Lakisääteisen tapaturmavakuutuksen tapauskohtaisten eläkkeiden tilastollinen ennustaminen

SHV-harjoitustyö (suppea)

Tiina Seppänen

11.9.2014

Sisältö

Abstract ... 1
1 Johdanto ... 2
2 Lakisääteisen tapaturmavakuutuksen korvauksista ... 2
2.1 Yleistä ... 2
2.2 Korvauslajit ... 3
2.2.1 Ohimenevät korvaukset ... 3
2.2.2 Pysyvät korvaukset ... 4
2.2.3 Jakojärjestelmäkorvaukset ... 5
3 Tuntemattomien ja tunnettujen eläkkeiden korvausvastuu ... 5
4 Logistinen regressiomalli ... 6
4.1 Yleistetyt lineaariset mallit ... 6
4.2 Mallin sopivuuden tarkastelu ... 8
5 Eläke-ennustemalli ... 10
5.1 Mallin esittely ... 10
5.2 Mallinnusaineisto ... 11
5.3 Tulokset ... 12
5.3.1 Esimerkkitapauksia ... 19
5.4 Mallin validointi ... 19
6 Yhteenveto ... 21
Lähteet ... 23

Abstract

Workers' compensation covers work-related accidents and occupational diseases. The purpose of the insurance is to compensate for loss of earnings, costs of medical treatment, permanent physical impairments and some other damages and costs. Compensations are determined according to the Employment Accidents Act. In this work we consider only occupational accidents, not diseases.

After an occupational accident employee's disability is compensated by daily allowance. Accident pension is paid if disability lasts for at least a year. Workers' compensation is heavily long-tailed insurance because accident pension can be paid until the death of an injured. Technical reserves are set up in respect of future payments. They constitute a large part of the insurer's liabilities. Reserves also play a significant role in rating of large companies' insurances because reserves, as a part of insured's claims costs, have an impact on premium in individual tariffs.

In this work a statistical probability model is built up for predicting and identifying individual accident pensions earlier than at present. The prediction is based on claims data like compensated daily allowances. Separate models are made for claims of different age because the amount of information in the data depends on the claim's age. The modelling is carried out in SAS using logistic regression.

Individual pension probabilities can be used in assessing the collective claims reserve for unknown pension cases. The probabilities can also be utilized in monitoring and reporting claims of large insureds. However, the parameter estimates cannot directly be applied to claims data of other insurance companies because the estimates depend on the situation and the nature of the modelling data at a certain point. Claims development and processes differ between companies.

1 Johdanto

Lakisääteisen tapaturmavakuutuksen korvauksilla työntekijän toimeentulo pyritään säilyttämään vastaavantasoisena kuin ennen työtapaturmaa tai ammattitautia. Vakuutuksesta maksettavat korvaukset ovat tapaturmavakuutuslaissa säädettyjä. Korvauksista suurimman osan muodostavat ansionmenetyskorvaukset ja sairaanhoitokulut.

Tapaturmaeläkettä maksetaan niin kauan kuin työtapaturma tai ammattitauti aiheuttaa työkyvyttömyyttä, pisimmillään vahingoittuneen koko lopun eliniän ajan vahinkotapahtumasta eteenpäin. Tulevan ajan korvauksia varten varattavat tapaturmaeläkkeiden rahastoidut pääoma-arvot muodostavat merkittävän osan vakuutusyhtiön korvausvastuusta. Näiden tunnettujen tapauskohtaisten varausten suuruus vaihtelee kymmenistä tuhansista jopa miljooniin euroihin, joten varausten muutokset voivat aiheuttaa yhtiön tulokseen huomattavaa heiluntaa. Tunnettujen vahinkojen eläkevastuussa esiintyvä heilunta on vaikeasti ennustettavissa. Lisäksi keskisuurten ja suurten asiakkaiden vakuutukset hinnoitellaan erikoismaksujärjestelmillä, joissa varausten pääoma-arvot vaikuttavat korvausmenona työnantajan vakuutusmaksuun.

Tapauskohtaisten eläkkeiden tilastollisen ennustamisen tarkoituksena on pyrkiä tunnistamaan ja varaamaan muutaman ensimmäisen kehitysvuoden aikana tapauskohtaisesti varattavat eläkkeet aiempaa nopeammin ja tasaisemmin. Työssä muodostetaan eläkkeen todennäköisyydelle ennustemalli käyttämällä logistista regressiota ja hyödyntämällä vahingosta olemassa olevaa informaatiota, kuten tietoa siitä, kuinka monta päivää vahingosta on maksettu päivärahaa.

Mallin antamia vahinkokohtaisia todennäköisyyksiä voidaan hyödyntää sekä kollektiivisen korvausvastuun laskennassa että erikoismaksujärjestelmiin kuuluvien asiakkaiden vahinkojen seurannassa.

Työssä rajoitutaan työtapaturmavahinkoihin ja jätetään ammattitaudit tarkastelun ulkopuolelle.

2 Lakisääteisen tapaturmavakuutuksen korvauksista

2.1 Yleistä

Lakisääteinen tapaturmavakuutus on osa sosiaaliturvaa ja maamme vanhin sosiaalivakuutus [10, s. 13]. Nykyinen tapaturmavakuutuslaki [17] on vuodelta 1948. Sen jälkeen lakiin on tehty monia muutoksia ja useita vuosia työn alla olleen lain kokonaisuudistuksen on tarkoitus toteutua vuonna 2016 [15].

Lakisääteinen tapaturmavakuutus perustuu tapaturmavakuutuslakiin sekä korvausten että vakuutusasioiden osalta. Tapaturmavakuutuslain mukaisilla korvauksilla turvataan työntekijän tai hänen omaistensa toimeentulo työkyvyttömyyden ja kuoleman varalta työtapaturman tai ammattitaudin jälkeen. Oikeus korvaukseen perustuu ensisijaisesti työsuhteeseen ja tapaturmavakuutuslain säännöksiin [10, s. 13]. Yksityisen ja kunnallisen sektorin työantajalla on velvollisuus ottaa työntekijöitään varten vakuutus tapaturmavakuutuslaitoksesta. Valtiolla ei ole vakuuttamisvelvollisuutta ja korvaukset hoitaa Valtiokonttori [10, s. 17]. Työ- ja virkasuhteisten lisäksi korvaukseen oikeutettuja ovat eräät laissa mainitut erityisryhmät, kuten maatalousyrittäjät sekä koululaiset ja opiskelijat tietyissä olosuhteissa [10, s. 18–19]. Pakollisen vakuutusturvan ulkopuolelle jäävät henkilöt, kuten yrittäjät, voidaan vakuuttaa vapaaehtoisella

vakuutuksella, joka vastaa turvaltaan pakollista vakuutusta. Sekä pakollisessa että vapaaehtoisessa vakuutuksessa voidaan vakuutusturva laajentaa kattamaan myös vapaa-ajan vahingot [10, s. 200].

Työtapaturmalla tarkoitetaan työntekijää työssä tai työstä johtuvissa olosuhteissa, kuten työmatkalla, tapahtunutta tapaturmaa, joka aiheuttaa vamman tai sairauden. Tapaturmalle on tyypillistä äkillisyys ja ennalta arvaamattomuus [10, s. 45–46]. Ammattitaudit sitä vastoin eivät ole tällaisia lyhyenä aikana syntyneitä vammoja vaan usean vuoden altistumisen seurauksena syntyviä terveydellisiä haittoja [10, s. 60]. Tässä työssä rajoitutaan työtapaturmavahinkoihin.

Lakisääteinen tapaturmavakuutus on ensisijainen muihin korvausjärjestelmiin, kuten sairausvakuutukseen, nähden. Ensisijaisuus tarkoittaa sitä, että mikäli kyseessä on tapaturmavakuutuslain mukaan korvattava tapahtuma, korvataan se ensin työtapaturmana tai ammattitautina. Vakuutustapahtuma voi myös kuulua kahden järjestelmän piiriin, kuten työmatkalla sattunut liikennevahinko. Tällöin liikennevakuutuksesta voidaan maksaa lisäkorvausta tapaturmakorvausten jälkeen. Tapaturmavakuutusyhtiöllä on oikeus tietyin edellytyksin saada maksamansa korvaus takaisin niin sanottuna regressinä liikennevakuutusyhtiöltä. Liikennevahinkojen lisäksi regressioikeus voi tulla kyseeseen myös vastuu- tai potilasvahingoissa, kuten vastuuvakuutuksen liukastumisvahingossa. [10, s. 66, 138–140]

2.2 Korvauslajit

Lakisääteisen tapaturmavakuutuksen korvaukset jaotellaan tyypillisesti ohimeneviin ja pysyviin korvauksiin. Koko korvausmenosta noin puolet syntyy ohimenevistä ja toinen puoli pysyvistä korvauksista. Ohimeneviä korvauksia maksetaan yleensä kaikista tapaturmista, mutta vain pieni osa vahingoista johtaa pysyviin korvauksiin. Pysyviin korvauksiin johtaneista vahingoista aiheutunut korvausmeno on kuitenkin hyvin suuri. [10, s. 188–189]

Ohimenevien ja pysyvien korvausten lisäksi lakisääteisen tapaturmavakuutuksen korvauksista erotetaan yleensä omaksi ryhmäkseen jakojärjestelmäkorvaukset. Toisin kuin ohimenevät ja pysyvät korvaukset, jakojärjestelmäkorvaukset ovat niin sanottuja ei-rahastoitavia korvauksia eli ne katetaan vakuutusmaksuilla sitä mukaa kuin korvauksia maksetaan. Kullakin vakuutuskaudella sattuneet vahingot on rahastoitavien korvausten osalta pyrittävä rahoittamaan kyseessä olevan kauden vakuutusmaksuilla. [10, s. 189–190], [8, s. 8–9]

2.2.1 Ohimenevät korvaukset

Ohimeneviä korvauksia ovat päiväraha, sairaanhoidon kustannukset, kuten lääkärinpalkkiot, lääkkeet ja matkakulut, fysikaalisen hoidon aiheuttama ansionmenetys, kodinhoidon lisäkustannukset enintään vuoden ajalta, proteesien ja muiden apuvälineiden hankkiminen, työtapaturman yhteydessä särkyneet silmälasit, kuulokojeet ja vastaavat. Sairaanhoitokuluja korvataan niin kauan kuin niitä aiheutuu eikä kulujen korvaamiselle ole euromääräisiä rajoja. [10, s. 99–101, 189]

Sairaanhoitokulujen korvaussääntöjen uudistus, niin sanottu täyskustannusvastuu eli TÄKY, koskee vuonna 2005 tai sen jälkeen sattuneita työtapaturmia. Uudistuksen tavoitteena oli työtapaturmien hoidon ja työhön paluun nopeutuminen. Sääntöjen mukaan hoitoon voi mennä niin julkiselle kuin yksityisellekin sektorille. Julkisen sektorin hoidosta korvataan asiakasmaksu ja lisäksi kunnalle tai kuntayhtymälle niin

sanotut todelliset kustannukset. Yksityisen sektorin hoidosta korvataan todelliset kustannukset. Isommissa tutkimuksissa ja toimenpiteissä, kuten magneettikuvauksissa ja leikkaushoidossa, vakuutusyhtiöllä on oikeus valita hoitopaikka niin yksityisellä kuin julkisellakin sektorilla. Näissä tilanteissa todellisten kustannusten korvaaminen edellyttää maksusitoumusta vakuutusyhtiöltä. Ennen vuotta 2005 sattuneiden työtapaturmien osalta lääkärikäynneistä ja toimenpiteistä korvataan yksityisellä sektorilla todelliset kustannukset, mutta julkisella sektorilla ainoastaan asiakasmaksu. [10, s. 101–106]

Päivärahaa maksetaan työkyvyttömyyden ajalta, kun työkyvyttömyys kestää vähintään kolme peräkkäistä päivää tapaturmapäivää lukuun ottamatta. Päivärahaa maksetaan enintään vuoden ajan tapaturmasta lukien. Tämän jälkeen mahdollista työkyvyttömyyttä korvataan tapaturmaeläkkeellä. [10, s. 72–73]

Päivärahan määrä perustuu vahingoittuneen omiin ansioihin. Jos työantaja maksaa vahingoittuneelle sairausajan palkkaa, on työnantajalla oikeus saada korvauksesta takaisin sairausajan palkkaa vastaava määrä. Ensimmäisten neljän viikon osalta päivärahan määrä perustuu joko sairausajan palkkaan tai tapaturmaa edeltävien neljän viikon ansioihin. Tämän jälkeen päivärahan määrä perustuu vahingoittuneen vuosityöansioon. Vuosityöansio määrätään tapaturman sattumisajankohdan mukaan. [10, s. 72–81]

Tässä työssä muilla ohimenevillä korvauksilla tarkoitetaan ohimeneviä korvauksia lukuun ottamatta päivärahaa. Muut ohimenevät korvaukset eivät sisällä myöskään jakojärjestelmäkorvauksia eikä niistä ole vähennetty regressinä saatavia korvauksia niissä vahingoissa, joissa on regressioikeus.

2.2.2 Pysyvät korvaukset

Pysyviä korvauksia ovat pysyvän työkyvyttömyyden ja kuolemantapausten johdosta maksettavat korvaukset, kuten tapaturmaeläke, haittaraha, haittalisä, vaate- ja opaskoiralisä, perhe-eläke ja hautausapu. [10, s. 189]

Tapaturmaeläkettä maksetaan, mikäli työtapaturma aiheuttaa työkyvyttömyyttä sen jälkeen, kun on kulunut vuosi tapaturman sattumisesta. Tapaturmaeläke voi olla määräaikainen tai kestää koko eliniän vahinkotapahtumasta eteenpäin. Määräaikaista eläkettä maksetaan usein esimerkiksi kuntoutuksen ajalta. Tapaturmista hyvin pieni osa, noin yksi prosentti, aiheuttaa pysyvän työkyvyttömyyden. Pääsäännön mukaan täysi tapaturmaeläke on 85 prosenttia vahingoittuneen vuosityöansiosta ja 65 ikävuoden jälkeen 70 prosenttia vuosityöansiosta. Osaeläke on työkyvyn alenemaa kuvaavan prosenttiluvun mukainen osa täydestä eläkkeestä. [10, s. 88–89]

Tapaturmaeläke on kuukausittain maksettava jatkuva korvaus ja tulevan ajan korvauksia vastaavan summan vakuutusyhtiö varaa rahastoituna pääoma-arvona korvausvastuuseen. Lakisääteisen tapaturmavakuutuksen hinnoittelussa keskisuurten ja suurten asiakkaiden erikoismaksujärjestelmissä vakuutuksenottajan oma vahinkotilasto vaikuttaa vakuutusmaksuun, joten näissä järjestelmissä eläkkeen pääoma-arvo vaikuttaa korvausmenona vakuutusmaksuun joko osittain tai kokonaan maksujärjestelmän luonteesta riippuen. [10, s. 89, 191–192]

Haittarahaa maksetaan työtapaturman aiheuttamasta pysyvästä yleisestä haitasta eli toimintakyvyn alentumisesta aikaisintaan vuoden kuluttua tapaturman sattumisesta. Haittarahan suuruus määräytyy haittaluokan mukaan lain taulukosta. [10, s. 113]

Haittalisää maksetaan, jos vamma tai sairaus aiheuttaa vahingoittuneelle poikkeuksellista haittaa tai hän on niin avuttomassa tilassa, ettei tule toimeen ilman toisen henkilön apua. Haittalisän määrä perustuu tapaturmavakuutuslaissa olevaan vamman tilakuvauksen mukaiseen luokitukseen. Vaatelisää maksetaan määrältään vakiokorvauksena proteesin, tukisidosten tai vastaavan käytöstä aiheutuvasta vaatteiden erityisestä kulumisesta. Opaskoiralisää maksetaan sairaanhoitokustannuksista korvatusta opaskoirasta aiheutuvien menojen kattamiseksi. [10, s. 111–112]

Kuolemantapauksissa maksetaan hautausapua ja perhe-eläkettä. Hautausavun määrä on kiinteä, vuosittain muuttuva summa. Perhe-eläkettä maksetaan leskelle ja lapsille ja sen enimmäismäärä on 70 prosenttia vahingoittuneen vuosityöansiosta. [10, s. 129–130]

2.2.3 Jakojärjestelmäkorvaukset

Jakojärjestelmäkorvauksia ovat eläkkeiden, haittarahojen ja lisien sekä eräiden muiden korvausten indeksikorotukset, vähintään kymmenen vuotta vanhoista vahingoista maksettavat sairaanhoito- ja lääkinnällisen kuntoutuksen korvaukset sekä eräät muut tapaturmavakuutuslaissa luetellut korvaukset. Lakisääteistä tapaturmavakuutusta harjoittavat vakuutusyhtiöt osallistuvat näiden korvausten rahoitukseen vuosittain vakuutusmaksutulon mukaisessa suhteessa kerättävillä jakojärjestelmämaksuilla. [10, s. 189–190], [8, s. 8, 44]

3 Tuntemattomien ja tunnettujen eläkkeiden korvausvastuu

Vahingon sattuessa vakuutusyhtiölle syntyy velvollisuus korvata siitä aiheutuvat kustannukset, vaikka vahinko ei vielä olisi yhtiön tiedossa. Vahingon ilmoittaminen yhtiöön, korvausten käsittely ja muut tekijät aiheuttavat viivettä vahingon sattumisen ja korvausten maksaminen välille. Erityisen pitkäksi viive muodostuu eläkevahingoissa, joissa pysyvän työkyvyttömyyden johdosta tapaturmaeläkettä voidaan maksaa vuosikymmenien ajan. Tulevaisuudessa maksettavat korvaukset yhtiön on tilinpäätöksessä sisällytettävä vastuuvelkaansa, tarkemmin korvausvastuuseen. [13, s. 1]

Korvausvastuun eläkevahinkoihin liittyvät erät koostuvat vahinkokohtaisista varauksista ja kollektiivisesta korvausvastuusta. Tunnettujen keskeneräisten vahinkojen maksamattomat pysyvät eläkemuotoiset korvaukset, kuten päivärahakauden jälkeen maksettavat ansionmenetyskorvaukset ja haittarahat sekä niihin liittyvät lisät, varataan vahinkokohtaisina arviovarauksina. Arviovaraus tehdään vahinkokohtaisten tietojen, kuten työkyvyttömyyden keston, työkyvyn aleneman ja vahingoittuneen vuosityöansion perusteella, jotta varaus vastaa mahdollisimman hyvin tulevaisuudessa maksettavaksi tulevia korvauksia. Sen jälkeen kun on tehty päätös toistaiseksi maksettavasta tapaturmaeläkkeestä ja varauksen määräämisessä käytetyt vahingoittunutta koskevat tiedot ovat vakuutusyhtiön arvion mukaan riittävissä määrin vakiintuneet, voidaan tapauskohtainen varaus vahvistaa lopulliseksi. Ennen päätöksen antamista vakuutusyhtiön on pyydettävä lausunto tapaturma-asiain korvauslautakunnalta [10, s. 162]. Keskeneräisten vahinkojen muut ohimenevät ja pysyvät korvaukset sekä tuntemattomien vahinkojen korvaukset varataan kollektiivisesti. [13, s. 5], [8, s. 9]

Kollektiivisella korvausvastuulla tarkoitetaan tietylle vahinkojen joukolle tilastollisin menetelmin yhteisesti estimoitua korvausvastuuta [13, s. 5]. Kollektiivivarausten määräämiseksi on olemassa useita erilaisia menetelmiä ja varausten arvioimisessa hyödynnetään vahingoista olemassa olevaa tilastoaineistoa. Vahinkojen raportoitumisen ja selviämisen tarkastelemista varten tilastoaineisto esitetään tyypillisesti korvauskolmion eli niin sanotun run-off – kolmion muodossa. Korvauskolmiossa maksetut korvaukset, maksetut korvaukset ja vahinkokohtaiset varaukset tai vahinkojen lukumäärät taulukoidaan sattumis- ja kehitysjakson mukaan. Kehitysjaksoilla tarkoitetaan niitä jaksoja $i, i + 1, i + 2, \dots, k$, joilta vakuutusyhtiöllä on vuoden k lopussa käytössään tilastoaineisto sattumisvuoden $i, i \leq k$, vahingoista. Jakson pituus voi olla esimerkiksi kuukausi, kvartaali tai vuosi [13, s. 8].

4 Logistinen regressiomalli

4.1 Yleistetyt lineaariset mallit

Klassisessa lineaarisessa mallissa vastemuuttujan Y_i riippuvuus selittävistä muuttujista x_{ij} on muotoa

$$\mathrm{E}(Y_i) = \mu_i = \sum_{j=1}^{p} x_{ij}\,\beta_j,$$

missä $\beta_1, \dots, \beta_p$ ovat mallin tuntemattomia parametreja, $x_{ij}, j = 1, \dots, p$, selittävien muuttujien arvot havaintoyksiköille $i = 1, \dots, n$, ja $Y_i \sim N(\mu_i, \sigma^2)$, $\mathrm{var}(Y_i) = \sigma^2$. Muotoa μ_i kutsutaan mallin systemaattiseksi osaksi. Lineaarisessa mallissa oletetaan, että vastemuuttuja Y_i voi saada mitä tahansa reaalilukuarvoja. Tämän vuoksi on monia käytännön tilanteita, kuten lukumäärä- ja binäärivasteet, joihin malli ei sovellu. Aina mallin systemaattisessa osassa riippuvuus tuntemattomista parametreista β_j ei myöskään ole lineaarista muotoa. [5, s. 5], [7, s. 5–6], [3, s. 45]

Binäärivasteen tilanteessa vastemuuttuja Y_i voi saada ainoastaan arvoja 1 tai 0. Esimerkiksi tieto siitä, päätyykö vahinko eläkkeeksi vai ei, on binäärinen vaste. Tällöin Y_i saa arvon 1, mikäli vahingosta tulee eläke, ja muussa tapauksessa arvon 0. Binäärivaste noudattaa Bernoulli-jakaumaa ja jakauman parametri on positiivisen vasteen todennäköisyys π_i, $0 < \pi_i < 1$, toisin sanoen todennäköisyys sille, että vahinko päätyy eläkkeeksi. Jakauman tiheysfunktio on

$$f_{Y_i}(y_i; \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{(1-y_i)}, \qquad y_i = 0{,}1,$$

odotusarvo π_i ja varianssi $\pi_i(1 - \pi_i)$. [5, s. 9], [2, s. 21]

Yleistetyt lineaariset mallit (generalized linear models, GLM) ovat klassisten lineaaristen mallien laajennus tilanteisiin, joissa klassista lineaarista mallia ei voida käyttää. Mallin systemaattinen osa on nyt muotoa

$$\eta_i = g(\mu_i) = \sum_{j=1}^{p} x_{ij}\,\beta_j,$$

missä g on reaaliarvoinen monotoninen ja derivoituva niin sanottu linkkifunktio. Termiä η_i kutsutaan lineaariseksi ennusteeksi ja varianssifunktio V, $\mathrm{var}(Y_i) = V(\mu_i)$, ei ole nyt välttämättä vakio, kuten lineaarisessa mallissa. [5, s. 9–10], [7, s. 6–7], [11, s. 27]

Vastemuuttujan Y_i jakauma kuuluu eksponenttiperheeseen ja sen tiheysfunktion oletetaan olevan muotoa

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{a_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y_i, \phi / a_i)\right),$$

missä $\theta_i, i = 1, \dots, n$, ovat tuntemattomia niin sanottuja kanonisia parametreja, ϕ on hajontaparametri (tunnettu tai tuntematon), a_i ovat havaintoyksiköihin $1, \dots, n$ liittyviä tunnettuja prioripainoja ja b ja c ovat tunnettuja funktioita. Eksponenttiperheeseen kuuluvan jakauman odotusarvolle ja varianssille pätee

$$\mathrm{E}(Y_i) = b'(\theta_i) = \mu_i$$

$$\mathrm{var}(Y_i) = \frac{b''(\theta_i)\phi}{a_i} = \frac{V(\mu_i)\phi}{a_i}.$$

Mallin parametrit estimoidaan suurimman uskottavuuden (maximum likelihood) menetelmällä, jolloin parametriestimaatit maksimoivat log-uskottavuusfunktion

$$l(\theta_1, \dots, \theta_n; \phi, \boldsymbol{a}, \boldsymbol{y}) = \sum_{i=1}^{n} \log f_{Y_i}(y_i; \theta_i, \phi).$$

Vastemuuttujan Y_i ollessa binäärinen odotusarvo μ_i voidaan rajata välille (0,1) logit-linkkifunktiolla

$$g(\mu_i) = \mathrm{logit}(\mu_i) = \log\left(\frac{\mu_i}{1 - \mu_i}\right) = \sum_{j=1}^{p} x_{ij}\,\beta_j,$$

missä $\mu_i = \mathrm{E}(Y_i) = \pi_i$ on positiivisen vasteen todennäköisyys. Nähdään, että tiheysfunktio on eksponenttiperheen muotoa valinnoilla $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = \log(1 + e^{\theta_i})$ ja $\phi = 1$. [5, s. 14–18], [7, s. 13–14], [2, s. 37, 67, 98]

Logistisessa mallissa positiivisen ja negatiivisen vasteen todennäköisyyksien suhde (odds) on

$$\frac{\mu_i}{1 - \mu_i} = \exp\left(\sum_{j=1}^{p} x_{ij}\,\beta_j\right)$$

ja positiivisen vasteen todennäköisyys

$$\mu_i = \pi_i = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{1 + \exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}.$$

Logistisen mallin parametrien $\beta_j, j = 1, \dots, p$, tulkinta nähdään seuraavasti. Jos k:nnen selittävän muuttujan, x_{ik}, arvoa kasvatetaan yhdellä muiden selittävien muuttujien pysyessä muuttumattomina,

$$x'_{ij} = \begin{cases} x_{ij} + 1, & \text{kun } j = k, \\ x_{ij}, & \text{muilla } j, \end{cases}$$

muuttuu positiivisen ja negatiivisen vasteen todennäköisyyksien suhde (odds) kertoimella

$$\frac{\mu'_i/(1-\mu'_i)}{\mu_i/(1-\mu_i)} = \exp\left[\sum_{j=1}^{p} \beta_j\left(x'_{ij} - x_{ij}\right)\right] = e^{\beta_k}.$$

Tätä suhdetta kutsutaan nimellä odds ratio. [5, s. 58–59], [2, s. 98]

Logistisessa mallissa vastemuuttujan ja selittävien tekijöiden välinen riippuvuus voi olla myös ei-lineaarinen, toisin sanoen malli on muotoa

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \sum_{j=1}^{p}\sum_{h=1}^{l} x_{ij}^{h}\,\beta_{j,h}, \qquad l \geq 1.$$

Esimerkiksi toisen tai kolmannen asteen polynomi voi kuvata muuttujien välistä riippuvuutta ensimmäisen asteen polynomia paremmin. Ei-lineaarisen riippuvuuden heikkoutena on se, että tällöin parametrien tulkinta ei ole yhtä suoraviivaista. Kun lineaarisessa mallissa positiivisen ja negatiivisen vasteen todennäköisyyksien suhde kasvaa kertoimella e^{β_k} k:nnen selittävän muuttujan kasvaessa yhdellä, riippuu vaikutus ei-lineaarisessa mallissa myös selittävän muuttujan arvosta x_{ik}. [2, s. 99–100]

Tutkittaessa selittävien tekijöiden ja vastemuuttujan välisen riippuvuuden muotoa on hyödyllistä piirtää vastemuuttuja kutakin selittävää tekijää vasten. Toisinaan selittävän tekijän jokin muunnos, kuten logaritmi, voi toimia mallissa alkuperäistä muuttujaa paremmin [2, s. 4–11]. Mikäli selittävälle muuttujalle tehdään logaritmimuunnos, vastaa logaritmoidun muuttujan yhden yksikön kasvu alkuperäisen muuttujan kertomista luonnollisen logaritmin kantaluvulla ($e \approx 2{,}7$). Tällöin odds ratio siis kertoo, kuinka paljon positiivisen ja negatiivisen vasteen todennäköisyyksien suhde (odds) muuttuu, kun selittävä muuttuja kasvaa noin 2,7-kertaiseksi.

4.2 Mallin sopivuuden tarkastelu

Selittävien tekijöiden välistä lineaarista riippuvuutta voidaan jatkuvien muuttujien tapauksessa tutkia Pearsonin korrelaatiokertoimen avulla. Kertoimen arvo on välillä $[-1{,}1]$, positiiviset arvot kertovat muuttujien välisestä positiivisesta korrelaatiosta ja negatiiviset arvot vastaavasti negatiivisesta korrelaatiosta. Kertoimen saadessa arvon 0 ei muuttujien välillä ole lineaarista riippuvuutta [12, s. 41–45]. Selittävien tekijöiden välistä riippuvuutta eli multikollineaarisuutta voidaan lisäksi tarkastella laskemalla niin sanottu VIF (variance inflation factor)

$$\mathrm{VIF}_k = \frac{1}{1 - R^2_{(k)}}$$

arvo. Tässä $R^2_{(k)}$ on mallin selitysaste sellaisessa mallissa, jossa k:nnetta tekijää selitetään muilla tekijöillä. VIF saa arvon 1, mikäli muuttuja ei korreloi muiden tekijöiden kanssa ja viittä suurempia arvoja voidaan pitää merkkinä siitä, että korrelaatio on niin vahva, että jokin korreloivista tekijöistä on syytä jättää mallista pois [3, s. 101–102].

Kuhunkin selittävään tekijään liittyvän parametriestimaatin β_j tilastollista merkitsevyyttä voidaan testata χ^2-jakaumaa yhdellä vapausasteella noudattavan Waldin testin

$$\frac{\hat{\beta}_j^2}{\mathrm{Var}(\hat{\beta}_j)} \sim \chi_1^2$$

ja siihen liittyvän p-arvon avulla. [2, s. 75]

Selittävien tekijöiden lisääminen malliin parantaa mallin ja aineiston yhteensopivuutta, mutta tarpeettomien, vaikkakin tilastollisesti merkitsevien, tekijöiden lisääminen toisaalta huonontaa parametriestimaattien tarkkuutta eli lisää epävarmuutta. Tavoitteena olisi löytää sellainen määrä tekijöitä, että malli soveltuu aineistoon hyvin, mutta parametriestimaattien keskihajonnat olisivat kuitenkin riittävän pieniä. Sopivan mallin valintaan voidaan käyttää esimerkiksi Akaiken informaatiokriteeriä (Akaike's Information Criterion)

$$\mathrm{AIC} \equiv -2l + 2p,$$

missä l on mallin log-uskottavuusfunktio ja p on selittävien muuttujien lukumäärä. Hyvin sopivassa mallissa uskottavuusfunktio saa suuren arvon, jolloin $-2l$ on pieni. Termi $2p$ puolestaan on selittävien tekijöiden eli parametrien lukumäärään liittyvä sakkotermi, joka siis kasvaa lisättäessä malliin selittäviä tekijöitä. Samaan aineistoon sovitetuista malleista kriteerin perusteella valitaan se, jolla on pienin AIC arvo. Pieninkään arvo ei kuitenkaan takaa sitä, että malli soveltuisi aineistoon hyvin, vaan yhteensopivuutta on lisäksi tarkasteltava muilla tavoin. [2, s. 62–63]

Logistisen regressiomallin käytännöllisyyttä arvioidaan usein tarkastelemalla mallin ennustavuutta ja erottelukykyä. Näissä tarkasteluissa eläke-ennustemallissa tietyn todennäköisyyden ylittävät tapaukset luokitellaan eläkkeiksi ja alittavat ei-eläkkeiksi. Mallin erottelukykyä voidaan graafisesti tarkastella niin sanotun ROC (receiver operating characteristic) – käyrän avulla. Siinä y-akselilla oleva mallin sensitiivisyys piirretään x-akselilla olevaa arvoa $1 -$ spesifisyys vasten muuttamalla rajana käytettyä todennäköisyyttä (threshold) vähän kerrallaan. Sensitiivisyydellä tarkoitetaan mallin herkkyyttä eli sitä osuutta todellisista eläkkeistä, jotka malli luokittelee eläkkeiksi, ja spesifisyydellä mallin tarkkuutta. $1 -$ spesifisyys kuvaa sitä osuutta ei-eläkkeistä, jotka malli virheellisesti luokittelee eläkkeiksi. Diagonaalilla kulkeva suora $y = x$ vastaa täysin hyödyttömän luokittelun ROC-käyrää ja ideaalitilanteessa ROC-käyrä kulkee pisteestä (0,0) pystysuorasti pisteeseen (0,1) ja siitä vaakasuorasti pisteeseen (1,1). [2, s. 108–110], [4, s. 1–2]

ROC-käyrän lisäksi mallin erottelukykyä kuvaa niin sanottu AUC (area under the curve) – tunnusluku. Tunnusluvun ollessa 1 mallin tarkkuus on paras mahdollinen ja tunnusluvun ollessa 0,5 mallin erottelukyky on hyödytön. Tunnusluvun arvo vastaa sitä todennäköisyyttä, jolla satunnaisesti valittu eläkevahinko saa mallissa korkeamman vastetodennäköisyyden kuin satunnaisesti valittu vahinko, joka ei ole eläketapaus. [2, s. 109], [4, s. 2]

Harvinaisen vasteen tilanteessa, jolloin positiivisen arvon saavia havaintoja on vähemmän, ROC-käyrän ja AUC-tunnusluvun informatiivisuus kuitenkin vähenee, koska ROC-käyrällä on taipumus nousta nopeasti kohti vasemman yläkulman (0,1)-pistettä ja AUC-tunnusluku saa tyypillisesti lähellä ykköstä olevia arvoja. [6]

Mallin luokittelukyvyn tarkkuutta voidaan tutkia myös esimerkiksi niin kutsutun F_1-mitan (F_1 score)

$$F_1 = 2 \cdot \frac{pr \cdot s}{pr + s}$$

avulla, missä s tarkoittaa mallin sensitiivisyyttä ja pr mallin tarkkuutta (precision) eli todellisten eläkkeiden osuutta kaikista mallin eläkkeiksi luokittelemista tapauksista tietyllä rajana käytetyllä todennäköisyydellä. F_1- mitta saa huonoimmillaan arvon 0 ja parhaimmillaan arvon 1. [6]

5 Eläke-ennustemalli

5.1 Mallin esittely

Eläke-ennustemallin tavoitteena on pyrkiä tunnistamaan muutaman ensimmäisen kehitysvuoden aikana varattavat eläketapaukset. Eläkkeen todennäköisyyttä mallinnetaan logistisella regressiolla. Mallissa vastemuuttujana on tieto siitä, löytyykö vahingolle työtapaturmaeläkkeen varausta tarkasteluhetkestä riippumatta.

Koska eri-ikäisistä vahingoista on aineistossa eri määrä informaatiota, muodostetaan eri-ikäisille vahingoille omat mallit. Ajan myötä lisääntyvä informaatio pyritään hallitsemaan erillisillä malleilla siten, etteivät eläketodennäköisyydet muutu liikaa mallista toiseen siirryttäessä.

Työssä muodostetaan omat mallit kolmen kuukauden, kuuden kuukauden, vuoden, puolentoista vuoden, kahden, kahden ja puolen vuoden sekä kolmen, neljän, viiden, kahdeksan ja kymmenen vuoden ikäisille vahingoille. Työssä korkeintaan vuoden ikäisten vahinkojen malleja kutsutaan nuorten vahinkojen malleiksi. Keski-ikäisten vahinkojen malleilla tarkoitetaan puolentoista vuoden, kahden sekä kahden ja puolen vuoden ikäisille vahingoille tehtyjä malleja. Tätä vanhempien vahinkojen malleja kutsutaan vanhojen vahinkojen malleiksi. Jotta mahdolliset eläketapaukset voitaisiin tunnistaa aiempaa varhaisemmassa vaiheessa, oleellisimpia malleista ovat nuorten ja keski-ikäisten vahinkojen mallit. Tätä vanhempien vahinkojen malleja käytetään lähinnä tarkasteltaessa, kuinka eläkkeen todennäköisyys muuttuu ajan kuluessa.

Selittävinä tekijöinä nuorten vahinkojen malleissa ovat päivärahakauden pituus eli tieto siitä, kuinka monta vuorokautta vahingosta on maksettu päivärahaa, maksetut muut ohimenevät korvaukset euroina, vahingoittuneen ikä vahinkohetkellä ja viimeisestä korvauksesta kulunut aika vuorokausina, jota työssä kutsutaan nimellä hiljaiselo. Keski-ikäisten ja vanhojen vahinkojen malleissa on selittävänä tekijänä edellä mainittujen lisäksi työkyvyttömyyskauden pituus eli tieto siitä, kuinka monta vuorokautta vahingosta on maksettu ansionmenetyskorvausta joko kuntoutuksen tai muun työkyvyttömyyden ajalta päivärahakauden jälkeen. Ennustemalli kunkin ikäisille vahingoille on siis muotoa

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \sum_{j=1}^{5}\sum_{h=1}^{l} x_{ij}^{h}\,\beta_{j,h},$$

missä μ_i on eläkkeen todennäköisyys, β_0 on vakiotermi, $\beta_{j,h}$ ovat parametriestimaatteja, x_{i1} = päivärahakauden pituus, x_{i2} = muut ohimenevät korvaukset, x_{i3} = vahingoittuneen ikä vahinkohetkellä, x_{i4} = hiljaiselo, x_{i5} = työkyvyttömyyskauden pituus, $l = 2$, kun $j = 4$, ja $l = 1$ muulloin. Nuorten vahinkojen malleissa $x_{i5} = 0$.

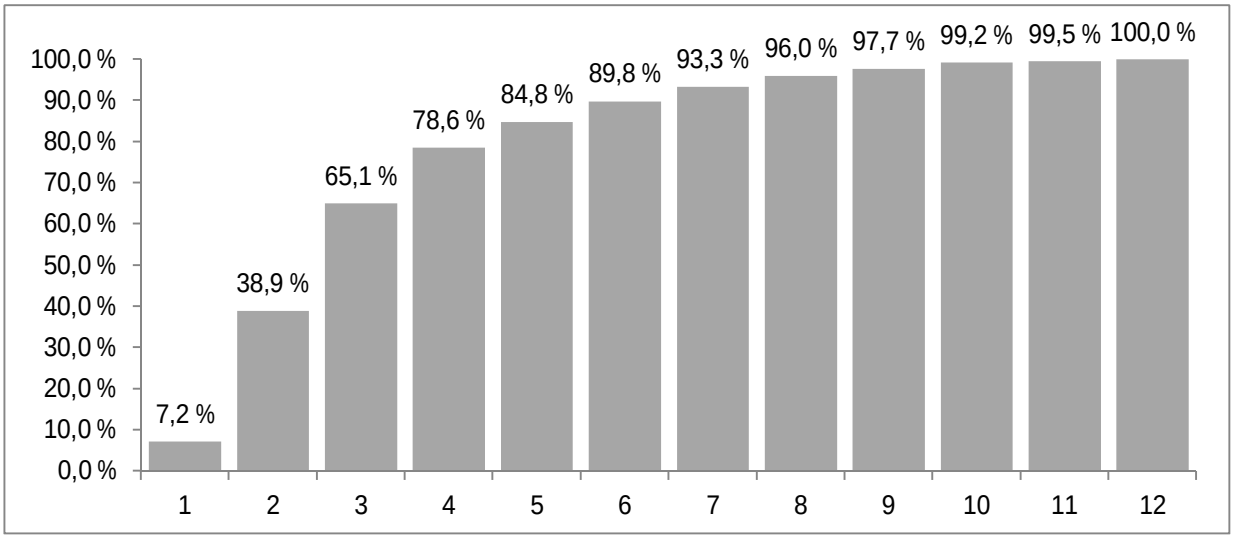
Parametrien estimointi on toteutettu SAS-ohjelmiston (versio 9.2) LOGISTIC-proseduurilla. GENMOD-proseduuri tuottaisi samat parametriestimaatit, mutta LOGISTIC-proseduurin etuna ovat sen logistiselle regressiomallille ominaiset tulosteet. Koska vastemuuttuja on harvinainen eläketodennäköisyyden ollessa mallinnusaineistossa 0,6 prosenttia, on LOGISTIC-proseduurissa käytetty FIRTH-optiota. Tällöin estimointi tehdään tavanomaisen suurimman uskottavuuden (maximum likelihood) sijaan niin sanotulla penalized likelihood – menetelmällä, jolloin vasteen harvinaisuudesta johtuvaa parametriestimaattien harhaa voidaan pienentää. [1], [9], [14]

5.2 Mallinnusaineisto

Kuvassa 1 on esitetty työtapaturmavahinkojen tapauskohtaisten eläkevarausten tietoontuloviive. Kuvasta nähdään, että viidessä vuodessa varataan noin 85 prosenttia työtapaturmaeläkkeiden kappaleista. Toisaalta kuluu noin 12 vuotta siihen, kunnes kaikki eläketapaukset on varattu. Koska ennustemallin tavoitteena ei ole tunnistaa vuosikausien viiveellä ilmeneviä eläkkeitä, pidetään viiden vuoden aikajaksoa riittävänä tunnistamaan suurin osa tapauksista.

Mallinnusaineistoon on valittu ne Pohjola Vakuutuksen ja A-vakuutuksen lakisääteisen tapaturmavakuutuksen vuosina 2000–2008 sattuneet työtapaturmavahingot, jotka ovat johtaneet vahingoittuneen työkyvyttömyyteen eli joista on korvattu päivärahaa. Tällöin tuoreimmiltakin sattumisvuosilta aineistossa on mukana viiden vuoden historia. Kahdeksan vuoden ikäisten vahinkojen mallinnusta varten aineisto on rajattu sattumisvuosiin 2000–2005 ja kymmenen vuoden ikäisten vahinkojen mallinnusta varten sattumisvuosiin 2000–2003. Vahingoista mallinnukseen on poimittu satunnaisesti 70 prosenttia ja loppua 30 prosentin osuutta on käytetty mallin sopivuuden ja luotettavuuden tarkasteluun. Sattumisvuosien 2000–2008 70 prosentin mallinnusaineistossa on mukana 110 559 vahinkoa, joista 650 eläketapausta.

Vuonna 2005 voimaantullut sairaanhoitokulujen korvaussääntöjen uudistus, TÄKY, näkyy aineistossa maksettujen ohimenevien korvausten tason nousuna vuodesta 2005 alkaen. Tämän ja korvausinflaation vuoksi maksetut ohimenevät korvaukset on korjattu sattumisvuosittain vuoden 2013 tasolle pitäen vuoden 2013 maksettujen korvausten mediaania perustasona.



Kuva 1. Eläkekappaleiden tietoontuloviive (vuosina).

Tarkastellaan esimerkkinä puolen vuoden ikäisten vahinkojen mallinnusaineiston muodostamista. Vuosina 2000–2008 sattuneille työtapaturmavahingoille haetaan asiakastiedoista vahingoittuneen syntymäaika ja siitä lasketaan vahingoittuneen ikä vahingon sattumishetkellä. Korvausaineistosta vahingoille haetaan tieto siitä, kuinka monta vuorokautta kustakin vahingosta on korvattu päivärahaa kuuden kuukauden ajalla vahingon sattumisesta. Päivärahajaksoissa huomioidaan myös maksupäivä siten, että mukaan otetaan ainoastaan ne päivärahajaksot, jotka on maksettu ennen kuin vahingon sattumisesta on kulunut puoli vuotta. Muussa tapauksessa ei yhtiöllä puolivuotishetkellä olisi tiedossa, että vahingoittunut on ollut työkyvyttömänä, vaikka todellisuudessa näin olisi ollutkin. Lisäksi aineistoon otetaan mukaan vahingosta puolen vuoden aikana maksetut muut ohimenevät korvaukset kuin päiväraha ja tieto siitä, kuinka monta vuorokautta on puolivuotispäivänä kulunut viimeisestä korvaussuorituksesta tai päivärahapäivästä, toisin sanoen kuinka kauan vahinko on ollut hiljaiselossa. Vastemuuttujaksi haetaan tieto siitä, löytyykö vahingolle varausaineistosta vahinkokohtaista työtapaturmaeläkkeen varausta. Tässä ei tehdä rajausta puolivuotispäivän eikä muunkaan ajanjakson suhteen.

5.3 Tulokset

Ennustemallin lähtökohtana oli muodostaa suhteellisen yksinkertainen malli, joka kuitenkin riittävän tarkasti pystyisi tunnistamaan potentiaaliset eläketapaukset. Alussa mahdollisia muita selittäviä tekijöitä malliin lopulta valittujen lisäksi olivat muun muassa jaottelu työ- ja vapaa-ajan vahinkoihin, työnlaatu (rakennus, liikenne, toimisto ja muut), vahingoittuneen sukupuoli ja työkyvyttömyysaste. Muut selittävät tekijät eivät kuitenkaan osoittautuneet tilastollisesti merkitseviksi 95 prosentin luottamustasolla, toisin sanoen niiden parametriestimaatteihin liittyvät p-arvot olivat suurempia kuin 0,05 ja odds ratio -estimaattien luottamusvälit sisälsivät arvon 1.

Kaikissa malleissa eläkkeen todennäköisyyttä selittää parhaiten työkyvyttömyyskauden pituus: sekä päivärahakauden pituus että päivärahakauden jälkeinen työkyvyttömyys. Työkyvyttömyyskauden kasvaessa eläkkeen todennäköisyys kasvaa, kun muut tekijät pysyvät muuttumattomina. Samansuuntainen vaikutus,

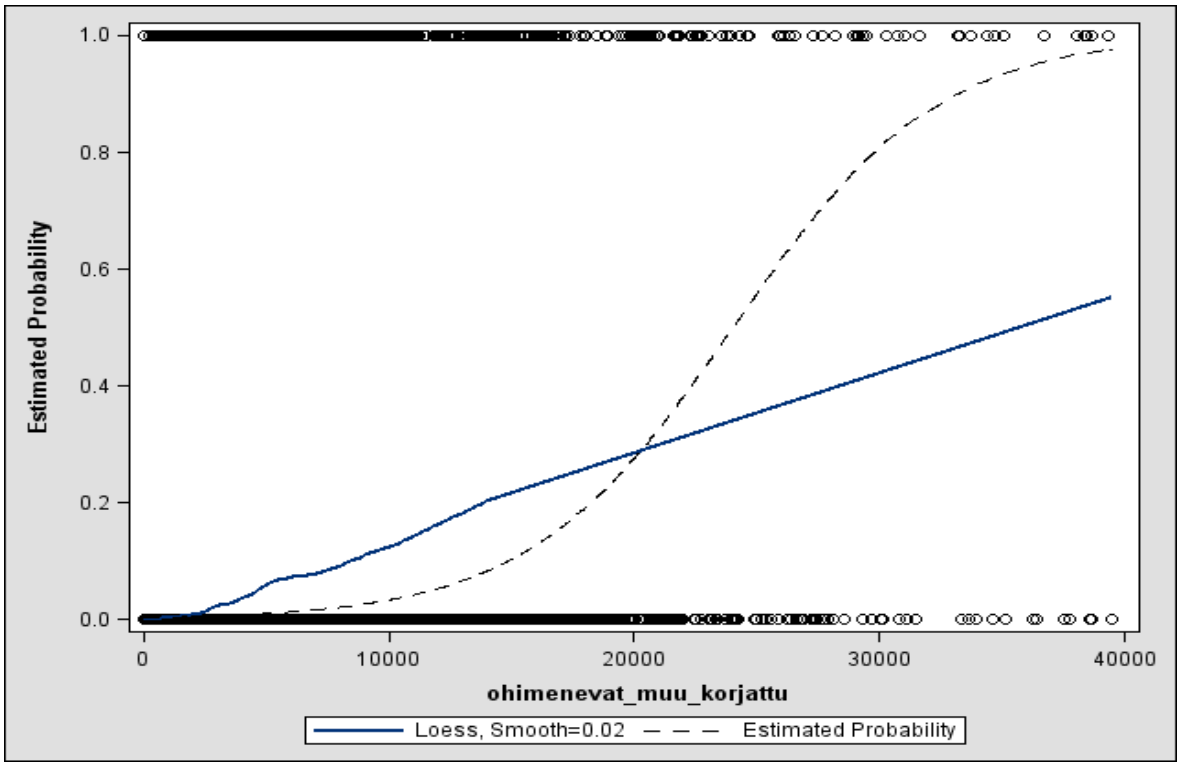
mutta lievempi, on maksetuilla muilla ohimenevillä korvauksilla ja vahingoittuneen iällä. Hiljaiselolla on päinvastainen vaikutus: mitä pidempään vahinko on ollut hiljaiselossa, sitä pienemmäksi eläkkeen todennäköisyys muuttuu.

Selittävien tekijöiden väliset korrelaatiot vaihtelevat hieman sen mukaan, minkä ikäisistä vahingoista on kyse. Vahingon iästä riippumatta korrelaatio on pienintä vahingoittuneen iän ja muiden selittävien tekijöiden välillä. Esimerkiksi vahingoittuneen iän ja päivärahakauden pituuden välinen Pearsonin korrelaatiokerroin on noin 0,11 ja vahingoittuneen iän ja päivärahakauden jälkeisen työkyvyttömyyskauden pituuden välinen korrelaatio noin 0,04. Suurinta korrelaatio on hiljaiselon ja päivärahakauden, hiljaiselon ja työkyvyttömyyskauden sekä päivärahakauden ja sen jälkeisen työkyvyttömyyskauden välillä. Suurimmillaan korrelaatiokerroin on noin $\pm 0{,}7$. Eri tekijöiden VIF-arvot vaihtelevat reilusta yhdestä hieman yli kahteen.

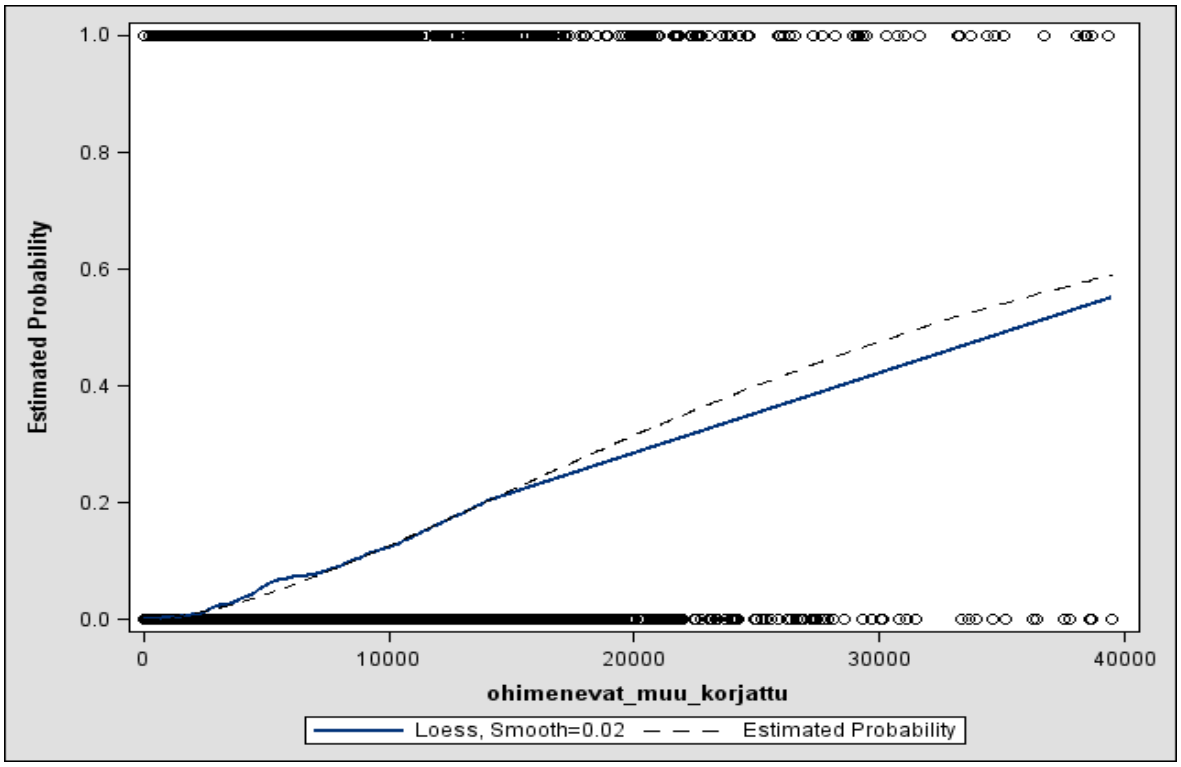
Tarkasteltaessa selittävien tekijöiden vaikutusta eläkkeen todennäköisyyteen ilmeni, että päivärahakauden pituus toimii malleissa sellaisenaan, ilman muunnosta. Vahingoittuneen ikää paremmin toimii sen logaritmimuunnos. Myös työkyvyttömyyskauden pituudessa ja muissa ohimenevissä korvauksissa logaritmimuunnos toimii alkuperäistä muuttujaa paremmin. Koska työkyvyttömyyskauden pituus ja muut ohimenevät korvaukset voivat kuitenkin saada arvokseen myös 0, päädyttiin tekemään muotoa $\log(1 + x)$ oleva muunnos, jolloin muuttujan saadessa arvon 0, myös muunnettu muuttuja on 0. Hiljaiselon ja eläkkeen todennäköisyyden välinen riippuvuus osoittautui olevan joko ensimmäistä tai toista astetta riippuen siitä, minkä ikäisten vahinkojen mallista on kyse.

Vastemuuttujan ja selittävien tekijöiden välisen yhteyden muotoa tutkittaessa voidaan logistisen regressiomallin tuottamia estimoituja todennäköisyyksiä verrata esimerkiksi SAS-ohjelmiston LOESS-proseduurilla tuotettuihin estimaatteihin. Loess (locally weighted scatterplot smoother) on parametriton tasoitusmenetelmä. Tasoitusparametrilla voidaan säädellä tasoituksen sileyttä: mitä pienempi parametrin arvo, sitä tarkemmin estimaatti mukailee aineistoa. Kuvassa 2 on esitetty loess-tasoituksen ja logistisen regression tuottamat estimaatit muiden ohimenevien korvausten vaikutuksesta eläkkeen todennäköisyyteen vuoden ikäisten vahinkojen mallissa, kun muut ohimenevät korvaukset on mallissa ainoa selittävä tekijä. Kuvassa 3 on vastaava tarkastelu, mutta siinä muille ohimeneville korvauksille on tehty muotoa $\log(1 + x)$ oleva muunnos ennen logistisen regression sovittamista. Nähdään, että logaritmoitu muoto kuvaa korvausten ja eläketodennäköisyyden välistä yhteyttä alkuperäistä muuttujaa huomattavasti paremmin. Kuvissa muut ohimenevät korvaukset on katkaistu 40 000 euroon, koska suurin osa tapauksista on sitä pienempiä. [14]

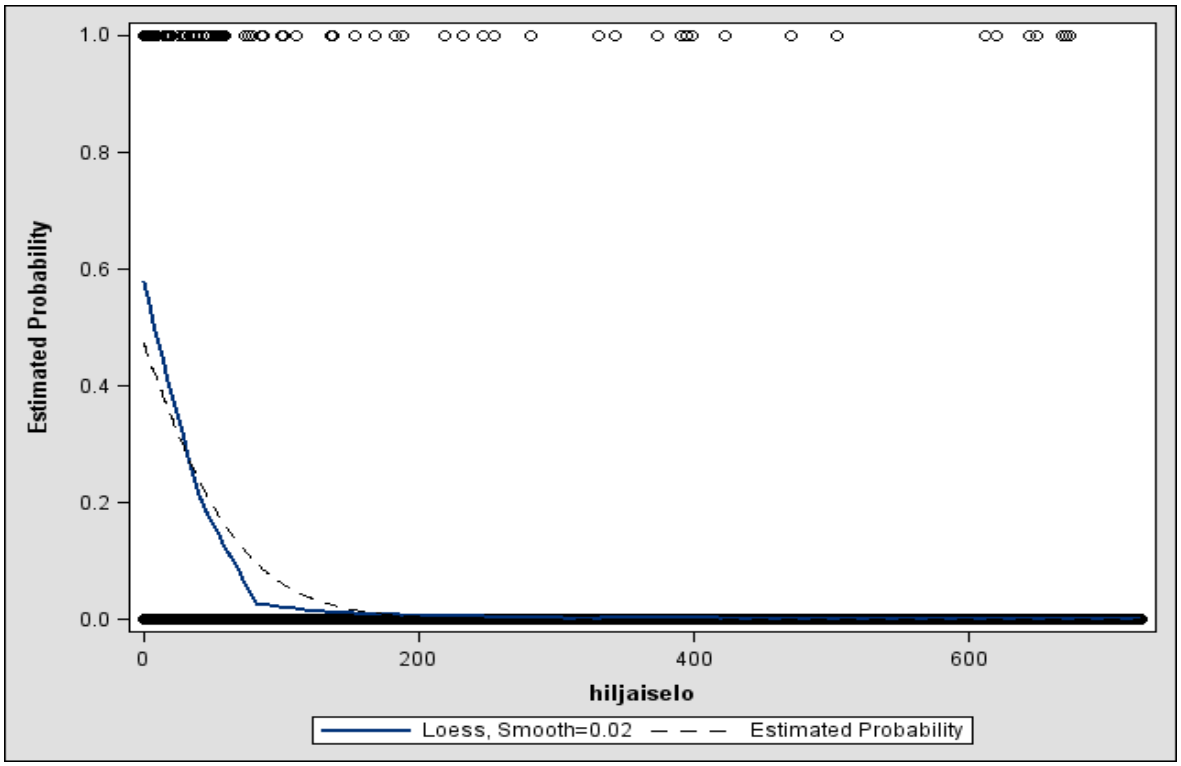
Kuvissa 4 ja 5 on hiljaiselon (vuorokausina) vaikutus eläkkeen todennäköisyyteen kahden vuoden ikäisillä vahingoilla loess-tasoituksella ja logistisella regressiolla estimoituna, kun hiljaiselo on mallin ainoa selittävä tekijä. Kuvassa 4 mallimuodon oletetaan olevan ensimmäistä astetta ja kuvassa 5 toista astetta. Toista astetta oleva muoto sopii aineistoon ensimmäistä astetta paremmin.



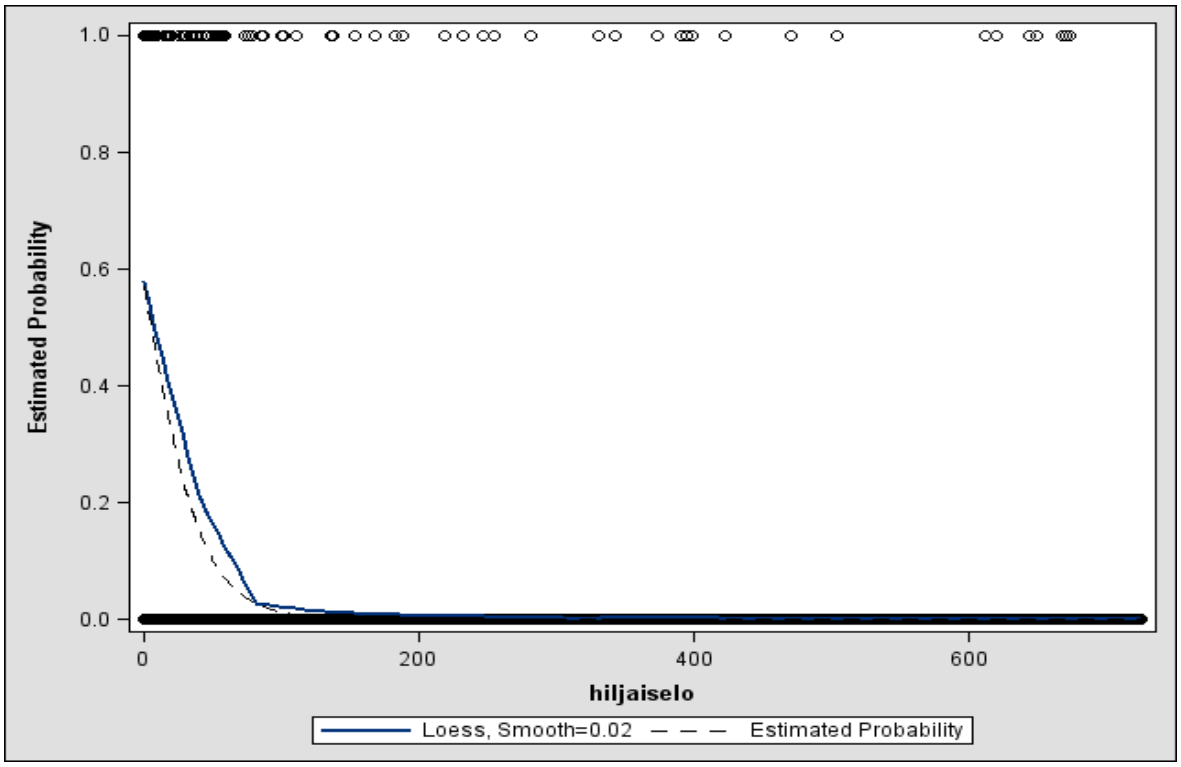
Kuva 2. Muiden ohimenevien korvausten vaikutus eläkkeen todennäköisyyteen loess-tasoituksella ja logistisella regressiolla estimoituna.



Kuva 3. Muiden ohimenevien korvausten vaikutus eläkkeen todennäköisyyteen loess-tasoituksella ja logistisella regressiolla estimoituna, kun logistisessa regressiossa käytetään logaritmoituja korvauksia.



Kuva 4. Hiljaiselon (vrk) vaikutus eläkkeen todennäköisyyteen loess-tasoituksella ja logistisella regressiolla estimoituna, kun vaikutuksen oletetaan olevan ensimmäistä astetta.



Kuva 5. Hiljaiselon (vrk) vaikutus eläkkeen todennäköisyyteen loess-tasoituksella ja logistisella regressiolla estimoituna, kun vaikutuksen oletetaan olevan toista astetta.

Taulukossa 1 on esitetty odds ratio – estimaatit eri-ikäisten vahinkojen malleissa. Esimerkiksi vuoden ikäisten vahinkojen mallissa päivärahakauden kasvaessa yhdellä vuorokaudella muiden tekijöiden pysyessä muuttumattomina kasvaa eläkkeen todennäköisyys suhteessa ei-eläkkeen todennäköisyyteen (odds) noin 1,4 prosenttia. Logaritmoitujen muuttujien osalta odds ratio – estimaatit kertovat, kuinka paljon eläkkeen todennäköisyys suhteessa ei-eläkkeen todennäköisyyteen muuttuu, kun alkuperäinen muuttuja (ilman logaritmimuunnosta) kasvaa noin 2,7-kertaiseksi (Neperin luku $e \approx 2{,}7$). Näin ollen esimerkiksi vahingoittuneen iän kasvaessa noin 2,7-kertaiseksi muiden tekijöiden pysyessä muuttumattomina kasvaa eläkkeen todennäköisyys suhteessa ei-eläkkeen todennäköisyyteen noin nelinkertaiseksi vuoden ikäisten vahinkojen mallissa.

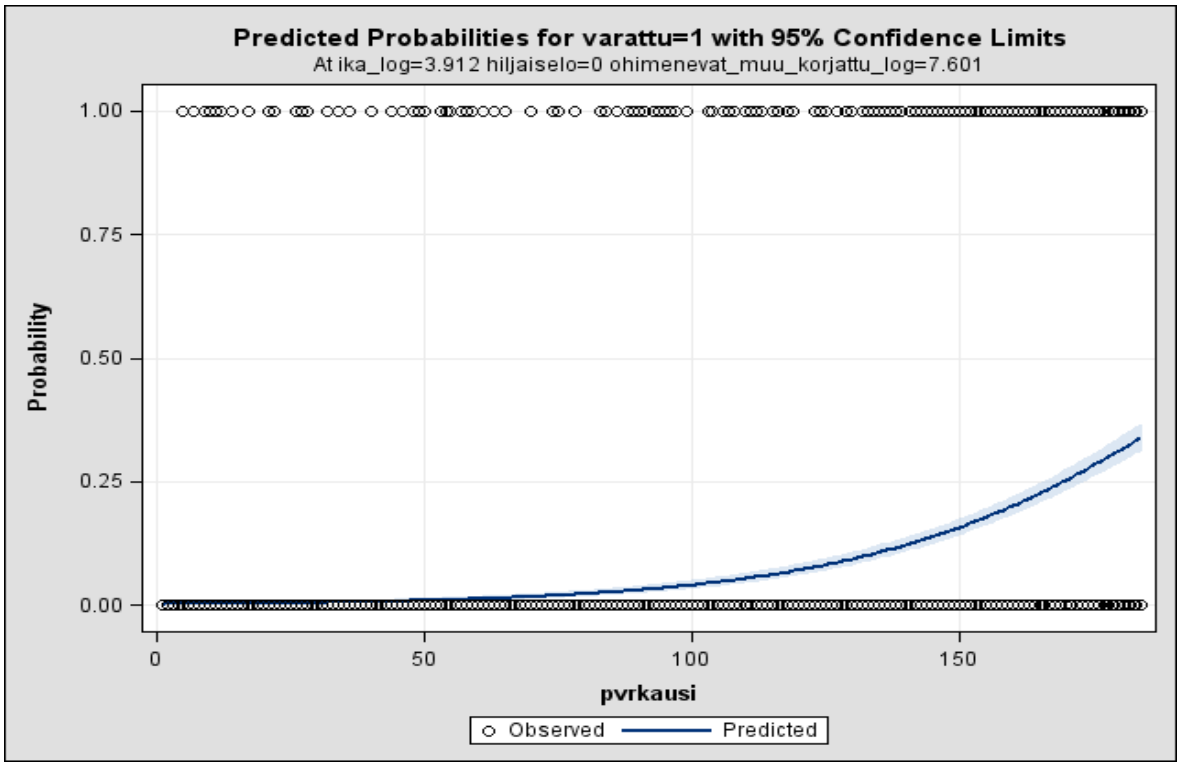
Taulukossa 1 hiljaiselon odds ratio – estimaatit puuttuvat niistä malleista, joissa hiljaiselon ja eläkkeen todennäköisyyden välinen riippuvuus on toista astetta, koska tällöin vaikutus riippuu edelleen myös hiljaiselon saamasta arvosta. Estimaatteja tarkasteltaessa on hyvä myös huomioida, että mallin selittävät tekijät ovat eri suureita, joten esimerkiksi sillä, että muut ohimenevät korvaukset kasvavat eurolla, on pienempi vaikutus eläkkeen todennäköisyyteen kuin sillä, että päivärahakausi kasvaa vuorokaudella.

	3 kk	6 kk	12 kk	1 v 6 kk	2 v	3 v	5 v
Päivärahakauden pituus	1,048	1,030	1,014	1,009	1,009	1,009	1,008
Muut ohimenevät korvaukset	1,252	1,185	1,190	1,172	1,120	1,163	1,037
Vahingoittuneen ikä	4,774	3,303	4,011	5,613	7,435	12,588	20,129
Hiljaiselo	0,970	0,978					
Työkyvyttömyyskauden pituus	-	-	-	1,360	1,387	1,378	1,431

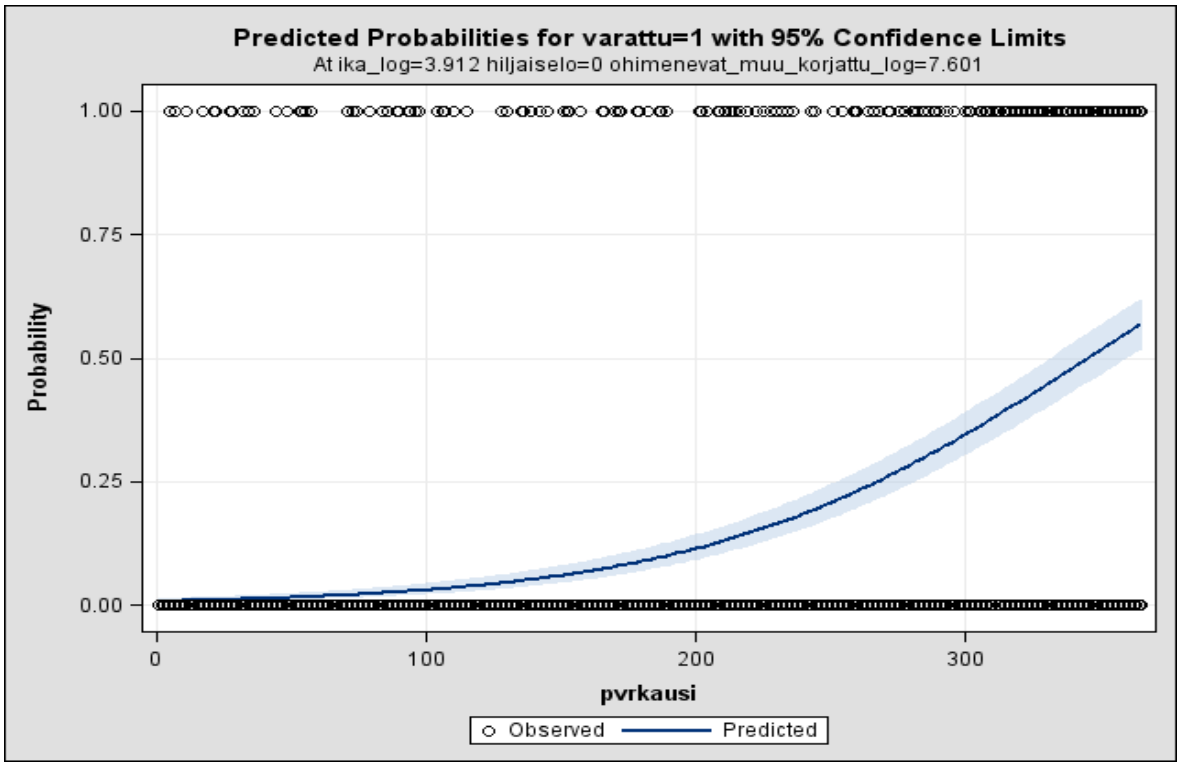
Taulukko 1. Selittävien muuttujien odds ratio – estimaatit eri-ikäisten vahinkojen malleissa.

Kun mallissa on mukana useita selittäviä tekijöitä, täytyy muiden tekijöiden arvot kiinnittää, jotta voidaan tarkastella vuorollaan kunkin tekijän vaikutusta eläkkeen todennäköisyyteen. Kuvassa 6 on esitetty päivärahakauden pituuden vaikutus eläkkeen todennäköisyyteen puolen vuoden ikäisten vahinkojen mallissa ja kuvassa 7 vuoden ikäisten vahinkojen mallissa. Molemmissa kuvissa maksetut muut ohimenevät korvaukset on kiinnitetty 2000 euroon ja vahingon oletetaan olevan aktiivinen, toisin sanoen hiljaiselon oletetaan olevan 0 vuorokautta. Vahingoittuneen iän oletetaan vahinkohetkellä olevan 50 vuotta.

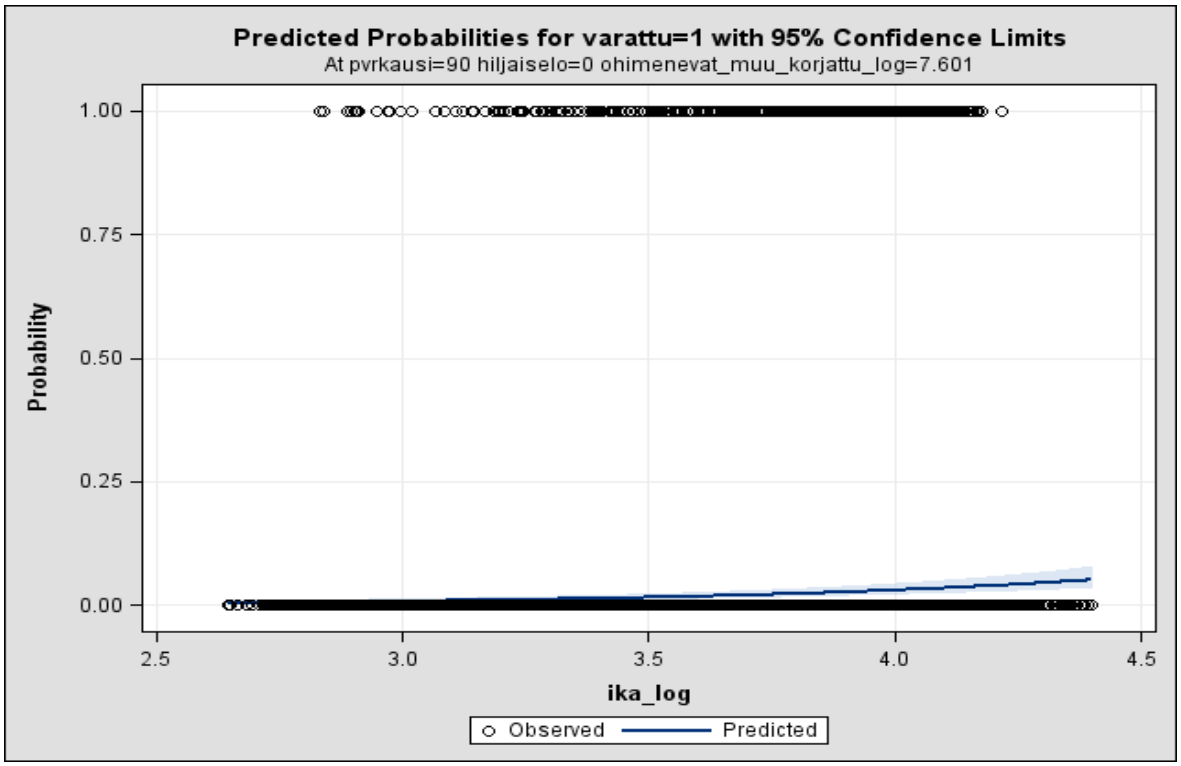
Kuvassa 8 on esitetty (logaritmoidun) vahingoittuneen iän vaikutus eläkkeen todennäköisyyteen vuoden ikäisten vahinkojen mallissa. Päivärahakauden pituudeksi on kiinnitetty 90 vuorokautta (noin kolme kuukautta), maksettujen muiden ohimenevien korvausten määräksi 2000 euroa ja hiljaiselon oletetaan olevan 0 vuorokautta. Kuvassa 9 on vastaava vaikutus, kun päivärahakauden pituuden oletetaan olevan vuosi. Iän vaihteluväli aineistossa on 14 vuodesta 81 vuoteen eli logaritmisella asteikolla 2,6:sta 4,4:ään. Iän keskiarvo on 41 vuotta ja mediaani 42 vuotta, logaritmisella asteikolla noin 3,7.



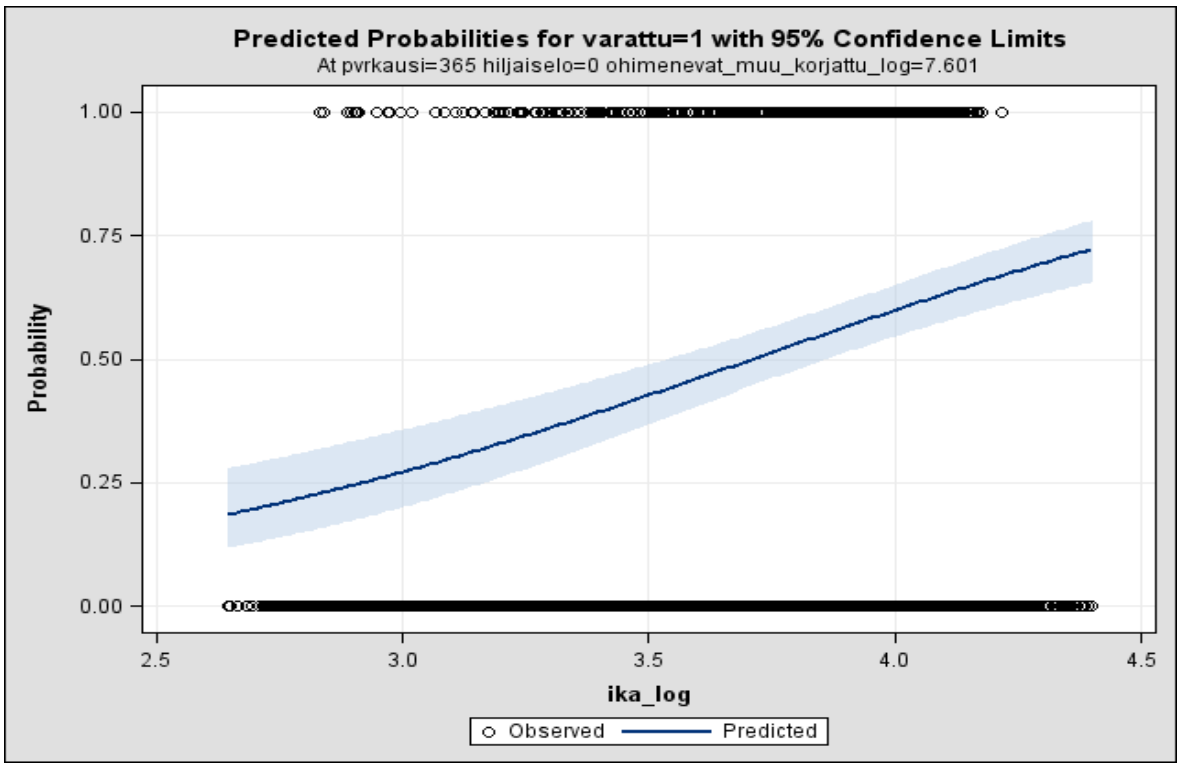
Kuva 6. Päivärahakauden pituuden (vrk) vaikutus eläkkeen todennäköisyyteen kuuden kuukauden ikäisten vahinkojen mallissa.



Kuva 7. Päivärahakauden pituuden (vrk) vaikutus eläkkeen todennäköisyyteen vuoden ikäisten vahinkojen mallissa.



Kuva 8. Vahingoittuneen iän (logaritmin) vaikutus eläkkeen todennäköisyyteen vuoden ikäisten vahinkojen mallissa, kun päivärahaa on korvattu 90 vuorokautta.



Kuva 9. Vahingoittuneen iän (logaritmin) vaikutus eläkkeen todennäköisyyteen vuoden ikäisten vahinkojen mallissa, kun päivärahaa on korvattu 365 vuorokautta.

5.3.1 Esimerkkitapauksia

Tarkastellaan esimerkkinä heinäkuussa 2012 sattunutta työtapaturmavahinkoa, jonka seurauksena 59-vuotiaan mekaanikon oikea ranne murtui. Ensimmäiset kolme kuukautta vahingoittunut oli työkyvytön ja päivärahan lisäksi vahingosta maksettiin muuta ohimenevää korvausta noin 4 800 euroa. Mallin antama eläkkeen todennäköisyys kolmen kuukauden kohdalla on 44 prosenttia. Kuuden kuukauden kuluttua vahingon sattumisesta vahingoittunut oli edelleen ollut koko ajan työkyvytön ja muuta ohimenevää korvausta oli maksettu tähän mennessä noin 7 000 euroa. Vahinko oli siis ollut aktiivisena koko ajan. Mallin antama eläkkeen todennäköisyys puolen vuoden kohdalla on 53 prosenttia. Vuoden kuluttua vahingon sattumisesta malli antaa eläkkeen todennäköisyydeksi 60 prosenttia. Tällöin päivärahaa oli korvattu 349 vuorokautta eli lähes vuosi, muuta ohimenevää korvausta oli maksettu reilut 12 700 euroa ja vahinko oli ollut edelleen aktiivisena. Puolentoista vuoden kohdalla mallin antama todennäköisyys on 82 prosenttia. Tällöin vahingoittunut oli ollut työkyvyttömänä myös päivärahakauden jälkeisen ajan ja muuta ohimenevää korvausta oli maksettu tähän mennessä yhteensä reilut 13 600 euroa. Vahingosta tehtiin 100 prosentin työkyvyn aleneman mukainen tapauskohtainen arviovaraus maaliskuussa 2014, jolloin siis oli kulunut hieman yli puolitoista vuotta vahingon sattumisesta. Voidaan sanoa, että eläke-ennustemalli tunnisti vahingon eläketapaukseksi jo varhaisessa vaiheessa.

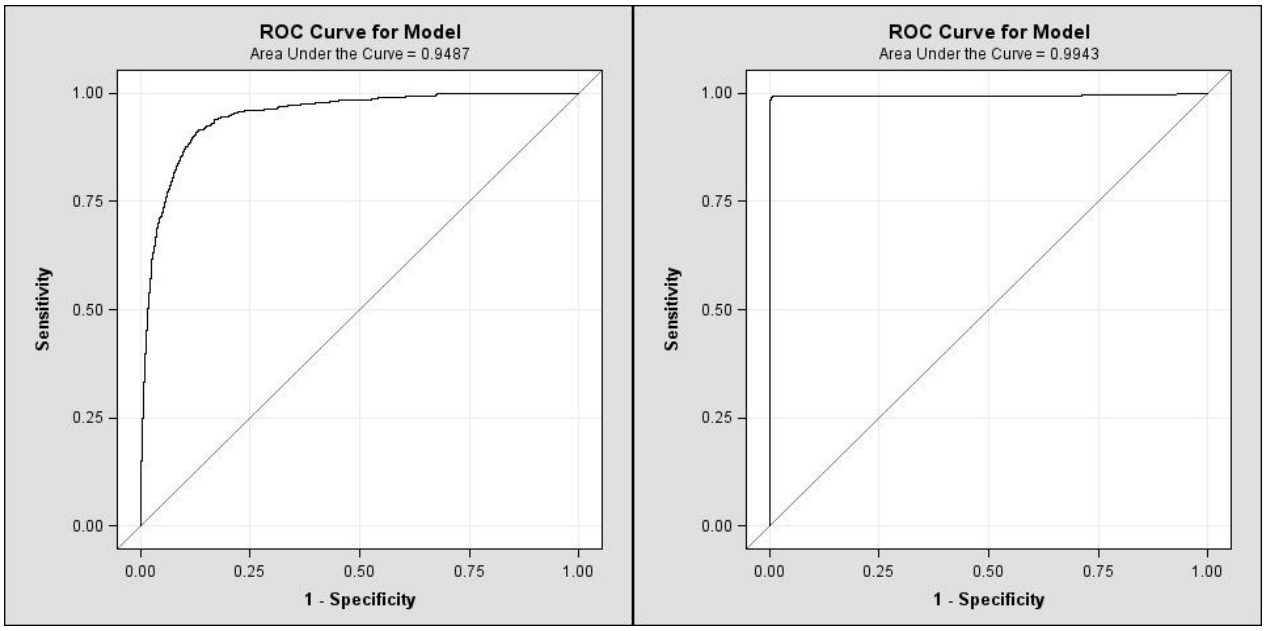
Niissä tapauksissa, joissa ennustemalli ei useankaan vuoden kuluttua tunnista vahinkoa eläkkeeksi, on yleensä viivettä korvausten raportoitumisessa, vaikka vahingoittunut on todellisuudessa ollut työkyvytön. Kyse on useimmiten poikkeustapauksista, joissa vahingon selvittely syystä tai toisesta kestää tavanomaista pidempään.

Mallin luokittelukyvyn havainnollistamiseksi tarkastellaan esimerkkinä vuoden ikäisten vahinkojen mallia. Luokitellaan vahinko eläkkeeksi, jos mallin antama todennäköisyys on vähintään 50 prosenttia ja ei-eläkkeeksi tätä pienemmillä todennäköisyyksillä. Vuosina 2000–2008 sattuneista 223 108 työtapaturmasta eläketapauksia on 1 304. Eläketapauksista malli luokittelee oikein 50,8 prosenttia ja väärin 49,2 prosenttia vuoden kuluttua vahingon sattumisesta. Niistä vahingoista, jotka eivät ole eläketapauksia, malli luokittelee oikein 99,8 prosenttia ja väärin 0,2 prosenttia. Kahden vuoden ikäisten vahinkojen mallissa vastaavat osuudet ovat eläketapauksilla 75,4 prosenttia ja 24,6 prosenttia ja ei-eläketapauksilla 99,9 prosenttia ja 0,1 prosenttia. Mikäli luokittelun rajana käytettävää todennäköisyyttä lasketaan 50 prosentista, kasvaa oikein luokiteltujen eläketapausten osuus, mutta oikein luokiteltujen ei-eläkkeiden osuus puolestaan pienenee. Mikäli rajana käytettävää todennäköisyyttä taas nostetaan, tapahtuu päinvastoin.

5.4 Mallin validointi

Eläke-ennustemalleissa mallin tarkkuutta kuvaavan AUC-tunnusluvun arvot kasvavat siirryttäessä nuorten vahinkojen malleista keski-ikäisten ja vanhojen vahinkojen malleihin. Nuorten vahinkojen malleissa AUC-tunnuslukujen arvot ovat välillä 0,949–0,992, keski-ikäisten vahinkojen malleissa välillä 0,991–0,995 ja vanhojen vahinkojen malleissa välillä 0,991–1,000. Tunnusluvun arvot ovat siis hyvin lähellä ykköstä jo nuorten vahinkojen malleissa, mutta tämä johtuu osittain vasteen harvinaisuudesta.

Kuvassa 10 on vasemmalla kolmen kuukauden ja oikealla viiden vuoden ikäisten vahinkojen mallien tuottamat ROC-käyrät. AUC-tunnusluvut saavat arvot 0,949 ja 0,994.



Kuva 10. Kolmen kuukauden ja viiden vuoden ikäisten vahinkojen mallien tuottamat ROC-käyrät.

Myös F_1-mitan arvot kasvavat siirryttäessä nuorten vahinkojen malleista vanhojen vahinkojen malleihin. Kolmen kuukauden ikäisten vahinkojen mallissa F_1 on suurimmillaan 0,25, puolen vuoden ikäisten vahinkojen mallissa 0,43 ja vuoden ikäisten vahinkojen mallissa 0,64 kasvaen siitä edelleen siten, että esimerkiksi kahden vuoden ikäisten vahinkojen mallissa F_1 on suurimmillaan 0,81 ja viiden vuoden ikäisten vahinkojen mallissa 0,93. F_1-mitan arvot kuvaavat AUC-tunnusluvun arvoja selkeämmin eroja eri-ikäisten vahinkojen mallien luokittelukyvyissä.

Taulukossa 2 on esitetty 30 prosentin testiaineistosta estimoidut odds ratio – estimaatit. Estimaatit ovat hyvin samansuuntaisia kuin varsinaisesta 70 prosentin mallinnusaineistosta estimoidut, joitain erojakin tosin löytyy. Vahingoittuneen iän estimaatit poikkeavat toisistaan jonkin verran eri aineistoista estimoituna. Muut ohimenevät korvaukset on tilastollisesti merkitsevä tekijä 30 prosentin aineistossa ainoastaan kolmen ja kuuden kuukauden ikäisten vahinkojen malleissa. Tämän vuoksi estimaatin luottamusvälit ovat leveät varsinkin vanhempien vahinkojen malleissa ja odds ratio on jopa alle yhden viiden vuoden ikäisten vahinkojen mallissa. AUC-tunnusluvun ja F_1-mitan arvot ovat lähellä 70 prosentin aineistosta laskettuja.

	3 kk	6 kk	12 kk	1 v 6 kk	2 v	3 v	5 v
Päivärahakauden pituus	1,050	1,029	1,015	1,011	1,009	1,008	1,007
Muut ohimenevät korvaukset	1,350	1,246	1,128	1,050	1,050	1,006	0,939
Vahingoittuneen ikä	6,295	4,205	4,321	6,135	8,117	8,928	11,895
Hiljaiselo	0,988	0,982					
Työkyvyttömyyskauden pituus	-	-	-	1,255	1,348	1,383	1,447

Taulukko 2. Selittävien muuttujien odds ratio – estimaatit eri-ikäisten vahinkojen malleissa 30 prosentin testiaineistosta estimoituna.

Ennustemalliin mukaan otettavia tekijöitä valittaessa on huomioitava sekä käytännön rajoitukset että mallinnuksesta saatava informaatio. Alussa mahdollisista selittävistä tekijöistä karsittiin pois ne, jotka eivät osoittautuneet tilastollisesti merkitseviksi ja joiden parametriestimaatteihin liittyvät luottamusvälit olivat leveät. Mukaan otettujen tekijöiden välisiä riippuvuuksia tutkittiin korrelaatioiden perusteella ja mallin ja aineiston yhteensopivuutta eri mallimuodoilla AIC-arvojen avulla.

Varsinaisen mallinnusaineiston ja 30 prosentin testiaineiston perusteella tarkimmin eläkkeen todennäköisyyttä ennustavat päiväraha- ja työkyvyttömyyskauden pituus sekä hiljaiselo. Esimerkiksi vuoden ikäisten vahinkojen mallissa päivärahakauden odds ratio – estimaatin 95 prosentin luottamusväli 70 prosentin aineistosta estimoituna on [1,013, 1,016] ja 30 prosentin aineistosta estimoituna [1,014, 1,016]. Muut ohimenevät korvaukset ei ole tilastollisesti merkitsevä tekijä kaikissa malleissa ja sen odds ratio -estimaattien luottamusvälit levenevät, kun siirrytään nuorten vahinkojen malleista vanhempien vahinkojen malleihin. Esimerkiksi kolmen kuukauden ikäisten vahinkojen mallissa muiden ohimenevien korvausten odds ratio – estimaatin 95 prosentin luottamusväli on 70 prosentin aineistosta estimoituna [1,180, 1,328] ja kahden vuoden ikäisten vahinkojen mallissa [0,979, 1,281]. Myös vahingoittuneen iän odds ratio -estimaatteihin liittyvät luottamusvälit ovat kaikissa malleissa suhteellisen leveät ja sitä leveämmät mitä vanhempien vahinkojen mallista on kyse. Esimerkiksi puolen vuoden ikäisten vahinkojen mallissa vahingoittuneen iän odds ratio – estimaatin 95 prosentin luottamusväli on 70 prosentin aineistosta estimoituna [2,310, 4,723] ja puolentoista vuoden ikäisten vahinkojen mallissa [3,569, 8,827]. Lisäksi 70 prosentin ja 30 prosentin aineistoista saaduissa estimaateissa on eroja, kuten edellä on todettu.

Parametriestimaatteihin liittyvän epävarmuuden perusteella vahingoittuneen iän ja muut ohimenevät korvaukset voisi jättää pois ainakin yksittäisistä malleista, vahingoittuneen iän mahdollisesti jopa kaikista malleista. Molemmilla tekijöillä kuitenkin on vaikutusta eläkkeen todennäköisyyteen ja muilla ohimenevillä korvauksilla erityisesti eläketapausten tunnistamisen kannalta oleellisimmissa nuorten vahinkojen malleissa. Koska lisäksi on toivottavaa, että eri-ikäisten vahinkojen mallit ovat muodoltaan mahdollisimman samanlaisia, jotta eläketodennäköisyydet eivät muutu liikaa mallista toiseen siirryttäessä, kannattaa epävarmuutta sisältävät tekijät joko pitää mukana kaikissa malleissa tai jättää kokonaan pois.

6 Yhteenveto

Lakisääteinen tapaturmavakuutus on lajina pitkähäntäinen, sillä vakavista pysyvää työkyvyttömyyttä aiheuttavista vahingoista syntyy vakuutusyhtiölle pisimmillään vuosikymmeniä kestävä korvausvelvollisuus. Työtapaturmaeläkkeitä varten tehtävillä vahinkokohtaisilla varauksilla on huomattava merkitys myös suuremmille vakuutuksenottajille, koska heillä varaukset vaikuttavat korvausmenona vakuutusmaksuun. Muutokset tapauskohtaisten varausten suuruuksissa sekä uudet ja päättyneet varaukset kirjataan tilinpäätöksessä yhtiön tulokseen korvausvastuun muutoksena. Varausten tekeminen oikea-aikaisesti ja oikean suuruisena on näin ollen tärkeää.

Useat tekijät, kuten vahinkojen selvittelystä johtuvat viiveet ja korvaustoiminnan prosessit, aiheuttavat heiluntaa tunnettujen vahinkojen eläkevastuuseen. Tilastollisella eläkkeiden ennustamisella tapauskohtaisesti varattavat eläkkeet pyritään tunnistamaan aiempaa nopeammin ja tasaisemmin

hyödyntämällä vahingoista olemassa olevaa tietoa. Esimerkiksi työkyvyttömyysaikaa kuvaavat tekijät, kuten päivärahakauden pituus, ennustavat eläkkeen todennäköisyyttä hyvin.

Oman haasteensa mallinnukseen tuo vastetapahtuman harvinaisuus, sillä mallinnusaineiston työtapaturmista alle prosentti päätyy eläkkeeksi. Aineistossa on kuitenkin yli 100 000 vahinkoa ja eläketapauksia näin ollen useita satoja, joten havaintoaineiston koon ollessa riittävä vasteen harvinaisuus ei muodosta todellista ongelmaa ja logistista regressiomallia voidaan käyttää.

Koska eri-ikäisistä vahingoista on mallinnusaineistossa eri määrä informaatiota, muodostetaan tietyn ikäisille vahingoille omat mallit. Mallimuotoja ja mukaan otettavia tekijöitä valittaessa pyritään siihen, että eri-ikäisten vahinkojen mallit eivät poikkea toisistaan liikaa ja mallista toiseen voidaan siirtyä interpoloimalla parametrien arvot muodostettujen mallien välillä. Malliparametrit riippuvat kuitenkin korvausaineiston luonteesta ja tilanteesta tietyllä hetkellä, joten parametreja ei voi sellaisenaan soveltaa toisen yhtiön vahinkoaineistoon. Muun muassa vahinkojen selviämisnopeudet ja korvaustoiminnan prosessit eroavat yhtiöiden välillä.

Eläke-ennustemallin tuottamia vahinkokohtaisia todennäköisyyksiä voidaan hyödyntää kollektiivisen korvausvastuun laskennassa. Todennäköisyyksistä saadaan eläkkeiden lukumääräennusteet ja todennäköisille eläketapauksille voidaan muodostaa tilastollinen tapauskohtainen varaus vahingon vakavuutta ja vahingoittunutta koskevien tietojen, kuten työkyvyttömyysasteen ja vuosityöansion, perusteella. Lisäksi vahinkokohtaisia eläketodennäköisyyksiä voidaan käyttää erikoismaksujärjestelmiin kuuluvien asiakkaiden vahinkojen seurannassa.

Lähteet

[1] Allison, P. (2012): Logistic Regression for Rare Events. Statistical Horizons. Viitattu 2.9.2014. http://www.statisticalhorizons.com/logistic-regression-for-rare-events

[2] de Jong, P., Heller, G.Z. (2008): Generalized Linear Models for Insurance Data. Cambridge University Press.

[3] Dobson, A.J., Barnett A.G. (2008): An Introduction to Generalized Linear Models, Third Edition. Chapman & Hall/CRC.

[4] Downer, R.G., Richardson, P.J. (2009): Illustrative Logistic Regression Examples using PROC LOGISTIC: New Features is SAS/STAT 9.2. Paper SP03-2009. Viitattu 2.9.2014. http://www.lexjansen.com/pharmasug/2009/sp/sp03.pdf

[5] Heikkinen, J. (2005): Yleistetyt lineaariset mallit. Matematiikan ja tilastotieteen laitos, Helsingin yliopisto. Viitattu 2.9.2014. http://www.rni.helsinki.fi/~jmh/glm05/glm05.pdf

[6] Huang, C. (2013): Data analysis: When ROC fails logistic regression for rare-event data. Viitattu 2.9.2014. http://www.sasanalysis.com/2013/11/when-roc-fails-logistic-regression-for.html

[7] Karvanen, J. (2009): Generalized linear models. University of Helsinki, April 30, 2009. Viitattu 2.9.2014. http://wiki.helsinki.fi/download/attachments/35917349/lectures.pdf

[8] Kauppi, M. (2005): Lakisääteisen tapaturmavakuutuksen maksujen määräytymisestä. SHV-työ, 13.7.2005.

[9] King, G., Zeng, L. (2001): Logistic Regression in Rare Events Data. Political Analysis 9: 137-163. Viitattu 2.9.2014. http://gking.harvard.edu/files/gking/files/0s.pdf

[10] Kukkonen, S., Karmavalo, T. (2010): Työtapaturmakirja. Työtapaturmien ja ammattitautien korvaus- ja vakuutusasiat. 12. uudistettu painos. FINVA.

[11] McCullagh, P., Nelder, J.A. (1989): Generalized Linear Models, Second Edition. Chapman & Hall/CRC.

[12] Mellin, I. (2007): Tilastolliset menetelmät. Osa 4: Lineaarinen regressioanalyysi. Tilastollinen riippuvuus ja korrelaatio. TKK. Viitattu 2.9.2014. http://math.aalto.fi/opetus/sovtoda/luennot/TILRI100.pdf

[13] Ropponen, S. (2010): Kollektiivinen korvausvastuu. 14.9.2010, päivitetty 31.1.2013. Viitattu 2.9.2014. http://www.actuary.fi/fi/liitteet/koulutus/Kollektiivinen_korvausvastuu.pdf

[14] SAS/STAT(R) 9.2 User's Guide, Second Edition. Viitattu 2.9.2014. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#titlepage.htm

[15] Sosiaali- ja terveysministeriö (2013): Luonnos hallituksen esitykseksi työtapaturma- ja ammattitautilaiksi ja eräiksi siihen liittyviksi laeiksi. 4.11.2013. Viitattu 2.9.2014. http://www.stm.fi/c/document_library/get_file?folderId=6719234&name=DLFE-27809.pdf

[16] Tapaturmavakuutuslaitosten liiton (TVL) www-sivut. Viitattu 2.9.2014. http://www.tvl.fi/fi/

[17] Tapaturmavakuutuslaki (1948). 20.8.1948/608.