, Michail Tsagris ja Ioannis Tsamardinos (2017). “Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets”. *Journal of Statistical Software* 80.7, 1–25.

J. Li ja L. Ji (2005). “Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix”. *Heredity* 95, lk. 221–227. DOI: https://doi.org/10.1038/sj.hdy.6800717.

S. J. Mason ja N. E. Graham (2002). “Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation”. *Quarterly Journal of the Royal Meteorological Society* 128.584, lk. 2145–2166. DOI: https://doi.org/10.1256/003590002320603584. eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1256/003590002320603584. URL: https://rmets.onlinelibrary.wiley.com/doi/abs/10.1256/003590002320603584.

P McCullagh (1989). “Generalized Linear Models (2nd ed.)” DOI: https://doi.org/10.1201/9780203753736.

A. Peluso, R. Glen ja T.M.D. Ebbels (2021). “Multiple-testing correction in metabolome-wide association studies”. *BMC Bioinformatics* 22, lk. 67. DOI: https://doi.org/10.1186/s12859-021-03975-2.

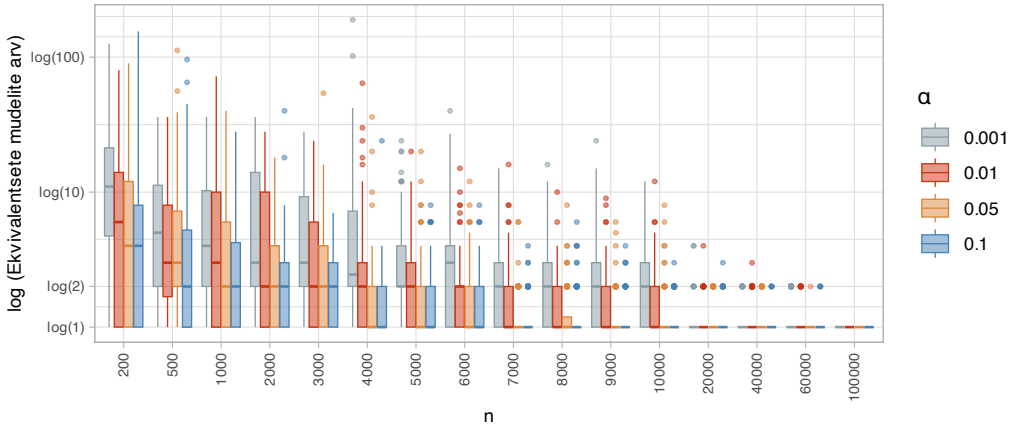
Robert Tibshirani (1996). “Regression Shrinkage and Selection via the Lasso”. *Journal of the Royal Statistical Society. Series B (Methodological)* 58.1, lk. 267–288. ISSN: 00359246. URL: http://www.jstor.org/stable/2346178 (vaadatud 17.05.2024).

TÜ genoomika instituut (kuupäev puudub). *Eesti geenivaramu*. URL: https://genomics.ut.ee/et/sisu/genoomika-instituudist (vaadatud 25.10.2021).

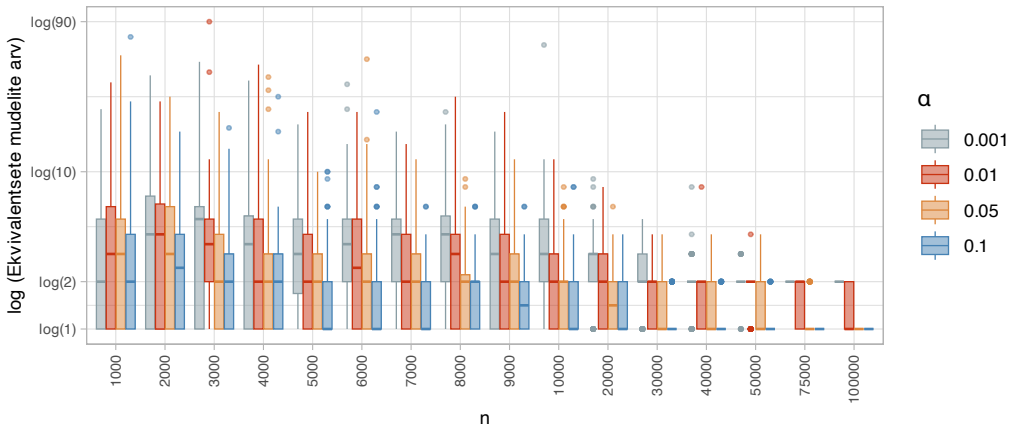
W. N. Venables ja B. D. Ripley (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: https://www.stats.ox.ac.uk/pub/MASS4/.

Ernst Wit, Edwin van den Heuvel ja Jan-Willem Romeijn (2012). “‘All models are wrong...’: an introduction to model uncertainty”. *Statistica Neerlandica* 66.3, lk. 217–236. DOI: https://doi.org/10.1111/j.1467-9574.2012.00530.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9574.2012.00530.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9574.2012.00530.x.

Lisa 1. Ekvivalentsete mudelite arv

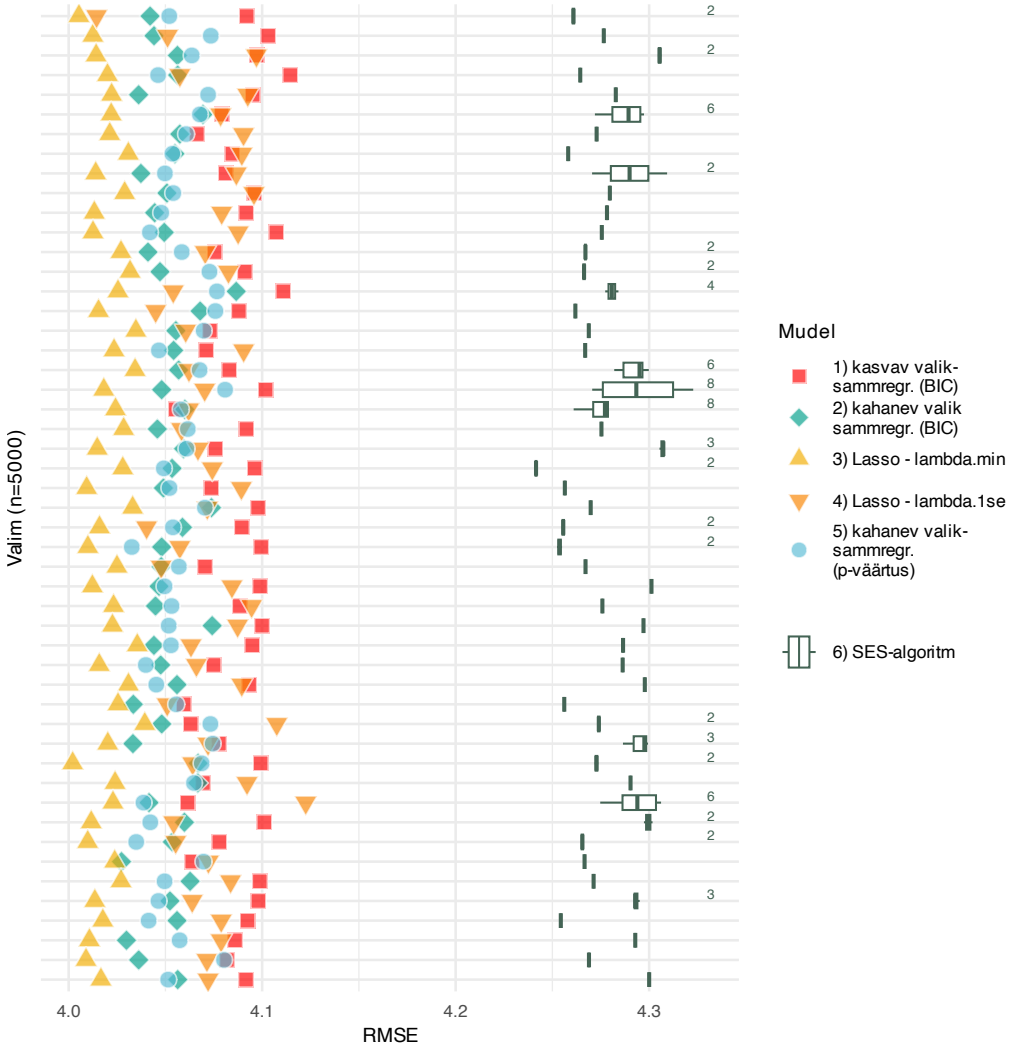


Joonis 20: Ekvivalentsete mudelite arv log-skaalal sõltuvalt olulisuse nivoost α ja valimi mahust n. Mudelid prognoosivad kehamassiindeksit.

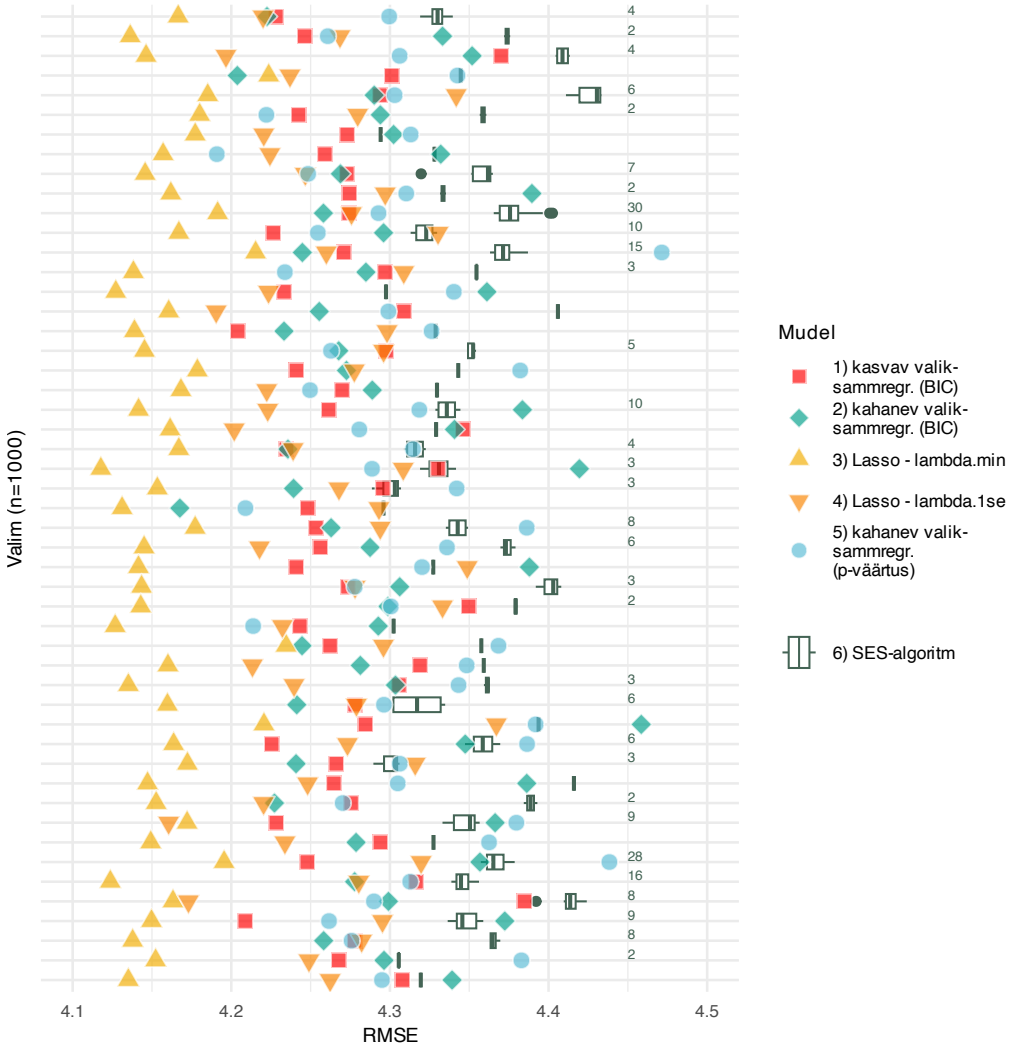


Joonis 21: Ekvivalentsete mudelite arv log-skaalal sõltuvalt olulisuse nivoost α ja valimi mahust n. Mudelid prognoosivad surma tõenäosust.

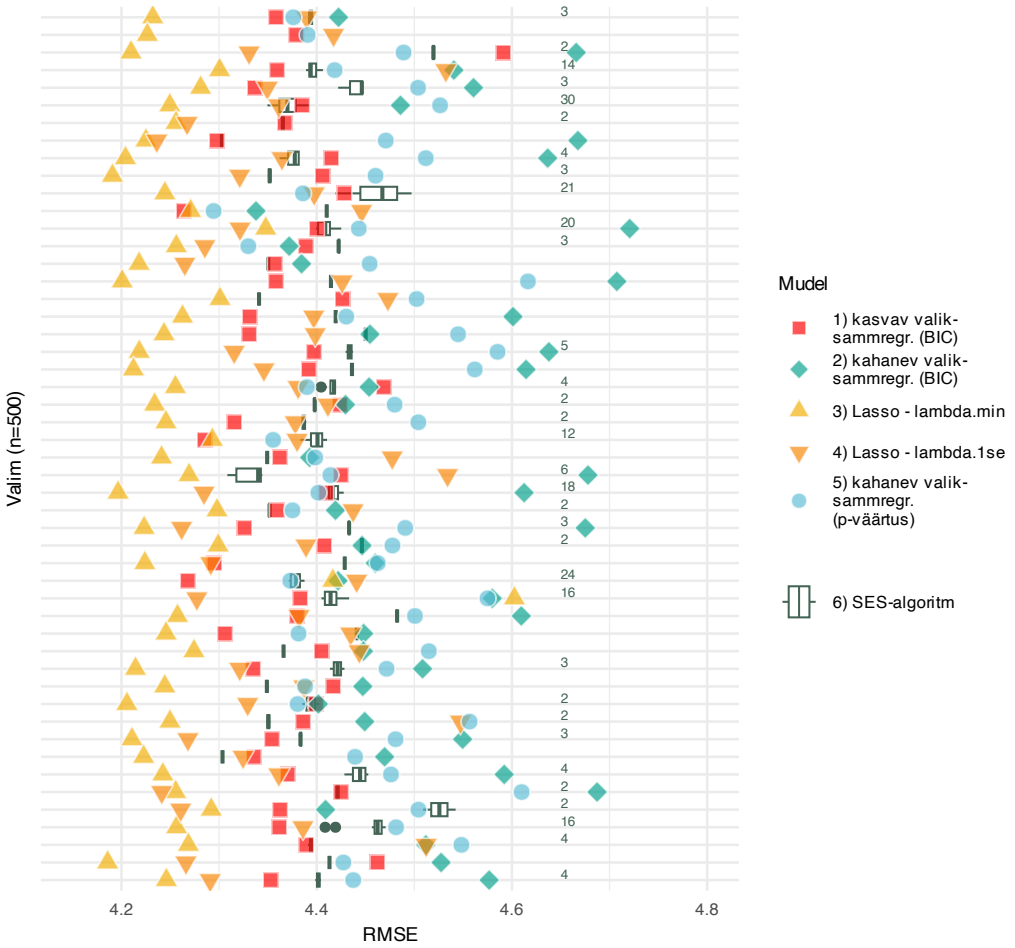
Lisa 2. Meetodite võrdlus valimite lõikes



Joonis 22: KMI mudelite võrdlus erinevate tunnuse valiku meetodite vahel RMSE väärtuse põhjal; mudelid hinnatud valimitel suurusega $n = 5000$; kui SES algoritm valis välja 2 või enam ekvivalentset mudelit, on nende arv märgitud tumerohelise kirjaga joonisele.



Joonis 23: KMI mudelite võrdlus erinevate tunnuse valiku meetodite vahel RMSE väärtuse põhjal; mudelid hinnatud valimitel suurusega $n = 1000$; kui SES algoritm valis välja 2 või enam ekvivalentset mudelit, on nende arv märgitud tumerohelise kirjaga joonisele.



Joonis 24: KMI mudelite võrdlus erinevate tunnuse valiku meetodite vahel RMSE väärtuse põhjal; mudelid hinnatud valimitel suurusega $n = 500$; kui SES algoritm valis välja 2 või enam ekvivalentset mudelit, on nende arv märgitud tumerohelise kirjaga joonisele.

Joonis 25: Logistilise regressiooni mudelite võrdlus erinevate tunnuse valiku meetodite vahel AUC väärtuse põhjal; mudelid hinnatud valimitel suurusega $n = 5000$; kui SES algoritm valis välja 2 või enam ekvivalentset mudelit, on nende arv märgitud tumerohelise kirjaga joonisele.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Hanna Sõnajalg,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Statistiliselt ekvivalentsete argumenttunnuste kogumite leidmine“, mille juhendajad on Krista Fischer ja Oliver Aasmets, 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 4.0, 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.

Hanna Sõnajalg
22.05.2024