Tampereen yliopisto

Walter Sirén

YLEISTETYT LINEAARISET MALLIT VAKUUTUSTEN HINNOITTELUSSA

Informaatioteknologian ja viestinnän tiedekunta
Pro gradu -tutkielma
Huhtikuu 2024

TIIVISTELMÄ

Walter Sirén: Yleistetyt lineaariset mallit vakuutusten hinnoittelussa
Pro gradu -tutkielma
Tampereen yliopisto
Master's Programme in Computing Sciences, Statistical Data Analytics
Huhtikuu 2024

Yleistetyt lineaariset mallit tarjoavat joustavan työkalun erilaisten vakuutusturvien riskimaksurakenteen muodostamiseksi. Tässä tutkielmassa perehdytään yleistettyjen lineaaristen mallien teoriaan sekä esitetään, kuinka vakuutettavan riskin vahinkotiheyttä ja keskivahinkoa voidaan mallintaa käyttämällä yleistettyjä lineaarisia malleja. Tutkielmassa käytetty aineisto on vuodelta 2021.

Yleistettyjä lineaarisia malleja voidaan käyttää myös osana koosteoppimismenetelmiä. Tutkielmassa pyritään parantamaan perinteisen menetelmän ennustetarkkuutta luomalla ensin useita yleistettyjä lineaarisia malleja, joista lopullinen ennuste muodostetaan ottamalla keskiarvo kaikista yksittäisistä ennusteista. Menetelmä ei onnistu parantamaan ennustetta testiaineistossa, kun käytettävänä mittarina on keskineliövirhe.

Seuraavassa vaiheessa muodostetaan testiaineistolle riskimaksu vahinkotiheysmallin sekä keskivahinkomallin tulona. Riskimaksun piste-estimaatin lisäksi näytetään, kuinka yksittäiselle riskimaksulle voidaan muodostaa ennusteväli. Lopuksi esitetään simulointimenetelmä riskimaksun varmuuslisän asettamiseksi.

Avainsanat: Yleistetty lineaarinen malli, riskimaksu, koosteoppiminen, varmuuslisä

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla.

1 JOHDANTO 1

1.1 ESIMERKKI RISKITEKIJÖIDEN VAIKUTUKSESTA VAHINGON ODOTUSARVOON 1
1.2 TYÖN RAKENNE 2

2 YLEISTETTYJEN LINEAARISTEN MALLIEN PERUSTEET 4

2.1 YLEISTETYN LINEAARISEN MALLIN RAKENNE 4
2.2 EKSPONENTIAALINEN JAKAUMAPERHE JA USKOTTAVUUSFUNKTIOT 5
2.3 USKOTTAVUUSYHTÄLÖIDEN JOHTAMINEN YLEISTETYILLE LINEAARISILLE MALLEILLE 7
2.4 NEWTON-RAPHSONIN MENETELMÄ 9
2.5 NEWTON-RAPHSONIN ALGORITMIN TOTEUTUS POISSON-REGRESSIOLLE 10

3 MALLINNUSPROSESSI YLEISTETYILLÄ LINEAARISILLA MALLEILLA 13

3.1 AKAIKEN INFORMAATIOKRITEERI (AIC) JA KESKINELIÖVIRHE (MSE) 13
3.2 KUVAILEVA ANALYYSI VAHINKOTIHEYDESTÄ 14
3.3 VAHINKOTIHEYDEN MALLINTAMINEN 16
3.4 KUVAILEVA ANALYYSI KESKIVAHINGOSTA 19
3.5 KESKIVAHINGON MALLINTAMINEN 21

4 KOOSTEOPPIMINEN ENNUSTEEN LUOMISESSA 24

4.1 MENETELMÄN KUVAUS 25
4.2 BOOTSTRAP 26
4.3 ETENEVÄ VALINTA – STEPAIC 26
4.4 ENNUSTEET YKSITTÄISILLÄ MALLEILLA JA KOOSTAMINEN 28
4.5 KOOSTEOPPIMISMALLIN VERTAILU PERINTEISEEN VAHINKOTIHEYSMALLIIN 28

5 RISKIMAKSUN MUODOSTAMINEN 30

5.1 ENNUSTEVÄLIN MUODOSTAMINEN YKSITTÄISELLE RISKIMAKSULLE 31
5.2 RISKIMAKSUN RIITTÄVYYS JA VARMUUSLISÄ 32

6 YHTEENVETO 34

LÄHTEET 36

LIITTEET 37

LIITE 1: NEWTON-RAPHSONIN MENETELMÄ PARAMETRIEN RATKAISEMISEKSI. R-KOODI. 37
LIITE 2: KOOSTEOPPIMISMENETELMÄN TOTEUTUS. R-KOODI. 38

1 Johdanto

Vakuutussopimuksen solmimalla vakuutettu siirtää taloudellista riskiä itseltään vakuutusyhtiölle. Vakuutusyhtiö sitoutuu korvaamaan mahdollisesti tulevaisuudessa tapahtuvan vahingon tehdyn vakuutussopimuksen ehtojen mukaisesti. Vastaavasti vakuutettu maksaa riskin siirtämisestä tai pienentämisestä vakuutusyhtiölle vakuutusmaksun. Vakuutusmaksun suuruus riippuu vakuutettavasta riskistä ja riskin suuruuteen vaikuttavista tekijöistä. Nämä tekijät vaihtelevat vakuutuksenottajilla, ja vakuutusyhtiön tavoitteena onkin asettaa jokaisen vakuutetun vakuutusmaksu riskiä vastaavalle tasolle.

Haitallisessa valikoitumisessa puolestaan on kyse epäsymmetrisestä informaatiosta vakuutuksenantajan ja vakuutetun välillä. Puutteellisella informaatiolla kasattu hinnoittelurakenne voi johtaa yhtiön sisäisiin kannattavuusongelmiin vakuutuskannan muodostuessa epätasapainoiseksi, kun vakuutettujen vaihtelevia riskiominaisuuksia ei tunnisteta tarpeeksi tehokkaasti [Zweifel et al. 2021]. Tilannetta voidaan selventää yksikertaisella esimerkillä.

1.1 Esimerkki riskitekijöiden vaikutuksesta vahingon odotusarvoon

Olkoot A ja B saman vakuutuskannan kaksi riskeiltään erilaista ryhmää. Ryhmälle A vahinkoja sattuu sata kappaletta tuhatta vakuutusvuotta kohden ja ryhmälle B sattumistiheys on viisikymmentä kappaletta tuhatta vakuutusvuotta kohden. Molempien ryhmien keskivahingon suuruus on yksinkertaisuuden vuoksi täysin sama, tuhat euroa. Ryhmien vahinkojen odotusarvoa merkitään seuraavasti:

$$\mathrm{E(A)} = \text{vahinkotiheys} \times \text{keskivahinko} = 0.1 \times 1000€ = 100€$$

$$\mathrm{E(B)} = \text{vahinkotiheys} \times \text{keskivahinko} = 0.05 \times 1000€ = 50€$$

Kuvitellaan tilanne, jossa vakuutuskannassa on yhtä paljon molempien ryhmien edustajia. Vakuutetun kannan odotusarvo vahinkomenolle on tässä tapauksessa:

$$\mathrm{E(A\ ja\ B)} = \frac{1}{2}\mathrm{E(A)} + \frac{1}{2}\mathrm{E(B)} = 0.5 \times 100€ + 0.5 \times 50€ = 75€$$

Mikäli ryhmien erilaista riskiä ei tunnistettaisi ja maksu päädyttäisiin asettamaan kaikille yhtä suureksi (75 €), maksaisi ryhmän B edustajat 25 euroa enemmän kuin heidän vahinkomenonsa odotusarvo on. Samalla hinnoittelurakenne suosii ryhmän A edustajia asettamalla maksun 25 euroa edullisemmaksi ryhmän vahingon odotusarvoon nähden.

1.2 Työn rakenne

Vakuutustuotteen hinnoittelussa pohjimmiltaan on siis kyse riskien luokittelusta ja riskin suuruuteen vaikuttavien tekijöiden löytämisestä. Vakuutusmaksussa huomioidaan syntyvän vahingon odotusarvon lisäksi esimerkiksi vahingon selvittelystä syntyvät kulut. Rantalan ja Kivisaaren [2020] mukaan vakuutusmaksu jaetaankin tavallisesti neljään osaan:

1. Riskimaksu
2. Hoitokulukuormitus
3. Riskilisä
4. Voittomarginaali

Tässä tutkielmassa tarkastellaan riskimaksun muodostamista tilastollisesti mallintamalla. Kuvatut menetelmät pyrkivät myös huomioimaan tilastollisten mallien epävarmuutta, jolloin lopullinen hinnoittelurakenne voidaan asettaa riittävän turvaavalle tasolle. Näin käsitellään mukautetuin osin myös riskilisän osuutta vakuutusmaksusta.

Riskimaksun muodostamiseksi on perinteisesti käytetty yleistettyjä lineaarisia malleja. Toisessa luvussa käydään läpi yleistettyjen lineaaristen mallien teoriapohjaa ja esitellään malliperheen rakennetta sekä mallien sovittamiseksi tarvittavien uskottavuusyhtälöiden johtamista. Lisäksi luvussa toteutetaan iteratiivinen algoritmi mallin parametrien ratkaisemiseksi.

Kolmannessa luvussa luodaan yleistetyt lineaariset mallit sekä vahinkotiheydelle että keskivahingolle. Mallinnus aloitetaan tekemällä molemmille vastemuuttujille kuvaileva analyysi. Mallien luomisessa hyödynnetään tunnettuja mittareita ja tunnuslukuja, mutta tehdään myös näkyväksi valintojen perustaminen kyseessä olevan aineiston asiantuntijaosaamiseen.

Neljännessä luvussa esitellään vaihtoehtoinen menetelmä vahinkotiheyden mallintamiseksi. Kyseessä on koosteoppimisen sovellus, jonka tavoitteena on luoda aineistosta useita malleja ennustetarkkuuden sekä yleistettävyyden parantamiseksi. Usean mallin ennusteet kerätään yhteen käyttämällä yksinkertaisesti keskiarvoa jokaisen yksittäisen mallin ennusteesta. Menetelmää voisi yhtä lailla hyödyntää vahinkotiheyden lisäksi keskivahingon mallinnuksessa.

Viidennessä luvussa yhdistetään muodostettujen mallien tulokset riskimaksun hinnoittelurakenteeksi. Muodostettu hintarakenne on tehty vain tämän työn esitystä varten, eikä se ole käytössä yhdenkään vakuutusyhtiön todellisena riskimaksuna. Lisäksi näytetään, kuinka muodostetun riskimaksun piste-estimaatin ohella voidaan tarkastella riskimaksun epävarmuuden suuruutta luomalla riskimaksulle ennusteväli. Riskimaksun varmuuslisäksi esitetään yksinkertaista menettelyä sopivan turvaavuuden varmistamiseksi. Kuudennessa luvussa käydään yhteenvetona tehdyn tutkimuksen vaiheet ja pohditaan jatkokehittämisen mahdollisuuksia.

2 Yleistettyjen lineaaristen mallien perusteet

Riskien luokittelussa, ja siten myös hinnoittelumallin rakentamisessa, tavoitteena on löytää kuinka vastemuuttuja Y (vahinkotiheys tai keskivahinko) riippuu yhdestä tai useammasta selittävästä tekijästä. Lineaarinen regressioanalyysi ei kuitenkaan usein ole kaikkein sopivin tilastollinen menetelmä vahinkovakuutusten mallintamiseen, sillä siinä virheiden oletetaan jakautuvan normaalisesti. Vahinkojen lukumäärä noudattaa diskreettiä todennäköisyysjakaumaa positiivisilla kokonaisluvuilla ja vahinkomeno ei-negatiivista, usein voimakkaasti oikealle vinoa, jatkuvaa jakaumaa. Näin ollen yleistetyt lineaariset mallit tarjoavat sopivamman menetelmän vastemuuttujien, vahinkotiheyden sekä vahinkomenon, mallintamiseen laajentamalla lineaarisen regression normaalijakaumaoletuksen muihin eksponentiaalisen jakaumaperheen jakaumiin.

2.1 Yleistetyn lineaarisen mallin rakenne

Yleistetty lineaarinen malli koostuu kolmesta komponentista [Agresti 2019].

1. *Satunnaiskomponentti* määrittää riippuvan muuttujan Y ja olettaa sille jonkin todennäköisyysjakauman. Satunnaismuuttujan Y havaintoyksikkö Y_i voi olla esimerkiksi binäärinen (0 tai 1), jolloin satunnaismuuttujan Y oletetaan noudattavan binomijakaumaa. Usealla sovellusalalla, kuten esimerkiksi vakuutusyhtiön mallintaessa vahinkokappaleita, vastemuuttuja voi saada arvoikseen ei-negatiivisia kokonaislukuja. Tällöin voidaan olettaa, että Y noudattaa esimerkiksi Poissonin jakaumaa. Toisaalta mallinnettaessa vahinkomenoa saa havaintoyksikkö Y_i ei-negatiivisia jatkuvia arvoja, jolloin on sopivaa olettaa vastemuuttujan noudattavan gammajakaumaa. Eri havaintoyksiköihin i ja i^* liittyvien satunnaismuuttujien Y_i ja Y_{i^*} oletetaan olevan toisistaan riippumattomia.

2. *Lineaarinen ennustin,* joka asettaa selittävät muuttujat $x_1, x_2, \dots\ x_p$ osaksi mallia. Tämä komponentti voidaan esittää muodossa:

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.1}$$

3. *Linkkifunktio* on sellainen funktio $g(\cdot)$, joka yhdistää selitettävän muuttujan Y odotusarvon μ mallin lineaariseen ennustimeen:

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.2}$$

Yksinkertaisin vaihtoehto linkkifunktiolle on niin sanottu identiteettilinkki $g(\mu) = \mu$. Tällöin malli voidaan esittää muodossa:

$$\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.3}$$

Lineaarisessa (tavallisessa) regressiomallissa oletetaan nimenomaan selitettävän muuttujan Y odotusarvon μ riippuvan lineaarisesti selittävistä tekijöistä $x_1, x_2, \ldots\ x_p$ kaavan (2.3) mukaisesti.

Yleistetty lineaarinen malli mahdollistaa myös toisenlaisten linkkifunktioiden käyttämisen. Lukumäärien mallinnuksessa hyödyllinen vaihtoehto on log-linkki $g(\mu) = \log(\mu)$, sillä μ ei voi olla negatiivinen. Tällaista mallia kutsutaan log-lineaariseksi malliksi ja se on muotoa:

$$log(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{2.4}$$

2.2 Eksponentiaalinen jakaumaperhe ja uskottavuusfunktiot

Alaluvussa 2.1 oletettiin havaintoyksikön Y_i noudattavan jotakin todennäköisyysjakaumaa. Yleistettyjen lineaaristen mallien tapauksessa todennäköisyysjakauma on aina peräisin eksponentiaalisesta jakaumaperheestä. Esimerkiksi Myers et al. [2012] määrittää kuinka jakaumaperheen kaikki todennäköisyysjakaumat, kuten Poisson-jakauma tai gammajakauma, voidaan esittää yleisessä muodossa yhtälön (2.5) mukaisesti

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}. \tag{2.5}$$

Parametria θ_i kutsutaan luonnolliseksi parametriksi ja parametria ϕ hajontaparametriksi. $a(\cdot), b(\cdot)$ ja $c(\cdot)$ ovat tiettyjä funktioita. Taulukossa 1 on listattu joitakin eksponentiaalisen jakaumaperheen todennäköisyysjakaumia ja niiden parametrit.

Taulukko 1: Eksponentiaalisen jakaumaperheen jakaumia ja parametreja

Jakauma	**θ**	**b(θ)**	$\boldsymbol{\phi}$	**E(y)**
Poisson (μ)	$\ln \mu$	e^{θ}	1	μ
Normaali (μ,σ^2)	μ	$\frac{1}{2}\theta^2$	σ^2	μ
Gamma (μ,ν)	$-\frac{1}{\mu}$	$-\ln(-\theta)$	$\frac{1}{\nu}$	μ
Käänteinen normaali (μ,σ^2)	$-\frac{1}{2\mu^2}$	$-\sqrt{-2\theta}$	σ^2	μ

Jokaisella satunnaismuuttujan Y_i realisaatioilla y_i on todennäköisyysfunktio f($y_i; \theta_i, \phi$). Havainnon todennäköisyys riippuu siis parametreista θ_i ja ϕ. Havainnot oletetaan tosistaan riippumattomaksi (kts. alaluku 2.1), jolloin niiden yhteistodennäköisyysjakauma voidaan esittää yksittäisten todennäköisyysfunktioiden tulona:

$$L^*(\boldsymbol{\theta} = (\theta_1, \theta_2, \dots \theta_n), \phi; y_1, y_2, \dots, y_n) = \prod_{i=1}^{n} \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right\}. \quad (2.6)$$

Yhtälöä (2.6) kutsutaan uskottavuusfunktioksi. Usein on kuitenkin laskennallisesti helpompi työstää edellisen sijaan sen logaritmoitua muotoa:

$$L(\boldsymbol{\theta}, \phi; y_1, y_2, \dots, y_n) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \quad (2.7)$$

Yhtälöä (2.7) kutsutaan log-uskottavuudeksi. Merkitään Agrestin [2015] tapaan L_i osoittamaan havainnon y_i vaikutusta log-uskottavuusfunktioon identiteetin $L = \sum_i L_i$ mukaisesti. Tällöin on voimassa:

$$L_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi). \tag{2.8}$$

Nyt voidaan yhtälöstä (2.8) johtaa ensimmäinen ja toinen osittaisderivaatta parametrin θ_i suhteen, yhden havainnon log-uskottavuusfunktiolle:

$$\frac{\partial L_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)}, \qquad \frac{\partial^2 L_i}{\partial \theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)}. \tag{2.9}$$

Tässä esityksessä otetaan annettuina Agrestin [2015] tapaan kaksi yleistä uskottavuus-funktioiden tulosta:

$$E\left(\frac{\partial L}{\partial \theta}\right) = 0, \qquad -E\left(\frac{\partial^2 L}{\partial^2 \theta^2}\right) = E\left(\frac{\partial L}{\partial \theta}\right)^2. \tag{2.10}$$

Yhdistämällä tulokset yhtälöistä (2.9) ja (2.10) johdetaan havainnon y_i odotusarvo:

$$E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right) = 0 \quad \text{jolloin} \quad E(y_i) = \mu_i = b'(\theta_i). \tag{2.11}$$

Varianssi puolestaan on:

$$\frac{-b''(\theta_i)}{a(\phi)} = E\left(\frac{y_i - b'(\theta_i)}{a(\phi)}\right)^2 = \frac{var(y_i)}{a(\phi)^2}$$

jolloin

$$var(y_i) = b''(\theta_i)(\phi). \tag{2.12}$$

2.3 Uskottavuusyhtälöiden johtaminen yleistetyille lineaarisille malleille

Alaluvussa 2.1 näytettiin, kuinka linkkifunktio $g(\cdot)$ yhdistää odotusarvon mallin lineaa-riseen ennustimeen. Merkitään matriisimuodossa:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}. \tag{2.13}$$

Yhtälössä (2.13) $\boldsymbol{X}$ on mallimatriisi, vektori $\boldsymbol{\beta}$ sisältää alkiot $\beta_0, \dots, \beta_p$ ja vektori $\boldsymbol{\mu}$ sisältää alkiot $\mu_1, \dots, \mu_n$. Esitetään yhtälön (2.7) log-uskottavuus nyt uudessa muodossa:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{2.14}$$

Merkintä $L(\boldsymbol{\beta})$ korostaa, kuinka luonnollinen parametri $\boldsymbol{\theta}$ riippuu yleistetyn lineaarisen mallin parametreista $\boldsymbol{\beta}$ [Agresti 2015]. Logaritmoidun uskottavuusfunktion osittaisderivaatat voidaan nyt esittää seuraavasti:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial L_i}{\partial \beta_j} = 0, \qquad \text{kun } j = 1, \dots, p. \tag{2.15}$$

Logaritmoidun uskottavuusfunktion osittaisderivaatan muodostamiseksi voidaan käyttää ketjusääntöä:

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{2.16}$$

Koska $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}$, niin on voimassa:

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}. \tag{2.17}$$

Nyt voidaan yhtälöiden (2.9), (2.11), (2.12) ja (2.17) tulokset sijoittaa yhtälöön (2.16). Yleistetyn lineaarisen mallin logaritmoitu uskottavuusyhtälö on:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i)x_{ij}}{var(y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0, \qquad j = 1, \dots, p. \tag{2.18}$$

Yhtälö (2.18) voidaan esittää myös matriisimuodossa

$$\boldsymbol{X}^T\boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})=\boldsymbol{0}, \tag{2.19}$$

missä D ja V ovat diagonaalimatriiseja:

$$\boldsymbol{D}=\begin{pmatrix}\frac{\partial\mu_1}{\partial\eta_1} & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & \frac{\partial\mu_n}{\partial\eta_n}\end{pmatrix} \text{ ja } \mathbf{V}=\begin{pmatrix}Var(y_1) & \cdots & 0\\ \vdots & \ddots & \vdots\\ 0 & \cdots & Var(y_n)\end{pmatrix}.$$

2.4 Newton-Raphsonin menetelmä

Alaluvussa 2.3 johdettiin logaritmoitu uskottavuusyhtälö yleistetyille lineaarisille malleille. Tavoitteena on johtaa yhtälöstä mallin parametrit $\boldsymbol{\beta}$ käyttäen sopivaa numeerista menetelmää, sillä usein yhtälölle ei löydy suljetun muodon ratkaisua [Agresti 2019]. Newton-Raphsonin menetelmällä yhtälöt voidaan kuitenkin ratkaista iteratiivisesti funktion maksimikohdan löytämiseksi. Menetelmä perustuu Taylorin sarjakehitelmään, jossa logaritmoitua uskottavuusfunktiota approksimoidaan jokaisessa iteraation t = 1, 2, ... vaiheessa seuraavasti:

$$L(\boldsymbol{\beta})\approx L(\boldsymbol{\beta}^t)+\boldsymbol{u}^{(t)T}(\boldsymbol{\beta}-\boldsymbol{\beta}^t)+\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}^t)^T\boldsymbol{H}^t(\boldsymbol{\beta}-\boldsymbol{\beta}^t) \tag{2.20}$$

missä

$$\boldsymbol{u}=\left(\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_1},\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_2},\dots,\frac{\partial L(\boldsymbol{\beta})}{\partial\beta_p}\right)^T \quad \text{ja} \quad \boldsymbol{H}=\begin{pmatrix}\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_1} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_1\partial\beta_p}\\ \vdots & \ddots & \vdots\\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_p\partial\beta_1} & \cdots & \frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_p\partial\beta_p}\end{pmatrix}.$$

Matriisilla **H** tarkoitetaan tässä Hessen matriisia. Parametrien $\boldsymbol{\beta}$ ratkaisemiseksi asetetaan Taylorin sarjana muodostetun uskottavuusyhtälön osittaisderivaatat nollaksi, josta seuraa:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \boldsymbol{u}^{(t)} + \boldsymbol{H}^{t}(\boldsymbol{\beta} - \boldsymbol{\beta}^{t}) = \mathbf{0}, \tag{2.21}$$

ja lopuksi

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\boldsymbol{H}^{(t)}\right)^{-1}\boldsymbol{u}^{(t)}. \tag{2.22}$$

Iteraatiot jatkuvat, kunnes peräkkäisten kierrosten *t ja t+1* muutos funktion arvossa $L(\boldsymbol{\beta}^{t})$ on riittävän pieni. Iteratiivisesti ratkaistun parametrivektorin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori $\widehat{\boldsymbol{\beta}}$ on asymptoottisesti harhaton. Lisäksi estimaattori noudattaa multinormaalijakaumaa

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}\{\boldsymbol{\beta}, (\boldsymbol{X}^{T}\boldsymbol{W}\boldsymbol{X})^{-1}\}, \tag{2.23}$$

missä **W** on diagonaalimatriisi, alkioilla $w_i = \frac{\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2}{var(y_i)}$.

2.5 Newton-Raphsonin algoritmin toteutus Poisson-regressiolle

Osoitetaan seuraavaksi, kuinka Newton-Raphsonin menetelmä estimoi mallin parametrit. Menetelmäkuvaus ja liitteessä 1 esitetty R-koodi noudattelee Liun [2023] toteutusta. Kun oletetaan vastemuuttujan noudattavan Poissonin jakaumaa, pistetodennäköisyysfunktio yhdelle havainnolle on muotoa:

$$f(y_i; \mu_i) = \frac{e^{-\mu_i}\mu_i^{\,y_i}}{y_i!}. \tag{2.24}$$

Kun havaintoja on n-kappaletta, voidaan logaritmoitu uskottavuusfunktio esittää seuraavasti:

$$L(\boldsymbol{\mu}; y_1, y_2, \dots, y_n) = -\sum_{i=1}^{n} \mu_i + \sum_{i=1}^{n} y_i \log(\mu_i) - \sum_{i=1}^{n} \log(y_i)\,!\,. \tag{2.25}$$

Seuraavaksi valitaan linkkifunktio, jolla yhdistetään mallin lineaarinen ennustin odotusarvoon $\boldsymbol{\mu}$. Poisson-regressiossa tavallinen valinta on log-linkki, jolloin havainnon i odotusarvo on

$$\mu_i = e^{\boldsymbol{x_i}^T \boldsymbol{\beta}}. \tag{2.26}$$

Yhtälössä (2.26) i viittaa havaintoon ja $\boldsymbol{x_i}$ on vektori alkioilla $x_{i0}, \dots, x_{ip}$. Yhdistämällä yhtälöt (2.25) ja (2.26) saadaan mallin lopullinen logaritmoitu uskottavuusfunktio

$$L(\boldsymbol{\mu}; y_1, y_2, \dots, y_n) = -\sum_{i=1}^{n} e^{\boldsymbol{x_i}^T \boldsymbol{\beta}} + \sum_{i=1}^{n} y_i \boldsymbol{x_i}^T \boldsymbol{\beta} - \sum_{i=1}^{n} \log(y_i)\,!\,. \tag{2.27}$$

Algoritmin toteuttamiseksi määritellään funktion ensimmäinen ja toinen derivaatta parametrivektorin $\boldsymbol{\beta}$ suhteen. Nämä voidaan esittää matriisimuodossa:

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \boldsymbol{X}^T\left(\boldsymbol{y} - e^{\boldsymbol{X\beta}}\right) \quad ja \quad \frac{\partial^2 L}{\partial^2 \boldsymbol{\beta}} = -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \tag{2.28}$$

missä **W** on

$$\begin{pmatrix} e^{\boldsymbol{x_1}^T \boldsymbol{\beta}} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\boldsymbol{x_n}^T \boldsymbol{\beta}} \end{pmatrix}. \tag{2.29}$$

Yhtälössä (2.22) esitetty parametrivektorin $\boldsymbol{\beta}$ ratkaisu iteratiivisen algoritmin avulla vaatii ensimmäiseen vaiheeseen alkuarvauksen parametrien arvoista. Hardin ja Hilbe [2001] ehdottavat ratkaisuksi vakiotermimallin käyttöä. Poisson-regression tapauksessa suurim-

man uskottavuuden estimaatti vakiotermille β_0 on logaritmi vastemuuttujan Y keskiarvosta. Nyt ensimmäisenä arvauksena voidaan käyttää vektoria $\boldsymbol{\beta} = (\beta_0, 0, \dots, 0)^T$. Algoritmi tarvitsee myös tiedon siitä, että kuinka suuri erotus parametrien arvoissa peräkkäisissä vaiheissa *t* ja *t+1* sallitaan. Liitteessä 1 toteutettu algoritmi loppuu, kun erotus on alle 0.00001. Taulukossa 2 havainnollistetaan parametrien muodostuminen eräälle testijoukolle.

Taulukko 2: Parametrien arvot jokaisessa iteraation t vaiheessa.

Algoritmin vaihe *t*	$\boldsymbol{\beta_0}$	$\boldsymbol{\beta_1}$
Alustus	-2.288708	0
1	-0.68560458	-0.03131349
2	-0.6021847	-0.350326
3	-0.58824920	-0.03545305
4	-0.58814531	-0.3545593
5	-0.58814531	-0.3545593

3 Mallinnusprosessi yleistetyillä lineaarisilla malleilla

Tässä luvussa muodostetaan aidolle esimerkkiaineistolle malli sekä vahinkotiheydelle että keskivahingolle. Keskivahingon mallinnuksessa testiaineistosta poimitaan vain ne havainnot, joille on maksettu korvausta. Käytettävä aineisto on vuodelta 2021 ja sisältää 36 765 havaintoa vahinkovakuutusyhtiön kotivakuutuksen asiakkaista.

Aineisto jaetaan satunnaisesti kahteen osaan niin, että 70 prosenttia havainnoista otetaan mallien opetusaineistoksi ja loput 30 prosenttia jätetään valittujen mallien testaamiseksi. Mallinnus aloitetaan selittävien muuttujien kuvailevalla analyysilla, jolloin voidaan muodostaa käsitys niiden marginaalisesta vaikutuksesta vastemuuttujaan [Agresti 2015]. Kuvaileva analyysi tehdään koko havaintojoukolle.

Yleistetyn lineaarisen mallin valinnassa pätee samat periaatteet kuin tavallisessa lineaarisessa regressiossa. Tavoitteena on valita malli, joka riittävän hyvin selittää tarkastelun kohteena olevaa ilmiötä tekemättä kuitenkaan mallista tarpeettoman monimutkaista ja vaikeasti tulkittavaa. Selittävien muuttujien valinta ei myöskään välttämättä ole aina puhtaasti tilastotieteellinen [de Jong & Heller 2008]. Tilastollisesti merkitsevä muuttuja voidaan esimerkiksi tiputtaa pois mallista liiketoiminnallisesta syystä.

Lopullisen mallin valinnassa voidaan käyttää useita eri kriteereitä. Seuraavassa esitellään mallien vertailussa käytettyjä mittareita.

3.1 Akaiken informaatiokriteeri (AIC) ja keskineliövirhe (MSE)

Akaiken informaatiokriteeriä voidaan käyttää kilpailevien mallien vertailussa. Kriteeri pyrkii hakemaan tasapainon mallin uskottavuuden ja estimoitujen parametrien lukumäärän välille. Uuden muuttujan lisääminen malliin käytännössä aina parantaa mallin uskottavuutta, mutta mallin tulisi silti sisältää vain merkityksellisimmät muuttujat. Kriteeriä tulkitaan niin, että vertaillessa malleja keskenään, pienempi on parempi. Akaike [1974] määrittelee kriteerin seuraavasti:

$$AIC = -2\log(L) + 2k. \tag{3.1}$$

Edellä esitetyssä kriteerissä L on havainnoista estimoitu uskottavuusfunktion arvo ja k mallin parametrien määrä. Usein mallien ennustevoiman arvioinnissa käytössä taas on

keskineliövirhe (*Mean Squared Error, MSE*). Mittarin tarkoituksena on arvioida, kuinka mallin tuottamat ennusteet poikkeavat havaituista arvoista. Pienempi keskineliövirhe on parempi vertaillessa vaihtoehtoisia malleja. Keskineliövirhe määritellään seuraavasti:

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}, \tag{3.2}$$

missä y_i on havaittu arvo, $\hat{y}_i$ on mallin tuottama ennustus ja n on havaintojen lukumäärä.

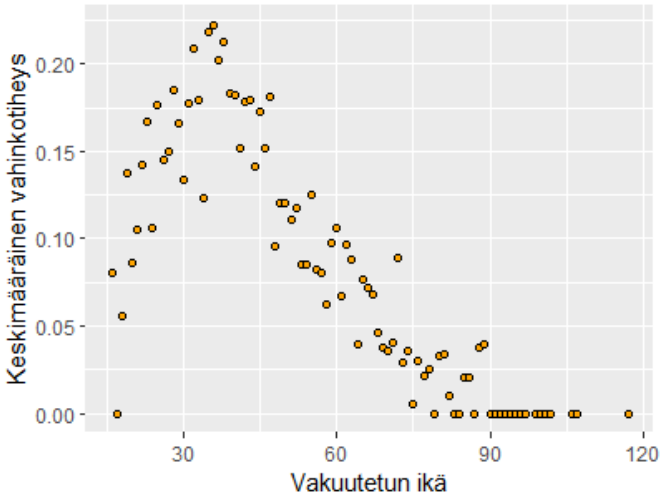
3.2 Kuvaileva analyysi vahinkotiheydestä

Mahdollisesti malliin päätyviä selittäviä muuttujia on yhteensä seitsemän kappaletta, joista yksi on vakuutuksenottajan ikä. Muiden kuuden muuttujan nimet on piilotettu tästä esityksestä. Muuttujien luokkien määrät ovat listattuna taulukossa 3.

Taulukko 3: Mahdollisesti selittävät muuttujat

Muuttuja	**Luokkien lukumäärä**
Ikä	jatkuva muuttuja
Muuttuja 1	5
Muuttuja 2	3
Muuttuja 3	2
Muuttuja 4	3
Muuttuja 5	2
Muuttuja 6	2

Jatkuvan muuttujan vaikutusta jatkuvaan vastemuuttujaan, tässä tapauksessa vahinkotiheyteen, on luontevinta tutkia hajontakuviolla [de Jong & Heller 2008]. Kuvassa 1 on esitetty hajontakuvio vakuutuksenottajan iästä ja vahinkotiheydestä.



Kuva 1: Hajontakuvio keskimääräisestä vahinkotiheydestä vakuutuksenottajan iän mukaan

Vahinkotiheyden tarkastelu vakuutuksenottajan iän mukaan paljastaa muuttujan epälineaarisen luonteen. Mallin valinnassa on siis syytä tarkastella ikätekijän tuomista malliin esimerkiksi toisen tai kolmannen asteen polynomina. Suoritetaan myös marginaalinen tarkastelu muista muuttujista. Vahinkotiheydet on esitetty taulukoissa 4, 5 ja 6.

Taulukko 4: Muuttujan 1 vahinkotiheys tasoilla 1–5

Muuttujan (1) taso	**Keskimääräinen vahinkotiheys**
1	0.08684
2	0.10856
3	0.16522
4	0.19682
5	0.30707

Muuttujan 1 tasoilla on luontainen järjestys. Vahinkotiheydessä on havaittavissa selkeä kasvava trendi, kun siirrytään ylöspäin tasolta toiselle. Tasolla 5 vahinkotiheys on kolminkertainen tasoon 1 nähden.

Taulukko 5: Muuttujien 2 ja 4 vahinkotiheydet tasoilla 1–3

Muuttujan (2 & 4) taso	Keskimääräinen vahinkotiheys (muuttuja 2)	Keskimääräinen vahinkotiheys (muuttuja 4)
1	0.12845	0.10551
2	0.12463	0.13068
3	0.11601	0.16357

Muuttujan 2 tasojen välillä vahinkotiheys ei vaikuta olevan merkittävästi erilainen. Muuttujan 4 tasoilla taas vahinkotiheys kasvaa siirryttäessä tasolta 1 tasolle 3 yli 50 prosenttia.

Taulukko 6: Muuttujien 3, 5 ja 6 vahinkotiheydet tasoilla 1 ja 2

Muuttujan (3, 5 & 6) taso	Keskimääräinen vahinkotiheys (muuttuja 3)	Keskimääräinen vahinkotiheys (muuttuja 5)	Keskimääräinen vahinkotiheys (muuttuja 6)
1	0.10915	0.12796	0.11273
2	0.31024	0.11609	0.13206

Muuttujan 3 tasojen ero vahinkotiheyksissä on huomattavan suuri, kun taas muuttujien 5 ja 6 tasoilla erot ovat vähemmän merkittävä.

3.3 Vahinkotiheyden mallintaminen

Edellä esitettyjen marginaalitarkastelujen avulla voidaan luoda käsitys muuttujien vaikutuksesta vahinkotiheyteen. Seuraava vaihe, yleistetyn lineaarisen mallin luominen, aloitetaan valitsemalla vastemuuttujalle sopiva jakauma ja linkkifunktio. Vahinkojen lukumäärää mallinnettaessa käytetään usein Poisson-jakaumaa ja log-linkkiä.

Vahinkotiheyttä mallinnettaessa on tärkeää ottaa huomioon myös riskille altistumisen aika jokaisella havaintoyksiköllä. Tämä tehdään ottamalla malliin mukaan offset-termi, joka kertoo kuinka kauan asiakkaan vakuutus on ollut voimassa tilastovuonna. Termi muodostetaan jakamalla voimassaolopäivien määrä koko tilastovuoden päivillä. Koko vuoden voimassa ollut vakuutus saa siis arvokseen 1.

Malli voidaan nyt esittää muodossa

$$log(\mu) = \log(exposure) + \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \,, \tag{3.3}$$

missä *exposure* on edellä mainittu offset-termi.

Malli voidaan muodostaa käyttämällä esimerkiksi etenevää valintaa, jolloin ensimmäinen malli sisältää pelkästään vakiotermin. Seuraavassa vaiheessa malliin lisätään selittäviä muuttujia niin, että eniten mallia parantava muuttuja otetaan mukaan. Tämä jatkuu, kunnes seuraava lisättävä muuttuja ei enää paranna mallia. Kriteerinä voidaan käyttää esimerkiksi AIC-informaatiokriteerin arvoa.

Vakiotermimallin AIC-arvo opetusaineistolle on 15 993. Taulukossa 7 on esitetty ensimmäisen vaiheen mukaiset AIC-arvot, kun malliin otetaan mukaan vakiotermin lisäksi kyseinen muuttuja.

Taulukko 7: AIC-arvot kun vakiotermin lisäksi mallissa mukana kyseinen muuttuja

Muuttuja	**AIC**
Ikä	15 744
Muuttuja 1	15 530
Muuttuja 2	15 995
Muuttuja 3	15 664
Muuttuja 4	15 920
Muuttuja 5	15 991
Muuttuja 6	15 976

Muuttujan 1 kohdalla AIC-arvo on pienin, jolloin se otetaan malliin mukaan. Näin malliin voidaan lisätä selittäviä muuttujia vaihe vaiheelta, kunnes AIC-arvo ei enää pienene. Muuttujien valinnassa päädytään opetusaineistolla vaiheeseen, jossa malliin on valittu järjestyksessä mukaan muuttuja 1, ikä, muuttuja 3, muuttuja 2, muuttuja 4 ja muuttuja 6.

Muuttujan 5 lisääminen malliin ei enää paranna AIC-arvoa, jolloin se ei päädy mukaan selittäväksi tekijäksi.

Vahinkotiheyden kuvailevassa analyysissa havaittiin ikätekijän epälineaarinen vaikutus. Seuraavaksi testataan paraneeko malli, kun ikä tuodaan malliin toisen tai kolmannen asteen polynomina. Tuotaessa polynominen termi mukaan malliin, on tapana sisällyttää myös alemman asteen termi mukana mallissa. Tässä vaiheessa olevan mallin AIC-arvo on 14 918. Toisen asteen polynomi ikätekijällä tiputtaa arvon lukemaan 14 911. Kolmannen asteen polynomi taas asettaa AIC-arvon lukemaan 14 913. Näin voidaan tulkita, että toisen asteen malli on sopivampi valinta.

Muuttuja 6 on vakuutetun vakuutusturvan sisältöön liittyvä muuttuja. Kuten aikaisemmin todettiin, voidaan muuttuja sisällyttää tai olla sisällyttämättä malliin myös muusta kuin tilastotieteellisestä syystä. Tässä tehdään valinta olla ottamatta kyseistä muuttujaa mukaan, vaikkakin sen sisällyttäminen parantaa AIC-arvoa. Lopullisen mallin parametrit ja keskihajonnat on esitetty taulukossa 8.

Taulukko 8: Vahinkotiheysmallin parametrit, estimaatit ja otoskeskivirheet

Parametrin nimi	**Parametri**	**Estimaatti**	**Otoskeskivirhe**
Vakiotermi	β_0	-2.7330536	0.2040264
Muuttuja 1: taso 2	β_1	0.2816157	0.0599515
Muuttuja 1: taso 3	β_2	0.4440590	0.0726048
Muuttuja 1: taso 4	β_3	0.5095122	0.0714926
Muuttuja 1: taso 5	β_4	0.9568031	0.0721357
Ikä	β_5	0.0360432	0.0100876
Ikä^2	β_6	-0.0006644	0.0001109
Muuttuja 3: taso 2	β_7	1.1544213	0.0621106
Muuttuja 2: taso 2	β_8	-0.3522023	0.0607297
Muuttuja 2: taso 3	β_9	-0.6773352	0.0801654
Muuttuja 4: taso 2	β_{10}	0.1629740	0.0532475
Muuttuja 4: taso 3	β_{11}	0.3706764	0.0625396

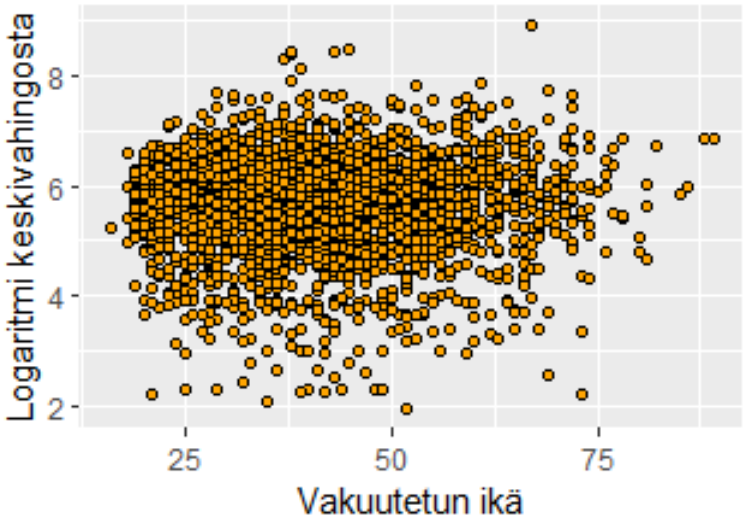
Ohjelmisto ei estimoi parametria muuttujien perustasoille, jotka tässä on valittu olemaan muuttujien ensimmäiset tasot. Esitetyistä parametrien arvoista on luontevaa ottaa eksponentti, jolloin kunkin muuttujan vaikutus voidaan esittää kertoimena. Vahinkotiheys voidaan nyt ennustaa esimerkiksi havaintoyksikölle, jonka muuttuja 1 on tasolla 3, ikä vuosissa on 50, muuttuja 3 on tasolla 1, muuttuja 2 tasolla 3 ja muuttuja 4 tasolla 3:

$$\hat{\mu} = \exp\left(\widehat{\beta_0} + \cdots + \widehat{\beta_{11}} x_{11}\right) = \exp(-2.7330536 + \cdots + 0.3706764) \approx 0.086$$

3.4 Kuvaileva analyysi keskivahingosta

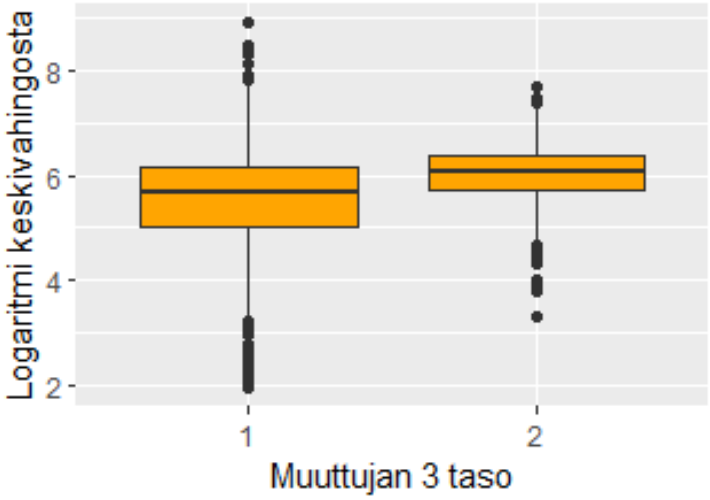
Keskivahingon kuvaileva analyysi voidaan suorittaa samaan tapaan kuin vahinkotiheydenkin osalta. Mahdollisina selittävinä muuttujina on edelleen taulukossa 3 esitetyt muuttujat, pois lukien muuttuja 2.

Vakuutetun iästä ja keskivahingosta muodostetaan jälleen hajontakuvio. Tässä tarkastelussa tehdään kuvaaja niistä havainnoista, joille on sattunut vahinko. Euromääräisen keskivahingon sijasta kuvataan y-akselilla logaritmi keskivahingosta, jolloin muuttujan vaikutuksen arviointi selkiytyy [de Jong & Heller 2008].



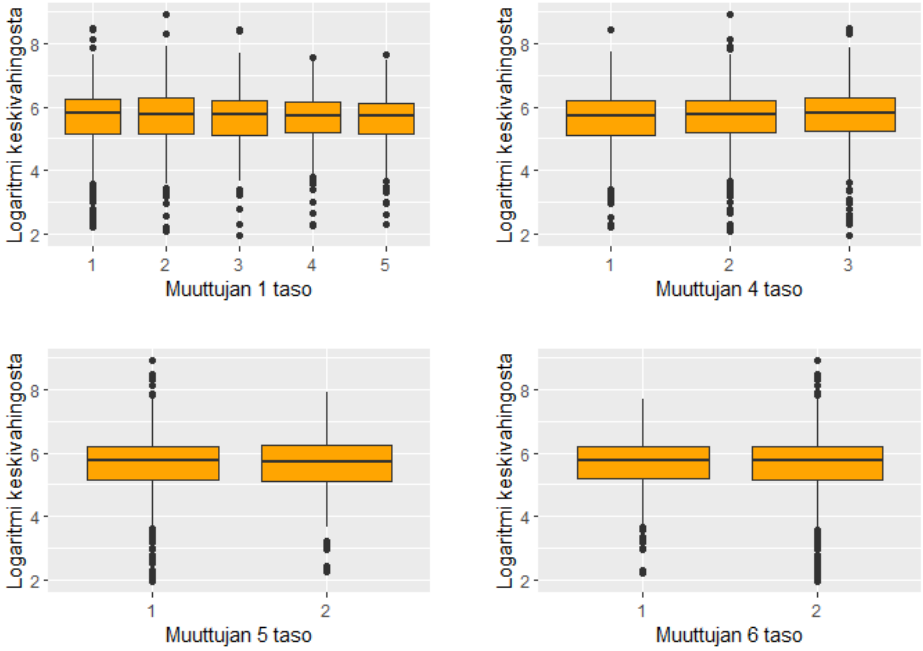
Kuva 2: Hajontakuvio keskivahingosta vakuutuksenottajan iän mukaan

Hajontakuvio iästä ja keskivahingon logaritmista vaikuttaa paljastavan sen, että vakuutetun iällä ei näytä olevan isoa merkitystä vahingon laajuuteen. Jatketaan tarkastelua, tekemällä kategorisille kandidaattimuuttujille laatikkokuviot keskivahingosta.



Kuva 3: Laatikkokuvaaja keskivahingosta muuttujan 3 tasojen mukaan

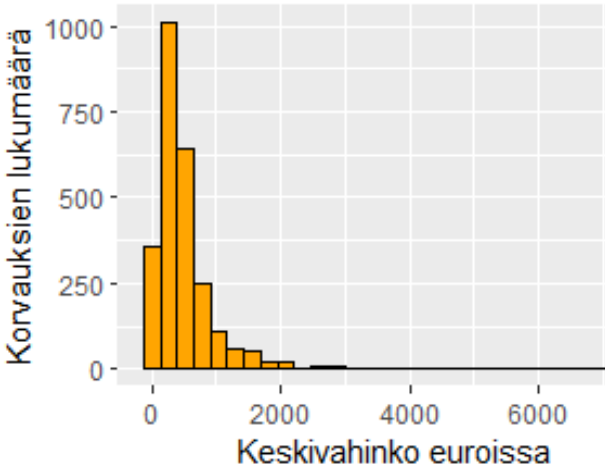
Muuttujan 3 laatikkokuvaaja näyttää erot keskivahinkojen suuruudessa tasoilla 1 ja 2. Muuttujilla 1, 4, 5 ja 6 erot tasojen välillä taas vaikuttavat kuvaajien perusteella lähes merkityksettömiltä.



Kuva 4: Laatikkokuvaajat muuttujista 1, 4, 5 ja 6

3.5 Keskivahingon mallintaminen

Keskivahinkoa mallinnetaan tavallisesti käyttäen gamma-jakaumaa ja log-linkkiä. Tällainen valinta voidaan perustella sillä, että keskivahinko on muuttujana ei-negatiivinen ja voimakkaasti oikealle vino. Näin havaitaan myös kuvassa 5, jossa on tarkasteltu eri suuruisten keskivahinkojen lukumäärää.



Kuva 5: Keskivahinkojen jakauma aineistossa

Myös keskivahingolle mallia valittaessa voidaan käyttää etenevää valintaa. Ensimmäisen mallin AIC-arvo on 31 456, kun mallissa on mukana vain vakiotermi. Malliin lisätään jälleen muuttujia, kunnes AIC-arvo ei uuden muuttujan lisäämisellä pienene. Keskivahinkomalliin valikoituneet muuttujat, parametrien estimaatit ja näiden keskivirheet ovat listattuna taulukossa 9.

Taulukko 9: Keskivahinkomallin parametrit, estimaatit ja otoskeskivirheet

Parametrin nimi	Parametri	Estimaatti	Otoskeskivirhe
Vakiotermi	β_0	5.86037	0.04803
Muuttuja 3: taso 2	β_1	0.27675	0.05684
Muuttuja 4: taso 2	β_2	0.09923	0.05513
Muuttuja 4: taso 3	β_3	0.17285	0.06382
Muuttuja 1: taso 2	β_4	0.06315	0.06350
Muuttuja 1: taso 3	β_5	-0.09630	0.07634
Muuttuja 1: taso 4	β_6	-0.11653	0.07388
Muuttuja 1: taso 5	β_7	-0.21831	0.07439

Muuttujasta 1 huomataan, että estimoitujen parametrien keskivirheet ovat kohtuullisen suuria suhteessa kertoimiin. Tehdään keskivahingolle kilpaileva malli, jossa muuttuja 1 ei ole selittävänä tekijänä. Kilpailevan mallin parametrien estimaatit ja keskivirheet ovat esitetty taulukossa 10.

Taulukko 10: Vaihtoehtoisen keskivahinkomallin parametrit, estimaatit ja otoskeskivirheet

Parametrin nimi	Parametri	Estimaatti	Otoskeskivirhe
Vakiotermi	β_0	5.84233	0.04235
Muuttuja 3: taso 2	β_1	0.24948	0.05810
Muuttuja 4: taso 2	β_2	0.07634	0.05471
Muuttuja 4: taso 3	β_3	0.11514	0.06157

Vaihtoehtoisessa mallissa muuttujan 4 tason 2 keskivirhe on myös suuri verrattuna parametrin kertoimeen. Tehdään kuitenkin valinta pitää se mallissa mukana, ja siirrytään vertailemaan malleja keskenään. Taulukossa 9 esitetyn, enemmän selittäviä tekijöitä sisältävän, mallin AIC-arvo on 31 388. Vaihtoehtoisen mallin AIC-arvo puolestaan on 31 415. Pelkästään AIC-arvon perusteella valinta kohdistuisi siis ensimmäisenä luotuun malliin. Testataan ennen lopullista valintaa, kuinka kumpikin malli onnistuu ennustamaan testiaineistossa havaintojen keskivahinkoja.

Testiaineistossa ensimmäisenä luodun mallin keskineliövirhe on 119 625. Toisena esitetyn yksinkertaisemman mallin keskineliövirhe on puolestaan 113 420. Opetusaineistoilla luoduista malleista yksinkertaisempi vaihtoehto ennustaa siis paremmin keskivahinkoa testiaineistossa, kun mittarina käytetään keskineliövirhettä.

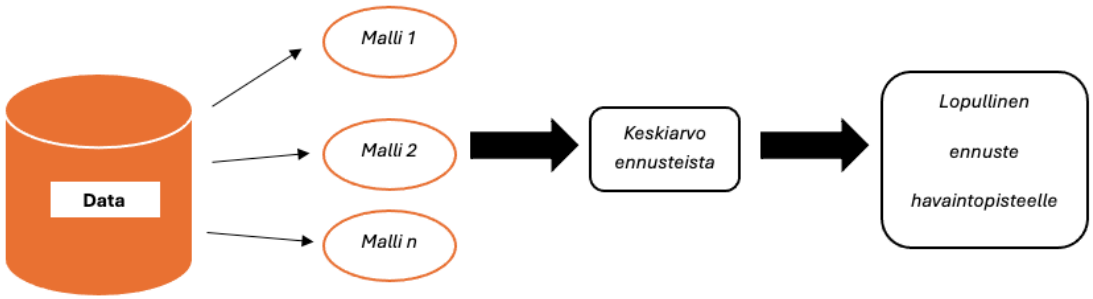
Aikaisemmassa vaiheessa todettiin, että mallin valinnassa tulee painottaa yksinkertaisempaa vaihtoehtoa. Testiaineistolla tehty tarkastelu mallin ennustuskyvystä vahvistaa myös taulukossa 10 esitetyn mallin valitsemista. Valitaan yksinkertaisempi vaihtoehto siis lopulliseksi keskivahinkomalliksi.

4 Koosteoppiminen ennusteen luomisessa

Luvussa 3 luotiin sekä vahinkotiheydelle että keskivahingolle mallit hyödyntäen yleistettyjä lineaarisia malleja. Kyseisen malliperheen etuna suhteessa muihin metodeihin on erityisesti tulosten tulkitsemisen helppous. Muuttujien (hinnoittelutekijöiden) väliset suhteet ja erot on helppo esittää esimerkiksi vahinkotiheyttä laskevana tai nostavana kertoimena. Yleistettyjen lineaaristen mallien käyttö vakuutusmaksun muodostamiseksi onkin ollut alan standardi monessa maassa jo 1990-luvulta alkaen [Ohlsson & Johansson 2010].

Tilastollisen mallin ennusteen tarkkuutta voidaan usein kuitenkin parantaa hyödyntämällä erilaisia koneoppimismenetelmiä. Käytetyistä menetelmistä mainittakoon esimerkiksi satunnaismetsät (random forest), neuroverkot ja tehostaminen (boosting). Vaikkakin muun muassa edellä mainitut menetelmät usein onnistuvat nostamaan ennustetarkkuutta, koneoppimismenetelmien hyödyntäminen laajemmassa mittakaavassa saattaa törmätä tulosten tulkitsemisen haasteeseen. Kyseessä on niin sanottu mustan laatikon (black box) ongelma, jolloin esimerkiksi tietyn ennusteen saaminen käytetyllä koneoppimismenetelmällä on jopa mahdotonta selittää.

Koneoppimista voidaan toteuttaa myös niin sanottuna koosteoppimisena (ensemble learning). Koosteoppimisen perusideana on luoda opetusaineistosta useita malleja, joiden tuloksia yhdistelemällä saavutetaan havaintopisteen ennuste [Zhou 2012]. Yksittäisiä malleja yhdistelemällä päästään usein parempaan lopputulokseen kuin mitä yksi malli itsessään onnistuu ennustamaan. Luokitteluongelmissa koosteoppimisen lopputulemaan päästään esimerkiksi käyttämällä enemmistöäänestystä yksittäisten mallien pohjalta: havaintopisteen ennustettu luokka on se, joka ennustetaan yksittäisissä malleissa kaikista useimmiten. Regressiossa ennuste taas voi olla jatkuva, jolloin lopullinen ennuste koosteoppimismenetelmällä hyödyntää usein keskiarvoa yksittäisistä malleista (kuva 6).



Kuva 6: Koosteoppimisen perusidea

Alaluvussa 4.1 toteutetaan yksi edellä kuvatun koosteoppimisen menetelmä, laukutus (bagging, bootstrap aggregation), vahinkotiheyden mallintamiseksi. Menetelmä toteutetaan samaan opetusaineistoon kuin luvussa 3 luotu yleistetty lineaarinen malli. Alaluvussa 4.5 vertaillaan näiden menetelmien eroja ennustetarkkuudessa.

4.1 Menetelmän kuvaus

Toteutettu menetelmä toimii pääasiassa samalla tavalla kuin Songin et al. [2013] luoma R-ohjelmiston paketti *randomGLM*. Toteutus on tässä tutkielmassa kuitenkin muokattu yksinkertaisemmaksi selittävien muuttujien valinnan osalta, koska mahdollisten muuttujien joukko satunnaisotoksessa on maksimissaan 7. Näin ollen voi olla perusteltua käyttää kaikkia mahdollisia muuttujia, eikä aina rajata joukkoa pienemmäksi ennen mallin rakentamista. Valmis ohjelmakirjasto, randomGLM, ei myöskään tarjoa yksiselitteisesti mahdollisuutta käyttää kaavan 3.3 mukaista offset-termiä mallissa. Algoritmin toteutus on esitetty liitteessä 2.

Menetelmä koostuu neljästä osasta:

1. Bootstrap-otosten luominen takaisinotannalla
2. Mallin luominen otokselle etenevää valintaa hyödyntäen
3. Ennusteen luominen yksittäisillä malleilla
4. Keskiarvon ottaminen yksittäisistä ennusteista lopullisen ennusteen tekemiseksi

4.2 Bootstrap

Ensimmäisessä vaiheessa poimitaan opetusaineistosta bootstrap-otos, josta yksittäinen malli luodaan myöhemmässä vaiheessa. Opetusaineistosta, jonka koko on *n,* poimitaan n-kappaletta havaintoja takaisinpalautuksella. Näin yksittäiseen otokseen voi tulla sama havainto useamman kerran ja toisaalta myös osa havainnoista ei tule poimituksi kertaakaan. Tällaisia otoksia muodostetaan *m*-kappaletta. Otosten (laukkujen) määrä m on usein vähintään 100 kappaletta. Tässä toteutuksessa laukkujen määräksi on valittu 100.

4.3 Etenevä valinta – stepAIC

Toisessa vaiheessa satunnaisotokselle luodaan yleistetty lineaarinen malli lähes samaan tapaan kuin alaluvussa 3.3. Toteutetussa menetelmässä ikää käsitellään kuitenkin kategorisena muuttujana niin, että ikävuosista 18–80 muodostetaan luokat kahdelle peräkkäiselle ikävuodelle kerrallaan. Esimerkiksi ikävuodet 18 ja 19 kuuluvat samaan luokkaan ja vuodet 20 ja 21 omaansa. Kaikki yli 80-vuotiaat vakuutetut kerätään omaan luokkaansa.

Yksittäisen mallin mahdollisten selittävien tekijöiden lukumäärää rajoitetaan myös satunnaisesti. Taulukko 3 listasi kaikki mahdolliset vahinkotiheyttä selittävät tekijät, joita oli yhteensä seitsemän kappaletta. Tässä toteutuksessa tehdään ohjelmistokirjaston randomGLM tapaan rajoitus siihen, että kuinka montaa selittävää tekijää yksittäinen malli voi käyttää. Jokaiselle bootstrap-otokselle valitaan satunnaisesti joko maksimissaan 5, 6 tai 7 ehdokasta selittäväksi tekijäksi. Vähäisen selittäjien lukumäärän vuoksi painotetaan kuitenkin mahdollisuutta käyttää kaikkia seitsemää tekijää.

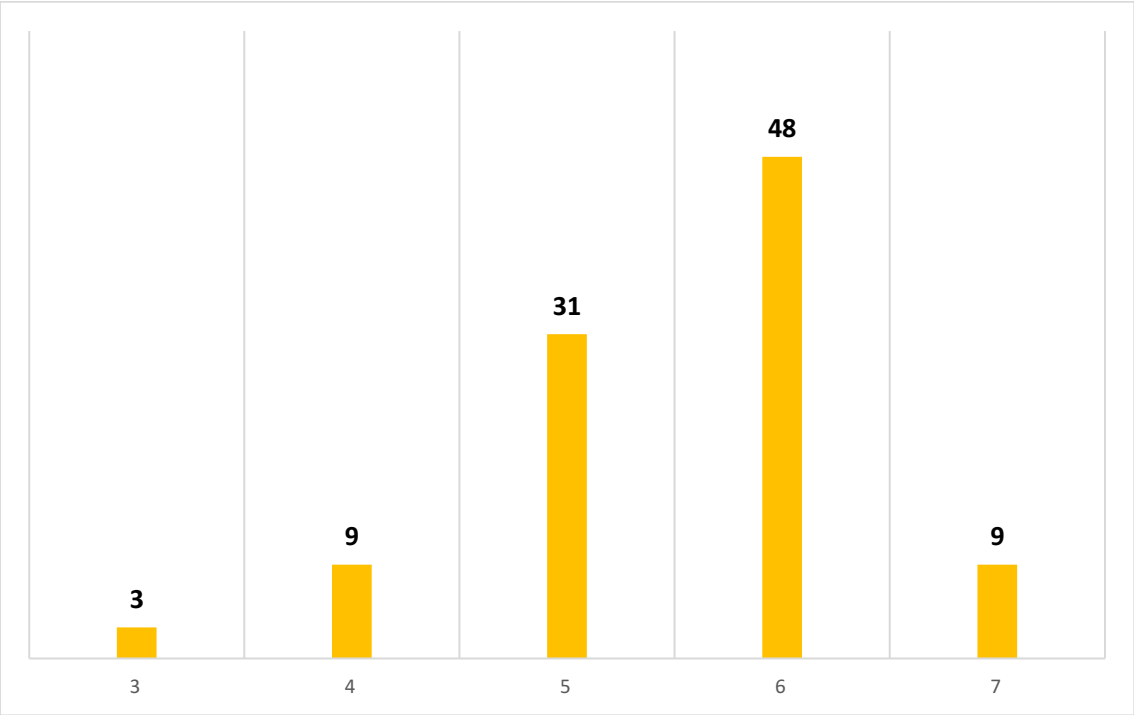
Sadalle otokselle on kaikille tässä vaiheessa valittu maksimäärä mahdollisia selittäviä muuttujia. Otoksista 13 voi käyttää maksimissaan viittä muuttujaa, 16 maksimissaan kuutta muuttujaa ja 71 otosta saa seuraavaan vaiheeseen kaikki 7 muuttujaa.

Seuraavaksi rakennetaan jokaiselle otokselle Poissonin log-lineaarinen malli käyttäen etenevää valintaa ja edellisen vaiheen mahdollisia selittäviä muuttujia. Etenevä valinta aloittaa vakiotermimallista ja lisää jokaisella askeleella ehdolla olevia selittäviä muuttujia niin kauan kunnes AIC-arvo ei enää parane. Kaikki ehdolla olevat muuttujat eivät siis

välttämättä päädy yksittäisen mallin selittäväksi tekijäksi. Valinta on täysin automaattinen, ja siihen voidaan käyttää ohjelmakirjaston *MASS* funktiota stepAIC. Taulukko 11 näyttää havainnollistavasti, kuinka eri otoksille valikoituu ehdokkaita malliin ja kuinka monta muuttujaa etenevä valinta lopulta valitsee. Kuvassa 7 on lopullisten muuttujien lukumäärän jakauma.

Taulukko 11: Laukkujen otosten koko, ehdokasmuuttujien lukumäärä ja lopullinen selittävien muuttujien lukumäärä

Otoksen numero	**Otoksen koko**	**Ehdolla olevien muuttujien lukumäärä**	**Selittävien muuttujien lopullinen lukumäärä**
1	25 809	7	6
2	25 809	5	5
3	25 809	5	3
.	25 809	.	.
.	25 809	.	.
.	25 809	.	.
100	25 809	7	7



Kuva 7: Mallien selittävien muuttujien lukumäärän jakauma

4.4 Ennusteet yksittäisillä malleilla ja koostaminen

Jokaisella alaluvussa 4.3 luodulla mallilla voidaan nyt tuottaa testiaineistoon havaintoyksikölle ennuste vahinkotiheydestä. Osa malleista siis käyttää kaikkia mahdollisia selittäviä muuttujia ennusteen tekemiseksi, ja toisaalta esimerkiksi yhdeksän mallia on päätynyt käyttämään vain neljää selittävää muuttujaa. Havaintoyksikkö saa nyt yhteensä 100 kappaletta ennusteita vahinkotiheydelleen. Kaikista ennusteista otetaan yksinkertaisesti keskiarvo lopullisen ennusteen koostamiseksi.

4.5 Koosteoppimismallin vertailu perinteiseen vahinkotiheysmalliin

Vertaillaan lopuksi taulukossa 8 esitetyn vahinkotiheysmallin ja koosteoppimismenetelmän eroja ennusteen tarkkuudessa. Molempien mallinnuksessa käytetyn opetusaineiston lisäksi saatavilla on testiaineisto, jonka havaintojen osalta mallien ennustetarkkuuksia ja toteumaa vertaillaan. Tarkkuuden mittarina käytetään keskineliövirhettä.

Ennen molempien vaihtoehtoisten mallien arviointia tuotetaan keskivahingolle ennusteet myös kaikista yksinkertaisimman mallin, vakiotermimallin avulla. Muita malleja arvioitaessa voidaan vakiotermimallia pitää eräänlaisena lähtötasona, jonka tarkkuutta edellä kehitetyt mallit pyrkivät parantamaan. Vakiotermimallissa ei ole siis yhtäkään selittävää tekijää, vaan kaikki havainnot saavat vahinkotiheyden ennusteeksi saman arvon. Tarkkuuden arvioinnissa otetaan kuitenkin edelleen huomioon vakuutusturvan voimassaolo offset-termillä. Koko vuoden voimassa olleen havainnon ennustettu vahinkotiheys on 0.125392. Testiaineiston keskineliövirhe vakiotermimallilla on 0.1373.

Taulukon 8 mukaisen yleistetyn lineaarisen mallin keskineliövirhe on testiaineistolla 0.13309. Vakiotermimalliin nähden mallin keskineliövirhe tippuu noin 3.1 prosenttia. Koosteoppimismenetelmällä saatu ennustus laskee keskineliövirheen arvoon 0.13311.

Koosteoppimismenetelmällä tuotettu ennuste ei tässä tapauksessa paranna ennustetarkkuutta. Laukutusalgoritmit tehokkuus syntyy usein siitä, että yksittäiset mallit ovat riittävän epästabiileja [Zhou 2012]. Songin et al. ohjelmakirjasto *randomGLM* pyrkii luomaan yksittäisille malleille myös vaihtelua ennusteeseen lisäämällä satunnaisuutta siihen,

kuinka monta muuttujaa annetaan etenevän mallinvalinnan syötteeksi. Tässä tutkimuksessa käytetty aineisto sisältää vain seitsemän mahdollista muuttujaa, joten yksittäisten mallien ennusteet ovat todennäköisesti keskenään liian samankaltaisia halutun vaihtelun aikaansaamiseksi. Laukutusalgoritmin ennustetehokkuutta voitaisiin kuitenkin pyrkiä parantamaan esimerkiksi kasvattamalla bootstrap-otoksen kokoa, laukkujen lukumäärää sekä lisäämällä muuttujille yhdysvaikutuksia.

Jälleen voidaan tehdä valinta ottamalla yksinkertaisempi ja helpommin tulkittava malli lopulliseksi vahinkotiheysmalliksi. Valitaan vahinkotiheysmalliksi siis alaluvussa 3.3 esitetty malli. Jatkoon valittu keskivahinkomalli puolestaan on esitetty alaluvussa 3.5.

5 Riskimaksun muodostaminen

Luvussa 3 luotiin mallit sekä vahinkotiheydelle että keskivahingolle. Luodaan seuraavaksi riskimaksu testiaineistolle yhdistämällä mallit hinnoittelurakenteeksi. Ohlssonin ja Johanssonin [2010] mukaan tavallinen tapa on mallintaa vahinkotiheyttä ja keskivahinkoa erikseen ja toteuttaa hintarakenne yksinkertaisesti näiden kertolaskuna. Taulukossa 12 on esitetty rakenne nimenomaan mallien tulona. Mallien kertoimista on otettu ennen kertolaskua eksponentti, jolloin hinnoittelutekijöiden vaikutus tulee helpommin tulkittavaksi.

Taulukko 12: Hinnoittelurakenne

Parametri	Estimaatti
Vakio	22.4073
Muuttuja 1: taso 2	1.3253
Muuttuja 1: taso 3	1.5590
Muuttuja 1: taso 4	1.6645
Muuttuja 1: taso 5	2.6034
Ikä	1.0367
Ikä^2	0.9993
Muuttuja 3: taso 2	4.0711
Muuttuja 2: taso 2	0.7031
Muuttuja 2: taso 3	0.5080
Muuttuja 4: taso 2	1.2704
Muuttuja 4: taso 3	1.6255

Ikätekijän kerroin muuttuu epälineaarisesti, ja yksittäiselle ikäpisteelle kerroin saadaankin kaavan $1.0367^{Ik\ddot{a}} \cdot 0.9993^{Ik\ddot{a}^2}$ mukaisesti. Esitetään vielä, kuinka riskimaksu voidaan laskea halutulle havaintopisteelle. Esimerkkitapauksen ikä on 30 vuotta, muuttuja 1 on tasolla 1, muuttuja 2 tasolla 2, muuttuja 3 tasolla 1 ja muuttuja 4 tasolla 3. Ikätekijän kerroin on tällöin $1.0367^{30} \cdot 0.9993^{900} = 1.569971$. Riskimaksu on tässä tapauksessa:

$$22.4073 \cdot 1.569971 \cdot 1 \cdot 0.7031 \cdot 1 \cdot 1.6255 = 40.21€$$

5.1 Ennustevälin muodostaminen yksittäiselle riskimaksulle

Kertomalla vahinkotiheysmallin sekä keskivahinkomallin tuottamat piste-estimaatit keskenään päädytään havainnon riskimaksuun taulukon 12 mukaisesti. Piste-estimaattien lisäksi voidaan muodostaa havainnolle myös ennusteväli molempien mallien ennusteiden epävarmuuden tulkitsemiseksi.

Ennusteet riippuvat mallien omista estimoiduista parametrivektoreista $\widehat{\boldsymbol{\beta}}$ ja esimerkiksi Agresti [2015] osoittaa, kuinka $\widehat{\boldsymbol{\beta}}$ noudattaa asymptoottisesti multinormaalijakaumaa yhtälön (2.23) mukaisesti. Nyt havainnon ennusteväli voidaan helposti muodostaa simuloimalla uusia havaintoja multinormaalijakaumasta. Toteutetaan ennustevälin simulointi edellisen vaiheen esimerkkitapaukselle seuraavasti:

1. Estimoidaan mallin linkkifunktion odotusarvo havainnolle $\eta_* = \boldsymbol{x}_*^T\widehat{\boldsymbol{\beta}}$.
2. Estimoidaan ennustimen η_* kovarianssimatriisi $var(\eta_*) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
3. Simuloidaan uusia ennustuksia 10 000 kappaletta multinormaalijakaumasta, jonka odotusarvo ja varianssi estimoitiin vaiheissa 1 ja 2.
4. Otetaan simuloiduista tuloksista linkkifunktion käänteisfunktio. Molemmissa malleissa käytettiin log-linkkiä, jonka käänteisfunktio on eksponenttifunktio.
5. Lasketaan simuloitujen havaintojen jakaumasta halutut tunnusluvut, kuten ala- ja yläkvartiili.

Vaiheet 1–5 toteutetaan erikseen vahinkotiheyden sekä keskivahingon osalta. Riskimaksun ennusteväli voidaan nyt esittää ennustevälien tulona. Alakvartiili esimerkin riskimaksulle on 33.7 euroa ja yläkvartiili 48 euroa.

5.2 Riskimaksun riittävyys ja varmuuslisä

Riskimaksun tulisi keskimäärin kattaa vakuutetun riskin odotettu korvausmeno. Vahinkomeno voi vakuutuslajista riippuen kuitenkin heilahdella suuresti ajan kuluessa. Vakuutusmaksuun tuleekin sisällyttää niin sanottu varmuuslisä, jolla turvataan vuosittaisten vaihteluiden aiheuttama epävarmuus korvausmenossa [Pellikka et al. 2020].

Simuloidaan seuraavaksi testiaineistosta erilaisia otoksia sopivan varmuuslisän määrittämiseksi. Jokaiselle testiaineiston havainnoille on laskettu riskimaksu taulukon 11 mukaisesti. Valitaan aineistosta kerrallaan 3000 havaintoa takaisinpalauttamalla, jolloin sama havainto voi tulla valituksi satunnaisotantajoukkoon useamman kerran. Tämän jälkeen satunnaisotantajoukosta lasketaan kokonaissumma sekä toteutuneelle vahinkomenolle että kerätylle riskimaksulle. Satunnaisotannan poiminta ja siitä suoritettava summaus vahinkomenolle ja riskimaksulle suoritetaan yhteensä 10 000 kertaa. Taulukossa 13 on listattuna esimerkinomaisesti muutamia simulointikierrosten tuloksia.

Taulukko 13: Simulointikierrosten tuloksia

Kierroksen numero	Otoskoko	Kokonaisvahinkomeno	Kerätty riskimaksu
1	3000	106 928	101 818
2	3000	114 425	99 642
3	3000	93 536	100 805
.	3000	.	.
.	3000	.	.
.	3000	.	.
9999	3000	89 991	102 938
10 000	3000	97 128	104 648

Taulukosta 13 huomataan, että kierroksesta riippuen, kerätty riskimaksu ei aina riitä kattamaan syntynyttä korvausmenoa. Taulukon 12 mukaan luodulla riskimaksulla, korvausmeno pystytään kattamaan noin 59 prosentilla simulointikierroksista.

Valitaan sopiva varmuuslisä hyvinkin yksinkertaisella menettelyllä. Riskimaksua korotetaan kaikilla havainnoilla kiinteän prosenttiosuuden verran. Taulukkoon 14 on listattu

kuinka kiinteä korotus vaikuttaa korvausmenon kattavien kierrosten lukumäärään simulointiaineistolla.

Taulukko 14: Varmuuslisän merkitys riskimaksun kattavuudessa

Riskimaksun korotus (%)	**Osuus simulointikierroksista, kun riskimaksu on riittävä**
0	59.4 %
1	63.4 %
2	67.1 %
3	70.1 %
4	73.8 %
5	76.8 %
10	88.8 %
20	98.5 %

Varmuuslisä voidaan näin valita kuvatun menetelmän avulla sopivan turvaavalle tasolle. Esimerkiksi 5 prosentin varmuuslisä jo aikaisemmin tuotetun riskimaksun lisäksi riittää kattamaan vahinkomenon yli 75 prosentilla simulaatiokierroksista.

6 Yhteenveto

Tutkielmassa näytettiin, kuinka vakuutusmaksun riskimaksuosa voidaan muodostaa hyödyntämällä yleistettyä lineaarisia malleja. Tutkielman mallit luotiin pitkälti perinteisen kehikon puitteissa. Malleja voidaan esimerkiksi rikastaa lisäämällä selittäviksi tekijöiksi yhdysvaikutuksia sekä hyödyntämällä paikkatietoa riskin maantieteellisen jakautumisen tunnistamisessa. Yleistetyt lineaariset mallit käsittävät myös joukon muita laajennuksia tässä tutkielmassa käytettyihin menetelmiin. Mainittakoon näistä ainakin tasoittavien splinikäyrien käyttö jatkuvien selittävien muuttujien osalta sekä satunnaisvaikutusten hyödyntäminen monitasoisilla kategorisilla muuttujilla.

Erilaisilla koneoppimisen menetelmillä on myös paikkansa riskin tunnistamisessa. Tässä tutkielmassa esitetty koosteoppimismenetelmä ei kuitenkaan onnistunut parantamaan vahinkotiheyden ennustetarkkuutta verrattuna perinteisempään menetelmään. Todennäköiseksi syyksi esitettiin yksittäisten mallien vakautta ja selittävien muuttujien pientä joukkoa, jolloin myös yksittäisten mallien ennusteissa ei ilmennyt riittävää vaihtelua. Menetelmän kykyä parantaa ennustetarkkuutta voisikin jatkossa tutkia myös toisenlaisella riskillä, jossa havaintojen lukumäärä olisi pienempi ja mahdollisten selittävien tekijöiden joukko suurempi.

Tutkielmassa esitettyä koosteoppimismenetelmää kokeiltiin tällä kertaa vain vahinkotiheyden mallinnuksessa. Jatkokehitysmahdollisuuksia voisi laajentaa myös keskivahingon ennustetarkkuuden parantamiseksi. Nyt luotu keskivahinkomalli muodostettiin vain niistä havainnoista, joille oli tilastovuonna 2021 sattunut vahinko. Havaintojoukko on siis huomattavasti pienempi kuin vahinkotiheyden mallinnusjoukko, jolloin koosteoppimismenetelmä saattaisi onnistua parantamaan ennustetarkkuutta. Ennustetarkkuuden parantaminen onkin menetelmän yksi keskeinen tavoite. Toisaalta myös ylisovittamisen välttämiseksi tutkimuskysymyksenä voisi olla menetelmän toteutus käyttämällä huomattavasti pienempää laukkujen lukumäärää.

Lopulliseen riskimaksuun päädyttiin vahinkotiheyden sekä keskivahingon tulona. Lisäksi näytettiin, kuinka yksittäisen havainnon riskimaksun piste-estimaatille voidaan luoda riskimaksun ennusteväli. Ennustevälin avulla voidaan tehdä tulkintoja odotetun vahinkomenon epävarmuudesta. Mielenkiintoinen jatkotutkimuskysymys voisi liittyä ennustevälin hyödyntämiseen riskimaksun varmuuslisän määrittämiseksi.

Sopivan varmuuslisän asettamiseksi ehdotettiin myös vaihtoehtoista menetelmää. Havaintojoukosta poimittiin satunnaisotoksia takaisinpalautuksella 10 000 kappaletta. Jokaisen muodostetun otoksen toteutunutta vahinkomenoa vertailtiin muodostettuun riskimaksuun. Lopuksi tutkittiin, kuinka riskimaksun kiinteä korotus tietyllä prosenttiosuudella pyrkii varmistamaan kerätyn maksun riittävyyttä otoksessa. Menetelmä on hyvinkin yksinkertainen ja mahdollisesti liikaa mallien ennustekykyyn nojautuva. Pyrkimys menetelmällä on kuitenkin intuitiivinen: luoda vaihtelevia satunnaisotoksia muodostuvasta vakuutuskannasta ja siihen kohdistuvista korvausvaatimuksista.

Lähteet

Agresti, A. (2015). *Foundations of linear and generalized linear models*. John Wiley & Sons Inc.

Agresti, A. (2019). *An introduction to categorical data analysis* (3. painos). Wiley.

Akaike, H. (1974). *A new look at the statistical model identification*. IEEE Transactions on Automatic Control, 19.6, s. 716–723.

de Jong, P. & Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data* (1. painos.). Cambridge University Press.

Hardin, J. N. & Hilbe, J. M. (2001). *Generalized linear models and extensions*. Stata Press.

Liu, E. (2023). *A detailed review on the Iterative re-weighted least squares (IRLS) algorithm using Poisson regression and Logistic regression as examples*. https://rpubs.com/enwuliu/1011024

Myers, R. H., Montgomery, D. C., Vining, G. G., & Robinson, T. J. (2012). *Generalized linear models with applications in engineering and the sciences* (2. painos.). Wiley.

Ohlsson, E. & Johansson, B. (2010). *Non-Life Insurance Pricing with Generalized Linear Models* (1. painos). Springer Berlin Heidelberg.

Pellikka, T., Peilimö, P., Puntari, P. & Vaitomaa, M. (2020). *Omaisuuden vakuuttaminen* (4. uudistettu painos.). Finva.

Rantala, J. & Kivisaari, E. (2020). *Vakuutusoppi* (13. painos.). FINVA.

Song, L., Langfelder, P. & Horvath, S. (2013). *Random generalized linear model: a highly accurate and interpretable ensemble predictor*. BMC Bioinformatics 14, 5. https://doi.org/10.1186/1471-2105-14-5

Zhou, Zhi-Hua. (2012). *Ensemble Methods : Foundations and Algorithms*. CRC Press.

Zweifel, P., Eisen, R. & Eckles, D. L. (2021). *Insurance economics* (2. painos.). Springer.

Liitteet

Liite 1: Newton-Raphsonin menetelmä parametrien ratkaisemiseksi. R-koodi.

```
# Algoritmi jatkuu, kunnes erotus on pienempi kuin asetettu lukema.
erotus <- 0.00001

# Vastemuuttuja.
y <- d$VAH_KPL

# Vakio ja yksi selittävä tekijä.
x <- cbind(1,d$IKA)

# Alkuarvaus parametrivektorille.
beta.alku <- c(log(mean(y)),0)

# Päivitettävä parametrivektori. Ensimmäisessä vaiheessa sama kuin alkuarvaus.
beta.loppu <- beta.alku

# Peräkkäisten vaiheiden muutos, joka päivittyy algoritmin jokaisella kierroksella.
muutos <- Inf

# Algoritmi.
while (muutos>erotus) {

  eta <- x %*% beta.alku

  # Diagonaalivektori.
  w <- diag(as.vector(exp(eta)), nrow = nrow(x), ncol = nrow(x))

  # Uusi parametrivektori.
  beta.loppu <- beta.alku + solve(t(x)%*%w%*%x)%*%t(x)%*%(y-exp(eta))

  # Päivitetään muutos.
  muutos <- sqrt(sum((beta.loppu - beta.alku)^2))

  # Asetetaan edellisen vaiheen parametrivektoriin uudet arvot.
  beta.alku <- beta.loppu
  print(beta.alku)
}
```

Liite 2: Koosteoppimismenetelmän toteutus. R-koodi.

```
# Tallennetaan opetusaineisto.
tdata <- traindata

# Laukkujen lukumäärä.
n_bag <- 100

# Maksimimäärä mallin muuttujien lukumääräksi.
max_vars <- c()

# Mallin lopullisten muuttujien lukumäärä.
n_vars <- c()

# Testijoukolle tehtyjen ennusteiden tallennus.
preds_step <- data.frame(matrix(nrow = nrow(testdata)))

# Algoritmin toteutus.
for (i in 1:n_bag) {
  print(i)
  tempIndex <- sample(1:nrow(tdata),nrow(traindata), replace = T)
  tempdata <- tdata[tempIndex,]

  # Mahdolliset muuttujat.
  muuttujat <- c("Muuttuja1","Muuttuja2","Muuttuja3",
              "Muuttuja4","Muuttuja5”,"Muuttuja6","IKALUOKKA")

  # Valitaan muuttujien lukumäärä etenevään. Painotetaan mahdollisuutta käyttää
  kaikkia muuttujia.
  lkm <- sample(5:7, 1, replace = F, prob = c(0.3, 0.3, 0.4))

  # Tallennetaan, kuinka monta muuttujaa malliin on maksimissaan mahdollista saada.
  max_vars <- append(max_vars, lkm)

  # Satunnaisesti valitut muuttujat.
  mjat <- sample(muuttujat, lkm, replace = F)

  # Kaava täydelle mallille
  formu <- as.formula(paste("VAH_KPL ~", paste(mjat, collapse = "+"),"+offset(log(DU-
  RAATIO))"))

  # Vakiotermimalli.
  base_freq <- glm(VAH_KPL ~ offset(log(DURAATIO)),
    family = poisson(link = "log"), data = tempdata)
```

```
# Täysi malli.
  fullmodel_freq <- glm(formula = formu,
                family = poisson(link = "log"), data = tempdata)

  # Etenevä mallivalinta. Lähtee vakiotermimallista ja lisää muuttujia yksi kerrallaan,
  kunnes AIC-arvo ei enää parane.
  tempfreq <- stepAIC(base_freq, scope=list(upper=fullmodel_freq, lower=base_freq),
                 direction = "forward")

# Tallennetaan selittävien muuttujien lopullinen lukumäärä.
  n_vars <- append(n_vars, length(names(tempfreq$xlevels)))

  # Testijoukon ennustukset.
  temppprediction <- predict(tempfreq, newdata = testdata, type = "response")
  temppprediction <- as.data.frame(temppprediction)
  colnames(temppprediction) <- paste0("PRED_", i)
  preds_step <- cbind(preds_step, temppprediction)

}

# Ennusteet.
preds_step <- preds_step[,-1]

# Otetaan ennustuksista keskiarvo.
testdata$PRED_mean_step <- rowMeans(preds_step)
```