Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Kadi Kilgi

Liikluskindlustuslepingute pikkuste prognoosimine Weibulli mudelite abil

Matemaatilise statistika eriala
Bakalaureusetöö (9 EAP)

Juhendajad: Prof. Krista Fischer, *Ph.D*
Raine Talvet, *MSc*

Tartu 2020

Liikluskindlustuslepingute pikkuste prognoosimine Weibulli mudelite abil

Bakalaureusetöö

Kadi Kilgi

Lühikokkuvõte. Käesoleva bakalaureusetöö eesmärk on analüüsida liikluskindlustuslepingute katkestamisi ning hinnata tunnused, mis iseloomustavad kliente, kes katkestavad kiiremini kui teised. Esmalt tutvustatakse elukestusanalüüsi mõisteid ning mudeldamiseks kasutatavaid võrdelise riski ja kiirendatud tõrkeaja mudeleid. Lisaks kirjeldatakse Weibulli jaotust ning rakendustarkvara R võimalusi elukestusandmetega töötamiseks. Seejärel puhastatakse andmed ja luuakse mudel. Mudeli põhjal leitakse sõltumatutele andmetele prognoosid, mille abil valideeritakse mudel. Veel tuuakse välja mudelist selgunud tulemused. Analüüsiks kasutatakse Ergo kindlustuse andmeid. Töö tulemusena valmib mudel, mille abil saab prognoosida uute lepingute pikkust.

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

Märksõnad: elukestusanalüüs, Kaplan-Meieri hinnang, Weibulli jaotus, võrdeliste riskide mudel, kiirendatud tõrkeaja mudel.

Estimating durations of motor insurance contracts with Weibull models

Bachelor's thesis

Kadi Kilgi

Abstract. The purpose of this thesis is to analyse possible reasons, why motor insurance contracts are ended and characterize clients, who are more likely to end their contracts. At the beginning, the concepts of survival analysis are introduced, including proportional hazard model, accelerated failure time model and models assuming Weibull distribution. In addition, the relevant functions of statistical software R are described. The final model is validated in an independent dataset, where also the properties of model-based predictions are studied. The data for this thesis is provided by Ergo Insurance SE. The result of the thesis is a model that can be used to predict durations for the new contracts in motor insurance.

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

Keywords: survival analysis, Kaplan-Meier estimation, Weibull distribution, proportional hazard model, accelerated failure time model.

Sisukord

Sissejuhatus *5*

1. Metoodika *6*

1.1. Elukestusanalüüsi põhimõisted 6

1.2. Kaplan-Meieri hinnang üleelamisfunktsioonile 7

1.3. Võrdeliste riskide mudel ja kiirendatud tõrkeaja mudel 9

1.4. Weibulli jaotus elukestusele ja mudelid Weibulli jaotuse eeldusel 10

1.5. Weibulli jaotusega mudeli hindamine rakendustarkvara R abil 13

1.6. Prognoosi täpsuse hindamine 14

2. Liikluskindlustuse andmete analüüs *17*

2.1. Andmete kirjeldus 17

2.2. Andmete puhastus 20

2.3. Andmete sobivus metoodikaga 21

2.4. Mudel 22

2.5. Tulemused mudelist 27

2.6. Prognoosid 31

Kokkuvõte *35*

Kasutatud kirjandus *37*

Lisad *38*

Lisa 1. Andmestiku lahutamine treening- ja testandmestikuks ning mudeli koostamine 38

Lisa 2. Parameetri beeta hinnangute illustreerimine usaldusvahemikega 38

Lisa 3. Mudeliga lepingu üleelamistõenäosuse hindamine 39

Lisa 4. Mudeliga lepingu pikkuse prognoosimine 40

Lisa 5. Prognoositud ja tegelike lepingu pikkuste jaotus 40

Lisa 6. Tegelikud ja prognoositud lepingu pikkused kategooriates „Kuni 1 aasta“ ja „Üle ühe aasta“ 40

Sissejuhatus

Maanteeametis on 29.02.2020 aasta seisuga registreeritud 676 638 sõidukit [1]. Eestis on kohustuslik omada liikluskindlustust mootorsõidukitel ja haagistel, mis on registreeritud liiklusregistris ning mida kasutatakse liikluses liiklemiseks. Liikluskindlustuse puhul katab kindlustusfirma, kelle juures on sõlmitud leping, kulud, mis on tekitatud kindlustatud sõiduvahendi poolt kolmandale isikule. Liikluskindlustuslepingute pikkuseks on kuni 1 aasta. Liikluskindlustus katab tekitatud kulud, kuid saab kindlustada ka enda sõidukile tekitatud kahjud, sellist kindlustusliiki nimetatakse sõidukikindlustuseks. [2]

Käesoleva bakalaureusetöö eesmärk on analüüsida liikluskindlustuslepingute katkestamisi ning hinnata tunnused, mis iseloomustavad kliente, kes katkestavad kiiremini kui teised. Töös kasutatakse Ergo kindlustusseltsi andmeid liikluskindlustuslepingute kohta aastatel 2015 kuni 2019. Ergo kindlustuses pole varasemalt antud teemat süvitsi analüüsitud ning seetõttu puudub väljatöötatud metoodika. Töö tulemusena valmib mudel, mis annab võimekuse prognoosida uute lepingute puhul nende pikkust. Kindlustusseltsile annab antud töö võimaluse ennetada klientide lahkumist ja hoida lojaalseid uusi kliente.

Töö koosneb kahest peatükist ja 12 alapeatükist. Esimeses peatükis kirjeldatakse metoodikat, mida kasutatakse andmete analüüsiks. Analüüsiks kasutatakse elukestusanalüüsi meetodeid, sealhulgas Kaplan-Meieri hinnang, Weibulli jaotus, võrdeliste riskide ja kiirendatud tõrkeaja mudelid. Teises peatükis tehakse esmalt ülevaade andmetest ja andmepuhastusest. Seejärel tutvustatakse leitud mudelit ning mudeli põhjal leitud hinnanguid lepingu pikkusele ja nende kooskõla reaalselt vaadelduga. Analüüsil on kasutatud rakendustarkvara R [3].

1. Metoodika

Järgnevas peatükis tutvustame elukestusanalüüsi mõisteid ja meetodeid, mida analüüsiks kasutame.

1.1. Elukestusanalüüsi põhimõisted

Elukestusanalüüsi kasutame andmetel, mida iseloomustab mingil ajahetkel toimuv huvipakkuv sündmus. Peamiselt pakub huvi aeg teatud algmomendist sündmuse toimumiseni (nii selle ajavahemiku jaotus kui seda mõjutavad tegurid). Seega on analüüsiks vaja teada vaatluse alla sattunud indiviidi algusaega, millal indiviid võetakse uurimise alla, ning lõppaega, millal oodatud sündmus toimub. Väga tihti kasutatakse meetodit kliinilistes uuringutes, mille käigus uuritakse näiteks inimese eluaega ning sellisel juhul on huvipakkuvaks sündmuseks surm. Samas esineb sarnaseid andmeid ka paljudes teistes valdkondades, näiteks elektroonikaseadmete elukestuse uurimine. [4, lk.1] Meie andmete puhul on vaatlusalusteks liikluskindlustuslepingud ja huvipakkuv sündmus on lepingu katkestamine.

Tsenseeritud vaatluseks loeme indiviidi, kelle puhul meil puudub informatsioon sündmuse toimumise aja kohta. See enamasti tähendab, et vaatlusperioodi jooksul oodatud sündmust ei toimunud. Järelikult puudub informatsioon vaatlusaluse kohta ning me ei tea kindlalt, kas vaatlusalune jõuab oodatud sündmuseni või kui ta jõuab, siis mis hetkel. [4, lk. 2–4] Meie andmete korral on tsenseeritud vaatlusteks lepingud, mida pole katkestatud meie vaatluse ajal ehk lepingud, mis vaatluse viimasel päeval veel kehtisid. Tsenseerimise korral vaatleme kahte juhuslikku suurust:

T_i – tähistab *i*-nda indiviidi elukestust (*i=1,2,...,n*) ehk aega sündmuse toimumiseni

C_i – tähistab *i*-nda indiviidi tsenseerimisaega [4, lk. 2–4].

Meie andmete korral T_i tähistab aega päevades alates lepingu sõlmimisest kuni lepingu katkestamiseni ja C_i tähistab aega päevades lepingu sõlmimisest kuni andmebaasist väljavõtte tegemiseni (30.10.2019). Seega on see viimane ajahetk, mille kohta on meil andmed teada. Kui $T_i \leq C_i$, siis on toimunud lepingu katkestamine ja saame leida lepingu elukestuse [4, lk. 2–4]. Vastasel juhul on tegemist tsenseeritud vaatlusega ehk lepingu lõpukuupäev on hetkel teada olevalt väljaspool meie vaatlusaega ning vaatlusajal käesolevat lepingut ei katkestatud.

Elukestusandmete korral huvitab meid üleelamistõenäosus – tõenäosus, et ajahetkel *t* ei ole indiviid veel jõudnud oodatud sündmuseni. See tõenäosus on määratud üleelamisfunktsiooniga. Olgu *T* juhuslik suurus, mis kirjeldab uuritava indiviidi elukestust ehk aega, mis kulus oodatud sündmuseni jõudmiseks. Üleelamisfunktsioon hindab tõenäosust, et oodatud sündmus ei toimu enne ajamomenti *t*:

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

kus $F(t)$ on *T* jaotusfunktsioon

$$F(t) = P(T < t) = \int_0^t f(u)du.$$

Kui $f(t)$ on *T* tihedusfunktsioon, siis

$$S(t) = 1 - \int_0^t f(u)du.$$

Lepingu katkestamistõenäosust hindame seega valemiga $1 - S(t)$. [4, lk. 10–12]

Lisaks üleelamisfunktsioonile kasutame ka riskifunktsiooni. Definitsiooni kohaselt on riskifunktsioon tõenäosus, et oodatav sündmus toimub ajamomendil t, kui on teada, et sündmus ei toimunud varasemalt. Pideva elukestuse aja puhul avaldub riskifunktsioon kujul:

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

Kui aega vaadelda diskreetsena $(t_0, t_1, t_2, \dots)$, siis avaldub riskifunktsioon kujul:

$$h_j = P(T = t_j | T \geq t_j).$$

Vaadeldes aega *t* päevades, iseloomustab riskifunktsioon tõenäosust, et indiviid, kes ei ole varasemalt oodatud sündmust kogenud, teeb seda t_j-ndal päeval. [4, lk. 12]

1.2. Kaplan-Meieri hinnang üleelamisfunktsioonile

Kui meie andmetes ei oleks tsenseeritud vaatlusi, siis saaksime *T* jaotust kirjeldada empiirilise jaotusfunktsiooni kaudu. Seega üleelamisfunktsioon oleks hinnatav kui:

$$\hat{S}(t) = \frac{indiviidide\ arv, kes\ elasid\ kuni\ ajahetkeni\ t\ või\ kauem}{indiviidide\ arv\ andmestikus},$$

kus $\hat{F}(t) = \frac{indiviidide\ arv, kes\ elasid\ kuni\ ajahetkeni\ t}{indiviidide\ arv\ andmestikus}$. [4, lk. 17]

Leiame Kaplan-Meieri hinnangu tsenseeritud andmete korral. Eeldame, et meil on teada indiviidide elukestusajad. Moodustame intervallid nii, et igas intervallis on üks sündmuse toimumise aeg ja see asub intervalli otspunktis. Seega kui $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ on järjestatud sündmuse toimumise ajad, kus *r* tähistab erineva sündmuse toimumise ajahetkega indiviidide

arvu. Sellisel juhul moodustame intervallid: $[t_{(1)}, t_{(2)}), [t_{(2)}, t_{(3)}), \dots, [t_{(r-1)}, t_{(r)})$, kus ühel sündmuse toimumise hetkel võib sündmus toimuda korraga mitmel indiviidil. Seega kui andmestikus on *n* indiviidi, siis $r \leq n$. Tähistame

t_j – tähistab ajahetke päevades, mil toimus oodatud sündmus,

d_j – tähistab ajahetkel t_j sündmuseni jõudnute arvu,

n_j – tähistab indiviidide arvu nn riskigrupis, kes ei ole sündmuseni jõudnud enne ajahetke t_j kaasaarvatud need indiviidid, kes jõuavad sündmuseni ajahetkel t_j, kus $j = 1,2, \dots, r$.

Vaatleme intervalli $[t_{(j)}, t_{(j+1)})$, siis tõenäosus, et riskigrupis olev indiviid jõuab sündmuseni selles ajavahemikus, $P(T < t_{(j+1)} | T \geq t_{(j)})$, on hinnatav kui $\frac{d_j}{n_j}$. Järelikult tõenäosus, et indiviid ei jõua sündmuseni, $P\left(T \geq t_{(j+1)} \middle| T \geq t_{(j)}\right)$, ehk indiviid elab edasi, on hinnatav kui

$$1 - \frac{d_j}{n_j} = \frac{n_j - d_j}{n_j}.$$

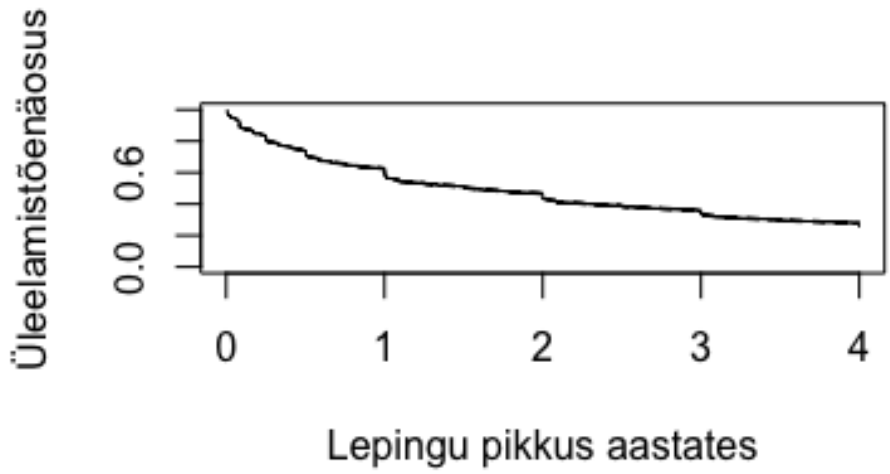
Kaplan-Meieri hinnang hindab tõenäosust, et indiviid elab üle intervalli $\left[t_{(k)}, t_{(k+1)}\right)$ ja kõik eelnevad intervallid, kus $k = 1,2, \dots, r$ ja $t_{(r+1)} = \infty$, mis avaldub kui:

$P\left(T \geq t_{(k)}\right) = P\left(T \geq t_{(1)}, T \geq t_{(2)}, \dots, T \geq t_{(k)}\right) = P\left(T \geq t_{(1)}\right) \cdot P\left(T \geq t_{(2)} \middle| T \geq t_{(1)}\right) \cdot \dots \cdot P\left(T \geq t_{(k)} \middle| T \geq t_{(k-1)}\right) = \prod_{j=1}^{k} P(T \geq t_{(j+1)} | T \geq t_{(j)})$.

Seega saame hinnangu kujul

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right), \text{ kus } t_{(k)} \leq t < t_{(k+1)}.$$

Lisaks teame, et kui $t < t_{(1)}$, siis mitte ükski indiviid ei ole jõudnud sündmuseni ja seega $\hat{S}(t) = 1$. Kui $t \geq t_{(r)}$ ja $n_r = d_r$, siis $\hat{S}(t) = 0$. Kaplan-Meieri hinnang üleelamisfunktsioonile on treppfunktsioon, mis on konstantne sündmuse toimumisaegade vahel ning väheneb ajal, mil toimub oodatud sündmus. [4, lk. 21–23]



Joonis 1. Kaplan-Meieri hinnang.

Näitena vaatame joonisel 1 kujutatud Kaplan-Meieri hinnangut, mis iseloomustab antud töös kasutatavaid andmeid. Näeme, et meie oodatud sündmust, lepingu katkestamist, toimub 4 aasta jooksul väga tihti. Katkestamisi toimub väga tihedalt esimese aasta jooksul ning üleelamistõenäosus väheneb märgatavalt. Lisaks on näha, et lepinguid ei katkestata ainult aasta möödudes.

1.3. Võrdeliste riskide mudel ja kiirendatud tõrkeaja mudel

Elukestusanalüüsi puhul soovime hinnata, millised tegurid mõjutavad riski, et vaatluse alla sattunud indiviid kogeb oodatud sündmust. Seetõttu esitatakse mudelite üldkuju tihti argumenttunnuste ja riskifunktsiooni seosena. Kui mudeli abil on saadud hinnang riskifunktsioonile, siis saame leida ka hinnangu üleelamisfunktsioonile.

Olgu meil *n* indiviidi ning soovime hinnata riskifunktsiooni *i*-ndale indiviidile, siis **võrdeliste riskide mudel** avaldub kujul

$$h_i(t) = \psi(\boldsymbol{x_i}) h_0(t),\ i = 1,2,\dots,n,$$

kus $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \dots, x_{pi})$ on tunnuste $\boldsymbol{X} = (X_1, X_2, \dots, X_p)$ väärtused. Kui *i*-nda indiviidi argumenttunnuste väärtused on kõik nullid, siis tema elukestus on T_0 ning riskifunktsioon $h_0(t)$, mida nimetatakse baasriskifunktsiooniks. Vastav baasüleelamisfunktsioon on $S_0(t)$. Kordaja $\psi(\boldsymbol{x_i})$ iseloomustab argumenttunnuste mõju riskifunktsioonile. Et $\psi(\boldsymbol{x_i})$ ei sõltu ajast *t*, siis see mudel eeldab, et argumenttunnuste erinevatele väärtustele vastavad riskid on

võrdelised. Kuna ψ peab olema positiivne, siis enamasti võetakse $\psi(\boldsymbol{x_i}) = e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}}$, kus $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$ on hinnatavad parameetrid. Seega saame, et *i*-nda indiviidi riskifunktsioon avaldub kujul

$$h_i(t) = e^{\beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}} h_0(t).$$ [4, lk. 58–59]

Kiirendatud tõrkeaja mudeli korral eeldame, et *i*-nda indiviidi üleelamisfunktsioon avaldub kujul

$$S_i(t) = S_0(\psi(\boldsymbol{x_i})t) \quad (1)$$

ning riskifunktsioon

$$h_i(t) = \psi(\boldsymbol{x_i}) h_0(\psi(\boldsymbol{x_i})t).$$

Tüüpiliselt eeldame, et $\psi(\boldsymbol{x_i}) = e^{-\eta_i}$, kus $\eta_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}$ ning $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ on hinnatavad parameetrid. [4, lk. 232–233]

1.4. Weibulli jaotus elukestusele ja mudelid Weibulli jaotuse eeldusel

Vaatleme juhtu, kus T on Weibulli jaotusega, $T \sim W(\lambda, \gamma)$. Weibulli jaotuse korral λ on skaalaparameeter ning γ on kujuparameeter. Kujuparameetri γ väärtus määrab, kas riskifunktsioon on kasvav või kahanev. Kui $\gamma = 1$, siis on tegemist eksponentjaotusega. Kui $\gamma < 1$, siis on tegemist kahaneva funktsiooniga, ning $\gamma > 1$ korral on tegemist kasvava funktsiooniga. Weibulli jaotuse korral avaldub riskifunktsioon kujul

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

kus $0 \leq t < \infty$ ja $\lambda, \gamma > 0$ ning üleelamisfunktsioon on kujul

$$S(t) = e^{-\lambda t^{\gamma}}.$$ [4, lk. 173–174] (2)

Kui logaritmime üleelamisfunktsiooni, siis

$$\ln S(t) = -\lambda t^{\gamma}.$$

Korrutades võrduse mõlemad pooled läbi -1 ja veelkord logaritmides saame, et

$$\ln(-\ln S(t)) = \ln \lambda t^{\gamma}.$$

Kasutades logaritmi omadust, et $\ln \lambda t = \ln \lambda + \ln t$, saame, et

$$\ln(-\ln S(t)) = \ln \lambda + \gamma \ln t.$$

Järelikult on tegemist lineaarse funktsiooniga $\ln t$ suhtes. Seega selleks, et hinnata, kas meie valim võiks olla Weibulli jaotusega, vaatleme hajuvusgraafikule kantud $\ln[-\log \hat{S}(t)]$, kus $\hat{S}(t)$ leitud Kaplan-Meieri hinnanguna, ja $\ln t$ punkte, mis Weibulli jaotuse korral peaksid asuma ühel sirgel. [4, lk. 177–178]

Soovides hinnata mudeleid Weibulli jaotusega elukestusele T, eeldame, et argumenttunnused mõjutavad skaalaparameetrit λ. Seega eeldame, et riskifunktsioon avaldub kui

$$h_i(t) = \lambda_i \gamma t^{\gamma-1},$$

kus $\lambda_i = e^{\beta_1 x_{1i} + \cdots + \beta_p x_{pi}} \lambda_0 = e^{\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}}$ ning $\lambda_0 = e^{\beta_0}$.

Seega baasriskifunktsiooni saame esitada kui

$$h_0(t) = \lambda_0 \gamma t^{\gamma-1}$$

ning tegemist on võrdeliste riskide mudeliga. Nüüd saame üldise üleelamisfunktsiooni (2) *i*-ndale indiviidile avaldub kui

$$S_i(t) = e^{-\lambda_i t^{\gamma}}, \tag{3}$$

kus elukestus $T_i \sim W(\lambda_i, \gamma)$. Mudeli parameetrid hindame suurima tõepära meetodil. [4, lk. 199–201]

Saab näidata, et Weibulli jaotuse korral kehtivad nii võrdeliste riskide kui ka kiirendatud tõrkeaja mudelid. Tähistame $l_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$, seega $\lambda_i = \lambda_0 e^{l_i}$. Kui argumenttunnuste väärtused on kõik nullid, siis saame baasüleelamisfunktsiooni:

$$S_0(t) = e^{-\lambda_0 t^{\gamma}}.$$

$$S_i(t) = e^{-\lambda_i t^{\gamma}} = e^{-\lambda_0 e^{l_i} t^{\gamma}} = e^{-\lambda_0 (e^{l_i} t^{\gamma})} = S_0(e^{\frac{l_i}{\gamma}} t)$$

Seega kehtib ka kiirendatud tõrkeaja mudel (1).

Järgnevalt näitame, et Weibulli jaotusega elukestuse T_i korral saab mudeli esitada ka kujul

$$\log T_i = \mu + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi} + \sigma \epsilon_i, \tag{4}$$

kus $\boldsymbol{x_i} = (x_{1i}, x_{2i}, \ldots, x_{pi})$ on argumenttunnuste väärtused *i*-nda indiviidi korral, μ, σ on parameetrid ning $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$ tähistab tundmatuid parameetreid, mis iseloomustavad argumenttunnuse mõju elukestusele ning juhuslik suurus ϵ_i on Gumbeli jaotusega. Üleelamisfunktsioon Gumbeli jaotuse korral on

$$S_{\epsilon_i}(\epsilon) = e^{-e^{\epsilon}}, \qquad -\infty < \epsilon < \infty.$$

Elukestusele T_i vastav üleelamisfunktsioon on siis

$$S_i(t) = e^{-e^{\frac{\log t - \mu - \alpha_1 x_{1i} - \alpha_2 x_{2i} - \cdots - \alpha_p x_{pi}}{\sigma}}}.$$

Saame selle viia kujule

$$S_i(t) = e^{-e^{\frac{\log t}{\sigma} - \frac{\mu}{\sigma} - \frac{\alpha_1 x_{1i} + \alpha_2 x_{2i} + \cdots + \alpha_p x_{pi}}{\sigma}}}.$$

Tähistades saadud üleelamisfunktsiooni järgmiselt:

$$S_i(t) = e^{-\lambda_i t^{1/\sigma}}, \tag{5}$$

kus $\lambda_i = e^{\frac{-\mu-\alpha_1 x_{1i}-\alpha_2 x_{2i}-\cdots-\alpha_p x_{pi}}{\sigma}}$. Valemist (3) järeldame, et elukestus T_i on Weibulli jaotusega, mille skaalaparameeter $\lambda_i = \lambda_0 e^{\frac{-\alpha_1 x_{1i}-\alpha_2 x_{2i}-\cdots-\alpha_p x_{pi}}{\sigma}}$, kus $\lambda_0 = e^{-\frac{\mu}{\sigma}}$, ja kujuparameeter $\gamma = \frac{1}{\sigma}$, ning parameetrid $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ on omavahel seotud võrdusega $\beta_j = -\frac{\alpha_j}{\sigma}$, kus $j = 1,2,\dots,p$. Seega mudel (4) avaldub kujule

$$\log T_i = \frac{1}{\gamma}\{-\log\lambda_0 - \beta_1 x_{1i} - \beta_2 x_{2i} - \cdots - \beta_p x_{pi} + \epsilon_i\} \tag{6}$$

ehk kiirendatud tõrkeaja mudel avaldub võrdeliste riskide mudeli kaudu. [4, lk. 174, 234, 236–238]

Meil on teada, et Weibulli jaotuse korral üleelamisfunktsioon on esitatav kujul (2). Saame hinnata elukestuse mediaani, mis leitakse üleelamisfunktsiooni pöördfunktsiooni kaudu, mida tähistame $t(p)$. Me soovime leida mediaani, seega otsitav väärtus on $t(50)$.

$$S\big(t(50)\big) = 0{,}5.$$

Kasutades üleelamisfunktsiooni (2) saame, et

$$e^{-\lambda(t(50))^\gamma} = 0{,}5$$

Avaldame nüüd tundmatu $t(50)$, et leida mediaan elukestusaeg:

$$-\lambda(t(50))^\gamma = \log 0{,}5 = -\log 2$$

$$t(50) = \sqrt[\gamma]{\left\{\frac{1}{\lambda}\log 2\right\}} = \left\{\frac{1}{\lambda}\log 2\right\}^{1/\gamma}. \tag{7}$$

Nüüd saame üldise valemi Weibulli jaotuse korral soovitud protsentiili leidmiseks:

$$t(p) = \left\{\frac{1}{\lambda}\log\left(\frac{100}{100-p}\right)\right\}^{1/\gamma}. \text{ [4, lk. 175]}$$

Seega kui meil on hinnatud mudeli (6) parameetrid $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_p, \hat{\lambda}_0, \hat{\gamma}$, siis saame leida *i*-ndale indiviidile prognoositud elukestuse vastavalt valemile (5) kui

$$\hat{S}_i(t) = e^{-e^{\widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_p} x_{pi}} \widehat{\lambda_0} t^{\hat{\gamma}}} \tag{8}$$

ja elukestuse mediaani vastavalt valemile (7) kui

$$\hat{t}(50) = \left\{\frac{\log 2}{\widehat{\lambda_0} e^{\widehat{\beta_1} x_{1i} + \widehat{\beta_2} x_{2i} + \cdots + \widehat{\beta_p} x_{pi}}}\right\}^{1/\hat{\gamma}}. \text{ [4, lk. 201]} \tag{9}$$

1.5. Weibulli jaotusega mudeli hindamine rakendustarkvara R abil

Rakendustarkvara R [3] kasutab elukestusandmete analüüsimisel paketti $survival$. Weibulli jaotusega mudeli hindamiseks kasutame funktsiooni

$$survreg(formula, data, dist = \text{“}weibull\text{“}, \dots),$$

kus $formula$ tähistab elukestusobjekti ($survival\ object$) ning $dist = \text{“}weibull\text{“}$ tähistab, et otsitav suurus on Weibulli jaotusega. [5]

Weibulli jaotuse parameetrite kindlaks tegemisel tuleb arvestada R eripära. Marili Zimmermanni magistritöös on välja toodud, et funktsiooni *rweibull* korral on kuju- ja skaalaparameeter tähistatud vastavalt *a* ja *b*. Lisaks on mainitud, et funktsioon *rweibull* kasutab Weibulli jaotuse parameetrite arvutamiseks seoseid: $\gamma = a$ ja $\lambda = (\frac{1}{b})^a$. [6, lk. 9] Funktsiooni *survreg* kasutades tuleb arvestada sellega, et funktsioon väljastab parameetrid $scale$ ja $intercept$, kus parameeter $scale$ on võrdne $\frac{1}{a}$ ning $intercept$ on võrdne $\log b$ [5].

Elukestusobjekti loob funktsioon $Surv(time, event)$, kus $time$ tähistab objekti elukestust ning $event$ tähistab indikaator tunnust, mille väärtus 0 tähistab tsenseeritud vaatlust ja 1 tähistab, et toimus sündmus parameetriga $time$ määratud hetkel. Anname elukestusobjektile juurde ka parameetrid, mida soovime mudelis kasutada: $Surv(time, event) \sim par1 + par2$. Kasutades leitud elukestusobjekti koos parameetritega funktsiooni *survfit* argumendina, saame Kaplan-Meieri hinnangu elukestuse mediaanile koos usaldusvahemikuga. Andes *survfit*'i ette funktsioonile *plot*, väljastatakse elukestuskõver. Lisaks saame funktsioonile *plot* ette anda ka parameetri *fun*. Märkides parameetri *fun* väärtuseks *„surv“*, saame tulemuseks tavalise elukestuskõvera. Valides väärtuseks *„log“*, *„cloglog“* või *„event“*, saame tulemuseks vastavalt logaritmitud elukestuskõvera, topelt logaritmitud elukestuskõvera, kus ka *x*-telg on logaritmitud, või $1 - \hat{S}(t)$. [5]

Funktsiooniga *survreg* mudeldades tuleb arvestada ka sellega, et mudeli koostamisel kasutatakse kiirendatud tõrkeaja mudelit ning hinnatakse parameetrite α_i-väärtused, kus *i=1,2,...,p* [7]. Eelnevalt nägime, et parameetrid $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ on omavahel seotud. Kasutades funktsiooni *WeibullReg*, mis vajab paketti *SurvRegCensCov* [8], hinnatakse võrdeliste riskide mudeli parameetrite β_i-väärtused, kus *i=1,2,...,p*. Lisaks leitakse ka Weibulli jaotuse kuju- ja skaalaparameetrid: λ_0 ja γ. Veel arvutab funktsioon *WeibullReg* välja ka riskisuhte (*hazard ratio*) $e^{\beta_i} = e^{-\frac{\alpha_i}{scale}}$ ja kiirenduse suhte (*event time ratio*) e^{α_i} koos usaldusvahemikega ning

samuti väljastatakse ka parameetrid $\boldsymbol{\alpha}$ ja $\boldsymbol{\beta}$ iga tunnuse tasemele koos standardhälbega ning *p*-väärtusega. [7]

Funktsioon $survreg$ tekitab *survreg*-objekti, millele saame omakorda rakendada funktsiooni *predict*. Viimase süntaks on *predict(object, newdata, type = c(„response“, „quantile“), p). Object* tähistab funktsiooni *survreg* poolt genereeritud mudelit ja *newdata* on andmestik, mille jaoks soovime prognoose leida. *Type*'iga täpsustame, millisel kujul antakse prognoos. Meie kasutame analüüsimiseks valikuid *response*, mis hindab igale objektile etteantud andmestikus prognoosi elueaks, ja *quantile*, mille korral tuleb lisaks ette anda ka parameeter *p*, mis väljendab kvantiile, mille järgi arvutatakse elukestus. [5] Näiteks *p=0,5* korral hinnatakse oodatud sündmuse toimumisaja mediaan, mis tähendab, et tõenäosusega 0,5 toimub oodatud sündmus enne prognoositud aega. Selleks, et saada hinnangud üleelamisfunktsioonile, anname parameetrile *p* ette arvud 0,01 kuni 0,99 ning *newdata*'le anname ette andmestiku koos fikseeritud mudeli parameetrite tasemetega. Seejärel saame iga kvantiili korral prognoosi elukestusele, mille saame kanda ka graafikule. Üleelamisfunktsiooni väärtuseid iseloomustades peame *y*-teljele kandma 1-*p* väärtused ja *x*-teljele leitud vastavad elukestused. [9]

1.6. Prognoosi täpsuse hindamine

Käesoleva töö eesmärk on töötada välja mudelipõhine meetod lepingu katkestamise prognoosimiseks. Enne selle meetodi praktikasse rakendamist on vaja teda valideerida sõltumatus andmestikus ja hinnata prognooside täpsust. Sõltumatu andmestik on vajalik, sest samas andmestikus mudeli väljatöötamisel ja prognooside testimisel on nn ülesobitamise oht – mudel kirjeldab osaliselt ka selles andmestikus leiduvat juhuslikku varieeruvust ja ei peegelda seaduspära. Seega on sellise töö puhul vaja jagada andmestik kaheks: testandmestik ja prognoosiandmestik. Testandmestiku abil hindame mudeli ja saadud mudeli põhjal prognoosime nn eluead prognoosiandmestiku jaoks.

Prognooside sobivuse kontrolliks on üheks võimaluseks seada mingi uuritava ajavahemiku piirmäär, millega võrrelda nii prognoosi kui vaadeldud aega, ja uurida, kui paljudel juhtudel langesid prognoos ja tegelikkus kokku (st mõlemad olid kas alla või üle selle piirmäära). Võtame selleks piiriks 366 päeva. Seejärel saame teada palju on meie mudeli puhul õigeid positiivseid tulemusi (*true positive*), õigeid negatiivseid tulemusi (*true negative*), valenegatiivseid tulemusi (*false negative*) ja valepositiivseid tulemusi (*false positive*). Sellisel

juhul meie andmete korral õige positiivne tulemus väljendab neid lepinguid, mis tegelikult kestsid alla aasta ja ka mudel hindas nende pikkuseks alla aasta. Õige negatiivne tulemus iseloomustab tulemusi, mis mõlemal juhul kestsid üle aasta. Valepositiivse tulemuse korral kehtib tegelikult leping üle aasta, kuid mudel hindab nende pikkuseks alla aasta, ning valenegatiivse tulemuse korral kehtib leping alla aasta, kuid mudel hindab pikkuse suuremaks kui 366 päeva. [10]

Mudeli täpsuse hindamisel hinnatakse kaks suurust: tundlikkus (*sensitivity*) ja spetsiifilisus (*specificity*). Tundlikkus hindab tõenäosuse, et kui vaatlusalusel leidub uuritav seisund, siis mudel annab positiivse tulemuse,

$$P(test\ on\ positiivne \mid objektil\ leidub\ uuritav\ seisund) = \frac{õige\ positiivsete\ arv}{objektide\ arv, kellel\ tegelikult\ leidub\ uuritav\ seisund},$$

ehk tõenäosus saada õige positiivseid tulemusi.

Spetsiifilisus hindab tõenäosust, et uuritava seisundi puudumisel annab ka mudel negatiivse tulemuse,

$$P(test\ on\ negatiivne \mid objektil\ ei\ leidu\ uuritavat\ seisundit) =$$

$$= \frac{õige\ negatiivsete\ arv}{objektid, kellel\ tegelikult\ ei\ leidu\ uuritavat\ seisndit},$$

ehk tõenäosus saada õige negatiivseid tulemusi.

Seega mudeli eksimist kirjeldavad valepositiivse,

$$P(test\ on\ positiivne \mid objektil\ ei\ leidu\ uuritavat\ seisundit) = 1 - spetsiifilisus,$$

ning valenegatiivse,

$$P(test\ on\ negatiivne \mid objektil\ leidub\ uuritav\ seisund) = 1 - tundlikkus,$$

tulemuste tõenäosused. [10]

Mudeli täpsuse graafiliseks väljendamiseks on kasutusel *ROC*-kõver. *ROC*-kõver väljendab tundlikkuse ja spetsiifilisuse väärtusi kõigi võimalike piirmäärade puhul. Võib tähele panna, et ühe parameetri suurendamisel vähendame teist. Viies tundlikkuse ligi 1 juurde, siis samal ajal suurendame valepositiivsete määra ning spetsiifilisust suurendades suurendame ka valenegatiivsete arvu. Seetõttu me soovime leida mudeli, mis võimalikult kõrge tundlikkuse ja spetsiifilisuse juures annab võimalikult vähe valenegatiivseid ja valepositiivseid tulemusi. Lisaks on joonisel diagonaal, mis iseloomustab olukorda, kus tõenäosus saada õige positiivne, õige negatiivne, valenegatiivne ja valepositiivne on kõigi korral 0,5. Olukorda iseloomustatakse sündmuse toimumise ennustamisega mündiviske teel ehk sellisel juhul loodud test ei oma eristusvõimet. [10]

Tarkvaraga R saame *ROC*-kõvera kasutades paketti *pROC* ning funktsiooni *roc*, millele anname ette binaarse tunnuse, mis iseloomustab, kas objektil leidub uuritav seisund või mitte. Valides parameetri väärtuseks $plot = TRUE$, mis väljastab meile *ROC*-kõvera. Funktsioon väljastab ka näitaja AUC (*Area Under the Curve*). [11]

AUC on mudeli headuse mõõt, mis halva mudeli korral on võrdne 0,5. Kokku on lepitud headuse piirid:

- $AUC \geq 0{,}9$, siis mudeli täpsus on suurepärane,
- $AUC \geq 0{,}8$, siis mudel hinnatakse heaks,
- $AUC \geq 0{,}7$, siis on mudel rahuldav, ning
- $AUC \geq 0{,}6$, siis on mudel kasin.

Alla 0,6 hinnatud AUCga mudeli prognoose ei ole mõistlik kasutada. [10]

2. Liikluskindlustuse andmete analüüs

Järgmises peatükis kirjeldame andmeid ja millisele kujule viime andmestiku enne mudeli koostamist. Seejärel kontrollime teooria sobivust andmetel ning koostame mudeli. Lisaks analüüsime mudelist saadud tulemusi. Mudeli valideerimiseks hindame prognoosiandmestiku peal prognooside täpsust. Andmete puhastamiseks kasutame rakendustarkvara R pakette *dplyr* [12] ja *tidyr* [13] ning tulemuste illustreerimiseks rakendustarkvara R paketti *ggplot2* [14].

2.1. Andmete kirjeldus

Meie andmete puhul on vaatlusalused Ergo kindlustuses sõlmitud lepingud aastatel 2015–2019, kus 2019 aasta andmed lõppevad oktoobrikuuga (30.10.2019), sest lõputöös analüüsitavad andmed võeti kasutusse novembris 2019. Algushetkena vaatleme me lepingu sõlmimise hetke ning kuna meid huvitav sündmus on lepingu katkestamine, siis me uurime aega kuni leping katkestatakse.

Liikluskindlustusleping koosneb lepingu komplektidest, kus komplekti pikkuseks on kuni 365 päeva ehk 1 aasta. Kui üks lepingu komplekt lõpetatakse ja sõlmitakse uus ning kui sinna vahele jääb maksimaalselt 70 päev, siis Ergo kindlustus loeb selle samaks lepinguks, mida on uuendatud. Vastasel juhul kui viimase lepingu komplekti lõpetamisest on möödunud enam kui 70 päeva ning sama klient on sõlminud samale sõiduvahendile uue lepingu, siis neid me sama lepingu jätkuna analüüsi arvesse ei võta.

Kasutatavas andmestikus on igal real kirjeldatud ühe komplekti tunnused ning uus komplekt avaldub uuel real. Andmestiku veergudes on lepingut, kindlustusvõtjat ja kindlustatavat sõiduvahendit iseloomustavad tunnused.

Lepingut iseloomustavad tunnused:

- „Lepingu alguskuupäev“ – esimese komplekti loomisekuupäev,
- „Lepingu komplekti alguskuupäev“,
- „Lepingu komplekti lõppkuupäev“,
- „Osamakse tüüp“ – ühekordne, kuu-, kvartali-, poolaasta- või aastamakse,
- „Komplekti sõlmimise kanal“ – jae-, e- või maaklerkanal.

Kindlustusvõtjat iseloomustavad tunnused:

- „ID“ – unikaalne kliendile loodud kood,
- „Vanus“,

- „Sugu“,
- „Isik“ – füüsiline või juriidiline isik,
- „Püsikliendi staatus“ – tava-, püsi- või ekspertklient,
- „BonusMalus“ – kliendi varasem liikluskindlustuse ajalugu, mida iseloomustavad kategooriad 1–14,
- „Maakond“ – maakond, kuhu kindlustaja on registreeritud,
- „Postiindeks“,
- „On soodustus“ – binaarne tunnus, mis iseloomustab, kas kliendile tehti lepingu sõlmimisel lisasoodustust.

Sõiduvahendit iseloomustavad tunnused:

- „VIN-kood“,
- „Mark“,
- „Sõiduvahendi kategooria“ – sõiduauto, veoauto, haagis, traktor, mootorratas või buss,
- „Sõiduvahendi vanus“,
- „Riskipiirkond“ – sõiduvahendi registreerimispiirkond.

Kasutatud andmestikus on kindlustaja keskmine vanus 53 aastat ning kõige noorem kindlustaja on 18-aastane ja vanim 102-aastane. Kindlustusvõtjatest 77% on mehed ning 23% naised. Kindlustusvõtjat iseloomustab veel tema varasem kokkupuude Ergo kindlustusega ehk kas tegemist on tavakliendiga või on tegemist juba varasema kindlustuse kliendiga. Meie andmetes on 91% klientidest tavakliendid ning 3% on püsikliendi ja 6% on ekspertkliendi staatuses. Andmetes leidub rohkem kliente, kes ei ole saanud lisasoodustust. Enamus analüüsitavatest lepingutest on aastamaksega või ühekordse maksega ja kõige vähem on kuumaksega lepinguid.

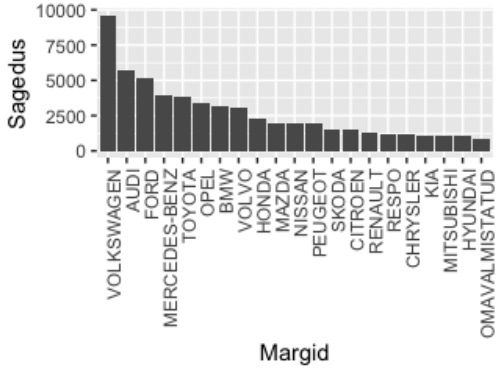
Tabel 1. Tunnuse „BonusMaluse“ jaotus andmestikus

„BonusMalus“ kategooria	**1–5**	**6–13**	**14**
Sagedus	49 745	24 576	307

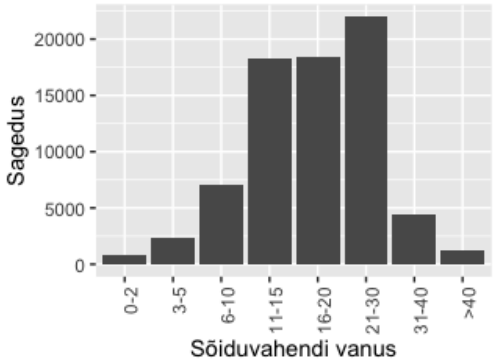
Kliendi varasemat liikluskindlustuse ajalugu kajastab tunnus „BonusMalus“, mis omab kategooriaid 1 kuni 14, kus 1 tähistab ilma kahjudeta klienti ja 14 rohkete kahjudega. Tunnuse „BonusMalus“ kategooriate jaotus on toodud tabelis 1, kus näeme, et kõige rohkem kliente on lepinguid luues väheste kahjudega.

Tabel 2. Sõiduvahendi kategooria jaotus andmetes

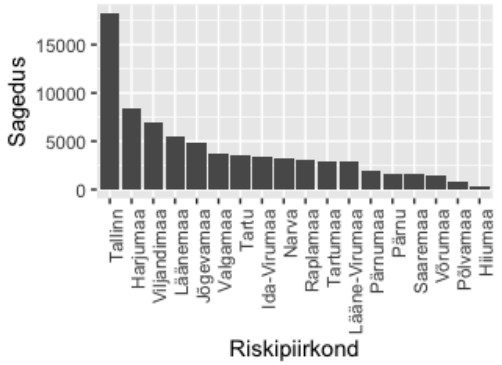
Sõiduvahendi kategooria	Sõiduauto	Haagis	Mootorratas	Traktor	Veoauto	Buss
Sagedus	60 549	6 516	4 514	1 872	1 052	125



Joonis 2. Sõiduvahendite markide jaotus andmetes



Joonis 3. Sõiduvahendi vanuse kategooriate jaotus andmetes



Joonis 4. Riskipiirkonna jaotus andmetes

Meie andmetes on kõige rohkem kindlustatud sõiduautosid ning kõige vähem busse (vt Tabel 2). Sõiduvahendite markide jaotus on kirjeldatud joonisel 2, kus on toodud 21 kõige populaarsemat marki. Võime näha, et kõige populaarseim mark on Volkswagen. Samuti on jõudnud kõige populaarsemate hulka ka omavalmistatud autod. Sõiduvahendi keskmine vanus on 18 aastat ning jooniselt 3 näeme, et enamus kindlustatud sõiduvahendeid jääb kategooriasse 11–30 aastat. Kõige vähem on andmetes uusi sõiduvahendeid ja üle 40 aasta vanuseid masinaid. Riskipiirkonna jaotus on kirjeldatud joonisel 4 ning näeme, et enamus kindlustatavatest sõiduvahenditest on registreeritud Tallinnas ja Harjumaal. Kõige vähem on kliente Hiiumaal.

2.2. Andmete puhastus

Soov on saada andmestik kujule, kus igale reale vastab üks objekt, kus objekt tähistab klienti ja talle kuuluvat sõidukit. Objekti unikaalne tunnus saadakse kliendi ID ja sõiduki VIN koodi liitmisel. Töö eesmärk on koostada mudel füüsilistele isikutele, kelle esimene lepingu komplekt on sõlmitud jaekanalis, mis tähendab, et leping sõlmiti Ergo kindlustuse müüjate poolt.

Andmestiku puhastamisel alustame sellest, et objekti kaupa leida aeg, mis jääb kahe komplekti vahele. Selleks tekitame andmestikku veeru, kuhu arvutame objekti kaupa lepingu komplektile järgneva ajavahemiku enne järgneva komplekti sõlmimist. Seejärel leiame need komplektid, mille korral komplektide vaheline aeg on suurem kui 70 päeva. Edasi jätame kõrvale need lepingu komplektid, millele eelnev ajavahe on suurem kui 70 päeva või talle eelnev komplekt on juba välja jäetud. See tähendab, et kui lepingu komplektide vahele on jäänud ühe korra rohkem kui 70 päeva, siis hilisemaid sõlmitud komplekte me analüüsiks enam ei kasuta, sest need ei kuulu uuendatud lepingute alla.

Antud töös analüüsime me füüsilisi isikuid, kelle jagame müügikanali järgi kolme rühma: jaekanalis, maaklerkanalis ja e-kanalis sõlmitud lepingud. Kanal võib aastatega muutuda, kuid analüüsiks kasutame me esimese lepingu komplekti sõlmimisel kasutatud kanalit. Andmetest jäetakse välja sootunnuseks „N/A“ märgitud isikud, sest need on reeglina välismaalased ning esindavad pigem erandlikke juhte.

Mudeli koostamiseks kasutame me teooria kohaselt esimese lepingu komplekti sõlmimisel kasutatud tunnuseid, mis näitavad, milliste tunnustega on klient Ergo kindlustusse asunud ning nende tunnuste pealt saame koostada prognoosimudeli. Lisaks klienti ja sõidukit iseloomustavatele tunnustele on vaja teada ka lepingu komplekti iseloomustavaid tunnuseid. Meid huvitab kaua kestavad lepingu komplektid objekti korral, et teada saada kui pikalt on objekt klient olnud. Lisaks kasutame lepingu lõpukuupäeva, mille järgi saame leida tsenseeritud lepingud. Tsenseeritud lepingud on lepingud, mis kestavad kauem kui 30.10.2019.

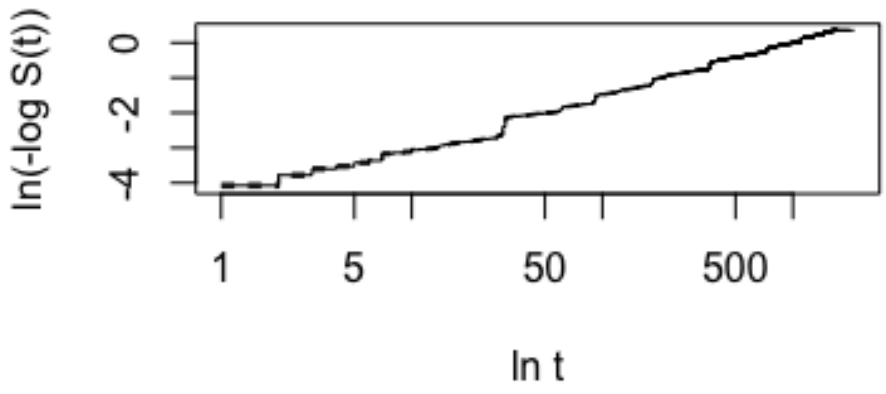
Eraldanud iga objekti kohta esimese lepingu komplekti, arvutame välja kliendi ja sõiduvahendi vanused. Sõiduvahendi vanused jagame kategooriatesse: $[0,2]$, $(2,5]$, $(5,10]$, $(10,15]$, $(15,20]$, $(20,30]$, $(30,40]$, $(40,\infty)$. Lisaks koondame kategooriatesse ka automargid, kus valime 21 populaarsemat: Volkswagen, Audi, Ford, Mercedes-Benz, Toyota, Opel, BMW,

Volvo, Honda, Mazda, Peugeot, Nissan, Citroen, Škoda, Renault, Chrysler, Mitsubishi, Kia, Respo, Hyundai ja Omavalmistatud. Veel tekitame sõiduki kategooriatest kuus gruppi: sõiduauto, mootorratas, haagis, traktor, buss ja veoauto.

Andmete korrastamise alguses soovisime mudeli tunnusena kasutada ka kindlustusvõtja maakonda, kuid selgus, et tunnuse andmekvaliteet on madal. Seejärel proovisime maakondi lisada postiindeksite kaudu, kuid selgus, et mingil osal lepingutest on postiindeks puudu ja mingil osal on valesti sisestatud. Seega otsustati kasutada riskipiirkonda, mis tähistab, millisesse maakonda on registreeritud sõiduk.

2.3. Andmete sobivus metoodikaga

Oleme viinud andmed soovitud kujule ning nüüd uurime andmete jaotust, et saaksime leida kõige paremini sobiva jaotuse. Funktsiooniga *survfit* Kaplan-Meieri hinnangu leides (vt Joonis 1), saame teada, et lepingu kestuse mediaan on 561 päeva koos usaldusvahemikuga (550; 576). Seega pooled lepingud lõppevad enne 561 päeva. Uurime, kas lepingu pikkused on kirjeldatavad Weibulli jaotusega. Teame, et sel juhul hajuvusgraafikule kantud $\ln[-\log \hat{S}(t)]$ ja $\ln t$ punktid peaksid asuma ühel sirgel (vt peatükk 1.4), kus *t* tähistab lepingu pikkust ja $\hat{S}(t)$ tähistab Kaplan-Meieri hinnangut üleelamisfunktsioonile. Kasutades *survfit* funktsiooni parameetriga *fun="cloglog"*, saame soovitud joonise.



Joonis 5. Hajuvusdiagramm Weibulli jaotuse korral.

Jooniselt 5 võime näha, et punktid asuvad enamvähem ühel sirgel, seega kasutame Weibulli jaotust mudeli hindamisel.

Järgmiseks hindame mudeli Weibulli jaotuse eeldusel. Esmalt hindame jaotuse parameetrid üldise mudeli korral, kuhu pole lisatud parameetreid. Saame tulemuseks, et $\lambda_0 = 0{,}01$ ja $\gamma = 0{,}66$. Vastavalt valemile (9) saame Weibulli jaotuse korral mediaaniks 615 päeva.

2.4. Mudel

Enne mudeli koostamist jagame oma andmestiku kaheks: test- ja prognoosiandmestikuks. Meie andmestikku jäi peale agregeerimist 74 628 objekti. Esiteks võtame välja mudeli tegemiseks 60 000 rida ning ülejäänud 14 628 rida jäävad prognoosiandmestikuks (vt Lisa 1), millele hindame hiljem prognoosid.

Lisame mudelisse tunnused:

- „Sugu“,
- „Osamakse tüüp“,
- „Püsikliendi staatus“,
- „BonusMalus“,
- „Mark“,
- „Sõiduvahendi vanus“,
- „Kindlustaja vanus“,
- „Riskipiirkond“,
- „On soodustus“ ja
- „Sõiduvahendi kategooria“.

Tabel 3. Tunnus „Püsikliendi staatus“ mudelis

Tunnus	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$	$\widehat{\boldsymbol{se}}(\boldsymbol{\beta}_{\boldsymbol{i}})$	$e^{\widehat{\beta}_i}$	***p*-väärtus**
Ekspertklient, ref=Tavaklient	-0,0103	0,024	0,99	0,6609
Püsiklient	-0,0300	0,030	0,97	0,3180

Tabelites 3 ja 4 on toodud nende parameetrite hinnangud saadud mudelist, mis vastavad faktortunnustele „Püsikliendi staatus“ ja „BonusMalus“. Esmalt näeme tabelist 3, et kliendile

omistatud staatuse korral tavakliendi risk katkestada ei erine oluliselt ekspertkliendist ja püsikliendist. Kuna tunnus osutus ebaoluliseks, jätame püsikliendi staatuse mudelist välja.

Tabel 4. Tunnus „BonusMalus“ mudelis

Tunnus	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$	$\widehat{\boldsymbol{se}}(\boldsymbol{\beta}_{\boldsymbol{i}})$	$\boldsymbol{e}^{\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}}$	***p*-väärtus**
„BonusMalus“ = 2 , ref=1	0,0161	0,0217	1,02	0,4588
„BonusMalus“ = 3	-0,0173	0,0296	0,98	0,5582
„BonusMalus“ = 4	0,0538	0,0236	1,06	0,0228
„BonusMalus“ = 5	0,0689	0,0294	1,07	0,0192
„BonusMalus“ = 6	0,1514	0,0231	1,16	< 0,0001
„BonusMalus“ = 7	0,2060	0,0222	1,23	< 0,0001
„BonusMalus“ = 8	0,3505	0,0211	1,42	< 0,0001
„BonusMalus“ = 9	0,2716	0,0161	1,31	< 0,0001
„BonusMalus“ = 10	0,2250	0,0494	1,25	< 0,0001
„BonusMalus“ = 11	0,2349	0,0677	1,26	0,0005
„BonusMalus“ = 12	0,3370	0,0340	1,40	< 0,0001
„BonusMalus“ = 13	0,4761	0,2186	1,61	0,0294
„BonusMalus“ =14	1,0381	0,0674	2,82	< 0,0001

Tunnus „BonusMalus“, mis iseloomustab kindlustusvõtjat, on järjestustunnus ning seetõttu ei pane me teda mudelisse pideva tunnusena. Kategooriat 1 ehk kliendid, kellel puuduvad eelnevad kahjud, võrdleme teistega. Tabelist 4 näeme, et „BonusMalus“ tasemed 1, 2 ja 3 ei erine oluliselt üksteisest. Vaadates, kuidas kasvab risk katkestada, siis näeme, et sarnased mõjud on tasemetel 1–5 ning 6–13. Eriliselt paistab silma tase 14 ehk rohkete kahjudega kliendid. Saadud informatsiooni põhjal jagame tunnuse „BonusMaluse“ gruppidesse: 1–5, 6–13 ja 14.

Hindame allesjäänud tunnustega mudeli, kasutades R-is funktsiooni *WeibullReg* (vt Lisa 1), mis koos parameetri $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$ väärtustega hindab parameetrid: $\lambda_0 = 0{,}0008$ ja $\gamma = 0{,}862$.

Tabel 5. Võrdeliste riskide mudeliga hinnatud parameetrite väärtused

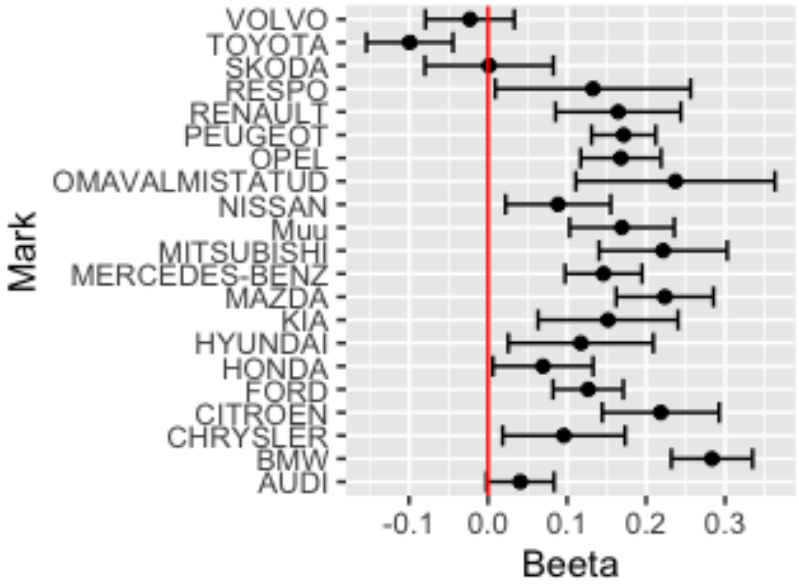
Tunnus	$\widehat{\boldsymbol{\beta}}_{\boldsymbol{i}}$	$\widehat{\boldsymbol{se}}(\boldsymbol{\beta}_{\boldsymbol{i}})$	$e^{\widehat{\beta}_i}$	***p*-väärtus**
„Sugu“ = Naine, ref = Mees	-0,043	0,012	0,96	0,0004
„Osamakse tüüp“ = Kuu, ref = Aasta	0,651	0,033	1,92	< 0,0001
„Osamakse tüüp“ = Ühekordne	1,775	0,014	5,90	< 0,0001
„Osamakse tüüp“ = Kvartal	0,341	0,017	1,41	< 0,0001
„Osamakse tüüp“ = Poolaasta	0,150	0,017	1,16	< 0,0001
„BonusMalus“ = 6–13, ref= 1–5	0,244	0,011	1,28	< 0,0001
„BonusMalus“ = 14	1,020	0,067	2,77	< 0,0001
„Mark“ = Audi, ref = Volkswagen	0,041	0,022	1,04	0,0609
„Mark“ = BMW	0,283	0,026	1,33	< 0,0001
„Mark“ = Chrysler	0,096	0,040	1,10	0,0151
„Mark“ = Citroen	0,219	0,038	1,24	< 0,0001
„Mark“ = Ford	0,127	0,022	1,14	< 0,0001
„Mark“ = Honda	0,069	0,032	1,07	0,0317
„Mark“ = Hyundai	0,117	0,047	1,12	0,0121
„Mark“ = Kia	0,152	0,045	1,16	0,0008
„Mark“ = Mazda	0,224	0,031	1,25	< 0,0001
„Mark“ = Mercedes-Benz	0,146	0,025	1,16	< 0,0001
„Mark“ = Mitsubishi	0,221	0,041	1,25	< 0,0001
„Mark“ = Nissan	0,088	0,034	1,09	0,0090
„Mark“ = Omavalmistatud	0,237	0,064	1,27	0,0002
„Mark“ = Opel	0,168	0,026	1,18	< 0,0001
„Mark“ = Peugeot	0,169	0,034	1,18	< 0,0001
„Mark“ = Muu	0,171	0,019	1,19	< 0,0001
„Mark“ = Renault	0,165	0,040	1,18	< 0,0001
„Mark“ = Respo	0,132	0,062	1,14	0,0349
„Mark“ = Škoda	0,001	0,041	1,00	0,9700
„Mark“ = Toyota	-0,099	0,028	0,91	0,0004
„Mark“ = Volvo	-0,023	0,029	0,98	0,4236
„Sõiduvahendi vanus“ = 3–5, ref = 0–2	0,625	0,112	1,87	< 0,0001
„Sõiduvahendi vanus“ = 6–10,	0,909	0,107	2,48	< 0,0001
„Sõiduvahendi vanus“ = 11–15	1,007	0,107	2,74	< 0,0001

Tunnus	$\widehat{\beta}_i$	$\widehat{se}(\beta_i)$	$e^{\widehat{\beta}_i}$	*p*-väärtus
„Sõiduvahendi vanus“ = 16–20	1,213	0,107	3,36	< 0,0001
„Sõiduvahendi vanus“ = 21–30	1,449	0,107	4,26	< 0,0001
„Sõiduki vanus“ = 31–40	1,584	0,108	4,88	< 0,0001
„Sõiduvahendi vanus“ = >40	1,715	0,113	5,56	< 0,0001
„Kindlustaja vanus“	-0,011	0,000	0,99	< 0,0001
„Riskipiirkond“ = Harjumaa, ref = Tallinn	-0,061	0,017	0,94	0,0004
„Riskipiirkond“ = Hiiumaa	-0,106	0,072	0,90	0,1427
„Riskipiirkond“ = Ida-Virumaa	-0,036	0,025	0,96	0,1474
„Riskipiirkond“ = Jõgevamaa	-0,205	0,023	0,81	< 0,0001
„Riskipiirkond“ = Lääne-Virumaa	-0,214	0,028	0,81	< 0,0001
„Riskipiirkond“ = Läänemaa	-0,178	0,021	0,84	< 0,0001
„Riskipiirkond“ = Narva	0,021	0,024	1,02	0,4014
„Riskipiirkond“ = Pärnu	-0,067	0,034	0,93	0,0485
„Riskipiirkond“ = Pärnumaa	-0,102	0,032	0,90	0,0015
„Riskipiirkond“ = Põlvamaa	-0,205	0,049	0,81	< 0,0001
„Riskipiirkond“ = Raplamaa	-0,166	0,027	0,85	< 0,0001
„Riskipiirkond“ = Saaremaa	-0,249	0,038	0,78	< 0,0001
„Riskipiirkond“ = Tartu	-0,139	0,025	0,87	< 0,0001
„Riskipiirkond“ = Tartumaa	-0,100	0,027	0,90	0,0002
„Riskipiirkond“ = Valgamaa	-0,191	0,024	0,83	< 0,0001
„Riskipiirkond“ = Viljandimaa	-0,111	0,019	0,89	< 0,0001
„Riskipiirkond“ = Võrumaa	-0,148	0,036	0,86	< 0,0001
„On soodustus“ = 1, ref = 0	-0,059	0,011	0,94	< 0,0001
„Sõiduki kategooria“ = Buss, ref = Sõiduauto	0,374	0,116	1,45	0,0012
„Sõiduki kategooria“ = Haagis	-0,661	0,029	0,52	< 0,0001
„Sõiduki kategooria“ = Mootorratas	0,296	0,022	1,34	< 0,0001
„Sõiduki kategooria“ = Traktor	-0,736	0,042	0,48	< 0,0001
„Sõiduki kategooria“ = Veoauto	0,206	0,040	1,23	< 0,0001

Tabelis 5 on välja toodud mudeli parameetrite hinnangud, mis vastavad võrdeliste riskide mudelile. Iga mudelis oleva tunnuse juures on välja toodud parameetri β_i hinnang koos standardhälbega ja olulisustõenäosusega. Lisaks on välja toodud ka riskisuhe $e^{\widehat{\beta}_i}$, mis hindab

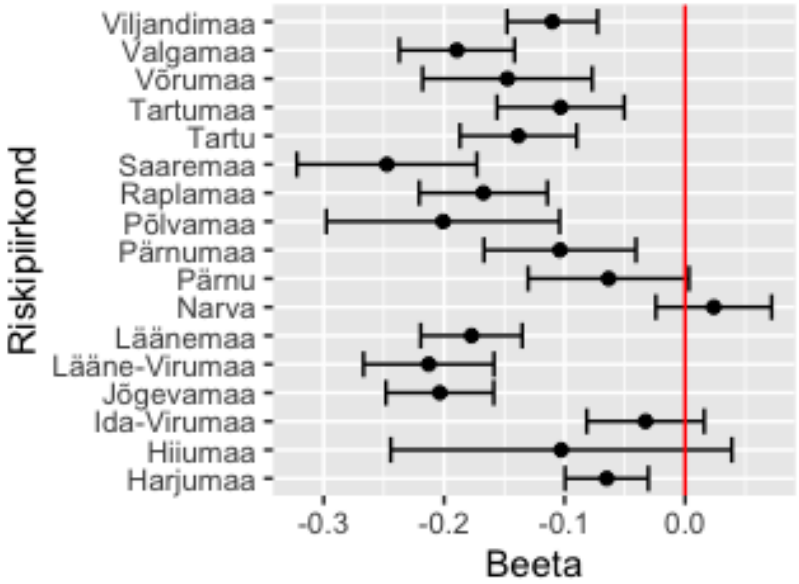
tunnuse puhul tema riski fikseeritud tasemel võrreldes aluseks oleva tasemega. Interpreteerimisel on oluline ka parameetri $\hat{\beta}_i$ märk, mis näitab negatiivse korral, et fikseeritud taseme korral risk väheneb võrreldes referentsiks võetud tasemega. Positiivse parameetri $\hat{\beta}_i$ korral, et fikseeritud taseme korral risk kasvab. Näiteks vaatame tunnust „Sugu“, kus on võrreldud mehi ja naisi. Näeme, et meeste puhul on risk katkestada 1,04 korda suurem kui naistel.

Mudelist loeb veel välja seda, et erinevate osamakse tüüpidega lepingute puhul on kõige riskantsemad ühekordse maksega lepingud, mille korral on risk katkestada mingil ajahetkel *t* peaaegu 6 korda suurem kui aastamaksega lepingu puhul. Lisaks on teistest erinev kuumaksega leping, millel on võrreldes aastase maksega lepingu korral kaks korda suurem risk katkestada. Veel paistab silma, et sõiduvahendi vanuse kasvades kasvab ka risk leping katkestada. Tulemus on ootuspärane, sest üle 40 aasta vanused sõidukid arvatavasti kantakse juba maha. Kindlustusvõtja vanuse puhul on näha, et vanuse kasvades risk katkestada väheneb. Seega võib järeldada, et vanemad inimesed ei vaheta nii tihti sõidukeid kui noored või on nad ühe sõiduvahendi puhul lojaalsemad kliendid. Uurides tunnuse „Sõiduvahendi kategooria“ tasemeid, on näha, et võrreldes sõiduautoga on traktoril ja haagisel poole väiksem risk katkestada.



Joonis 6. Parameetri beeta hinnangud koos usaldusvahemike tunnuse „Mark“ korral (vt Lisa 2)

Joonisele 6 on kantud tunnuse „Mark“ erinevatele tasemetele hinnatud parameetri β_i hinnangud koos usaldusvahemikega. Punane joon tähistab $x = 0$ ning tasemeid võrreldakse Volkswageniga. Näeme, et võrreldes Volkswageniga on madalam risk katkestada ainult Toyota omanikel. Kõige rohkem sarnanevad Volkswagenile Audi, Volvo ja Škoda. Kõige kõrgem risk katkestada, võrreldes Volkswageniga, on BMW omanikel. Lisaks on kõrgem risk ka omavalmistatud masina puhul, kuid neil on ka kõige suurem hajuvus, mis võib tuleneda sellest, et neid on andmetes vähem kui teisi (vt Joonis 2).



Joonis 7. Parameetri beeta hinnangud koos usaldusvahemike tunnuse „Riskipiirkond“ korral

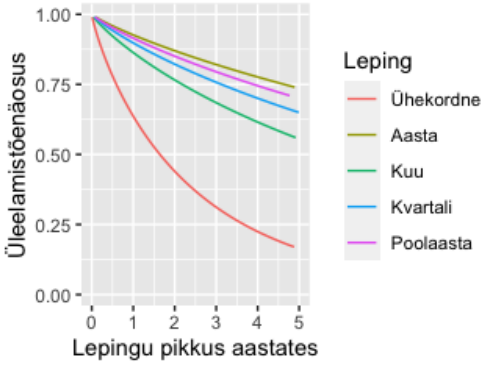
Joonisel 7 on kujutatud parameetri β_i hinnangud koos usaldusvahemikega tunnusele “Riskipiirkond“. Punane joon tähistab $x = 0$ ning tasemeid võrreldakse Tallinnaga. Näeme, et Tallinnaga sarnane risk on Pärnus, Hiiumaal, Ida-Virumaal ja Narvas. Teiste maakondade ja linnade korral on riskitase madalam. Kõige madalam risk, võrreldes Tallinnaga, on Saaremaal ning kõige suurema hajuvusega on Hiiumaa, mis on andmetes vähe esindatud piirkond (vt Joonis 4).

2.5. Tulemused mudelist

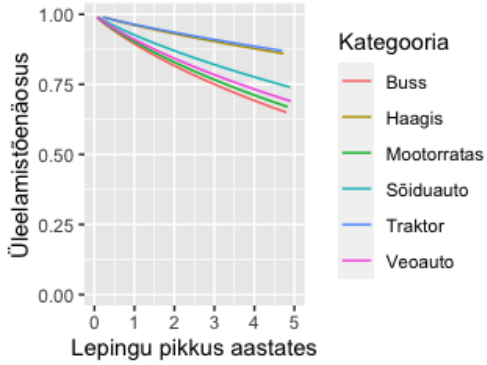
Hinnatud Weibulli mudel võimaldab leida prognoose üleelamisfunktsioonile, vastavalt valemile (8) ja kasutades tarkvara R funktsiooni *predict* (vt peatükk 1.5). Vaatleme nüüd prognoositavaid üleelamistõenäosuseid mudelis kajastuvate tunnuste kaupa. Iga tunnuse puhul on hinnatud üleelamisfunktsiooni lepingul, millel kõigi teiste tunnuste väärtused on fikseeritud baastasemele:

- „Sugu“ = Mees,
- „Osamakse tüüp“ = Aasta,
- „BonusMalus“ = 1–5,
- „Mark“ = Volkswagen,
- „Sõiduvahendi vanus“ = 0–2,
- „Vanus“ = 53,
- „Riskipiirkond“ = Tallinn,
- „On soodustus“ = 0 ehk klient ei ole saanud soodustust,
- „Sõiduvahendi kategooria“ = sõiduauto (vt Tabel 6).

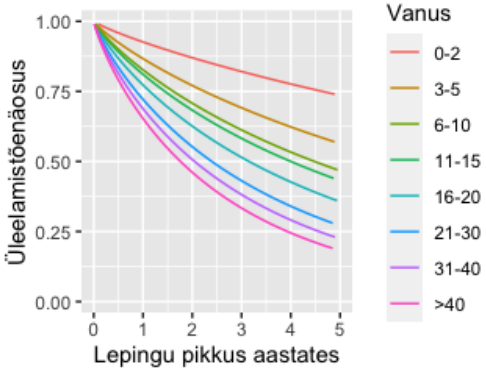
Järgnevatel joonistel on näha üleelamistõenäosused erinevate mudelis olevate tunnuste lõikes ning võrreldakse ühe tunnuse tasemete erinevust. Eraldi pole käsitletud tunnuseid „Riskipiirkond“ ja „Mark“, kuna need tunnused illustreeriti eelnevalt joonistel 6 ja 7.



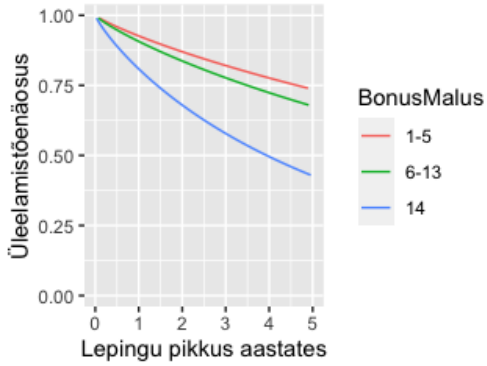
Joonis 8. Prognoosid tunnuse „Osamakse tüüp“ tasemetel (vt Lisa 3)



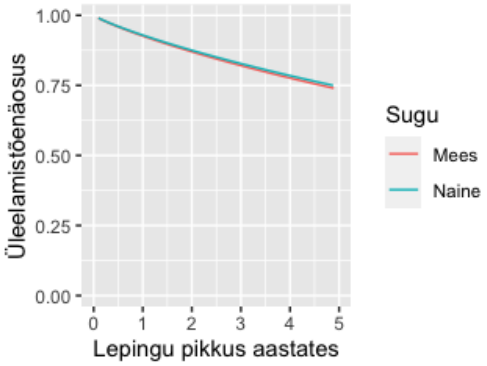
Joonis 9. Prognoosid tunnuse „Sõiduvahendi kategooria“ tasemetel



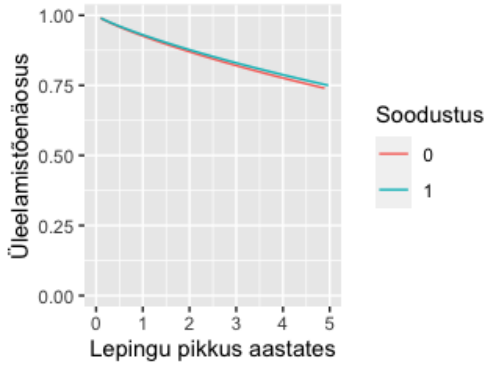
Joonis 10. Prognoosid tunnuse „Sõiduki vanus“ tasemetel



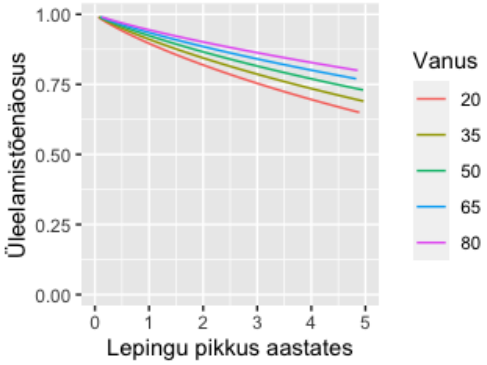
Joonis 11. Prognoosid tunnuse „BonusMalus“ tasemetel



Joonis 12. Prognoosid meeste ja naiste puhul



Joonis 13. Prognoosid tunnuse „On soodustus“ korral



Joonis 14. Prognoosid kindlustusvõtja vanuse tasemetel

Jooniselt 8 näeme, et kõige pikema elueaga on aastased lepingud, mille korral tõenäosus, et leping kestab üle 4 aasta on 0,75. Lisaks on näha, et mudel prognoosib kõige madalamat lepingu pikkust ühekordse maksega lepingute puhul. Nende lepingute korral tõenäosus katkestada tõuseb juba esimese aasta jooksul märgatavalt ning tõenäosus kesta üle 4 aasta on ligikaudu 0,25. Samal ajal tõenäosus katkestada enne 4 aastat on ligikaudu 0,75. Veel paistab silma, et poolaastamaksega lepingud ei erine oluliselt aastamaksega lepingutest. Aastamaksega võrreldes on suurem tõenäosus katkestada peale aasta möödumist ka kvartali- ja kuumaksega lepingute puhul. Seega võime järeldada, et ühekordse maksega kliendid on pigem lühemaajalised kliendid ning aasta- ja poolaastamaksega on pikaajalisemad kliendid.

Joonisel 9 võime näha üleelamistõenäosuseid erinevate sõiduvahendi kategooriate kaupa. Näeme, et kõige kindlamad ja pikaajalisemad lepingud on traktoritel ja haagistel, kellel ka viie aasta möödudes ei ole üleelamistõenäosus langenud alla 0,75. Lisaks paistab jooniselt 9, et kõige madalamad prognoosid saavad bussid, mootorrattad ja veoautod, kuid ka nende puhul tõenäosus kesta üle nelja aasta on ligikaudu 0,7. Andmestikus oli kõige rohkem lepinguid sõlmitud sõiduautodega (vt Tabel 2) ning selgub, et tõenäosus kesta üle 4 aasta on enam kui

0,75. Seega kõige lojaalsemad kliendid on traktorite ja haagiste omanikud, kuid ka sõiduautode puhul ei ole katkestamistõenäosus väga suur.

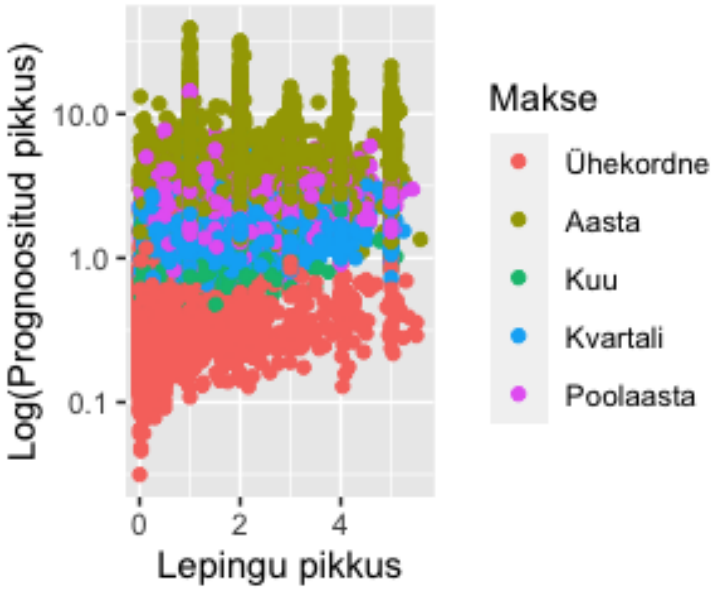
Uurides tunnuse „Sõiduvahendi vanus“ kategooriatele prognoositud tõenäosuseid, siis jooniselt 10 on näha ootuspäraselt, et sõiduvahendi vanuse kasvades suureneb ka tõenäosus katkestada leping. Tuletame meelde, et hetkel on tunnus „Sõiduvahendi kategooria“ fikseeritud tasemele „Sõiduauto“. Näeme, et uutele autodele, ehk vanusega kuni kahe aastased, prognoositakse kõige pikemad lepinguid ning tõenäosus kesta üle 4 aasta on enam kui 0,75. Samas 3–5 aastaste autode seas on näha langus, võrreldes kuni kahe aastaste autodega, kuid prognoosid on pigem pikemaajalised. Madalamad prognoosid saavad endale üle 20 aasta vanused autod. Üle 40 aasta vanuste autode puhul on näha, et tõenäosus katkestada enne 4 aastat on 0,75. Veel paistab silma, et üle 6 aastaste autode puhul katkestamistõenäosus kasvab kiiresti. Seega võime järeldada, et kõige pikemate lepingutega on uued autod, millel vanust on kuni 2 aastat, ning kõige lühemad lepingud on üle 20 aastastel autodel.

Tuletame meelde, et tunnus „BonusMalus“ kirjeldab kindlustusvõtja varasemat ajalugu. Jooniselt 11 näeme, et üleelamistõenäosused langevad kategooriate järjestuses ehk kõige pikemad lepingud prognoositakse väheste juhtumitega klientidele ja kõige lühemad prognoositakse rohkete juhtumitega klientidele. Näeme veel, et kategooriaga 14 hinnatud klientide puhul on tõenäosus kesta üle 4 aasta 0,5. Järelikult lühemad lepingud on pigem rohkete varasemate kahjudega klientidel.

Jooniselt 12 näeme, et naiste ja meeste puhul on langustrend üleelamistõenäosusel praktiliselt sama. Võib väita, et meeste ja naiste korral suurt erinevust lepingute pikkuste osas ei leidu. On näha, et üle 4 aasta on üleelamistõenäosus 0,75. Eelnevalt fikseeritud teiste tasemete puhul ei mõjuta sugu lepingu pikkust tugevalt. Samasugust tulemust näeme ka joonisel 13, kus on kirjeldatud tunnust „On soodustus“. Tegemist on binaarse tunnusega, kus 1 tähistab, et klient on saanud lepingu sõlmides soodustust, ja 0 tähistab, et soodustust ei saadud. Nagu mudelist nägime (vt Tabel 6), et riskisuhe soodustuse saamisel ja mitte saamisel on peaaegu võrdsed ning ka jooniselt 13 näeme, et soodustuse saamisel on prognoosid veidi kõrgemad, kuid prognoositud üleelamistõenäosused on väga sarnased. Järelikult otseselt selle tunnuse põhjal ei suuda me lepinguid eristada. Kindlustusvõtja vanuse puhul nägime tabelist 6, et hinnatud β_i väärtus oli negatiivne, seega vanuse kasvades risk katkestada väheneb. Ka jooniselt 14 näeme, et prognoositud üleelamistõenäosused langevad vanuse vähenedes.

2.6. Prognoosid

Soovime hinnatud mudelit kasutades prognoosida uutele lepingutele nende kestust. Kasutades prognoosiandmestikku hindame igale lepingule üleelamisfunktsiooni mediaani vastavalt valemile (9) ja kasutades R funktsiooni *predict* (vt Lisa 4). Lisaks mediaanile uurisime, kas mõne teise kvantiili kasutamine, st parameetri *p* varieerimine annaks paremaid tulemusi, kuid selgus, et kõige parem kokkulangevus prognoositud lepingu pikkuse ja tegeliku lepingu pikkusega saavutatakse siiski mediaani, ehk p=0,5 korral.



Joonis 15. Prognooside ja lepingute pikkus. (vt Lisa 5)

Teeme prognooside ja tegelike pikkuste illustreerimiseks graafiku. Joonisel 15 on kujutatud prognoosid logaritmitud skaalal ja lepingu pikkus aastates. Näeme, et aastase maksega lepingute puhul prognoositakse nende pikkused kõige pikemaks. Samal ajal tegelikest pikkustest selgub, et ka aastaste lepingute puhul ilmneb katkestamist enne aastat. Kõige madalamad prognoosid saavad ühekordse maksega lepingud, kuid tegelikult leidub üksikud lepingud isegi üle 4 aastase pikkusega. Üldiselt on näha ka trend prognoosimisel: aastasemaksega saavad kõige pikemad prognoosid, siis poolaasta-, kvartali-, kuumaksega, ning kõige madalamad ühekordse maksega lepingud, seda trendi nägime ka joonisel 8.

Tabel 6. Tegelikud ja prognoositud lepingu pikkused kategooriates „Kuni 1 aasta“ ja „Üle 1 aasta“

Tegelik \ Prognoositud pikkus	**Kuni 1 aasta**	**Üle 1 aasta**	**Kokku**
Kuni 1 aasta	4203	2996	7199
Üle 1 aasta	560	6869	7429
Kokku	4763	9865	14 628

Tabelis 6 võrdleme tegelikke ja prognoositud lepingu pikkuseid, kasutades lõikepunktina ühte aastat (vt Lisa 6) – st vaatame tegelikkuse ja prognoosi kokkulangevust selles osas, kas leping kestis ja oli prognoositud kestma kuni 1 aasta või üle selle. Näeme, et ligikaudu 76% juhtudel lähevad prognoosid ja tegelikkus kokku ning 24% juhtudest hinnatakse mudeli poolt valesti. Tegelikest üle aasta kestnud lepingutest 92% prognoositakse samuti üle aastasteks. Samal ajal kui leping kestab kuni 1 aasta, siis 58% puhul prognoositakse ka lepingu kestuseks alla aasta, kuid 42% puhul on prognoositud kestus üle aasta. Järelikult on pikemate lepingute korral leitud prognoosid päris head ning lühikeste lepingute korral hindab mudel rohkem valesti.

Kui vaadelda antud prognoosi kui testi, kas leping võidaks katkestada aasta jooksul, siis selle testi tundlikkus oleks 0,58 ja spetsiifilisus 0,92. Kõigist lepingutest, mille kestuseks on prognoositud üle aasta, kestab tegelikult üle aasta ligikaudu 70% ning kõigist kuni 1 aasta kestma prognoositud lepingutest, katkestatakse reaalsuses aasta jooksul 88%. Seega kui leping prognoositakse kestma alla aasta, siis 12% puhul kestab leping üle aasta ning kui leping prognoositakse kestma üle aasta, siis 30% puhul kestab leping alla aasta. Järelikult meie mudeli poolt hinnatavad prognoosid kirjeldavad päris hästi reaalsust ning lepingu sõlmimisel leitud prognoos annab aimu, kas uue kliendiga võiks olla lühi- või pikaajaline leping.

Tabel 7. Tegelikud ja prognoositud lepingu pikkused kategooriates „Kuni 1 aasta“, „1–2 aastat“ ja „üle 2 aasta“

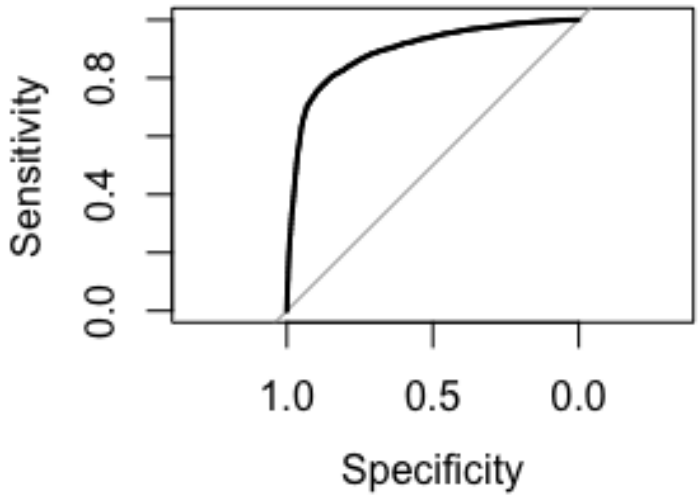
Tegelik\Prognoositud pikkus	**Kuni 1 aasta**	**1– 2 aastat**	**Üle 2 aasta**	**Kokku**
Kuni 1 aasta	4203	1126	1870	7199
Aasta – 2 aastat	279	579	1472	2330
Üle 2 aasta	281	810	4008	5099
Kokku	4763	2515	7350	14 628

Tabelist 7 näeme, et tegelikkuses kuni aasta pikkuste lepingute korral on ka prognoos 58% juhtudest õige, 16% puhul prognoositakse lepingu kestuseks 1–2 aastat ning 26% puhul üle 2 aasta. Reaalsuses üle 2 aasta kestnud lepingute puhul on prognoos samasugune 79% juhtudest ning vaid 6% puhul on neid hinnatud alla aastasteks. Kõige ebatäpsem on prognoos lepingute puhul, mille pikkus jääb aasta ja kahe aasta vahele. Selliste lepingute korral prognoositakse õigesti ainult 25%. Samas ei hinnata neid ka alla aastasteks, vaid 63% juhtudest hinnatakse pikaajalisteks lepinguteks. Järelikult lühiajalised ning pikaajalised lepingud hinnatakse üsna täpselt ning ebatäpsust leidub 1–2 aastaste lepingute korral.

Kui leping on prognoositud kestma alla aasta, siis näeme, et reaalsuses on see ka nii läinud 88% juhtudest, kusjuures 6% puhul on leping tegelikkuses kestnud 1–2 aastat ja 6% puhul veelgi kauem. 1–2 aastat kestma prognoositud lepingute puhul kestab reaalsuses leping sama kaua vaid 23% juhtudest, kusjuures tervelt 45% juhtudest katkestatakse leping enne aastat ning 32% juhtudest kestab leping üle 2 aasta. Lepingutest, mille kestuseks on prognoositud enam kui 2 aasta, kestab reaalsuses 55% üle 2 aasta, kuid 25% katkestatakse aasta jooksul ning 20% 1–2 aasta jooksul peale sõlmimist.

Tabeleid 6 ja 7 koos vaadates saame kokku võtta, et 9865 lepingu puhul, mille prognoositud pikkus on üle aasta kestab 30% tegelikkuses alla aasta, 21% 1–2 aastat ning 49% üle 2 aasta. Järelikult leitud mudel aitab üsna hästi leida kliente, kes suure tõenäosusega katkestavad aasta jooksul. Samuti eristada pikemaajalisi kliente kuid täpsemate hinnangute andmisel, näiteks kas leping kestab pigem 1–2 aastat või üle 2 aasta, on eksimise tõenäosus suhteliselt suur.

Hindame leitud prognooside täpsust ka ROC-kõvera ning näitaja AUC abil. Seame ette piiri 367 päeva ehk etteantav tunnus *roc* funktsioonile kirjeldab, kas tegelik lepingu pikkus on väiksem kui 367 (tähistatult 1) või suurem (tähistatult 0).



Joonis 16. ROC-kõver.

Joonisel 16 on *x*-teljele kantud spetsiifilisus ja *y*-teljele tundlikkus. Näeme, et tundlikkuse 0,6 korral on spetsiifilisus ligikaudu 0,9 ning tundlikkuse 0,8 korral on spetsiifilisus umbes 0,8. Ka kõvera järgi näeme, et leitud mudelil on eristusvõime. Vaatleme ka funktsiooni *roc* poolt väljastatud näitajat AUC, mis on meie andmete korral võrdne 0,8084. Seega võime lugeda oma mudeli heaks (vt peatükk 1.6).

Kokkuvõte

Bakalaureusetöö eesmärk oli analüüsida lepingute katkestamist liikluskindlustuse andmetel ning iseloomustada kliente, kellel on suurem tõenäosus katkestada leping. Töö käigus kasutati elukestusanalüüsi meetodeid, mida kirjeldati 1. peatükis. Seejärel puhastati ja agregeeriti andmed ning tehti kindlaks, et analüüsil võib kasutada Weibulli jaotuse eeldust. Puhastatud andmetele loodi mudel, mis iseloomustab füüsilisest isikust kliente, kes sõlmisid lepingud jaekanalis. Mudeli testimiseks leiti testandmestikule prognoosid, millega hinnati ka mudeli täpsust.

Töö käigus selgus, et kõik valitud tunnused, välja arvatud „Püsikliendi staatus“, kirjeldasid riski katkestada. Saadi teada, et lühiajalised lepingud on ühekordse maksega, võrreldes aastase maksega lepingutega. Selgelt tuli välja, et noored on riskantsemad kliendid. Kliendid, kellel on varasemalt rohkelt kahjusid tekitatud liikluses, on samuti pigem lühiajalised kliendid võrreldes väheste kahjudega klientidega. Veel selgus, et võrreldes Volkswageni omanikega on suurim risk katkestada leping BMW omanikel ning väiksema riskiga lõpetada leping on Toyota omanikel. Ootuspäraselt on kõrgema katkestamise riskiga sõiduvahendite registreerimispiirkonnaks Tallinn.

Töö tulemusena valminud prognoosimudel võimaldab hinnata potentsiaalset lepingu pikkust uute klientide puhul. Prognoosiandmestiku analüüsist selgus, et kui saadud mudeliprognoosi kohaselt on tõenäoline lepingu katkestamine aasta jooksul, siis 88% juhtudest see nii ka läheb. Kui aga uue kliendi puhul prognoositakse pigem üle aasta kestvat lepingut, siis reaalsuses kestabki leping enam kui aasta 70% juhtudest. Seega on saadud mudeli prognoosiväärtus üsna hea just selleks, et hinnata võimalikku katkestamist aasta jooksul. Detailsemate prognooside korral, mis eristasid lepingu hinnangulist kestust: 1–2 aasta ja üle 2 aasta, enam nii head kokkulangevust ei leitud.

Antud töö on kindlustuse jaoks esmane taoline analüüs nende andmete korral, seetõttu on tehtud ka mõningaid lihtsustusi – näiteks võttes üheks objektiks ühe sõiduki. Edasi võiks kindlasti uurida andmeid kliendi baasil, mis enamasti tähendab, et kliendil on mitu sõiduvahendit kindlustatud. Veel võiks uurida, kuidas mõjutavad tekkinud kahjud katkestamistõenäosust. Antud töö käigus uuriti füüsilisi isikuid, kes on jaekanali kliendid. Võimalik on saadud tulemusi võrrelda maaklerkanali ja e-kanali klientide puhul ning ka

juriidiliste isikute korral. Antud teemale on võimalik läheneda ka erinevate katkestamisliikide põhjal, see eeldab, et andmete korral suudetakse eristada katkestamisliigid nagu näiteks sõiduvahendi mahakandmine, omaniku vahetus ja kindlustuse vahetus.

Kasutatud kirjandus

[1] Maanteeameti koduleheküIg (2020). https://www.mnt.ee/et/ametist/statistika/soidukite-statistika [02.05.2020]

[2] Eesti kindlustusseltside liidu koduleheküIg (2020). https://www.lkf.ee/et/kindlustamise-tavad/kohustuslik-liikluskindlustus#section0 [02.05.2020]

[3] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/ [05.05.2020]

[4] Collett, D. (2015). Modelling Survival Data in Medical Research Third Edition. Suurbritannia: CRC Press.

[5] Therneau, T. M., Lumley, T., Atkinson, E. ja Crowson, C. (2020). Package „survival". https://cran.r-project.org/web/packages/survival/survival.pdf [30.04.2020]

[6] Zimmermann, M. (2018). Elukestusanalüüs vasakult tõkestatud andmete ning ajast sõltuva argumenttunnuse korral TÜ Eesti geenivaramu kohordi näitel. Magistritöö. Tartu Ülikool: matemaatika ja statistika instituut.

[7] Haile, S. R. (2015). Weibull AFT Regression Functions in R. https://cran.r-project.org/web/packages/SurvRegCensCov/vignettes/weibull.pdf [03.02.2020]

[8] Hubeaux, S. ja Rufibach, K. (2015). Package „SurvRegCensCov". https://cran.r-project.org/web/packages/SurvRegCensCov/SurvRegCensCov.pdf [30.04.2020]

[9] Wollschlaeger, D. (2013). Survival analysis: Parametric proportional hazards models. http://dwoll.de/rexrepos/posts/survivalParametric.html [27.03.2020]

[10] Kaart, T. (2012). Binaarsete tunnuste analüüsimeetodid. Õpiobjekt. Eesti Maaülikool. http://www.eau.ee/~ktanel/bin_tunnuste_analyys/pt35.php [08.04.2020]

[11] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, I., Sanchez, J.-C., Müller, M., Siegert, S. ja Doering, M. (2020). Package „pROC". https://cran.r-project.org/web/packages/pROC/pROC.pdf [05.05.2020]

[12] Wickham, H., Francois, R, Henry, L. ja Müller, K. (2020). dplyr: A grammar of Data Manipulation. https://cran.r-project.org/web/packages/dplyr/index.html [05.05.2020]

[13] Wickham, H. ja Henry, L. (2020). tidyr: Tidy Messy Data. https://cran.r-project.org/web/packages/tidyr/index.html [05.05.2020]

[14] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org. [05.05.2020]

Lisad

Lisa 1. Andmestiku lahutamine treening- ja testandmestikuks ning mudeli koostamine

```
library(survival)
library(dplyr)
d <- data.frame(andmed)
d1 <- sample_n(d,60000,replace = F) #testandmestik, millel
loome mudeli
l <- d$Objekt %in% d1$Objekt
d2 <- d[!l,] #prognoosiandmestik

#Mudel
mudel <- WeibullReg(Surv(aeg,1- tsens) ~ Sugu +
as.factor(Osamakse_tyyp) +
relevel(as.factor(BonusMalus),ref="1-5") +
relevel(as.factor(Mark),ref="VOLKSWAGEN") + Soiduvahendi_vanus
+ vanus + relevel(as.factor(Riskipiirkond),ref="Tallinn") +
as.factor(On_Soodustus) +
relevel(as.factor(Soiduvahendi_kategooria),ref="S6iduauto"),
data=d1)
```

Lisa 2. Parameetri beeta hinnangute illustreerimine usaldusvahemikega

```
m.coef <- mudel$coef
beeta.margid <- as.numeric(as.character(m.coef[10:29,1]))
sd.margid <- as.numeric(as.character(m.coef[10:29,2]))
cim <-
data.frame(cbind(c("Audi","BMW","CHRYSLER","CITROEN","FORD",
"HONDA","HYUNDAI","KIA","MAZDA","MERCEDES-BENZ",
"MITSUBISHI","NISSAN","OMAVALMISTATUD","OPEL","Muu","RENAULT",
"RESPO","SKODA","TOYOTA","VOLVO"),
as.numeric(as.character(beeta.margid)),
beeta.margid-1.96*sd.margid,beeta.margid+1.96*sd.margid))

margid <- data.frame(cim$X1,as.numeric(as.character(cim$X2)),
                     as.numeric(as.character(cim$X3)),
as.numeric(as.character(cim$X4)))

colnames(margid) <- c("Mark","Beeta","Alumine","Ylemine")
library(ggplot2)
ggplot(margid, aes(Mark,Beeta))+
  geom_errorbar(aes( ymin=Alumine,ymax=Ylemine)) +
  geom_point()+
  geom_hline(yintercept = 0, color="red")+
  coord_flip()
```

Lisa 3. Mudeliga lepingu üleelamistõenäosuse hindamine

```
fiks_andmed <- data.frame(Sugu=c("Mees", "Mees", "Mees",
"Mees", "Mees"),
Osamakse_tyyp=factor(c("Annually","Monthly", "One_payment",
"Quarterly", "Semi-annually"), levels =
levels(d1$Osamakse_tyyp)),
BonusMalus=factor(c("1-5","1-5", "1-5", "1-5", "1-5"), levels
= levels(d1$BonusMalus)),
Margid=c("VOLKSWAGEN", "VOLKSWAGEN",
"VOLKSWAGEN","VOLKSWAGEN","VOLKSWAGEN"),
Soiduvahendi_vanus =factor(c("0-2","0-2", "0-2","0-2","0-2"),
levels = levels(d1$Soiduvahendi_vanus)),
 vanus=c(53,53,53,53,53),
                    Riskipiirkond=c("Tallinn", "Tallinn",
"Tallinn","Tallinn","Tallinn"),
                    On_Soodustus=c(0,0,0,0,0),

Soiduvahendi_kategooria=factor(c("S6iduauto", "S6iduauto",
"S6iduauto","S6iduauto","S6iduauto"),
                                          levels =
levels(d1$Soiduvahendi_kategooria)))
percs <- (1:99)/100
FWeib <- predict(testmudel, newdata=fiks_andmed,
type="quantile", p=percs, se=TRUE)

m <- as.data.frame(FWeib$fit/365.25)
m$kategooria <-
c("Aasta","Kuu","Ühekordne","Kvartali","Poolaasta")
colnames(m) <- c(seq(0.01,0.99,0.01),"Leping")

s <- m %>%
  gather(key = toen, value = value, colnames(m)[1:ncol(m)-1])

s %>%
  ggplot(aes(x=value,y=1-as.numeric(as.character(toen)),
colour=Leping))+
  geom_line()+
  xlab("Lepingu pikkus aastates")+
  ylab("Üleelamistõenäosus")+
  xlim(c(0,5))
```

Lisa 4. Mudeliga lepingu pikkuse prognoosimine

```
testmudel <- survreg(Surv(aeg,1- tsens) ~ Sugu +
as.factor(Osamakse_tyyp) +
relevel(as.factor(BonusMalus),ref="1-5") +

relevel(as.factor(Mark),ref="VOLKSWAGEN") + Soiduvahendi_vanus
+ vanus + relevel(as.factor(Riskipiirkond),ref="Tallinn") +
                        as.factor(On_Soodustus) +
relevel(as.factor(Soiduvahendi_kategooria),ref="S6iduauto"),da
ta=d1)

test_pr.q <-
predict(testmudel,newdata=d2,type="quantile",p=c(0.5))
```

Lisa 5. Prognoositud ja tegelike lepingu pikkuste jaotus

```
koos <-
data.frame(cbind(as.vector(d2$aeg),as.vector(test_pr.q)))
koos$Makse <- as.character(d2$Osamakse_tyyp)

f <- as.factor(koos$Makse)
nimed <- c("Aasta","Kuu","Ühekordne","Kvartali","Poolaasta")
levels(f) <- nimed
koos$Makse <- as.character(f)

jaotus <- koos %>%
  ggplot(aes(x=X1/365.25,y=X2/365.25, colour=Makse))+
  geom_point()+
  scale_y_log10()+
  xlab("Lepingu pikkus")+
  ylab("Log(Prognoositud pikkus)")
```

Lisa 6. Tegelikud ja prognoositud lepingu pikkused kategooriates „Kuni 1 aasta“ ja „Üle ühe aasta“

```
aeg2 <- cut(as.numeric(d2$aeg), c(-1,366,Inf))
prog2 <- cut(test_pr.q, c(-1,366,Inf))
prognoosid.3 <- as.matrix(table(aeg2,prog2))
dimnames(prognoosid.3) <- list(c("Aasta", ">1a "),
                               c("Aasta", ">1a"))
names(dimnames(prognoosid.3)) <- c("Tegelik aeg","Prognoositud
aeg")
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kadi Kilgi

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Liikluskindlustuslepingute pikkuste prognoosimine Weibulli mudelite abil“, mille juhendajad on Krista Fischer ja Raine Talvet, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

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

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

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

Kadi Kilgi
13.05.2020