Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Perttu Narvik

Sõidukite kindlustuskahjude sageduse hindamine mitmemõõtmelise adaptiivse regressioonsplaini abil

Kindlustus- ja finantsmatemaatika eriala
Magistritöö (30 EAP)

Juhendajad: Meelis Käärik
Tõnis Maldre (Salva kindlustus)

Tartu 2019

Sõidukite kindlustuskahjude sageduse hindamine mitmemõõtmelise adaptiivse regressioonsplaini abil

Magistritöö eesmärgiks on tutvustada mitmemõõtmelist adaptiivset regressioonisplaini ning rakendada seda meetodit Eesti liikluskindlustuse andmetele. Töö esimeses osas antakse ülevaade üldistatud lineaarsest mudelist ja üldistatud aditiivsest mudelist loendusandmete korral ning kirjeldatakse mitmemõõtmelise adaptiivse regressioonisplaini kasutamist nii parameetrilisel kui ka mitteparameetrilisel viisil. Töö teises osas sobitatakse tutvustatud meetodide põhjal mudelid, mis prognoosivad kindlustuskahjude sagedust ning selgitatakse välja parim mudel.

Märksõnad: *kindlustusmatemaatika, sõidukikindlustus, statistilised mudelid, üldistatud lineaarsed mudelid*

P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika

Estimation of Claim Frequency for Vehicle Insurance by Using Multivariate Adaptive Regression Spline

The aim of this thesis is to introduce multivariate adaptive regression spline and apply it to Estonian vehicle insurance data. In the first section an overview is given about generalized linear model and generalized additive model for count data and multivariate adaptive regression spline is described for both non-parametric and parametric models. In the second section the models are fitted to predict claim frequency by using the methods that are introduced in the first section and the best model is chosen.

Keywords: *actuarial mathematics, vehicle insurance, statistical models, generalized linear models*

P160 Statistics, operation research, programming, actuarial mathematics

Sisukord

Sissejuhatus

Kindlustusreservide hindamisel on väga oluline prognoosida võimalikult täpselt nii kahjude suurust kui ka kahjude esinemise sagedust. Väga levinud meetod kahjude sageduse hindamiseks on üldistatud lineaarne mudel, kuid mitte-lineaarsete seoste modelleerimiseks on välja töötatud ka alternatiivseid meetodeid. J. Friedman tutvustas 1991. aastal mitmemõõtmelist adaptiivset regressioonisplaini (edaspidi ka MARS), mille abil on võimalik leida mitte-lineaarseid seoseid. MARSi algoritm teostab ka muutujate valikut.

Magistritöö eesmärgiks on tutvustada mitmemõõtmelist adaptiivset regressioonisplaini, rakendada seda meetodit Eesti liikluskindlustuse andmetele ning võrrelda saadud tulemusi üldistatud lineaarsete mudelitega.

Töö esimene pool koosneb kolmest teoreetilisest peatükist. Esimene peatükk annab ülevaate lihtsamatest üldistatud lineaarsetest mudelitest loendusandmete prognoosimiseks. Töö teises peatükis kirjeldatakse üldistatud aditiivset mudelit ja selle mudeli hindamise algoritmi Poissoni jaotuse korral. Kolmandas peatükis tutvustatakse mitmemõõtmelist adaptiivset regressioonisplaini ning algoritmi, kuidas selliseid mudeleid hinnatakse.

Töö teine pool sisaldab esimeses osas kirjeldatud meetodite rakendamist Eesti liikluskindlustuse andmetele. Mudelid sobitatakse kasutades logaritmseosefunktsiooni ning Poissoni ja negatiivse binoomjaotuse eeldusi. Peatükis kirjeldatakse R-i statistikapaketti `earth`, mille abil on võimalik MARSi mudeleid hinnata.

Magistritöö on vormistatud tekstitöötlusprogrammi LaTeXveebiversioonis *Over-*

leaf. Andmete analüüsiks, mudelite sobitamiseks ja jooniste tegemiseks on kasutatud statistikatarkvara R versiooni 3.5.3.

Autor tänab juhendajaid Tõnis Maldret huvitava teema tõstatamise ja praktiliste nõuannete eest ning dotsent Meelis Käärikut toetava ja samaaegselt nõudliku suhtumise eest.

1 Üldistatud lineaarne mudel

Järgnev peatükk põhineb teostel „Generalized Linear Models" (McCullagh ja Nelder, 1989) ning „Introduction to Linear Models" (Montgomery jt., 2012). Selles peatükis antud tähistused kehtivad ka kõigis järgnevates peatükkides.

Olgu uuritavaks tunnuseks juhuslik suurus Y, mille jaotus kuulub jaotuste eksponentsiaalsesse perre. Üldistatud lineaarseks mudeliks nimetatakse mudelit, mis avaldub kujul

$$g[\mu(\boldsymbol{X})] = \eta(\boldsymbol{X}), \tag{1}$$

kus

$$\eta(\boldsymbol{X}) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{2}$$

- $\eta(\boldsymbol{X})$ on lineaarne prediktor,
- j on argumenttunnuse indeks, p on argumenttunnuste arv, $j = 1, \ldots, p$,
- X_j on juhuslik suurus, mis tähistab j-ndat argumenttunnust,
- $\boldsymbol{X}$ on juhuslike suuruste X_j vektor, $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$,
- β_0, β_j on tundmatud parameetrid, $j = 1, ..., p$,
- $\mu(\boldsymbol{X})$ on uuritava tunnuse Y tinglik keskväärtus, $\mu(\boldsymbol{X}) = E(Y|\boldsymbol{X})$,
- $g(\cdot)$ on seosefunktsioon.

Töös kasutatakse ka muutujat i, mis tähistab vaatluse indeksit. Vaatluste arvu tähistab N, seega i võimalikud väärtused on $i = 1, \ldots, N$.

Seosefunktsiooniks nimetatakse funktsiooni, mille argumendiks on uuritava

tunnuse tinglik keskväärtus ning see funktsioon seob omavahel lineaarse prediktori ning tingliku keskväärtuse. Kõige enam kasutatud seosefunktsioonid on

- identsusseos, mille korral

$$g[\mu(\boldsymbol{X})] = \mu(\boldsymbol{X}), \text{ seega } \mu(\boldsymbol{X}) = \eta(\boldsymbol{X}),$$

- log-seos, mille korral

$$g[\mu(\boldsymbol{X})] = \ln \mu(\boldsymbol{X}), \text{ seega } \mu(\boldsymbol{X}) = e^{\eta(\boldsymbol{X})},$$

- logit-seos, mille korral

$$g[\mu(\boldsymbol{X})] = \ln \frac{\mu(\boldsymbol{X})}{1-\mu(\boldsymbol{X})}, \text{ seega } \mu(\boldsymbol{X}) = \frac{e^{\eta(\boldsymbol{X})}}{1+e^{\eta(\boldsymbol{X})}}.$$

Logit-seost kasutatakse juhul, kui uuritav tunnus on Bernoulli jaotusega $Y \sim Be(\mu)$ ning soovitakse hinnata sündmuse esinemise tõenäosust, kuna logit-seose mudeli prognoosid on alati nulli ja ühe vahel. Log-seost kasutatakse sageli juhul, kui uuritavaks tunnuseks on loendusandmed, kuna log-seose rakendamisel on mudeli prognoosid alati mittenegatiivsed.

Üldistatud lineaarne mudel avaldub log-seose korral järgmiselt:

$$\ln[\mu(\boldsymbol{X})] = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \tag{3}$$

seega tinglik keskväärtus $\mu(\boldsymbol{X})$ avaldub kujul

$$\mu(\boldsymbol{X}) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j X_j) = \exp(\beta_0) \cdot \prod_{j=1}^{p} \exp(\beta_j X_j). \tag{4}$$

Tuntuim jaotus loendusandmete modelleerimiseks on jaotuste eksponentsiaalsesse perre kuuluv Poissoni jaotus, mille tõenäosusfunktsioon avaldub kujul

$$f(y;\mu) = \frac{e^{-\mu}\mu^y}{y!},$$

kus $y = 0, 1, 2, ...$ ning parameeter $\mu > 0$.

Poissoni jaotuse korral on juhusliku suuruse Y keskväärtus ja dispersioon võrdsed jaotuse parameetriga,

$$EY = \mu \text{ ja } DY = \mu.$$

Kui mudeli sobitamisel kasutatakse Poissoni jaotuse eeldust, tuleb kontrollida, kas antud tingimus kehtib. Hajuvust kirjeldatakse sageli skaalaparameetri $\varphi = DY/EY$ kaudu. Kui mudel on korrektne, siis $\varphi \approx 1$. Kui $\varphi < 1$ ehk $DY < EY$, siis on tegemist alahajuvusega, vastupidisel juhul, kui $\varphi > 1$ ehk $DY > EY$, on tegemist ülehajuvusega. Alahajuvuse esinemine pole oluline probleem, kuid ülehajuvusele tuleb tähelepanu pöörata. Ülehajuvus võib tekkida, kui Poissoni protsess toimub juhusliku pikkusega intervallis. Samuti võib esineda ülehajuvust, kui andmetes on liigselt nulle või nullid üldse puuduvad. Tugeva ülehajuvuse korral tuleks Poissoni jaotuse asemel kasutada teisi jaotusi.

Alternatiivina kasutatakse loendusandmete modelleerimisel negatiivset binoomjaotust, mille tõenäosusfunktsioon avaldub kujul

$$f(y; k,\pi) = \frac{\Gamma(k+y)}{y!\Gamma(k)}\pi^k(1-\pi)^y = \binom{k+y-1}{y}\pi^k(1-\pi)^y, \qquad (5)$$

kus $y = 0, 1, 2, ...$, parameetri π väärtus on vahemikus $0 < \pi < 1$ ja parameeter $k > 0$ on täisarv. Juhuslikku suurust Y võib vaadelda kui ebaõnnestumiste arvu kuni k-nda õnnestumiseni Bernoulli protsessis ning π tähistab õnnestumise tõenäosust ühel katsel.

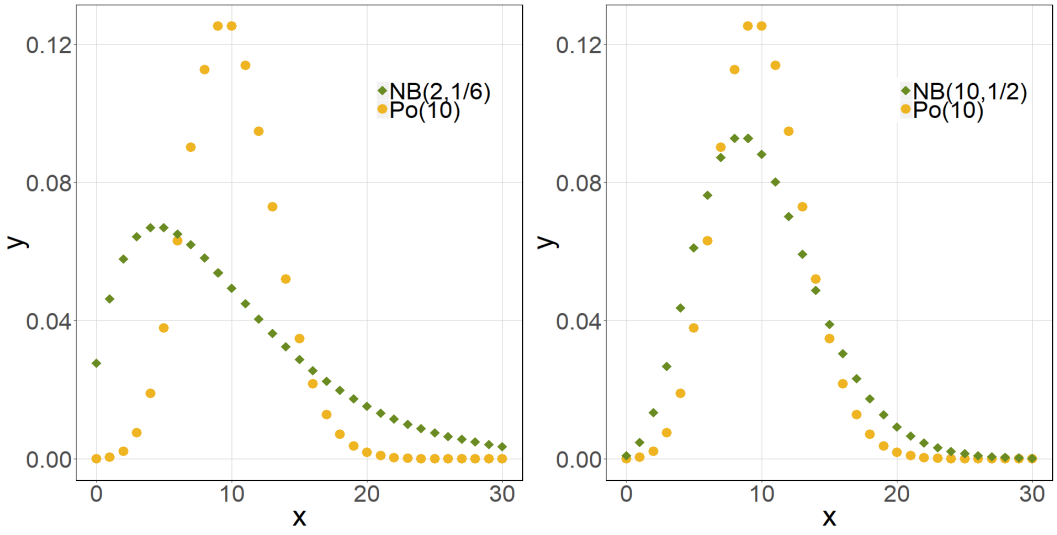
Negatiivse binoomjaotuse korral juhusliku suuruse Y keskväärtus avaldub kujul

$$EY = \mu = \frac{k(1-\pi)}{\pi}$$

ning dispersioon avaldub kujul

$$DY = \frac{k(1-\pi)}{\pi^2} = \mu + \frac{\mu^2}{k}.$$

Negatiivse binoomjaotuse tõenäosusfunktsioon on raskema sabaga kui Poissoni tõenäosusfunktsioon ning nullide esinemise tõenäosus on suurem. Joonisel 1 on kujutatud Poissoni tõenäosusfunktsiooni kollaste punktidega ja negatiivse binoomjaotuse tõenäosusfunktsiooni roheliste rombidega. Kõigil juhtudel on keskväärtuseks $\mu = 10$. Vasakpoolsel joonisel on negatiivse binoomjaotuse parameetri k väärtuseks 2 ning parempoolsel 10. Mida suurem väärtus parameetrile k omistatakse, seda sarnasemaid tulemusi Poissoni jaotus ja negatiivne binoomjaotus annavad.



Joonis 1. Poissoni ja negatiivse binoomjaotuse tõenäosusfunktsioonid, $\mu = 10$

1.1 Üldistatud lineaarse mudeli hindamine

Üldistatud lineaarse mudeli hindamiseks kasutatakse suurima tõepära meetodit. Tõepärafunktsioon avaldub kujul

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{i=1}^{N} f_i(y_i; \boldsymbol{\beta}),$$

kus

- y_i on juhusliku suuruse Y realisatsioon,
- $\boldsymbol{y} = (y_1, ..., y_N)$ on juhusliku suuruse Y realisatsioonide vektor,
- $\boldsymbol{\beta}$ on hinnatavate parameetrite vektor, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)$,
- $f_i(y_i; \boldsymbol{\beta})$ on juhusliku suuruse Y tõenäosusfunktsioon.

Suurima tõepära meetodi eesmärk on leida parameetrite väärtused selliselt, et tõepärafunktsioon saavutab maksimaalse väärtuse. Kuna logaritmfunktsioon on monotoonne, siis log-tõepärafunktsioon saavutab maksimaalse väärtuse samas kohas kui tõepärafunktsioon. Seetõttu maksimeeritakse praktikas sageli tõepärafunktsiooni asemel log-tõepärafunktsiooni, mis avaldub kujul

$$l(\boldsymbol{\beta}; \boldsymbol{y}) = \ln L(\boldsymbol{\beta}; \boldsymbol{y}) = \ln \prod_{i=1}^{N} f_i(y_i; \boldsymbol{\beta}) = \sum_{i=1}^{N} \ln f_i(y_i; \boldsymbol{\beta}).$$

Olgu $l_i(\boldsymbol{\beta}; y_i) = \ln f_i(y_i; \boldsymbol{\beta})$, skoorifunktsioon on defineeritud kui

$$s(\boldsymbol{\beta}; \boldsymbol{y}) = \sum_{i=1}^{N} \frac{\partial l_i(\boldsymbol{\beta}; y_i)}{\partial \boldsymbol{\beta}}.$$

Maksimaalne tõepära saavutatakse siis, kui $s(\boldsymbol{\beta}; \boldsymbol{y}) = 0$. (Hastie jt., 2017, lk 265-267)

Kui eeldatakse, et juhuslik suurus Y on tinglikult Poissoni jaotusega, siis tõepärafunktsioon avaldub kujul

$$L(\boldsymbol{\beta}; \boldsymbol{y}) = \prod_{i=1}^{N} f_i(y_i; \boldsymbol{\beta}) = \prod_{i=1}^{N} \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}$$

ning log-tõepärafunktsioon kujul

$$\begin{aligned} l(\boldsymbol{\beta};\boldsymbol{y}) = & \sum_{i=1}^{N} \ln \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!} = \sum_{i=1}^{N}(-\mu_i + y_i \ln \mu_i - \ln(y_i!)) = \\ = & \sum_{i=1}^{N} y_i \ln \mu_i - \sum_{i=1}^{N} \mu_i - \sum_{i=1}^{N} \ln(y_i!) \end{aligned}$$

kus $\mu_i = g^{-1}(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})$ ja x_{ij} on suuruse X_j realisatsioon indeksiga i. (Montgomery jt., 2012, lk 445)

Negatiivse binoomjaotuse tõenäosusfunktsioon parameetritega π ja k on antud valemiga 5. Jaotuse keskväärtuse valemist $\mu = \frac{k(1-\pi)}{\pi}$ järeldub $\pi = \frac{k}{\mu+k}$, seega valimi tõepärafunktsiooni saab avaldada kujul

$$L(\boldsymbol{\beta}, k; \boldsymbol{y}) = \prod_{i=1}^{N} \frac{\Gamma(k+y_i)}{y_i!\Gamma(k)} \left(\frac{\mu_i}{k} + 1\right)^{-k} \left(1 + \frac{k}{\mu_i}\right)^{-y_i}$$

ning log-tõepärafunktsioon on sellisel juhul

$$\begin{aligned} l(\boldsymbol{\beta},k; \boldsymbol{y}) = & \sum_{i=1}^{N} \ln \left[\frac{\Gamma(k+y_i)}{y_i!\Gamma(k)} \left(\frac{\mu_i}{k} + 1\right)^{-k} \left(1 + \frac{k}{\mu_i}\right)^{-y_i} \right] = \\ = & \sum_{i=1}^{N} \left[\ln \Gamma(k+y_i) - \ln \Gamma(k) - \ln(y_i!) - k \ln \left(\frac{\mu_i}{k} + 1\right) - y_i \ln \left(1 + \frac{k}{\mu_i}\right) \right]. \end{aligned}$$

Kuna tõepärafunktsioon sisaldab lisaks μ_i-le ka parameetrit k, siis tuleb hinnangud leida nii kordajatele $\boldsymbol{\beta}$ kui ka parameetrile k. Olgu $\boldsymbol{\theta} = (\beta_0, ..., \beta_p, k)$. Negatiivse binoomjaotuse korral saavutatakse maksimaalne tõepära, kui skoorifunktsioon $s(\boldsymbol{\theta}; \boldsymbol{y}) = 0$.

2 Üldistatud aditiivne mudel

Üldistatud aditiivne mudel on edasiarendus üldistatud lineaarsest mudelist. Sarnaselt lineaarsele mudelile kehtib aditiivsus, kuid erinevus seisneb selles, et üldistatud aditiivse mudeli korral on võimalik igale argumenttunnusele rakendada mitte-lineaarseid funktsioone. Üldistatud aditiivset mudelit on võimalik rakendada nii kvalitatiivsele kui ka kvantitatiivsele uuritavale tunnusele. (James jt., 2017, lk 282)

Üldistatud aditiivse mudeli korral on uuritava tunnuse tinglik keskväärtus seotud aditiivse prediktoriga seosefunktsiooni abil. Mudeli üldkuju avaldub järgmiselt:

$$g[\mu(\boldsymbol{X})] = \beta_0 + \sum_{j=1}^{p} f_j(X_j) = \beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p), \qquad (6)$$

kus f_j, $j = 1, .., p$ on sile mitteparameetriline funktsioon, $\beta_0 + \sum_{j=1}^{p} f_j(X_j)$ on aditiivne prediktor ning ülejäänud tähistused on defineeritud peatükis 1. Funktsiooni nimetatakse siledaks, kui see on määramispiirkonnas lõpmatult diferentseeruv ning mistahes järku tuletis on pidev. (Hastie jt., 2017, lk 296-297)

2.1 Aditiivse mudeli hindamine

Käesoleva peatüki aluseks on teos „Generalized Additive Models: Some Applications" (Hastie ja Tibshirani, 1987, lk 372-373).

Mittelineaarsete efektide modelleerimise aluseks on hajuvusdiagrammi silu-

jad. Silumisfunktsioon on selline funktsioon, mis annab tulemuseks funktsiooni f hinnangu $\hat{f}$. Silumisfunktsioon tuleb enne mudeli hindamist fikseerida.

Olgu meil identsusseos $g[\mu(\boldsymbol{X})] = \mu(\boldsymbol{X})$, siis aditiivne mudel on kujul

$$\mu(\boldsymbol{X}) = \beta_0 + \sum_{j=1}^{p} f_j(X_j). \tag{7}$$

Sellise aditiivse mudeli hindamiseks kasutatakse üldist *backfitting* algoritmi. Aditiivse mudeli hindamise eesmärk tuleb tinglikust keskväärtusest. Korrektse mudeli korral kehtib

$$E\Big(Y - \beta_0 - \sum_{j \neq k} f_j(X_j)\Big|X_k\Big) = f_k(X_k) \quad \forall k,\, k = 1, ..., p.$$

Backfitting algoritm leiab lahendi sellisele võrrandisüsteemile iteratiivselt. Igal tsükli sammul asendatakse tinglik osajääkide keskväärtus ühemõõtmelise silujaga.

Sageli kasutatakse silumisfunktsioonina kuupsplaini silujat. *Backfitting* algoritm langeb sellisel juhul kokku karistatud vähimruutude meetodiga (*penalized least squares method*), kus minimeeritakse suurust

$$\sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} f_j(x_{ij}) \right)^2 + \sum_{j=1}^{p} \lambda_j \int f_j''(t_j)^2 dt_j, \tag{8}$$

kus $\lambda_j \geq 0$ on silumisparameeter, $y_1, y_2, ..., y_N$ on valimi väärtused uuritava tunnuse kohta ning $x_{1j}, x_{2j}, ..., x_{Nj}$ on valimi väärtused j-nda seletava tunnuse kohta, $j = 1, ..., p$.

Esmalt leitakse vabaliikme β_0 hinnang, milleks on valimi väärtuste y_1,...y_N aritmeetiline keskmine ning see mudeli hindamise protsessis rohkem ei muutu. Tsükli esimesel sammul leitakse y_i osajäägid e_i iga $j = 1, ..., p$ korral, et eemaldada kõigi teiste tunnuste mõju uuritavale tunnusele. Teisel sammul rakendatakse silujat $\mathcal{S}$ osajääkidele e_i iga $j = 1, ..., p$ korral, et hinnata j-nda

Algoritm 1 *Backfitting* algoritm (Hastie ja Tibshirani, 1987, lk 373)

Samm 1. Olgu $\hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i$, $\hat{f}_j^{(0)}(x_j) = 0\,\forall x_j, \forall j$, $j = 1, ..., p$, $r = 0$.

Samm 2. Iga $j = 1, ..., p$ korral leiame osajäägid

$$e_{ij} = y_i - \hat{\beta}_0 - \sum_{\substack{k=1 \\ k\neq j}}^{p} \hat{f}_k^{(r)}(x_{ik}),\ i = 1, ..., N$$

ning seejärel leiame hinnangu

$$\hat{f}_j^{(r+1)}(x_j) = \mathcal{S}\Big[e_{1j}, e_{2j}, ..., e_{Nj}\Big].$$

Samm 3. Kui $||f_j^{(r+1)} - f_j^{(r)}|| < \delta$, siis lõpetame, vastasel juhul $r = r + 1$ ning liigume sammu 2.

tunnuse mõju uuritavale tunnusele. Kuna teiste tunnuste mõju uuritavale tunnusele pole esialgu täpselt teada, toimub hindamine iteratiivselt. Tsükkel lõppeb siis, kui erinevus funktsioonide vahel on väiksem kui etteantud lävend δ.

2.2 Üldistatud aditiivse mudeli hindamine

See peatükk toetub R. Tibshirani ja T. Hastie teosele „Generalized Additive Models". Eelnevas peatükis tutvustatud *backfitting* algoritmist piisab juhul, kui üldistatud aditiivse mudeli korral kasutatakse seosefunktsioonina identsusseost. Loobume sellest eeldusest ning olgu nüüd mudel taas üldkujul, mis on antud valemiga 6, tähistame

$$g[\mu(\boldsymbol{X})] = \beta_0 + \sum_{j=1}^{p} f_j(X_j).$$

Üldistatud aditiivse mudeli hindamiseks kasutatakse lokaalskooringu protseduuri. Algoritmis 2 kirjeldatakse lokaalskooringu algoritmi Poissoni jaotuse

eelduse ning log-seose korral. Algoritmis kasutatakse kaalutud *backfitting* algoritmi, mis tähendab, et kasutatav silumisfunktsioon peab olema selline, mis võtab arvesse vaatluste kaale.

Algoritm 2 Lokaalskooringu algoritm Poissoni jaotuse ja log-seose korral (Hastie ja Tibshirani, 1990, lk 139-141)

Samm 1. Olgu $\hat{\beta}_0 = g\Big[\frac{1}{N}\sum_{i=1}^{N} y_i\Big]$, $\hat{f}_j^{(0)}(x_j) = 0\,\forall x_j, \forall j,\ j = 1, ..., p,\ r = 0$.

Samm 2. Defineerime $\hat{\eta}^{(r)}(x_1, ..., x_p) = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{f}_j^{(r)}(x_j)$ ja

$\hat{\mu}^{(r)}(x_1, ..., x_p) = \exp\Big(\hat{\eta}^{(r)}(x_1, ..., x_p)\Big)$.

Samm 2.1. Konstrueerime kohandatud muutuja ning leiame selle väärtused z_i realiseerunud valimi väärtuste $x_{i1}, ..., x_{ip}$ põhjal, $i = 1, ..., N$,

$$z_i = \hat{\eta}^{(r)}(x_{i1}, ..., x_{ip}) + \frac{y_i - \hat{\mu}^{(r)}(x_{i1}, ..., x_{ip})}{\hat{\mu}^{(r)}(x_{i1}, ..., x_{ip})},\ i = 1, ..., N.$$

Samm 2.2. Moodustame kaalud w_i, $i = 1, ..., N$,

$$w_i = \hat{\mu}^{(r)}(x_{i1}, ..., x_{ip}),\ i = 1, ..., N. \tag{9}$$

Samm 2.3. Hindame kaalutud *backfitting* meetodiga aditiivse mudeli kohandatud tunnuse z_i jaoks kasutades kaale w_i, $i = 1, ..., N$, tulemuseks on hinnangud $\hat{f}_j^{(r+1)}\ \forall j,\ j = 1, ..., p$.

Samm 2.4. Arvutame koondumiskriteeriumi

$$\Delta\Big(\eta^{(r+1)},\eta^{(r)}\Big) = \frac{\sum_{j=1}^{p} \parallel f_j^{(r+1)} - f_j^{(r)} \parallel}{\sum_{j=1}^{p} \parallel f_j^{(0)} \parallel}.$$

Samm 2.5. Kui $\Delta\Big(\eta^{(r+1)},\eta^{(r)}\Big) < \delta$, siis lõpetame, vastasel juhul $r = r + 1$ ning liigume sammu 2.

3 Mitmemõõtmeline adaptiivne regressioonisplain

See peatükk põhineb T. Hastie, R. Tibshirani ja J. Friedmani teosel „The Elements of Statistical Learning: Data Mining, Inference and Prediction". Mitmemõõtmeline adaptiivne regressioonisplain ehk MARS (*Multivariate Adaptive Regression Splines*) on regressioonanalüüsi vorm. Idee poolest sarnaneb see üldistatud aditiivsele mudelile, kuna argumenttunnustele rakendatakse teatud tüüpi funktsioone. Nii üldistatud aditiivset mudelit kui ka MARSi mudelit kasutatakse mittelineaarsuste modelleerimiseks. Mitmemõõtmelist adaptiivset regressioonisplaini saab rakendada nii mitteparameetrilisel kui ka parameetrilisel viisil.

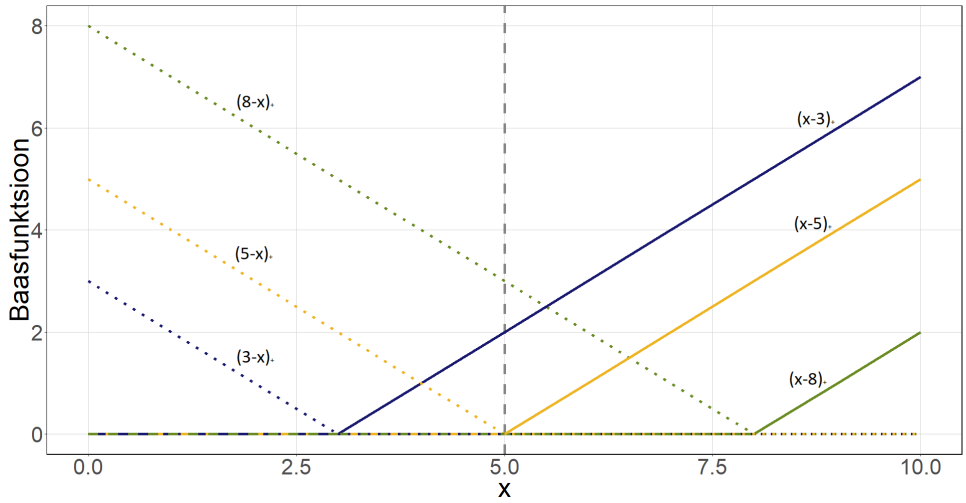
Mitmemõõtmeline adaptiivne regressioonisplain koosneb tükiti lineaarsetest funktsioonidest $h_1(x) = (x-t)_+$ ning $h_2(x) = (t-x)_+$, mis on defineeritud kujul

$$(x-t)_+ = \max\{0, x-t\} = \begin{cases} x-t, & \text{kui } x > t \\ 0, & \text{vastasel juhul,} \end{cases} \tag{10}$$

$$(t-x)_+ = \max\{0, t-x\} = \begin{cases} t-x, & \text{kui } x < t \\ 0, & \text{vastasel juhul.} \end{cases} \tag{11}$$

Funktsioone $(x-t)_+$ ja $(t-x)_+$ nimetatakse baasfunktsioonideks ning punkti t nimetatakse baasfunktsiooni sõlmeks. Samast sõlmest moodustunud kahte baasfunktsiooni nimetatakse peegeldatud funktsioonipaariks. Joonisel 2 on

kujutatud peegeldatud funktsioonipaare erinevate sõlmede väärtustega, $t \in \{3, 5, 8\}$.



Joonis 2. MARSi baasfunktsioonid

MARSi idee on moodustada peegeldatud funktsioonipaarid igast argumenttunnusest X_j sõlmedega x_{ij}, $j = 1, ..., p$, $i = 1, ..., N$, kus x_{ij} on tunnuse X_j realiseerunud väärtused. Sellised funktsioonid moodustavad hulga

$$\mathcal{C} = \{(X_j - t)_+, (t - X_j)_+\}, t \in \{x_{1j}, \ldots, x_{Nj}\}, j = 1, \ldots, p.$$

Kokku on p argumenttunnust, unikaalseid sõlmi iga argumenttunnuse korral on maksimaalselt N ning iga sõlme korral tekib kaks baasfunktsiooni, seega hulgas $\mathcal{C}$ on maksimaalselt $2Np$ baasfunktsiooni. Kui mudel on rangelt aditiivne, siis funktsioon $h_m(\boldsymbol{X})$ tähistab funktsiooni hulgast $\mathcal{C}$. Kui mudelis on lubatud ka koosmõjud, siis funktsioon $h_m(\boldsymbol{X})$ tähistab hulgast $\mathcal{C}$ ühte funktsiooni või mitme funktsiooni korrutist. Olgu M mudelisse kaasatud funktsioonide $h_m(\boldsymbol{X})$ arv ning $h_0(\boldsymbol{X}) = 1$, seega kokku on mudelis $M+1$ hinnatavat parameetrit. MARSi mitteparameetrilise vormi korral on mudel

kujul

$$f(\boldsymbol{X}) = \sum_{m=0}^{M} \beta_m h_m(\boldsymbol{X}) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{X}) \tag{12}$$

ning parameetrilise vormi korral on mudel kujul

$$g[\mu(\boldsymbol{X})] = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(\boldsymbol{X}). \tag{13}$$

3.1 MARSi mudeli hindamine

Käesolev peatükk põhineb T. Hastie, R. Tibshirani ning J. Friedmani teosel „The Elements of Statistical Learning: Data Mining, Inference, and Prediction".

MARSi mudeli hindamine on idee poolest sarnane kasvava valikuga sammregressioonile, kuid algkujul argumenttunnuste asemel kaastakse mudelisse peegeldatud funktsioonipaare hulgast $\mathcal{C}$. Kordajate hinnangud $\hat{\beta}_m$ saadakse mitteparameetrilisel juhul vähimruutude meetodil ning parameetrilisel juhul suurima tõepära meetodil.

Olgu mudelisse kuuluvate funktsioonide $h_m(\boldsymbol{X})$ hulk $\mathcal{M}$, kuhu esialgu kuulub vaid funktsioon $h_0(\boldsymbol{X}) = 1$, seega $\mathcal{M} = \{h_0(\boldsymbol{X})\}$. Igal sammul lisatakse mudelisse üks peegeldatud funktsioonipaar ning seetõttu kasvab hulk $\mathcal{M}$ igal sammul kahe võrra. Kumbki funktsioon lisatakse mudelisse korrutisena, kus üheks teguriks on lisatav baasfunktsioon ning teiseks teguriks on funktsioon hulgast $\mathcal{M}$, kusjuures kummagi korrutise puhul on teine tegur samasugune. Rangelt aditiivsel juhul on alati teiseks teguriks $h_0(\boldsymbol{X}) = 1$.

Algoritm 3 MARS mudeli hindamine mitteparameetrilisel juhul (Hastie jt., 2017, lk 321-326)

Samm 1. Olgu $\hat{f}(\boldsymbol{X}) = \hat{\beta}_0, \quad \hat{\beta}_0 = \frac{1}{N}\sum_{i=1}^{N} y_i,\ \mathcal{M} = \{h_0(\boldsymbol{X})\},\ M = 0,\ r = 1.$

Samm 2. Olgu $RSS_r = \infty$

Samm 2.1. Iga $i = 1, ..., N$, $j = 1, ..., p$ ja $\ell = 1, ..., M$ korral vaatame mudelit kujul

$$f(\boldsymbol{X};\boldsymbol{\beta}) = \sum_{m=0}^{M} \beta_m h_m(\boldsymbol{X}) + \quad \beta_{M+1} h_\ell(\boldsymbol{X}) \cdot (X_j - x_{ij})_+ + \\ + \quad \beta_{M+2} h_\ell(\boldsymbol{X}) \cdot (x_{ij} - X_j)_+$$

ning lahendame järgmise minimeerimiseülesande:

$$\min_{\boldsymbol{\beta}} \sum_{k=1}^{N} \Big(y_k - f(x_{k1}, x_{k2}, ..., x_{kp}; \boldsymbol{\beta})\Big)^2.$$

Saadud lahendi $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_{M+2})$ korral leiame

$$RSS_{ij\ell} = \sum_{k=1}^{N} \Big(y_k - f(x_{k1}, x_{k2}, ..., x_{kp}; \hat{\boldsymbol{\beta}})\Big)^2.$$

Kui $RSS_{ij\ell} < RSS_r$:

$h_1^*(\boldsymbol{X}) = h_\ell(\boldsymbol{X}) \cdot (X_j - x_{ij})_+$

$h_2^*(\boldsymbol{X}) = h_\ell(\boldsymbol{X}) \cdot (x_{ij} - X_j)_+$

$RSS_r = RSS_{ij\ell}$

Samm 2.2. Lisame $h_1^*(\boldsymbol{X})$ ja $h_2^*(\boldsymbol{X})$ hulka $\mathcal{M}$, $M = M + 2$.

Samm 2.3. Kui $M \geq M_{max}$, siis lõpetame, vastasel korral $r = r + 1$ ja liigume sammu 2.

Koosmõjude lubamisel mudelisse kehtivad mõned piirangud. Teiseks teguriks ei sobi funktsioonid, mis sisaldavad juba sama argumenttunnust, millel antud peegeldatud funktsioonipaar põhineb. Kui mudelisse on lubatud kuni q-ndat

järku koosmõjud, siis teiseks teguriks saab olla vaid funktsioon, mis omakorda koosneb maksimaalselt $(q-1)$-st erinevast tegurist (mis ei ole võrdsed ühega).

Iga peegeldunud funktsioonipaari ja võimaliku teguri hulgast $\mathcal{M}$ korral leitakse jääkide ruutude summa ning kõige väiksema jääkide ruutude summa taganud korrutised lisatakse hulka $\mathcal{M}$. Parameetrilisel juhul otsitakse maksimaalse tõepärafunktsiooni väärtuse taganud funktsioonide korrutist. Tsükkel kestab seni, kuni hulgas $\mathcal{M}$ on etteantud arv M_{max} funktsioone.

Selguse mõttes esitame näite algoritmi põhimõttest, kus mudelisse lubatakse kaasata esimest järku koosmõjusid. Lõpetame tsükli, kui hulgas $\mathcal{M}$ on viis elementi ehk $M = 4$. Kuna esialgu on selles hulgas vaid konstantne funktsioon, siis tsükli esimesel sammul saab olla $h_\ell(\boldsymbol{X})$ väärtuseks vaid $h_\ell(\boldsymbol{X}) = 1$ ning seetõttu hindame mudelit kujul

$$f(\boldsymbol{X}) = \beta_0 + \beta_1(X_j - t) + \beta_2(t - X_j).$$

Oletame, et minimaalse jääkide ruutude summa tagavad korrutised $h_1^*(\boldsymbol{X}) = (X_4 - 1)_+$ ning $h_2^*(\boldsymbol{X}) = (1 - X_4)_+$, seega hulga $\mathcal{M}$ elementideks on nüüd $\mathcal{M} = \{h_0(\boldsymbol{X}), h_1(\boldsymbol{X}), h_2(\boldsymbol{X})\} = \{1, (X_4-1)_+, (1-X_4)_+\}$ ning $M = 2$. Tsükli teisel sammul hindame mudelit kujul

$$f(\boldsymbol{X}) = \beta_0+\beta_1(X_4-1)_+ + \beta_2(1-X_4)_+ + \beta_3 h_\ell(\boldsymbol{X})\cdot(X_j-t)_+ + \beta_4 h_\ell(\boldsymbol{X})\cdot(t-X_j)_+,$$

kus $h_\ell(\boldsymbol{X})$ on üks kolmest funktsioonist:

$$\begin{aligned} h_\ell(\boldsymbol{X}) &= 1, \\ h_\ell(\boldsymbol{X}) &= (X_4 - 1)_+, \\ h_\ell(\boldsymbol{X}) &= (1 - X_4)_+ \end{aligned}$$

Oletame nüüd, et parima mudeli saame juhul, kui $h_\ell(\boldsymbol{X}) = (X_4 - 1)_+$ ning baasfunktsioonideks on $(X_9 + 5)_+$ ning $(-5 - X_9)_+$. Lisame hulka $\mathcal{M}$ funktsioonid $h_1^*(\boldsymbol{X}) = (X_4 - 1)_+ \cdot (X_9 + 5)_+$ ja $h_2^*(\boldsymbol{X}) = (X_4 - 1)_+ \cdot (-5 - X_9)_+$.

Nüüd on $M = 4$ ehk hulgas $\mathcal{M}$ on viis elementi ning lõpetame tsükli. Seega hinnatud mudel avaldub kujul

$$\begin{aligned}\hat{f}(\boldsymbol{X}) = \quad & \hat{\beta}_0 + \hat{\beta}_1(X_4 - 1)_+ + \hat{\beta}_2(1 - X_4)_+ + \\ + \quad & \hat{\beta}_3(X_4 - 1)_+ \cdot (X_9 + 5)_+ + \hat{\beta}_4 h(X_4 - 1)_+ \cdot (-5 - X_9)_+.\end{aligned}$$

Sellise tsükli tulemuseks on sageli ületreenitud mudel, mistõttu tuleb rakendada tagasivaatelist funktsioonide eemaldamist. Kui mudeli hindamise esimeses etapis lisatakse funktsioone mudelisse paarikaupa, siis teises etapis hinnatakse iga funktsiooni olulisust mudelis eraldi. Funktsioon, mille eemaldamisel jääkide ruutude summa kasvab kõige väiksemal määral, jäetakse mudelist kõrvale. Selliselt leitakse iga $\ell = 0, 1, ..., M$ kohta parim mudel, kus ℓ on mudelisse kaasatud funktsioonide $h_m(\boldsymbol{X})$ arv.

Viimases etapis tuleb hinnata, mitmest funktsioonist $h_m(\boldsymbol{X})$ koosnev mudel annab parima tulemuse. Selleks võib kasutada harilikku ristvalideerimist kui ka üldistatud ristvalideerimist.

Üldistatud ristvalideerimisel leitakse iga ℓ korral suurus

$$\mathrm{GCV}(\ell) = \frac{\sum_{i=1}^{N}(y_i - \hat{f}_\ell\big(x_{i1}, x_{i2}, ..., x_{ip})\big)^2}{\Big(1 - \frac{M(\ell)}{N}\Big)^2}, \tag{14}$$

kus $M(\ell) = r + cK$, r on hinnatavate parameetrite arv mudelis, K on mudelisse kaasatud sõlmede arv ning c on karistusparameeter. Kui mudel on rangelt aditiivne ehk mudelis ei ole baasfunktsioonide korrutisi, siis karistusparameetri väärtuseks on $c = 2$, vastasel juhul on $c = 3$. Tulemuseks on mudel $f_\ell(\boldsymbol{X})$, mille korral $\mathrm{GCV}(\ell)$ on kõige väiksem.

4 Liikluskindlustuse kahjusageduse hindamine

4.1 Andmestik

Andmestik koosneb Eesti liikluskindlustuse ületuruandmete valimist. Valimisse kuuluvad liikluskindlustuste katted, mille kehtivuse algus jääb vahemikku 01.01.2015-31.12.2017. Lisaks kehtivad katete kohta järgnevad tingimused:

- sõiduki kategooria on M1 (sõiduauto),
- sõiduk on arvel autoregistris,
- sõiduki kindlustuskohustuslik isik on eraisik,
- katteklass on tavapärane,
- katte tüüp on poliis,
- kindlustuse kehtinud päevade arv on positiivne.

Magistritöö jaoks on moodustatud sellistele tingimustele vastavatest koguturuandmetest lihtsa tagasipanekuta juhusliku valiku meetodil valim, mis moodustab 5% üldkogumist. Andmestiku suurus on 138 642 katet. Esmases andmestikus on järgnevad tunnused: toimunud kahjude arv, katte alguse aasta, teenitud kindlustuspäevade arv, kindlustusvõtja isikukoodi kolm esimest numbrit, sõiduki võimsus, registrimass, keretüüp, kütus ning sõiduki esmase registreerimise aasta.

4.1.1 Andmestiku korrastamine

Andmestikus on esindatud 13 erinevat keretüüpi, mis on välja toodud tabelis 1. Selle tabeli põhjal võib väita, et levinumad keretüübid on luukpära, sedaan, mahtuniversaal ning universaal. Teiste keretüüpidega katteid on oluliselt vähem ning mitme grupi korral pole tekkinud ühtegi õnnetusjuhtumit. Mitmed andmestikus esinevad keretüübid pole sõiduautode puhul tavapärased, näiteks limusiin ja võistlusauto, kuigi katte klass määrab sõiduki kategooriaks sõiduauto. Sellel põhjusel kaasame mudeli treenimise andmestikku vaid need katted, mille korral on keretüüp kuulub levinumate keretüüpide hulka.

Tabel 1. Katete sagedus ja kahjude arv keretüübi põhjal

Keretüüp	**Sagedus**	**Kahjude arv**
Buss	1	0
Elamu	152	2
Kaubik	3	0
Kupee	1 905	34
Lahtine	686	5
Limusiin	16	0
Luukpära	32 451	666
Mahtuniversaal	15 921	354
Pikap	4	0
Sedaan	41 625	859
Sihtotstarbeline	18	1
Universaal	45 828	905
Võistlusauto	32	0

Andmestikus leidub seitse erinevat kütusetüüpi. Levinumate keretüüpidega katete seas leidub viis katet, mille korral on kütusetüübiks märgitud diisel-

hübriid ning ühegi sellise katte puhul kahju ei tekkinud. Seetõttu on diiselhübriidi kütusetüübiga katted treeningandmestikust välja jäetud. Levinumate keretüüpidega katete seas leidub kate, mille korral on märgitud mootori võimsuseks seitse kilovatti. Kuna see vaatlus erineb ülejäänud katetest ning nii madal võimsus pole sõiduautodele tavapärane, jätame ka selle katte andmestikust välja.

Levinumad katte kehtivuse pikkused on üks, kolm, kuus kuud ning üks aasta. Andmestikus leidub katteid, mille pikkus on lühem kui üks päev. Sellised katted on andmestikust kõrvale jäetud.

Kindlustusvõtja isikukoodi kolme esimese numbri põhjal on võimalik leida isiku sugu ja vanus katte alguse aastal. Need katted, mille korral numbrite asemel esineb muid sümboleid, jätame valimist välja. Samuti leidub katteid, kus arvutatud vanus on äärmuslik ning seega kuuluvad andmestikku vaid katted, mille korral kindlustusvõtja vanus katte alguse aastal oli vahemikus $18 - 85$.

Sõiduki esmase registreerimise aasta põhjal arvutame sõiduki vanuse katte alguse aastal. Kõige vanem kindlustatud auto oli kindlustuskatte alguse aastal 77 aastat vana. Väga vanad autod võivad käituda kindlustuskahjude suhtes erinevalt võrreldes uuemate autodega ning seetõttu on andmestikku jäetud vaid katted, mille kehtivuse alguse aastal oli sõiduki vanuseks kuni 25 aastat.

Lõplikku andmestikku kuulub 130 990 katet. Saadud andmestikust eraldame 30% testandmestikuks ning ülejäänud katteid kasutame mudeli treenimiseks. Seega treeningandmestiku maht on 91 650 katet ning testandmestikku kuulub 39 340 katet. Ülevaate puhastatud andmetest leiab tabelist 2.

Tabel 2. Tunnuste kirjeldus

Tunnus	**Kirjeldus**	**Miinimum**	**Maksimum**	**Keskmine**	**Standardhälve**
loss_count	kindlustuskahjude arv	0	4	0,021	0,149
days	teenitud kindlustuspäevade arv	1	366,7	138,5	119,4
gender	kindlustusvõtja sugu: - mees (66,7%), 1 - naine (33,3%)				
person_age	kindlustusvõtja vanus aastates katte alguse aastal	18	85	44,0	14,2
vehicle_engine_power	sõiduki mootori võimsus (kW)	28	430	96,3	32,6
vehicle_reg_mass	sõiduki registrimass (kg)	1 040	4 218	1 946	300
vehicle_frame_type	sõiduki keretüüp: luukpära (24,0%), mahtuniversaal (12,0%), sedaan (29,6%), universaal (34,4%)				
cl_vehicle_fuel	sõiduki kütuse tüüp: bensiin (0,8%), diisel (39,3%), bensiin-hübriid (0,4%), bensiin-kat (59,3%), elekter (0,1%), surugaas (0,1%)				
vehicle_age	sõiduki vanus katte alguse aastal	0	25	12,0	6,1

4.2 R-i pakett `earth`

Selles peatükis tutvustatakse R-i dokumentatsiooni põhjal statistikapaketti „*Multivariate Adaptive Regression Splines*" (Milborrow, 2019). Pakett `earth` põhineb funktsioonil `mars`, mis asub paketis `mda` ning mille autorid on Trevor Hastie ning Robert Tibshirani. Selle töö raames on eelistatud paketti `earth`, kuna selles paketis on implementeeritud oluliselt rohkem mudeliga seonduvaid funktsioone ning hinnatud mudeli tulemusi on võimalik väljastada kasutajasõbralikumal kujul kui paketis `mda`.

Järgnevalt tutvustame lähemalt funktsiooni `earth`, mille abil on võimalik hinnata mitmemõõtmelisi adaptiivseid regressioonisplaine. Selle funktsiooni olulisemad argumendid on:

- `formula` - määrab mudeli sõltuva tunnuse ning argumenttunnused;
- `degree` - maksimaalne baasfunktsioonide arv, millest $h_m(X)$ koosneda võib (vaikimisi `degree`=1);
- `nk` - maksimaalne funktsioonide $h_m(X)$ arv mudelis;
- `thresh` - lävend, mille saavutamisel edasivaateline tsükkel lõppeb (vaikimisi `thresh`=0.001)
- `penalty` - karistusparameeter c (vaikimisi `penalty`=2, kui `degree`=1, muidu `penalty`=3);
- `prune` - väärtuseks `TRUE`, kui soovitakse teostada tagasivaatelist funktsioonide eemaldamist, `FALSE` vastasel juhul (vaikimisi väärtuseks `TRUE`)
- `glm` - võimalik hinnata üldistatud lineaarne mudel, milleks tuleb määrata jaotus- ning seosefunktsioon.

Loendusandmete puhul on võimalik kasutada Poissoni jaotust. Kuna kahjude

hulka on mõistlik hinnata ajaühiku kohta, siis `formula` osas on võimalik anda ette lepingu kestus märksõnaga `offset()`.

4.3 Tulemused

Mudelite hindamisel peame silmas, et lõpptulemus oleks loogiliselt selgitatav ning võimalikult lihtne interpreteerida. Nendel põhjustel on mudelitesse kaasatakse ainult esimest järku koosmõjusid. Üldistatud lineaarsete mudelite korral kasutatakse olulisusnivoona $\alpha = 0{,}05$ ning mudelid luuakse tagasivaatelise sammregressiooni abil. Mudelitesse jäetakse alles vaid tunnused, mis on antud olulisusnivool statistiliselt olulised.

Kõikides mudelites kasutame seosefunktsioonina logaritmfunktsiooni ning anname mudelites ette `offset`$= \ln(days)$, kus tunnus *days* määrab, millise ajaühiku jooksul kahjude arvu vaadeldi. Seega saame tulemuseks mudelid, mis prognoosivad kahjude arvu päevas. Pidevad tunnused kaasatakse mudelisse logaritmteisendustena, kuna siis on argumenttunnused samal skaalal kui uuritav tunnus. Teose „Generalized Linear Models for Insurance Rating" autorid väidavad, et log-seose korral peaks pideva tunnuse logaritmitud vormi käsitlema kui algset vormi ning logaritmimata väärtust kui teisendatud kuju, mida on mõistlik kasutada vaid erijuhtudel (Goldburd jt., 2016, lk 11-12).

Mudeli headust hindame Akaike informatsioonikriteeriumi alusel, mis avaldub valemina

$$AIC = -2 \ln L + 2k,$$

kus $\ln L$ on treenitud mudeli logaritmiline tõepära ning k on parameetrite arv mudelis. Kuna katsetatavad mudelid ei ole kõik omavahel hierarhilised, kasutame paariviisiliseks võrdlemiseks Vuongi testi. Test põhineb mudelite prognoosidel ning sisukas hüpotees väidab, et üks mudelitest on teisest parem

(Jackman, 2017). Lisaks leiame testandmestikul iga mudeli korral jääkide ruutude summad, mille alusel hindame mudeli prognoosivõimet.

Esmalt hindame üldistatud lineaarse mudeli, kus uuritava tunnuse tinglikuks jaotuseks eeldame Poissoni jaotust. Sellise eeldusega mudeli sobitamisel ei olnud ükski esimest järku koosmõju oluline ning seeõttu koosneb mudel vaid peamõjudest. Parimas Poissoni mudelis mõjutab kindlustuskahjude sagedust katte kehtivuse pikkus, sõiduki registrimass, vanus ja kütusetüüp, kindlustusvõtja sugu ja vanus. Programmi R väljundi leiab lisast 1 ning mudel avaldub kujul

$$\begin{aligned}\frac{loss\_count}{days} = \; & \exp(-10{,}59 - 0{,}0025 \cdot days + 0{,}51 \cdot \ln(vehicle\_reg\_mass) + \\ + \; & 0{,}61 \cdot I[cl\_vehicle\_fuel \in \{bensiin - hübriid, surugaas\}] + \\ + \; & 1{,}35 \cdot I[cl\_vehicle\_fuel = elekter] + \\ - \; & 0{,}013 \cdot vehicle\_age + 0{,}14 \cdot I[gender = 1] + \\ - \; & 0{,}058 \cdot person\_age + 0{,}00054 \cdot person\_age^2).\end{aligned} \tag{15}$$

Kuna Poissoni jaotuse korral on teoreetiline keskväärtus ja dispersioon omavahel võrdsed, siis tuleb selle mudeli korral kontrollida ka hajuvust. Korrektse mudeli korral on nii hälbimuse vigade kui ka Pearsoni vigade põhjal leitud skaalaparameetri hinnangud ligikaudu võrdsed ühega. Antud mudeli korral on hälbimuse vigade põhjal leitud skaalaparameetri hinnang $\varphi_D = 0{,}16$, Pearsoni jääkide põhjal leitud hinnang on $\varphi_P = 1{,}26$. Hälbimuse skaalaparameetri hinnang viitab tugevale alahajuvusele, kuid Pearsoni skaalaparameetri hinnang viitab sellele, et mudelis hajuvusega probleeme ei ole. Enamasti alahajuvust ei peeta mudeli juures suureks probleemiks ning seetõttu leiame, et Poissoni jaotus sobib neid andmeid kirjeldama.

Teisena hindame samuti üldistatud lineaarse mudeli, kuid uuritava tunnuse tinglikuks jaotuseks eeldame negatiivset binoomjaotust. Ka negatiivse bi-

noomjaotuse mudeli korral ei ole võimalik lõplikusse mudelisse kaasata statistiliselt olulisi koosmõjusid. Negatiivse binoomjaotuse parimas mudelis osutuvad statistiliselt oluliseks samad tunnused, mis on ka Poissoni mudelis, ning regressioonikordajate hinnangud on väga sarnased. Programmi R väljundi leiab lisast 2 ning mudel avaldub kujul

$$\begin{aligned}\frac{loss\_count}{days} = & \exp(-10{,}64 - 0{,}0025 \cdot days + 0{,}51 \cdot \ln(vehicle\_reg\_mass) + \\ + & \ 0{,}60 \cdot I[cl\_vehicle\_fuel \in \{bensiin - hübriid, surugaas\}] + \\ + & \ 1{,}35 \cdot I[cl\_vehicle\_fuel = elekter] + \\ - & \ 0{,}013 \cdot vehicle\_age + 0{,}14 \cdot I[gender = 1] + \\ - & \ 0{,}058 \cdot person\_age + 0{,}00054 \cdot person\_age^2).\end{aligned} \tag{16}$$

MARS mudeli hindamisel kasutame Poissoni jaotuse eeldust, kuna funktsioonis `earth` on võimalik kasutada negatiivse binoomjaotust ainult juhul, kui parameetri k väärtus on fikseeritud. MARS mudeli korral mõjutab kindlustuskahjude sagedust katte kehtivuse pikkus ning sõiduki vanus. Programmi R väljundi leiab lisast 3 ning mudel avaldub kujul

$$\begin{aligned}\frac{loss\_count}{days} = & \exp(-8{,}28 - 0{,}025 \cdot (56{,}2 - days)_+ + \\ - & \ 0{,}0044 \cdot (days - 56{,}2)_+ + 0{,}0034 \cdot (days - 210)_+ + \\ - & \ 0{,}0040 \cdot (56{,}2 - days)_+ \cdot (vehicle\_age - 17)_+ + \\ - & \ 0{,}0028 \cdot (56{,}2 - days)_+ \cdot (17 - vehicle\_age)_+).\end{aligned} \tag{17}$$

Kuna MARS mudelisse jääb alles vaid kaks seletavat tunnust, siis teeme teeme tagasivaatelise sammregressiooni abil mudeli, mis kasutab MARSi mudelisse kaasatud tunnuseid tükiti ning ülejäänud tunnuseid lineaarsena. Sellised mudelid teeme nii Poissoni jaotuse kui ka negatiivse binoomjaotuse eeldusel. Mõlemal juhul jäävad mudelisse samad tunnused, mis varasemalt üldistatud lineaarsete mudelite puhul. Programmi R väljundi leiab vastavalt lisadest 4

ja 5, Poissoni mudel avaldub kujul

$$\begin{aligned}\frac{loss\_count}{days} = & \exp(-11{,}51 + 0{,}023 \cdot (56{,}2 - days)_+ - 0{,}0039 \cdot (days - 56{,}2)_+ + \\ + & \ 0{,}0029 \cdot (days - 210)_+ + 0{,}57 \cdot \ln(vehicle\_reg\_mass) + \\ + & \ 0{,}71 \cdot I[cl\_vehicle\_fuel \in \{bensiin - h\ddot{u}briid, surugaas\}] + \\ + & \ 1{,}48 \cdot I[cl\_vehicle\_fuel = elekter] + \\ - & \ 0{,}0036 \cdot (56{,}2 - days)_+ \cdot (vehicle\_age - 17)_+ + \\ - & \ 0{,}0026 \cdot (56{,}2 - days)_+ \cdot (17 - vehicle\_age)_+) + \\ + & \ 0{,}16 \cdot I[gender = 1] - 0{,}047 \cdot person\_age + \\ + & \ 0{,}00043 \cdot person\_age^2) \end{aligned} \tag{18}$$

ning negatiivse binoomjaotuse mudel avaldub kujul

$$\begin{aligned}\frac{loss\_count}{days} = & \exp(-11{,}56 + 0{,}023 \cdot (56{,}2 - days)_+ - 0{,}0039 \cdot (days - 56{,}2)_+ + \\ + & \ 0{,}0030 \cdot (days - 210)_+ + 0{,}57 \cdot \ln(vehicle\_reg\_mass) + \\ + & \ 0{,}71 \cdot I[cl\_vehicle\_fuel \in \{bensiin - h\ddot{u}briid, surugaas\}] + \\ + & \ 1{,}49 \cdot I[cl\_vehicle\_fuel = elekter] + \\ - & \ 0{,}0036 \cdot (56{,}2 - days)_+ \cdot (vehicle\_age - 17)_+ + \\ - & \ 0{,}0026 \cdot (56{,}2 - days)_+ \cdot (17 - vehicle\_age)_+) + \\ + & \ 0{,}16 \cdot I[gender = 1] - 0{,}047 \cdot person\_age + \\ + & \ 0{,}00043 \cdot person\_age^2). \end{aligned} \tag{19}$$

Kõikide mudelite Akaike informatsioonikriteeriumi väärtused, jääkide ruutude summad ning summaarne prognoositud kahjude arv on tabelis 3.

Tabel 3. Akaike kriteerium ja jääkide ruutude summa erinevate mudelite korral

	Poisson	**Negatiivne binoomjaotus**	**MARS**	**MARSiga Poisson**	**MARSiga negatiivne binoomjaotus**
AIC	18 143	18 080	18 180	18 123	18 060
RSS	863,396	863,392	863,714	863,560	863,596

Tabelist 3 on näha, et testandmestikul leitud jääkide ruutude summad on kõikide mudelite korral väga sarnased. Akaike informatsioonikriteeriumis on näha väikeseid erinevusi ning selle põhjal parim mudel on MARSi funktsioone kasutanud üldine lineaarne mudel, kus eeldatakse negatiivset binoomjaotust. Kontrollime Vuongi testi abil hüpoteesi, kas MARSiga negatiivne binoomjaotuse mudel on parem kui tavaline negatiivne binoomjaotuse mudel. Vuongi testi p-väärtuseks on 0,044, mis on veidi väiksem, kui lubatud piir. Seetõttu võime võtta vastu sisuka hüpoteesi, mis ütleb, et MARSi tunnustega mudel on parem kui tavaline üldistatud lineaarne mudel negatiivse binoomjaotuse eeldusel.

Tabelis 4 on antud testandmestiku tegelik summaarne kahjude arv ning summaarsed kahjude arvu prognoosid katte kehtivuse gruppide põhjal. Üle kõigi gruppide annavad mudelid sarnaseid tulemusi, erinevus tegelikust on kuni kolm kahjut. Tabelist on näha, et Poissoni ja negatiivse binoomjaotuse mudeli prognoos kõige lühema kestusega katetele on täpne, kuid keskmise kehtivusega katete korral prognoositakse kahjude arvu märgatavalt suuremaks ning pikema kehtivusega katte korral prognoositakse kahjude arvu oluliselt väiksemaks. MARSi mudeli muutujaid kasutavad mudelid prognoosivad

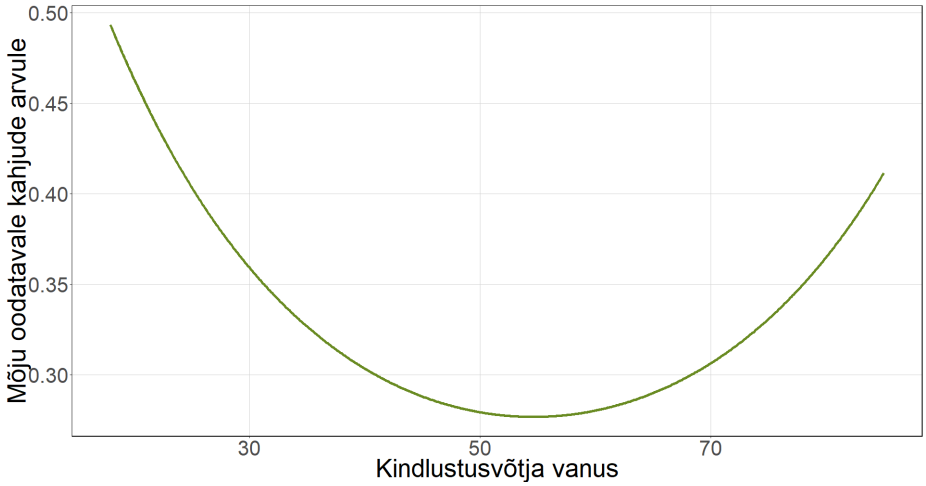
kõige lühema kehtivusega katete korral kahjude arvu mõnevõrra kõrgemaks, kuid teiste gruppide korral on prognoosid tegelikkusele lähemal kui tavalise Poissoni ja negatiivse binoomjaotuse mudelite korral.

Tabel 4. Summaarne kahjude arv testandmestikul katte kehtivuse lõikes

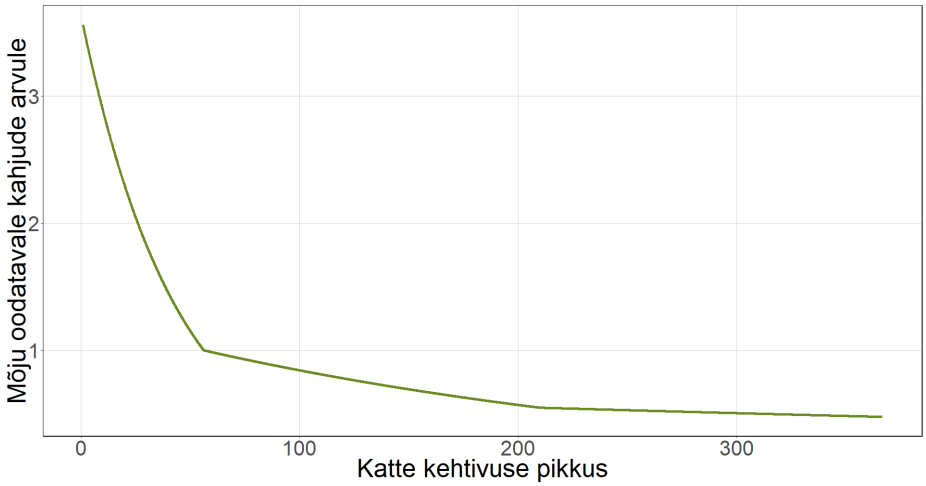
Katte kehtivus päevades	**Tegelik**	**Poisson**	**Negatiivne binoomjaotus**	**MARS**	**MARSiga Poisson**	**MARSiga negatiivne binoomjaotus**
[1, 56]	93	93,2	93,4	102,57	102,50	102,65
(56, 210]	388	422,39	422,93	405,93	406,54	406,34
(210, 367]	342	304,46	303,93	312,01	311,54	311,50
Kokku	823	820,05	820,26	820,51	820,58	820,49

Järgnevalt kirjeldame, kuidas erinevad argumenttunnused mõjutavad kahjude tekkimise sagedust MARSi tunnustega negatiivse binoomjaotuse mudeli korral, kui ülejäänud tunnuste väärtused on fikseeritud. Naiste kindlustuskahjude arv on $\exp(0{,}16) \approx 1{,}17$ korda ehk 17% võrra suurem kui meestel. Võrreldes bensiin-, diisel- ja bensiin-katalüsaatormootoriga autodega on elektriautode kahjude arv $\exp(1{,}49) \approx 4{,}4$ korda suurem ning surugaasil töötava või bensiin-hübriidmootoriga autode kahjude arv on $\exp(0{,}71) \approx 2$ korda suurem. Sõiduki massi 10% tõusmisel kasvab prognoositav kindlustuskahjude arv $1{,}1^{0{,}57} \approx 1{,}06$ korda ehk kahjude arvu prognoos on 6% võrra kõrgem. Kindlustusvõtja vanus on mudelis sees ruutpolünoomina, kõige madalam prognoositav kindlustuskahjude arv on inimestel vanuses 55 aastat. Joonisel 3 on kujutatud kindlustusvõtja vanuse mõju kindlustuskahjude arvule. Jooniselt 4 on näha, et katte kehtivuse pikenedes oodatav kahjude arv kahaneb. Kuni kehtivuse pikkuseni 56 päeva kahaneb oodatav kahjude arv järsult, edasi kuni 210 päevani kanaheb oodatav kahjude arv aeglasemalt

ning lõpus kahaneb oodatav kahjude arv kõige aeglasemalt. Sõiduki vanus mõjutab oodatavat kahjude arvu vaid juhul, kui katte pikkus on kuni 56 päeva. Kuni sõiduki vanuseni 17 oodatav kahjude arv kasvab vanuse suurenedes ning vanemate autode puhul oodatav kahjude arv vanuse suurenedes kahaneb.



Joonis 3. Kindlustusvõtja vanuse mõju oodatavale kahjude arvule



Joonis 4. Katte kehtivuse pikkuse mõju oodatavale kahjude arvule

Tabelis 4 on antud testandmestiku tegelikud kahjude arvud ning kõigi mudelite summaarsed prognoosid katte kehtivuse pikkuse lõikes. Vaatleme MARSi liikmetega negatiivse binoomjaotuse mudeli korral tegelikke ja summaarseid prognoose ka teiste tunnuste lõikes. Tabelis 5 on näha tegelik ja prognoositud summaarne kahjude arv kütuseliikide lõikes. Antud testandmestikul tundub, et baasist erinevates gruppides on oodatav kahjude arv umbes kaks korda kõrgem kui tegelik.

Tabel 5. Summaarne kahjude arv testandmestikul kütusetüübi lõikes

Kütusetüüp	**Tegelik**	**MARSiga negatiivne binoomjaotus**
Bensiin, bensiin-katalüsaator, diisel	817	808,9
Surugaas, bensiin-hübriid	1	2,5
Elekter	5	9,1

Tabelis 6 on toodud tegelik ja prognoosid summaarne kahjude arv kindlustusvõtja soo lõikes. Mudel prognoosib naistele vaid tegelikust kaks kahju rohkem ning meestele viis kahju vähem.

Tabel 6. Summaarne kahjude arv testandmestikul kindlustusvõtja soo lõikes

Kindlustusvõtja sugu	**Tegelik**	**MARSiga negatiivne binoomjaotus**
Mees	539	534,3
Naine	284	286,2

Jagame katted kindlustustusvõtja vanuse põhjal kvintiil-gruppidesse. Tabelis 7 on näha, et vanusegruppides (30,39] ja (39,47] on prognoosid tegelikule väga lähedal. Äärmistes vanusegruppides on prognoosid väga suurel määral erinevad tegelikust kahjude arvust. Noorte kindlustusvõtjate puhul

hindame oodatava kahjude arvu umbes 20 võrra suuremaks ning vanemate kindlustusvõtjate puhul hindame oodatava kahjude arvu umbes 30 võrra väiksemaks. Kuna erinevused on väga suured, siis oleks mõistlik võrdluseks sobitada mudel, kus vanuse ruutpolünoomi asemel kasutatakse mudelis vanusegruppe.

Tabel 7. Summaarne kahjude arv testandmestikul kindlustusvõtja vanuse lõikes

Kindlustusvõtja vanus	**Tegelik**	**MARSiga negatiivne binoomjaotus**
[18,30]	147	170,6
(30,39]	180	180,2
(39,47]	150	149,3
(47,57]	143	150,9
(57,85]	203	169,4

Kokkuvõte

Töö eesmärgiks oli tutvustada mitmemõõtmelist adaptiivset regressioonisplaini ning võrrelda sellel meetodil leitud mudeleid üldistatud lineaarsete mudelitega. Selleks andsime töö esimeses pooles ülevaate üldistatud lineaarsest mudelist, üldistatud aditiivsest mudelist ning mitmemõõtmelisest adaptiivsest regressioonisplainist. Töö teine pool sisaldab esimeses osas kirjeldatud meetodite rakendamist, mudelite võrdlemist ning interpreteerimist. Mudelid sobitasime Eesti liikluskindlustuse andmetele.

Esmalt hindasime kahjude sagedusele üldistatud lineaarsed mudelid kasutades log-seost ning Poissoni ja negatiivse binoomjaotuse eeldusi. Mõlemas mudelis mõjutavad kahjude sagedust katte kehtivuse pikkus, sõiduki registrimass, kütusetüüp ja vanus ning kindlustusvõtja sugu ja vanus, kusjuures Kordajate hinnangud on mudelites väga sarnased.

Seejärel hindasime MARSi mudeli kasutades log-seost ja Poissoni jaotuse eeldust. Selle põhjal mõjutavad kahjude sagedust vaid katte kehtivuse pikkus ning sõiduki vanus. Katte kehtivuse korral kasutatakse kahte sõlme ning sõiduki vanuse korral ühte sõlme. Akaike informatsioonikriteeriumi põhjal on üldistatud lineaarsed mudelid paremad kui MARSi mudel. Seetõttu otsustasime hinnata tagasivaatelise sammregressiooni abil üldistatud lineaarsed mudelid nii Poissoni kui ka negatiivse binoomjaotuse eeldusel, kuid katte kehtivuse pikkus ja sõiduki vanus on asendatud MARSi mudeli transformatsioonidega. Kahjude sagegdust mõjutavad nendes mudelites samad tunnused, mis mõjutavad kahjude sagedust tavalistes üldistatud lineaarsetes mudelites.

Selgus, et Akaike informatsioonikriteeriumi põhjal parim mudel on MARSi liikmetega negatiivse binoomjaotuse mudel. Testandmestikul uurisime lähemalt, kui täpselt see mudel prognoosib summaarset kahjude arvu erinevate tunnuste lõikes võrreldes tegeliku kahjude arvuga. Nägime, et sooliselt suudab mudel prognoosida väga sarnaseid tulemusi, kuid kindlustusvõtja vanuse korral on äärmiste väärtuste korral suured erinevused. Nooremate kindlustusvõtjatele prognoosib mudel oluliselt kõrgemat kahjude arvu ning vanemate kindlustusvõtjate korral prognoosib tegelikust madalamat kahjude arvu. See tähendab, et alternatiivina võiks luua mudeli, kus kindlustusvõtja vanus on eelnevalt rühmitatud ning kaasatud mudelisse kategoorilise tunnusena.

Kasutatud kirjandus

Goldburd, M., Khare, A. ja Tevet, A. (2016). *Generalized Linear Models for Insurance Rating*. Casualty Actuarial Society, Virginia.

Hastie, T. ja Tibshirani, R. (1987). Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82.

Hastie, T. ja Tibshirani, R. (1990). *Generalized Linear Models*. Chapman & Hall, London, first edition.

Hastie, T., Tibshirani, R. ja Friedman, J. (2017). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, second edition, 12th printing.

Jackman, S. (2017). *Package 'pscl': Political Science Computational Laboratory*. https://cran.r-project.org/web/packages/pscl/pscl.pdf [Kasutatud: 08.05.2019].

James, G., Witten, D., Hastie, T. ja Tibshirani, R. (2017). *An Introduction to Statistical Learning with Applications in R*. Springer, New York, 8th printing.

McCullagh, P. ja Nelder, J. (1989). *Generalized Linear Models*. Chapman & Hall, London, second edition.

Milborrow, S. (2019). *Package 'earth': Multivariate Adaptive Regression Splines*. https://cran.r-project.org/web/packages/earth/earth.pdf [Kasutatud: 23.04.2019].

Montgomery, D., Peck, E. ja Vining, G. (2012). *Introduction to Linear Regression Analysis*. John Wiley Sons, Inc., Hoboken, 5th edition.

Lisad

Lisa 1. Poissoni mudeli väljund

```
##
## Call:
## glm(formula = LOSS_COUNT ~ DAYS + log(VEHICLE_REG_MASS) + CL_VEHICLE_FUEL_group +
##     VEHICLE_AGE + GENDER + PERSON_AGE + I(PERSON_AGE^2), family = "poisson",
##     data = sample_train, offset = log(DAYS))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6332  -0.2436  -0.1942  -0.1309   4.9861
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.059e+01  1.203e+00  -8.800  < 2e-16
## DAYS                            -2.471e-03  1.960e-04 -12.604  < 2e-16
## log(VEHICLE_REG_MASS)            5.091e-01  1.566e-01   3.252 0.001146
## CL_VEHICLE_FUEL_groupElekter     1.345e+00  3.808e-01   3.531 0.000414
## CL_VEHICLE_FUEL_groupHübriid_CNG 6.078e-01  2.320e-01   2.620 0.008794
## VEHICLE_AGE                     -1.337e-02  4.042e-03  -3.307 0.000941
## GENDER1                          1.409e-01  4.941e-02   2.851 0.004358
## PERSON_AGE                      -5.810e-02  9.405e-03  -6.177 6.52e-10
## I(PERSON_AGE^2)                  5.397e-04  9.709e-05   5.559 2.71e-08
##
## (Intercept)                     ***
## DAYS                            ***
## log(VEHICLE_REG_MASS)           **
## CL_VEHICLE_FUEL_groupElekter    ***
## CL_VEHICLE_FUEL_groupHübriid_CNG **
## VEHICLE_AGE                     ***
## GENDER1                         **
## PERSON_AGE                      ***
```

```
## I(PERSON_AGE^2)                  ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 14695  on 91649  degrees of freedom
## Residual deviance: 14403  on 91641  degrees of freedom
## AIC: 18143
##
## Number of Fisher Scoring iterations: 6
```

Lisa 2. Negatiivse binoomjaotuse mudeli väljund

```
##
## Call:
## glm.nb(formula = LOSS_COUNT ~ DAYS + log(VEHICLE_REG_MASS) +
##     CL_VEHICLE_FUEL_group + VEHICLE_AGE + GENDER + PERSON_AGE +
##     I(PERSON_AGE^2) + offset(log(DAYS)), data = sample_train,
##     init.theta = 0.5136094733, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5815  -0.2402  -0.1927  -0.1304   4.4525
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.064e+01  1.237e+00  -8.602  < 2e-16
## DAYS                            -2.487e-03  2.013e-04 -12.356  < 2e-16
## log(VEHICLE_REG_MASS)            5.148e-01  1.609e-01   3.200  0.00137
## CL_VEHICLE_FUEL_groupElekter     1.351e+00  4.231e-01   3.192  0.00141
## CL_VEHICLE_FUEL_groupHübriid_CNG 6.049e-01  2.453e-01   2.466  0.01365
## VEHICLE_AGE                     -1.317e-02  4.146e-03  -3.176  0.00149
## GENDER1                          1.408e-01  5.079e-02   2.772  0.00557
## PERSON_AGE                      -5.799e-02  9.675e-03  -5.994 2.05e-09
## I(PERSON_AGE^2)                  5.404e-04  9.989e-05   5.410 6.30e-08
##
```

```
## (Intercept)                       ***
## DAYS                              ***
## log(VEHICLE_REG_MASS)             **
## CL_VEHICLE_FUEL_groupElekter      **
## CL_VEHICLE_FUEL_groupHübriid_CNG *
## VEHICLE_AGE                       **
## GENDER1                           **
## PERSON_AGE                        ***
## I(PERSON_AGE^2)                   ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.5136) family taken to be 1)
##
##     Null deviance: 12237  on 91649  degrees of freedom
## Residual deviance: 11960  on 91641  degrees of freedom
## AIC: 18080
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.5136
##           Std. Err.:  0.0912
##
##  2 x log-likelihood:  -18060.4340
```

Lisa 3. MARSi mudeli väljund

```
##
## Call:
## earth(formula = LOSS_COUNT ~ DAYS + log(VEHICLE_ENGINE_POWER) +
##     log(VEHICLE_REG_MASS) + VEHICLE_FRAME_TYPE + CL_VEHICLE_FUEL_group +
##     VEHICLE_AGE + GENDER + PERSON_AGE + I(PERSON_AGE^2) + offset(log(DAYS)),
##     data = sample_train, glm = list(family = "poisson"),
##     degree = 2, penalty = 3, thresh = 0.001)
## GLM coefficients
##                                    LOSS_COUNT
```

```
## (Intercept)                          -8.2826566
## h(56.1646-DAYS)                       0.0252340
## h(DAYS-56.1646)                      -0.0044025
## h(DAYS-210)                           0.0033860
## h(56.1646-DAYS) * h(VEHICLE_AGE-17) -0.0040288
## h(56.1646-DAYS) * h(17-VEHICLE_AGE) -0.0028478
##
## GLM (family poisson, link log):
##  nulldev    df       dev    df   devratio     AIC iters converged
##  14694.5 91649   14446.5 91644     0.0169   18180     7         1
##
## Earth selected 6 of 6 terms, and 2 of 12 predictors
## Termination condition: RSq changed by less than 0.001 at 6 terms
## Importance: DAYS, VEHICLE_AGE, log(VEHICLE_ENGINE_POWER)-unused, ...
## Offset: log(DAYS) with values log(364.284), log(91.99931), log(91.9...
## Number of terms at each degree of interaction: 1 3 2
## Earth GCV 0.05512956    RSS 5051.135    GRSq 0.95033    RSq 0.9503436
```

Lisa 4. MARSi tunnustega Poissoni mudeli väljund

```
##
## Call:
## glm(formula = LOSS_COUNT ~ DAYS56 + DAYS56_v + DAYS210 + DAYS56_v:V_AGE17 +
##     DAYS56_v:V_AGE17_v + GENDER + CL_VEHICLE_FUEL_group + PERSON_AGE +
##     I(PERSON_AGE^2) + log(VEHICLE_REG_MASS), family = "poisson",
##     data = sample_train, offset = log(DAYS))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6419  -0.2356  -0.1968  -0.1420   4.9261
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -1.151e+01  1.174e+00  -9.802  < 2e-16
## DAYS56                         -3.892e-03  6.434e-04  -6.049 1.46e-09
## DAYS56_v                        2.300e-02  4.740e-03   4.852 1.22e-06
## DAYS210                         2.927e-03  1.018e-03   2.877 0.004020
```

```
## GENDER1                             1.593e-01  4.910e-02   3.244 0.001178
## CL_VEHICLE_FUEL_groupElekter        1.482e+00  3.795e-01   3.907 9.36e-05
## CL_VEHICLE_FUEL_groupHübriid_CNG    7.139e-01  2.311e-01   3.089 0.002011
## PERSON_AGE                         -4.733e-02  9.422e-03  -5.023 5.08e-07
## I(PERSON_AGE^2)                     4.337e-04  9.712e-05   4.466 7.97e-06
## log(VEHICLE_REG_MASS)               5.654e-01  1.543e-01   3.664 0.000248
## DAYS56_v:V_AGE17                   -3.580e-03  1.788e-03  -2.002 0.045293
## DAYS56_v:V_AGE17_v                 -2.556e-03  5.451e-04  -4.688 2.75e-06
##
## (Intercept)                        ***
## DAYS56                             ***
## DAYS56_v                           ***
## DAYS210                            **
## GENDER1                            **
## CL_VEHICLE_FUEL_groupElekter       ***
## CL_VEHICLE_FUEL_groupHübriid_CNG   **
## PERSON_AGE                         ***
## I(PERSON_AGE^2)                    ***
## log(VEHICLE_REG_MASS)              ***
## DAYS56_v:V_AGE17                   *
## DAYS56_v:V_AGE17_v                 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 14695  on 91649  degrees of freedom
## Residual deviance: 14377  on 91638  degrees of freedom
## AIC: 18123
##
## Number of Fisher Scoring iterations: 7
```

Lisa 5. MARSi tunnustega negatiivse binoomjaotuse mudeli väljund

```
##
## Call:
## glm.nb(formula = LOSS_COUNT ~ DAYS56 + DAYS56_v + DAYS210 + DAYS56_v:V_AGE17 +
##     DAYS56_v:V_AGE17_v + GENDER + CL_VEHICLE_FUEL_group + PERSON_AGE +
##     I(PERSON_AGE^2) + log(VEHICLE_REG_MASS) + offset(log(DAYS)),
##     data = sample_train, init.theta = 0.5120736096, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5897  -0.2324  -0.1950  -0.1414   4.1146
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.156e+01  1.207e+00  -9.577  < 2e-16
## DAYS56                          -3.921e-03  6.581e-04  -5.957 2.56e-09
## DAYS56_v                         2.303e-02  4.794e-03   4.803 1.56e-06
## DAYS210                          2.964e-03  1.044e-03   2.840 0.004513
## GENDER1                          1.600e-01  5.046e-02   3.172 0.001516
## CL_VEHICLE_FUEL_groupElekter     1.492e+00  4.248e-01   3.513 0.000444
## CL_VEHICLE_FUEL_groupHübriid_CNG 7.130e-01  2.448e-01   2.912 0.003590
## PERSON_AGE                      -4.713e-02  9.687e-03  -4.865 1.15e-06
## I(PERSON_AGE^2)                  4.338e-04  9.988e-05   4.344 1.40e-05
## log(VEHICLE_REG_MASS)            5.706e-01  1.586e-01   3.598 0.000321
## DAYS56_v:V_AGE17                -3.565e-03  1.800e-03  -1.981 0.047647
## DAYS56_v:V_AGE17_v              -2.559e-03  5.492e-04  -4.659 3.17e-06
##
## (Intercept)                     ***
## DAYS56                          ***
## DAYS56_v                        ***
## DAYS210                         **
## GENDER1                         **
## CL_VEHICLE_FUEL_groupElekter    ***
## CL_VEHICLE_FUEL_groupHübriid_CNG **
## PERSON_AGE                      ***
```

```
## I(PERSON_AGE^2)                  ***
## log(VEHICLE_REG_MASS)            ***
## DAYS56_v:V_AGE17                 *
## DAYS56_v:V_AGE17_v               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.5121) family taken to be 1)
##
##     Null deviance: 12232  on 91649  degrees of freedom
## Residual deviance: 11929  on 91638  degrees of freedom
## AIC: 18060
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.5121
##           Std. Err.:  0.0907
##
##  2 x log-likelihood:  -18034.1170
```

Lisa 6. R-i kood

```
library(ggplot2)
library(gridExtra)
#MARSi baasfunktsioonid
t<-3
y<-c(pmax(x[1:11]-t,0),pmax(t-x[1:11],0))
data2<-data.frame(x=x,y=y,jrk=rep(c("A","B"),each=length(x)/2))
t<-8
y<-c(pmax(x[1:11]-t,0),pmax(t-x[1:11],0))
data3<-data.frame(x=x,y=y,jrk=rep(c("A","B"),each=length(x)/2))
data4<-data.frame(x=seq(0,3,length.out = 22),jrk="A",y=0)
data5<-data.frame(x=seq(3,5,length.out = 22),jrk="A",y=0)
ggplot(data, mapping=aes(x=x,y=y,group=jrk,linetype=jrk))+
  geom_line(size=2,color="goldenrod2")+
  theme(legend.position="none",text = element_text(size=40),
  panel.background =element_blank(), panel.border = element_rect(size=1,fill=NA),
```

```
  panel.grid.major = element_line(colour="lightgray"))+
  scale_linetype_manual(values=c("solid","dotted"))+
  geom_vline(xintercept=5,size=2,linetype="dashed",colour="gray52")+
  ylab("Baasfunktsioon")+geom_line(data3,mapping=aes(x=x,y=y),color="olivedrab4",
  size=2)+geom_line(data2,mapping=aes(x=x,y=y),color="midnightblue",size=2)+
  geom_line(size=2,color="goldenrod2")+geom_line(data4,mapping=aes(x=x,y=y),
  linetype="longdash",color="midnightblue", size=2)+geom_line(data4,
  mapping=aes(x=x,y=y),linetype="dashed",color="goldenrod2",size=2)+
  geom_line(data4,mapping=aes(x=x,y=y),linetype="dashed",color="olivedrab4",
  size=2)+geom_line(data5,mapping=aes(x=x,y=y),linetype="dashed",
  color="olivedrab4",size=2)+geom_line(data5,mapping=aes(x=x,y=y),
  linetype="dotted",color="midnightblue",size=2)

#joonis poissoni ja NB jaotuste kohta
t<-0:30
mu=10
k=2
pi=k/(k+mu)
data11<-data.frame(x=c(t,t),y=c(dpois(t,mu),dnbinom(t,mu=mu,size=2)),
  grupp=c(rep("Po(10)",31),rep("NB(2,1/6)",31)))
data12<-data.frame(x=c(t,t),y=c(dpois(t,mu),dnbinom(t,mu=mu,size=10)),
  grupp=c(rep("Po(10)",31),rep("NB(10,1/2)",31)))
g1<-ggplot(data11,mapping=aes(x=x,y=y,group=grupp,shape=grupp,color=grupp))+
  geom_point(size=6)+theme(legend.position = c(0.8, 0.8),
  legend.title=element_blank(),text=element_text(size=40),
  panel.background=element_blank(),
  panel.grid.major = element_line(colour="lightgray"),
  panel.border = element_rect(size=1,fill=NA))+
  scale_color_manual(values = c("olivedrab4","goldenrod2"))+
  scale_shape_manual(values=c(18,19))
g2<-ggplot(data12,mapping=aes(x=x,y=y,group=grupp,shape=grupp,color=grupp))+
  geom_point(size=6)+theme(legend.position = c(0.8, 0.8),
  legend.title=element_blank(),text = element_text(size=40),
  panel.background =element_blank(),
  panel.grid.major = element_line(colour="lightgray"),
  panel.border = element_rect(size=1,fill=NA))+
  scale_color_manual(values = c("olivedrab4","goldenrod2"))+
  scale_shape_manual(values=c(18,19))
grid.arrange(g1, g2, ncol=2)
```

```
#praktiline osa
library(car)
library(MASS)
library(earth)
library(pscl)
library(ggplot2)
library(gridExtra)
#andmestiku sisselugemine, korrastamine
load("C:/users/perttu/documents/magister/valim.Rdata")
nrow(valim)
#andmestiku suurus 138 642
sample<-valim
#fikseerin keretüübi
sample<-sample[sample$VEHICLE_FRAME_TYPE%in%c("LUUKPÄRA", "MAHTUNIVERSAAL",
  "SEDAAN","UNIVERSAAL"),]
sample$CL_VEHICLE_FUEL_group<-sample$CL_VEHICLE_FUEL
sample$CL_VEHICLE_FUEL_group[sample$CL_VEHICLE_FUEL=="Bensiin"]<-
  "Bensiin_diisel_kat"
sample$CL_VEHICLE_FUEL_group[sample$CL_VEHICLE_FUEL=="Diisel"]<-
  "Bensiin_diisel_kat"
sample$CL_VEHICLE_FUEL_group[sample$CL_VEHICLE_FUEL=="Bensiin-Kat"]<-
  "Bensiin_diisel_kat"
sample$CL_VEHICLE_FUEL_group[sample$CL_VEHICLE_FUEL=="Bensiin-Hübriid"]<-
  "Hübriid_CNG"
sample$CL_VEHICLE_FUEL_group[sample$CL_VEHICLE_FUEL=="Surugaas (CNG)"]<-
  "Hübriid_CNG"
#jätame kõrvale kõik vaatlused, mille isikukoodi algus ei ole numbriline -
#pole tegemist eesti isikukoodiga, ei saa eraldada sugu ega vanust
sample<-sample[!(is.na(as.numeric(sample$KK_SSID))),]
#sugu
GENDER_temp<-substring(sample$KK_SSID,1,1)
#0 - male, 1 - female
sample$GENDER<-GENDER_temp
sample$GENDER[GENDER_temp%in%c(3,5)]<-'0'
sample$GENDER[GENDER_temp%in%c(4,6)]<-'1'
#vanus
BIRTHYEAR<-as.numeric(substring(sample$KK_SSID,2,3))
BIRTHYEAR[GENDER_temp%in%c(5,6)]<-2000+BIRTHYEAR[GENDER_temp%in%c(5,6)]
```

```
BIRTHYEAR[GENDER_temp%in%c(3,4)]<-1900+BIRTHYEAR[GENDER_temp%in%c(3,4)]
sample$PERSON_AGE<-sample$START_YEAR-BIRTHYEAR
#auto vanus
sample$VEHICLE_AGE<-sample$START_YEAR-sample$VEHICLE_PRIME_REGISTRATION
#lõplik andmestik
sample_final<-sample[sample$PERSON_AGE>=18 & sample$PERSON_AGE<=85 &
  sample$DAYS>=1 & !(sample$CL_VEHICLE_FUEL=="Diisel-Hübriid") &
  sample$VEHICLE_ENGINE_POWER>7 & sample$VEHICLE_AGE<=25,]
sample_final$CL_VEHICLE_FUEL_group<-
  relevel(as.factor(sample_final$CL_VEHICLE_FUEL_group),
  ref="Bensiin_diisel_kat")
nrow(sample_final)
summary(sample_final)
sd(sample_final$DAYS);sd(sample_final$LOSS_COUNT);
sd(sample_final$VEHICLE_ENGINE_POWER);sd(sample_final$VEHICLE_REG_MASS);
sd(sample_final$PERSON_AGE);sd(sample_final$VEHICLE_AGE)
round(prop.table(table(sample_final$GENDER))*100,1)
round(prop.table(table(sample_final$VEHICLE_FRAME_TYPE))*100,1)
round(prop.table(table(sample_final$CL_VEHICLE_FUEL))*100,1)

#jaotamine treening- ja  testandmestikuks (70-30)
set.seed(85)
ss <- sample(1:2,size=nrow(sample_final),replace=TRUE,prob=c(0.7,0.3))
sample_train<-sample_final[ss==1,]
sample_test<-sample_final[ss==2,]
nrow(sample_train);nrow(sample_test)

#Poissoni mudeli loomine
mP<-glm(LOSS_COUNT~DAYS+log(VEHICLE_ENGINE_POWER)+log(VEHICLE_REG_MASS)+
  VEHICLE_FRAME_TYPE+CL_VEHICLE_FUEL_group+VEHICLE_AGE+GENDER+PERSON_AGE+
  I(PERSON_AGE^2), family="poisson",data=sample_train, offset=log(DAYS))
summary(mP) #AIC = 18 143
#Mootori võimsus välja (p=0,69)
mP<-glm(LOSS_COUNT~DAYS+log(VEHICLE_REG_MASS)+VEHICLE_FRAME_TYPE+
  CL_VEHICLE_FUEL_group+VEHICLE_AGE+GENDER+PERSON_AGE+I(PERSON_AGE^2),
  family="poisson",data=sample_train, offset=log(DAYS))
summary(mP) #AIC = 18 141
#keretüüp ka välja
mP<-glm(LOSS_COUNT~DAYS+log(VEHICLE_REG_MASS)+CL_VEHICLE_FUEL_group+
```

```
  VEHICLE_AGE+GENDER+PERSON_AGE+I(PERSON_AGE^2),
  family="poisson",data=sample_train, offset=log(DAYS))
summary(mP) #AIC = 18 143

#Negatiivse binoomjaotuse mudeli loomine
mNB<-glm.nb(LOSS_COUNT~DAYS+log(VEHICLE_ENGINE_POWER)+log(VEHICLE_REG_MASS)+
  VEHICLE_FRAME_TYPE+CL_VEHICLE_FUEL_group+VEHICLE_AGE+GENDER+PERSON_AGE+
  I(PERSON_AGE^2)+offset(log(DAYS)), data=sample_train)
summary(mNB) #AIC = 18 081
#mootori võimsus välja
mNB<-glm.nb(LOSS_COUNT~DAYS+log(VEHICLE_REG_MASS)+VEHICLE_FRAME_TYPE+
 CL_VEHICLE_FUEL_group+VEHICLE_AGE+GENDER+PERSON_AGE+I(PERSON_AGE^2)+
  offset(log(DAYS)), data=sample_train)
summary(mNB) #AIC = 18 080
#keretüüp välja
mNB<-glm.nb(LOSS_COUNT~DAYS+log(VEHICLE_REG_MASS)+CL_VEHICLE_FUEL_group+
  VEHICLE_AGE+GENDER+PERSON_AGE+I(PERSON_AGE^2)+offset(log(DAYS)),
  data=sample_train)
summary(mNB) #AIC = 18 080

###MARS
mMARS<-earth(LOSS_COUNT~DAYS+log(VEHICLE_ENGINE_POWER)+log(VEHICLE_REG_MASS)+
  VEHICLE_FRAME_TYPE+CL_VEHICLE_FUEL_group+VEHICLE_AGE+GENDER+PERSON_AGE+
  offset(log(DAYS)), data=sample_train, degree=2, penalty = 3,thresh = 0.001,
  glm=list(family="poisson"))
summary(mMARS)

#teeme glmi tarbeks uued muutujad, alustan täiesti esimesest mudelist
sample_train$DAYS56<-
  (sample_train$DAYS-56.1646)*((sample_train$DAYS>56.1646)*1)
sample_train$DAYS56_v<-
  (56.1646-sample_train$DAYS)*((sample_train$DAYS<56.1646)*1)
sample_test$DAYS56<-
  (sample_test$DAYS-56.1646)*((sample_test$DAYS>56.1646)*1)
sample_test$DAYS56_v<-
  (56.1646-sample_test$DAYS)*((sample_test$DAYS<56.1646)*1)
sample_train$DAYS210<-
  (sample_train$DAYS-210)*((sample_train$DAYS>210)*1)
sample_test$DAYS210<-
```

```
  (sample_test$DAYS-210)*((sample_test$DAYS>210)*1)
sample_train$V_AGE17<-
  (sample_train$VEHICLE_AGE-17)*((sample_train$VEHICLE_AGE>17)*1)
sample_train$V_AGE17_v<-
  (17-sample_train$VEHICLE_AGE)*((sample_train$VEHICLE_AGE<17)*1)
sample_test$V_AGE17<-
  (sample_test$VEHICLE_AGE-17)*((sample_test$VEHICLE_AGE>17)*1)
sample_test$V_AGE17_v<-
  (17-sample_test$VEHICLE_AGE)*((sample_test$VEHICLE_AGE<17)*1)

mMARS_glm_po<-glm(LOSS_COUNT~DAYS56+DAYS56_v+DAYS210+DAYS56_v:V_AGE17+
  DAYS56_v:V_AGE17_v,offset=log(DAYS),data=sample_train,family="poisson")
#AIC18 180
summary(mMARS_glm_po)
#Mars(Po)+manuaalne
mMARS_glm_po_add<-glm(LOSS_COUNT~DAYS56+DAYS56_v+DAYS210+DAYS56_v:V_AGE17+
  DAYS56_v:V_AGE17_v+GENDER+CL_VEHICLE_FUEL_group+PERSON_AGE+I(PERSON_AGE^2)+
  log(VEHICLE_REG_MASS), offset=log(DAYS),data=sample_train,family="poisson")
summary(mMARS_glm_po_add) #AIC = 18 123
#Mars(NB)+manuaalne
mMARS_glm_nb_add<-glm.nb(LOSS_COUNT~DAYS56+DAYS56_v+DAYS210+DAYS56_v:V_AGE17+
  DAYS56_v:V_AGE17_v+GENDER+CL_VEHICLE_FUEL_group+PERSON_AGE+I(PERSON_AGE^2)+
  log(VEHICLE_REG_MASS)+offset(log(DAYS)),data=sample_train)
summary(mMARS_glm_nb_add) #AIC = 18 060

#AIC
AIC(mP, mNB, mMARS_glm_po, mMARS_glm_po_add,mMARS_glm_nb_add)
#RSS
#poisson
mP_pred<-predict(mP,newdata=sample_test,type="response")
sum((sample_test$LOSS_COUNT-mP_pred)^2) #863,396
#NB
mNB_pred<-predict(mNB,newdata=sample_test,type="response")
sum((sample_test$LOSS_COUNT-mNB_pred)^2) #863,392
#MARS
mMARS_glm_po_pred<-predict(mMARS_glm_po,newdata=sample_test,type="response")
sum((sample_test$LOSS_COUNT-mMARS_glm_po_pred)^2) #863,714
#MARS+Po
mMARS_glm_po_add_pred<-predict(mMARS_glm_po_add,newdata=sample_test,
```

```
  type="response")
sum((sample_test$LOSS_COUNT-mMARS_glm_po_add_pred)^2) #863,560
#MARS+NB
mMARS_glm_nb_add_pred<-predict(mMARS_glm_nb_add,newdata=sample_test,
  type="response")
sum((sample_test$LOSS_COUNT-mMARS_glm_nb_add_pred)^2) #863,596

#vuong
vuong(mNB,mMARS_glm_nb_add)
vuong(mP,mNB) #nb parem

#prognoositud kahjude arv
DAYS_g<-cut(sample_test$DAYS,breaks=c(min(sample_test$DAYS),56.2,210,
  max(sample_test$DAYS)),include.lowest = TRUE,right=TRUE)
sample_test2<-data.frame(LOSS_COUNT=sample_test$LOSS_COUNT, mP_pred, mNB_pred,
  mMARS_glm_po_pred, mMARS_glm_po_add_pred,mMARS_glm_nb_add_pred, DAYS_g)
aggregate(.~DAYS_g, data=sample_test2,FUN="sum")
colSums(sample_test2[,1:6])
#päevade arvu mõju
days_x<-seq(min(sample_train$DAYS),max(sample_train$DAYS),by=.1)
h<-function(x){
  return(0.023*pmax(56.2-x,0)-0.0039*pmax(x-56.2,0)+0.003*pmax(x-210,0))
}
days_y<-h(days_x)
data13<-data.frame(x=days_x,y=exp(days_y))
ggplot(data13,mapping=aes(x=x,y=y))+geom_line(size=2,color="olivedrab4")+
  theme(legend.position = "none",legend.title=element_blank(),
  text = element_text(size=40),panel.background =element_blank(),
  panel.grid.major = element_line(colour="lightgray"),
  panel.border = element_rect(size=1,fill=NA))+xlab("Katte kehtivuse pikkus")+
  ylab("Mõju oodatavale kahjude arvule")
#vanuse mõju
vanus_x<-seq(min(sample_train$PERSON_AGE),max(sample_train$PERSON_AGE),by=0.1)
vanus_y<--0.047*vanus_x+0.00043*vanus_x^2
data15<-data.frame(x=vanus_x, y=exp(vanus_y)
ggplot(data15,mapping=aes(x=x,y=y))+geom_line(size=2,color="olivedrab4")+
  theme(legend.position = "none",text = element_text(size=40),
  panel.background =element_blank(),
  panel.grid.major = element_line(colour="lightgray"),
```

```
  panel.border = element_rect(size=1,fill=NA))+xlab("Kindlustusvõtja vanus")+
  ylab("Mõju oodatavale kahjude arvule")

#loendame kütusetüübi järgi
data_fuel<-data.frame(CL_VEHICLE_FUEL_group=sample_test$CL_VEHICLE_FUEL_group,
  LOSS_COUNT=sample_test$LOSS_COUNT, mMARS_glm_nb_add_pred)
aggregate(.~CL_VEHICLE_FUEL_group, data=data_fuel,FUN="sum")
#sugu
data_gender<-data.frame(GENDER=sample_test$GENDER,LOSS_COUNT=sample_test$LOSS_COUNT,
  mMARS_glm_nb_add_pred)
aggregate(.~GENDER, data=data_gender,FUN="sum")
#isiku vanus
person_quant<-quantile(sample_test$PERSON_AGE,prob=seq(0,1,by=.2))
data_age<-data.frame(PERSON_AGE_g=cut(sample_test$PERSON_AGE,
  breaks=as.numeric(person_quant),
  include.lowest = TRUE,right=TRUE), LOSS_COUNT=sample_test$LOSS_COUNT,
  mMARS_glm_nb_add_pred)
aggregate(.~PERSON_AGE_g, data=data_age,FUN="sum")
#registrimass
reg_mass_quant<-quantile(sample_test$VEHICLE_REG_MASS,prob=seq(0,1,by=.2))
data_mass<-data.frame(REG_MASS_g=cut(sample_test$VEHICLE_REG_MASS,
  breaks=as.numeric(reg_mass_quant),include.lowest = TRUE,right=TRUE),
  LOSS_COUNT=sample_test$LOSS_COUNT, mMARS_glm_nb_add_pred)
aggregate(.~REG_MASS_g, data=data_mass,FUN="sum")
```

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Perttu Narvik,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Sõidukite kindlustuskahjude sageduse hindamine mitmemõõtmelise adaptiivse regressioonsplaini abil", mille juhendajad on Meelis Käärik ja Tõnis Maldre, 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.

Perttu Narvik

15.05.2019