Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Stein-Marten Pool

Statistilised meetodid bioloogilise vanuse hindamiseks

Matemaatiline statistika

Bakalaureusetöö (9 EAP)

Juhendaja: Prof. Krista Fischer

Tartu 2023

STATISTILISED MEETODID BIOLOOGILISE VANUSE HINDAMISEKS

Bakalaureusetöö

Stein-Marten Pool

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärgiks on leida statistiline meetod, mille abil hinnata TÜ Eesti Geenivaramu andmestikus olevatele isikutele bioloogiline vanus. Töös on kasutatud üle 200 000 geenidoonori andmeid, kelle kohta on andmestikus üle 250 bioloogilise, füsioloogilise ja muu tunnuse. Esmalt käsitletakse töös laialt kasutatavat mitmese lineaarse regressioonanalüüsi mudelit. Töös tuuakse välja antud mudeli abil bioloogilise vanuse hindamisel kaasnevad negatiivsed aspektid ning näidatakse neid ka nii simuleeritud, kui Geenivaramu andmetel. Töö teises pooles käsitletakse elukestusanalüüsi meetodeid ning avaldatakse bioloogilise vanuse arvutamise valem, kasutades vanust, sugu ja biomarkeritest kokku pandud NMR-skoori. Edasi hinnatakse juhuslikule valimile bioloogiline vanus elukestusanalüüsi meetodil saadud valemiga ning analüüsitakse tulemusi.

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

Märksõnad: Bioloogiline vanus, elukestusanalüüs, matemaatiline statistika, geenidoonorid.

STATISTICAL METHODS FOR BIOLOGICAL AGE ESTIMATION

Bachelor thesis

Stein-Marten Pool

Abstract

The aim of this bachelor's thesis is to find a statistical method for estimating the biological age of individuals in the dataset of the University of Tartu Estonian Biobank. The study uses data from over 200,000 gene donors, with more than 250 biological, physiological, and other characteristics in the dataset. Firstly, the widely used multiple linear regression analysis model is discussed. The negative aspects associated with the estimation of biological age using this model are presented, and demonstrated using both simulated and Biobank data. In the second part of the thesis, survival analysis methods are discussed, and a formula for estimating biological age is derived using age, gender, and a score compiled from NMR-biomarkers. The biological age is then estimated for a random sample using the formula obtained by the survival analysis method, and the results are visualised.

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

Keywords: Biological age, survival analysis, mathematical statistics, gene donors.

Sisukord

Sissejuhatus

Käesoleva bakalaureusetöö eesmärk on leida statistiline meetod, mille abil hinnata TÜ Eesti Geenivaramu andmestikus olevatele isikutele bioloogiline vanus. Eesti Geenivaramu andmebaasis on andmed enam kui 200 000 geenidoonori kohta, kes on liitunud geenivaramuga aastatel 2002-2022. Andmestik sisaldab infot isikute bioloogiliste, füsioloogiliste ning teiste näitajate kohta.

Bioloogiliseks vanuseks nimetatakse isiku kronoloogilisest vanusest ehk elatud aastatest eraldiseisvat suurust, mis sõltub isiku füsioloogilistest ja bioloogilistest näitajatest. Näiteks, kronoloogilisest vanusest suurem bioloogiline vanus tähendab, et isiku tervisenäitajad on sarnasemad temast vanemate isikute keskmiste tervisenäitajatega. Bioloogilise vanuse abil saab anda üldist tagasisidet isiku tervisliku seisundi kohta. Samuti võib bioloogilisest vanusest teha järeldusi, kas isiku oodatav eluiga on keskmisest oodatavast elueast suurem või väiksem.

Töö esimeses pooles tuuakse bioloogilise vanuse hindamisel esmalt välja lineaarne regressioonanalüüsi mudel, mis on lihtsasti kasutatav ning bioloogilise vanuse arvutamisel levinud meetod. Seejärel selgitatakse, millised probleemid võivad tekkida sel moel saadud hinnangute tõlgendamisel ja põhjendatakse, miks tasuks antud mudeli asemel kasutada teistsugust lähenemist.

Töö teises pooles kasutatakse bioloogilise vanuse arvutamiseks elukestusanalüüsi meetodeid ning tuletatakse hinnang bioloogilisele vanusele Weibulli jaotuse eeldusel. Saadud meetodit rakendatakse TÜ Eesti Geenivaramu andmetel, hindamaks geenidoonorite bioloogilist vanust tuumamagnetresonants-meetodil (TMR) leitud biomarkeri-profiili põhjal.

1 Andmestiku kirjeldus

Käesoleva töö keskmes olev andmestik sisaldab geenidoonorite andmeid, kes on Tartu Ülikooli Eesti Geenivaramuga liitunud aastatel 2002-2022. Andmestik sisaldab 207 480 vereproovi andmeid, millest osa on samade inimeste korduvad proovid. Antud töös kasutatakse vaid iga isiku kõige varasema vereproovi andmeid. Kuna andmestikus leidub ka isikuid, kelle vanus proovide andmise hetkel oli alla 18, siis need vaatlused eemaldati Geenivaramu reeglite tõttu. Andmestikku jäi alles 205 023 isikut, kelle kohta on andmestikus isikut tähistavad tunnused, isiku kohta käivad tunnused, mida saab koguda lihtsate mõõtmiste või küsitluste teel ning lisaks veel ka erinevad biomarkerid.

Andmestikus olevad biomarkerite väärtused on määratud Soome ettevõtte Nightingale Health poolt, tuumamagnetresonants-spektroskoopia meetodil (Nightingale Health, 2020). Erinevaid biomarkereid on andmestikus 249, mille seas on näiteks nii glükoosi, albumiini ja Omega-6 tase, kui ka palju teisi biomarkereid (väga suur osa iseloomustab verelipiidide profiili). Isiku kohta on andmestikus veel ka sünniaasta, proovide andmise kuupäev, sugu, kehamassiindeks, haridustase, suitsetamist näitav tunnus ning surmaga seotud tunnused.

Andmestikus olevad surmaga seotud andmed on 09.01.2022 seisuga ja need on saadud geenidoonorite andmebaasi linkimisel Eesti surma põhjuste registri andmetega. Surmasid on andmestikus 7903, mis jagunesid üle kogu uurimisperioodi ning surnute vanused varieerusid 19-106 vahel, kus keskmine suremise vanus oli 72 ning mediaan 75. Üldine vanuseline jaotus oli proovide andmise hetkel 18-105 vahel, kus kõigi isikute keskmine vanus oli 50 ja mediaanvanus 48.

Naiste ja meeste osakaalud andmestikus olid vastavalt 65,5% ja 34,5%. Uuringu tegemise hetkel suitsetasid 20,5% isikutest ning 23,1% isikutest olid endised suitsetajad. Kõige sagedasemad kõrgeimad omandatud haridustasemed olid „Kutseõpe keskhariduse baasil" ja „Magister või sellega võrdsustatud haridus", mis moodustasid vastavalt 25,8% ja 25,2% kogu andmestikust.

2 Mitmene lineaarne regressioonanalüüsi mudel

Oma lihtsuse tõttu on mitmene lineaarne regressioonanalüüsi mudel bioloogilise vanuse arvutamisel üks enim kasutatavaid mudeleid. Mitmese lineaarse regressioonanalüüsi mudelit (MLR) kasutades leitakse treeningandmestiku pealt mudel, mis kõige väiksema veaga hindab isiku tervise ja muude näitajate põhjal isiku kronoloogilist vanust. Seega kasutatakse isiku kronoloogilist vanust antud mudelis sõltuva tunnusena. Selle mudeli prognoosi nimetatakse bioloogiliseks vanuseks. (Krut'ko *et al.*, 2000)

Mitmese lineaarse regressioonanalüüsi mudeli kuju indiviidi bioloogilise vanuse leidmiseks on

$$BA_i = b_0 + \sum_{j=1}^{n} b_j X_{ji}, \tag{1}$$

kus BA_i on indiviidi bioloogiline vanus, $b_0, b_1, b_2, ..., b_n$ on mudelist saadavad koefitsiendid ning $X_{1i}, X_{2i}, ..., X_{ni}$ vastavad indiviidi mudelis olevate tunnuste väärtustele. Seega see mudel eeldab, et

$$BA_i = E(CA_i | X_1, ..., X_n),$$

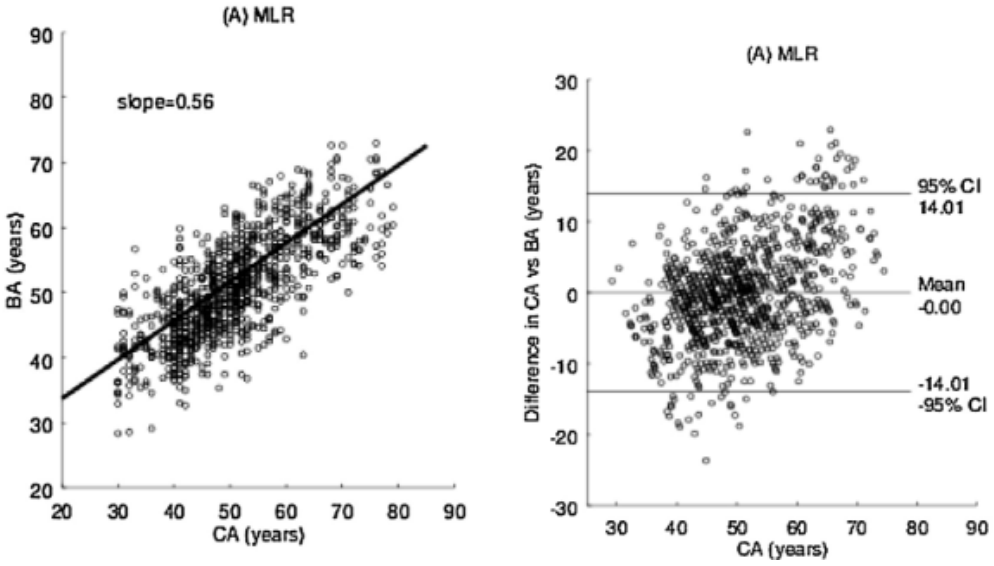
mis omakorda avaldub teatud argumenttunnuste lineaarse kombinatsioonina. Järelikult bioloogiline vanus

$$BA_i = prognoositud\ CA_i,$$

kus CA_i tähistab indiviidi kronoloogilist vanust. (Cho, Park ja Lim, 2010)

2.1 MLR kasutamise negatiivsed aspektid

Kuigi tegu on lihtsa ja laialt kasutatava mudeliga, leidub tavapäraselt antud mudeli rakendamisel saadud tulemustes aspekte, mille tõttu MLR-i kasutamine bioloogilise vanuse arvutamiseks ei pruugi olla kuigi hea variant. Bioloogilise vanuse prognoosimisel MLR-ga jäävad bioloogilise vanuse hinnangud vanematel inimestel sageli kronoloogilisest vanusest madalamaks ning noorematel inimestel vastupidiselt kõrgemaks. (Li *et al.*, 2023)



Joonis 1: Kronoloogilise vanuse ja bioloogilise vanuse graafik regressioonisirgega (vasakul) ning kronoloogilise vanuse ja vea ε_i graafik (paremal).

Seda tendentsi võib näha ka Jee ja Parki uuringus, kus joonisel 1 nähtavatel graafikutel on regressioonisirge silmnähtavalt erinev $BA = CA$ graafikust ning samuti on ka näha, et viga ε_i, mis on defineeritud kui

$$\varepsilon_i = CA_i - BA_i, \tag{2}$$

on vanematel inimestel suurel määral positiivne ning noorematel inimestel enam-

jaolt negatiivne. (Jee ja Park, 2017)

Antud probleemi saab näidata ka teoreetiliselt. Nimelt, MLR definitsioonist saab, et valemis (2) defineeritud vea saab kirjutada kujul $\varepsilon_i \sim \mathcal{N}(0; \sigma)$. Edasi saab valemist (2) tuletada

$$CA_i = BA_i + \varepsilon_i,$$

mille omakorda saab kirjutada kui

$$CA_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i \ ; \ \sigma).$$

Tulemust mõjutamata võib oletada, et $\beta_1 > 0$ ning $x_{min} = \min(X_i)$ ehk

$$(CA_i | X_i = x_{min}) \sim \mathcal{N}(\beta_0 + \beta_1 x_{min} \ ; \ \sigma),$$

millest saab tuletada, et

$$P(CA_i < \beta_0 + \beta_1 x_{min} | X_i = x_{min}) = 0.5.$$

Viimasest võrdusest saab järeldada, et isikutel, kelle kronoloogiline vanus on võrdne kohordi minimaalse kronoloogilise vanusega, on MLR-ga arvutatav bioloogiline vanus alati kõrgem, kui kronoloogiline vanus. Sarnaselt saab ka tõestada, et isikutel, kelle kronoloogiline vanus on võrdne kohordi maksimaalse kronoloogilise vanusega, on MLR-ga arvutatav bioloogiline vanus alati väiksem, kui kronoloogiline vanus.

Peale ebaühtlase vigade jaotuse võivad ka MLR-st saadavad koefitsiendid olla reaalses elus ebaloogilised. Näiteks võib kohati märgata tendentsi, et suitsetamise tunnuse koefitsient MLR-s on negatiivne. See tähendab, et kahe samade näitajatega inimese, kellest üks suitsetab ja teine mitte, bioloogiline vanus erineb nii, et suitsetaja bioloogiline vanus on väiksem.

2.2 Simulatsioon

Eelnevas alapeatükis kirjeldatud tendentside kirjeldamiseks loodi töö raames simulatsioon, kus genereeriti fiktiivsed isikud koos vanusega ning mõningate tunnustega. Simulatsioon viidi läbi kasutades rakendustarkvara R.

Esmalt loodi andmestik suurusega 4000, kus iga rida tähistab ühe fiktiivse isikuga seotud infot. Igale isikule määrati reaalarvuline kronoloogiline vanus, mis genereeriti jaotusest $CA \sim \mathcal{N}(45;\, 10)$, kus CA tähistab kronoloogilist vanust. Edasi määrati igale isikule binaarne tunnus „suits" näitamaks, kas isik suitsetab või mitte. Kuni 40aastastel määrati suitsetaja olemise tõenäosuseks 50%, 40-50aastastel 30% ning üle 50aastastel 10%. Keskmine vanus simulatsioonis on suitsetajatel 40,08 ning mittesuitsetajatel 47,16. Suitsetajaid on andmestikus kokku 1181 ehk $29,5\%$. Peale suitsetamise ja vanuse loodi andmestikku ka tunnused X_1 ja X_2 (n-ö „biomarkerid"), mis defineeriti kui

$$X_1 = 2 + 0.15CA + \varepsilon_A,$$
$$X_2 = 4 + suits + \varepsilon_B,$$

kus $\varepsilon_A \sim \mathcal{N}(0;\, 1)$ ja $\varepsilon_B \sim \mathcal{N}(0;\, 0,75)$.

Järgmiseks jagati andmestik kaheks, test- ja treeningandmestikuks. Testandmestikku kuulus 1000 isikut ehk 25% algsest andmestikust ja treeningandmestikku 3000 isikut ehk 75% algsest andmestikust. Seejärel arvutati treeningandmestikus isikute kronoloogilise vanuse prognoos ehk bioloogiline vanus. Selleks kasutati MLR-i, kus sõltuva tunnusena kasutati kronoloogilist vanust ning sõltumatute tunnustena X_1 ja X_2 ehk $CA \sim X_1 + X_2$.

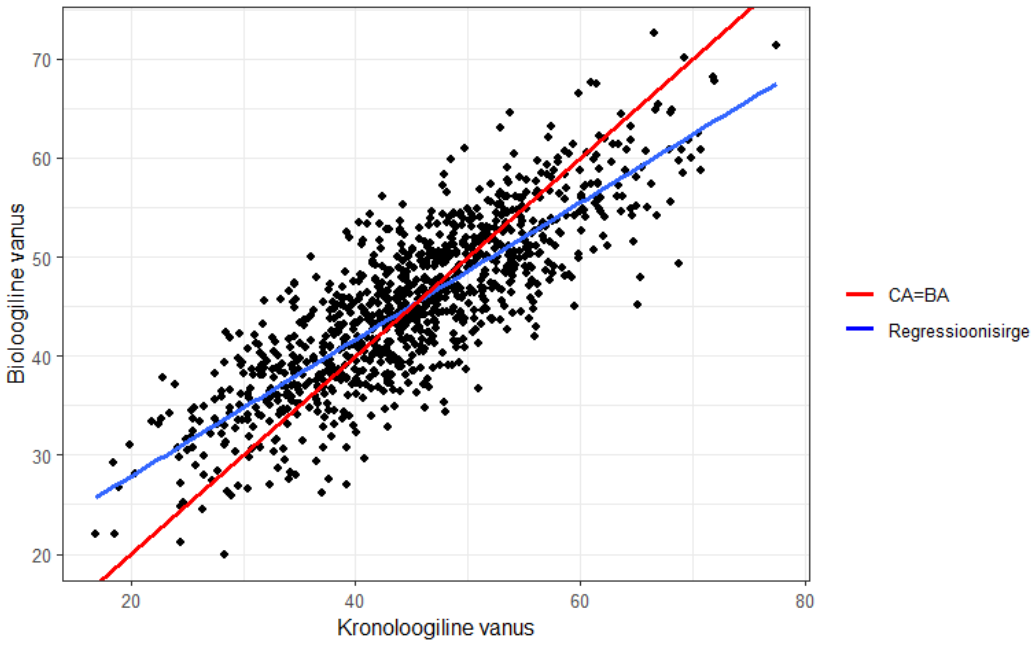
Mudelis osutusid nii X_1, kui X_2 mõju statistiliselt oluliseks ning mõlema tunnuse p-väärtus oli väiksem kui 0,001. Mudel ise oli samuti statistilisel oluline ($p < 0{,}001$) ning mitmene determinatsioonikordaja R^2 oli ligikaudu 0,7 ehk 70%. Seega võib

öelda, et X_1 ja X_2 kirjeldasid 70% kronoloogilise vanuse varieeruvusest. Mudeli kuju oli

$$BA = prognoositud\ CA = 8,4 + 4,48X_1 - 0,59X_2,$$

millest võib tähele panna, et mudelist saadud X_2 kordaja oli negatiivne. Teades, et X_2 tunnuse väärtus on suurem siis, kui isik suitsetab, siis sellest järeldub, et suitsetale prognoositakse keskmiselt väiksemat bioloogilist vanust, kui samade ülejäänud tunnustega mittesuitsetajale.

Testandmestikule bioloogilist vanust arvutades võib samuti märgata, et regressioonisirge tõus on väiksem kui 1 (vt joonis 2). See kinnitab väidet peatükist 2.1, et vanematel inimestel on bioloogiline vanus hinnatud kronoloogilisest vanusest enamjaolt väiksemaks ning noorematel inimestel seevastu suuremaks.



Joonis 2: Testandmestiku BA ja CA graafik, kus üks punkt tähistab ühte isikut.

2.3 MLR kasutamine Geenivaramu andmetel

Eelnevalt välja toodud MLR-i kasutamise negatiivsete aspektide tõttu bioloogilise vanuse arvutamisel, ei olnud põhjust antud töö eesmärki arvesse võttes leida parimat võimalikku MLR-i bioloogilise vanuse arvutamiseks. Efektiivsemaid bioloogilise vanuse arvutamise meetodeid kasutatakse järgmises peatükis, kuid siiski tasub eelnevalt mainitud negatiivsete aspektide olemasolu tõestada ka Geenivaramu andmetel. Simulatsioonis tehtud eeldus, et nooremad inimesed suitsetavad rohkem, kehtib ka Geenivaramu andmetel, kus alla 50-aastaste seas oli uuringus osalemise ajal suitsetajate osakaal 24% ning vähemalt 50-aastaste seas 15%.

Tendentside tõestamiseks loodi Geenivaramu andmestikust treening- ja testandmestikud, kus mõlemasse valiti juhuslikult pool algsest andmestikust. Seejärel kasutati treeningandmestiku peal MLR-i, mille valemi kuju oli

$$Vanus\,liitudes \sim suitsetamine + ApoB,$$

kus uuriti liitumisel olnud vanuse sõltuvust suitsetamise (1-suitsetaja, 0-mittesuitsetaja) ja ühe biomarkeri, praegusel juhul *Apolipoprotein B* (*ApoB*), tasemega. Mõlema tunnuse p-väärtuse puhul kehtis $p < 0{,}001$ ehk nii suitsetamine, kui ka valitud biomarker, olid mudelis statistiliselt olulised. Mudelist saadud suitsetamise tunnuse koefitsient on $-6{,}1$, mis tähendab, et sama *ApoB* tasemega inimeste seas on suitsetaja bioloogiline vanus 6,1 aasta võrra väiksem, kui mittesuitsetajal. Samuti oli märgata ka tendentsi, et nooremate inimeste bioloogiline vanus oli enamjaolt kõrgem, kui nende kronoloogiline vanus ning vanemate inimeste bioloogiline vanus suures osas madalam, kui nende kronoloogiline vanus.

3 Elukestusanalüüs

3.1 Statistiline metoodika

Elukestusanalüüsiks nimetatakse uurimisharu, kus uuritakse aja pikkust mingi sündmuse juhtumiseni. Kuigi suur osa elukestusanalüüsist uurib, nagu nimigi ütleb, elu pikkust, ehk sündmusena kasutatakse surma, siis ei ole elukestusanalüüs piiritletud elu kestusega. Näiteks kasutatakse elukestusanalüüsi meetodeid ka haiguste tekkimise ja erinevate elu sündmuste uurimiseks, samuti ka kindlustusmatemaatikas ja tööstuses toodete kvaliteedi kontrollil. Elukestusanalüüsi puhul tuleb meeles pidada, et tihti ei ole infot iga vaadeldava objekti uuritava sündmuse toimumisaja kohta. See tähendab, et kui uuritava sündmusena käsitletakse näiteks surma, siis uuringu tegemise hetkel ei ole kas mõned isikud veel surnud või pole surma aja kohta infot mõnel muul põhjusel. (Kartsonaki, 2016)

Sellisel juhul, kui uuritav sündmus ei ole toimunud, kasutatakse elukestusanalüüsis tsenseerimise mõistet. Tsenseerimine tähendab sellisel juhul, et teadaolev uuringus oldud aeg jääb väiksemaks sündmuse toimumise ajast. Nii kasutataksegi elukestusanalüüsis tunnuste paari (Y_i, δ_i), kus

$$Y_i = min(C_i, T_i) \quad ja \quad \delta_i = \begin{cases} 1, \ kui \ C_i \geq T_i \\ 0, \ kui \ C_i < T_i \end{cases} \tag{3}$$

ning kus omakorda C_i on tsenseerimise aeg ja T_i sündmuse toimumise aeg. Enamasti eeldatakse, et C_i ja T_i on sõltumatud. (Zimmermann, 2018)

Elukestusanalüüsis on kesksel kohal kaks funktsiooni - üleelamisfunktsioon ja riskifunktsioon. Juhusliku suuruse T üleelamisfunktsiooniks nimetatakse tõenäosust, et elukestus on ajalises mõttes pikem kui t ning defineeritakse kujul

$$S(t) = 1 - F(t),$$

kus $F(t)$ on juhusliku suuruse T jaotusfunktsioon. Juhusliku suuruse T riskifunktsiooniks nimetatakse tõenäosust, et kui elukestus on suurem kui t, siis uuritav sündmus toimub lõpmata väikeses ajavahemikus $[t; t + \delta t]$. T riskifunktsioon defineeritakse kujul

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t \,|\, T \geq t)}{\delta t} = \frac{f(t)}{S(t)},$$

kus $f(t)$ on juhusliku suuruse T tihedusfunktsioon. (Collett, 2015)

3.2 Weibulli jaotus

Weibulli jaotust hakkasid enne selle praeguse nime saamist esmalt kasutama P. Rosin ja E. Rammler, kes kirjeldasid sellega pulbristatud kivisöe peenust. Jaotus, millega praegusel ajal kirjeldatakse pigem eluiga, sai Weibulli nime pärast seda, kui Werody Weibull kasutas seda materjalide eluea uurimiseks. Peale elututе objektide, kirjeldatakse Weibulli jaotusega ka elus objektide, samuti ka inimeste eluiga. Pärast teist maailmasõda kasutati Weibulli jaotust ka näiteks Hiroshima katastroofis ellujäänud isikute edasise eluea pikkuse kirjeldamiseks. (Taketomi *et al.*, 2022)

Weibulli jaotuse jaotusfunktsioon on sarnane eksponentjaotuse jaotusfunktsiooniga, kuid kui eksponentjaotusel on vaid üks, siis Weibulli jaotusel on kaks parameetrit, λ ja γ. Tegelikult ongi eksponentjaotus üks võimalik Weibulli jaotuse vorm, kus parameeter $\gamma = 1$. Weibulli jaotuse üleelamisfunktsioon on kujul

$$S(t) = 1 - F(t) = 1 - (1 - \exp(-\lambda t^{\gamma})) = \exp(-\lambda t^{\gamma}) \tag{4}$$

ja riskifunktsioon kujul

$$h(t) = \lambda \gamma t^{\lambda - 1},$$

kus $\lambda > 0$ nimetatakse skaalaparameetriks ja $\gamma > 0$ nimetatakse kujuparameetriks ning kus $T \sim W(\lambda, \gamma)$ on Weibulli jaotusest juhulik suurus. (Zimmermann, 2018)

Kuigi tihti kasutatakse Weibulli jaotuse puhul valemis (4) olevat üleelamisfunktsiooni kuju, siis võib eri allikates märgata ka teistsuguseid definitsioone. Näiteks rakendustarkvara R kasutades võib tähele panna, et sealses *stats* paketis olevates Weibulli funktsioonides on jaotusfunktsioon kirjeldatud kui

$$F(t) = 1 - e^{-\left(\frac{t}{\lambda'}\right)^{\gamma}}$$

ehk üleelamisfunktsioon avaldub siit kujul

$$S(t) = 1 - F(t) = 1 - \left(1 - e^{-\left(\frac{t}{\lambda'}\right)^{\gamma}}\right) = e^{-\left(\frac{t}{\lambda'}\right)^{\gamma}},$$

kus samuti $\lambda' > 0$. Sellisel juhul $\lambda = \left(\frac{1}{\lambda'}\right)^{\gamma}$. (R Documentation, 2019)

3.3 Coxi võrdeliste riskide mudel

Coxi poolt 1972. aastal tutvustatud võrdeliste riskide mudel, mis kannab ka Coxi nime, on üks enim kasutatud mudeleid elukestusanalüüsis. Coxi mudeli riskifunktsioon on kujul

$$h_i(t) = h_0(t)\ \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik}),$$

kus $\beta_1, \beta_2, ..., \beta_k$ on mudelist saadavad koefitsiendid, $h_0(t)$ on baasriskifunktsioon ning $X_{i1}, X_{i2}, ..., X_{ik}$ vastavad indiviidi mudelis olevate tunnuste väärtustele. Defineerides juhtude i ja i', mille X_i väärtused erinevad, jaoks

$$\eta_i = \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_k X_{ik},$$
$$\eta_{i'} = \beta_1 X_{i'1} + \beta_2 X_{i'2} + ... + \beta_k X_{i'k},$$

saame kirjutada Coxi võrdeliste riskide mudeli kujul

$$\frac{h_i(t)}{h_{i'}(t)} = \frac{h_0(t)\, e^{\eta_i}}{h_0(t)\, e^{\eta_{i'}}} = \frac{e^{\eta_i}}{e^{\eta_{i'}}}.$$

Siit võib tähele panna, et Coxi võrdeliste riskide mudelis jäetakse $h_0(t)$ välja, mis teeb täieliku tõepärafunktsiooni arvutamise keeruliseks. Seetõttu tuli Cox välja osalise tõepärafunktsiooni ideega, kus ei käsitleta aeg t täpseid väärtuseid, vaid nende suurusjärjestust. Sellest tulenevalt nimetatakse Coxi proportsionaalset riskimudelit ka poolparameetriliseks. (Fox ja Weisberg, 2023)

3.4 Kaplan-Meieri hinnang

Kaplan-Meieri hinnangut kasutatakse üleelamisfunktsiooni hindamiseks. Tsenseerimist sisaldavate andmete puhul avaldub Kaplan-Meieri hinnang üleelamisfunktsioonile kujul

$$\hat{S}(t) = \prod_{j=1, t_j \leq t}^{k} \left(\frac{n_j - d_j}{n_j} \right), \tag{5}$$

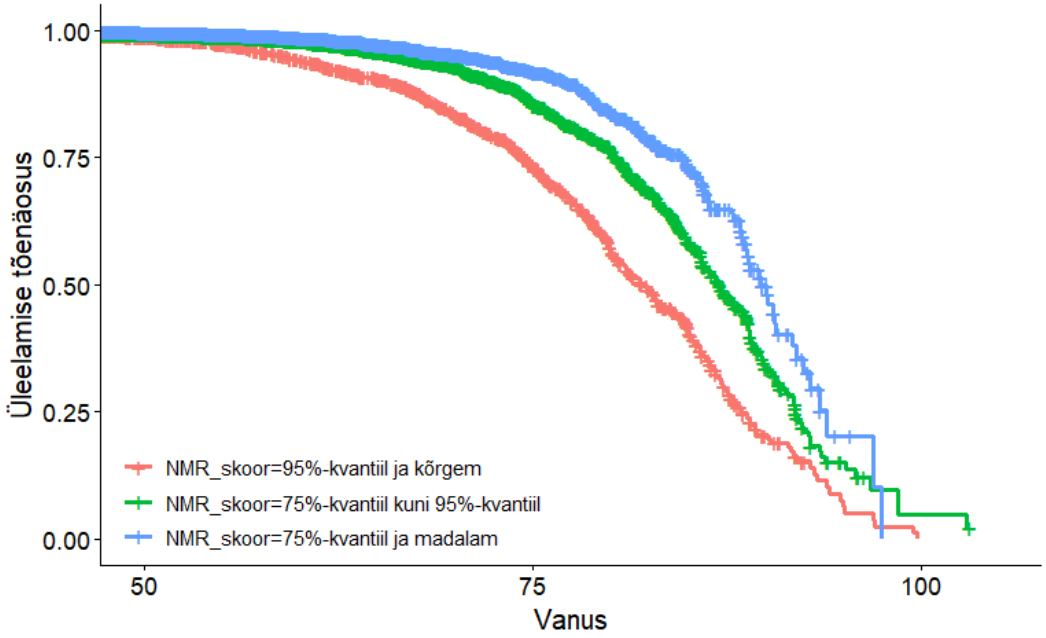
kus n_j on ajahetkel t riskigrupis olevate inimeste arv. Teisisõnu on n_j inimeste arv, kes on ajahetkel t uuringus, kuid kellega ei ole uuritavat sündmust ajahetkeks t veel juhtunud. Valemis (5) tähistab d_j inimeste arvu, kelle puhul uuritav sündmuse toimub ajahetkel t. (Zimmermann, 2018)

4 Bioloogilise vanuse arvutamine elukestusanalüüsi meetoditega

Bioloogilise vanuse arvutamiseks kasutati sugu, vanust liitudes, vanust uuringu lõpus või surma korral ning erinevaid biomarkereid. Kuna biomarkerite seas esines puuduvaid väärtuseid, eemaldati need enne mudeli tegemist andmestikust ning saadi 153 895 inimesega andmestik. Seejärel jagati andmestik kaheks, millest pool (76 947 inimest) läks treeningandmestikku ning teine pool (76 948 inimest) testandmestikku.

4.1 NMR-skoori arvutamine Coxi võrdeliste riskide mudeliga

Saadud treeningandmestikule rakendati rakendustarkvara R funktsiooni *coxph* ehk Coxi võrdeliste riskide mudelit, et luua biomarkerite kombinatsiooni põhjal kokku panduna üks pidev tunnus, mida edaspidi nimetame NMR-skooriks. Uuritava tunnusena lisati vanus liitudes, vanus uuringu lõpus või surma korral ning tsenseerimise tunnus. Sõltumatute tunnustena lisati mudelisse kokku 20 biomarkeri tunnust, mille kõigi p-väärtus oli Bonferroni meetodil ($\alpha = 0,05$ korral) statistiliselt olulised ehk iga biomarkeri korral $p < \frac{\alpha}{20} = 0{,}0025$. Saadud mudel ja biomarkerite lühendid on nähtaval lisas 1 (Lisa 1. Coxi võrdeliste riskide mudel). Seejärel arvutati testandmestikus saadud mudeli põhjal välja igale isikule vastav NMR-skoor.



Joonis 3: Meeste Kaplan-Meier funktsioonid eri NMR-skoori vahemikes.

Joonisel 3 on kujutatud testandmestikus olevate meessoost isikute Kaplan-Meier funktsioone eri NMR-skooride vahemikes. Graafikult võib märgata, et suurema NMR-skooriga meestel on suurem risk varakult surra. Graafiku loomisel on vaadeldud vaid meessoost isikute andmeid seetõttu, et naiste ja meeste keskmine oodatav eluiga erinevad. Sama tendents kehtib ka naiste puhul.

4.2 Weibulli jaotusega elukestusmudeli rakendamine

Varem defineeritud testandmestik jagati tulemuste kuvamiseks taas kaheks, millest 1000 isikut lisati antud andmestikust loodud testandmestikku ning ülejäänud 75 948 lisati antud andmestiku treeningandmestikku. Uuele treeningandmestikule rakendati esmalt Weibulli jaotusega elukestusmudelit (R-s *survreg*), kus uuritavaks tunnuseks oli valemis (3) tunnuste paar, kus aja tunnusena on kasutatud uuringus oldud aja pikkust. Sõltumatute tunnustena lisati mudelisse sugu, vanus liitudes ja NMR-skoor (edaspidi Mudel 1). Lisaks tehti veel teine sarnane mudel (edaspidi

Mudel 2), kuid kus sõltumatu tunnusena lisati vaid vanus liitudes. Mõlemas mudelis olid kõikide tunnuste p-väärtused $p < 0{,}001$ ehk tunnused olid statistiliselt olulised. Mudelid on nähtavad lisades (Lisa 2. Mudel 1 ja Lisa 3. Mudel 2).

4.3 Bioloogilise vanuse arvutamine

Olgu $S(t) = P(T > t)$ uuritava üldkogumi keskmine üleelamisfunktsioon ehk tõenäosus, et juhuslikult valitud isik elab vähemalt vanuseni t ning olgu $S_i(t) = P(T_i > t) = P(T > t \,|X_{i1}, ..., X_{in})$ indiviidi üleelamisfunktsioon, mis sõltub tunnustest $X_{i1}, ..., X_{in}$. Järgnevalt defineerime bioloogilise vanuse kui uuritavas populatsioonis selliste isikute keskmise vanuse, kelle puhul $S_i(CA_i) = S(BA_i)$.

Weibulli jaotusega elukestusmudeli abil bioloogilise vanuse arvutamiseks, tuleb esmalt avaldada bioloogilise vanuse tunnus. Bioloogilise vanuse avaldamiseks tuleb esmalt avaldada $\ln(\lambda_0)$, kus λ_0 on Mudel 2-st saadud λ-parameeter. Edasi saab Mudel 1 ja Mudel 2 üleelamisfunktsiooni kasutades $S_i(t) = S_0(t)$ kirjutada kui $\exp(-\lambda_i t^{\gamma_1}) = \exp(-\lambda_0 t^{\gamma_0})$, millest saab tuletada $\ln(\lambda_0) = \frac{\gamma_1}{\gamma_0} \ln(\lambda_i)$. Tähistades $\ln(\lambda_0) = f(BA)$, saab bioloogilise vanuse kirjutada kui

$$BA_i = f^{-1}\left(\frac{\gamma_1}{\gamma_0} \ln(\lambda_i)\right), \tag{6}$$

kus defineerides $\eta_i = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$, on $\lambda_i = \exp(-\eta_i)$ isiku andmetel arvutatud Mudel 1-st saadud λ-parameeter. $\beta_0, \beta_1, ..., \beta_k$ on Mudel 1-st saadud koefitsiendid ja $X_1, X_2, ..., X_k$ on isikule vastavad mudelis olevate tunnuste väärtused. γ_1 ja γ_0 on vastavalt Mudel 1-st ja Mudel 2-st saadud γ-parameetrid. Eelpool defineeritud Mudel 2 puhul $\lambda_0 = \exp(\beta_0' + \beta_1' X_1)$ ehk $f(BA) = \beta_0' + \beta_1' X_1$, kus $\beta_0', \beta_1', ..., \beta_k'$ on Mudel 2-st saadud koefitsiendid ja X_1 on isiku vanus liitudes. Siit saab valemis (6) oleva pöördfunktsiooni kirjutada kui

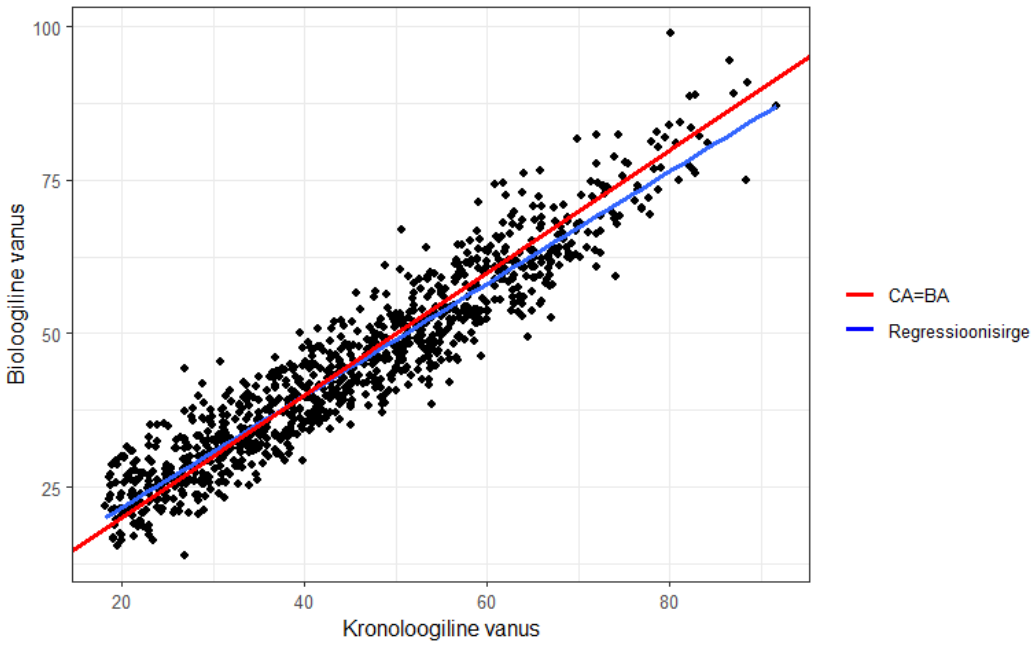
$$f^{-1}(x) = \frac{x - \beta_0'}{\beta_1'}$$

ning edasi saab bioloogilise vanuse valemi kirjutada kujul

$$BA_i = \frac{\frac{\gamma_1}{\gamma_0}\ln(\lambda_i) - \beta_0'}{\beta_1'}. \tag{7}$$

4.4 Tulemused sugu arvestava mudeliga

Kasutades valemit (7), arvutati algse andmestiku testandmestikust võetud uues testandmestikus igale isikule bioloogiline vanus. Joonisel 4 oleval graafikul on kujutatud antud andmestikus olevate isikute kronoloogiline ja bioloogiline vanus, kus iga punkt vastab ühele isikule.

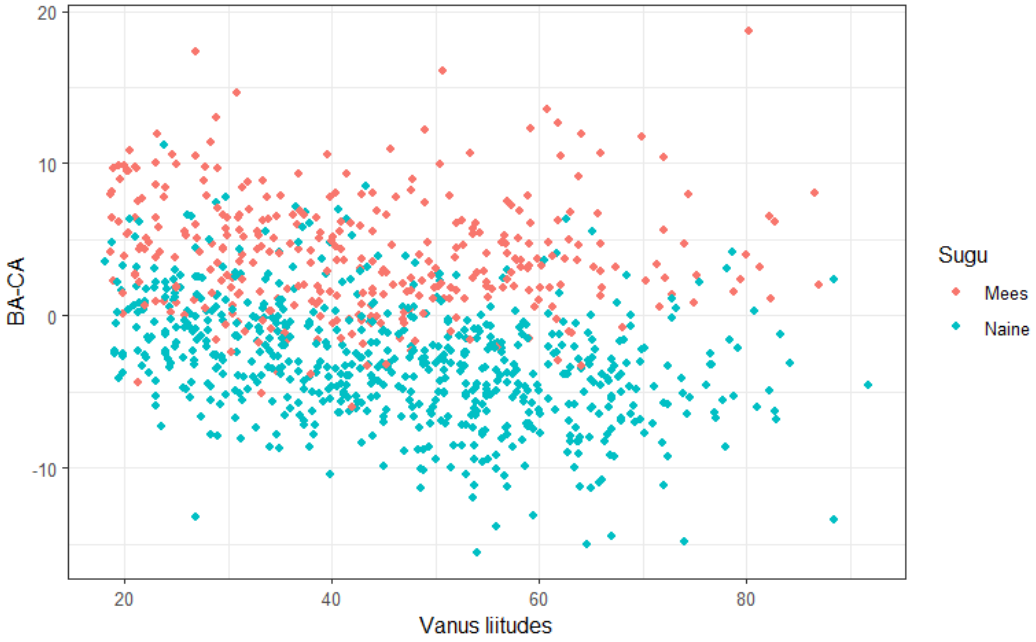


Joonis 4: Uue andmestiku testandmestiku BA ja CA graafik, kus üks punkt tähistab ühte isikut.

Graafikult võib näha, et regressioonisirge ja $BA = CA$ sirge on ühtlasema tõusuga, kui peatükis 2 nähtavatel graafikutel. See tähendab, et kuigi isiku bioloogiline vanus võib olla nii suurem kui ka väiksem, kui kronoloogiline vanus, siis keskmiselt

on ühe kronoloogilise vanusega inimeste keskmine bioloogiline vanus kronoloogilise vanusega võrdne või ligilähedane. Regressioonisirge on joonisele lisatud R-s visualiseerimiseks ning ei pärine mudelist.

Uurides valemis (2) defineeritud ε_i suuruseid, võib jooniselt 5 märgata, et meestel on bioloogilise ja kronoloogilise vanuse vahe ε_i enamjaolt positiivne ning naistel negatiivne. Sellist nähtust võib seletada see, et naiste keskmine oodatav eluiga on suurem, kui meestel, mistõttu tuleks meeste ja naiste bioloogilise vanuse arvutamiseks luua kaks eri mudelit.



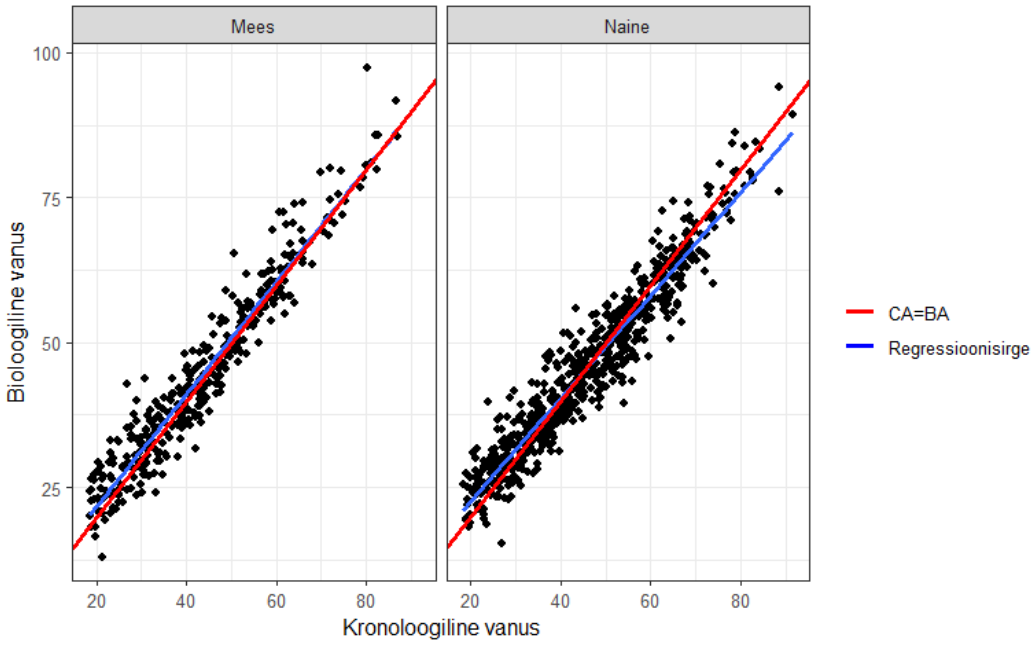
Joonis 5: Uue andmestiku testandmestiku $BA - CA$ graafik, kus üks punkt tähistab ühte isikut.

4.5 Tulemused eraldi soopõhiste mudelitega

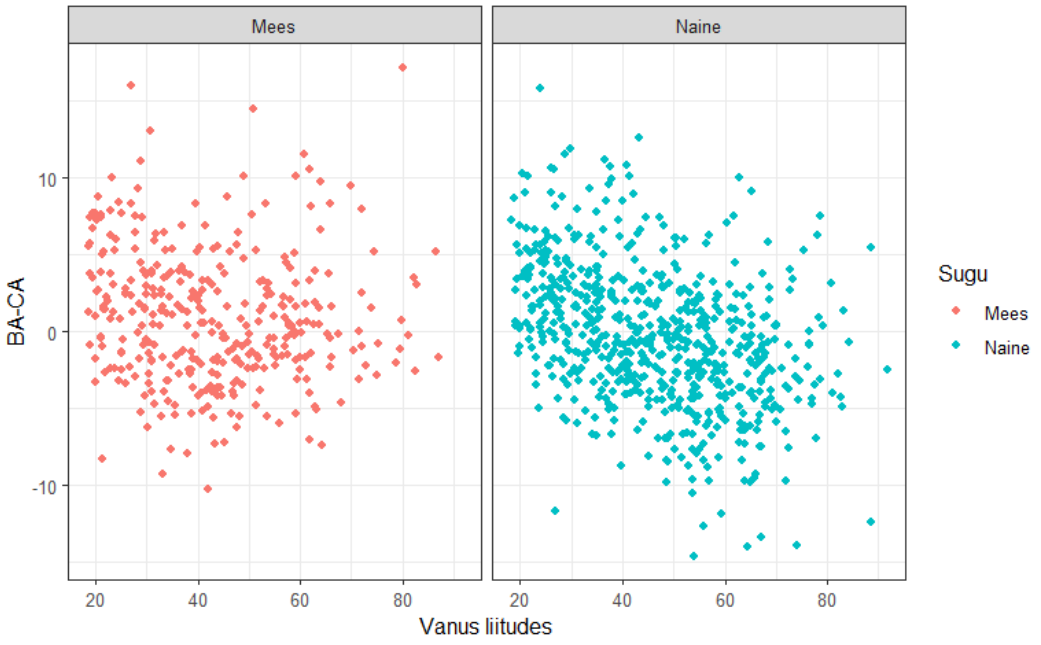
Soopõhiste mudelite loomiseks eraldati käsitletav testandmestik kaheks eraldi andmestikuks, kus ühes olid mehed (346 isikut) ja teises naised (654 isikut). Mõlema

andmestiku peal korrati peatükkides 4.2 ja 4.3 tehtud samme, ainsa erinevusega, et sugu Weibulli jaotusega Mudel 1 ei lisatud.

Naistele ja meestele arvutati erinevate koefitsientidega bioloogilised vanused. Jooniselt 6 võib näha, et nii meeste, kui naiste puhul on regressioonisirge ning $CA = BA$ sirge sarnase tõusuga ehk võib märgata peatükis 4.4 saadud tulemustega sarnast tulemust. Siiski on jooniselt 7 näha peatükis 4.4 täheldatud tendentsi, kus meestel on bioloogiline vanus enamjaolt kronoloogilisest vanusest suurem ja naistel vastupidi, enam märgata ei ole.

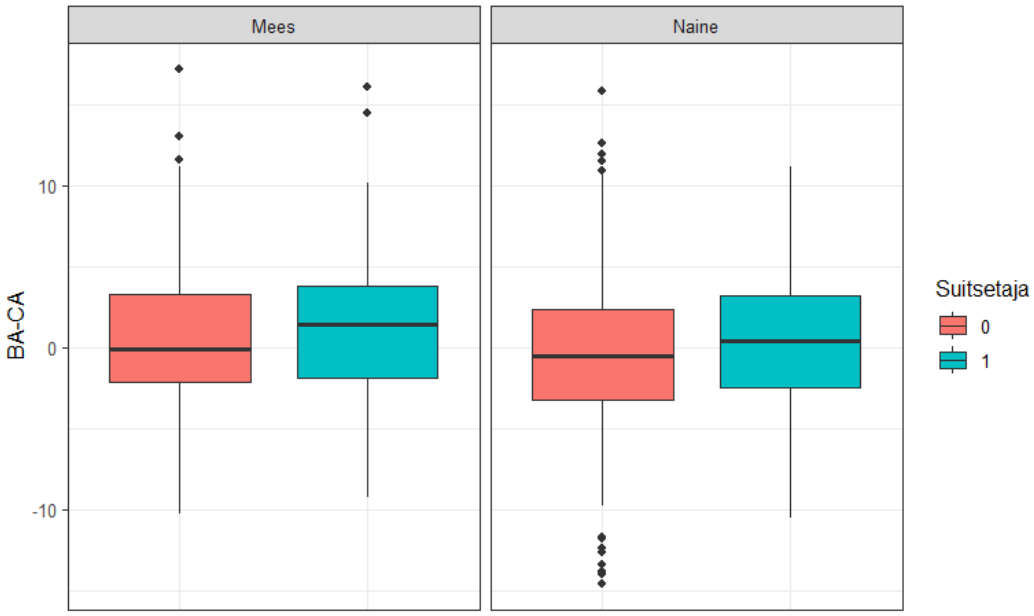


Joonis 6: Meeste ja naiste bioloogilise vanuse ning kronoloogilise vanuse graafik, kus iga punkt tähistab ühte isikut.



Joonis 7: Meeste ja naiste $BA-CA$ graafik, kus iga punkt tähistab ühte isikut.

Peatükis 2 toodi korduvalt välja, et suitsetavate inimeste bioloogiline vanus on MLR-ga arvutades madalam, kui mittesuitsetajate bioloogiline vanus. Uurides suitsetajate ja mittesuitsetajate bioloogilise vanuse vahet kronoloogilise vanusega, võib jooniselt 8 tähele panna, et elukestusanalüüsi meetoditel arvutatud mudeliga on suitsetajate keskmine bioloogiline vanus kronoloogilisest vanusest suurem. Nii meeste kui naiste puhul võib jooniselt märgata, et antud andmestikus olevate mittesuitsetajate keskmine $BA–CA$ vahe on negatiivne ehk mittesuitsetajate kronoloogiline vanus on suurem, kui bioloogiline vanus.



Joonis 8: Meeste ja naiste $BA-CA$ karpdiagramm, kus eristav tunnus on suitsetamine (1-suitsetaja, 0-mittesuitsetaja).

Saab väita, et elukestusanalüüsiga arvutatud bioloogiline vanus annab loogilisemaid tulemusi, kui MLR-ga arvutatud bioloogiline vanus. Elukestusanalüüsi meetodid väldivad peatükis 2 välja toodud negatiivseid aspekte ning muudavad bioloogilise vanuse keskmiselt ligilähedasemaks kronoloogilisele vanusele. Siiski ei ole kronoloogiline vanus isikupõhiselt võrdne bioloogilise vanusega, mis näitab, et biomarkeritel on suur mõju inimese bioloogilisele vanusele.

5 Kokkuvõte

Antud bakalaureusetöö eesmärgiks oli leida statistiline meetod, mille abil arvutada bioloogiline vanus Eesti Geenivaramu andmestikus olevatele isikutele.

Esmalt käsitleti mitmese lineaarse regressioonmudeliga bioloogilise vanuse arvutamise meetodit. Antud meetodi puhul toodi välja mõningad negatiivsed aspektid, mille tõttu on soovituslik bioloogilist vanust arvutada mõne muu statistilise meetodi abil. Põhilisteks vigadeks oli nii MLR-st saadud koefitsientide ebaloogilisus, eelkõige suitsetamise osas, kui ka regressioonisirge kallutatus. Kuna regressioonisirge tõus erineb silmnähtavalt $BA = CA$ graafikust, saab väita, et noorematel inimestel ülehinnatakse ning vanematel inimestel alahinnatakse bioloogilist vanust.

MLR-ga bioloogilise vanuse arvutamisel esinevate negatiivsete aspektide tõttu käsitleti bioloogilise vanuse arvutamisel elukestusanalüüsi meetodeid. Bioloogilise vanuse arvutamise genereeriti Coxi võrdeliste riskide mudeli abil igale isikule NMR-skoor, mis saadi kasutades 20 statistiliselt olulist biomarkerit. Edasi tuletati Weibulli jaotusega elukestusmudeli valemitest bioloogilise vanuse arvutamise valem, milles kasutati uuritava tunnuse defineerimisel uuringus oldud aja pikkust ning tsenseerimise tunnust. Sõltumatute tunnustena kasutati bioloogilise vanuse arvutamiseks vanust liitudes, NMR-skoori ja sugu.

Kasutades mudelit, kus sugu oli sõltumatu tunnus, saadi tulemuseks bioloogilise vanuse arvutamise valem, mis ülehindas meeste ja alahindas naiste bioloogilist vanust. Seetõttu loodi kaks eri mudelit mõlema soo jaoks eraldi ning saadi tulemused, mille puhul kaotati kõik töös eelnevalt loetletud negatiivsed aspektid.

Kasutatud allikad

Cho, Il Haeng, Kyung S. Park ja Chang Joo Lim (2010). *An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI)*. URL: https://www-sciencedirect-com.ezproxy.utlib.ut.ee/science/article/pii/S0047637409001675?ref=cra_js_challenge&fr=RR-1 (vaadatud 15.02.2023).

Collett, David (2015). *Modelling survival data in medical research*. Second Edition. Chapman & Hall/CRC.

Fox, John ja Sanford Weisberg (31. jaanuar 2023). *Cox Proportional-Hazards Regression for Survival Data in R*. An Appendix to An R Companion to Applied Regression, third edition. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/appendices/Appendix-Cox-Regression.pdf (vaadatud 03.04.2023).

Jee, Haemi ja Jaehyun Park (2017). *Selection of an optimal set of biomarkers and comparative analyses of biological age estimation models in Korean females*. URL: https://www-sciencedirect-com.ezproxy.utlib.ut.ee/science/article/pii/S0167494317300213?ref=cra_js_challenge&fr=RR-1 (vaadatud 15.02.2023).

Kartsonaki, Christiana (2016). *Survival analysis*. URL: https://www.sciencedirect.com/science/article/pii/S1756231716300639?ref=cra_js_challenge&fr=RR-1 (vaadatud 12.03.2023).

Krut'ko, V. N., T. M. Smirnova, V. I. Dontsov ja S. E. Borisov (2000). *Diagnosing Aging: I. Problem of Reliability of Linear Regression Models of Biological Age*. URL: https://link-springer-com.ezproxy.utlib.ut.ee/content/pdf/10.1023/A:1012941413535.pdf (vaadatud 15.02.2023).

Li, Zhe, Weiguang Zhang, Yuting Duan, Yue Niu, Yan He, Yizhi Chen, Xiaomin Liu, Zheyi Dong, Ying Zheng, Xizhao Chen, Zhe Feng, Yong

Wang, Delong Zhao, Xuefeng Sun, Guangyan Cai, Hongwei Jiang ja Xiangmei Chen (2023). *Biological age models based on a healthy Han Chinese population.* URL: https://www-sciencedirect-com.ezproxy.utlib.ut.ee/science/article/pii/S0167494322002928?ref=cra_js_challenge&fr=RR-1 (vaadatud 15.02.2023).

Nightingale Health (2020). *Our technology.* https://research.nightingalehealth.com/technology. (Vaadatud 12.04.2023).

R Documentation (2019). *Weibull: The Weibull Distribution.* https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/Weibull. (Vaadatud 03.04.2023).

Taketomi, Nanami, Kazuki Yamamoto, Christophe Chesneau ja Takeshi Emura (21. oktoober 2022). *Parametric Distributions for Survival and Reliability Analyses, a Review and Historical Sketch.* URL: https://www.mdpi.com/2227-7390/10/20/3907 (vaadatud 18.03.2023).

Zimmermann, Marili (2018). *Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel.* URL: https://dspace.ut.ee/bitstream/handle/10062/61020/zimmermann_marili_msc_2018.pdf?sequence=1&isAllowed=y.

Lisad

Lisa 1. Coxi võrdeliste riskide mudel

	coef	exp(coef)	se(coef)	z	p
Omega_6_by_Omega_3	0.20315	1.22525	0.02343	8.671	< 0.0000000000000002
Gly	-0.13527	0.87348	0.02238	-6.044	0.000000001505283
Tyr	-0.08648	0.91715	0.02364	-3.658	0.000254
Glucose	0.21800	1.24358	0.01970	11.063	< 0.0000000000000002
S_HDL_PL_pct	-0.15557	0.85593	0.04489	-3.465	0.000530
S_HDL_FC_pct	0.20191	1.22373	0.03781	5.340	0.000000093050314
MUFA_pct	0.20684	1.22979	0.04914	4.209	0.000025637047464
Total_C	-1.99758	0.13566	0.22343	-8.941	< 0.0000000000000002
Sphingomyelins	0.63763	1.89199	0.08604	7.411	0.000000000000126
non_HDL_C	2.87252	17.68157	0.46980	6.114	0.000000000969770
His	-0.07614	0.92668	0.02402	-3.169	0.001528
L_LDL_C_pct	0.17744	1.19416	0.03881	4.572	0.000004826840707
Clinical_LDL_C	-1.36464	0.25547	0.31501	-4.332	0.000014770724897
XL_VLDL_TG_pct	-0.14327	0.86652	0.02982	-4.804	0.000001552926332
L_VLDL_CE_pct	-0.30208	0.73928	0.05460	-5.533	0.000000031519422
VLDL_size	-0.48877	0.61338	0.10274	-4.757	0.000001960463282
Acetone	0.03404	1.03462	0.01106	3.077	0.002088
Creatinine	0.06346	1.06551	0.01927	3.293	0.000991
XS_VLDL_L	-0.31598	0.72907	0.10616	-2.976	0.002916
XL_HDL_FC_pct	-0.12601	0.88161	0.03960	-3.182	0.001461

Likelihood ratio test=862.9 on 20 df, p=< 0.00000000000000022
n= 76947, number of events= 2331

Joonis 9: 20 biomarkeriga Coxi võrdeliste riskide mudel treeningandmestikul.

Lisa 2. Mudel 1

```
Call:
survreg(formula = Surv(time, status) ~ vanus_liitudes + sex_W +
    nmr_score, data = data_test_treening, dist = "weibull")
                Value Std. Error     z      p
(Intercept)     6.61582    0.08250  80.2 <2e-16
vanus_liitudes -0.05112    0.00109 -46.8 <2e-16
sex_W           0.32809    0.02485  13.2 <2e-16
nmr_score      -0.42874    0.01929 -22.2 <2e-16
Log(scale)     -0.55320    0.01425 -38.8 <2e-16

Scale= 0.575

Weibull distribution
Loglik(model)= -10710.5   Loglik(intercept only)= -13497.5
        Chisq= 5574 on 3 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 10
n= 74948
```

Joonis 10: Weibulli jaotusega loodud mudel kolme sõltumatu tunnusega.

Lisa 3. Mudel 2

```
Call:
survreg(formula = Surv(time, status) ~ vanus_liitudes, data = data_test_treening,
    dist = "weibull")
                Value Std. Error     z      p
(Intercept)     6.8339     0.0821  83.2 <2e-16
vanus_liitudes -0.0545     0.0011 -49.7 <2e-16
Log(scale)     -0.5786     0.0136 -42.7 <2e-16

Scale= 0.561

Weibull distribution
Loglik(model)= -11121.5   Loglik(intercept only)= -13497.5
        Chisq= 4752.08 on 1 degrees of freedom, p= 0
Number of Newton-Raphson Iterations: 10
n= 74948
```

Joonis 11: Weibulli jaotusega loodud mudel ühe sõltumatu tunnusega.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Stein-Marten Pool,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose Statistilised meetodid bioloogilise vanuse hindamiseks, mille juhendaja on Prof. Krista Fischer, 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.

Stein-Marten Pool
09.05.2023