

TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
MATEMAATIKA JA STATISTIKA INSTITUUT

Kadi-Liis Kuum
**Pseudovaatluste kasutamine elukestusanalüüsis
liiklusõnnetuste ravikulude hüvitamise näitel**

Kindlustus- ja finantsmatemaatika (162777)

Magistritöö (30 EAP)

Juhendaja: MSc Anastassia Kolde

TARTU 2022

PSEUDOVAATLUSTE KASUTAMINE ELUKESTUSANALÜÜSIS LIIKLUSÕNNETUSTE RAVIKULUDE HÜVITAMISE NÄITEL

Magistritöö

Kadi-Liis Kuum

Lühikokkuvõte

Magistritöö eesmärk on tutvustada pseudovaatluste kasutamist elukestusanalüüsis. Pseudovaatlused on *jackknife* meetodil arvutatud hinnangud üleelamistõenäosusele, mis näitavad ühe vaatluse mõju valimi keskmisele, võttes arvesse andmetes esinevat tsenseerimist.

Töö teoreetilises osas tehakse ülevaade elukestusanalüüsi mõistetest ja meetoditest, keskendudes Coxi võrdeliste riskide mudelile. Seejärel tutvustatakse pseudovaatlusi ja kolme pseudovaatluste arvutamise juhtu: hinnangut üleelamistõenäosusele, hinnangut tõkestatud keskmisele elueale ja konkureerivate riskide kumulatiivsele avaldumisele.

Töö praktilises osas kasutatakse Coxi võrdeliste riskide meetodit ja pseudovaatlusi, et hinnata liiklusõnnetuse tagajärjel tekkinud ravikulude hüvitamise kestust mõjutavaid tegureid.

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

Märksõnad: Elukestusanalüüs, Coxi võrdeliste riskide mudel, pseudovaatlused, kahjukindlustus.

PSEUDO-OBSERVATIONS IN SURVIVAL ANALYSIS ON THE EXAMPLE OF REIMBURSEMENT OF MEDICAL EXPENSES FOR TRAFFIC ACCIDENTS

Master thesis

Kadi-Liis Kuum

Abstract

The aim of this thesis is to introduce pseudo-observations in survival analysis. Pseudo-observations are estimations of survival probability, calculated by *jackknife* method, that show the effect of one observation to the sample mean, taking into account censoring.

The theoretical part of the thesis gives an overview of survival analysis and it's methods focusing on the Cox proportional-hazards model. It is followed by an introduction of pseudo-observations and three instances of calculating pseudo-observations: estimation on survival probability, estimation on restricted mean survival time and cumulative incidence of competing risks.

In the practical part of the thesis Cox model and pseudo-observations are used to estimate the factors that have an effect on the duration of reimbursement of medical expenses caused by traffic accidents.

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

Key Words: Survival analysis, Cox proportional hazards model, pseudo-observations, non-life insurance.

Sisukord

Sissejuhatus	5
1 Elukestusanalüüs	7
1.1 Olulisemad mõisted	8
1.2 Kaplan-Meieri hinnang	11
1.3 Coxi võrdeliste riskide mudel	13
1.3.1 Riskide proportsionaalsuse eelduse kontrollimine	13
1.3.2 Stratifitseeritud Coxi mudel ja ajas muutuvad koefitsiendid	15
2 Pseudovaatlused	17
2.1 Pseudovaatluste arvutamine	18
2.1.1 Hinnang üleelamistõenäosusele	18
2.1.2 Tõkestatud keskmine	19
2.1.3 Konkureerivate riskide mudel	20
2.2 Mudelite sobitamine	22
2.2.1 Riskimäärade suhte hindamine üleelamistõenäosusele	23
2.3 Pseudovaatlused R-is	25
2.4 Pseudovaatlused simulatsiooni näitel	27
2.4.1 Näide ilma tsenseerimiseta	27
2.4.2 Näide tsenseerimisega	29
2.4.3 Näide tõkestatud keskmisest	31
3 LKFi andmete näitel	35
3.1 Andmete kirjeldus	35

3.2 Tulemused	37
4 Arutelu	42
Kokkuvõte	44
Kasutatud allikad	46
Lisad	49

Sissejuhatus

Elukestusanalüüs on analüüsi tüüp, millega uuritakse aega huvipakkuva sündmuse toimumiseni. Elukestusanalüüsi kasutatakse palju epidemioloogiliste uuringute puhul, kus uuritakse aega surmani, haiguse tekkeni või näiteks operatsioonist taastumise aega. Kestusandmed on kasutusel ka näiteks tööstuses (aeg mõne seadme tõrkeni), panganduses (aeg laenuvõtja maksejõuetuseni) ja paljudes muudes elusfäärides. Elukestusanalüüsi omapäraks on puuduvad ehk tsenseeritud vaatlused. Tsenseerimine toimub enamasti siis, kui huvipakkuv sündmus ei ole vaadeldava aja jooksul toimunud ja vaadeldava elukestuse puhul on teada aeg tsenseerimiseni, kuid mitte tegelik elukestus. Tsenseerimise arvesse võtmiseks tuleb kasutada elukestusanalüüsi meetodeid. (Collett, 2015)

Üks levinumaid elukestusanalüüsi meetodeid on Coxi võrdeliste riskide mudel. See mudel hindab riskide suhet. Coxi mudeli kasutamise korral peab arvestama olulise eeldusega, et riskide suhe on võrdeline. See tähendab, et riskide suhe on aja t jooksul muutumatu. See eeldus ei ole aga tihti peale päriselu andmete puhul täidetud. Üks võimalik lahendus mitte-võrdeliste riskide mudeldamiseks on pseudovaatluste kasutamine.

Pseudovaatluste kasutamise idee seisneb selles, et *jackknife* meetodil hinnatakse ühe vaatluse mõju valimi keskmisele, võttes seejuures arvesse ka tsenseeritud vaatlusi ning teadmist, et tsenseeritud vaatluse elukestus ei ole teada, ning saadakse seeläbi uued vaatlused, milles ei esine tsenseerimist. Sellistele andmetele saab aga rakendada erinevaid statistilisi meetodeid, mida tavaliste elukestusandmete puhul ei saa, näiteks regressioonmudeleid või masinõppe meetodeid. (Andersen ja Perme, 2010)

Pseudovaatluste kasutamine elukestusanalüüsis on võrdlemisi uus meetod, mille pakkusid esmakordselt välja Andersen, Klein ja Rosthøj (2003). Alates sellest ajast on pseudovaatlusi kasutatud erinevate elukestusandmete uurimiseks. Siin töös kirjeldatakse kolme võimalikku viisi pseudovaatluste kasutamiseks: tavalist hinnan-

gut üleelamistõenäosusele, tõkestatud keskmise hinnangut üleelamistõenäosusele ja hinnangut konkureerivate riskide kumulatiivsele avaldumisele. Sellega ei ole aga pseudovaatluste kasutamine piiratud.

Käesoleva magistritöö üheks eesmärgiks on uurida pseudovaatluste kasutamist elukestusanalüüsis, kuidas pseudovaatlusi arvutada ja mudelites kasutada ning mis on pseudovaatluste kasutamise eelised ja puudused.

Töö teine eesmärk on hinnata liiklusõnnetuste ravikulude hüvitamise kestust mõjutavaid tegureid nii Coxi võrdeliste riskide mudeli kui ka pseudovaatluste abil. Uuritavad andmed on pärit Eesti Liikluskindlustuse Fondist (LKF). Tegemist on 2016.-2021. aasta andmetega Eestis toimunud liiklusõnnetuste kohta, kus üks või mitu osapoolt vajasisid või vajavad endiselt ravikulude hüvitamist. Andmestikus esineb ka tsenseeritud juhtumeid, mis antud juhul tähendab seda, et ravikulude hüvitamine ei ole 31.12.2021 seisuga lõppenud ning kogu hüvitamise pikkus ei ole teada.

Töö esimeses peatükis tehakse ülevaade elukestusanalüüsi olulisematest mõistetest ja meetoditest. Tutvustatakse Coxi võrdeliste riskide mudelit, proportsionaalsete riskide eeldust ja selle kontrollimist ning tuuakse välja kaks meetodit, millega mitte-proportsionaalseid riske hinnata. Teises peatükis tutvustatakse pseudovaatlusi. Pseudovaatluste kohta tehakse üldine ülevaade ning seejärel kirjeldatakse kolme pseudovaatluse arvutamise juhtu. Teises peatükis tutvustatakse ka üldistatud hindamismudeleid (*GEE*) ja kirjeldatakse üleelamistõenäosuste pseudovaatluste mudeldamist *GEE*-ga tarkvaras R ning tuuakse näiteid pseudovaatluste arvutamisest väikse simulatsioonandmestiku põhjal. Kolmandas peatükis kirjeldatakse LKF-i andmestikku, sellele sobitatud mudeleid ja saadud tulemusi.

Andmete analüüsimiseks kasutati statistikatarkvara R 4.0.5 (R Core Team, [2021](#)).

1 Elukestusanalüüs

Peatükk 1 ja selle alapeatükid põhinevad raamatul "*Modelling Survival Data in Medical Research*" (Collett, 2015), kui ei ole viidatud teisiti.

Elukestusanalüüs on analüüsi tüüp, kus uuritavaks suuruseks on aeg huvipakkuva sündmuse toimumiseni. Elukestusanalüüsi kasutatakse palju meditsiiniliste või epidemioloogiliste uuringute korral, kus uuritakse aega sündmuseni, näiteks surmani, haiguse tekkimiseni, paranemiseni, ja seda aega mõjutavaid faktoreid. Lisaks inimeste uurimisele saab elukestusanalüüsi kasutada ka erinevate objektide uurimiseks, näiteks elektroonikaseadmete tööaja hindamiseks. Siin töös uuritakse liiklusõnnetuse tagajärjel tekkinud ravikulude hüvitamise kestust ja seda mõjutavaid tegureid.

Elukestusanalüüsi omapäraks on tsenseerimine. Vaatlus on tsenseeritud, kui vaadeldava inimese või objekti puhul ei ole huvipakkuv sündmus vaadeldava aja jooksul toimunud ning me ei tea selle tegelikku väärtust. Kõige levinum tsenseerimise tüüp on paremalt tsenseerimine. Enamasti toimub paremalt tsenseerimine seetõttu, et huvipakkuv sündmus toimus pärast vaadeldava ajaperioodi lõppu, kuid vaatlus on paremalt tsenseeritud ka siis, kui näiteks vaadeldav inimene on uuringus osalemisest loobunud või mõne muu, mitte huvipakkuva sündmuse tõttu katkestanud. Kui kõik vaatlused oleks teada, saaks nende uurimiseks rakendada erinevaid tavapäraseid statistilisi meetodeid (Andersen ja Perme, 2010), kuid tsenseerimise arvesse võtmiseks kasutatakse elukestusanalüüsi meetodeid.

Lisaks paremalt tsenseerimisele võib elukestusanalüüsis tulla ette ka vasakult tõkestatust ja intervall-tsenseerimist. Vasakult tõkestatus tähendab seda, et huvipakkuv sündmus on toimunud enne vaadeldava ajaperioodi algust. Intervall-tsenseerimine on justkui kombinatsioon paremalt ja vasakult tsenseerimisest, kuna ei ole teada täpset sündmuse toimumise aega, kuid on teada intervall, kus sündmus toimus.

Oluline eeldus elukestusanalüüsi ja hiljem ka pseudovaatluste puhul on mitte-informatiivne tsenseerimine (*independent censoring*). See tähendab, et tsenseerimine ei sõltu ühestki parameetrist või mehhanismist ja ei anna sündmuse toimumise kohta mingit infot. Mitte-informatiivse tsenseerimise korral ei erine tsenseeritud isik millegi poolest mitte-tsenseeritud isikust. Mitte-informatiivne tsenseerimine on oluline, kuna paljud elukestusanalüüsi ja ka pseudovaatluste meetodid põhinevad sellel eeldusel ning informatiivse või ka sõltuva tsenseerimise korral peab kasutama teistsuguseid meetodeid, mida siin töös ei käsitleta.

Elukestusanalüüsi puhul on oluline määrata ajaskaala, mille jooksul elukestust vaadatakse. Tüüpilise elukestusuuringu puhul ei alga vaatlemine kõigil objektidel või inimestel samal ajal ning kõigi vaadeldava puhul ei juhtu ka vaadeldavat sündmust uuringu ajal ehk toimub tsenseerimine. Epidemioloogiliste uuringute puhul võib vaadeldavaks suuruseks olla näiteks inimese vanus – vanus uuringuga liitumise ajal, vanus uuritava haiguse tekkimise ajal või ka ainult vaatlusaeg. Siin töös on elukestusobjektiks vaatlusaeg päevades, mis on arvutatud ravikulude hüvitamise lõppkuupäeva ja alguskuupäeva vahena. Kui lõppkuupäeva teada ei ole, on tegemist tsenseeritud vaatlusega.

1.1 Olulisemad mõisted

Siin peatükis toome ülevaate olulisematest mõistetest ja funktsioonidest, mida elukestusanalüüsis kasutatakse. Need mõisted on vajalikud, et edaspidi kirjeldada analüüsi meetodeid, mida töös kasutati.

Üleelamisfunktsioon ehk *survival function* näitab vaadeldava inimese või objekti puhul tõenäosust, et sündmust ei toimu ajahetkeni t . Inimese elukestuse puhul näitab see tõenäosust, et inimene on elus või haiguse vaba ajahetkeni t .

Olgu T juhuslik suurus, mis näitab elukestust. Sellisel juhul avaldub üleelamis-

funktsioon kujul

$$S(t) = \mathbb{P}(T \geq t) = 1 - F(t),$$

kus $F(t)$ on elukestuse T jaotusfunktsioon.

Riskifunktsioon ehk *hazard function* näitab vaadeldava inimese või objekti riski, et sündmus toimub ajahetkel t , kui selle hetkeni sündmust toimunud ei ole. Elukestuse uurimise korral tähendab see riski, et inimene sureb ajahetkel t , kui ta on kuni hetkeni t elanud.

Elukestuse T korral, kus T on pideva jaotusega juhuslik suurus, on riskifunktsioon

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (1)$$

Kasutades tinglikku tõenäosust, saame valemi 1 viia kujule

$$\begin{aligned} h(t) &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t \cdot P(T \geq t)} = \\ &= \lim_{\Delta t \rightarrow 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} \cdot \frac{1}{S(t)} = \\ &= \frac{f(t)}{S(t)}, \end{aligned} \quad (2)$$

kus $f(t)$ on elukestuse T tihedusfunktsioon.

Saadud võrdusest 2 saame järeldada

$$h(t) = -\frac{d}{dt} \log(S(t)),$$

millest saame ellujäämisfunktsiooni avaldada kujul

$$S(t) = \exp(-H(t)), \quad (3)$$

kus $H(t)$ on kumulatiivne riskifunktsioon valemiga

$$H(t) = \int_0^t h(u)du,$$

mis näitab kumulatiivset riski, et huvipakkuv sündmus on toimunud ajahetkeks t . Kumulatiivset riski saame leida ka üleelamisfunktsiooni kaudu (valemi 3 põhjal)

$$H(t) = -\log(S(t)). \quad (4)$$

Elukestusanalüüsi puhul, nagu ka teiste statistiliste analüüsitüüpide puhul, pakub meile huvi erinevate parameetrite ehk seletavate tunnuste (*explanatory variables*) mõju elukestusele. Selle uurimiseks kasutatakse meetodeid, mis keskenduvad vaadeldava objekti riskifunktsiooni hindamisele erinevate parameetrite väärtuste korral. Järgnevalt vaatame võrdeliste riskide mudelit.

Olgu meil p seletavat tunnust $X_1, X_2, \dots, X_p$, mille väärtustest $\mathbf{x}$, nii et $\mathbf{x} = (x_1, x_2, \dots, x_p)'$, sõltub vaadeldava objekti risk. Kui vaadeldava objekti puhul on kõik $\mathbf{x}$ väärtused nullid, siis selle objekti riskifunktsiooni $h_0(t)$ nimetatakse baasriskiks. Selle põhjal saab leida i -nda objekti riski nii, et

$$h_i(t) = \psi(\mathbf{x}_i)h_0(t),$$

kus $\psi(\mathbf{x}_i)$ on funktsioon, mis koosneb i -nda vaatluse seletavate tunnuste väärtustest $x_{1i}, \dots, x_{pi}$. Funktsiooni $\psi(\mathbf{x}_i)$ saab käsitleda kui riskifunktsioonide suhet (*hazard ratio*) kahe vaadeldava vahel, kellest ühel on seletavate tunnuste väärtused $\mathbf{x}_i$ ja teisel $\mathbf{x} = \mathbf{0}$.

Kuna riskide suhe ei saa olla negatiivne, kasutatakse tihti $\psi(\mathbf{x}_i)$ asemel kuju

$$\exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}) = \exp(\boldsymbol{\beta}' \mathbf{x}_i),$$

kus $\psi(\cdot) = \exp(\cdot)$ ja $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ tähistab parameetrite kordajaid. Seega saame

võrdeliste riskide mudeli kirjutada kujul

$$\begin{aligned} h_i(t) &= h_0(t) \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}) \\ &= h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i). \end{aligned}$$

Kahe objekti riskide suhe, kus argumenttunnused on vastavalt $\mathbf{x}_1$ ja $\mathbf{x}_2$ ning riskifunktsioonid $h_1(t)$ ja $h_2(t)$, on

$$\frac{h_1(t)}{h_2(t)} = \frac{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_1)}{h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_2)} = \exp(\boldsymbol{\beta}' (\mathbf{x}_1 - \mathbf{x}_2))$$

ehk näeme, et kahe objekti riskide suhe ei sõltu ajast t ja baasriskist.

Elukestusanalüüsi meetodid saab jagada kolme kategooriasse: mitte-parameetrilised, poolparameetrilised ja parameetrilised. Siin töös tutvustatakse mitte-parameetrilist Kaplan-Meieri hinnangut ja poolparameetrilist Coxi võrdeliste riskide mudelit.

1.2 Kaplan-Meieri hinnang

Kaplan-Meieri hinnang on mitte-parameetiline meetod, millega hinnatakse üleelamisfunktsiooni $S(t)$.

Kui andmetes ei esineks tsenseerimist ehk kõikide vaatluste väärtus T oleks teada, saaksime üleelamisfunktsiooni hinnata empiirilise funktsiooniga

$$\hat{S}(t) = \frac{\text{Inimeste arv, kes elavad kauem kui } t}{\text{Kõigi inimeste arv}},$$

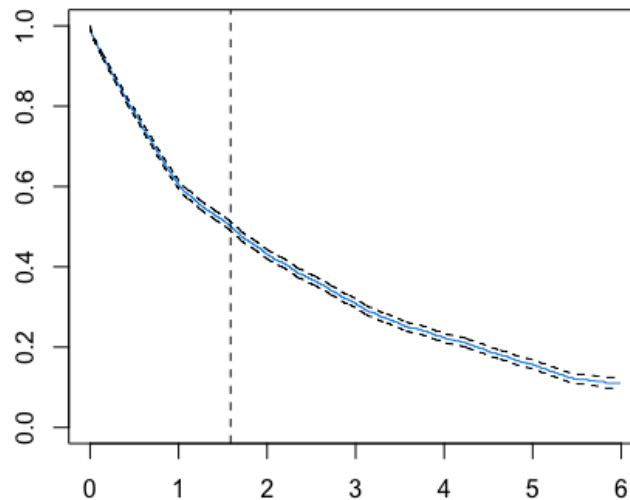
kusjuures $\hat{S}(t) = 1$ enne esimest surma või sündmuse toimumist ja $\hat{S}(t) = 0$ peale viimast surma või sündmuse toimumist. Saadud funktsioon $\hat{S}(t)$ on konstantne treppfunktsioon, mis väheneb ajas t iga surma või sündmuse toimumise korral.

Olgu nüüd andmestik n vaadeldavaga, kelle puhul on teada nende elukestused $t_1, \dots, t_n$ ja andmetes võib esineda ka tsenseerimist. Lisaks võivad mõned elukestused t_i ja t_j , $i, j \in [1, \dots, n]$ olla samad. Seega on andmestikus r sündmu-

se toimumise aega, nii et $r \leq n$. Järjestame sündmuse toimumise ajad nii, et $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. Tähistagu nüüd d_j , $j = 1, \dots, r$ vaadeldavaid, kes surevad ajahetkel $t_{(j)}$ ehk nende puhul toimub sündmus sel ajahetkel ja n_j neid vaadeldavaid, kes on vahetult enne ajahetke $t_{(j)}$ elus ehk indiviide, kellel on risk ajahetkel $t_{(j)}$ sündmust kogeda. Saame üleelamisfunktsiooni hinnangu leida kujul

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

kus $t_k \leq t < t_{(k+1)}$, $k = 1, \dots, r$. Saadud hinnangut nimetatakse Kaplan-Meieri hinnanguks.



Joonis 1: Kaplan-Meieri hinnang ravikulude hüvitamise pikkusele.

Joonisel 1 on näha Kaplan-Meieri hinnang üleelamistõenäosusele ravikulude hüvitamise kohta töös kasutatava andmestiku näitel. Sinise pidevjoonega on toodud hinnang ja mustade katkendjoontega hinnangu 95% usaldusintervall. Jooniselt saame näha tõenäosust, et ravikulude hüvitamine ei ole lõppenud aja t jooksul, mis

joonisel on esitatud aastates. Vertikaalse katkendjoonega on toodud mediaanaeg ehk hetk, mille toimumise ajaks on pooled elukestused lõppenud. Antud näite puhul on mediaanajaks veidi rohkem kui 1,5 aastat, mis tähendab, et pooled ravikulude hüvitamised lõppevad pooleteist aasta jooksul.

1.3 Coxi võrdeliste riskide mudel

Coxi võrdeliste riskide mudel on edasiarendus võrdeliste riskide mudelist nii, et hinnatakse parameetrid β , kuid baasriskifunktsiooni $h_0(t)$ ei määrata. Seetõttu nimetatakse Coxi võrdeliste riskide mudelit poolparameetriliseks meetodiks.

Olgu taas n vaadeldavat ja r sündmuse toimumise aega. Tähistame sündmuse toimumise ajad $t_{(1)} < t_{(2)} < \dots < t_{(r)}$. Tähistagu $R(t_{(j)})$ vaadeldavaid, kes ei ole vahetult enne ajahetke $t_{(j)}$ sündmust kogenud ehk neil on ajahetkel $t_{(j)}$ risk sündmust kogeda. Parameetrite $\beta = (\beta_1, \dots, \beta_p)'$ hindamiseks kasutatakse suurima tõepära meetodit, nii et

$$\mathbf{L}(\beta) = \prod_{j=1}^r \frac{\exp(\beta' \mathbf{x}_j)}{\sum_{l \in R(t_j)} \exp(\beta' \mathbf{x}_l)}. \quad (5)$$

Seda nimetatakse osalise tõepära funktsiooniks, kuna see kasutab väärtuste asemel nende astakuid. Võrrandi 5 lahendamiseks kasutatakse numbrilisi meetodeid, enamasti Newton-Raphsoni meetodit.

1.3.1 Riskide proportsionaalsuse eelduse kontrollimine

Coxi mudelil on oluline eeldus, et riskid peavad olema ajas võrdelised ehk proportsionaalsed. See tähendab, et kahe erineva vaadeldava A ja B riskide $h_a(t)$ ja $h_b(t)$ suhe on aja t jooksul samasugune.

Coxi mudeli eelduste kontrollimiseks on olemas erinevad jääkide arvutamisel põhinevad meetodid, näiteks Cox-Snelli jäägid, martingaalijäägid, jm. Siin töös kasu-

tame Schoenfeldi jääke, millega saab hinnata, kas sobitatud Coxi mudeli puhul on riskide võrdelisuse eeldus täidetud. Schoenfeldi jääkide meetodi puhul arvutatakse igale vaadeldavale nii mitu jääki, kui mitu kirjeldavat tunnust on sobitatud Coxi mudelis.

Olgu meil n vaadeldavat ja p kirjeldavat tunnust $X_1, \dots, X_p$. Leiame i -ndale vaatlusele, $i = 1, \dots, n$ ja j -ndale tunnusele, $j = 1, \dots, p$ vastava jäägi valemiga

$$r_{s_{ji}} = \delta_i \left\{ x_{ji} - \frac{\sum_{l \in R(t_i)} x_{jl} \exp(\hat{\beta}' \mathbf{x}_l)}{\sum_{l \in R(t_i)} \exp(\hat{\beta}' \mathbf{x}_l)} \right\},$$

kus δ_i tähistab sündmuse toimumise indikaatorit ja $R(t_i)$ inimeste hulka, kellel on ajahetkel t_i risk sündmust kogeda.

Schoenfeldi jääkide edasiarendus, kaalutud Schoenfeldi jäägid (Grambsch ja Therneau, 1994) on leitavad valemiga

$$r_{s_i}^* = d\text{var}(\hat{\beta}) r_{s_i},$$

kus $r_{s_i}^*$ on i -nda vaatluse Schoenfeldi jääkide vektor ja d tähistab sündmuste arvu n inimese seas. On leitud, et kaalutud Schoenfeldi jäägid toimivad jääkide hindamises efektiivsemalt, kui tavalised Schoenfeldi jäägid. Ka siin töös kasutatud funktsioon hindab jääke kaalutud Schoenfeldi jääkide kaudu.

Grambsch ja Thernau näitasid, et i -ndale vaatlusele ja j -ndale tunnusele vastava kaalutud Schoenfeldi jäägi $r_{s_{ji}}^*$ korral kehtib

$$E(r_{s_{ji}}^*) \approx \beta_j(t_i) - \hat{\beta}_j, \quad (6)$$

kus $\hat{\beta}_j$ on j -nda tunnuse hinnang sobitatud Coxi mudelist ja $\beta_j(t_i)$ on j -nda tunnuse ajast sõltuva koefitsiendi väärtus sel ajahetkel, kui toimub i -nda vaatluse sündmus. Siit tuleneb üks võimalus võrdelisuse eelduse kontrollimiseks, milleks on punktide $r_{s_{ji}}^* + \hat{\beta}_j$ graafiline esitus. Kui proportsionaalsuse eeldus on täidetud, on j -nda

tunnuse ajast sõltuva koefitsiendi $\beta_j(t)$ väärtus ajas konstantne ja vaadeldavad punktid asuvad horisontaalsel joonel.

Teine, enam levinud meetod proportsionaalsuse hindamiseks on *Grambsch ja Thernau* või ka *zph* test, mis põhineb samamoodi seosel (6). Kui j -s tunnus X_j on ajast lineaarselt sõltuv, saab seda sõltuvust kirjeldada seosega

$$\beta_j(t_i) = \beta_j + \nu(t_i - \bar{t}) \quad \Rightarrow \quad E(r_{s_{ji}}^*) = \nu(t_i - \bar{t}),$$

kus ν_j on regressiooni koefitsient. Kui tõus $\nu_j = 0$, on j -s tunnus ajast sõltumatu ja proportsionaalsuse eeldus täidetud. Olgu meil n vaadeldavat ja tähistagu $r_1, \dots, r_d$ ajahetki, kui toimus sündmus. Saame ajast sõltuvust hinnata teststatistikuga kujul

$$\frac{\{\sum_{i=1}^d (r_i - \bar{r}) r_{s_{ji}}^*\}^2}{d \text{var}(\hat{\beta}_j) \sum_{i=1}^d (r_i - \bar{r})^2},$$

kus $\bar{r} = \frac{1}{d} \sum_{i=1}^d r_i$ on toimunud sündmuste keskmine toimumise aeg. Nullhüpoteesi kehtivuse korral tingimusele, et tõus on 0, on see teststatistik χ^2 jaotusega. Proportsionaalsuse eeldus ei ole täidetud väga suurte teststatistiku väärtuste korral.

Tarkvaraprogrammis R saab seda testi läbi viia funktsiooniga *cox.zph* ja tulemust hinnata p-väärtuste abil. Suurte p-väärtuste korral on võrdelisuse eeldus täidetud.

1.3.2 Stratifitseeritud Coxi mudel ja ajas muutuvad koefitsiendid

Kui proportsionaalsuse eeldus ei ole täidetud, siis ei ole Coxi mudeli hinnangud usaldusväärsed. Üks võimalus mitte-proportsionaalsuse arvesse võtmiseks on stratifitseerimine. See tähendab, et tunnus, mille riskide suhe on ajas muutuv, jaotatakse väiksematesse osadesse ning leitakse siis Coxi hinnangud nendele osadele eraldi. Tüüpiliselt tähendab väiksematesse osadesse jaotamine seda, et kategoorilise tunnuse puhul hinnatakse kõigile tunnuse tasemetele j erinev baasrisk $h_{0j}(t)$. Stratifitseerimise puhul peab aga arvestama sellega, et stratifitseeritud

tunnus ei ole enam mudelis hinnatav parameeter. Stratifitseerimise miinus on see, et seda on keeruline kasutada pideva tunnuse või mitme tunnuse puhul.

Teine levinud meetod mitte-võrdeliste riskide korral Coxi mudeli kasutamiseks on ajas muutuvate koefitsientide hindamine (*time-varying coefficients*). See meetod toimib nii, et vaadeldav aeg jaotatakse erinevateks mittelõikuvateks ajavahemikeks ning seejärel leitakse parameetrite hinnangud igas lõigus eraldi. Kui vaadeldava ajalõigu jooksul on riskimäärade suhe proportsionaalne, on selles vahemikus Coxi võrdelisuse eeldus täidetud ja mudeli tulemus usaldusväärne. Ajas muutuvate koefitsientide kasutamise puhul peaks arvestama sellega, et mudeli tulemuse tõlgendamine võib olla keeruline. Kuna parameetrid hinnatakse igas ajalõigus eraldi, siis tuleb ka tulemusi arvestada ja tõlgendada ajalõikude põhiselt. Lisaks peab arvestama, et aja lõikepunktide valik võib analüüsi tulemust oluliselt mõjutada. (Zhang *et al.*, [2018](#))

Lisaks stratifitseerimisele ja ajas muutuvate koefitsientide hindamisele on võimalik riskide mitte-proportsionaalsust arvesse võtta ka pseudovaatluste meetodi abil. Seda meetodit kirjeldatakse põhjalikult järgmises peatükis.

2 Pseudovaatlused

Klassikalistel elukestusanalüüsi meetoditel on erinevaid puudujääke. Näiteks on elukestusandmete puhul masinõppe meetodite kasutamine piiratud. Eluskestusanalüüsi levinuim meetod, Coxi mudel sobib andmete analüüsimiseks hästi, kuid omab olulist eeldust, et riskid on proportsionaalsed, mis päriselu andmete puhul ei ole alati täidetud. Lisaks on tihtipeale päriselu uuringute, eriti epidemioloogiliste uuringute puhul vaja arvestada rohkem kui ühe riski ja sündmusega. Sellisel juhul kasutatakse konkureerivate riskide (*competing risks*) või mitme olekuga (*multi-state*) mudeleid, millel on omad puudujäägid. Nende probleemide lahendamiseks pakuvad Andersen, Klein ja Rosthøj (2003) ning Klein ja Andersen (2005) pseudovaatlusi. Pseudovaatluste põhimõte seisneb selles, et olemasolevad vaatlused, nii tsenseerimist sisaldavad kui ka mitte-sisaldavad, asendatakse pseudovaatlustega ning edasi analüüsitakse uusi, tsenseerimist mitte-sisaldavaid vaatlusi.

Peatükk 2 põhineb Anderseni ja Perme'i (2010) artiklil "*Pseudo-observations in survival analysis*", kui ei ole viidatud teisiti.

Pseudovaatlused (ka pseudo-observatsioonid või pseudoväärtused) on autori eestikeelne tõlge terminist *pseudo-observations*. Pseudovaatlused põhinevad jackknife meetodil, kus iga vaatluse kohta leitakse valimi teatud parameeter ilma üht vaatlust arvestamata. Sedasi saab näha ühe vaatluse mõju kogu valimile.

Olgu $T_i, i = 1, \dots, n$ juhuslik suurus, mis näitab elukestust. Kui kõik elukestused oleks teada ehk andmetes ei esineks tsenseerimist, saaks suvalise funktsiooni $f(\cdot)$ korral hinnata keskmist valemiga

$$E(f(T)) = \frac{1}{n} \sum_i^n f(T_i). \quad (7)$$

Eeldame nüüd, et andmetes esineb ka tsenseerimist ehk kõik elukestused T_i ja seega ka väärtused $f(T_i)$ ei ole teada, kuid leidub piisavalt lähedane hinnang $\hat{\theta}$ keskmisele $\theta = E(f(T))$ valemist (7). Sellisel juhul defineerime väärtuse $f(T_i), i = 1, \dots, n$

i -nda pseudovaatluse kui

$$\hat{\theta}_i = n \cdot \hat{\theta} - (n - 1) \cdot \hat{\theta}^{-i},$$

kus $\hat{\theta}^{-i}$ on hinnang, mis on leitud valimi põhjal, kust on välja jäetud i -s vaatlus.

Pseudovaatluste kasutamiseks peame seega leidma hinnangu $\hat{\theta}$, mis on piisavalt lähedane keskmisele $E(f(T))$.

2.1 Pseudovaatluste arvutamine

Järgnevalt leiame keskmisele $\theta = E(f(T))$ piisavalt lähedase hinnangu $\hat{\theta}$, millest saab arvutada pseudovaatlusi. Käsitleme kolme juhtu: hinnang üleelamistõenäosusele tavalise paremalt tsenseeritud andmestiku korral, tõkestatud keskmise (*restricted mean survival time*) korral ja konkureerivate riskide mudeli korral.

2.1.1 Hinnang üleelamistõenäosusele

Vaatame elukestust T ning üleelamisfunktsiooni $S(t)$. Tavalise paremalt tsenseeritud andmestiku korral saame leida üleelamisfunktsiooni väärtuse ajahetkel t_j , $j = 0, \dots, k$ valemiga

$$S(t_j) = E(I(T > t_j)),$$

kus I on indikaatorfunktsioon väärtustega 0 või 1. Siit saame võrduse

$$S(t) = E(I(T > t)) = E(f(T)) = \theta,$$

kus $f(T) = E(I(T > t))$ ja saame, et $\theta = S(t)$.

Järgmiseks võtame kasutusse tähistuse D_i , $i = 1, \dots, n$, mis näitab, kas i -s elukestus T_i on tsenseeritud ($D_i = 0$) või vaadeldav sündmus toimus ja väärtus on meile teada ($D_i = 1$). Tähistame elukestust $\tilde{X}_i$, nii et $\tilde{X}_i$ on võrdne tõelise elukestusega

T_i , kui sündmus toimus ja T_i on teada ($D_i = 1$) ning $\tilde{X}_i$ on võrdne tsenseerimise ajaga, kui tõeline elukestus ei ole teada ($D_i = 0$).

Nüüd leiame vahemikus $[0, t]$ toimunud sündmuste arvu

$$N(t) = \sum_i^n I(\tilde{X}_i \leq t, D_i = 1)$$

ja objektide arvu, mille korral ajahetkel t ei ole sündmust veel toimunud ehk neil on risk kogeda sündmust ajahetkel t

$$Y(t) = \sum_i^n I(\tilde{X}_i \geq t).$$

Huvipakkuva suuruse $\theta = S(t)$ hinnanguks sobib Kaplan-Meieri hinnang

$$\hat{S}(t) = \prod_{u \leq t} \left(1 - \frac{dN(u)}{Y(u)}\right). \quad (8)$$

Mitte-informatiivse tsenseerimise korral on saadud hinnang nihketa suurusele $S(t)$ ehk oleme leidnud otsitava hinnangu $\hat{\theta} = \hat{S}(t)$ (Andersen *et al.*, 1993). Kuna Kaplan-Meieri hinnang on pidev treppfunktsioon, mille väärtus muutub iga sündmuse toimumise aja $t_{(j)}$ korral, siis on ka läbi selle hinnangu arvutatud pseudovaatlused ajas pidevad väärtused, mis muutuvad iga sündmuse toimumise aja $t_{(j)}$ korral. Ajahetkel $t_{(j)}$, $j = 0, \dots, t$ saame i -nda pseudovaatluse leida valemiga

$$\hat{S}_i(t_{(j)}) = n \cdot \hat{S}(t_{(j)}) - (n-1) \cdot \hat{S}^{-i}(t_{(j)}).$$

2.1.2 Tõkestatud keskmine

Tõkestatud keskmise mudelit tuleks kasutada siis, kui leidub fikseeritud väärtus $\tau > 0$, millest suurem ei saa uuritav elukestus olla. Need mudelid on sarnased eelnevas peatükis kirjeldatud mudelitele, kuid tuleb arvesse võtta seda, et elukestus T_i ei saa olla suurem kui τ . (Andersen ja Klein, 2007)

Vaatame elukestust T , fikseeritud $\tau > 0$ ja τ -tõkestatud keskmist valemiga

$$\mu_\tau = E(T \wedge \tau),$$

kus $T \wedge \tau$ tähistab miinimumi T -st ja τ -st. Saame kirjutada

$$\mu_\tau = E(T \wedge \tau) = E(f_\tau(T)) = E(f(T)) = \theta,$$

kust saame, et $\mu_\tau = \theta$.

Kuna kehtib, et

$$T \wedge \tau = \int_0^{T \wedge \tau} 1 dt = \int_0^\tau I(T > t) dt,$$

siis saame θ leida läbi üleelamistõenäosuse

$$\theta = \mu_\tau = E(T \wedge \tau) = \int_0^\tau E(I(T > t)) dt = \int_0^\tau S(t) dt.$$

Tõkestatud keskmise korral avaldub nihketa hinnang keskmisele $\theta = E(T \wedge \tau)$ integraalina Kaplan-Meieri hinnangust ehk

$$\hat{\mu}_\tau = \int_0^\tau \hat{S}(u) du.$$

Sellisel juhul saab i -nda pseudovaatluse ajahetkel τ leida valemiga

$$\hat{\mu}_{\tau i} = n \int_0^\tau \hat{S}(t) dt - (n-1) \int_0^\tau \hat{S}^{-i}(t) dt.$$

2.1.3 Konkureerivate riskide mudel

Konkureerivate riskide mudeli puhul on vaadeldaval inimesel või objektil risk rohkem kui ühe sündmuse toimumiseks. Sellisel juhul on lisaks elukestusele oluline ka see, mis sündmuse tõttu elukestus lõppes. (Andersen, Klein ja Rosthøj, 2003)

Järgmiseks vaatame konkureerivate riskide mudelit, kus on kaks võimalikku sünd-

must A ja B ning ajahetkel t on seega kolm võimalikku olekut: 0 - vaadeldav on elus ajahetkel t , 1 - vaadeldav koges sündmust A enne hetke t , 2 - vaadeldav koges sündmust B enne hetke t . Olgu tõenäosus sündmuse A kogemiseks $\alpha_1(t)$ ja sündmuse B kogemiseks $\alpha_2(t)$ ning tähistagu D seda, kas toimus sündmus A ($D = 1$) või sündmus B ($D = 2$). Saame leida esimese sündmuse kumulatiivse avaldumise $C_1(t)$ valemiga

$$C_1(t) = E(I(T \leq t, D = 1)) = \int_0^t S(u-) \alpha_1(u) du,$$

kus T on elukestus ja $S(u) = P(T > u)$ tähistab üldist üleelamistõenäosust $P(T(u) = 0)$ ehk tõenäosust, et ajahetkel u ei ole juhtunud kumbagi sündmustest A ja B . Saame siit sündmuse A ($D = 1$) jaoks kirjutada võrduse

$$C_1(t) = E(I(T(t) = 1)) = E(f_t(T(\cdot))) = E(f(T(\cdot))) = \theta$$

ehk $\theta = C_1(t)$.

Leiame hinnangu $\hat{\theta}$, millega hinnata keskmist $C_1(t)$. Vaatame elukestusi $(\hat{X}_i, D_i)$, $i = 1, \dots, n$, kus $\hat{X}_i$ tähistab elukestust, mis võib olla teada või tsenseeritud (nagu peatükis 2.1.1) ja D_i tähistab sündmuse A ($D_i = 1$) või B ($D_i = 2$) toimumist ning $D_i = 0$ tähistab tsenseerimist (vaatlus ei saa olla tsenseeritud, kui on toimunud sündmus A või B). Saame vahemikus $[0, t]$ toimunud sündmuse j arvu loendada valemiga

$$N_j(t) = \sum_i I(\hat{X}_i \leq t, D_i = j), \quad j = 1, 2.$$

$Y(t)$ näitab ajahetkel t ühtegi sündmust mitte kogenud vaadeldavate arvu ehk neid, kellel on ajahetkel t risk kogada mõnda sündmust. Saame $Y(t)$ leida samamoodi nagu peatükis 2.1.1 valemiga

$$Y(t) = \sum_i I(\tilde{X}_i \geq t).$$

Sündmuse j kumulatiivse avaldumise funktsioon on hinnatav Aalen-Johanseni hinnanguga

$$\hat{C}_j(t) = \int_0^t \hat{S}(u-) d\hat{A}_j(u),$$

kus $\hat{A}_j(t)$ valemiga

$$\hat{A}_j(t) = \int_0^t \frac{N_j(u)}{Y(u)} du$$

on Nelson-Aaleni hinnang kumulatiivsele riskile sündmuse j toimumiseks

$$A_j(t) = \int_0^t \alpha_j(u) du.$$

Konkureerivate riskide mudeli puhul saame i -nda pseudovaatluse sündmuse j korral leida valemiga

$$\hat{C}_{ji}(t) = \int_0^t \hat{S}_i(u-) d\hat{A}_j(u).$$

Konkureerivate riskide mudel ja mitme olekuga mudel on komplekssemad ja ei vasta selles töös kasutatavatele andmetele, nii et nende uurimine pseudovaatluste abil ei ole selle töö eesmärk. Küll aga tasub märkida, et just nende mudelite puhul on pseudoväärtuste kasutamine andnud märgatavalt häid tulemusi teiste meetodite ees (Sachs *et al.*, 2019; Klein ja Andersen, 2005).

2.2 Mudelite sobitamine

Tsenseerimata andmete analüüsimiseks on olemas suur valik nii klassikalisi, kui ka masinõppe mudeleid. Tsenseeritud andmete korral on aga mudelite valik enamasti piiratud poolparameetrilise Coxi võrdeliste riskide mudeliga. Pseudovaatluste kasutamise eelis on see, et saame tsenseerimisest andmetes vabaneda, mis omakorda võimaldab kasutada laiemat mudelite valikut. Pseudovaatluste omapärast tulenevalt ei saa me kasutada klassikalisi *GLM* mudeleid. Näiteks üleelamistõenäosuse hindamise korral vastab andmestikus ühele uuritavale mitu rida, täpsemalt iga arvutatud pseudovaatluse kohta üks rida. Seega peame kasutama meetodeid, mis ar-

vestavad vaatluste omavahelist seost. Pseudovaatluste mudeldamiseks on Andersen ja Perme (2010) välja pakkunud üldistatud hinnanguvõrrandite (*GEE*, *generalized estimating equations*) kasutamise.

Üks võimalus korduvate vaatluste või klastrite analüüsimiseks on mitmetasemelised lineaarsed mudelid (*GLMM*, *generalized linear mixed models*). *GEE* ja *GLMM* mudelite põhimõtteline erinevus on see, et *GEE* mudeliga leitakse üldine populatsiooni keskmine võttes arvesse korduvaid vaatlusi, kuid *GLMM* mudeli puhul keskendutakse ka objektipõhisele mõjule. *GEE* mudelite tööpõhimõte on lihtsam kui *GLMM* oma, näiteks ei kasuta *GEE* meetod tõepära hindamist. (Ford, 2021)

Üldistatud hinnanguvõrrandite puhul hinnatakse uuritava suuruse $T = (T_1, \dots, T_m)$, kus m võib olla objektiti erinev, puhul keskmine $\mu = E(T)$ ning variatsioon $\nu(\mu)$, mis näitab, kuidas $var(T)$ sõltub keskmisest μ . Seejuures ei hinnata suuruse T jaotust ja ei täpsustata tõepära hinnangut. See tähendab, et *GEE* mudeli puhul ei saa hinnata mudeli sobivust, kuid saab leida parameetrite hinnangud $\hat{\beta}$ koos standardvigadega $se(\hat{\beta})$. (Agresti, 2002)

Korduvate vaatluste mõju arvesse võtmiseks tuleb hinnata ka sama vaatluse erinevate korduste vahelist kovariatsiooni. Selleks hinnatakse *GEE* mudelis vaatluste $(T_1, \dots, T_m)$ variatsiooni ja paarikaupa korrelatsiooni. Seejuures tasub märkida, et *GEE* mudeli hinnangud on usaldusväärsed isegi siis, kui korrelatsioonistruktuur on halvasti määratud. (Agresti, 2002)

2.2.1 Riskimäärade suhte hindamine üleelamistõenäosusele

Regressioonanalüüsi üldkuju on

$$g(E(f(T)|Z)) = \beta_0 + \sum \beta_j Z_j,$$

kus g on sobiv linkfunktsioon, $\theta = E(f(T))$ on vaadeldav suurus ja $Z = (Z_1, \dots, Z_p)$ on kovariaadid, mille seost suurusega θ me tahame leida.

Vaatame mudelit ajahetkel t , kus $f(T) = I(T > t)$ ehk $\theta = S(t)$ ja $g(x) = \text{cloglog}(x) = \log(-\log(1-x))$ ehk linkfunktsiooniks on täiend-log-log funktsioon. Saame mudeli

$$\log(-\log(S(t))) = \log A_0(t) + \sum_j \beta_j Z_j, \quad t > 0.$$

Tegemist on Coxi regressioonmudeliga, kus $A_0(t) = \int_0^t h_0(u) du$ on kumulatiivne baasrisk Coxi mudelist

$$h(t|Z) = h_0(t) \exp\left(\sum_j \beta_j Z_j\right).$$

Linkfunktsiooniks võib lisaks täiend-log-log funktsioonile olla ka näiteks *identity* link $g(x) = x$, *logit* link logistilise regressiooni jaoks $g(x) = \log(\frac{x}{1-x})$, jne. Järgmiseks vaatame, kuidas hinnata parameetreid β GEE mudeli põhjal.

Selleks, et saaksime mudelites arvesse võtta tsenseerimist, saame kasutada pseudovaatlusi. Pseudovaatlused arvutatakse ajapunktide $t_1, \dots, t_k$ korral, kus k võib maksimaalselt olla võrdne toimunud sündmuste arvuga. Sellisel juhul arvutatakse pseudovaatlused iga sündmuse toimumise ajal, kuna nende väärtus muutub iga sündmuse toimumise ajahetkel. Anderseni ja Perme'i (2010) kohaselt piisab tegelikult ka vähematest, 5-10 ajas võrdselt jaotatud ajapunktist. Graw *et al.* (2009) aga leiavad, et mõnel juhul ei pruugi 10 ajapunktist piisata ja pseudovaatluste arvutamise periood sõltub andmetest.

Olgu i -nda objekti pseudovaatluste väärtused $\hat{\theta}_i = (\hat{\theta}_{ij}, j = 1, \dots, k)$ ehk i -nda vaatluse igal ajahetkel $t_1, \dots, t_k$ arvutatud pseudovaatlused. Sellisel juhul $\hat{\theta}_{ij} = f_{t_j}(T_i)$ ja vaadeldav mudel on

$$g(E(f_{t_j}(T_i)|Z_i)) = \beta' Z_{ij}^*,$$

kus $Z_{ij}^* = (I(t_l = t_j), l = 1, \dots, k; Z_i)$ sisaldab infot nii ajapunktide kui kovariaa-

tide kohta ja g on valitud linkfunktsioon. Saame leida parameetrite β hinnangud valemist

$$\sum_i \left(\frac{\delta}{\delta\beta} g^{-1}(\beta' Z_{ij}^*) \right)' V_i^{-1} (\hat{\theta}_i - g^{-1}(\beta' Z_{ij}^*)) = \sum_i U_i(\beta) = U(\beta) = 0,$$

kus V_i on pseudovaatlusele $\hat{\theta}_i$ vastav *working covariance matrix* kirjeldades ühe objekti pseudovaatluste omavahelist korrelatsiooni.

Tähistagu

$$I(\beta) = \sum_i \left(\frac{\delta g^{-1}(\beta' Z_{ij}^*)}{\delta\beta} \right)' V_i^{-1} \left(\frac{\delta g^{-1}(\beta' Z_{ij}^*)}{\delta\beta} \right)$$

ja

$$\widehat{var}(U(\beta)) = \sum_i U_i(\hat{\beta})' U_i(\hat{\beta}).$$

Saame $\hat{\beta}$ dispersiooni hinnata valemiga

$$\widehat{var}(\hat{\beta}) = I(\hat{\beta})^{-1} \widehat{var}(U(\beta)) (I(\hat{\beta})^{-1}),$$

mida nimetatakse *sandwich* hinnanguks. Selle hinnangu kasutamise eeliseks on, et see hindab $\hat{\beta}$ variatsiooni asümptootiliselt hästi ka siis, kui hinnatud $var(T)$ ei ole täpne. (Agresti, 2002)

Järgmistes peatükkides selgitatakse, kuidas pseudovaatlusi praktikas kasutada ja arvutada tarkvaras R.

2.3 Pseudovaatlused R-is

Andmete analüüsimiseks ja pseudovaatluste arvutamiseks kasutati siin töös tarkvarapaketti R. Pseudovaatluste arvutamiseks on R-is loodud pakett *pseudo*. Põhjalikuma tutvustuse pakettis leiduvatest funktsioonidest ja nende algoritmist koos

konkreetsete näidetega leiab (Klein *et al.*, 2008). Tasub märkida, et see artikkel on kirjutatud aastal 2008 ning sellest ajast saati on funktsioonide väljund veidi muutunud (Maja Pohar Perme ja Gerster, 2017). Funktsioonide arvutuspõhimõtted on jäänud siiski samaks.

Nagu Anderseni ja Perme'i (2010) artiklis on toodud kolm pseudovaatluste juhtumit, on ka R-i paketi *pseudo* kolm funktsiooni nendele juhtudele vastavate pseudovaatluste arvutamiseks. Tavalisele hinnangule üleelamistõenäosusele ehk siin töös kasutatavale juhule vastab funktsioon *pseudosurv*, tõkestatud keskmise juhule *pseudomean* ja konkureerivate riskide mudelile *pseudoci*. Kõigi kolme funktsiooni ülesehitus on sarnane. Käsitleme siin põhjalikumalt esimest funktsiooni *pseudosurv*, mida kasutati analüüsiks siin töös.

Funktsioon *pseudosurv(time, event, tmax)* kasutab sisendiks kolme argumenti. Argument *time* on uuritava suuruse elukestus ehk siin töös hüvitamise pikkus päevades, *event* tähistab tsenseerimist samamoodi nagu muude elukestusanalüüsi funktsioonide korral ja *tmax* tähistab ajapunkti või -punkte, mille korral pseudovaatlused arvutatakse. Kui argument *tmax* tühjaks jätta, arvutab funktsioon pseudovaatlused iga ajapunkti *time* korral. Pärast arvutamist on funktsiooni väljundiks kaks listi, millest *\$time* tähistab ajapunkte, mille korral pseudovaatlused arvutati, ning teises listis *\$pseudo* on arvutatud pseudoväärtused.

Tõkestatud keskmise korral saab samuti ette anda erinevad ajahetked, mille korral pseudovaatlusi arvutada, kusjuures tõkestatud ajaks τ määrab programm suurima antud aegadest. Kui *tmax* tühjaks jätta, võtab programm automaatselt tõkestatud ajaks τ kõige suurema vaadeldud kestuse ehk suurima väärtuse argumendist *time*. Sellisel juhul väljastab funktsioon iga vaatluse kohta vaid ühe pseudovaatluse, mis on leitud ajahetkel τ ehk hetkel, millest suurem ei saa ükski elukestus olla.

Kui pseudovaatlused on arvutatud, tuleb esialgne andmestik ja pseudovaatlused ühendada pikaks andmestikuks, kus iga vaatluse kohta on nii mitu rida, kui mitu pseudovaatlust ühele vaatlusele arvutati. Saadud andmestiku peal saab läbi viia

GEE mudeli sobitamist, nii et mudel võtab iga vaatluse jaoks arvesse pseudovaatluste muutust ajas. *GEE* mudeli sobitamist saab teha paketi *geepack* funktsiooniga *geese*, mis toimib sarnaselt *glm* funktsioonile. Funktsioonis tuleb määrata uuritav suurus ja kovariaadid, mille mõju me uuritavale suurusele leida tahame. Argument *mean.link* määrab linkfunktsiooni, mida arvutamiseks kasutada. Regressioonanalüüsi läbi viimiseks sobib linkfunktsioon "*identity*", mis on ka vaikimisi valitud funktsioon. Kui *mean.link* väärtuseks määrata "*cloglog*" saab läbi viia Coxi regressioonanalüüsi.

2.4 Pseusovaatlused simulatsiooni näitel

Järgmiseks tehti lihtne simulatsioonandmestik, mille peal pseudovaatluste arvutamist ja kasutamist selgitada.

2.4.1 Näide ilma tsenseerimiseta

Vaatame esmalt näidet, kus andmestikus ei esine tsenseerimist ehk kõik elukestused T_i on teada. Näite jaoks simuleeritud andmestikus on 20 vaatlust ja 19 sündmuse toimumise aega ehk kahe vaadeldava puhul toimus sündmus samal ajal. Sündmuse toimumise ajad päevades on toodud tabelis 1.

Tabel 1: Ajahetked, kus toimus sündmus

27	30	38	187	239	240	247	251	275	339
392	408	453	636	806	1036	1059	1094	1686	

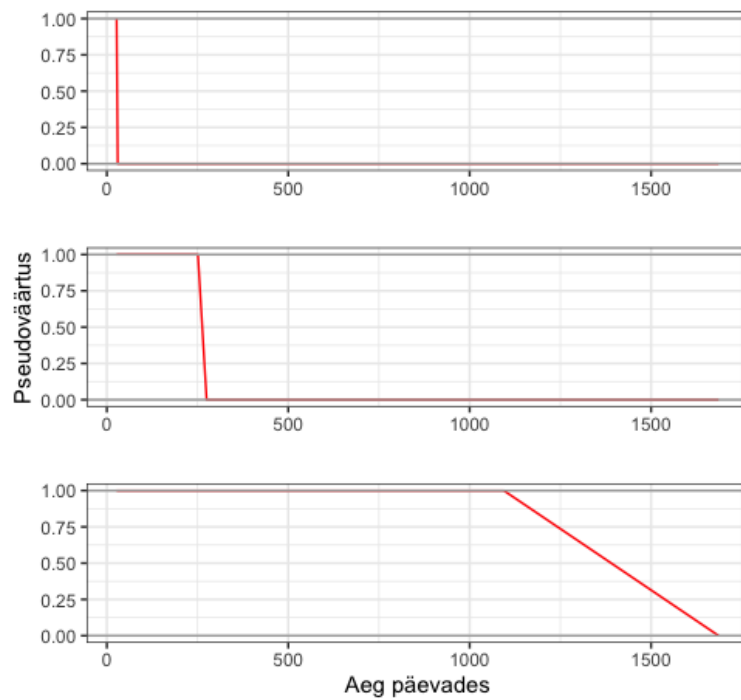
Järgmiseks arvutame kõigile 20 vaadeldavale pseudovaatlused kõigil sündmuste toimumise ajahetkedel. Kuna pseudovaatlused on arvutatud Kaplan-Meieri hinnangu põhjal, mis muutub iga sündmuse toimumise ajal ja on pidev vahepeal, siis kehtib sama ka pseudovaatluste puhul. Tabel 2 näitab ühe vaadeldava pseudovaatluste väärtusi.

Tabel 2: Ühe vaadeldava pseudovaatlused ajahetkedel t

Ajahetk	27	30	38	187	239	240	247	251	275
Pseudovaatlus	1	1	1	1	1	1	1	1	-1,776e-15
339	392	408	453	636	806	1036	1059	1094	1686
0	-8,882e-16	0	0	-8,882e-16	-4,441e-16	0	-2,220e-16	-1,110e-16	0

Selle vaadeldava puhul on elukestuse T pikkus 275 päeva. Seda on näha ka tabelist 2, kus hetkeni 275 on pseudovaatluste väärtus 1 ning peale seda 0 või selle lähedane. Tuletame meelde, et vaatame praegu andmestikku, kus ei esine tsenseerimist ja kõik elukestused on teada. Sellisel juhul taandub pseudovaatluste arvutamine tavalise elukestuse $S(t)$ hindamisele.

Vaatame graafiliselt kolme vaadeldava pseudovaatlusi.



Joonis 2: Kolme vaadeldava pseudovaatlused aja t jooksul.

Tabelis 1 ja 2 toodud vaadeldav on joonisel 2 näha keskmisel graafikul. Ülemisel graafikul on näha vaadeldav, kelle elukestus T on 30 päeva. Näeme, et tema elukestus langeb kiiresti nulli ajahetkede 27 ja 30 vahel ning on 0 või selle lähedane

kõikide järgnevate ajahetkede jooksul. Keskmisel graafikul näeme graafilist esitust tabelist 2, kus pseudovaatlused langevad nulli ajahetkeks 275. Alumisel graafikul on näha vaadeldav, kelle elukestus lõpeb viimasel vaadeldaval ajahetkel 1686. Tema pseudovaatluste väärtus on 1 kuni eelviimase ajahetkeni ja langeb nulli ajahetkeks 1686. Kuna viimase kahe ajahetke 1094 ja 1686 vahe on suur, siis on alumise graafiku langus sujuvam, kui kahe ülemise graafiku oma, kus vaadeldavad ajahetked on tihedamad. Epidemioloogiliste uuringute puhul saab sellest mõelda, kui *follow-up* ajast. Mida tihedamini me vaadeldava või patsiendi seidundit kontrollime, seda täpsemini saame määrata sündmuse toimumise aja.

Vaadatud näited olid illustratiivsed pseudovaatluste kasutamise kohta. Kuna pseudovaatluste kasutamise headus seisneb just tsenseerimise arvesse võtmises, siis vaatame järgmiseks näidet simulatsioonandmestikust, kus esineb ka tsenseerimist.

2.4.2 Näide tsenseerimisega

Vaatame nüüd andmestikku, kus on taas 20 vaadeldavat, kuid andmetes esineb ka tsenseerimist. Siin simulatsioonis on 15 vaatlust, kus sündmus toimus, ja 5 tsenseeritud vaatlust. Tabelis 3 on esitatud 15 sündmuse toimumise aega ja 5 aega, kui toimus vaatluse tsenseerimine (esile tõstetud kollase taustaga).

Tabel 3: Ajahetked, kus toimus sündmus

10	37	56	63	86	268	284	312	315	333
364	435	442	506	622	633	738	1069	1117	1197

Arvutame kõigile vaadeldavatele pseudovaatlused igal sündmuse toimumise hetkel. Kuna pseudovaatluste väärtus muutub iga sündmuse toimumise ajal, kuid mitte tsenseerimise ajal, siis saame iga vaatluse kohta 15 pseudovaatlust. Tabelis 4 on toodud näide vaatlusest A , mille elukestus on 622 päeva.

Näeme, et pseudovaatlused on 1 ja rohkem kuni ajahetkeni 622 ehk vaadeldud elukestuseni. Sel hetkel muutuvad pseudovaatlused negatiivseks ja jäävadki negatiivseks, lähenedes aja suurenedes nullile.

Tabel 4: Vaatluse A pseudovaatluste väärtused.

Ajahetk	10	37	56	63	86	312	315
Pseudovaatlus	1	1	1	1	1	1,013	1,026
364	435	506	622	633	738	1069	1117
1,053	1,080	1,142	-0,309	-0,247	-0,185	-0,124	-0,062

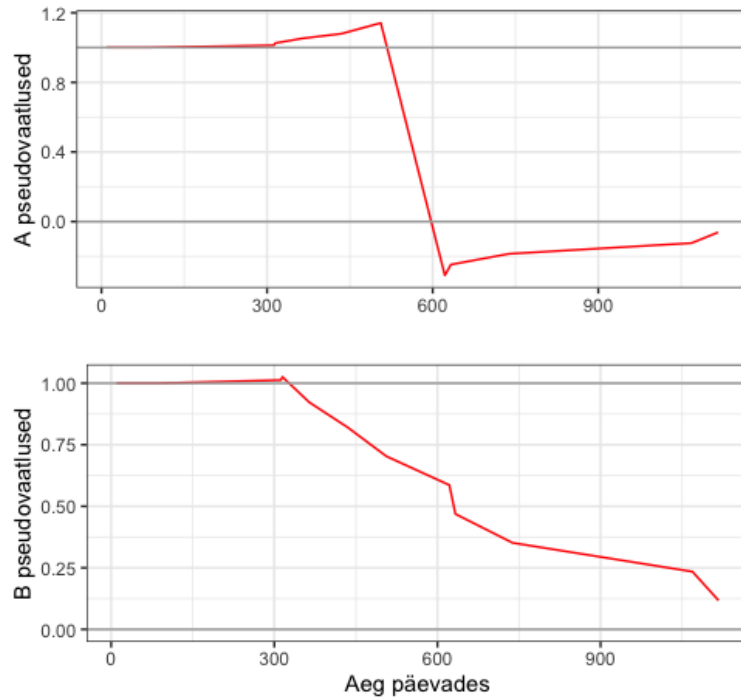
Järgmiseks vaatame vaatlust B , mille elukestus ei ole teada ehk vaatlus on tsenseeritud ja tsenseerimine toimus ajahetkel 333.

Tabel 5: Vaatluse B pseudovaatluste väärtused.

Ajahetk	10	37	56	63	86	312	315
Pseudovaatlus	1	1	1	1	1	1,013	1,026
364	435	506	622	633	738	1069	1117
0,923	0,821	0,703	0,586	0,469	0,352	0,234	0,117

Tabelis 5 on näha vaatluse B pseudovaatlused. Näeme, et pseudovaatlused on 1 ja rohkem kuni hakkavad vähenema ja langevad alla 1 ajahetkede 315 ja 364 vahel ehk vaatluse tsenseerimise ajal. Pärast tsenseerimist hakkavad pseudovaatlused järjest vähenema, kuid ei lähe ühelgi ajahetkel negatiivseks nagu mitte-tsenseeritud vaatluse korral.

Vaatame ka mõlema vaatluse pseudovaatluste muutust ajas graafiliselt. Ka joonisel 3 on näha, et vaatluse A korral muutub pseudovaatluste väärtus negatiivseks ajahetkeks 633 ning läheneb nullile sealt edasi. Vaatluse B korral on aga näha langemist alla 1 peale ajahetke 315 (viimane hetk enne tsenseerimist) ning pseudovaatluste vähenemist ajas.

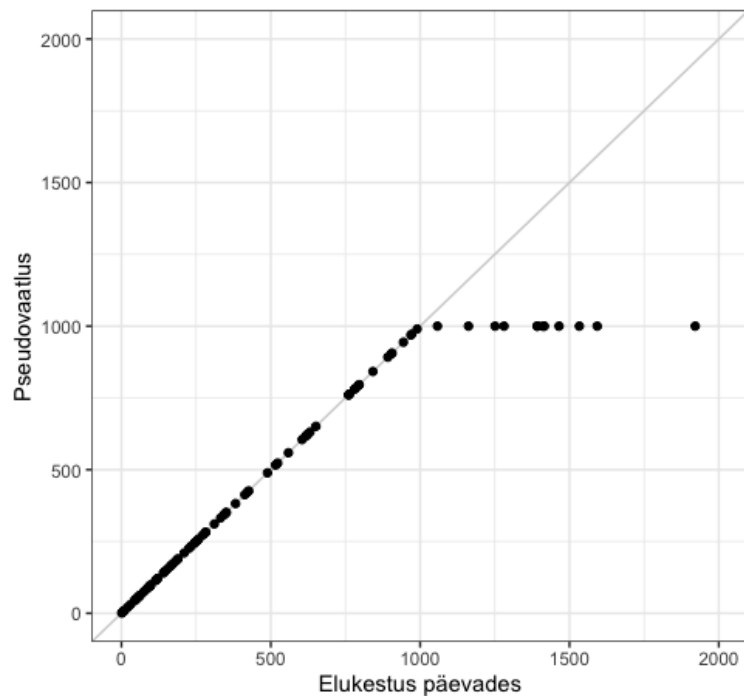


Joonis 3: Vaatluse A ja B pseudovaatlused aja t jooksul.

2.4.3 Näide tõkestatud keskmisest

Kolmanda näitena vaatame tõkestatud keskmise juhtu. Simuleerime andmestiku, kus on 100 vaadeldavat ja andmestikus ei esine tsenseerimist ehk kõik elukestused T_i on teada. Võtame tõkestatud keskmiseks $\tau = 1000$ päeva. Kui andmestikus ei esine tsenseerimist, siis taandub pseudovaatluste arvutamine lihtsale arvutusele, kus kõik elukestused, mis on suuremad kui $\tau = 1000$ asendatakse τ väärtusega.

Joonisel 4 on näha vaadeldavate tõelised elukestused ja pseudovaatlused. Kuna andmetes puudub tsenseerimine, on kõik pseudovaatluste väärtused samad, mis tõelised elukestused kuni τ väärtuseni. Sealt edasi on pseudovaatluste väärtus $\tau = 1000$, kuid tõelised elukestused võivad ka sellest suuremad olla.



Joonis 4: Elukestused ja pseudovaatlused ilma tsenseerimiseta tõkestatud keskmise $\tau = 1000$ korral.

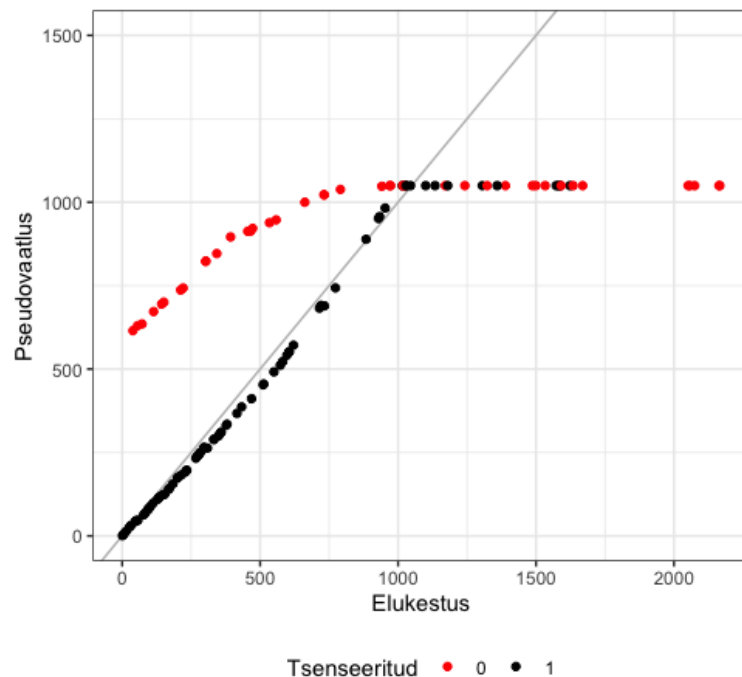
Vaatame nüüd tõkestatud keskmise juhtu, kus andmetes esineb ka tsenseerimist. Simuleerime taaskord 100 vaadeldavat, nii et osad vaatlused on tsenseeritud. Võtame taaskord tõkestatud keskmiseks $\tau = 1000$ päeva ja arvutame pseudovaatlused sel ajahetkel. Tõkestatud keskmise korral arvutatakse pseudovaatlus vaid ühel ajahetkel, milleks on τ .

Tabel 6: 8 näidet vaadeldavate elukestusest ja arvutatud pseudovaatlustest

Vaatlus	Tsenseeritud	Vaadeldud elukestus	Pseudovaatlus
1	1	324	251,53
2	1	853	831,25
3	0	604	974,48
4	0	253	764,83
5	1	1583	1087,50
6	1	1888	1087,50
7	0	2011	1087,50
8	0	1289	1087,50

Tablis 6 on toodud näide mõnest vaadeldavast. Vaatlused 1 ja 2 on mitte-tsenseeritud ehk nende tegelik elukestus on teada ja see on väiksem kui 1000. Näeme, et nende vaatluste puhul on pseudovaatlused väiksemad kui tegelik elukestus. Vaatlused 3 ja 4 on tsenseeritud, nii et nende tsenseerimise aeg ehk vaadeldud elukestus on väiksem kui 1000. Nende vaatluste puhul on näha, et pseudovaatlused on vaadeldud elukestusest märgatavalt suuremad. Pseudovaatlused võtavad arvesse seda, et nende vaatluste tegelik elukestus ei olnud tsenseerimise ajaks lõppenud ja on suurem, kui vaadeldud elukestus. Vaatlused 5-8 on kõik suurema vaadeldava elukestusega kui 1000. Näeme, et sellisel juhul ei ole vahet, kas vaatlus on tsenseeritud või mitte, sest mõlemal juhul on pseudovaatluste väärtuseks 1087,5.

Vaatame seda andmestikku ka graafiliselt.



Joonis 5: Elukestused ja pseudovaatlused tõkestatud keskmise $\tau = 1000$ korral, kui andmetes esineb tsenseerimist.

Joonisel 5 tähistavad mustad punktid vaatlusi, mille tegelik elukestus on teada, ja punased punktid vaatlusi, mis on tsenseeritud. Näeme, et teada olevate elukestuste

puhul, mis on väiksemad kui τ , on pseudovaatluste väärtus veidi väiksem kui tegelik elukestus. Tsenseeritud vaatluste puhul on aga pseudovaatluste väärtus suurem, kui vaadeldud aeg tsenseerimiseni. Nii tsenseeritud kui mitte-tsenseeritud vaatluste, mis on suuremad kui τ , pseudovaatluse väärtus on 1085,5.

Peatükis 2 anti ülevaade pseudovaatlustest ja nende kasutamisest. Järgmises peatükis kasutame pseudovaatlusi, lisaks teistele meetoditele, et analüüsida päriselu andmeid.

3 LKFi andmete näitel

Siin peatükis räägime LKF-i andmete peal tehtud analüüsist. Esmalt kirjeldame andmete struktuuri ja seejärel tulemusi, mis me saime kasutades eelnevates peatükkides kirjeldatud analüüsimeetodeid.

3.1 Andmete kirjeldus

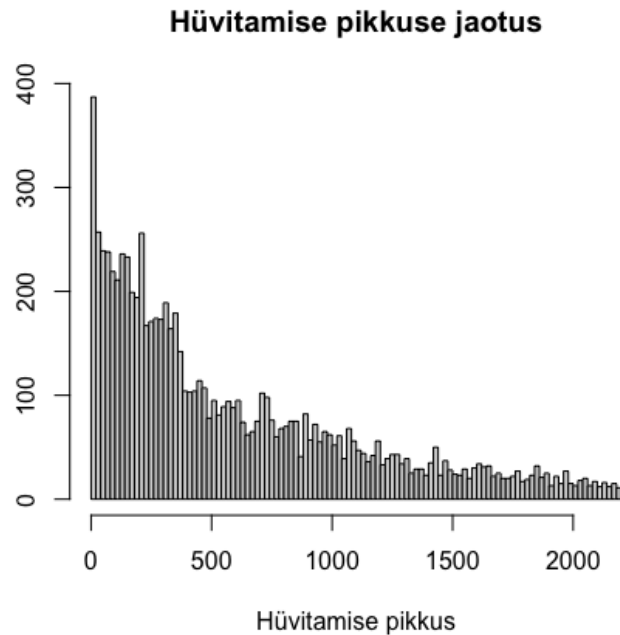
Töös kasutatavad andmed on pärit Eesti Liikluskindlustuse Fondist. Andmed on inimeste kohta, kes vajavad liiklusõnnetuse tagajärjel ravi, mille kulusid hüvitab liikluskindlustus. Andmestikus on juhtumid 2016. aasta algusest kuni 2021. aasta lõpuni ja andmed on 8260 hüvitamise juhtumi kohta.

Andmestikus on juhtumeid, mille hüvitamine ei ole 31.12.2021 seisuga lõppenud ehk käsitlemise lõpu kuupäev on puudu. Need on juhtumid, mis selles uuringus on tsenseeritud ehk hüvitamise kogupikkus ei ole teada. Tsenseeritud vaatlusi on kokku 2743 ehk umbes 33% juhtumitest on tsenseeritud.

Andmestikus on toodud hüvitamise algus- ja lõpukuupäev, mille põhjal on arvutatud hüvitamise pikkus päevades. Tsenseeritud vaatluste puhul on hüvitamise pikkuseks aeg hüvitamise algusest tsenseerimise ajani ehk kuupäevani 31.12.2021. Kõige pikem vaadeldud elukestus on 2188 päeva ehk peaaegu 6 aastat. Hüvitamise pikkuse jaotus on esitatud joonisel [6](#).

Lisaks hüvitamisele on teada ka õnnetuse toimumise kuupäev. Sellest kuupäevast tekitame tunnuse *aastaaeg*, mis näitab aastaaega, kuna õnnetus toimus.

Andmestikus on juhud nii Eestis kui välismaal toimunud juhtumite kohta. Suurem enamus (92%) juhtumitest toimus Eestis. Välismaal toimunud õnnetused on toimunud põhiliselt Euroopas, kõige enam Soomes. Andmete analüüsimiseks grpeerime õnnetuse asukoha ümber kahe väärtusega tunnuseks *asukoht*, mille tasemed on "Eesti" või "välisriik".



Joonis 6: Ravikulude hüvitamise kestuse jaotus.

Andmestikus on info õnnetuse põhjustanud sõiduki kohta ja kannatanu sõiduki kohta. Teada on sõiduki kategooria, nt sõiduauto, veoauto, mootorratas. Grupeerime need tunnused nii, et õnnetuse põhjustanud sõidukil on neli gruppi:

- "sõiduauto, väikeveok",
- "motikas"- mootorrattad ja mopeedid,
- "suursõiduk"- buss, veoauto, traktor, haagis,
- "teadmata"

ja kannatanu sõidukil on kuus gruppi:

- "jalakäija",
- "sõiduauto, väikeveok",

- "motikas"- mootorrattad ja mopeedid,
- "suursõiduk"- buss, veoauto, traktor, haagis,
- "mitu",
- "teadmata".

Hüvitise kohta on teada hüvitamise kogukahju suurus eurodes 01.04.2022 seisuga. Tsenseeritud vaatluste puhul arvestatakse kogukahju suuruseks juba välja makstud kahjusumma ja eraldise summat.

Hüvitist saava inimese kohta on teada ta sugu ja vanus. Uuritavate sooline jaotus on selline, et 52% uuritavatest on naised ja 48% mehed. Kõige noorem andmestikus esinev kannatanu on noorem kui 1-aastane ja kõige vanem 98-aastane. Keskmine vanus on 44 eluaastat.

Veel on teada, kas tegemist on õnnetuse põhjustanud isiku ravikuluga või õnnetuses kannatada saanud isiku omaga. Selle tunnuse nimi on *Võla alamliik*.

3.2 Tulemused

Pärast esialgset andmete puhastamist ja ümber kodeerimist viidi läbi erinevad analüüsid, mis on kirjeldatud peatükkides [1](#) ja [2](#).

Coxi võrdeliste riskide mudel

Esmalt vaatame Coxi mudeli tulemusi, mis on esitatud tabelis [7](#).

Näeme, et Coxi mudelis tulevad statistiliselt olulised tunnused *sugu*, *kogukahju*, *vastutava sõiduki kategooria*, *kannataja sõiduki kategooria* ja *võla alamliik*.

Sugu on oluline $\alpha = 0,05$ juures ja hinnanguks on $-0,0603$. Mudel võrdles meeste ravikulude hüvitamise kestust naiste omaga ehk saame hinnangu järgi öelda, et meeste ravikulude hüvitamise kestus on pikem kui naistel. Kuna meie uuritav

Tabel 7: Coxi mudeli tulemused

	$\hat{\beta}$	$se(\hat{\beta})$	p-väärtus
Sugu (mees)	-0,0603	0,0301	0,04497
Vanus	0,0000318	0,000717	0,96461
Kogukahju	-0,0000928	0,00000668	<0,0001
Asukoht (välisriik)	-0,00942	0,165	0,95457
Aastaaeg (sügis)	0,00998	0,0417	0,811
Aastaaeg (suvi)	0,00599	0,0404	0,882
Aastaaeg (talv)	0,00845	0,0413	0,838
Vastutav sõiduk (motikas)	0,192	0,0847	0,02304
Vastutav sõiduk (suursõiduk)	0,207	0,0488	<0,0001
Vastutav sõiduk (teadmata)	0,0481	0,252	0,84832
Kannataja sõiduk (sõiduauto, väikeveok)	0,118	0,0359	0,00101
Kannataja sõiduk (suursõiduk)	0,00628	0,127	0,96056
Kannataja sõiduk (mootorratas, mopeed)	0,0406	0,0747	0,58628
Kannataja sõiduk (mitu)	-0,278	0,0690	<0,0001
Kannataja sõiduk (teadmata)	-0,387	0,225	0,08562
Võla alamliik (põhjustaja ravikulu)	0,338	0,0410	<0,0001

suurus on ravikulude hüvitamise kestus, tähendab negatiivne hinnang seda, et hüvitamise lõppemise risk on meestel väiksem kui naistel.

Võla alamliik tuleb mudelis oluline hinnanguga 0,338. Selle tunnuse puhul võrreldi põhjustaja ravikulu kannatanu ravikuluga, mis tähendab positiivse hinnangu korral seda, et kui ravikulu on õnnetuse põhjustaja oma, on tal suurem risk hüvitamise lõppemisele kui õnnetuses kannatada saanud isikul.

Nii vastutava kui kannatanu sõiduki kategooria puhul leidsid tasemeid, mis on mudelis olulised. Vastutava sõiduki kategooria baastase oli "sõiduauto, väikeveok". Selle tasemega võrreldes on hinnang riskile oluliselt erinev tasemel "motikas" ja ka tasemel "suursõiduk". Mõlema taseme puhul on risk ravikulude hüvitamise lõppemiseks suurem kui baastasemel "sõiduauto, väikeveok". Kannatanu sõiduki puhul oli baastasemeks "jalakäija". Jalakäijaga võrreldes on mudelis oluliselt erinevad tasemed "sõiduauto, väikeveok" ja tase "mitu", mis tähistab seda, et õnnetuses oli mitut liiki kannatanud sõidukeid.

Mudelis osutusid ebaoluliseks tunnused *vanus*, *asukoht* ja õnnetuse toimumise *aastaaeg*.

Proportsionaalsuse eeldus

Järgmiseks hinnati, kas kasutatud Coxi mudeli proportsionaalsuse eeldus on täidetud. Selleks vaadati kaalutud Schoenfeldi jääke *zph* testiga.

Tabel 8: *Zph* testi tulemus

	p-väärtus
Sugu	0,103
Vanus	<0,0001
Kogukahju	<0,0001
Asukoht	0,662
Aastaaeg	0,027
Vastutav sõiduk	<0,0001
Kannataja sõiduk	0,002
Võla alamliik	0,210

Tabelist 8 näeme, et võrdelisuse eeldus on täidetud parameetrite *sugu*, *asukoht* ja *võla alamliik* korral. Ülejäänud tunnuste korral, sealhulgas *kogukahju* ja *vastutava sõiduki* ning *kannatanu sõiduki* kategooria korral, ei ole proportsionaalsuse eeldus täidetud ehk saadud Coxi mudeli tulemus ei ole usaldusväärne.

Et riskide mitte-võrdelisust arvesse võtta, prooviti andmete analüüsimiseks ajas muutuvate koefitsientide meetodit. Kuna vaadeldav aeg on 6 aastat pikk, siis prooviti esmalt jaotada vaatlusaeg kuueks mittelõikuvaks ajavahemikuks ning viia läbi Coxi analüüs nende kuue ajavahemiku korral. Tulemuseks oli mudel, kus iga ajast sõltuv tunnus on jaotatud kuueks ning igale osale on leitud eraldi hinnatav parameeter.

Analüüsitavas andmestikus on kolm kategoorilist tunnust, mille puhul ei olnud proportsionaalsuse eeldus täidetud: *aastaaeg*, *vastutav sõiduk*, *kannatanu sõiduk*. Üldiselt saab ajas muutuvate koefitsientide analüüsis kasutada ka kategoorilist tunnust. Tulemuseks on igale tunnuse tasemele (v.a. baastasemele) iga valitud ajalõigu

raames leitud hinnang. See tähendab, et tunnuse *aastaaeg* korral saame mudelist $(4 - 1) \cdot 6 = 18$ erinevat hinnangut. Ravikulude andmestiku puhul on aga tunnuste *vastutava sõiduki* ja *kannatanu sõiduki* osad tasemed vähem esindatud kui teised ning kuue leitud ajavahemiku korral jäävad mõned taseme ja ajavahemiku kombinatsioonid vaatlustest tühjaks. See tähendab, et mudel ei saa neid tasemeid ning tunnuse üldist proportsionaalsust hinnata. Seetõttu jäid vastutava ja kannatanu sõiduki tunnused lõplikust mudelist välja. See on üks oluline kitsaskoht ajas muutuvate koefitsientide analüüsi puhul, kuna mudelist jäid välja tunnused, millel on potentsiaalne mõju uuritavale suurusele, antud juhul hüvitamise pikkusele.

Teiste tunnuste puhul viidi läbi ajas muutuvate koefitsientide hindamine ning seejärel hinnati saadud mudeli riskide proportsionaalsust.

Tabel 9: Ajas muutuvate koefitsientide proportsionaalsus

	p-väärtus
Sugu	0,0586
Vanus	<0,00232
Kogukahju	<0,00026
Asukoht	0,329
Aastaaeg	<0,0001
Võla alamliik	0,0992

Tablis 9 on näha, et tunnused *vanus*, *kogukahju* ja *aastaaeg* ei täida võrdeliste riskide eeldust ka pärast läbi viidud ajas muutuvate koefitsientide mudeldamist. Lisaks sellele kadusid mudelist tunnused *vastutav sõiduk* ja *kannataja sõiduk*, mille mõju ravikulude hüvitamisele me hinnata soovime. Seega võib öelda, et ajas muutuvate koefitsientide hindamine ei andnud ravikulude hüvitamise uurimise korral soovitud tulemust. Järgmiseks prooviti pseudovaatluste kasutamist, kuna see meetod ei tee võrdeliste riskide eeldust.

Pseudovaatlused

Pseudovaatlused arvutati iga veerandaasta kohta ehk ühe aasta kohta vaadati nelja võrdselt jaotatud ajahetke. See tähendab, et igale vaatlusele arvutati kokku 24

pseudovaatlust. Pärast pseudovaatluste arvutamist ja andmestiku pikale kujule viimist sobitati *GEE* mudel *cloglog* linkfunktsiooniga. Mudeli tulemus on toodud lisas 1 tabelis 11.

Võrdleme Coxi ja *cloglog* lingiga *GEE* mudeli tulemusi tabelis 10.

Tabel 10: Coxi ja pseudovaatluste kasutamise tulemuste võrdlus

	Cox		GEE : cloglog	
	Hinnang	p-väärtus	Hinnang	p-väärtus
Sugu (mees)	-0,0603	0,04497	-0,0650	0,0344
Vanus	0,0000318	0,96461	0,0005	0,4897
Kogukahju	-0,0000928	<0,0001	-0,000095	<0,0001
Asukoht (välisriik)	-0,00942	0,95457	-0,0328	0,8598
Aastaaeg (sügis)	0,00998	0,811	0,0293	0,4962
Aastaaeg (suvi)	0,00599	0,882	0,0177	0,6746
Aastaaeg (talv)	0,00845	0,838	-0,00182	0,9678
Vastutav sõiduk (mootorratas, mopeed)	0,192	0,02304	0,113	0,2058
Vastutav sõiduk (suursõiduk)	0,207	<0,0001	0,228	<0,0001
Vastutav sõiduk (teadmata)	0,0481	0,84832	0,1204	0,2586
Kannataja sõiduk (sõiduauto, väikeveok)	0,118	0,00101	0,101	0,0074
Kannataja sõiduk (suursõiduk)	0,00628	0,96056	-0,0204	0,8777
Kannataja sõiduk (mootorratas, mopeed)	0,0406	0,58628	0,0565	0,4365
Kannataja sõiduk (mitu)	-0,278	<0,0001	-0,314	<0,0001
Kannataja sõiduk (teadmata)	-0,387	0,08562	-0,401	0,1056
Võla alamliik (põhjustaja ravikulu)	0,338	<0,0001	0,368	<0,0001

Coxi ja *cloglog* lingiga *GEE* mudeli puhul näeme, et tunnuste hinnangud on üsna sarnased, mõne erandiga tunnuste puhul, kus proportsionaalsuse eeldus ei olnud täidetud. Kõik tunnused, mis olid olulised Coxi mudelis, tulid olulised ka *GEE cloglog* mudelis, kusjuures viimane ei eelda riskide proportsionaalsust. Suur muudatus p-väärtustes tuli *vastutava sõiduki* tasemel "motikas", mis Coxi mudelis oli oluline, kuid *GEE* mudelis mitte. See muutus võis tuleneda sellest, et *vastutava sõiduki* korral ei olnud võrdeliste riskide eeldus täidetud ja seega oli Coxi tulemus ebausaldusväärne. Mõlema analüüsi tulemusel saame aga öelda, et liiklusõnnetuse tagajärjel tekkinud ravikulude hüvitamise kestust mõjutavad kannatanu sugu, kogukahju suurus, õnnetuses vastutava sõiduki kategooria, kannatanu sõiduki kategooria ja see, kas tegemist on õnnetuse põhjustaja või kannatanu ravikuluga.

4 Arutelu

Ravikulude hüvitamise analüüsist selgus, et pseudovaatluste meetod andis sarnaseid tulemusi Coxi võrdeliste riskide mudeliga, kusjuures pseudovaatluste meetod ei eelda ajas võrdelisi riske. Kuigi ka antud juhul andsid pseudovaatlused häid tulemusi, seisneb selle meetodi tõeline eelis olukordades, kus standardseid elukestusanalüüsi meetodeid kasutusel ei ole (Andersen ja Perme, 2010). Selline olukord on näiteks tõkestatud keskmise üleelamistõenäosuse hindamisel või konkureerivate riskide mudeldamisel. On mitmeid näiteid kirjandusest, kus pseudovaatluste kasutamine konkureerivate riskide mudelis on andnud häid tulemusi (näiteks Sachs *et al.* (2019), Mortensen *et al.* (2018)) .

Pseudovaatluste suur potentsiaalne eelis seisneb ka laialdasemas masinõppe meetodite rakendamise võimaldamises. Näiteks konkureerivate riskide puhul pole häid masinõppe algoritme, mis oskaksid arvesse võtta tsenseerimist. Kuna pseudovaatlused kaotavad andmetest ära tsenseerimise, saab neid vaatlusi kasutada erinevates statistilistes mudelites, sealhulgas masinõppes (Sachs *et al.*, 2019).

Pseudovaatluste kasutamine on veel võrdlemisi uus meetod. Üks pseudovaatluste negatiivseid külgi on arvutusmahukus. Kuna pseudovaatlused arvutatakse mitmel ajahetkel (teoorias igal sündmuse toimumise ajahetkel, v.a. tõkestatud keskmise korral), siis on suure andmestiku korral tegemist mahuka arvutusprotsessiga. Kuigi Anderseni ja Perme'i (2010) kohaselt piisab praktikas vähematest, 5-10 pseudovaatlusest, siis tundub, et see väide ei ole täielikku kinnitust leidnud. Autorid leidsid, et rohkemate pseudovaatluste arvutamine ei andnud eeliseid või ei mõjutanud tulemusi oluliselt, kuid Graw *et al.* (2009) toovad välja, et vajalik punktide hulk võib sõltuda ka andmetest.

Teine negatiivne külg pseudovaatluste kasutamise juures on tulemuste tõlgendamine. Kui klassikalised elukestusanalüüsi meetodid hindavad enamasti riski või üleelamistõenäosust, siis pseudovaatlused võtavad üleelamistõenäosuse, kuid muudavad selle omaette suuruseks, mida ei saa intuiitiivselt tõlgendada. Küll aga saab

pseudovaatlusi kasutada erinevate tunnuste mõju hindamiseks elukestusele ja selles nende kasutamise eesmärk seisnebki - et oleks võimalik elukestusandmete peal kasutada rohkemaid analüüsimeetodeid, kui siiani klassikalises elukestusanalüüsis.

Kokkuvõte

Töö eesmärk oli aru saada, mis on pseudovaatlused, mis on nende kasutamise eelised ja kuidas neid arvutada ning praktikas rakendada. Töö teine eesmärk oli mõista, mis mõjutab liiklusõnnetuse tagajärjel tekkinud ravikulude hüvitamise kestust LKF-i andmete näitel. Selleks kasutati nii Coxi proportsionaalsete riskide mudelit kui ka pseudovaatluste põhjal arvutatud *GEE* mudelit *cloglog* lingiga, mille tulemusi saab võrrelda Coxi mudeli tulemustega.

Töö esimeses peatükis anti ülevaade elukestusanalüüsist ja selle olulisematest mõistetest. Elukestusanalüüsi meetoditest keskenduti Coxi võrdeliste riskide mudelile ning kirjeldati võrdeliste riskide eelduse täidetust. Eelduse mitte täidetuse lahendusena pakuti stratifitseerimist ja ajas muutuvate koefitsientide mudelit.

Töö teises peatükis tutvustati pseudovaatlusi. Esmalt tehti üldine ülevaade nende olemusest ning seejärel vaadati täpsemalt kolme juhtu: pseudovaatlused tavalise hinnanguna üleelamistõenäosusele, tõkestatud keskmise üleelamistõenäosusele ja konkureerivate riskide mudelile. Kõigi kolme juhu puhul kirjeldati meetodeid, kuidas ja mis ajahetkedel pseudovaatlusi arvutada. Peale seda kirjeldati *GEE* meetodit pseudovaatluste mudeldamiseks. Kuna näiteks üleelamistõenäosuse pseudovaatluste arvutamise korral saadakse andmestik, mis sisaldab ühe vaatluse kohta mitut rida, tuleb andmetes sõltuvuse arvesse võtmiseks eelistada *GEE* meetodit *glm* mudelite asemel. Veel kirjeldati pseudovaatluste ja *GEE* mudelite rakendamist statistikatarkvaras R. Viimaks näitlikustati lihtsa simulatsioonandmestiku peal pseudovaatluste olemust üleelamistõenäosuse ja tõkestatud keskmise eluea korral.

Töö kolmas peatükk oli praktilise osa kokkuvõte. Töö praktilise osa eesmärk oli leida ravikulude hüvitamist mõjutavad tegurid. Selleks kasutati esmalt Coxi võrdeliste riskide mudelit, prooviti rakendada ka ajas muutuvate koefitsientide mudelit ning viimaks pseudovaatlusi. Pseudovaatlused arvutati iga veerandaasta kohta ja seejärel sobitati pseudovaatluste abil andmestikule *GEE* mudel täiend-log-log lingiga.

Töö praktilises osas leiti, et ravikulude hüvitamise kestust mõjutab inimese sugu, hüvitamise kogukahju, kannatanu sõiduki kategooria (sh jalakäija), põhjustaja sõiduki kategooria ja see, kas hüvitise puhul on tegemist õnnetuse põhjustaja või õnnetuses kannatanu ravikuluga. Need tunnused osutusid oluliseks nii Coxi mudeli kui ka GEE mudeli korral, kusjuures tunuste hinnangud tulid mõlema mudeli puhul sarnased.

Kasutatud allikad

- Agresti, Alan (2002). *Categorical Data Analysis*. 2. väljaanne. John Wiley & Sons, Inc.
- Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill ja Niels Keiding (1993). *Statistical models based on counting processes*. Springer-Verlag.
- Andersen, Per Kragh ja John P. Klein (2007). “Regression Analysis for Multistate Models Based on a Pseudo-value Approach, with Applications to Bone Marrow Transplantation Studies”. *Scandinavian Journal of Statistics* 34.1, lk. 3–16. DOI: <https://doi.org/10.1111/j.1467-9469.2006.00526.x>.
- Andersen, Per Kragh, John P. Klein ja Susanne Rosthøj (2003). “Generalised linear models for correlated pseudo-observations, with applications to multi-state models”. *Biometrika* 90.1, lk. 15–27. DOI: [10.1093/biomet/90.1.15](https://doi.org/10.1093/biomet/90.1.15).
- Andersen, Per Kragh ja Maja Pohar Perme (2010). “Pseudo-observations in survival analysis”. *Statistical Methods in Medical Research* 19.1, lk. 71–99. DOI: [10.1177/0962280209105020](https://doi.org/10.1177/0962280209105020).
- Collett, David (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC. DOI: <https://doi.org/10.1201/b18041>.
- Ford, Clay (2021). *Getting Started with Generalized Estimating Equations*. URL: [https://data.library.virginia.edu/getting-started-with-generalized-estimating-equations/#:~:text=Generalized%20Estimating%20Equations%2C%20or%20GEE,\(ie%2C%20model%20coefficients\)](https://data.library.virginia.edu/getting-started-with-generalized-estimating-equations/#:~:text=Generalized%20Estimating%20Equations%2C%20or%20GEE,(ie%2C%20model%20coefficients)).
- Grambsch, Patricia M. ja Terry M. Therneau (september 1994). “Proportional hazards tests and diagnostics based on weighted residuals”. *Biometrika* 81.3, lk. 515–526. DOI: [10.1093/biomet/81.3.515](https://doi.org/10.1093/biomet/81.3.515).

- Graw, Frederik, Thomas A. Gerds ja Martin Schumacher (2009). “On pseudo-values for regression analysis in competing risks models”. *Lifetime Data Analysis* 15.2, lk. 241–255. DOI: [10.1007/s10985-008-9107-z](https://doi.org/10.1007/s10985-008-9107-z).
- Klein, John P. ja Per Kragh Andersen (2005). “Regression Modeling of Competing Risks Data Based on Pseudovalues of the Cumulative Incidence Function”. *Biometrics* 61.1, lk. 223–229. DOI: <https://doi.org/10.1111/j.0006-341X.2005.031209.x>.
- Klein, John P., Mette Gerster, Per Kragh Andersen, Sergey Tarima ja Maja Pohar Perme (2008). “SAS and R functions to compute pseudo-values for censored data regression”. *Computer Methods and Programs in Biomedicine* 89.3, lk. 289–300. ISSN: 0169-2607. DOI: <https://doi.org/10.1016/j.cmpb.2007.11.017>.
- Maja Pohar Perme ja Mette Gerster (2017). *pseudo: Computes Pseudo-Observations for Modeling*. R package version 1.4.3. URL: <https://cran.r-project.org/web/packages/pseudo/pseudo.pdf>.
- Mortensen, Lotte Maxild, Camilla Plambeck Hansen, Kim Overvad, Søren Lundbye-Christensen ja Erik T. Parner (2018). “The Pseudo-Observation Analysis of Time-To-Event Data. Example from the Danish Diet, Cancer and Health Cohort Illustrating Assumptions, Model Validation and Interpretation of Results”. *Epidemiologic Methods* 7.1, lk. 20170015. DOI: [doi:10.1515/em-2017-0015](https://doi.org/10.1515/em-2017-0015).
- R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: <https://www.R-project.org/>.
- Sachs, Michael C., Andrea Discacciati, Åsa H. Everhov, Ola Olén ja Erin E. Gabriel (2019). “Ensemble prediction of time-to-event outcomes with competing risks: a case-study of surgical complications in Crohn’s disease”.

Journal of the Royal Statistical Society: Series C (Applied Statistics) 68.5, lk. 1431–1446. DOI: <https://doi.org/10.1111/rssc.12367>.

Zhang, Zhongheng, Jaakko Reinikainen, Kazeem Adedayo Adeleke, Marcel E Pieterse ja Catharina G M Groothuis-Oudshoorn (aprill 2018). “Time-varying covariates and coefficients in Cox regression models.” *Annals of translational medicine* 6. DOI: [10.21037/atm.2018.02.12](https://doi.org/10.21037/atm.2018.02.12).

Lisad

Lisa 1. *GEE* mudeli tulemus kasutades pseudovaatlusi

Pseudovaatluste *GEE cloglog*-lingiga mudeldamise tulemus.

Tabel 11: Pseudovaatluste põhjal arvutatud *GEE* mudeli tulemused

	Hinnang	sd	p-väärtus
Sugu (mees)	-0,0650	0,0307	0,0344
Vanus	0,0005	0,0007	0,4897
Kogukahju	-0,000095	0,000072	<0,0001
Asukoht (välisriik)	-0,0328	0,1856	0,8598
Aastaaeg (sügis)	0,0293	0,0431	0,4962
Aastaaeg (suvi)	0,0177	0,0422	0,6746
Aastaaeg (talv)	-0,00182	0,0449	0,9678
Vastutav sõiduk (mootorratas, mopeed)	0,113	0,0893	0,2058
Vastutav sõiduk (suursõiduk)	0,228	0,0484	<0,0001
Vastutav sõiduk (teadmata)	0,1204	0,107	0,2586
Kannataja sõiduk (sõiduauto, väikeveok)	0,101	0,0377	0,0074
Kannataja sõiduk (suursõiduk)	-0,0204	0,1327	0,8777
Kannataja sõiduk (mootorratas, mopeed)	0,0565	0,0726	0,4365
Kannataja sõiduk (mitu)	-0,314	0,0628	<0,0001
Kannataja sõiduk (teadmata)	-0,401	0,248	0,1056
Võla alamliik (põhjustaja ravikulu)	0,368	0,0446	<0,0001

Lisa 2. Kasutatud R-i kood pseudovaatluste arvutamiseks

```
# Tekitame ajahetked, kuna pseudovaatlusi arvutada
veerandaastad <- c(91,183,274,365,
                   456,548,639,730,
                   821,913,1004,1095,
                   1186,1278,1369,1460,
                   1551,1643,1734,1825,
                   1916,2008,2099,2188)

# Arvutame pseudovaatlused igal veerandaastal
ps_values_12 <- pseudosurv(lkf_andmestik_12$pikkus,
                           lkf_andmestik_12$tsenseeritud, veerandaastad)

# cloglog mudeli jaoks
ll <- 1-ps_values_12$pseudo
ps_values_12$pseudo <- ll

# Tekitame pika andmestiku
b_12 <- NULL
for(it in 1:length(ps_values_12$time)){
  b_12 <- rbind(b_12,cbind(lkf_andmestik_12,
                           ps_values=ps_values_12$pseudo[,it],
                           tpseudo = ps_values_12$time[it],
                           id=1:nrow(lkf_andmestik_12)))
}
b_12 <- b_12[order(b_12$id),]
# Andmestik b_12 on pikal kujul andmestik, kus iga vaatlus esineb
# nii mitu korda, kui mitu pseudovaatlust arvutati.
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kadi-Liis Kuum,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Pseudovaatluste kasutamine elukestusanalüüsis liiklusõnnetuste ravikulude hüvitamise näitel", mille juhendaja on Anastassia Kolde, 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 Commons litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.
3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.
4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Kadi-Liis Kuum

17.05.2022