Tampereen yliopisto

Väinö Anttalainen

TIELIIKENNEONNETTOMUUKSIEN VAKAVUUTEEN VAIKUTTAVIEN TEKIJÖIDEN TUTKIMINEN LOGISTISELLA REGRESSIOLLA

Kandidaatintyö
Tekniikan ja luonnontieteiden tiedekunta
Tarkastaja: Jussi Kangas
Tammikuu 2023

TIIVISTELMÄ

Väinö Anttalainen: Tieliikenneonnettomuuksien vakavuuteen vaikuttavien tekijöiden tutkiminen logistisella regressiolla
Kandidaatintyö
Tampereen yliopisto
Teknis-luonnontieteellinen koulutusohjelma
Tammikuu 2023

Tieliikenneonnettomuudet aiheuttavat kärsimystä osallisille ja kustannuksia yhteiskunnalle. Koska liikenneturvallisuus vaihteilee maittain suuresti, on tärkeää löytää ne tekijät, jotka vaikuttavat onnettomuuksien vakavuuteen eniten juuri Suomessa. Tällöin liikenneturvallisuutta voidaan parantaa vaikuttavilla ratkaisuilla.

Onnettomuuksien vakavuus ilmoitetaan yleensä muutamalla kategorialla, kuten ei henkilövahinkoa, loukkaantuminen tai kuolema. Tällaisen muuttujan mallintamiseen käytetään yleensä logistista regressiomallia. Tässä työssä käytetään binääristä logistista regressiota vakavuuteen vaikuttavien tekijöiden tarkastelemiseen.

Aineisto sisältää Suomessa vuosina 2017–2021 sattuneita tieliikenneonnettomuuksia. Aluksi aineisto siistitään ja siitä muodostetaan kaksi erilaista mallia. Ensimmäisellä mallilla tarkastellaan henkilövahingon riskiin liittyviä tekijöitä. Toisella mallilla tarkastellaan kuoleman riskiin liittyviä tekijöitä. Mallit muodostetaan askeltavalla menetelmällä (stepwise method).

Luotujen mallien suorituskykyä arvioidaan kahdella eri tavalla. Mallien hyvyyttä tarkastellaan Hosmer–Lemeshow-testillä ja mallien erottelukykyä tarkastellaan ROC-kuvaajalla. Kumpikaan malli ei osoita merkkejä huonosta hyvyydestä, ja kummankin mallin erottelukyky on hyväksyttävä.

Onnettomuuksien vakavuuteen vaikuttavia tekijöitä tutkitaan tulkitsemalla malleihin valikoituneiden selittäjien vetosuhteita. Osa tuloksista on odotettuja, kuten että kevyen liikenteen osallisuus tai korkea nopeusrajoitus kasvattaa onnettomuuden vakavuutta. Suurin osa tämän työn tuloksista on myös linjassa aikaisempien tutkimusten kanssa. Aiempien tutkimusten ja tämän työn tulosten vertailussa ilmenee, että jäisen ja lumisen kelin vaikutusta sekä valoisuuden vaikutusta liikenneturvallisuuteen voisi tutkia tarkemmin.

Avainsanat: logistinen regressio, liikenneonnettomuudet, askeltava menetelmä, R, odds ratio

SISÄLLYSLUETTELO

1. JOHDANTO

Tässä työssä tutkitaan tieliikenneonnettomuuksien vakavuuteen vaikuttavia tekijöitä. Tieliikenneonnettomuudet aiheuttavat paitsi kärsimystä osallisille, myös taloudellisia kustannuksia yhteiskunnalle. Tieliikenneonnettomuuksissa sattuneiden henkilövahinkojen taloudellinen kustannus Suomessa vuonna 2021 oli 1,2 miljardia euroa [15]. Jotta liikenneonnettomuuksiin liittyvää kärsimystä ja kustannuksia voidaan vähentää, on tärkeää tunnistaa niiden vakavuuteen vaikuttavia tekijöitä.

Liikenneturvallisuus vaihtelee maittain suuresti. Esimerkiksi Suomessa liikennekuolemia tapahtuu Ruotsiin ja Norjaan verrattuna kaksinkertaisesti väkilukuun suhteutettuna [26]. EU-komissio on julkaissut tieliikenneturvallisuuteen liittyvän linjauksen, joka tähtää liikennekuolemien ja vakavien onnettomuuksien puolittamiseen vuoteen 2030 mennessä [20]. Tutkimusten avulla olisikin tärkeää löytää ne tekijät, jotka vaikuttavat onnettomuuksien vakavuuteen eniten juuri Suomessa, jotta liikenneturvallisuutta voidaan parantaa vaikuttavilla ratkaisuilla.

Onnettomuuksien vakavuutta kuvataan usein muutamalla kategorialla. Esimerkiksi tässä työssä käytetyssä aineistossa vakavuus on kuvattu kolmella kategorialla: ei henkilövahinkoa, loukkaantuminen tai kuolema. Tällaisen riippuvan muuttujan mallintamiseen voidaan käyttää logistista regressiomallia. Logistisen regressiomallin avulla voidaan ennustaa tulevia tapahtumia, kun mallissa olevat riippumattomat selittäjät tiedetään. Jos esimerkiksi tiedetään onnettomuuden olosuhteet, mallin avulla voidaan ennustaa onnettomuuden vakavuus. Tässä työssä logistista regressiomallia käytetään kuitenkin onnettomuuksien vakavuuteen vaikuttavien tekijöiden löytämiseen vakavuuden ennustamisen sijaan.

Tässä työssä käytetään binäärisiä logistisia regressiomalleja, joissa riippuva muuttuja saa kahta erilaista arvoa. Luvussa 2 esitetään logistisen regression teoria, joka alkaa binäärisen logistisen regressiomallin yleisen muodon esittämisestä. Alaluvuissa 2.2 ja 2.3 esitetään, miten aineiston avulla muodostetaan logistinen regressiomalli ja millaisia numeerisia ongelmia tässä voi syntyä. Seuraavaksi alaluvuissa 2.4 ja 2.5 esitetään menetelmä mallin valitsemiselle, kun mahdollisia riippumattomia muuttujia on useita. Mallin valinnan jälkeen alaluvussa 2.6 esitetään kaksi erilaista menetelmää mallin suorituskyvyn arvioimiseksi. Lopuksi alaluvussa 2.7 esitetään, miten luotua mallia tulkitaan.

Luvussa 3 logistista regressiota sovelletaan aineistoon, joka sisältää Suomessa vuosina 2017–2021 sattuneita tieliikenneonnettomuuksia. Koska aineistossa vakavuus on kuvattu kolmella kategorialla ja työssä käytetään binääristä logistista regressiota, on aineisto koodattava uudelleen. Alaluvussa 3.1 aineisto siistitään ja siitä muodostetaan kaksi erilaista binääristä versiota. Alaluvussa 3.2

aineistojen pohjalta luodaan kaksi erilaista mallia ja mallien suorituskykyä arvioidaan alaluvussa 3.3. Alaluvussa 3.4 vakavuuteen vaikuttavia tekijöitä tutkitaan tulkitsemalla muodostettuja malleja. Lisäksi saatuja tuloksia vertaillaan aiempien tutkimusten tuloksiin.

2. TEORIA

Regressioanalyysi on tilastollinen menetelmä, jonka avulla tutkitaan yhden tai useamman riippumattoman muuttujan vaikutusta yhteen riippuvaan muuttujaan [8, s. 1]. Tässä työssä riippumattomista muuttujista käytetään nimitystä *selittäjät* ja riippuvasta muuttujasta *vaste*.

Jatkuvan vasteen tapauksessa käytetään lineaarista regressiota. Vaste voi olla myös kategorinen, jolloin vasteen ja selittäjien välille ei voida muodostaa lineaarista regressiomallia. Tällaisessa tapauksessa voidaan käyttää *logistista regressiomallia*. [1, s. 3] Tässä työssä rajoitutaan binäärisiin logistisiin regressiomalleihin, joissa vaste voi saada vain kahta arvoa. Yleisesti vasteelle käytetään arvoa 1, kun tietty tapahtuma realisoituu, ja arvoa 0, kun kyseinen tapahtuma ei realisoidu [8, s. 2].

2.1 Logit-muunnos

Olkoon selittäjiä k kappaletta, jolloin ne voidaan esittää vektorissa $\boldsymbol{x} = [x_1, x_2, \ldots, x_k]^T$. Selittäjät voivat olla jatkuvia, diskreettejä tai kategorisia [1, s. 98]. Kategoriset selittäjät ilmaistaan binääristen *apumuuttujien* (dummy variable) avulla [9, s. 36]. Merkitään $\pi(\boldsymbol{x}) = P(Y = 1 \mid \boldsymbol{x})$, jolloin $\pi(\boldsymbol{x})$ kuvaa ehdollista todennäköisyyttä, joka kertoo millä todennäköisyydellä vaste realisoituu arvoksi 1, kun havaintovektori $\boldsymbol{x}$ on realisoitunut.

Logistinen regressio perustuu logit-muunnokseen

$$\operatorname{logit}(\boldsymbol{x}) = \ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{2.1}$$

[9, s. 35] Yhtälöstä (2.1) voidaan ratkaista vasteen realisoitumisen todennäköisyys

$$\begin{aligned}
\ln\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \\
\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} &= \mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \\
\pi(\boldsymbol{x}) &= (1-\pi(\boldsymbol{x}))\,\mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \\
\pi(\boldsymbol{x}) &= \mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} - \pi(\boldsymbol{x})\mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \\
\pi(\boldsymbol{x})\left(1+\mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}\right) &= \mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k} \\
\pi(\boldsymbol{x}) &= \frac{\mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}}{1+\mathrm{e}^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k}} \\
\pi(\boldsymbol{x}) &= \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_k x_k)}}.
\end{aligned} \tag{2.2}$$

2.2 Suurimman uskottavuuden estimointi

Logistisen regressiomallin parametrit $\boldsymbol{\beta} = [\beta_0, \beta_1, \ldots, \beta_k]^T$ estimoidaan havaittujen selittäjien ja vasteen arvojen avulla käyttämällä *suurimman uskottavuuden estimointia* (maximum likelihood estimation) eli SU-estimointia. Kun selittäjiä on k kappaletta ja riippumattomia havaintoja n kappaletta, voidaan datamatriisi X ja vastevektori $\boldsymbol{y}$ kirjoittaa muodossa

$$\mathrm{X} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{2.3}$$

[24, s. 51, 65] SU-estimointi perustuu havaintojen yhteisjakauman tiheysfunktion arvon maksimointiin [24, s. 46]. Yhteisjakauman tiheysfunktio on tulo

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{1-y_i}, \tag{2.4}$$

missä vektori $\boldsymbol{x}_i$ vastaa datamatriisin X riviä i. Funktiota (2.4) kutsutaan *uskottavuusfunktioksi* (likelihood function). On kuitenkin helpompi laskea parametrien estimaatit summamuotoisesta *log-uskottavuusfunktiosta* (log-likelihood function)

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(\boldsymbol{x}_i)] + (1 - y_i) \ln[1 - \pi(\boldsymbol{x}_i)]\}. \tag{2.5}$$

Parametrien estimaatit $\hat{\boldsymbol{\beta}}$ valitaan SU-estimoinnissa niin, että ne maksimoivat funktion (2.5) arvon. [9, s. 9, 37]

Parametrien estimaatteja ei voida ratkaista analyyttisesti, vaan estimaatit lasketaan numeerisilla iterointimenetelmillä. Tällaisia ovat esimerkiksi Newton–Raphson-algoritmi ja Fisher Scoring -menetelmä. [18, liite A] Yleensä ohjelmistot laskevat parametrien estimaatit automaattisesti. R-ohjelmisto käyttää parametrien estimointiin Fisher Scoring -menetelmää.

2.3 Numeeriset ongelmat

SU-estimoinnin yhteydessä voi nousta esiin numeerisia ongelmia, jotka ilmenevät usein poikkeavan suurista estimoitujen keskivirheiden arvoista. Numeerisia ongelmia aiheuttavat aineiston *kollineaarisuus* (collinearity), *separoituminen* (separation) ja *solun nolla-arvot* (zero cell count). [9, s. 150] Kollineaarisuus johtuu selittäjien korrelaatiosta keskenään [18, s. 126]. Solun nolla-arvo tarkoittaa sitä, että esimerkiksi kategorisen selittäjän jollekin apumuuttujalle ei koskaan realisoidu tapahtumaa $y = 1$ [9, s. 145]. Separoitumista taas tapahtuu silloin, kun selittäjän arvot ennustavat vasteen täydellisesti [1, s. 136].

Separaatiota tai solun nolla-arvoja sisältävää aineistoa voidaan käsitellä erilaisten menetelmien

avulla [6, 7]. Kollineaarisuudelle on usein on vaikea tehdä mitään [3, s. 49]. Lievä kollineaarisuus ei kuitenkaan aiheuta suuria ongelmia, ja se voidaan jättää huomiotta [3, s. 40]. Kollineaarisuus, separoituminen tai solujen nolla-arvot eivät aiheuta virheellisiä kertoimien estimaatteja, vaikka suuren estimoidun keskivirheen takia kertoimien estimaateilla on suuri epätarkkuus [18, s. 132]. Alauvussa 2.7 esitetään lyhyesti, miten keskivirheiden estimaatit lasketaan. Ohjelmistot laskevat keskivirheiden estimaatit kuitenkin automaattisesti. Tässä työssä numeerisia ongelmia aiheuttavat selittäjät tyydytään vain poistamaan mallista.

2.4 Mallin valinta

Jos mahdollisia selittäjiä on useita, tarvitaan menetelmä mallin valitsemiselle. Mallissa tulisi olla tarpeeksi selittäjiä, jotta se sopisi aineistoon hyvin. Toisaalta liian suuri määrä selittäjiä vaikeuttaa mallin tulkintaa. [1, s. 123] Mallin valinta voi tapahtua pelkästään tilastollisten suureiden avulla, mutta valinnassa voidaan käyttää myös omaa harkintaa ja tutkittavan alan tietämystä. Malliin voidaan esimerkiksi ottaa mukaan selittäjä, jonka tiedetään olevan tutkimuksen kannalta oleellinen, vaikkei se olisikaan tilastollisesti merkitsevä. [9, s. 89, 90, 93] Tässä työssä selittäjät valitaan vain tilastollisen merkitsevyyden mukaan.

Tässä alaluvussa esitellään menetelmät mallin ja selittäjien merkitsevyyden testaamiselle. Nämä muodostavat teoreettisen pohjan *askeltavalle menetelmälle* (stepwise method), jota tullaan käyttämään menetelmänä mallin valitsemiselle. Askeltavaa menetelmää käsitellään alaluvussa 2.5.

Mallin merkitsevyyttä voidaan arvioida vertaamalla sitä muihin malleihin. Vertaaminen voidaan suorittaa *uskottavuussuhdetestillä* (likelihood-ratio test). Vertailtavien mallien tulee olla sisäkkäisiä (nested). Sisäkkäiset mallit eroavat toisistaan niin, että toisesta mallista puuttuu osa selittäjistä, jotka ovat toisessa mallissa. Olkoon mallissa M_0 selittäjiä m kappaletta ja mallissa M_1 selittäjiä $m + l$ kappaletta. Uskottavuussuhdetestin testisuure G saadaan laskemalla

$$G = 2\left(L_1 - L_0\right), \tag{2.6}$$

missä L_1 on mallin M_1 log-uskottavuusfunktio (2.5) ja L_0 on mallin M_0 log-uskottavuusfunktio (2.5). Testisuure noudattaa isoilla otoskooilla jakaumaa $G \sim \chi^2((m + l) - m)$ eli $G \sim \chi^2(l)$. Suuri testisuure (2.6) ja sen pieni p-arvo viittaisivat siihen, että laajempi malli M_1 sopii aineistoon paremmin kuin karsitumpi malli M_0. Pieni testisuure (2.6) ja sen suuri p-arvo viittaisivat puolestaan siihen, että karsitumpi malli M_0 sopii aineistoon paremmin kuin laajempi malli M_1. [1, s. 80, 81]

Uskottavuussuhdetestin avulla voidaan tarkastella myös vain yksittäisen selittäjän merkitsevyyttä. Tällöin mallit ovat muuten samanlaiset, mutta malli M_1 sisältää myös tarkasteltavan selittäjän. Jos tarkasteltava selittäjä on kategorinen, sisältää malli M_1 kaikki selittäjän esittämiseen käytetyt apumuuttujat [9, s. 41]. Yksittäisten apumuuttujien merkitsevyyksiä ei siis tarkastella.

2.5 Askeltava menetelmä

Askeltava menetelmä on mallinvalintamenetelmä, joka perustuu pelkästään tilastollisten suureiden tarkasteluun. Se koostuu *poistovalinnasta* (backward elimination) ja *eteenpäinvalinnasta* (forward selection). [9, s. 93] Tässä työssä mallin valinta päädyttiin tekemään askeltavalla menetelmällä, koska menetelmä on aikoinaan ollut hyvin suosittu [9, s. 125] ja koska automatisoidun luonteensa ansiosta se on helppo toteuttaa.

Poistovalinnassa malliin valitaan aluksi kaikki ne selittäjät, joilla uskotaan olevan vaikutusta vasteeseen. Mallista poistetaan kierroksittain aina vähiten merkitsevä selittäjä, kunnes malli sisältää vain merkitseviä selittäjiä. [9, s. 129] Selittäjän merkitsevyyttä voidaan tarkastella uskottavuussuhdetestin (2.6) avulla [9, s. 125].

Eteenpäinvalinnassa aloitetaan tyhjästä mallista. Jokaisesta selittäjästä, jolla uskotaan olevan vaikutusta vasteeseen, muodostetaan oma mallinsa. Seuraavaksi lasketaan jokaisen mallin uskottavuussuhde (2.6) tyhjän mallin kanssa. Mikäli merkitsevimmän mallin p-arvo on alle valitun riskitason, valitaan kyseinen selittäjä malliin. Tämän jälkeen luodaan uudet mallit jäljelle jääneistä selittäjistä. Uudet mallit sisältävät nyt aiemman malliin valitun merkitsevän selittäjän ja uuden tarkasteltavan selittäjän. Verrataan uusia malleja malliin, jossa on vain aiemmin valittu merkitsevä selittäjä. Lisätään taas merkitsevin riskitason alittava selittäjä malliin ja muodostetaan uudet mallit vertailua varten. Tätä jatketaan, kunnes kaikkien valitsematta jääneiden selittäjien p-arvo on suurempi kuin käytetty riskitaso tai kunnes kaikki mahdolliset selittäjät ovat mallissa. Tällöin mitään selittäjää ei voida enää lisätä malliin. [9, s. 126, 127]

Koska R-ohjelmistossa ei ole valmista funktiota uskottavuussuhteeseen perustuvalle askeltavalle menetelmälle, esitetään yksi esimerkki sellaisen toteutuksesta. Tässä työssä käytettävä askeltava menetelmä on mukailtu algoritmista, jonka ovat esittäneet Hosmer et al. kirjassa [9, s. 126–128]. Ohessa on esitetty tässä työssä käytetyn menetelmän algoritmi. Sitä noudattava ohjelmakoodi löytyy liitteestä A.

1. Suoritetaan kierros eteenpäinvalintaa, jolloin tyhjään malliin lisätään merkitsevin selittäjä. Mikäli selittäjää ei voida lisätä, algoritmin suoritus päättyy.
2. Suoritetaan uusi kierros eteenpäinvalintaa, jolloin malliin koitetaan lisätä uusi selittäjä.
3. Suoritetaan kierros poistovalintaa.
4. Kohtia 2 ja 3 jatketaan niin kauan, kunnes kaikki mahdolliset selittäjät on valittu malliin. Suoritus päättyy myös silloin, kun mitään selittäjää ei pystytä poistamaan mallista eikä mitään selittäjää pystytä lisäämään malliin.

Lee ja Koval [12] suosittelevat yleisenä ohjeena riskitasoksi eteenpäinvalinnassa arvoa $0,15 \leq \alpha_{\mathrm{E}} \leq 0,20$. Poistovalinnan riskitasoksi valitaan eteenpäinvalinnan riskitasoa isompi arvo eli $\alpha_{\mathrm{R}} > \alpha_{\mathrm{E}}$. Tällöin vältytään tilanteelta, jossa algoritmi poistaa mallista juuri lisätyn selittäjän. [9, s. 127] Tässä työssä tullaan käyttämään selittäjän lisäykselle riskitasoa $\alpha_{\mathrm{E}} = 0,15$ ja selittäjän poistolle riskitasoa $\alpha_{\mathrm{R}} = 0,20$.

2.6 Mallin hyvyys ja erottelukyky

Mallin muodostamisen jälkeen halutaan arvioida sen suorituskykyä. Tässä alaluvussa esitetään kaksi erilaista tapaa arvioida sitä. Suorituskykyä voidaan arvioida tarkastelemalla miten hyvin mallin ennusteet vertautuvat havaittuihin tuloksiin. Havaituilla tuloksilla tarkoitetaan tässä työssä aineistossa esiintyviä vasteen arvoja. Tätä kutsutaan *mallin hyvyyden* (goodness of fit) tarkasteluksi [9, s. 153]. Mallin suorituskykyä voidaan arvioida myös tarkastelemalla, miten hyvin malli erottaa tapahtuman, jolle realisoituu $y = 1$ tapahtumasta, jolle realisoituu $y = 0$. Tätä kutsutaan mallin *erottelukyvyksi* (discrimination). [2]

Mallin hyvyyttä voidaan tutkia *Hosmer–Lemeshow-testin* avulla [2]. Siinä havainnot järjestetään nousevaan järjestykseen estimoidun todennäköisyyden $\hat{\pi}(\boldsymbol{x})$ mukaan, jonka jälkeen havainnot jaetaan ryhmiin esimerkiksi kvantiileittain [9, s. 157]. Kvantiilien avulla voidaan laskea mallin hyvyydestä kertova Hosmer–Lemeshow-testisuure

$$\hat{C} = \sum_{i=1}^{g} \frac{(o_{1i} - n_i \bar{\pi}_i)^2}{n_i \bar{\pi}_i (1 - \bar{\pi}_i)}, \tag{2.7}$$

missä o_{1i} on ryhmässä i olevien havaintojen $y = 1$ lukumäärä, n_i on havaintojen lukumäärä ryhmässä i ja $\bar{\pi}_i$ on keskimääräinen estimoitu todennäköisyys ryhmälle i. Usein ryhmien määräksi valitaan $g = 10$. Testisuure noudattaa arviolta jakaumaa $\chi^2(g - 2)$. [9, s. 158-160] Testin nollahypoteesi on, että mallilla on hyvä yhteensopivuus [9, s. 161]. Hyvä yhteensopivuus tarkoittaa sitä, että estimoitujen ja havaittujen arvojen määrät eivät poikkea toisistaan missään ryhmässä [2]. Jos testisuure (2.7) saa suuren arvon, saa jakauma pienen p-arvon. Merkitsevä tulos viittaa siihen, että mallilla ja aineistolla on huono yhteensopivuus. Tämä tarkoittaa sitä, että estimoidut ja havaitut arvot poikkeavat jossain ryhmässä toisistaan enemmän, kuin pelkän sattuman kannalta olisi syytä olettaa [2].

Mallin hyvyyttä voidaan tarkastella visuaalisesti taulukon avulla. Taulukon sarakkeina ovat ryhmän numero väliltä 1–10, estimoitu todennäköisyysväli kyseiselle ryhmälle, havaittujen tapahtumien $y = 0$ lukumäärä kyseisessä ryhmässä, havaittujen tapahtumien $y = 1$ lukumäärä kyseisessä ryhmässä, estimoitujen tapahtumien $\hat{y} = 0$ lukumäärä kyseisessä ryhmässä ja estimoitujen tapahtumien $\hat{y} = 1$ lukumäärä kyseisessä ryhmässä. Estimoitujen tapahtumien $\hat{y} = 1$ lukumäärä ryhmässä i vastaa kaavan (2.7) tuloa $n_i \bar{\pi}_i$. Estimoitujen tapahtumien $\hat{y} = 0$ lukumäärä ryhmässä i saadaan puolestaan laskettua tulolla $n_i\,(1 - \bar{\pi}_i)$. Tällainen taulukko auttaa hahmottamaan mallin yhteensopivuutta tarkemmin eri ryhmissä. Taulukoita käytetään alaluvussa 3.3, jossa luotujen mallien hyvyyttä tarkastellaan.

Malli erottelee havainnon tahtumaan $\hat{y} = 1$ tai $\hat{y} = 0$ vertaamalla estimoitua todennäköisyyttä (2.2) valittuun kynnysarvoon (cutpoint). Kynnysarvona tapahtuman erottelulle voitaisiin käyttää esimerkiksi arvoa 0,5. Mikäli malli estimoi tapahtumalle $\hat{\pi}(\boldsymbol{x}) < 0,5$, voidaan merkitä, että $\hat{y} = 0$, ja mikäli malli estimoi tapahtumalle $\hat{\pi}(\boldsymbol{x}) \geq 0,5$, voidaan merkitä, että $\hat{y} = 1$. [9, s. 170] Mikäli kynnysarvoa kasvatetaan, väärien positiivisten lukumäärä vähenee, mutta useampi oikea positiivinen jää huomaamatta. Vastaavasti matala kynnysarvo johtaa suurempaan oikeiden positiivisten lukumäärään, mutta myös väärien positiivisten lukumäärä kasvaa. [2]

Taulukko 2.1. *AUC-arvojen kuvaukset [9, s. 177].*

AUC	Kuvaus
$0,5$	tulos satunnainen, ei erottelukykyä
$0,5-0,7$	huono erottelukyky, muttei satunnainen
$0,7-0,8$	hyväksyttävä erottelukyky
$0,8-0,9$	erinomainen erottelukyky
$0,9-1$	ilmiömäinen erottelykyky

Mallin erottelukykyä voidaan tarkastella *ROC-kuvaajan* (receiver operating characteristic) avulla. Siinä kuvaajaan piirtyy oikeiden positiivisten suhde (sensitivity) ja väärien positiivisten suhde (1−specificity) kaikille mahdollisille kynnysarvoille. [9, s. 174] Oikeiden positiivisten suhde tarkoittaa oikeiden positiivisten osuutta kaikista havaituista positiivisista tapauksista. Väärien positiivisten suhde tarkoittaa väärien positiivisten osuutta kaikista havaituista negatiivisista tapauksista. [9, s. 171] ROC-kuvaajan alle jäävää alaa voidaan merkitä termillä AUC (area under the ROC curve). AUC saa arvoja väliltä $0,5-1,0$ ja sen voidaan katsoa noudattavan taulukossa 2.1 esitettäviä kuvauksia [9, s. 177]. Usein AUC-arvot ovat väliltä $0,6-0,9$ [8, s. 258].

R-ohjelmistolla ei ole valmiita komentoja Hosmer–Lemeshow-testille eikä ROC-kuvaajalle. Tässä työssä Hosmer–Lemeshow-testin laskemiseen käytetään R-ohjelmiston pakettia ResourceSelection. ROC-kuvaajan piirtoon ja AUC-arvon laskemiseen käytetään R-ohjelmiston pakettia ROCR.

Mallin hyvyyden ja erottelukyvyn testaaminen samalla aineistolla, jolla malli on luotu, johtaa yleensä liian optimistisiin tuloksiin [9, s. 202][23]. Mallin arviointi uudella aineistolla antaa realistisemman kuvan mallin hyvyydestä ja erottelukyvystä. Tätä kutstutaan mallin validoinniksi ja se on erityisen tärkeää silloin, kun mallia käytetään tulevien tapausten luokitteluun [9, s. 202]. Koska tässä työssä luotuja malleja käytetään vain onnettomuuksien vakavuuteen vaikuttavien selittäjien tutkimiseen eikä uusien tapausten luokitteluun, ei malleille suoriteta validointia.

2.7 Vetosuhde

Kun malli on muodostettu ja sen suorituskyky todettu riittäväksi, on aika tulkita mallia. Mallin tulkinta tapahtuu tässä työssä *vetosuhteita* (odds ratio, OR) käyttämällä. Vetosuhde on yleisesti käytetty suure, joka kertoo vasteen ja tarkasteltavan selittäjän välisestä riippuvuudesta [9, s. 52]. Vetosuhteet lasketaan parametrien $\boldsymbol{\beta}$ avulla. Selittäjän x_i vetosuhteeksi saadaan yleisessä tapauksessa

$$\begin{aligned}
\mathrm{OR}(a,b) &= \frac{\frac{\pi(x_i=a)}{1-\pi(x_i=a)}}{\frac{\pi(x_i=b)}{1-\pi(x_i=b)}} \\
&= \frac{\mathrm{e}^{(\beta_0+\beta_1 x_1+\cdots+a\beta_i+\cdots+\beta_k x_k)}}{\mathrm{e}^{(\beta_0+\beta_1 x_1+\cdots+b\beta_i+\cdots+\beta_k x_k)}} \\
&= \mathrm{e}^{(\beta_0+\beta_1 x_1+\cdots+a\beta_i+\cdots+\beta_k x_k)-(\beta_0+\beta_1 x_1+\cdots+b\beta_i+\cdots+\beta_k x_k)} \\
&= \mathrm{e}^{(a-b)\beta_i},
\end{aligned} \tag{2.8}$$

missä $i = 1, \ldots, k$ ja $a, b \in \mathbb{R}$. Osoittaja $\pi(x_i = a)/(1 - \pi(x_i = a))$ vastaa *vetoa* (odds) tilanteessa, jossa tapahtuma $y = 1$ realisoituu, kun tarkasteltava selittäjä saa arvon $x_i = a$. Nimittäjä $\pi(x_i = b)/(1 - \pi(x_i = b))$ vastaa vetoa taas siinä tilanteessa, jossa tapahtuma $y = 1$ realisoituu, kun tarkasteltava selittäjä saa arvon $x_i = b$. Muiden selittäjien arvot oletetaan pysyvän vakiona. [9, s. 51, 55][22, s. 67] Vetosuhde saa arvoja väliltä $(0, \infty)$ [1, s. 32]. Seuraavaksi esitetään vetosuhteiden tarkemmat tulkinnat kategorisille selittäjille, jonka jälkeen esitetään vetosuhteiden tarkemmat tulkinnat jatkuville ja diskreeteille selittäjille.

Kategoriset selittäjät tuodaan malliin käyttämällä apumuuttujia. Kunkin kategorisen selittäjän apumuuttujan vetosuhde saadaan sijoittamalla kaavaan (2.8) $a = 1$ ja $b = 0$. Tällöin saadaan

$$\mathrm{OR} = \mathrm{e}^{\beta_i}, \tag{2.9}$$

missä β_i on tarkasteltavan apumuuttujan kerroin. Kategorisen selittäjän tapauksessa apumuuttujan vetosuhde (2.9) kertoo kyseisen apumuuttujan vedon suhteesta referenssimuuttujan vetoon [21, s. 25]. Vetosuhteen arvo OR = 1 tarkoittaa, että tarkasteltavan apumuuttujan ja referenssimuuttujan vedot eivät eroa toisistaan. Kun OR $>$ 1, on tarkasteltavan apumuuttujan tapahtuman $y = 1$ veto suurempi kuin referenssimuuttujan. Tämä tarkoittaa, että tapahtuma $y = 1$ realisoituu todennäköisemmin tarkasteltavalle apumuuttujalle kuin referenssimuuttujalle. Kun OR $<$ 1, on tarkasteltavan apumuuttujan tapahtuman $y = 1$ veto pienempi kuin referenssimuuttujan. Tällöin tapahtuma $y = 1$ realisoituu todennäköisemmin referenssimuuttujalle kuin tarkasteltavalle apumuuttujalle. [1, s. 32]

Jatkuvan tai diskreetin selittäjän tapauksessa vetosuhde (2.8) kertoo vetojen suhteen muutoksesta, kun selittäjän arvo muuttuu $a - b$ yksikköä [9, s. 63]. Merkitään $a - b = c$, jolloin c yksikön muutoksen vetosuhde saadaan muotoon

$$\mathrm{OR}(c) = \mathrm{e}^{c\beta_i}. \tag{2.10}$$

Jatkuvan tai diskreetin selittäjän tapauksessa vetosuhteen arvo OR = 1 tarkoittaa, että tarkasteltava selittäjä ja vaste ovat riippumattomia. Kun OR $>$ 1, tapahtuman $y = 1$ veto kasvaa selittäjän c yksikön muutoksessa. [1, s. 32] Tällöin voidaan sanoa, että selittäjällä on tapahtuman $y = 1$ todennäköisyyttä lisäävä vaikutus. Kun OR $<$ 1, tapahtuman $y = 1$ veto puolestaan vähenee selittäjän c yksikön muutoksessa [1, s. 32]. Tällöin voidaan sanoa, että selittäjällä on tapahtuman $y = 1$ todennäköisyyttä vähentävä vaikutus.

Vetosuhteiden estimaatit $\widehat{\mathrm{OR}}$ saadaan laskettua parametriestimaattien $\hat{\boldsymbol{\beta}}$ ja kaavojen (2.9) ja (2.10) avulla

$$\widehat{\mathrm{OR}} = \mathrm{e}^{\hat{\beta}_i}, \qquad \widehat{\mathrm{OR}}(c) = \mathrm{e}^{c\hat{\beta}_i} \tag{2.11}$$

[9, s. 54, 63]. Johdetaan seuraavaksi kaavat vetosuhteiden 95 % luottamusvälin laskemiseen.

Log-uskottavuusfunktion (2.5) avulla voidaan laskea havaittu informaatiomatriisi (observed information matrix)

$$\mathrm{I}(\boldsymbol{\beta}) = -\left[\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_i \partial\beta_j}\right], \tag{2.12}$$

missä $i, j = 0, 1, \ldots, k$. Matriisin koko on $(k + 1) \times (k + 1)$. Estimoitujen parametrien varianssit

ja kovarianssit saadaan havaitun informaatiomatriisin (2.12) käänteismatriisista

$$\mathrm{Var}(\boldsymbol{\beta}) = \mathrm{I}^{-1}(\boldsymbol{\beta}). \tag{2.13}$$

Varianssien ja kovarianssien estimaatit $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ saadaan sijoittamalla kaavaan (2.13) parametrien estimaatit $\hat{\boldsymbol{\beta}}$. Kunkin parametrin β_i varianssin estimaatti $\widehat{\mathrm{Var}}(\hat{\beta}_i)$ saadaan siis matriisin $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ diagonaalilta. Estimoitujen parametrien estimoidut keskivirheet saadaan estimoitujen varianssien avulla laskemalla

$$\widehat{\mathrm{SE}}(\hat{\beta}_i) = \left[\widehat{\mathrm{Var}}(\hat{\beta}_i)\right]^{1/2}, \tag{2.14}$$

missä $i = 0, 1, \ldots, k$. [9, s. 37, 38] Yleensä ohjelmistot laskevat estimoitujen parametrien estimoidut keskivirheet automaattisesti, eikä niitä tarvitse itse laskea.

Estimoitujen keskivirheiden avulla voidaan laskea luottamusvälit estimoiduille parametreille sekä estimoitujen parametrien vetosuhteille. Parametrin β_i estimaatin $100(1-\alpha)\%$ luottamusväli saadaan laskemalla

$$\hat{\beta}_i \pm z_{1-\alpha/2} \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i) \tag{2.15}$$

[9, s. 59]. Kun riskitasoksi valitaan $\alpha = 0,05$, on silloin $z_{1-0,05/2} = 1,96$ [9, s. 54]. Siispä parametrin β_i estimaatin 95 % luottamusväliksi saadaan

$$\hat{\beta}_i \pm 1,96 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i). \tag{2.16}$$

Parametriestimaatin $(a-b)\hat{\beta}_i$ estimoiduksi keskivirheeksi saadaan $\widehat{\mathrm{SE}}\left[(a-b)\hat{\beta}_i\right] = |a-b|\widehat{SE}\left(\hat{\beta}_i\right)$. Tällöin yleiseksi parametrin β_i estimaatin $a-b$ yksikön muutoksen 95 % luottamusväliksi saadaan

$$\exp[\hat{\beta}_i(a-b) \pm 1,96 \cdot |a-b|\widehat{\mathrm{SE}}(\hat{\beta}_i)]. \tag{2.17}$$

[9, s. 56, 63] Kun kaavaan (2.17) sijoitetaan $a = 1$ ja $b = 0$, saadaan kategorisen selittäjän apumuuttujan parametrin estimaatin 95 % luottamusväliksi

$$\exp[\hat{\beta}_i \pm 1,96 \cdot \widehat{\mathrm{SE}}(\hat{\beta}_i)]. \tag{2.18}$$

Kun kaavaan (2.17) sijoitetaan $a - b = c$, saadaan diskreetin ja jatkuvan selittäjän parametrin estimaatin c yksikön muutoksen 95 % luottamusväliksi

$$\exp[c\hat{\beta}_i \pm 1,96 \cdot |c|\widehat{\mathrm{SE}}(\hat{\beta}_i)]. \tag{2.19}$$

3. LOGISTISEN REGRESSION SOVELTAMINEN AINEISTOON

Tässä luvussa tarkastellaan kahden erilaisen binäärisen logistisen regressiomallin sovittamista tieliikenneonnettomuusaineistoon. Aluksi aineisto siistitään ja siitä muodostetaan kaksi erilaista versiota. Tämän jälkeen logistiset regressiomallit muodostetaan askeltavalla menetelmällä. Muodostettujen mallien suorituskykyä arvioidaan Hosmer–Lemeshow-testin ja ROC-kuvaajan avulla. Onnettomuuksien vakavuuteen vaikuttavia tekijöitä tutkitaan tulkitsemalla muodostettujen mallien vetosuhteita. Lopuksi saatuja tuloksia vertaillaan aiempien tutkimusten tuloksiin.

3.1 Aineisto ja sen siistintä

Tässä työssä käytetty aineisto sisältää tietoa Suomen tieliikenneonnettomuuksista. Aineisto on koottu vuosien 2017–2021 havainnoista. Aineistoja ylläpitää Väylävirasto, ja ne ovat saatavissa osoitteessa `https://ava.vaylapilvi.fi/ava/Tie/Tieliikenneonnettomuudet` lisenssillä CC 4.0 BY (tarkistettu 10.12.2022). Samasta osoitteesta on saatavissa myös aineiston tietosisällön kuvaus. Aineisto sisältää merkinnän onnettomuuden vakavuudesta sarakkeessa Vakavuusko. Tämä tulee olemaan muodostettujen mallien vaste. Muut aineiston sarakkeet ovat mahdollisia selittäjiä.

Aluksi mahdollisista selittäjistä karsitaan pois kaikki epäolennaiset. Tässä työssä selittäjä on epäolennainen, jos se sisältää vain kirjallisen kuvauksen toisesta selittäjästä (esimerkiksi Ontyypsel), yli puolet sen havainnoista on tyhjiä arvoja tai jos selittäjän kuvausta ei löydy tietosisällön kuvauksesta. Mahdollisiksi selittäjiksi jää Tienpit, Tie, Aosa, Aet, Ajr, Kk, ELY, Poliisipri, Tunti, Vkpv, Ontyyppi, Onluokka, Osallkm, Nopraj, Taajmerk, Pinta, Valoisuus, Sää, Onnpaikka, Maakunta, Kunta, Taajama, Toimluokka, Kvl, Raskaskvl, Tienlev, Tietyö, Päällyste, Lämpötila, Risteys, Talvhoitlk ja Poikkileik.

Koska selittäjä Tie sisältää 4769 kategoriaa, selittäjä Aosa sisältää 474 kategoriaa, selittäjä Kunta sisältää 295 kategoriaa ja selittäjä Ontyyppi sisältää 76 kategoriaa, jätetään ne pois mahdollisista selittäjistä. Näin iso määrä kategorioita vaikeuttaisi mallin tulkintaa. Koska selittäjä Tienpit sisältää siistimisen jälkeen vain yhtä arvoa, myös se jätetään pois mahdollisista selittäjistä. Selittäjät Poikkileik ja Ajr näyttäisivät sisältävän saman informaation. Myös selittäjät Taajmerk ja Taajama näyttävät sisältävän saman informaation. Koska kaksi samanlaista selittäjää mallissa johtaa vahvaan kollineaarisuuteen ja täten myös numeerisiin ongelmiin, jätetään aineistoon vain toinen selittäjistä. Jätetään aineistoon Ajr ja Taajmerk.

Koodataan selittäjän Tunti havainnot uudelleen niin, että tunnin sijaan ilmoitetaan aikaväli, jolloin

Taulukko 3.1. *Selittäjien Tunti ja Talvhoitlk uudet koodaukset ja koodien kuvaukset.*

Koodi	Tunti	Talvhoitlk
0	-	Liukkaudentorjunta ilman toimenpideaikaa
1	00–03	Tie, joka on normaalisti aina paljaana
2	03–06	Tie, joka on normaalisti paljaana
3	06–09	Tie, joka on osan talvea lumipintainen
4	09–12	Pääosin hiekoitettava tie
5	12–15	Pääosin lumipintainen tie
6	15–18	Tie, joka hiekoitetaan vain pahimmissa tilanteissa
7	18–21	Talvihoidettu kevyen liikenteen väylä
8	21–00	-

onnettomuus on sattunut. Tämä vähentää mallissa olevien apumuuttujien määrää ja selkeyttää mallin tulkintaa. Myös selittäjä Talvhoitlk koodataan uudelleen, koska eri vuosien koodaukset poikkeavat toisistaan. Uudelleenkoodattujen selittäjien koodit ja kuvaukset on esitetty taulukossa 3.1.

Onnettomuuden vakavuus on merkitty arvolla 0, jos onnettomuudessa ei ole sattunut henkilövahinkoa, arvolla 1, jos onnettomuus on johtanut kuolemaan ja arvolla 2, jos onnettomuus on johtanut loukkaantumiseen. Koska vakavuus, eli tässä työssä vaste, on ilmoitettu kolmella kategorialla ja tässä työssä käytetään binäärisiä logistisia regressiomalleja, on vakavuus koodattava uudelleen. Luodaan tätä varten aineistosta kaksi erilaista versiota. Ensimmäisessä versiossa vakavuus koodataan niin, että se saa arvon 0, jos onnettomuus ei ole johtanut henkilövahinkoon ja arvon 1, jos onnettomuus on johtanut henkilövahinkoon. Henkilövahinko tarkoittaa tässä työssä sitä, että onnettomuus on johtanut loukkaantumiseen tai kuolemaan. Tämän koodauksen pohjalta luotuun malliin viitataan vastaisuudessa nimityksellä malli 1. Toisessa versiossa vakavuus koodataan niin, että se saa arvon 0, jos onnettomuus on johtanut loukkaantumiseen ja arvon 1, jos onnettomuus on johtanut kuolemaan. Tämän koodauksen pohjalta luotuun malliin viitataan vastaisuudessa nimityksellä malli 2. Mallin 1 avulla tarkastellaan siis selittäjiä, jotka erottelevat henkilövahinkoon johtavan onnettomuuden onnettomuudesta, joka ei johda henkilövahinkoon. Mallin 2 avulla tarkastellaan puolestaan selittäjiä, jotka erottelevat kuolemaan johtavan onnettomuuden loukkaantumiseen johtavasta onnettomuudesta.

On yleistä, että osa etenkin lievemmistä onnettomuuksista jää raportoimatta. Tämä saattaa johtaa vääristyneisiin parametrien estimaatteihin. [25] Voidaan siis olettaa, että tässä työssä käytetty aineisto ei kata kaikkia Suomessa aikavälillä 2017–2021 sattuneita tieliikenneonnettomuuksia. Tämä saattaa vaikuttaa työn tuloksiin, mutta aiheeseen ei tässä kiinnitetä enempää huomiota.

3.2 Mallien muodostus

Muodostetaan mallit käyttämällä askeltavaa menetelmää, jonka toteutus on esitetty liitteessä A. Riskitasona selittäjän lisäämiselle käytetään $\alpha_{\mathrm{E}} = 0,15$. Riskitasona selittäjän poistamiselle käyte-

Taulukko 3.2. *Diskreettien selittäjien estimaatit c yksikön vetosuhteille.*

Selittäjä	c	$\widehat{\mathrm{OR}}(c)$	95 % luottamusväli
Kvl (malli 1)	1000	0,988	[0,984; 0,993]
Kvl (malli 2)	1000	0,973	[0,959; 0,987]
Lämpötila (malli 1)	10	1,239	[1,174; 1,308]

tään $\alpha_{\mathrm{R}} = 0,20$.

Malliin 1 valikoituu selittäjiksi Onluokka, Lämpötila, Maakunta, Nopraj, Kvl, Osallkm, Pinta, Talvhoitlk, Ajr, Onnpaikka, Sää, ELY, Päällyste, Valoisuus, Tunti ja Taajmerk. Mallin tuloste on esitetty liitteessä B.1. Malliin valikoituu selittäjä Maakunta, joka vaikuttaisi sisältävän osittain samaa informaatiota, kuin ELY. Molemmat kuvaavat missä päin Suomea onnettomuus on tapahtunut. Usealla selittäjän Maakunta apumuuttujalla on normaalia suurempi estimoitu keskihajonta, mikä viittaisi johonkin numeeriseen ongelmaan. Syynä voisi olla esimerkiksi kollineaarisuus selittäjän ELY kanssa. Selittäjällä Maakunta on paljon enemmän apumuutujia kuin selittäjällä ELY, joten poistetaan Maakunta mallista. Yksinkertaisempia malleja on helpompi tulkita, joten suositaan yksinkertaisempia selittäjiä. Karsitumman mallin tuloste on esitetty liitteessä B.2.

Malliin 2 valikoituu selittäjiksi Onluokka, Pinta, Tunti, Kvl, Toimluokka, Onnpaikka, Risteys, Taajmerk, Lämpötila ja Valoisuus. Mallin tuloste on esitetty liitteessä B.3. Selittäjien Risteys ja Onnpaikka useilla apumuuttujilla on poikkeavan suuria keskihajonnan arvoja, mikä viittaa numeerisiin ongelmiin. Poistetaan selittäjät mallista. Karsitumman mallin tuloste on esitetty liitteessä B.4.

Molempiin malleihin valikoituu diskreettejä selittäjiä. Tarkastellaan selittäjän Kvl arvojen suuruusluokkaa alkuperäisen aineiston pohjalta. Yhden yksikön muutos vuoden keskimääräisessä vuorokausiliikenteessä vastaa yhden moottoriajoneuvon muutosta. Selittäjän Kvl minimiarvo on 0, keskiarvo 9066 ja maksimi 103623. Lisäksi ensimmäisen kvantiilin arvo on 1176, mediaanin 3477 ja kolmannen kvantiilin arvo on 8613. Tämän perusteella voidaan todeta, että yhden moottoriajoneuvon tarkastelun sijaan olisi järkevää tarkastella esimerkiksi 1000 moottoriajoneuvon muutosta. Selittäjää Lämpötila tarkastellaan 10 asteen muutoksina. Selittäjää Osallkm tarkastellaan yhden osallisen muutoksina. Lasketaan selittäjille Kvl ja Lämpötila uudet estimoidut vetosuhteet ja estimoitujen vetosuhteiden 95 % luottamusvälit kaavojen (2.11) ja (2.19) avulla. Tulokset on esitetty taulukossa 3.2. Niitä tulkitaan myöhemmin luvussa 3.4.

3.3 Mallien hyvyyden ja erottelukyvyn tarkasteleminen

Tarkastellaan aluksi mallien hyvyyttä laskemalla Hosmer–Lemeshow-testisuure (2.7) paketin ResourceSelection avulla. Käytetään ryhmien määränä 10. Tulokset malleille on esitetty taulukossa 3.3. Koska mallien p-arvot eivät ole merkitseviä yleisesti käytetyillä riskitasoilla, voidaan olettaa, ettei malleilla ole huonoa yhteensopivuutta. Voidaan siis todeta, että havaittujen ja estimoitujen tapausten lukumäärät vastaavat toisiaan jokaisessa ryhmässä.

Havainnollistetaan mallien hyvyyttä vielä taulukoilla, joista kerrottiin alaluvussa 2.6. Niissä esi-

Taulukko 3.3. *Hosmer–Lemeshow-testistä saadut tunnusluvut.*

Malli	$\hat{\mathbf{C}}$	**Vapausasteet**	***p*-arvo**
1	12,764	8	0,120
2	9,432	8	0,307

Taulukko 3.4. *Hosmer–Lemeshow-testiin liittyvät ryhmät mallille 1.*

Ryhmä	**Väli**	y_0	y_1	$\hat{y}_0$	$\hat{y}_1$
1	[2,58e-05; 0,0875]	1671	121	1675,94	116,06
2	(0,0875; 0,138]	1596	195	1592,37	198,63
3	(0,138; 0,195]	1510	281	1493,38	297,62
4	(0,195; 0,244]	1398	393	1397,95	393,05
5	(0,244; 0,285]	1363	428	1316,63	474,37
6	(0,285; 0,328]	1209	582	1242,02	548,98
7	(0,328; 0,376]	1136	655	1161,67	629,33
8	(0,376; 0,436]	1069	722	1065,52	725,48
9	(0,436; 0,536]	918	873	932,87	858,13
10	(0,536; 1]	546	1245	537,65	1253,35

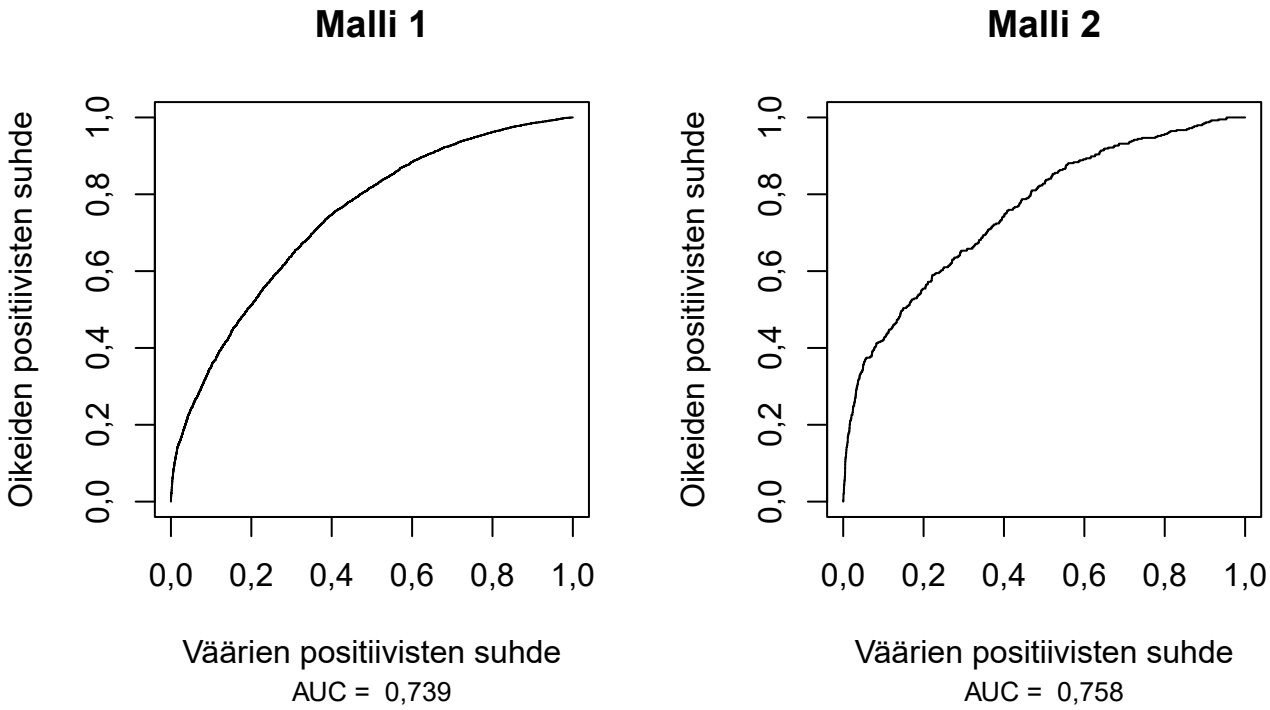
tetään jokaiselle ryhmälle estimoitujen ja havaittujen tapahtumien lukumäärät. Mallin 1 tulokset on esitetty taulukossa 3.4 ja mallin 2 tulokset on esitetty taulukossa 3.5. Huomataan, että pienillä estimoidun todennäköisyyden $\hat{\pi}(\boldsymbol{x})$ arvoilla sekä havaittujen tapahtumien $y = 0$ että estimoitujen tapahtumien $\hat{y} = 0$ lukumäärät ovat selvästi suurempia kuin havaittujen tapahtumien $y = 1$ ja estimoitujen tapahtumien $\hat{y} = 1$ lukumäärät. Näin on kummassakin mallissa. Estimoidun todennäköisyyden kasvaessa havaittujen tapahtumien $y = 0$ ja estimoitujen tapahtumien $\hat{y} = 0$ lukumäärät vähenevät ja havaittujen tapahtumien $y = 1$ ja esimoitujen tapahtumien $\hat{y} = 1$ lukumäärät kasvavat. Tämä on odotettu tulos, sillä logistisen regressiomallin halutaan estimoivan tapahtumille $y = 0$ pieniä todennäköisyyksiä $\hat{\pi}(\boldsymbol{x})$ ja tapahtumille $y = 1$ suuria todennäköisyyksiä $\hat{\pi}(\boldsymbol{x})$. Kummassakin mallissa estimoidut ja havaitut arvot näyttävät vastaavan suhteellisen hyvin toisiaan. Joissakin ryhmissä arvot eroavat toisistaan hieman enemmän kuin muissa ryhmissä, mutta kuten taulukon 3.3 tuloksista huomattiin, erot eivät ole niin isoja, etteivätkö ne voisi olla sattumaa.

Tarkastellaan sitten mallien erottelukykyä piirtämällä ROC-kuvaajat paketin ROCR avulla. Kuvassa 3.1 on esitetty molempien mallien ROC-kuvaajat ja ilmoitettu niiden alle jäävä ala eli AUC. Molempien mallien erottelukyky on taulukon 2.1 mukaan hyväksyttävä.

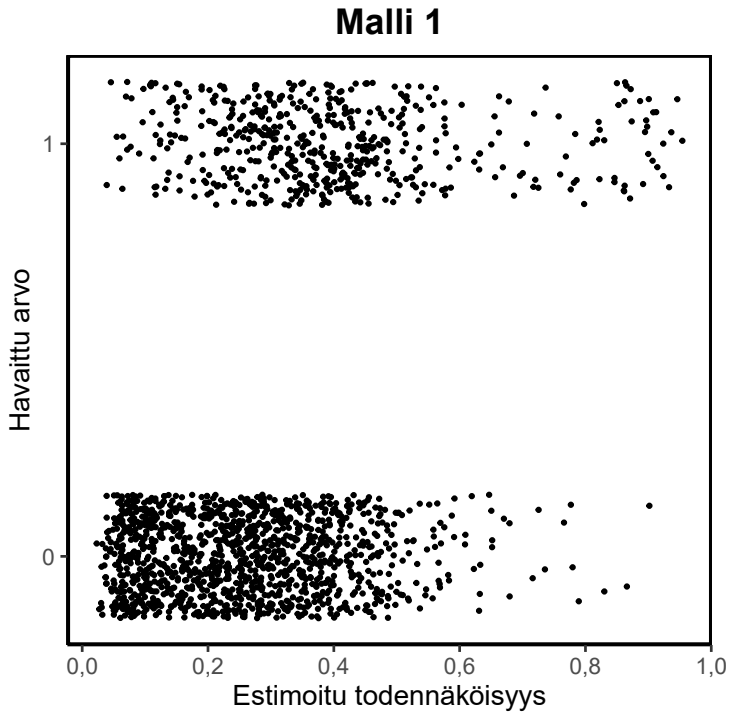
Havainnollistetaan mallien erottelukykyä myös pistekaaviolla, jossa x-akselilla on mallin estimoima todennäköisyys $\hat{\pi}(\mathbf{x})$ ja y-akselilla vasteen havaittu arvo. Kun y-akselin arvoihin lisätään hieman kohinaa, pisteet ja pisteiden ryhmittymät näkyvät kuvasta selkeämmin. Pistekaavioon piirtyy kaksi kaistaa. Ylemmällä kaistalla on pisteet, joilla havaitaan arvo $y = 1$. Alemmalla kaistalla on pisteet, joilla havaitaan arvo $y = 0$. Jos mallilla olisi hyvä erottelukyky, ylempi kaista olisi oikealla ja alempi kaista vasemmalla. Tällöin malli estimoi tapahtumille $y = 1$ suuria todennäköisyyksiä ja tapahtumille $y = 0$ pieniä todennäköisyyksiä. Täydellisen erottelukyvyn tapauksessa ylemmällä ja

Taulukko 3.5. *Hosmer–Lemeshow-testiin liittyvät ryhmät mallille 2.*

Ryhmä	Väli	y_0	y_1	$\hat{y}_0$	$\hat{y}_1$
1	[1,81e-07; 0,0144]	543	7	545,07	4,93
2	(0,0144; 0,0225]	537	12	538,78	10,22
3	(0,0225; 0,0297]	540	10	535,69	14,31
4	(0,0297; 0,0377]	532	17	530,46	18,54
5	(0,0377; 0,0453]	521	29	527,12	22,88
6	(0,0453; 0,0559]	514	35	521,56	27,44
7	(0,0559; 0,0697]	513	36	514,66	34,34
8	(0,0697; 0,0904]	509	41	506,24	43,76
9	(0,0904; 0,155]	498	51	486,00	63,00
10	(0,155; 0,602]	393	157	394,43	155,57



Kuva 3.1. *Mallien ROC-kuvaajat ja niiden alle jäävä ala AUC.*



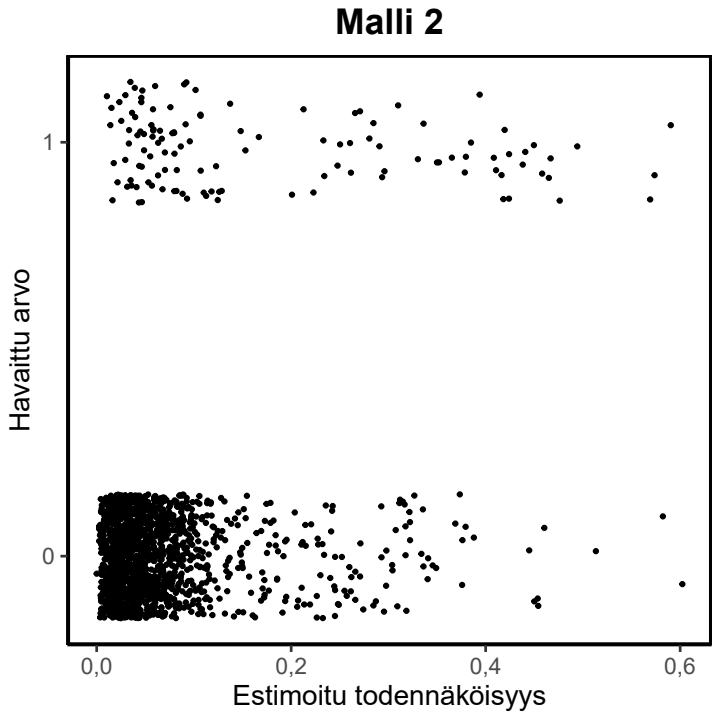
Kuva 3.2. *Mallin 1 erottelukykyä kuvaava pistekaavio.*

alemmalla kaistalla ei ole päällekkäisyyttä [23]. Koska mallissa 1 havaintoja on 17911 ja mallissa 2 havaintoja on 5495, jätetään osa pisteistä piirtämättä. Tällöin pistekaavioista on helpompi saada selvää. Mallista 1 piirretään 10 % havainnoista ja mallista 2 piirretään 30 % havainnoista. Mallin 1 havainnot ovat pistekaaviossa 3.2 ja mallin 2 havainnot ovat pistekaaviossa 3.3.

Pistekaaviosta 3.2 huomataan, että mallin 1 kaistoilla on paljon päällekkäisyyttä. Alemmalla kaistalla on vahva ryhmittymä vasemmalla ja kaista ohenee huomattavasti estimoidun todennäköisyyden kasvaessa yli arvon 0,5. Ylemmällä kaistalla ei ole näin vahvaa ryhmittymää, vaan pisteet ovat jakautuneet tasaisemmin. Ylempi kaista ei ole vasemmalta kuitenkaan yhtä tiheä kuin alempi kaista. Silmämääräisesti näyttää kuitenkin siltä, ettei mallilla 1 ole erottelukykyä kovinkaan paljon.

Pistekaaviosta 3.3 huomataan, että mallissa 2 pisteitä arvolla 1 on huomattavasti vähemmän, kuin pisteitä arvolla 0. Tästä voidaan päätellä, että kuolemaan johtanut onnettomuus on paljon harvinaisempi kuin loukkaantumiseen johtanut onnettomuus. Ylemmällä kaistalla ei ole havaittavissa kovin vahvaa ryhmittymää. Alemmalla kaistalla on hyvin vahva ryhmittymä vasemmalla estimoitujen todennäköisyyksien 0,0 ja 0,1 välissä. Silmämääräisesti näyttää siltä, ettei myöskään mallilla 2 ole erottelukykyä kovinkaan paljon.

Kuten aiemmin mainittiin, mallien testaus samalla aineistolla, jolla ne on muodostettu, antaa yleensä liian optimistisen kuvan mallien hyvyydestä ja erottelukyvystä. Niinpä saatuihin tuloksiin täytyy suhtautua varauksella. Jos testit suoritettaisiin toisella aineistolla, voisi käydä niin, että mallit näyttäisivät merkkejä huonosta yhteensopivuudesta ja erottelukyvystä.



Kuva 3.3. *Mallin 2 erottelukykyä kuvaava pistekaavio.*

3.4 Mallien tulkinta

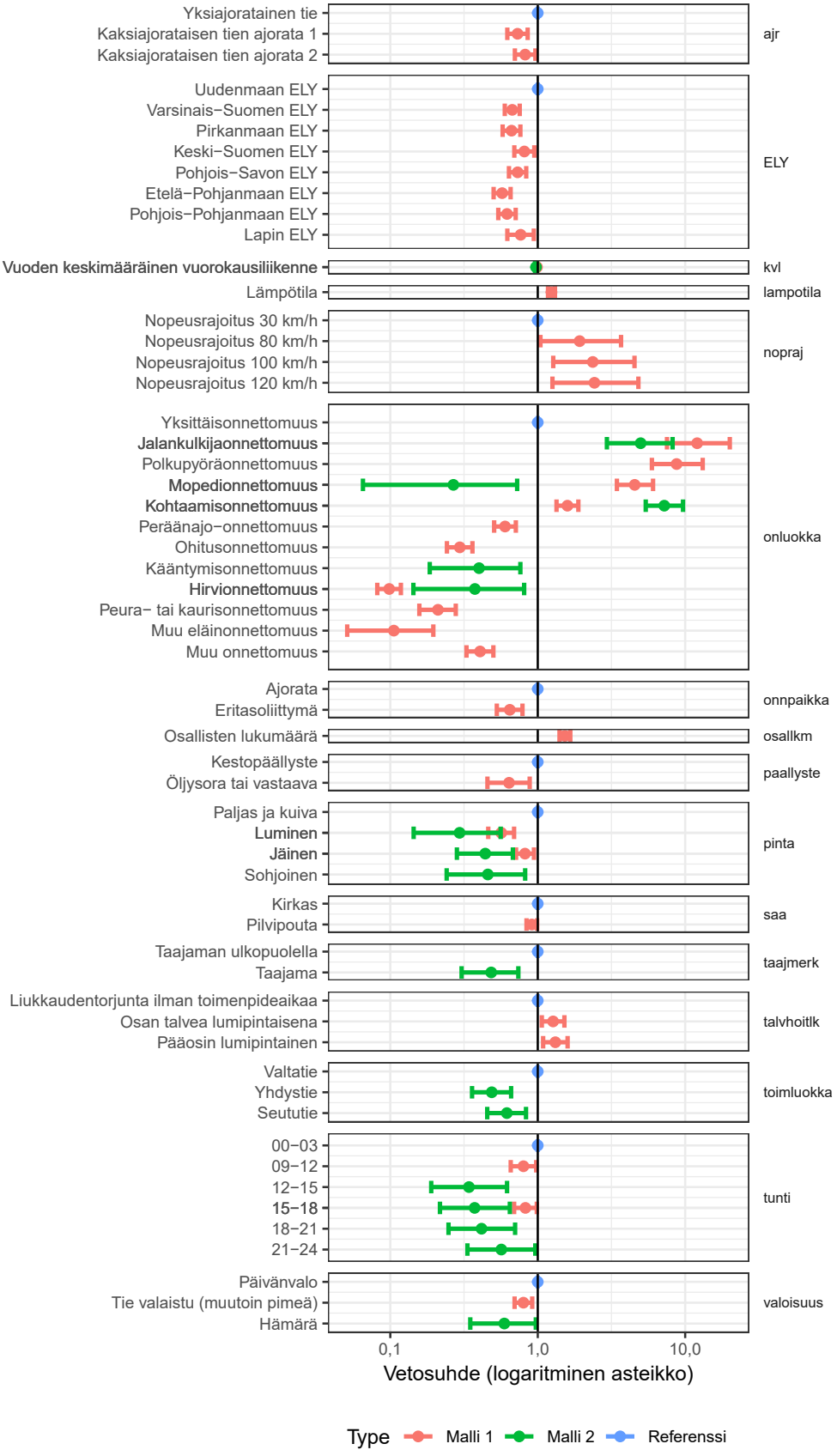
Tutkitaan onnettomuuksien vakavuuteen vaikuttavia tekjöitä tarkastelemalla mallien tuottamia vetosuhteita. Erityisen kiinnostavia ovat selittäjät, jotka ovat merkitseviä ja joilla on poikkeavan suuri tai pieni vetosuhde. Poikkeavan suuret tai pienet vetosuhteen arvot kertovat selittäjästä, jolla on vahva vaikutus onnettomuuden vakavuuteen.

Käytetään selittäjien merkitsevyyden tarkastelussa riskitasoa $\alpha = 0,05$. Kuvaan 3.4 on piirretty merkitsevien selittäjien vetosuhteet ja vetosuhteiden 95 % luottamusvälit molemmista malleista. Kuvaan on piirretty myös diskreettien selittäjien c yksikön muutoksen vetosuhteet taulukosta 3.2. Kategoristen selittäjien referenssimuuttuja on esitetty sinisellä pallolla. Referenssimuuttujan vetosuhde on aina 1. Seuraavissa alaluvuissa tulkitaan kummankin mallin merkitseviä selittäjiä. Kunkin selittäjän vetosuhteen estimaatti ja vetosuhteen 95 % luottamusvälin estimaatti on ilmoitettu suluissa.

3.4.1 Malli 1

Ensimmäisen mallin tulosteesta B.2 ja kuvasta 3.4 huomataan, että muuttujista erityisesti onnettomuusluokka (Onluokka) vaikuttaa kiinnostavalta. Se sisältää apumuuttujia, joilla on poikkeavan suuria ja pieniä arvoja. Selittäjän referenssimuuttuja on yksittäisonnettomuus.

Tarkastellaan aluksi onnettomuusluokan selittäjiä, joilla on suuri vetosuhde. Näissä henkilövahingon riski on suurempi kuin yksittäisonnettomuudessa. Erityisen suuret vetosuhteet esiintyvät kevyen liikenteen onnettomuuksissa. Henkilövahinkoon johtavan onnettomuuden veto on jalakul-



Kuva 3.4. *Mallien merkitsevien* ($\alpha = 0,05$) *selittäjien vetosuhteet ja niiden* 95 % *luottamusvälit.*

kijaonnettomuudessa noin 12-kertainen (12,06, [7,394; 19,65]) ja polkupyöräonnettomuudessa noin 9-kertainen (8,745, [5,884, 13,00]) verrattuna yksittäisonnettomuuteen. Mopedionnettomuudessa henkilövahinkoon johtavan onnettomuuden veto on noin 4-kertainen (4,546, [3,429; 6,028]) verrattuna yksittäisonnettomuuteen. Kohtaamisonnettomuudessa henkilövahinkoon johtavan onnettomuuden veto on noin 59 % (1,590, [1,341; 1,886]) suurempi kuin yksittäisonnettomuudessa.

Tarkastellaan sitten onnettomuusluokan selittäjiä, joilla on pieni vetosuhde. Näissä henkilövahingon riski on pienempi kuin yksittäisonnettomuudessa. Erityisen pienet vetosuhteet ovat eläinonnettomuuksissa. Henkilövahinkoon johtavan onnettomuuden veto on hirvionnettomuudessa noin 90 % (0,09830, [0,08176; 0,1180]) ja peura- tai kaurisonnettomuudessa noin 79 % (0,2105, [0,1584; 0,2800]) pienempi kuin yksittäisonnettomuudessa. Muussa eläinonnettomuudessa henkilövahinkoon johtavan onnettomuuden veto on noin 89 % (0,1057, [0,05433; 0,2060]) pienempi kuin yksittäisonnettomuudessa. Myös ohitusonnettomuudessa (0,2961, [0,2426; 0,3610]), peräänajoonnettomuudessa (0,6001, [0,5066; 0,7110]) ja muussa onnettomuudessa (0,4063, [0,3294; 0,5010]) on yksittäisonnettomuutta pienempi veto.

Toinen selittäjä, jolla on suuria vetosuhteita on nopeusrajoitus (Nopraj). Selittäjän referenssimuuttujana on 30 km/h nopeusrajoitus. Henkilövahinkoon johtavan onnettomuuden veto nopeusrajoituksella 80 km/h on noin 92 % (1,924, [1,028; 3,601]) suurempi kuin nopeusrajoituksella 30 km/h. Luottamusväli on melko lähellä arvoa 1, joten voi olla, etteivät vedot eroakaan toisistaan. Nopeusarjoituksilla 100 km/h (2,362, [1,255; 4,444]) ja 120 km/h (2,425 [1,242; 4,733]) henkilövahinkoon johtavan onnettomuuden veto on noin 2,4-kertainen verrattuna nopeusrajoitukseen 30 km/h. Näilläkin nopeusrajoituksilla on laajat luottamusvälit, joten vetosuhteiden oikeat arvot voivat vaihdella suhteellisen laajalla alueella. On kuitenkin varmaa, että yli 80 km/h nopeusrajoituksilla henkilövahingon riski on suurempi kuin 30 km/h nopeusrajoituksella.

Selittäjä ajorata (Ajr) on esitetty kolmella kategorialla, joista yksiajoratainen tie on referenssimuuttuja. Kaksiajorataisen tien ajorata 1 tarkoittaa tien oikeanpuoleista ajorataa tieosoitteen kasvusuunnassa. Ajorata 2 tarkoittaa tien vasemmanpuoleista ajorataa tieosoitteen kasvusuunnassa. Henkilövahinkoon johtavan onnettomuuden veto on ajoradalla 1 noin 27 % (0,7294, [0,6222; 0,8550]) ja ajoradalla 2 noin 18 % (0,8218, [0,7000; 0,9650]) pienempi kuin yksiajorataisella tiellä. Voidaan siis sanoa, että kaksiajorataisella tiellä henkilövahingon riski on pienempi kuin yksiajorataisella tiellä.

Selittäjän ELY referenssimuuttujana on Uudenmaan ELY-keskus. Henkilövahinkoon johtavan onnettomuuden veto on Varsinais-Suomessa noin 33 % (0,6716, [0,5977; 0,7550]), Pirkanmaalla noin 34 % (0,6642, [0,5783; 0,7630]), Keski-Suomessa noin 19 % (0,8114, [0,6942; 0,9480]), Pohjois-Savossa noin 27 % (0,7303, [0,6382; 0,8360]), Etelä-Pohjanmaalla noin 43 % (0,5736, [0,5024; 0,6550]), Pohjois-Pohjanmaalla noin 38 % (0,6189, [0,5402; 0,7090]) ja Lapissa noin 23 % (0,7658, [0,6228; 0,9420]) pienempi kuin Uudellamaalla. Uudenmaan ELY-keskuksen alueella henkilövahingon riski näyttäisi olevan suurempi kuin muiden ELY-keskusten alueilla.

Selittäjä vuoden keskimääräinen vuorokausiliikenne (Kvl) on merkitsevä ensimmäisessä mallissa vaikkei se näy kunnolla kuvassa 3.4. Kyseessä on diskreetti selittäjä ja sen 1000 yksikön muutoksen vetosuhde on esitetty taulukossa (3.2). Tuhannen moottoriajoneuvon lisäys vuoden keskimääräi-

seen vuorokausiliikenteeseen vähentää henkilövahinkoon johtavan onnettomuuden vetoa noin 1,2 % (0,9884, [0,9840; 0,9927]). Vuoden keskimääräisen vuorokausiliikenteen kasvulla näyttäisi olevan siis henkilövahingon riskiä vähentävä vaikutus. Luottamusväli on kuitenkin lähellä arvoa 1, joten voi olla, ettei tuhannen moottoriajoneuvon muutos vuoden keskimääräisessä vuorokausiliikenteessä vaikuta henkilövahinkoon johtavan onnettomuuden vetoon.

Lämpötila on myös diskreetti selittäjä ja sen 10 yksikön muutoksen vetosuhde on esitetty taulukossa (3.2). Kymmenen asteen lämpeneminen kasvattaa henkilövahinkoon johtavan onnettomuuden vetoa noin 24 % (1,2394, [1,1742; 1,3083]). Lämpötilan nousulla näyttäisi olevan henkilövahigon riskiä lisäävä vaikutus.

Selittäjän onnettomuuspaikka (Onnpaikka) referenssimuuttujana on ajorata. Henkilövahinkoon johtavan onnettomuuden veto on eritasoliittymässä noin 35 % (0,6462, [0,5289; 0,7890]) pienempi kuin ajoradalla. Henkilövahingon riski on suurempi ajoradalla kuin eritasoliittymässä.

Osallisten lukumäärä (Osallkm) on diskreetti selittäjä ja sen vetosuhde on 1,527 ([1,404; 1,661]). Tämä tarkoittaa sitä, että jokainen osallinen kasvattaa henkilövahinkoon johtavan onnettomuuden vetoa noin 53 %. Osallisten lukumäärällä on siis henkilövahingon riskiä lisäävä vaikutus.

Selittäjän päällyste referenssimuuttujana on kestopäällyste. Henkilövahinkoon johtavan onnettomuuden veto on öljysoraisella tai vastaavalla tiellä noin 36 % (0,6389, [0,4595; 0,8880]) pienempi kuin kestopäällystetyllä tiellä. Henkilövahingon riski on suurempi kestopäällysteellä kuin öljysoralla tai vastaavalla päällysteellä.

Selittäjä pinta kuvaa tien pintaa onnettomuushetkellä. Referenssimuuttujana on paljas ja kuiva tien pinta. Henkilövahinkoon johtavan onnettomuuden veto on lumisella tiellä noin 43 % (0,5656, [0,4622; 0,6920]) ja jäisellä tiellä noin 18 % (0,8196, [0,7126; 0,9430]) pienempi kuin paljaalla ja kuivalla tiellä. Henkilövahingon riski näyttäisi olevan pienempi lumisella sekä jäisellä tien pinnalla kuin kuivalla ja paljaalla tien pinnalla.

Selittäjän sää referenssimuuttujana on kirkas sää. Henkilövahinkoon johtavan onnettomuuden veto on pilvipoutaisella säällä noin 9 % (0,9116, [0,8392; 0,9900]) pienempi kuin kirkkaalla säällä. Henkilövahingon riski näyttäisi olevan suurempi kirkkaalla säällä kuin pilvipoutaisella säällä. Luottamusvälin nojalla on kuitenkin mahdollista, ettei kirkkaan ja pilvipoutaisen sään veto juurikaan eroa toisistaan.

Selittäjän talvihoitoluokka (Talvhoitlk) referenssimuuttujana on tie, jossa suoritetaan liukkaudentorjuntaa ilman toimenpideaikaa. Henkilövahinkoon johtavan onnettomuuden veto tiellä, joka on osan talvea lumipintaisena, on noin 27 % (1,271, [1,065; 1,516]) ja tiellä, joka on pääosin lumipintainen, on noin 32 % (1,317, [1,088; 1,594]) suurempi kuin tiellä, jossa suoritetaan liukkaudentorjuntaa ilman toimenpideaikaa. Henkilövahingon riskiä näyttäisi vähentävän tehokas liukkaudentorjunta.

Selittäjä tunti kuvaa kellonaikaa onnettomuushetkellä. Selittäjän referenssimuuttujana on tunnit 00–03. Henkilövahinkoon johtavan onnettomuuden veto on aikavälillä 09–12 noin 20 % (0,7998, [0,6554; 0,9760]) ja aikavälillä 15–18 noin 17 % (0,8251, [0,6911; 0,9850]) pienempi kuin aikavälillä 00–03. Henkilövahingon riskiä näyttäisi lisäävän aikavälillä 00–03 sattunut onnettomuus.

Apumuuttujien luottamusvälit ovat melko lähellä arvoa 1, joten voi olla, että näiden vedot eivät juurikaan eroa tuntien 00–03 vedosta.

Selittäjän valoisuus referenssimuuttujana on päivänvalo. Henkilövahinkoon johtavan onnettomuuden veto tiellä, joka on pimeän aikaan valaistu, on noin 20 % (0,7997, [0,6967; 0,9180]) pienempi kuin päivänvalossa. Henkilövahingon riski näyttäisi olevan pienempi pimeän aikaan valaistulla tiellä kuin päivänvalossa.

3.4.2 Malli 2

Myös toisessa mallissa onnettomuusluokka (Onnluokka) vaikuttaisi erityisen kiinnostavalta. Suurimmat vetosuhteet ovat kohtaamis- ja jalankulkijaonnettomuuksilla. Näissä kuoleman riski on suurempi kuin yksittäisonnettomuudessa. Kuolemaan johtavan onnettomuuden veto on kohtaamisonnettomuudessa noin 7-kertainen (7,213, [5,393; 9,649]) ja jalankulkijaonnettomuudessa noin 5-kertainen (4,989, [2,993; 8,315]) verrattuna yksittäisonnettomuuteen. Kohtaamisonnettomuuden vetosuhde on huomattavasti suurempi mallissa 2 (7,2) kuin mallissa 1 (1,6). Tämä tarkoittaa sitä, että kohtaamisonnettomuus johtaa yksittäisonnettomuutta "hieman"todennäköisemmin henkilövahinkoon, mutta henkilövahinko on kohtaamisonnettomuudessa yksittäisonnettomuutta paljon todennäköisemmin kuolema kuin loukkaantuminen.

Onnettomuusluokan pienet vetosuhteet ovat mopedi-, hirvi- ja kääntymisonnettomuuksissa. Näissä kuoleman riski on pienempi kuin yksittäisonnettomuudessa. Kuolemaan johtavan onnettomuuden veto on mopedionnettomuudessa noin 73 % (0,2678, [0,08376; 0,8560]), hirvionnettomuudessa noin 63 % (0,3749, [0,1602; 0,8770]) ja kääntyisonnettouudessa noin 60 % (0,3999, [0,1987; 0,8050]) pienempi kuin yksittäisonnettomuudessa. Mopedionnettomuuden vetosuhde on mallissa 2 (0,27) huomattavasti pienempi kuin mallissa 1 (4,5). Tämä tarkoittaa sitä, että mopedionnettomuus johtaa yksittäisonnettomuutta todennäköisemmin henkilövahinkoon, mutta henkilövahinko on yksittäisonnettomuutta todennäköisemmin loukkaantuminen kuin kuolema. Näillä apumuuttujilla on laajat luottamusvälit, joten vetosuhteiden oikeat arvot voivat vaihdella suhteellisen laajalla alueella.

Vuoden keskimääräisen vuorokausiliikenteen 1000 yksikön muutoksen vetosuhde on esitetty taulukossa 3.2. Tuhannen moottoriajoneuvon lisäys vuoden keskimääräiseen vuorokausiliikenteeseen vähentää kuolemaan johtavan onnettomuuden vetoa noin 2,7 % (0,9731, [0,9595; 0,9869]). Vuoden keskimääräisen vuorokausiliikenteen kasvulla näyttäisi olevan siis kuoleman riskiä vähentävä vaikutus. Luottamusväli on kuitenkin lähellä arvoa 1, joten voi olla, ettei tuhannen moottoriajoneuvon muutos vuoden keskimääräisessä vuorokausiliikenteessä vaikuta kuolemaan johtavan onnettomuuden vetoon.

Selittäjä pinta on merkitsevä myös mallissa 2. Kuolemaan johtavan onnettomuuden veto on lumisella tiellä noin 70 % (0,2951, [0,1503; 0,5790]), jäisellä tiellä noin 56 % (0,4407, [0,2848; 0,6820]) ja sohjoisella tiellä noin 54 % (0,4586, [0,2493; 0,8440]) pienempi kuin paljaalla ja kuivalla tiellä. Kuoleman riski näyttäisi olevan pienempi lumisella, jäisellä sekä sohjoisella tien pinnalla kuin kuivalla ja paljaalla tien pinnalla.

Selittäjän taajama (Taajmerk) referenssimuuttujana on onnettomuus, joka on sattunut taajaman ulkopuolella. Kuolemaan johtavan onnettomuuden veto on taajamassa noin 52 % (0,4833, [0,3108; 0,7520]) pienempi kuin taajaman ulkopuolella. Kuolemaan johtavan onnettomuuden riski on suurempi taajaman ulkopuolella kuin taajamassa.

Selittäjän tien toiminnallinen luokka (Toimluokka) referenssimuuttujana on onnettomuus, joka on sattunut valtatiellä. Kuolemaan johtavan onnettomuuden veto on seututiellä noin 38 % (0,6174, [0,4563; 0,8350]) ja yhdystiellä noin 51 % (0,4876, [0,3595; 0,6610]) pienempi kuin valtatiellä. Kuolemaan johtavan onnettomuuden riski on suurempi valtatiellä kuin yhdys- tai seututiellä.

Selittäjä tunti on merkitsevä myös mallissa 2. Kuolemaan johtavan onnettomuuden veto on aikavälillä 12–15 noin 66 % (0,3416, [0,1891; 0,6170]), aikavälillä 15–18 noin 63 % (0,3732, [0,2165; 0,6440]), aikavälillä 18–21 noin 58 % (0,4158, [0,2475; 0,6990]) ja aikavälillä 21–24 noin 43 % (0,5664, [0,3339; 0,9610]) pienempi kuin aikavälillä 00–03. Sen lisäksi, että aikavälillä 00–03 sattunut onnettomuus lisää henkilövahingon riskiä, se lisää myös kuoleman riskiä.

Selittäjä valoisuus on merkitsevä myös mallissa 2. Kuolemaan johtavan onnettomuuden veto on hämärässä noin 41 % (0,5934, [0,3564; 0,9880]) pienempi kuin päivänvalossa. Kuoleman riski näyttäisi olevan pienempi hämärässä kuin päivänvalossa. Luottamusväli on kuitenkin melko lähellä arvoa 1, joten voi olla, etteivät hämärässä ja päivänvalossa sattuneiden kuolemaan johtaneiden onnettomuuksien vedot juurikaan eroa toisistaan.

3.4.3 Aiemmat tutkimukset

Tieliikennonnettomuuksien vakavuutta on tutkittu laajasti useilla erilaisilla menetelmillä, kuten tässäkin työssä käytetyllä binäärisellä logistisella regressiolla [17][25]. Tässä alaluvussa tämän työn tuloksia vertaillaan aiempiin tutkimuksiin, joissa on tutkittu tieliikenneonnettomuuksien vakavuuteen vaikuttavia tekijöitä. Tarkasteltavia tutkimuksia ei välttämättä ole toteutettu logistisen regression avulla.

Tässä työssä kohtaamisonnettomuuksissa on suurempi vetosuhde kuin muissa onnettomuusluokissa, joissa osallisena on jokin muu moottoriajoneuvo kuin mopo. Tämä tarkoittaa sitä, että kohtaamisonnettomuudet johtavat muita moottoriajoneuvoihin liittyviä onnettomuusluokkia todennäköisemmin sekä henkilövahinkoon että kuolemaan. Tämä vastaa tutkimusten [10][14] tuloksia, joiden mukaan kohtaamisonnettomuus johtaa usein vakavampaan onnettomuuteen kuin muut onnettomuusluokat. Kohtaamisonnettomuudet ovat olleet Suomessakin yksi pääteiden suurimmista kuolemaan johtavista onnettomuustyypeistä [27].

Tässä työssä suurilla nopeusrajoituksilla on henkilövahingon riskiä lisäävä vaikutus. Tämä vastaa tutkimusten [5][14][19] tuloksia, joissa suurempi nopeusrajoitus oli yhteydessä vakavampaan onnettomuuteen. Lee ja Li toteavat tutkimuksessaan [14], että suuremmalla nopeusrajoituksella ajoneuvolla on suurempi vauhti, mikä johtaa vakavampaan törmäykseen onnettomuustilanteessa. Tämä puolestaan johtaa todennäköisemmin vakavampaan onnettomuuteen. Myös Suomessa on raportoitu nopeusrajoitusten lisäävän onnettomuuden vakavuutta [27].

Tässä työssä aikavälillä 00–03 sattuneilla onnettomuuksilla havaitaan olevan sekä henkilövahingon

että kuoleman riskiä lisäävä vaikutus. Tämä vastaa tutkimusten [10][14] tuloksia, joiden mukaan onnettomuuden vakavuus on suurempi yöllä kuin päivällä. Tutkimusten mukaan tulos voi johtua huonommasta valaistuksesta ja vauhdikkaammasta ajosta yöllä, kun muuta liikennettä ei ole.

Tässä työssä tiellä, joka on pimeän aikaan valaistu, sattuneella onnettomuudella on pienempi henkilövahingon riski kuin päivänvalossa sattuneella onnettomuudella. Lisäksi hämärässä sattuneella onnettomuudella on pienempi kuoleman riski kuin päivänvalossa sattuneella onnettomuudella. Tämä vastaa tutkimuksen [5] tulosta, jonka mukaan päivänvalo kasvattaa henkilövahingon ja kuoleman riskiä. Toisaalta tutkimuksen [19] tulokset ovat päinvastaiset. Sen mukaan päivänvalo vähentää onnettomuuksien vakavuutta.

Tässä työssä lumisella sekä jäisellä tiellä on kuivaa ja paljasta tietä pienempi henkilövahingon ja kuoleman riski. Myös sohjoisella tiellä on kuivaa ja paljasta tietä pienempi kuoleman riski. Khattak ja Knapp esittävät tutkimuksessaan [13], että vaikka tieliikenneonnettomuuksien määrät saattavat kasvaa lumisella tiellä verrattuna kuivaan tiehen, niiden vakavuus vähenee. Myös tutkimusten [5][10][19] mukaan luminen ja jäinen tien pinta ovat yhteydessä lievempään onnettomuuteen. Tutkimuksissa arvellaan, että autoilijat reagoivat huonoon ajokeliin hidastamalla vauhtia ja ajamalla varovaisemmin, mikä johtaa lievempiin onnettomuuksiin. Malmivuo ja Kärki kertovat tutkimuksessa [16] vuodelta 2002, että Suomessa jäisen kelin onnettomuusriski (kelillä tapahtuneet onnettomuudet / kelin liikennesuorite) olisi huomattavasti korkeampi kuin kuivalla ja paljaalla tiellä. Tutkimuksessa ei kuitenkaan käsitellä onnettomuuksien vakavuutta.

Tässä työssä vuoden keskimääräisen vuorokausiliikenteen kasvulla on sekä henkilövahingon että kuoleman riskiä lievästi vähentävä vaikutus. Tulos on linjassa tutkimusten [5][10][13] kanssa, joissa keskimääräinen vuorokausiliikenteen kasvu on yhteydessä lievempään onnettomuuteen. Tämän on arveltu johtuvan siitä, että liikennenopeudet ovat hitaampia ruuhkaisemmilla teillä, mikä johtaa lievempiin onnettomuuksiin.

Tässä työssä kaksiajorataisella tiellä on pienempi henkilövahingon riski kuin yksiajorataisella tiellä. Tämä on linjassa tutkimuksen [10] kanssa. Sen mukaan yksiajorataisella tiellä onnettomuuksien vakavuus kasvaa verrattuna muihin tietyyppeihin. Myös Suomessa kaksiajorataisen tien on havaittu vähentävän onnettomuuksien vakavuutta [27]. Ruotsissa vakavia liikenneonnettomuuksia ja kuolemia on onnistuttu vähentämään tehokkaasti erottamalla ajosuunnat toisistaan keskikaiteilla [11].

Tässä työssä lämpötilan nousulla on henkilövahingon riskiä lisäävä vaikutus. Tulos on linjassa tutkimuksen [10] kanssa. Sen mukaan lämpötilan nousulla saattaa olla onnettomuuden vakavuutta lisäävä vaikutus.

Tässä työssä osallisten lukumäärällä on henkilövahingon riskiä lisäävä vaikutus. Tämä vastaa tutkimusten [4][19] tuloksia, joiden mukaan osallisten lukumäärän lisääntyessä onnettomuuden vakavuus kasvaa verrattuna kahden ajoneuvon törmäykseen. Chen et al. arvelevat tutkimuksessa [4], että tulos voisi johtua siitä, että usean ajoneuvon onnettomuudessa sattuu useampi törmäys, mikä johtaa suurempaan loukkaantumisriskiin.

Tässä työssä eläinonnettomuuksissa on muita onnettomuusluokkia pienempi henkilövahingon riski. Tämä vastaa tutkimuksen [4] tulosta, jonka mukaan eläinonnettomuuksissa on muita onnetto-

muusluokkia pienempi riski vakavaan onnettomuuteen.

Tässä työssä jalankulkijaonnettomuuksissa on korkea sekä henkilövahingon että kuoleman riski. Tämä vastaa tutkimuksen [19] tulosta, jonka mukaan jalankulkijan läsnäolo lisää onnettomuuden vakavuutta. Tulos ei ole kovin yllättävä. Lisäksi tutkimuksen [10] mukaan jalankulkijat ovat pyöräilijöitä alttiimpia vakavammille onnettomuuksille, mikä vastaa myös tämän työn tulosta. Jalankulkijaonnettomuuksissa on polkupyöräonnettomuuksia suurempi vetosuhde mallissa 1.

4. YHTEENVETO

Tässä työssä tutkittiin Suomessa vuosina 2017–2021 sattuneiden tieliikenneonnettomuuksien vakavuuteen vaikuttavia tekijöitä kahden erilaisen binäärisen logistisen regressiomallin avulla. Mallin 1 avulla tarkasteltiin tekijöitä, jotka erottelevat henkilövahinkoon johtavan tieliikenneonnettomuuden tieliikenneonnettomuudesta, joka ei johda henkilövahinkoon. Mallin 2 avulla tarkasteltiin tekijöitä, jotka erottelevat kuolemaan johtavan tieliikenneonnettomuuden tieliikenneonnettomuudesta, joka johtaa loukkaantumiseen.

Mallit muodostettiin askeltavalla menetelmällä käyttämällä selittäjän sisällyttämiseen riskitasoa $\alpha_{\mathrm{E}} = 0,15$ ja selittäjän poistamiseen riskitasoa $\alpha_{\mathrm{R}} = 0,20$. Malleista karsittiin selittäjät, joissa ilmeni numeerisia ongelmia.

Mallien hyvyyttä tutkittiin Hosmer–Lemeshow-testin avulla. Kumpikaan malli ei näyttänyt merkkejä huonosta hyvyydestä. Mallien erottelukykyä tutkittiin puolestaan ROC-kuvaajan ja AUC-arvon avulla. Sen lisäksi erottelukykyä havainnollistettiin pistekaavion avulla. Mallille 1 AUC = 0,739 ja mallille 2 AUC = 0,758, joten kummankin mallin erottelukyky oli AUC-arvon mukaan hyväksyttävä. Pistekaavioista ei kuitenkaan näkynyt merkkejä kovin hyvästä mallien erottelukyvystä, mutta lievää erottelukykyä oli havaittavissa. Mikäli malleja käytettäisiin onnettomuuksien vakavuuden ennustamiseen, tulisi mallien validointi suorittaa uudella aineistolla.

Vakavuuteen vaikuttavia tekjöitä tutkittiin tulkitsemalla mallien merkitsevien ($\alpha = 0,05$) selittäjien estimoituja vetosuhteita. Erityisesti onnettomuusluokalla näyttäisi olevan suuri vaikutus onnettomuuden vakavuuteen kummassakin mallissa. Henkilövahingon riski on suurin jalankulkija- ja polkupyöräonnettomuuksissa. Jalankulkijaonnettomuus lisää vahvasti myös kuoleman riskiä. Kevyen liikenteen alttius vakavammille onnettomuuksille on varsin oletettu tulos. Mopedionnettomuudet lisäävät henkilövahingon riskiä enemmän kuin yksittäisonnettomuudet, mutta henkilövahinko on mopedionnettomuudessa yksittäisonnettomuutta todennäköisemmin loukkaantuminen kuin kuolema. Tulokseen voisi vaikuttaa esimerkiksi mopojen hitaat ajonopeudet, mutta tulosta voisi tutkia tarkemmin. Kohtaamisonnettomuudet lisäävät henkilövahingon riskiä ja ne lisäävät kuoleman riskiä huomattavasti. Kohtaamisonnettomuudet ovat olleet merkittävässä roolissa Suomen tieliikennekuolemissa, joten tulos on odotettu.

Tässä työssä tiellä, joka on pimeän aikaan valaistu, on pienempi henkilövahingon riski kuin päivänvalossa. Lisäksi hämärässä kuoleman riski on pienempi kuin päivänvalossa. Aiemmissa tutkimuksissa on saatu samankaltaisia, mutta myös päinvastaisia tuloksia. Valoisuuden ja onnettomuuden vakavuuden välistä suhdetta voisi tutkia tarkemmin.

Tässä työssä lumisella ja jäisellä tien pinnalla oli kuivaa ja paljasta tien pintaa pienempi henkilöva-

hingon ja kuoleman riski. Kuoleman riskiä pienensi myös sohjoinen tien pinta. Talvisilla keleillä on esitetty olevan onnettomuuksien vakavuutta vähentävä vaikutus myös aiemmissa tutkimuksissa. Toisaalta onnettomuuksien määrän on raportoitu kasvavan talvisilla keleillä. Suomessa talvikelien on raportoitu kasvattavan onnettomuusriskiä huomattavasti. Suomen talvikelien onnettomuusriskiä ja onnettomuuksien vakavuutta sekä niiden välistä suhdetta voisi tutkia tarkemmin.

Koska saadut tulokset ovat suurimmalta osin linjassa aiempien tutkimusten kanssa, näyttäisivät luodut mallit soveltuvan työn aiheen tutkimiseen. Malleja voitaisiin kuitenkin parantaa esimerkiksi sisällyttämällä malleihin selittäjien välisiä yhdystermejä, jolloin vakavuuteen vaikuttavat yksityiskohtaisemmat tekijät tulisivat esille. Mallien muodostuksessa voitaisiin käyttää myös esimerkiksi tieturvallisuusalan tietämystä pelkän tilastollisen tarkastelun sijaan. Mallit sisälsivät tiehen ja olosuhteisiin liittyviä tekijöitä, mutta myös kuskiin ja ajoneuvoon liittyvät tekijät, kuten kuskin kokemus ja ajoneuvon ikä, saattavat vaikuttaa onnettomuuden vakavuuteen.

Lisäksi mallien selittäjien välisiin riippuvuuksiin voitaisiin kiinnittää enemmän huomioita. Esimerkiksi taajamissa on hiljaisemmat nopeusrajoitukset kuin taajamien ulkopuolella. Toisin sanoen alhainen nopeusrajoitus on todennäköisempi taajamassa kuin taajaman ulkopuolella eivätkä korkeat nopeusrajoitukset ole edes mahdollisia taajamissa. Täten nopeusrajoitus ja onnettomuuden sattuminen taajamassa eivät ole riippumattomia muuttujia. Tällaiset riippuvuudet saattavat vääristää mallista saatavia tuloksia.

LÄHTEET

[1] A. Agresti. *An introduction to categorical data analysis*. 3rd ed. Wiley series in probability and mathematical statistics. Newark: Wiley, 2019.

[2] A. C. Alba, T. Agoritsas, M. Walsh, S. Hanna, A. Iorio, P. J. Devereaux, T. McGinn ja G. Guyatt. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *JAMA : the journal of the American Medical Association* 318.14 (2017), pp. 1377–1384.

[3] W. D. Berry. *Multiple regression in practice*. Quantitative applications in the social sciences ; no. 07-050. Newbury Park, [Calif.] ; SAGE, 1985.

[4] C. Chen, G. Zhang, H. Huang, J. Wang ja R. A. Tarefder. Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. *Accident analysis and prevention* 96 (2016), pp. 79–87.

[5] K. Haleem, S. Azam, U. Manepalli ja M. Mays. Identifying and comparing the injury severity risk factors on rural freeways in different states in the United States. *International journal of injury control and safety promotion* 26.4 (2019), pp. 343–353.

[6] G. Heinze. A comparative investigation of methods for logistic regression with separated or nearly separated data: COMPARATIVE INVESTIGATION OF METHODS FOR LOGISTIC REGRESSION. *Statistics in medicine* 25.24 (2006), pp. 4216–4226.

[7] G. Heinze ja M. Schemper. A solution to the problem of separation in logistic regression. *Statistics in medicine* 21.16 (2002), pp. 2409–2419.

[8] J. Hilbe. *Logistic regression models*. Chapman & Hall/CRC texts in Statistical Science Series. Boca Raton: Chapman & Hall/CRC, 2009.

[9] D. W. Hosmer, S. Lemeshow ja R. X. Sturdivant. *Applied logistic regression, third edition*. 3rd ed. Wiley series in probability and statistics. Hoboken, NJ: John Wiley ja Sons, 2013.

[10] S. Hyodo ja K. Hasegawa. Factors Affecting Analysis of the Severity of Accidents in Cold and Snowy Areas Using the Ordered Probit Model. *Asian transport studies* 7 (2021).

[11] K. Hytönen ja H. Peltola. Kustannustehokkaat keskikaideratkaisut. *Liikenneviraston tutkimuksia ja selvityksiä* 2/2016 (2016).

[12] K. In Lee ja J. J. Koval. Determination of the best significance level in forward stepwise logistic regression. *Communications in statistics. Simulation and computation* 26.2 (1997), pp. 559–575.

[13] A. Khattak ja K. Knapp. Interstate highway crash injuries during winter snow and nonsnow events: SAFETY AND HUMAN PERFORMANCE. *Transportation research record*. TRANSPORTATION RESEARCH RECORD-SERIES 1746 (2001), pp. 30–36.

[14] C. Lee ja X. Li. Analysis of injury severity of drivers involved in single- and two-vehicle crashes on highways in Ontario. *Accident analysis and prevention* 71 (2014), pp. 286–295.

[15] *Liikenneonnettomuuksista aiheutuneet taloudelliset vahingot ja niiden yksikköhinnat*. Traficom, 2022. Saatavissa (viitattu 21.11.2022): `https://liikennefakta.fi/fi/turvallisuus/`

tieliikenne/liikenneonnettomuuksista-aiheutuneet-taloudelliset-vahingot-ja-niiden.

[16] M. Malmivuo ja O. Kärki. Ajokeliin liittyvä riski. *Tiehallinnon selvityksiä* 39/2002 (2002).

[17] F. L. Mannering ja C. R. Bhat. Analytic methods in accident research: Methodological frontier and future directions. *Analytic methods in accident research* 1 (2014), pp. 1–22.

[18] S. W. Menard. *Logistic regression from introductory to advanced concepts and applications*. Thousand Oaks, Calif. ; SAGE, 2010.

[19] P. Michalaki, M. A. Quddus, D. Pitfield ja A. Huetson. Exploring the factors affecting motorway accident severity in England using the generalised ordered logistic regression model. *Journal of safety research* 55 (2015), pp. 89–97.

[20] *Next steps towards 'Vision Zero' : EU road safety policy framework 2021-2030*. European Commission, Directorate-General for Mobility ja Transport, 2020.

[21] F. C. Pampel. *Logistic regression : a primer*. 2nd ed. Quantitative applications in the social sciences. Los Angeles, CA: SAGE Publications, Inc., 2021.

[22] P. Pere. *Tilastomenetelmien perusteet*. Luentotiivistelmä. 2020.

[23] P. Royston ja D. G. Altman. Visualizing and assessing discrimination in the logistic regression model. *Statistics in medicine* 29.24 (2010), pp. 2508–2520.

[24] K. Ruohonen. *Tilastomatematiikka*. Luentotiivistelmä. 2011.

[25] P. T. Savolainen, F. L. Mannering, D. Lord ja M. A. Quddus. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident analysis and prevention* 43.5 (2011), pp. 1666–1676.

[26] *Suomen tieliikenteen turvallisuus kansainvälisessä vertailussa*. Traficom, 2022. Saatavissa (viitattu 21.11.2022): https://tieto.traficom.fi/fi/tilastot/suomen-tieliikenteen-turvallisuus-kansainvalisessa-vertailussa.

[27] *Turvallinen päätieverkko - uusia ratkaisuja kustannustehokkaasti*. Tiehallinto, 2007.

LIITE A: ASKELTAVAN MENETELMÄN OHJELMAKOODI

```
stepwise_LRT <- function(model, alpha_E, alpha_R) {
    # Luo logistisen regressiomallin kutsumalla vuorotellen
    # funktioita forward_LRT (eteenpäinvalinta) ja backward_LRT
    # (poistovalinta).
    #
    # Args:
    #   model: Logistinen regressiomalli, jolle halutaan
    #   suorittaa askeltava menetelmä.
    #   alpha_E: Eteenpäinvalinnan riskitaso.
    #   alpha_R: Poistovalinnan riskitaso.
    #
    # Returns:
    #   Logistinen regressiomalli, jolle on suoritettu
    #   askeltava menetelmä.
    #
    original <- model
    model <- forward_LRT(model, alpha_E, 1)
    while (!identical(original, model)) {
        original <- model
        model <- forward_LRT(model, alpha_E, 1)
        model <- backward_LRT(model, alpha_R, 1)
    }
    return(model)
}
```

```
forward_LRT <- function(model, alpha, num_of_iter = 999) {
    # Suorittaa eteenpäinvalintaa halutun määrän kertoja
    # käyttäen uskottavuussuhdetestiä merkitsevyyden
    # testaamiseen.
    #
    # Args:
    #   model: Logistinen regressiomalli, jolle halutaan
    #   suorittaa eteenpäinvalinta.
```

```
    #   alpha: Eteenpäinvalinnan riskitaso.
    #   num_of_iter: Suorituskertojen lukumäärä. Oletuksena
    #   suorittaa eteenpäinvalinnan loppuun asti. Askeltavassa
    #   menetelmässä tälle annetaan arvoksi 1.
    #
    # Returns:
    #   Kutsuu rekursiivisesti itseään niin kauan kuin
    #   mahdollista. Lopuksi palauttaa logistisen
    #   regressiomallin, jolle on suoritettu
    #   eteenpäinvalintaa haluttu määrä kertoja.
    #
    lowest_p <- 1
    var_name <- ""
    var_index <- 0
    dataf <- model$data
    form <- as.formula(paste("~ . +",
        paste(colnames(dataf[2:length(dataf)]), collapse = "+")))
    add <- add1(model, scope = form, test = "LRT")
    for (i in 2:(length(add$Pr))) {
        if ((!is.na(add$Pr[i])) && (add$Pr[i] < lowest_p)) {
            lowest_p <- add$Pr[i]
            var_name <- colnames(dataf[!(colnames(dataf) %in%
                attr(model$terms, "term.labels"))])[i]
            var_index <- i
        }
    }
    if ((lowest_p < alpha) && (num_of_iter > 0)) {
        model <- update(model, as.formula(paste(". ~ . +", var_name)))
    }
    else {
        return(model)
    }
    return(forward_LRT(model, alpha, num_of_iter-1))
}

backward_LRT <- function(model, alpha, num_of_iter = 999) {
    # Suorittaa poistomenetelmää halutun määrän kertoja
    # käyttäen uskottavuussuhdetestiä merkitsevyyden
    # testaamiseen.
    #
    # Args:
    #   model: Logistinen regressiomalli, jolle halutaan
    #   suorittaa poistovalinta.
```

```
9       #   alpha: Poistovalinnan riskitaso.
10      #   num_of_iter: Suorituskertojen lukumäärä. Oletuksena
11      #   suorittaa poistovalinnan loppuun asti. Askeltavassa
12      #   menetelmässä tälle annetaan arvoksi 1.
13      #
14      # Returns:
15      #   Kutsuu rekursiivisesti itseään niin kauan kuin
16      #   mahdollista. Lopuksi palauttaa logistisen
17      #   regressiomallin, jolle on suoritettu
18      #   poistovalintaa haluttu määrä kertoja.
19      #
20      highest_p <- 0
21      var_name <- ""
22      var_index <- 0
23      dataf <- model$data
24      drop = drop1(model, test = "LRT")
25      for (i in 2:length(drop$Pr)) {
26          if (drop$Pr[i] > highest_p) {
27              highest_p <- drop$Pr[i]
28              var_name <- attr(model$terms, "term.labels")[i-1]
29              var_index <- i-1
30          }
31      }
32      if ((highest_p > alpha) && (num_of_iter > 0)) {
33          model <- update(model, paste("~ . -", var_name))
34      }
35      else {
36          return(model)
37      }
38      return(backward_LRT(model, alpha, num_of_iter-1))
39  }
```

LIITE B: MALLIEN TULOSTEET

R-ohjelmiston komentoa `summary` on muokattu niin, että se tulostaa kunkin selittäjän vetosuhteen estimaatin ja vetosuhteen 95 % luottamusvälin estimaatin.

B.1 Alkuperäinen malli 1

```
Call:
glm(formula = aineisto1$y ~ onluokka + lampotila + maakunta +
    nopraj + kvl + osallkm + pinta + talvhoitlk + ajr + onnpaikka +
    saa + ELY + paallyste + valoisuus + tunti + taajmerk,
    family = binomial(link = "logit"), data = aineisto1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4165  -0.8409  -0.5498   1.0349   2.6893

Coefficients:
               Estimate  Std. Error          OR       lower      upper Pr(>|z|)
(Intercept)  -2.829e+00   8.574e-01   5.906e-02   1.100e-02  3.170e-01 0.000967 ***
onluokka2    -1.599e-01   8.889e-02   8.522e-01   7.160e-01  1.014e+00 0.072089 .
onluokka3    -1.216e+00   1.018e-01   2.965e-01   2.429e-01  3.620e-01  < 2e-16 ***
onluokka4    -1.856e-02   8.403e-02   9.816e-01   8.326e-01  1.157e+00 0.825225
onluokka5     4.633e-01   8.740e-02   1.589e+00   1.339e+00  1.886e+00 1.15e-07 ***
onluokka6    -5.067e-01   8.670e-02   6.025e-01   5.083e-01  7.140e-01 5.10e-09 ***
onluokka7     1.504e+00   1.446e-01   4.501e+00   3.391e+00  5.976e+00  < 2e-16 ***
onluokka8     2.174e+00   2.030e-01   8.797e+00   5.909e+00  1.310e+01  < 2e-16 ***
onluokka9     2.442e+00   2.494e-01   1.150e+01   7.054e+00  1.875e+01  < 2e-16 ***
onluokka10   -2.317e+00   9.426e-02   9.857e-02   8.194e-02  1.190e-01  < 2e-16 ***
onluokka11   -1.460e+00   1.462e-01   2.322e-01   1.743e-01  3.090e-01  < 2e-16 ***
onluokka12   -2.270e+00   3.398e-01   1.033e-01   5.307e-02  2.010e-01 2.39e-11 ***
onluokka13   -8.990e-01   1.072e-01   4.070e-01   3.298e-01  5.020e-01  < 2e-16 ***
lampotila     2.154e-02   2.770e-03   1.022e+00   1.016e+00  1.027e+00 7.41e-15 ***
maakunta1     1.638e+00   8.157e-01   5.147e+00   1.040e+00  2.546e+01 0.044581 *
maakunta2     9.310e-01   5.934e-01   2.537e+00   7.929e-01  8.118e+00 0.116650
maakunta4     6.615e-01   5.953e-01   1.938e+00   6.033e-01  6.223e+00 0.266475
maakunta5     1.566e+00   8.193e-01   4.788e+00   9.611e-01  2.386e+01 0.055935 .
maakunta6    -1.295e+01   1.515e+02   2.386e-06  2.524e-135 2.255e+123 0.931910
maakunta7     1.535e+00   8.191e-01   4.642e+00   9.321e-01  2.312e+01 0.060904 .
maakunta8    -1.323e+01   3.247e+02   1.790e-06  6.688e-283 4.789e+270 0.967495
maakunta9    -1.286e+01   3.247e+02   2.606e-06  9.738e-283 6.973e+270 0.968417
maakunta10   -4.287e-01   8.650e-01   6.514e-01   1.195e-01  3.549e+00 0.620167
maakunta11   -5.123e-01   8.632e-01   5.991e-01   1.103e-01  3.253e+00 0.552836
```

```
maakunta12   -3.242e-01   8.660e-01   7.231e-01   1.325e-01  3.948e+00 0.708128
maakunta13   -8.954e-01   6.318e-01   4.084e-01   1.184e-01  1.409e+00 0.156391
maakunta14    1.018e+00   1.148e+00   2.768e+00   2.919e-01  2.626e+01 0.374995
maakunta15   -1.712e-02   1.149e+00   9.830e-01   1.035e-01  9.339e+00 0.988111
maakunta16    4.662e-01   1.154e+00   1.594e+00   1.661e-01  1.530e+01 0.686165
maakunta17    1.102e+01   2.209e+02   6.114e+04  5.668e-184 6.595e+192 0.960209
maakunta18    1.062e+01   2.209e+02   4.081e+04  3.783e-184 4.403e+192 0.961667
maakunta19   -1.538e+00   1.178e+00   2.148e-01   2.133e-02  2.163e+00 0.191848
nopraj40     -1.278e-01   3.287e-01   8.801e-01   4.620e-01  1.676e+00 0.697515
nopraj50     -1.739e-03   3.199e-01   9.983e-01   5.332e-01  1.869e+00 0.995663
nopraj60      3.510e-01   3.195e-01   1.421e+00   7.594e-01  2.657e+00 0.271934
nopraj70      6.163e-01   3.421e-01   1.852e+00   9.472e-01  3.621e+00 0.071624 .
nopraj80      6.004e-01   3.195e-01   1.823e+00   9.746e-01  3.409e+00 0.060177 .
nopraj100     8.130e-01   3.223e-01   2.255e+00   1.199e+00  4.241e+00 0.011652 *
nopraj120     8.428e-01   3.411e-01   2.323e+00   1.190e+00  4.533e+00 0.013486 *
kvl          -1.254e-05   2.340e-06   1.000e+00   1.000e+00  1.000e+00 8.33e-08 ***
osallkm       4.171e-01   4.297e-02   1.518e+00   1.395e+00  1.651e+00  < 2e-16 ***
pinta2       -6.542e-02   6.168e-02   9.367e-01   8.300e-01  1.057e+00 0.288866
pinta3        8.968e-02   2.066e-01   1.094e+00   7.295e-01  1.640e+00 0.664302
pinta4       -5.458e-01   1.034e-01   5.794e-01   4.731e-01  7.090e-01 1.29e-07 ***
pinta5       -6.562e-02   1.106e-01   9.365e-01   7.540e-01  1.163e+00 0.552949
pinta6       -1.783e-01   7.180e-02   8.367e-01   7.268e-01  9.630e-01 0.013006 *
pinta7        1.160e-01   1.326e-01   1.123e+00   8.660e-01  1.456e+00 0.381832
talvhoitlk1   2.491e-02   7.628e-02   1.025e+00   8.828e-01  1.191e+00 0.744001
talvhoitlk2   1.407e-01   9.824e-02   1.151e+00   9.494e-01  1.395e+00 0.152155
talvhoitlk3   2.278e-01   9.157e-02   1.256e+00   1.049e+00  1.503e+00 0.012860 *
talvhoitlk4   2.170e-02   1.355e-01   1.022e+00   7.836e-01  1.333e+00 0.872786
talvhoitlk5   2.671e-01   9.900e-02   1.306e+00   1.076e+00  1.586e+00 0.006984 **
talvhoitlk6   2.197e-01   1.245e-01   1.246e+00   9.760e-01  1.590e+00 0.077592 .
ajr1         -3.220e-01   8.188e-02   7.247e-01   6.172e-01  8.510e-01 8.40e-05 ***
ajr2         -2.001e-01   8.245e-02   8.187e-01   6.965e-01  9.620e-01 0.015228 *
onnpaikka2   -9.419e-02   2.323e-01   9.101e-01   5.773e-01  1.435e+00 0.685079
onnpaikka3   -2.764e-01   7.441e-01   7.585e-01   1.764e-01  3.261e+00 0.710262
onnpaikka4   -2.727e-01   4.250e-01   7.613e-01   3.310e-01  1.751e+00 0.521102
onnpaikka6   -4.120e-02   1.626e-01   9.596e-01   6.977e-01  1.320e+00 0.800041
onnpaikka8   -4.369e-01   1.033e-01   6.460e-01   5.276e-01  7.910e-01 2.34e-05 ***
onnpaikka9    4.175e-01   6.476e-01   1.518e+00   4.266e-01  5.402e+00 0.519156
saa2         -1.036e-01   4.251e-02   9.016e-01   8.295e-01  9.800e-01 0.014817 *
saa3         -1.845e-01   1.777e-01   8.315e-01   5.870e-01  1.178e+00 0.299114
saa4         -1.542e-01   8.086e-02   8.571e-01   7.315e-01  1.004e+00 0.056486 .
saa5          1.312e-01   9.273e-02   1.140e+00   9.508e-01  1.368e+00 0.156948
saa6         -2.099e-01   1.372e-01   8.106e-01   6.195e-01  1.061e+00 0.126075
ELY2          3.977e-01   7.228e-01   1.488e+00   3.610e-01  6.137e+00 0.582157
ELY3          1.466e+01   3.247e+02   2.328e+06  8.684e-271 6.242e+282 0.963992
ELY4          1.413e+01   1.515e+02   1.368e+06 1.442e-123 1.299e+135 0.925706
ELY8          1.718e+00   1.182e+00   5.574e+00   5.491e-01  5.657e+01 0.146205
ELY9          2.273e+00   1.027e+00   9.712e+00   1.298e+00  7.266e+01 0.026818 *
ELY10         5.374e-01   1.406e+00   1.712e+00   1.089e-01  2.690e+01 0.702195
ELY12        -9.820e+00   2.209e+02   5.434e-05 5.023e-193 5.879e+183 0.964541
ELY14         2.841e+00   1.428e+00   1.713e+01   1.042e+00  2.816e+02 0.046727 *
paallyste2   -4.122e-01   1.692e-01   6.622e-01   4.753e-01  9.230e-01 0.014833 *
paallyste3   -2.243e-01   1.202e-01   7.991e-01   6.314e-01  1.011e+00 0.061978 .
```

```
paallyste4   -1.319e+01   3.247e+02   1.872e-06  6.995e-283 5.008e+270 0.967605
paallyste5    3.966e-01   1.431e+00   1.487e+00   8.992e-02  2.458e+01 0.781715
paallyste6    2.792e-01   6.183e-01   1.322e+00   3.935e-01  4.442e+00 0.651550
valoisuus2   -6.284e-02   7.523e-02   9.391e-01   8.104e-01  1.088e+00 0.403572
valoisuus3   -6.816e-02   6.842e-02   9.341e-01   8.169e-01  1.068e+00 0.319164
valoisuus4   -2.242e-01   7.068e-02   7.992e-01   6.958e-01  9.180e-01 0.001517 **
tunti2       -3.270e-02   1.069e-01   9.678e-01   7.849e-01  1.193e+00 0.759722
tunti3       -1.183e-01   9.604e-02   8.884e-01   7.360e-01  1.072e+00 0.217906
tunti4       -2.166e-01   1.021e-01   8.053e-01   6.592e-01  9.840e-01 0.033907 *
tunti5       -5.378e-02   9.871e-02   9.476e-01   7.809e-01  1.150e+00 0.585892
tunti6       -1.919e-01   9.082e-02   8.254e-01   6.908e-01  9.860e-01 0.034601 *
tunti7       -9.754e-02   8.670e-02   9.071e-01   7.653e-01  1.075e+00 0.260564
tunti8        3.414e-03   8.863e-02   1.003e+00   8.434e-01  1.194e+00 0.969273
taajmerk1     1.087e-01   7.313e-02   1.115e+00   9.659e-01  1.287e+00 0.137314
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22085  on 17910  degrees of freedom
Residual deviance: 18993  on 17821  degrees of freedom
AIC: 19173

Number of Fisher Scoring iterations: 11
```

B.2 Karsittu malli 1

```
Call:
glm(formula = aineisto1$y ~ onluokka + lampotila + nopraj + kvl +
    osallkm + pinta + talvhoitlk + ajr + onnpaikka + saa + ELY +
    paallyste + valoisuus + tunti + taajmerk, family = binomial(link = "logit"),
    data = aineisto1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3860  -0.8437  -0.5592   1.0512   2.6031

Coefficients:
              Estimate  Std. Error          OR       lower     upper Pr(>|z|)
(Intercept)  -1.310e+00   3.455e-01   2.699e-01   1.371e-01 5.310e-01 0.000150 ***
onluokka2    -1.570e-01   8.852e-02   8.547e-01   7.186e-01 1.017e+00 0.076160 .
onluokka3    -1.217e+00   1.016e-01   2.961e-01   2.426e-01 3.610e-01  < 2e-16 ***
onluokka4    -2.659e-02   8.369e-02   9.738e-01   8.264e-01 1.147e+00 0.750655
onluokka5     4.639e-01   8.701e-02   1.590e+00   1.341e+00 1.886e+00 9.77e-08 ***
onluokka6    -5.107e-01   8.645e-02   6.001e-01   5.066e-01 7.110e-01 3.48e-09 ***
onluokka7     1.514e+00   1.440e-01   4.546e+00   3.429e+00 6.028e+00  < 2e-16 ***
onluokka8     2.169e+00   2.022e-01   8.745e+00   5.884e+00 1.300e+01  < 2e-16 ***
onluokka9     2.489e+00   2.494e-01   1.206e+01   7.394e+00 1.965e+01  < 2e-16 ***
onluokka10   -2.320e+00   9.398e-02   9.830e-02   8.176e-02 1.180e-01  < 2e-16 ***
onluokka11   -1.558e+00   1.449e-01   2.105e-01   1.584e-01 2.800e-01  < 2e-16 ***
onluokka12   -2.247e+00   3.395e-01   1.057e-01   5.433e-02 2.060e-01 3.61e-11 ***
```

onluokka13	-9.006e-01	1.071e-01	4.063e-01	3.294e-01	5.010e-01	< 2e-16	***
lampotila	2.147e-02	2.760e-03	1.022e+00	1.016e+00	1.027e+00	7.32e-15	***
nopraj40	-6.708e-02	3.291e-01	9.351e-01	4.906e-01	1.783e+00	0.838516	
nopraj50	6.195e-02	3.204e-01	1.064e+00	5.677e-01	1.994e+00	0.846702	
nopraj60	4.144e-01	3.199e-01	1.514e+00	8.084e-01	2.833e+00	0.195193	
nopraj70	6.549e-01	3.423e-01	1.925e+00	9.841e-01	3.766e+00	0.055719	.
nopraj80	6.542e-01	3.199e-01	1.924e+00	1.028e+00	3.601e+00	0.040840	*
nopraj100	8.594e-01	3.225e-01	2.362e+00	1.255e+00	4.444e+00	0.007705	**
nopraj120	8.857e-01	3.413e-01	2.425e+00	1.242e+00	4.733e+00	0.009451	**
kvl	-1.171e-05	2.257e-06	1.000e+00	1.000e+00	1.000e+00	2.12e-07	***
osallkm	4.235e-01	4.292e-02	1.527e+00	1.404e+00	1.661e+00	< 2e-16	***
pinta2	-7.622e-02	6.142e-02	9.266e-01	8.215e-01	1.045e+00	0.214592	
pinta3	5.251e-02	2.059e-01	1.054e+00	7.039e-01	1.578e+00	0.798735	
pinta4	-5.699e-01	1.030e-01	5.656e-01	4.622e-01	6.920e-01	3.11e-08	***
pinta5	-8.286e-02	1.100e-01	9.205e-01	7.420e-01	1.142e+00	0.451136	
pinta6	-1.989e-01	7.140e-02	8.196e-01	7.126e-01	9.430e-01	0.005331	**
pinta7	8.739e-02	1.322e-01	1.091e+00	8.423e-01	1.414e+00	0.508484	
talvhoitlk1	5.770e-02	7.493e-02	1.059e+00	9.147e-01	1.227e+00	0.441274	
talvhoitlk2	1.335e-01	9.668e-02	1.143e+00	9.455e-01	1.381e+00	0.167374	
talvhoitlk3	2.396e-01	8.997e-02	1.271e+00	1.065e+00	1.516e+00	0.007735	**
talvhoitlk4	4.993e-02	1.342e-01	1.051e+00	8.081e-01	1.367e+00	0.709848	
talvhoitlk5	2.750e-01	9.745e-02	1.317e+00	1.088e+00	1.594e+00	0.004773	**
talvhoitlk6	2.314e-01	1.226e-01	1.260e+00	9.912e-01	1.603e+00	0.059046	.
ajr1	-3.155e-01	8.112e-02	7.294e-01	6.222e-01	8.550e-01	0.000101	***
ajr2	-1.963e-01	8.181e-02	8.218e-01	7.000e-01	9.650e-01	0.016447	*
onnpaikka2	-1.278e-01	2.312e-01	8.800e-01	5.594e-01	1.384e+00	0.580399	
onnpaikka3	-2.985e-01	7.692e-01	7.419e-01	1.643e-01	3.351e+00	0.697966	
onnpaikka4	-3.193e-01	4.253e-01	7.266e-01	3.157e-01	1.672e+00	0.452693	
onnpaikka6	-4.473e-02	1.622e-01	9.563e-01	6.958e-01	1.314e+00	0.782758	
onnpaikka8	-4.367e-01	1.022e-01	6.462e-01	5.289e-01	7.890e-01	1.92e-05	***
onnpaikka9	4.075e-01	6.522e-01	1.503e+00	4.186e-01	5.396e+00	0.532122	
saa2	-9.252e-02	4.226e-02	9.116e-01	8.392e-01	9.900e-01	0.028564	*
saa3	-1.702e-01	1.776e-01	8.435e-01	5.956e-01	1.195e+00	0.337698	
saa4	-1.397e-01	8.045e-02	8.696e-01	7.428e-01	1.018e+00	0.082543	.
saa5	1.421e-01	9.240e-02	1.153e+00	9.617e-01	1.382e+00	0.124165	
saa6	-2.034e-01	1.366e-01	8.159e-01	6.243e-01	1.066e+00	0.136529	
ELY2	-3.981e-01	5.946e-02	6.716e-01	5.977e-01	7.550e-01	2.16e-11	***
ELY3	-5.596e-02	8.839e-02	9.456e-01	7.952e-01	1.124e+00	0.526648	
ELY4	-4.092e-01	7.066e-02	6.642e-01	5.783e-01	7.630e-01	7.03e-09	***
ELY8	-3.143e-01	6.881e-02	7.303e-01	6.382e-01	8.360e-01	4.94e-06	***
ELY9	-2.091e-01	7.956e-02	8.114e-01	6.942e-01	9.480e-01	0.008599	**
ELY10	-5.559e-01	6.762e-02	5.736e-01	5.024e-01	6.550e-01	< 2e-16	***
ELY12	-4.799e-01	6.935e-02	6.189e-01	5.402e-01	7.090e-01	4.53e-12	***
ELY14	-2.668e-01	1.055e-01	7.658e-01	6.228e-01	9.420e-01	0.011397	*
paallyste2	-4.481e-01	1.681e-01	6.389e-01	4.595e-01	8.880e-01	0.007684	**
paallyste3	-2.137e-01	1.197e-01	8.076e-01	6.387e-01	1.021e+00	0.074079	.
paallyste4	-1.091e+01	1.195e+02	1.819e-05	3.685e-107	8.975e+96	0.927205	
paallyste5	2.623e-01	1.453e+00	1.300e+00	7.529e-02	2.244e+01	0.856795	
paallyste6	3.549e-01	6.186e-01	1.426e+00	4.242e-01	4.794e+00	0.566214	
valoisuus2	-6.547e-02	7.499e-02	9.366e-01	8.086e-01	1.085e+00	0.382678	
valoisuus3	-6.668e-02	6.812e-02	9.355e-01	8.186e-01	1.069e+00	0.327611	
valoisuus4	-2.236e-01	7.032e-02	7.997e-01	6.967e-01	9.180e-01	0.001476	**

```
tunti2       -2.299e-02   1.065e-01   9.773e-01   7.932e-01 1.204e+00 0.829042
tunti3       -1.237e-01   9.565e-02   8.836e-01   7.326e-01 1.066e+00 0.195927
tunti4       -2.233e-01   1.016e-01   7.998e-01   6.554e-01 9.760e-01 0.028001 *
tunti5       -6.716e-02   9.831e-02   9.350e-01   7.712e-01 1.134e+00 0.494530
tunti6       -1.922e-01   9.046e-02   8.251e-01   6.911e-01 9.850e-01 0.033566 *
tunti7       -9.354e-02   8.633e-02   9.107e-01   7.689e-01 1.079e+00 0.278601
tunti8       -1.460e-03   8.829e-02   9.985e-01   8.399e-01 1.187e+00 0.986803
taajmerk1     9.803e-02   7.282e-02   1.103e+00   9.563e-01 1.272e+00 0.178259
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22085  on 17910  degrees of freedom
Residual deviance: 19116  on 17839  degrees of freedom
AIC: 19260

Number of Fisher Scoring iterations: 9
```

B.3 Alkuperäinen malli 2

```
Call:
glm(formula = aineisto2$y ~ onluokka + pinta + tunti + kvl +
    toimluokka + onnpaikka + risteys + taajmerk + lampotila +
    valoisuus, family = binomial(link = "logit"), data = aineisto2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3744  -0.3820  -0.2859  -0.1932   3.6118

Coefficients:
              Estimate Std. Error          OR      lower  upper Pr(>|z|)
(Intercept) -1.163e+00  2.935e-01  3.124e-01  1.758e-01  0.555 7.39e-05 ***
onluokka2   -7.412e-01  3.774e-01  4.766e-01  2.274e-01  0.999 0.049552 *
onluokka3    2.271e-01  3.355e-01  1.255e+00  6.502e-01  2.422 0.498472
onluokka4    3.868e-01  3.166e-01  1.472e+00  7.915e-01  2.739 0.221890
onluokka5    1.936e+00  1.488e-01  6.932e+00  5.178e+00  9.279  < 2e-16 ***
onluokka6   -4.145e-01  2.663e-01  6.606e-01  3.920e-01  1.113 0.119558
onluokka7   -1.197e+00  5.983e-01  3.022e-01  9.354e-02  0.976 0.045499 *
onluokka8    8.833e-01  3.033e-01  2.419e+00  1.335e+00  4.383 0.003593 **
onluokka9    1.798e+00  2.731e-01  6.040e+00  3.536e+00 10.314 4.53e-11 ***
onluokka10  -1.030e+00  4.342e-01  3.568e-01  1.523e-01  0.836 0.017644 *
onluokka11  -4.385e-01  6.024e-01  6.450e-01  1.981e-01  2.100 0.466621
onluokka12  -1.383e+01  7.462e+02  9.828e-07  0.000e+00    Inf 0.985211
onluokka13   4.818e-01  3.159e-01  1.619e+00  8.716e-01  3.007 0.127240
pinta2      -2.106e-01  1.589e-01  8.101e-01  5.933e-01  1.106 0.185009
pinta3      -1.476e+00  1.039e+00  2.285e-01  2.981e-02  1.751 0.155349
pinta4      -1.240e+00  3.448e-01  2.895e-01  1.473e-01  0.569 0.000325 ***
pinta5      -7.856e-01  3.119e-01  4.558e-01  2.473e-01  0.840 0.011779 *
pinta6      -8.225e-01  2.234e-01  4.393e-01  2.835e-01  0.681 0.000232 ***
pinta7      -3.024e-01  3.588e-01  7.390e-01  3.658e-01  1.493 0.399294
```

```
tunti2      -2.364e-01  3.054e-01  7.894e-01  4.339e-01  1.436 0.438848
tunti3      -2.332e-01  2.784e-01  7.920e-01  4.589e-01  1.367 0.402293
tunti4      -5.171e-01  3.025e-01  5.962e-01  3.296e-01  1.079 0.087317 .
tunti5      -1.048e+00  3.029e-01  3.507e-01  1.937e-01  0.635 0.000542 ***
tunti6      -9.318e-01  2.787e-01  3.938e-01  2.281e-01  0.680 0.000829 ***
tunti7      -8.356e-01  2.656e-01  4.336e-01  2.576e-01  0.730 0.001656 **
tunti8      -5.086e-01  2.703e-01  6.013e-01  3.540e-01  1.021 0.059903 .
kvl         -2.820e-05  7.310e-06  1.000e+00  1.000e+00  1.000 0.000115 ***
toimluokka2 -2.433e-01  1.859e-01  7.841e-01  5.447e-01  1.129 0.190640
toimluokka3 -5.033e-01  1.566e-01  6.045e-01  4.447e-01  0.822 0.001311 **
toimluokka4 -7.750e-01  1.582e-01  4.607e-01  3.379e-01  0.628 9.66e-07 ***
onnpaikka2  -1.162e+00  5.201e-01  3.129e-01  1.129e-01  0.867 0.025493 *
onnpaikka3  -1.395e+01  8.116e+02  8.753e-07  0.000e+00    Inf 0.986288
onnpaikka4  -1.402e+01  7.176e+02  8.182e-07  0.000e+00    Inf 0.984418
onnpaikka6   1.694e-01  4.403e-01  1.185e+00  4.998e-01  2.808 0.700512
onnpaikka8  -1.315e+00  5.257e-01  2.684e-01  9.577e-02  0.752 0.012346 *
onnpaikka9   3.302e-01  1.195e+00  1.391e+00  1.338e-01 14.468 0.782268
risteys1    -1.372e+01  4.050e+02  1.099e-06  0.000e+00    Inf 0.972971
risteys2    -2.013e-01  2.545e-01  8.176e-01  4.965e-01  1.347 0.428966
risteys3    -1.491e-01  3.809e-01  8.615e-01  4.083e-01  1.818 0.695453
risteys4    -2.193e+00  1.023e+00  1.115e-01  1.503e-02  0.828 0.031971 *
risteys5    -3.317e-01  3.901e-01  7.177e-01  3.341e-01  1.542 0.395064
risteys6    -1.314e+01  1.634e+03  1.958e-06  0.000e+00    Inf 0.993582
risteys8    -1.375e+01  2.400e+03  1.067e-06  0.000e+00    Inf 0.995428
taajmerk1   -4.234e-01  2.321e-01  6.548e-01  4.154e-01  1.032 0.068160 .
lampotila   -1.513e-02  8.324e-03  9.850e-01  9.690e-01  1.001 0.069210 .
valoisuus2  -5.290e-01  2.611e-01  5.892e-01  3.531e-01  0.983 0.042772 *
valoisuus3  -3.062e-01  2.115e-01  7.362e-01  4.864e-01  1.114 0.147706
valoisuus4  -3.063e-02  2.302e-01  9.698e-01  6.176e-01  1.523 0.894159
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2840.7  on 5494  degrees of freedom
Residual deviance: 2405.4  on 5447  degrees of freedom
AIC: 2501.4

Number of Fisher Scoring iterations: 15
```

B.4 Karsittu malli 2

```
Call:
glm(formula = aineisto2$y ~ onluokka + pinta + tunti + kvl +
    toimluokka + taajmerk + lampotila + valoisuus,
    family = binomial(link = "logit"), data = aineisto2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3580  -0.3821  -0.2883  -0.2080   3.0848
```

```
Coefficients:
               Estimate  Std. Error          OR       lower      upper Pr(>|z|)
(Intercept)  -1.180e+00   2.916e-01   3.072e-01   1.734e-01  5.440e-01 5.19e-05 ***
onluokka2    -9.164e-01   3.569e-01   3.999e-01   1.987e-01  8.050e-01 0.010230 *
onluokka3     2.019e-01   3.344e-01   1.224e+00   6.354e-01  2.357e+00 0.545951
onluokka4     1.786e-01   2.354e-01   1.196e+00   7.536e-01  1.897e+00 0.448124
onluokka5     1.976e+00   1.484e-01   7.213e+00   5.393e+00  9.649e+00  < 2e-16 ***
onluokka6    -5.061e-01   2.645e-01   6.028e-01   3.590e-01  1.012e+00 0.055663 .
onluokka7    -1.317e+00   5.931e-01   2.678e-01   8.376e-02  8.560e-01 0.026333 *
onluokka8     4.083e-01   2.800e-01   1.504e+00   8.690e-01  2.604e+00 0.144746
onluokka9     1.607e+00   2.606e-01   4.989e+00   2.993e+00  8.315e+00 6.99e-10 ***
onluokka10   -9.812e-01   4.338e-01   3.749e-01   1.602e-01  8.770e-01 0.023700 *
onluokka11   -4.067e-01   6.019e-01   6.659e-01   2.047e-01  2.166e+00 0.499260
onluokka12   -1.178e+01   2.742e+02   7.677e-06  2.887e-239 2.041e+228 0.965743
onluokka13    4.483e-01   3.131e-01   1.566e+00   8.477e-01  2.892e+00 0.152126
pinta2       -2.511e-01   1.577e-01   7.780e-01   5.711e-01  1.060e+00 0.111248
pinta3       -1.475e+00   1.036e+00   2.288e-01   3.001e-02  1.744e+00 0.154649
pinta4       -1.220e+00   3.443e-01   2.951e-01   1.503e-01  5.790e-01 0.000393 ***
pinta5       -7.796e-01   3.110e-01   4.586e-01   2.493e-01  8.440e-01 0.012196 *
pinta6       -8.194e-01   2.227e-01   4.407e-01   2.848e-01  6.820e-01 0.000234 ***
pinta7       -2.842e-01   3.563e-01   7.526e-01   3.744e-01  1.513e+00 0.425014
tunti2       -2.718e-01   3.041e-01   7.620e-01   4.198e-01  1.383e+00 0.371575
tunti3       -2.968e-01   2.774e-01   7.432e-01   4.315e-01  1.280e+00 0.284689
tunti4       -5.664e-01   3.011e-01   5.676e-01   3.146e-01  1.024e+00 0.059948 .
tunti5       -1.074e+00   3.017e-01   3.416e-01   1.891e-01  6.170e-01 0.000370 ***
tunti6       -9.855e-01   2.779e-01   3.732e-01   2.165e-01  6.440e-01 0.000391 ***
tunti7       -8.774e-01   2.647e-01   4.158e-01   2.475e-01  6.990e-01 0.000918 ***
tunti8       -5.685e-01   2.696e-01   5.664e-01   3.339e-01  9.610e-01 0.034946 *
kvl          -2.726e-05   7.201e-06   1.000e+00   1.000e+00  1.000e+00 0.000154 ***
toimluokka2  -2.272e-01   1.847e-01   7.968e-01   5.548e-01  1.144e+00 0.218519
toimluokka3  -4.822e-01   1.542e-01   6.174e-01   4.563e-01  8.350e-01 0.001770 **
toimluokka4  -7.183e-01   1.555e-01   4.876e-01   3.595e-01  6.610e-01 3.85e-06 ***
taajmerk1    -7.270e-01   2.254e-01   4.833e-01   3.108e-01  7.520e-01 0.001255 **
lampotila    -1.496e-02   8.252e-03   9.852e-01   9.693e-01  1.001e+00 0.069892 .
valoisuus2   -5.219e-01   2.601e-01   5.934e-01   3.564e-01  9.880e-01 0.044818 *
valoisuus3   -2.812e-01   2.107e-01   7.549e-01   4.995e-01  1.141e+00 0.181899
valoisuus4   -1.108e-01   2.270e-01   8.951e-01   5.736e-01  1.397e+00 0.625553
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2840.7  on 5494  degrees of freedom
Residual deviance: 2440.5  on 5460  degrees of freedom
AIC: 2510.5

Number of Fisher Scoring iterations: 13
```