Tilastotieteen pro gradu -tutkielma

Logistista regressioanalyysiä ajokokeen tuloksiin vaikuttavista tekijöistä

Ville Vauhkonen

Jyväskylän yliopisto
Matematiikan ja tilastotieteen laitos
Tilastotiede
11.11.2009

Tiivistelmä

Ville Vauhkonen: *Logistista regressioanalyysiä ajokokeen tuloksiin vaikuttavista tekijöistä*
Tilastotieteen pro gradu -tutkielma, Jyväskylän yliopisto, 11.11.2009. Sivuja 27.

Tässä tutkielmassa tarkastellaan osaa Ajoneuvohallintokeskuksen (AKE) vuonna 2006 keräämästä ajokoe-aineistosta. Aineistossa on tietoja ensimmäistä kertaa henkilöauton ajokoetta Keski-Suomessa suorittavista, heidän opettajistaan, opetustavoista ja ajokokeen vastaanottajista. Aineistoon kuuluvat ne 2779 oppilasta, jotka ovat ajaneet vähintään 30 ajokertaa ja joiden ajokokeiden vastaanottajat ovat ottaneet vastaan vähintään 206 ajokoetta (6 ajokokeen vastaanottajaa).

Tutkimusongelmana on selvittää, vaikuttaako oppilaan sukupuoli, ikä, opetustapa (autokoulu/opetuslupa) tai ajokokeen vastaanottaja ajokokeen tulokseen (hyväksytty/hylätty). Ongelmaa lähestytään sovittamalla aineistoon logistinen regressiomalli. Lähteenä on käytetty Hosmer & Lemeshow'n (1989) teosta *Applied logistic regression*.

Suurin osa ajamista harjoittelevista (85.2%) valitsi opetusmuodokseen autokoulun. Selvästi yli puolet (67.3%) oppilaista myös suoritti ajokortin heti 18-vuotiaana. Oppilaan sukupuoli ja opetustapa paljastuivat kaikkein voimakkaimmin ajokokeen lopputulokseen vaikuttaviksi muuttujiksi. Miehet selviytyivät naisia paremmin ja autokoululaiset opetuslupalaisia paremmin ajokokeesta. Hyväksymisprosentit ensimmäisellä suorituskerralla olivat miehillä 70.9%, naisilla 59.2%, autokoululaisilla 66.7% ja opetuslupalaisilla 51.9%.

Oppilaan ikä vaikutti ajokokeen tulokseen niin, että yli 18-vuotiaat selviytyivät kokeesta hieman huonommin kuin 18-vuotiaat. Enemmän ajo-opetusta tarvinneet selviytyivät vähemmän ajaneita huonommin. Ainakin autokoululaisten osalta tämä saattaa selittyä oppilasaineksen valikoitumisella. Paljon ajo-opetusta tarvinneiden joukosta ovat jo karsiutuneet pois parhaat oppilaat. Opetuslupalaisten menestymiseen ajomäärillä ei ollut vaikutusta. Ajokokeen vastaanottajista vain vastaanottajalla numero 4 (vao(4)) oli vaikutusta ajokokeen lopputulokseen. Jostain syystä ne opetuslupalaiset, joiden ajokokeet hän oli ottanut vastaan, menestyivät poikkeuksellisen hyvin. Muut ajokokeen vastaanottajat eivät eronneet toisistaan siinä, miten he arvostelivat ajokokeet.

Sisältö

1 Johdanto

Ajatus tämän tutkielman kirjoittamisesta lähti liikkeelle Helsingin Sanomien artikkelista *Inssiajojen reputusmäärät vaihtelevat maakuntien ja sukupuolten välillä* (14.1.2007). Tutkimusaineisto on saatu Ajoneuvohallintokeskukselta (AKE). Se on osa vuoden 2006 ajokoe-aineistoa. Aineisto koostuu tiedoista ajokokelaista, heidän opettajistaan, opetustavoista ja ajokokeen vastaanottajista. Tässä tutkielmassa tutkimusaineisto rajataan pääpiirteissään niin, että se sisältää tietoja vain Keski-Suomessa ensimmäistä kertaa henkilöauton ajokokeeseen osallistuneista.

Tutkimusongelmana on selvittää, miten ajo-oppilaiden taustatiedot selittävät läpäisytodennäköisyyden vaihtelua. Halutaan tietää, vaikuttaako ajo-oppilaan sukupuoli, ikä tai opetustapa ajokokeen tulokseen ja mikä merkitys ajokokeen vastaanottajalla on. Ihannetilanteessahan pelkästään oppilaan taidot ratkaisevat sen, miten hän menestyy ajokokeessa. Jos oppilaan iällä tai sukupuolella on jotain vaikutusta, niin ehkä eri ikäisten tai eri sukupuolta olevien oppilaiden tulisi saada erilaista opetusta. Sekin voi olla kiinnostava tieto, menestyvätkö autokoulussa opetusta saaneet tai opetusluvalla ajamista harjoitelleet ajokokeessa muita paremmin. Heikompaa opetustapaa tulisi silloin kehittää. Erityisen tärkeää on se, onko ajokokeen vastaanottajalla vaikutusta kokeen lopputulokseen. Ajokoe ei ole uskottava, jos oppilaan osaamisen arviointi vaihtelee voimakkaasti sen mukaan, kuka ajokokeen on ottanut vastaan.

Tutkimusongelmaa lähestytään logistisen regression keinoin. Regressiomenetelmiä käytetään yleisesti kuvailtaessa vastemuuttujan ja yhden tai useamman selittävän muuttujan suhdetta toisiinsa. Tavoitteena on löytää parhaiten sopiva, mahdollisimman niukka ja mahdollisimman järkevä malli kuvaamaan vastemuuttujan ja selittävän tai selittävien muuttujien (kovariaattien) välistä yhteyttä. (Hosmer & Lemeshow (1989), s. 1 ja s. 25)

Logistisessa regressiossa vastemuuttuja on binäärinen eli dikotominen (voi saada vain kahta eri arvoa) (Hosmer & Lemeshow (1989), s. 1 ja s. 25). Tässä tapauksessa vastemuuttujana on ajokokeen tulos, joka saa arvon 0, kun koe on hylätty ja arvon 1, kun kokelas läpäisee kokeen. Hosmer ja Lemeshow toteavat, että kun vastemuuttujan binäärisyys otetaan huomioon, noudattavat logistisen regression menetelmät samoja yleisiä periaatteita kuin lineaarisessa regressiossakin. Keskeistä on muuttujien kerrointen estimointi ja kerrointen merkitsevyyden testaaminen. (Hosmer & Lemeshow (1989), s. 1 ja s. 25)

Lähteenä on käytetty Hosmer & Lemeshow'n teosta *Applied logistic regression* (1989) ja erityisesti sen lukuja 1, 2 ja 3. Jokaisen tekstikappaleen loppuun on merkitty ne Hosmer & Lemeshow'n sivut, joihin kyseinen teksti perustuu. Kaikki analyysit on suoritettu SPSS-ohjelmistolla.

2 Logistinen regressiomalli

2.1 Vastemuuttujan odotusarvo

Käsitellään riippumattomien muuttujien (p kpl) joukkoa ja merkitään sitä vektorilla $\boldsymbol{x}' = (x_1, x_2, \ldots, x_p)$. Oletetaan, että jokainen näistä muuttujista on vähintään välimatka-asteikollinen.

Vastemuuttujan odotusarvoa, kun riippumaton muuttuja on annettu, kutsutaan ehdolliseksi odotusarvoksi ja merkitään $E(Y|\boldsymbol{x})$, missä Y on vastemuuttuja ja $\boldsymbol{x}$ on riippumattoman muuttujan arvo. Kun vastemuuttuja Y on dikotominen, on ehdollisen odotusarvon oltava väliltä 0–1 eli $0 \leq E(Y|\boldsymbol{x}) \leq 1$.

Ehdollista todennäköisyyttä, että vastemuuttuja Y saa arvon yksi ehdolla $\boldsymbol{x}$ (eli, että vastemuuttujan edustama ominaisuus esiintyy) merkitään merkintöjen yksinkertaistamiseksi $P(Y = 1|\boldsymbol{x}) = \pi(\boldsymbol{x})$. Vastemuuttujan Y arvo, kun $\boldsymbol{x}$ on annettu, on $Y = \pi(\boldsymbol{x}) + \epsilon$. Virhettä on merkitty symbolilla ϵ.

Logaritminen muunnos monimuuttujaisen logistisen regression mallista on

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p \tag{1}$$

jolloin monimuuttujainen logistisen regression malli on

$$\pi(\boldsymbol{x}) = \frac{e^{g(\boldsymbol{x})}}{1 + e^{g(\boldsymbol{x})}} \tag{2}$$

(Hosmer & Lemeshow (1989), s. 5 ja s. 25–26)

2.2 Design-muuttuja D

Jos jokin riippumaton muuttuja on epäjatkuva ja luokitteluasteikollinen (esim. ajokokeen vastaanottaja), sitä ei voi sisällyttää malliin samalla tavalla kuin välimatka-asteikollista muuttujaa. Näin on, koska luvut, jotka edustavat epäjatkuvan muuttujan eri tasoja, ovat vain tasojen tunnisteita eivätkä mitenkään muuten kuvaa kyseisiä tasoja. Tällaisessa tilanteessa on käytettävä design- eli dummy-muuttujia D.

On olemassa useita eri menetelmiä design-muuttujien muodostamiseen. Design-muuttujien koodaus, jossa käytetään vertailuluokkaa (referenssiluokkaa), on kaikkein yleisimmin käytetty menetelmä. Ensisijainen syy menetelmän yleiseen käyttöön on halu estimoida altistusluokan riskiä suhteessa kontrolliluokan riskiin.

Asetetaan ensin yksi muuttujan luokista vertailuluokaksi. Seuraavaksi asetetaan kaikki design-muuttujat nolliksi vertailuluokalle ja sen jälkeen ai-

na yksittäinen design-muuttuja ykköseksi kullekin muulle luokalle. Tämä on havainnollistettu Taulukossa (1).

Taulukko 1: Neliluokkaisen muuttujan design-muuttujat, kun luokka 1 on valittu vertailuluokaksi.

	Design-muuttujat		
muuttuja(koodi)	D_1	D_2	D_3
luokka(1)	0	0	0
luokka(2)	1	0	0
luokka(3)	0	1	0
luokka(4)	0	0	1

Yleistettynä, jos luokitteluasteikollinen muuttuja voi saada k arvoa niin tarvitaan $k-1$ design-muuttujaa. Oletetaan, että j:nennellä riippumattomalla muuttujalla x_j on k_j tasoa. Desing-muuttujia on silloin $k_j - 1$ kappaletta. Design-muuttujaa merkitään D_{ju} ja näiden design-muuttujien kertoimia merkitään β_{ju}, $u = 1, 2, \ldots, k_j - 1$.

Logaritmista logistisen regression mallia, jossa on p-muuttujaa ja jossa j:s muuttuja on epäjatkuva, merkitään

$$g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{u=1}^{k_j-1} \beta_{ju} D_{ju} + \beta_p x_p$$

Kun kyseessä on dikotominen (eli kaksi luokkainen) riippumaton muuttuja (esim. sukupuoli), voidaan joko käyttää design-muuttujaa tai koodata muuttuja saamaan arvon 0 tai 1 ja liittää se malliin sellaisenaan ilman design-muuttujan käyttöä. Hosmer & Lemeshow pitävät jälkimmäistä tapaa suositeltavampana. Tavan valinta saattaa vaikuttaa jonkin verran estimoituun kertoimeen. (Hosmer & Lemeshow (1989), s. 26–27 ja s. 50)

2.3 Muuttujien kerrointen estimointi suurimman uskottavuuden menetelmällä

Muuttujien kertoimet estimoidaan käyttäen apuna suurimman uskottavuuden menetelmää. Oletetaan riippumattomien havaintoparien (x_i, y_i) otos (n kpl), $i = (1, \ldots, n)$, missä y_i on dikotomisen vastemuuttujan arvo ja x_i on i:nnenteen havaintoon liittyvän riippumattoman muuttujan arvo. Oletetaan lisäksi, että vastemuuttuja Y on koodattu saamaan arvon 0 tai 1 sen mukaan, onko muuttujan kuvaama piirre poissa vai läsnä. Logistisen regressio-

mallin sovittaminen (yhtälö (2), s. 2) vaatii, että vektorin $\boldsymbol{\beta'} = (\beta_0, \ldots, \beta_p)$ tuntemattomille parametreille saadaan estimaatit.

Käytetään suurimman uskottavuuden menetelmää. On rakennettava uskottavuusfunktio. Yhtälön (2) (s. 2) mukainen $\pi(\boldsymbol{x})$ antaa (mielivaltaisille $\boldsymbol{\beta'}$ -vektorin arvoille, $\boldsymbol{\beta'} = (\beta_0, \ldots, \beta_p)$) ehdollisen todennäköisyyden "todennäköisyys, että $Y = 1$, kun $\boldsymbol{x}$ on annettu", merkitään $P(Y|\boldsymbol{x})$. Tästä seuraa, että yhtälö $1 - \pi(\boldsymbol{x})$ antaa ehdollisen todennäköisyyden $P(Y = 0|\boldsymbol{x})$. Siten pareja (x_i, y_i), joille $y_i = 1$, vastaa uskottavuusfunktio $\pi(x_i)$ ja pareja, joille $y_i = 0$, vastaa uskottavuusfunktio $1 - \pi(x_i)$, missä $\pi(x_i)$ vastaa $\pi(\boldsymbol{x})$:n arvoa pisteessä x_i. Kätevä tapa ilmaista uskottavuusfunktion yhteys pareihin (x_i, y_i) on yhtälö:

$$\zeta(x_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i} \tag{3}$$

Kun havainnot oletetaan riippumattomiksi, uskottavuusfunktio voidaan esittää yhtälön (3) merkinnöin seuraavasti:

$$l(\boldsymbol{\beta}) = \prod_{i=1}^{n} \zeta(x_i) \tag{4}$$

Uskottavuusfunktio ilmaisee havaitun datan todennäköisyyden tuntemattomien parametrien funktiona. Näiden parametrien suurimman uskottavuuden estimaattoreiksi on valittu ne arvot, jotka maksimoivat funktion. Monimuuttujaisessa tapauksessa uskottavuusfunktio on lähes sama kuin yhtälössä (4) paitsi, että $\pi(\boldsymbol{x})$ on nyt määritelty kuten yhtälössä (2) (s. 2).

Matemaattisesti on kuitenkin helpompi työskennellä logaritmisen uskottavuusfunktion kanssa. Se voidaan esittää muodossa:

$$L(\boldsymbol{\beta}) = \ln[l(\boldsymbol{\beta})] = \sum_{i=1}^{n} \{y_i \ln[\pi(x_i)] + (1 - y_i) \ln[1 - \pi(x_i)]\} \tag{5}$$

Derivoimalla logaritmista suurimman uskottavuuden funktiota kunkin $p+1$:n kertoimen suhteen erikseen saadaan $p + 1$ uskottavuusyhtälöä. Saadut uskottavuusyhtälöt voidaan esittää seuraavasti:

$$\sum_{i=1}^{n} [y_i - \pi(x_i)] = 0 \tag{6}$$

ja

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(x_i)] = 0 \tag{7}$$

kun $j = 1, 2, \ldots, p$. Yhtälöt ovat epälineaarisia parametrien β_j suhteen ja niiden ratkaiseminen vaatii erityisiä iteratiivisia menetelmiä. Uskottavuusyhtälöiden ratkaisuna saatua vektoria $\hat{\boldsymbol{\beta}}$ kutsutaan suurimman uskottavuuden

estimaatiksi (SU-estimaatiksi). $\hat{\boldsymbol{\beta}}$ on se vektorin $\boldsymbol{\beta}$ arvo, joka maksimoi logaritmisen uskottavuusfunktion $L(\boldsymbol{\beta})$. (Hosmer & Lemeshow (1989), s. 8–10 ja s. 27–28)

2.4 Kerrointen merkitsevyyden testaaminen

Kun malliin mukaan otettujen muuttujien kertoimet on saatu estimoitua, arvioidaan seuraavaksi kerrointen merkitsevyyttä. Yritetään selvittää, mitkä mallin riippumattomista muuttujista vaikuttavat merkitsevästi vastemuuttujaan.

Kertooko malli, joka sisältää kyseisen muuttujan, enemmän vastemuuttujasta kuin malli, josta kyseinen muuttuja puuttuu? Kysymykseen etsitään vastausta vertailemalla vastemuuttujan havaittuja arvoja arvoihin, jotka on ennustettu mallilla, jossa kyseinen muuttuja on mukana ja mallilla, josta kyseinen muuttuja puuttuu. Jos mallilla, jossa muuttuja on mukana, ennustetut arvot ovat parempia tai jossain mielessä tarkempia kuin muuttujaa sisältämättömällä mallilla ennustetut arvot, niin voidaan ajatella, että tämä muuttuja on merkitsevä.

Logistisessa regressiossa havaittujen ja ennustettujen arvojen vertailu perustuu logaritmiseen uskottavuusfunktioon. On helpompi ymmärtää tämä vertailu, jos sekä vastemuuttujan havaittujen että ennustettujen arvojen ajatellaan olevan peräisin saturoidusta mallista. Saturoidussa mallissa on yhtä monta parametria kuin havaintoyksikköäkin. (Hosmer & Lemeshow (1989), s. 12–14)

2.5 Waldin testi W kerrointen merkitsevyyden testaamisessa

Waldin testi on eräs uskottavuussuhdetesti. Waldin testisuure saadaan jakamalla kertoimen β_j suurimman uskottavuuden estimaatti $\hat{\beta}_j$ sen estimoidulla keskivirheellä ($\hat{SE}$). Waldin testisuure on

$$W_j = \frac{\hat{\beta}_j}{\hat{SE}(\hat{\beta}_j)}$$

Waldin testisuureen arvot osoittavat, mitkä mallin muuttujista ovat merkitseviä ja mitkä eivät. Jos kriittisenä arvona käytetään lukua 2, joka vastaa merkitsevyystasoa $\alpha = 0.05$, niin muuttuja on merkitsevä, jos sen Waldin testisuureen arvo $W \leq -2$ tai $W \geq 2$.

Waldin testisuureen kaksisuuntainen p–arvo on $p = P(|Z| > W)$, missä W on Waldin testisuureen arvo ja Z on standardinormaalijakaumasta saatu satunnaismuuttuja. Hosmer & Lemeshow'n mukaan Waldin testi saattaa

kuitenkin ehdottaa kertoimen hylkäämistä, vaikka kerroin todellisuudessa olisikin merkitsevä. (Hosmer & Lemeshow (1989), s. 14–17 ja s. 31–33)

2.6 Regressiomallin kerrointen tulkitsemisesta yleisesti

Sovitetun mallin riittävyyden arviointi olisi tehtävä aina ennen kuin mallia yritetään tulkita. Logistisen regressiomallin tapauksessa menetelmät sopivuuden arvioimiseen ovat melko teknisiä. Oletetaan, että logistinen regressiomalli on sovitettu aineistoon, että mallin muuttujat ovat merkitseviä sovellettavan tieteen alan kannalta tai tilastotieteellisessä mielessä ja että malli sopii aineistoon myös joidenkin tilastollisten mallin sopivuutta mittaavien mittarien mukaan.

Minkä tahansa aineistoon sovitetun mallin tulkinta edellyttää, että kyetään yhdistämään estimoidut kertoimet malliin. Mitä mallin estimoidut kertoimet kertovat tässä tutkimuksessa tutkittavasta asiasta? Useimmissa malleissa kiinnostuksen kohteena ovat riippumattomien muuttujien estimoidut kertoimet. Joskus myös vakiotermin kerroin voi olla kiinnostava. Mallien tulkinta sisältää siis kaksi kysymystä: riippuvan muuttujan ja riippumattoman muuttujan välisen funktionaalisen suhteen määrittämisen ja riippumattoman muuttujan muutosyksikön määrittämisen. (Hosmer & Lemeshow (1989), s. 38–39)

2.7 Kerrointen tulkinta, kun riippumaton muuttuja on dikotominen eli kaksi luokkainen

Oletetaan, että riippumaton muuttuja x on koodattu niin, että se voi saada arvon 0 tai 1. Kun x on asetettu näin, niin $\pi(x)$ ja $1 - \pi(x)$ voivat myös molemmat saada kaksi eri arvoa.(ks. Taulukko (2)).

Taulukko 2: Logistisen regresiomallin arvot, kun riippumaton muuttuja on dikotominen.

		Riippumaton muuttuja x	
		$x = 1$	$x = 0$
vastemuuttuja	$y = 1$	$\pi(1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
Y	$y = 0$	$1 - \pi(1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$1 - \pi(0) = \frac{1}{1+e^{\beta_0}}$
	Yhteensä	1.0	1.0

Riippuvan muuttujan eli vastemuuttujan y vedonlyöntisuhde (odds) on $\pi(1)/[1 - \pi(1)]$, kun $x = 1$. Kun $x = 0$, on vastemuuttujan vedonlyöntisuhde

$\pi(0)/[1-\pi(0)]$. Logaritmiset vedonlyöntisuhteet ovat tässä tapauksessa

$$g(1) = \ln\{\pi(1)/[1-\pi(1)]\}$$

ja

$$g(0) = \ln\{\pi(0)/[1-\pi(0)]\}$$

Ristitulosuhdetta (odds ratio) merkitään kirjaimella ψ. Se on vedonlyöntisuhteiden $x = 1$ ja $x = 0$ välinen suhde.

$$\psi = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} \tag{8}$$

Logaritminen ristitulosuhde on

$$\ln(\psi) = \ln\left[\frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}\right] = g(1) - g(0) = \ln\left(e^{\beta_1}\right) = \beta_1$$

Käyttämällä taulukon (2) (s. 6) merkintöjä logistiselle regressiomallille saadaan ristitulosuhteelle

$$\psi = \frac{\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\frac{1}{1+e^{\beta_0}}}{\frac{e^{\beta_0}}{1+e^{\beta_0}}\frac{1}{1+e^{\beta_0+\beta_1}}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Ristitulosuhde on mitta, joka arvioi, miten paljon todennäköisempi (tai epätodennäköisempi) tutkittu vaste on niiden joukossa, joilla $x = 1$ kuin niiden joukossa, joilla $x = 0$. Useissa tapauksissa, joissa logistista regressiota käytetään, on jatkuva muuttuja tehty dikotomiseksi katkaisemalla se jostain sovellettavan tieteenalan kannalta mielekkäästä kohdasta. $\hat{\psi}$ on suurimman uskottavuuden estimaatti ristitulosuhteelle (odds ratio).

Dikotomisen muuttujan tapauksessa kiinnostava parametri on ristitulosuhde. Tämän parametrin estimaatti saadaan estimoimalla logistisen regression kerroin. Logistisella regressiolla saadun kertoimen ja ristitulosuhteen välinen yhteys muodostaa perustan kaikkien logistisen regression tulosten tulkitsemiseen. (Hosmer & Lemeshow (1989), s. 39–41)

2.8 Kerrointen tulkinta riippumattoman muuttujan ollessa moniluokkainen

Oletetaan, että riippumaton muuttuja onkin k-luokkainen. Tämä muuttuja voi saada vain joitakin tiettyjä epäjatkuvia arvoja ja se on mitta-asteikoltaan luokitteluasteikollinen. On muodostettava joukko design-muuttujia edustamaan muuttujan luokkia.

Kun verrataan luokkaa (2) luokkaan (1) saadaan:

$$\begin{aligned}
\ln[\hat{\psi}(\text{luokka}(2), \text{luokka}(1))] &= \hat{g}(\text{luokka}(2)) - \hat{g}(\text{luokka}(1)) \\
&= [\hat{\beta}_0 + \hat{\beta}_{11} \times (D_1 = 1) + \hat{\beta}_{12} \times (D_2 = 0) \\
&\quad + \hat{\beta}_{13} \times (D_3 = 0)] - [\hat{\beta}_0 + \hat{\beta}_{11} \times (D_1 = 0) \\
&\quad + \hat{\beta}_{12} \times (D_2 = 0) + \hat{\beta}_{13} \times (D_3 = 0)] \\
&= \hat{\beta}_{11}
\end{aligned}$$

missä $\hat{\psi}$ on suurimman uskottavuuden estimaatti ristitulosuhteelle. (Hosmer & Lemeshow (1989), s. 47 ja s. 50)

2.9 Yhdysvaikutus ja sekoittaminen

Ajatellaan aluksi, että mallissa on vain yksi dikotominen muuttuja, jota kutsutaan dikotomiseksi riskitekijämuuttujaksi. Lisätään seuraavaksi malliin toinen, jatkuva muuttuja. Jos riskitekijämuuttujan kerroin muuttuu nyt voimakkaasti, on Hosmer & Lemeshow'n mukaan kyse lisätyn uuden muuttujan sekoittamisvaikutuksesta. Lisätty uusi muuttuja on siis sekoittaja (confounder). Uusi muuttuja vaikuttaa sekä riskitekijämuuttujaan että vastemuuttujaan ja näin sekoittaa niiden välistä yhteyttä.

Lisätyn muuttujan ja riskitekijämuuttujan välinen yhdysvaikutus tarkoittaa sitä, että uusi muuttuja muuttaa jollain tavalla riskitekijämuuttujan vaikutusta. Jos riskitekijämuuttujan ja uuden muuttujan välinen yhteys on lineaarista, on riskitekijämuuttujan eri tasoja kuvaavilla suorilla eri kulmakertoimet. Tällöin uusi muuttuja on Hosmer & Lemeshow'n mukaan vaikutuksen määrittäjä (effect modifier).

Uusi muuttuja voi olla pelkkä sekoittaja tai pelkkä vaikutuksen määrittäjä. Muuttuja voi myös olla yhtä aikaa sekä sekoittaja että vaikutuksen määrittäjä. Tai muuttuja voi olla vaikutuksen määrittäjä vaikka ei olisikaan sekoittaja. Yhdysvaikutustermin lisääminen malliin muuttaa usein selvästi riskitekijämuuttujan kerrointa. Siksi Hosmer & Lemeshow'n mukaan uuden muuttujan sekoittamisvaikutusta ei voida arvoida riskitekijämuuttujan kertoimen muuttumisen kautta, kun yhdysvaikutustermi on mukana mallissa. Jos uusi muuttuja on sekä sekoittaja että vaikutuksen määrittäjä on sen status sekoittajana Hosmer & Lemeshow'n mukaan toissijaista, koska riskitekijämuuttujan vaikutuksen estimaatti riippuu uuden muuttujan arvosta.

Jos riskitekijämuuttujan estimoitu kerroin muuttuu millä tahansa tarkasteltavan tieteen alan kannalta tärkeällä tavalla, kun uusi muuttuja liitetään malliin, on uutta muuttujaa pidettävä Hosmer & Lemeshow'n mukaan sekoittajana ja se on liitettävä malliin, vaikka sen estimoitu kerroin ei olisikaan merkitsevä. Sen sijaan Hosmer & Lemeshow pitävät uutta muuttujaa

vaikutuksen määrittäjänä vain, jos yhdysvaikutustermi on sekä tarkasteltavan tieteen alan kannalta järkevä että tilastollisesti merkitsevä. (Hosmer & Lemeshow (1989), s. 63–67)

3 Logistisen regressiomallin sovittaminen ajokoe-aineistoon

3.1 Tutkimusongelma

Helsingin Sanomien artikkelissa (14.1.2007) kerrottiin, että ajokokeen läpäisytodennäköisyys vaihtelee ilmeisesti alueellisesti ja mahdollisesti myös ajokokeen vastaanottajasta riippuen.

Tutkimuksen tavoitteena on selvittää, miten ajo-oppilaiden taustatiedot selittävät läpäisytodennäköisyyden vaihtelua. Tutkimus perustuu Ajoneuvohallintokeskuksen (AKE) keräämiin tietoihin ajokokeen suorittajista, ajokokelaiden opettajista, opetustavoista ja ajokokeen vastaanottajista. Tässä tutkimuksessa aineisto rajataan vuonna 2006 Keski-Suomessa ensimmäistä kertaa henkilöauton ajokokeeseen osallistuneisiin.

Tutkimusmenetelmänä käytetään logistista regressiota, jossa vastemuuttujana on ajokokeen tulos (tulos=hyväksytty/hylätty). Tavoitteena on saada vastaus mm. kysymykseen, onko eri ikäisillä mies- ja naiskokelailla erilainen läpäisytodennäköisyys, kun otetaan muut asiaan vaikuttavat tekijät huomioon. Onko autokoulun käyneillä ja opetusluvalla opiskelleiden välillä eroja? Edelleen etsitään vastausta kysymykseen, vaihteleeko läpäisytodennäköisyys kokeen vastaanottajien välillä merkitsevästi myös sen jälkeen, kun mahdolliset erot kokeen vastaanottajissa on otettu huomioon (esim. vaikuttaako kokeen vastaanottajan sukupuoli ja koulutus).

3.2 Lisärajauksia aineistoon

Alkuperäisenä tavoitteena oli tutkia vuonna 2006 Keski-Suomessa ensimmäistä kertaa henkilöauton ajokokeeseen osallistuneita. Tuntui kuitenkin tarkoituksenmukaiselta tehdä aineistoon joitakin lisärajoituksia.

Ensimmäiseksi aineistoon kuuluvista kolmestatoista ajokokeenvastaanottajasta rajataan seitsemän pois sillä perusteella, että he ovat ottaneet vastaan vain muutamia ajokokeita (15–145). Jäljelle jää tiedot 3019 oppilaasta ja kuudesta vastaanottajasta, joista jokainen on ottanut vastaan vähintään 224 ajokoetta. Näistä ajokokeen kuudesta vastaanottajasta kaksi on katsastusmiehiä (vastaanottajat vao(1) ja vao(2)) ja loput neljä liikenneopettajia.

Toiseksi aineistosta rajataan pois ne oppilaat, joilla ajokertoja on alle 30. Ajo-oppilaallahan on oltava takanaan vähintään 30 ajokertaa ennen kuin hän saa osallistua ajokokeeseen ellei hänellä sitten ole jo ennestään jokin muu ajokortti, josta voi saada hyvitystä ajokertojen määrään. Tämän rajauksen avulla aineistosta pitäisi tippua pois ainakin ne, joilla on ollut jokin ajokortti ennen B-ajokokeen suorittamista ja jotka ovat aiemmasta kortistaan jol-

Taulukko 3: Ajokoe-aineiston muuttujat

muuttujan koodi	**selitys**	**luokittelu**
luok.ikä	luokiteltu oppilaan ikä	0=18v., 1=yli 18v.
sp	oppilaan sukupuoli	0=nainen, 1=mies
o.sp	opettajan sukupuoli	0=nainen, 1=mies
o.ikä	opettajan ikä	jatkuva
opetus	opetustapa	0=opetuslupa, 1=autokoulu
atulos	ajokokeen tulos	0=hylätty, 1=hyväksytty
suorkk	ajokokeen suorituskuukausi	kuukauden numero
paikka	toimipaikan koodi	numerokoodi
koulu	autokoulun koodi	numerokoodi
vao	tutkinnon vastaanottajan koodi	numerokoodi
kokemus	vastaanottajan koulutus	0=kats.mies, 1=liik.opett.
pimeä	pimeällä ajamisen opetuskerrat	0=ei ajoa, 1=2 ajoa
ajoyht	ajo-opetus yht. ilman pim. ajoa	0=30, 1=31–34, 2=35–

lain tapaa hyötyneet. Ajattelen, että aiempi ajokortti helpottaa B-ajokokeen suorittamista. Aineistossa pitäisi siis olla jäljellä enää vain sellaisia oppilaita, joilla ei ole ollut ennestään mitään ajokorttia tai jotka ovat aiemmasta kortistaan huolimatta tarvinneet vähintään 30 ajokertaa ennen B-ajokokeeseen pääsemistään.

Tehtyjen rajausten jälkeen aineistoon kuuluu 2779 ajo-oppilasta, joista jokainen on ajanut vähintään 30 ajokertaa. Heidän ajokokeensa on ottanut vastaan kuusi eri vastaanottajaa. Vastaanottajista vao(2) on ottanut vastaan vähiten ajokokeita eli 206 koetta.

3.3 Muuttujien tarkastelua ja muuttujien luokittelu

Tähän ajokoe-aineiston osaan kuuluu 13 muuttujaa. Tarkemmat tiedot muuttujista voi katsoa Taulukosta (3) (tämän sivun yläreunassa).

Ajokokeen vastaanottajia on kuusi ja heillä kokelaita yhteensä 2779. Tietoja vastaanottajien iästä ja sukupuolesta ei ole käytettävissä.

Muuttujat *opettajan sukupuoli* ja *opettajan ikä* ovat ongelmallisia siksi, että tiedot ovat olemassa vain opetusluvalla opettaneilta, mutta ei autokoulun opettajilta. Opetusluvalla saa opettaa vain yhtä oppilasta, kun taas autokoululaisella saattaa olla useampiakin ajo-opettajia. Puuttuvien tietojen vuoksi näitä muuttujia ei ole mahdollista käyttää tulevissa malleissa.

Muuttujassa *autokoulun koodi* on 39 autokoulun koodit, koodi puuttuvalle tiedolle ja lisäksi vielä yksi koodi opetuslupalaisia varten. Muuttujasta tekee huonon se, että se on liian moniluokkainen ja lisäksi useissa sen luokissa on hyvin vähän oppilaita.

Oppilaiden ikäjakauma on erittäin vino. Oppilaista 67.3 prosenttia on kahdeksantoistavuotiaita. Monien kokeilujen jälkeen päätin muodostaa uuden muuttujan *luokiteltu oppilaan ikä*, jossa oppilaat on luokiteltu kahteen luokkaan iän perusteella (0=18v. (67.3%), 1=yli 18v. (32.7%)). Tällaista luokittelua tukee se, että ajokortti ajetaan yleensä heti 18-vuotiaana.

Ajoneuvohallintokeskuksen (AKE) ylitarkastajan Marita Koivukosken mukaan muuttuja *ajoyht* tarkoittaa oppilaan saamaa ajo-opetusmäärää. Minimi määrä on noin 15 tuntia eli 30×25 min. Muuttujan yksikkö on ajokerta ja yksi ajokerta kestää 25 minuuttia. Jos oppilaalla jo on A-luokan (moottoripyörä) kortti, hän saa korvattua sillä joitakin B-ajotunteja. Jos muuttuja *ajo-opetus yhteensä* saa arvon 0, on se luultavasti puuttuva tieto tai oppilas on voinut päästä ajokokeeseen kokonaan ilman opetusta. Tällaisella oppilaalla voi olla esim. ulkomailla myönnetty vastaava kortti, jolla pääsee suoraan tutkintoon Suomessa. Sitä ei kerrottu, paljonko jo olemassa olevasta A-kortista annetaan maksimissaan hyvitystä. Ilmeisesti ajotunteja otetaan tarpeen mukaan lisää ja muita syitä kuin A-kortti, ulkomailla myönnetty B-kortti tai puuttuva tieto ei ole sille, että oppilaalla on alle 30 ajokertaa.

Ajo-opetusmäärää kuvaava muuttuja *ajo-opetus yhteensä* luokitellaan kolmiluokkaiseksi muuttujaksi. Ensimmäisen luokan muodostavat oppilaat, joilla on 30 ajokertaa. Toiseen luokkaan tulevat 31–34 kertaa ajaneet ja kolmanteen luokkaan yli 34 kertaa ajaneet. Tälläisen luokittelun taustalla on halu saada mahdollisimman saman kokoiset luokat. 49.2% oppilaista on ajanut tasan 30 ajokertaa. 31–34 kertaa (26.2%) ja yli 34 kertaa ajaneita (24.6%) on taas suurin piirtein saman verran. Kuten Taulukko (4) sivulla 13 osoittaa, tavoite saman kokoisista luokista onnistuu kuitenkin hyvin vain autokoululaisten osalta. Taulukosta (9) sivulta 19 voidaan nähdä, että ajo-muuttujan luokat 1 ja 2 eroavat molemmat merkitsevästi luokasta 3 (kun siis vertailuluokkana on luokka 3 ja merkitsevyystasona pidetään viittä prosenttia). Luokat eroavat merkitsevästi toisistaan myös silloin, jos vertailuluokkana käytetään luokkaa 1.

3.4 Tutkittavat ajo-oppilaat

Kun aineistosta rajataan pois alle 30 ajokertaa ajaneet, jää jäljelle 2779 ajo-oppilasta 3019 oppilaasta. Heistä 45.1% (48.4%) on miehiä. (Suluissa ovat vastaavat prosentit ennen rajausta.) Autokoulun käyneitä on 85.2% (86%), joista miehiä 44.0% (47.8%). Opetusluvalla harjoitelleita on 14.8% (14%) ja

heistä 51.7% (52.0%) miehiä. Kaikista ajo-oppilaista 64.5% suoritti ajokokeensa hyväksytysti. Autokoululaisista hyväksyttiin 66.7% ja opetuslupalaisista 51.9%. Miehistä hyväksyttiin 70.9% ja naisista 59.2%.

Alle 30 ajokertaa ajaneet poikkeavat voimakkaasti muista ajo-oppilaista. Heitä on 221 ja heistä miehiä peräti 88.2%. Ajo-oppilaista autokoululaisia on 97.3%, joista miehiä 88.4%. Opetuslupalaisia on vain kuusi ja vain yksi heistä on nainen. Alle 30 kertaa ajaneista hyväksyttiin 85.1%. Autokoululaisista hyväksyttiin 86.0% ja opetuslupalaisista 50.0%. Miehistä hyväksyttiin 87.2% ja naisista 69.2%.

Näitä prosentteja ja lukumääriä on esitelty tarkemmin Taulukossa(5) ja Taulukossa(6) sivulla 14.

3.5 Sekoittajat ja vaikutuksen määrittäjät

Kaksimuuttujaisilla malleilla testataan, muuttuuko muuttujan kerroin/ kertoimet ja merkitsevyys, kun malliin lisätään toinen muuttuja eli paljastuuko jokin muuttujista *luokiteltu oppilaan ikä*, *oppilaan sukupuoli*, *opetustapa*, *vastaanottaja*, *ajo-opetus yhteensä* tai *ajokokeen suorituskuukausi* sekoittavaksi tekijäksi tai vaikutuksen määrittäjäksi. Hosmer & Lemeshow kertovat käsityksensä sekoittavasta tekijästä ja vaikutuksen määrittäjästä teoksensa *Applied logistic regression* (1989) sivuilla 63–68.

Muuttuja *ajo-opetus yhteensä* vaikuttaa näistä muuttujista kaikkein voimakkaimmin muihin muuttujiin. Sillä on voimakkain vaikutus muuttujan *opetustapa* kertoimeen. *Opetustapa* -muuttujan β-kerroin muuttuu 0.615:stä 0.082:ksi (muutos -0.533). Muuttujien *opetustapa* ja *ajo-opetus yhteensä* välillä on myös yhdysvaikutusta ($p = 0.046$) eli muuttuja *ajo-opetus yhteensä* olisi tässä tapauksessa vaikutuksen määrittäjä. Tosin tämä yhdysvaikutus katoaa ja vain voimakas sekoittamisvaikutus säilyy, kun malliin lisätään myöhemmin uusia muuttujia.

Taulukko 4: Muuttujan *ajo-opetus yhteensä* luokittelu. Taulukko kertoo, monellako prosentilla kaikista aineistoon kuuluvista, autokoululaisista ja opetuslupalaisista oli ajokertoja 30, 31–34 tai enemmän kuin 34.

ajokertoja	kaikki	autokoululaiset	opetuslupalaiset
30	49.2%	57.1%	3.6%
31–34	26.2%	28.4%	13.3%
35–	24.6%	14.5%	83.0%

Taulukko 5: Aineistosta lasketut hyväksymisprosentit oppilaille ennen rajausta ja rajauksen jälkeen sekä myös poisrajatuiksi tulleille oppilaille.

		oppilaat ennen rajausta	oppilaat rajauksen jälkeen	poisrajatut oppilaat
kaikki	kaikki	66.0%	64.5%	85.1%
	miehet	73.2%	70.9%	87.2%
	naiset	59.3%	59.2%	69.2%
auto-	kaikki	68.3%	66.7%	86.0%
koulu	miehet	75.6%	73.3%	87.9%
	naiset	61.7%	61.5%	72.0%
opetus-	kaikki	51.8%	51.9%	50.0%
lupa	miehet	59.2%	59.2%	60.0%
	naiset	43.8%	44.2%	00.0%

Taulukko 6: Oppilaiden lukumäärät ennen rajausta ja rajauksen jälkeen sekä myös poisrajatuiksi tulleiden lukumäärät.

		oppilaat ennen rajausta	oppilaat rajauksen jälkeen	poisrajatut oppilaat
kaikki	kaikki	3019	2779	221
	miehiä	1462	1254	195
	miehet%	48.4	45.1	88.2
auto-	kaikki	2600	2367	215
koulu	miehiä	1244	1041	190
	miehet%	47.8	44.0	88.4
opetus-	kaikki	419	412	6
lupa	miehiä	218	213	5
	miehet%	52.0	51.7	83.3

Muuttujan *ajo-opetus yhteensä* vaikutuksesta muuttuvat myös muuttujan *vastaanottaja* vastaanottajien vao(2), vao(3) ja vao(4) kertoimet. Vastaanottajan vao(2) kerroin muuttuu 0.226:sta 0.094:ksi (muutos −0.132), vastaanottajan vao(3) kerroin −0.022:sta −0.160:ksi (muutos −0.138) ja vastaanottajan vao(4) kerroin 0.282:sta 0.165:ksi (muutos −0.117). *Oppilaan sukupuoli* -muuttujan kerroin muuttuu 0.517:stä 0.391:ksi (kertoimen muutos −0.126) ja muuttujan *luokiteltu oppilaan ikä* kerroin −0.335:stä −0.219:ksi (muutos +0.116). Näiden muuttujien ja muuttujan *ajo-opetus yhteensä* väliltä ei löydy yhdysvaikutuksia, mutta *ajo-opetus yhteensä* vaikuttaa sekoittavasti niiden kaikkien suhteeseen vastemuuttujaan. Sen sijaan muuttujan *ajo-opetus yhteensä* kertoimet eivät muutu minkään muuttujan vaikutuksesta.

Muut muuttujat kuin muuttuja *ajo-opetus yhteensä* eivät vaikuta lähes ollenkaan toistensa kertoimiin. Suurin muutos tapahtuu muuttujien *opetustapa* ja *luokiteltu oppilaan ikä* välillä. Muuttujan *luokiteltu oppilaan ikä* lisääminen muuttujan *opetustapa* sisältävään malliin kasvattaa muuttujan *opetustapa* kerrointa 0.698:sta 0.743:een (muutos +0.045). Toisin päin tehtynä muuttujan *luokitelu oppilaan ikä* kerroin muuttuu −0.396:sta −0.44:ksi (muutos −0.044).

Aiemmin havaitun yhdysvaikutuksen *ajo-opetus yhteensä* × *opetustapa* ($p = 0.046$) lisäksi kahden muuttujan malleissa löytyy vain yksi muu yhdysvaikutus. Tämä yhdysvaikutus on muuttujien *vastaanottaja* ja *opetustapa* välillä ($p = 0.003$). Se, kuka ajokokeen ottaa vastaan, määrittää opetustavan vaikutusta ajokokeen läpäisytodennäköisyyteen. Muuttuja *vastaanottaja* on siis vaikutuksen määrittäjä.

3.6 Mallin rakentaminen

Jo ongelmanasettelu määrää pitkälti sen, mitkä muuttujat mallissa on ainakin oltava mukana. Koska ajokokeen vastaanottajan vaikutus kiinnostaa erityisen paljon, on mielestäni järkevintä alkaa sovittaa malliin muuttujaa *vastaanottaja* (vao). Se kuvaa vastaanottajan vaikutusta parhaiten. Muuttujat *vastaanottajan koulutus* (kokemus) ja *toimipaikan koodi* (paikka) häivyttäisivät vastaanottajan vaikutusta. Muuttuja *vastaanottajan koulutus* rajaisi vastaanottajat vain katsastusmiehiksi ja liikenneopettajiksi. Kuudesta vastaanottajasta neljä on liikenneopettajia ja kaksi katsastusmiehiä (vastaanottajat vao(1) ja vao(2)). Toimipaikoissa on töissä useita eri vastaanottajia ja toisaalta monet vastaanottajat työskentelevät useissa toimipaikoissa.

Muita tutkimusongelman kannalta tärkeitä muuttujia ovat *luokiteltu oppilaan ikä* (luok.ikä), *oppilaan sukupuoli* (sp) ja *opetustapa* (opetus). Näiden lisäksi halutaan testata myös muuttujien *ajo-opetus yhteensä* (ajoyht), *ajokokeen suorituskuukausi* (suorkk) ja *pimeällä ajamisen opetuskerrat* (pimeä)

merkitystä.

Ristiintaulukoin muuttujia vastemuuttujan kanssa nähdäkseni, onko niillä vaikutusta toisiinsa. Vähiten merkitseviä Pearsonin χ^2-arvoja saavat muuttujat *vastaanottajan koulutus* ($p = 0.955$) ja *pimeällä ajamisen opetuskerrat* ($p = 0.444$). Sovitan aineistoon myös yksimuuttujaisia logistisen regression malleja. Malleissa muuttujien kertoimille lasketut Waldin testisuureiden p-arvot kertovat saman, mitä ristiintaulukointikin. Muuttujat *vastaanottajan koulutus* ($p = 0.955$) ja *pimeällä ajamisen opetuskerrat* ($p = 0.444$) eivät vaikuta ajokokeen tulokseen. Jäljellä ovat vielä muuttujat *luokiteltu oppilaan ikä*, *oppilaan sukupuoli*, *opetustapa*, *ajo-opetus yhteensä*, *vastaanottaja* ($p = 0.052$) ja *ajokokeen suorituskuukausi* ($p = 0.193$).

Kootaan mallia niin, että aloitetaan yhden muuttujan mallista ja lisätään malliin yksi kerrallaan uusi muuttuja, jos sen Waldin testisuureen arvo on merkitsevä (noin 10 prosentin merkitsevyystasolla). Lähtökohtana ovat muuttujat *luokiteltu oppilaan ikä*, *oppilaan sukupuoli*, *opetustapa*, *vastaanottaja*, *ajo-opetus yhteensä* ja *ajokokeen suorituskuukausi*. Saman voisi tehdä myös vertaamalla suppeampaa ja laajempaa mallia uskottavuussuhdetestillä G (Testistä kerrotaan tarkemmin Hosmer & Lemeshow'n (1989) *Applied logistic regression* sivulla 15). SPSS-ohjelma ei kuitenkaan laske tätä tunnuslukua automaattisesti. Toisaalta lisättäessä malliin muuttujia yksi muuttuja kerrallaan myös Waldin testistä näkee helposti, tuoko uusi muuttuja mitään parannusta malliin. Siksi tyydytään käyttämään pelkkää Waldin testiä. Kun kaikki kuusi muuttujaa ovat mukana mallissa, saadaan malli, joka on esitelty Taulukossa (7) sivulla 17.

Seuraavaksi sovitetaan malli, joka sisältää kaikki jäljellä olevat muuttujat (myös muuttujat *ajokokeen suorituskuukausi* ($p = 0.106$) ja *vastaanottaja* ($p = 0.117$)) ja kaikki kahden muuttujan väliset yhdysvaikutukset (15 kpl.). Pudotetaan yhdysvaikutuksia yksi kerrallaan pois mallista vähiten merkitsevästä aloittaen. Yhdysvaikutukset tippuvat pois Taulukon (8) osoittamassa järjestyksessä (s. 18).

Kahden muuttujan välisistä yhdysvaikutuksista jäävät jäljelle *opetustapa* $\times$ *vastaanottaja* ($p = 0.000$), *vastaanottaja* $\times$ *ajo-opetus yhteensä* ($p = 0.014$) ja *oppilaan sukupuoli* $\times$ *vastaanottaja* ($p = 0.027$). Nyt siis päädytään hieman erilaisiin yhdysvaikutuksiin kuin kappaleessa **3.5 Sekoittajat ja vaikutuksen määrittäjät**, jossa etsittiin yhdysvaikutuksia kaksimuuttujaisissa malleissa. Muuttuja *ajokokeen suorituskuukausi* ei edelleenkään ole merkitsevä (nyt $p = 0.085$), joten se voidaan jättää pois mallista. Jotta malli säilyisi yksinkertaisena ja tulkinnat selkeinä, jätetään mallista pois myös yhdysvaikutus *vastaanottaja* $\times$ *ajo-opetus yhteensä*, vaikka se onkin merkitsevä. Samalla yhdysvaikutus *oppilaan sukupuoli* $\times$ *vastaanottaja* menettää merkitsevyytensä ($p = 0.027 \rightarrow 0.148$), joten sekin voidaan jättää pois mallista.

Taulukko 7: Kuuden muuttujan malli

muuttuja	β	SE	Wald	df	p-arvo	exp(β)
sp	0.406	0.086	22.3	1	0.000	1.50
luok.ikä	−0.224	0.091	6.1	1	0.014	0.80
opetus	0.266	0.137	3.8	1	0.052	1.31
vao			8.8	5	0.117	
vao(1)	−0.084	0.143	0.3	1	0.556	0.92
vao(2)	0.102	0.180	0.3	1	0.573	1.11
vao(3)	−0.214	0.124	3.0	1	0.085	0.81
vao(4)	0.128	0.135	0.9	1	0.342	1.14
vao(5)	−0.088	0.140	0.4	1	0.532	0.92
ajoyht			46.5	2	0.000	
ajoyht(1)	0.759	0.124	37.4	1	0.000	2.14
ajoyht(2)	0.221	0.124	3.2	1	0.074	1.25
suorkk			17.2	11	0.101	
suorkk(1)	0.084	0.225	0.1	1	0.707	1.09
suorkk(2)	−0.080	0.206	0.1	1	0.699	0.92
suorkk(3)	−0.172	0.210	0.7	1	0.414	0.84
suorkk(4)	−0.480	0.203	5.6	1	0.018	0.62
suorkk(5)	−0.424	0.195	4.7	1	0.029	0.65
suorkk(6)	−0.155	0.194	0.6	1	0.423	0.86
suorkk(7)	0.020	0.207	0.0	1	0.924	1.02
suorkk(8)	−0.147	0.199	0.5	1	0.461	0.86
suorkk(9)	0.077	0.208	0.1	1	0.710	1.08
suorkk(10)	−0.261	0.203	1.7	1	0.198	0.77
suorkk(11)	−0.192	0.201	0.9	1	0.338	0.83
vakio	0.061	0.190	0.1	1	0.749	1.06

Taulukko 8: Poistetut yhdysvaikutukset

yhdysvaikutus	p-arvo
sp × suorkk	0.904
sp × opetus	0.820
luok.ikä × opetus	0.667
opetus × suorkk	0.503
opetus × ajoyht	0.642
vao × luok.ikä	0.394
vao × suorkk	0.410
ajoyht × suorkk	0.282
sp × ajoyht	0.219
ajoyht × luok.ikä	0.113
suorkk × luok.ikä	0.094
sp × luok.ikä	0.069

Tässä tutkielmassa ei myöskään selvitetä kolmen tai useamman muuttujan välisiä yhdysvaikutuksia.

Loppujen lopuksi malliin tulevat siis muuttujat *oppilaan sukupuoli*, *luokiteltu oppilaan ikä*, *opetustapa*, *vastaanottaja* ja *ajo-opetus yhteensä* sekä yhdysvaikutus *opetustapa* × *vastaanottaja*. Tämä malli voidaan esittää muodossa:

$$\begin{aligned} g(\textit{ajokokeen tulos}) \;=\; & \beta_0 + \beta_1 \textit{oppilaan sukupuoli}_i \\ & + \beta_2 \textit{luokiteltu oppilaan ikä}_j + \beta_3 \textit{opetustapa}_k \\ & + \beta_4 \textit{vastaanottaja}_m + \beta_5 \textit{ajo-opetus yhteensä}_l \\ & + \beta_6 \textit{opetustapa}_k \times \textit{vastaanottaja}_m \end{aligned}$$

missä $i, j, k = 0, 1$; $l = 1, 2, 3$; $m = 1, 2, \ldots, 6$ ja luokiteltujen muuttujien viimeistä luokkaa on käytetty vertailuluokkana. Sama malli on myös Taulukossa (9) sivulla 19.

3.7 Mallin tulkintaa

Oppilaan sukupuoli vaikuttaa ajokokeen lopputulokseen. Miehet selviytyvät ajokokeesta naisia paremmin. Heidän ristitulosuhteensa naisiin verrattuna on 1.482.

Oppilaan ikä vaikuttaa lievästi ja negatiivisesti ajokokeen läpäisyyn. Yli 18-vuotiaat pärjäävät ajokokeessa hieman heikommin kuin 18-vuotiaat. Yli 18-vuotiaiden ristitulosuhde 18-vuotiaisiin verrattuna on 0.808.

Taulukko 9: Lopullinen malli

muuttuja	β	SE	Wald	df	p-arvo	exp(β)
sp	0.393	0.086	21.0	1	0.000	1.482
luok.ikä	−0.214	0.089	5.8	1	0.016	0.808
opetus	0.570	0.237	5.8	1	0.016	1.769
vao			20.4	5	0.001	
vao(1)	0.450	0.337	1.8	1	0.182	1.569
vao(2)	0.206	0.369	0.3	1	0.576	1.229
vao(3)	−0.392	0.308	1.6	1	0.204	0.676
vao(4)	1.158	0.343	11.4	1	0.001	3.185
vao(5)	0.299	0.326	0.8	1	0.358	1.349
ajoyht			45.3	2	0.000	
ajoyht(1)	0.763	0.124	38.1	1	0.000	2.144
ajoyht(2)	0.253	0.123	4.2	1	0.040	1.288
opetus × vao			16.9	5	0.005	
op. × vao(1)	−0.606	0.371	2.7	1	0.103	0.546
op. × vao(2)	−0.134	0.422	0.1	1	0.752	0.875
op. × vao(3)	0.229	0.336	0.5	1	0.496	1.257
op. × vao(4)	−1.173	0.371	10.0	1	0.002	0.309
op. × vao(5)	−0.496	0.359	1.9	1	0.168	0.609
vakio	−0.367	0.209	3.1	1	0.079	0.693

Muuttuja *ajo-opetus yhteensä* ja varsinkin sen luokkaan 1=”30 ajokertaa” kuuluminen vaikuttaa voimakkaasti ajokokeen tulokseen. Yli 34 ajokertaa ajaneet (luokka 3) selviytyvät ajokokeesta huonommin kuin 31–34 kertaa (luokka 2) tai 30 kertaa (luokka 1) ajaneet. Kaikkein parhaiten selviytyvät vähiten eli 30 kertaa ajaneet. Heidän riskinsä selviytyä ajokokeesta on 2.141–kertainen verrattuna yli 34 kertaa ajaneisiin. Kuten myöhemmin havaitaan, autokoululaisilla erot luokkien välillä saattavat johtua oppilaiden valikoitumisesta. Hyvät ajo-oppilaat oppivat ajamaan 30 ajokerran aikana. Nämä parhaat oppilaat jäävät pois niiden oppilaiden joukosta, jotka tarvitsevat lisä-ajokertoja. Tästä syystä enemmän harjoitelleet selviytyvät ajokokeesta heikommin kuin vähemmän ajaneet. Myöhemmin käy myös ilmi, että jos tämä sama malli sovitetaan aineistoon, joka sisältää vain opetuslupalaiset, niin heille muuttujalla *ajo-opetus yhteensä* ei ole merkitystä.

Muuttujien *opetustapa* ja *vastaanottaja* välillä on yhdysvaikutusta ($p = 0.005$). Sitä on havainnollistettu Kuvassa(1) (s. 21) ja Kuvassa(2) (s. 22). Kuvassa(1) kaikkien muiden vastaanottajien suhtautumista ajo-oppilaisiinsa

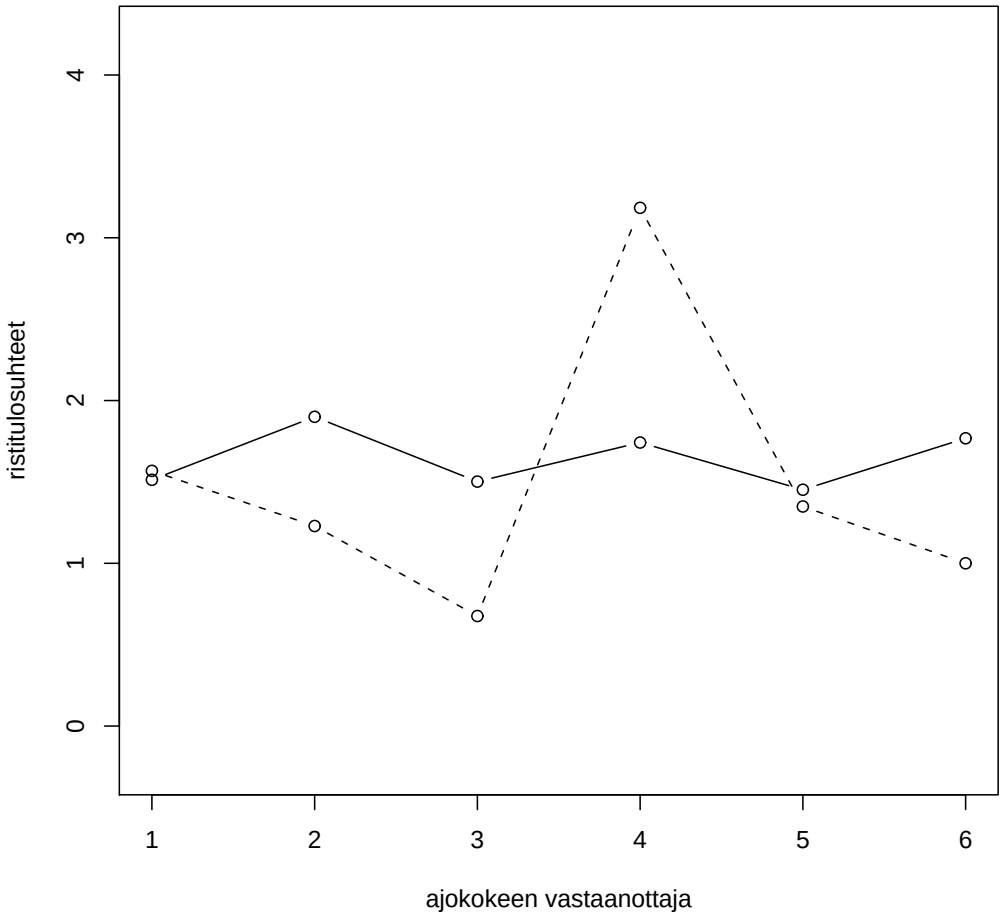
verrataan vastaanottajan vao(6) suhtautumiseen omiin opetuslupalaisiinsa, joiden riskiä selviytyä ajokokeesta kuvataan luvulla 1. Esim. vastaanottajan vao(2) kohdalla autokoululaisen riski selviytyä ajokokeesta on noin 1.900 ($= e^{0.642} = e^{(0.570+0.206-0.134)-(0+0+0)}$) verrattuna vastaanottajan vao(6) vastaanottaman opetuslupalaisen riskiin selviytyä siitä.

Kuvassa(2) (s. 22) vertailu tapahtuu kunkin vastaanottajan sisällä. Siinä on ilmoitettu vastaanottajan autokoululaiselleen antama etu verrattuna saman vastaanottajan opetuslupalaiseen. Esim. vastaanottajalla vao(3) autokoululaisen riski selviytyä ajokokeesta on 2.223 ($= e^{0.799} = e^{(0.570-0.392+0.229)} \times e^{(-(0-0.392+0))}$) verrattuna saman vastaanottajan opetuslupalaiseen.

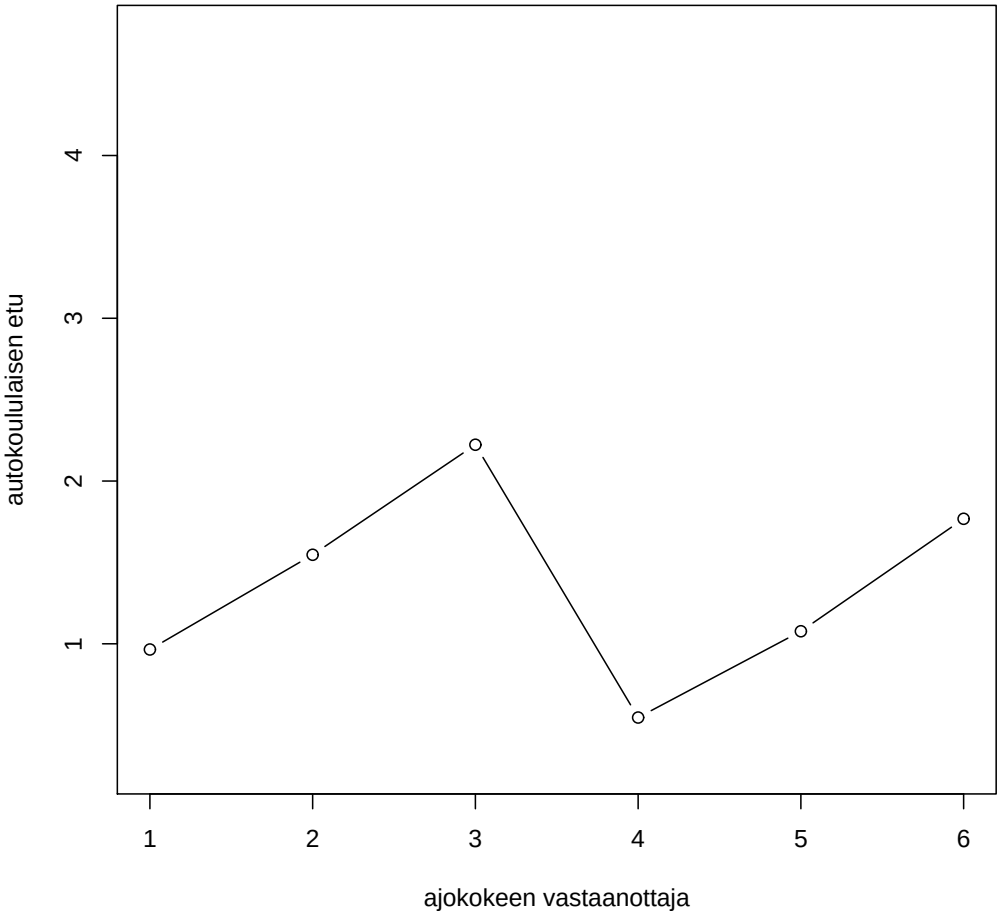
Kun yleisesti autokoululaiset selviytyvät ajokokeesta opetuslupalaisia paremmin (riskitulosuhde 1.769), niin se, kuka ajokokeen vastaanottaja oppilaalle on sattunut, muuttaa jonkin verran tätä suhdetta. Vastaanottajalla on vaikutusta. Muita vastaanottajia verrataan vastaanottajaan numero 6 (vao(6)). Eniten vastaanottajasta vao(6) poikkeaa vastaanottaja vao(4). Kun vastaanottajan vao(6) ajattama oppilas saa β-kertoimen 0, niin vastaanottajan vao(4) oppilas saa kertoimen 1.158 ($p = 0.001$). Jostain syystä vastaanottajan vao(4) ajattamat opetuslupalaiset menestyvät ajokokeessa poikkeuksellisen hyvin ja jopa paremmin kuin yhdenkään vastaanottajan autokoululaiset (Ks. Kuva(1), s. 21). Kuvasta (1) näkyy myös, että autokoululaiset menestyvät ajokokeessa suurin piirtein yhtä hyvin riippumatta siitä, kuka ajokokeen on ottanut vastaan. Sen sijaan opetuslupalaisten menestyminen vaihtelee enemmän eri vastaanottajien välillä.

Tarkastellaan tarkemmin muuttujien *opetustapa* ja *vastaanottaja* välistä yhdysvaikutusta. Sovitetaan aineistoon sama malli kuin aikaisemmin, mutta tarkastellaan vain opetusluvalla harjoitelleita. Tulokset näkyvät Taulukosta (10) sivulta 23. Opetuslupalaisille ajokokeenvastaanottajalla on merkitystä. Erityisesti ne oppilaat, joiden ajokokeen on ottanut vastaan vastaanottaja vao(4), ovat menestyneet kokeessa poikkeuksellisen hyvin. Sen sijaan muuttujalla *ajo-opetus yhteensä* ei ole merkitystä opetuslupalaisille. Opetuslupalaiset menestyvät ajokokeessa yhtä hyvin riippumatta siitä, minkä verran heillä on ajokertoja.

Taulukossa (11) sivulla 24 on malli, jossa ovat mukana vain autokoululaiset. Heidän menestymiseensä ajokokeessa ei vaikuta *vastaanottaja* eikä myöskään muuttuja *luokiteltu oppilaan ikä*. Sen sijaan muuttujalla *ajo-opetus yhteensä* on vaikutusta autokoululaisten menestymiseen. Opetuslupalaisten menestymiseenhän tällä muuttujalla ei ollut vaikutusta. Se, että muuttuja *ajo-opetus yhteensä* selittää autokoululaisten menestymistä ajokokeessa mutta ei selitä opetuslupalaisten menestymistä, voi johtua siitä, että autokouluissa saatetaan olla paljon tarkempia ajokertojen määrästä. Autokouluissa pyritään siihen, että oppilas oppii ajamaan autoa riittävän hyvin 30 ajokerran



Kuva 1: Opetustavan ja vastaanottajan yhdysvaikutus sovitetun mallin perusteella (y-akselilla ristitulosuhteet verrattuna vastaanottajan vao(6) opetuslupalaisiin, x-akselilla vastaanottajat vao(1–6)). Yhtenäinen viiva kuvaa autokoululaisia ja katkoviiva opetuslupalaisia.



Kuva 2: Kunkin vastaanottajan autokoululaiselle antama etu verrattuna saman vastaanottajan opetuslupalaiseen. (y-akselilla autokoululaisten vastaanottajakohtaiset edut, x-akselilla vastaanottajat vao(1–6)). Vastaanottajan vao(6) kohdalla pelkkä autokoulun ajo-oppilaalle antama etu.

Taulukko 10: Malli, jossa ovat vain opetuslupalaiset (412 oppilasta)

muuttuja	β	SE	Wald	df	p-arvo	exp(β)
sp	0.516	0.209	6.084	1	0.014	1.675
luok.ikä	−0.622	0.247	6.364	1	0.012	0.537
vao			20.823	5	0.001	
vao(1)	0.469	0.343	1.874	1	0.171	1.599
vao(2)	0.175	0.373	0.221	1	0.639	1.191
vao(3)	−0.421	0.312	1.821	1	0.177	0.656
vao(4)	1.166	0.348	11.235	1	0.001	3.210
vao(5)	0.275	0.330	0.694	1	0.405	1.316
ajoyht			0.038	2	0.981	
ajoyht(1)	−0.106	0.549	0.038	1	0.846	0.899
ajoyht(2)	−0.013	0.308	0.002	1	0.967	0.987
vakio	−0.257	0.238	1.169	1	0.280	0.773

aikana. Lisäajokertoja annetaan vain, jos se on välttämätöntä. Opetusluvalla harjoitteleville on tärkeää oppia vain ajamaan eikä ajokertojen määrään kiinnitetä niin paljon huomiota. He saattavat ajaa muutaman ylimääräisen ajokerran ihan vain varmuuden vuoksi ja kuten Taulukosta(4) (s. 13) näkyy, he ajavatkin selkeästi enemmän kuin autokoululaiset ja heidän ajomäärissään on enemmän vaihtelua. Luultavasti opetuslupalaisten ajokertoja ei myöskään merkitä ylös niin tarkasti kuin autokoululaisten.

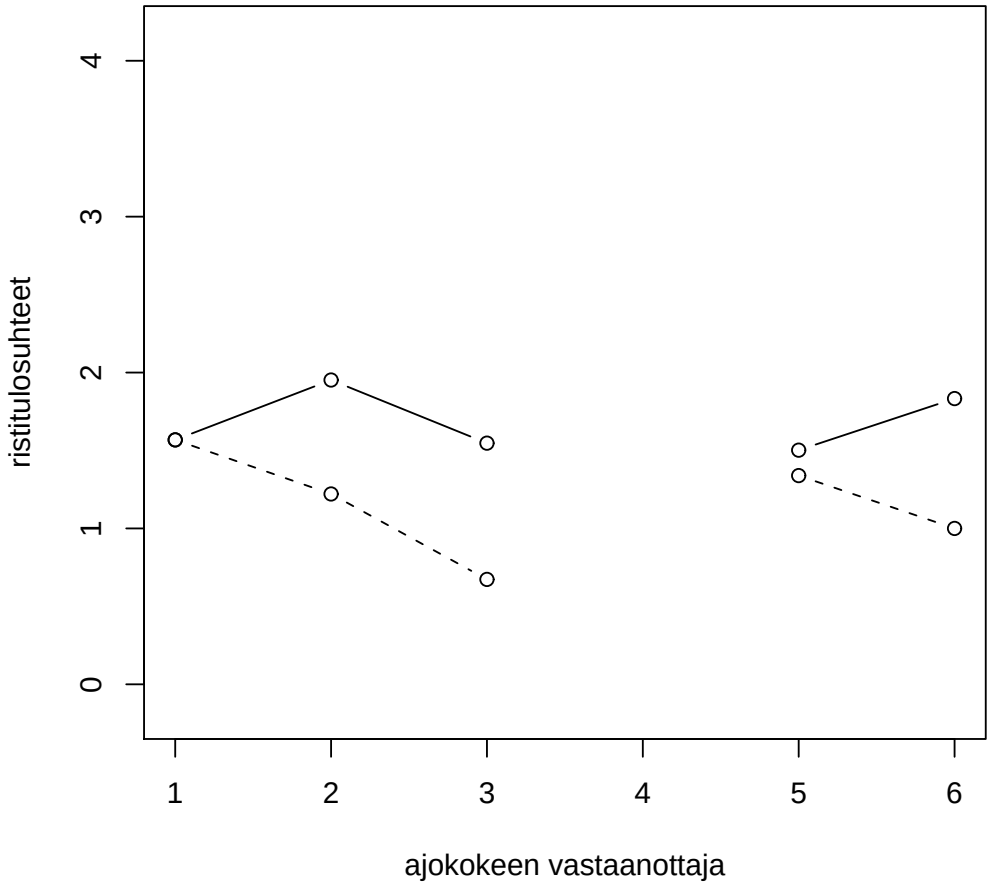
Taulukossa (12) sivulla 24 on malli, josta on jätetty pois vastaanottaja vao(4). Sivulla 25 on Kuva (3), joka esittää *opetustavan* ja *vastaanottajan* yhdysvaikutusta ilman vastaanottajaa vao(4). Nyt vastaanottajalla ei olekaan vaikutusta ajokokeessa menestymiseen eikä muuttujien *opetustapa* ja *vastaanottaja* välillä ole yhdysvaikutusta. Tästä voidaan päätellä, että vain vastaanottajalla vao(4) oli vaikutusta ajokokeen lopputulokseen ja tämäkin vaikutus ilmeni vain opetuslupalaisilla.

Taulukko 11: Malli, jossa ovat vain autokoululaiset (2367 oppilasta)

muuttuja	β	SE	Wald	df	p-arvo	exp(β)
sp	0.352	0.095	13.862	1	0.000	1.422
luok.ikä	−0.139	0.096	2.095	1	0.148	0.870
vao			3.909	5	0.563	
vao(1)	−0.155	0.155	0.999	1	0.317	0.857
vao(2)	0.071	0.206	0.119	1	0.730	1.073
vao(3)	−0.162	0.133	1.482	1	0.224	0.850
vao(4)	−0.019	0.141	0.019	1	0.891	0.981
vao(5)	−0.196	0.153	1.649	1	0.199	0.822
ajoyht			51.229	2	0.000	
ajoyht(1)	0.871	0.132	43.260	1	0.000	2.390
ajoyht(2)	0.348	0.136	6.554	1	0.010	1.416
vakio	0.108	0.149	0.518	1	0.472	1.114

Taulukko 12: Malli ilman vastaanottajaa vao(4)

muuttuja	β	SE	Wald	df	p-arvo	exp(β)
sp	0.436	0.095	20.944	1	0.000	1.547
luok.ikä	−0.250	0.099	6.436	1	0.011	0.779
opetus	0.606	0.239	6.410	1	0.011	1.833
vao			7.044	4	0.134	
vao(1)	0.450	0.338	1.771	1	0.183	1.568
vao(2)	0.200	0.369	0.295	1	0.587	1.222
vao(3)	−0.396	0.309	1.643	1	0.200	0.673
vao(5)	0.292	0.326	0.803	1	0.370	1.339
ajoyht			36.259	2	0.000	
ajoyht(1)	0.733	0.136	28.830	1	0.000	2.081
ajoyht(2)	0.195	0.135	2.088	1	0.148	1.216
opetus × vao			6.563	4	0.161	
op. × vao(1)	−0.606	0.372	2.660	1	0.103	0.545
op. × vao(2)	−0.137	0.423	0.105	1	0.746	0.872
op. × vao(3)	0.226	0.337	0.451	1	0.502	1.254
op. × vao(5)	−0.491	0.360	1.863	1	0.172	0.612
vakio	−0.368	0.210	3.072	1	0.080	0.692



Kuva 3: Opetustavan ja vastaanottajan yhdysvaikutus mallissa, josta on jätetty pois vastaanottaja vao(4) (y-akselilla ristitulosuhteet verrattuna vastaanottajan vao(6) opetuslupalaisiin, x-akselilla vastaanottajat vao(1–3, 5–6)). Yhtenäinen viiva kuvaa autokoululaisia ja katkoviiva opetuslupalaisia.

4 Pohdintaa

Oppilaan sukupuoli ja opetustapa paljastuivat kaikkein voimakkaimmin ajokokeen lopputulokseen vaikuttaviksi muuttujiksi. Miehet selviytyvät ajokokeesta naisia paremmin ja autokoululaiset paremmin kuin opetusluvalla harjoitelleet. Sukupuolieroista osa voi selittyä sillä, että miehillä saattaa olla ollut enemmän kiinnostusta autoja ja autoilua kohtaan sekä myös enemmän kokemusta autolla ajamisesta ennen ajokortin suorittamista kuin naisilla.

Taulukon (10) (s. 23) ja Taulukon (11) (s. 24) perusteella näyttäisi siltä, että autokoululla saattaisi olla menestymiseroja tasoittava vaikutus. Autokoulussa oppilaan ikä ei vaikuta ajokokeessa menestymiseen ja sukupuolenkin vaikutus heikkenee (autokoulussa $\beta = 0.352$, opetusluvalla $\beta = 0.516$). Toisin sanoen naisten ja miesten välinen ero ajokokeessa menestymisessä ei ole niin suuri autokoulun käyneillä kuin opetusluvalla harjoitelleilla. Samalla autokoulu myös lisäsi sekä miesten että naisten menestymistä ajokokeessa. Ehkä autokoulussa saa näin ollen laadukkaampaa opetusta kuin opetusluvalla. Useinhan opetusluvan haltija opettaa vain yhden oppilaan ajamaan autolla, kun autokoulun opettajalla on ehkä takanaan vuosien kokemus opettamisesta ja satoja ajokortin saaneita oppilaita.

Ajokokeen vastaanottajat eivät arvioineet oppilaiden ajokokeita mitenkään toisistaan poikkeavasti ellei vastaanottajaksi sattunut vastaanottaja vao(4), jolloin opetuslupalaiset menestyivät poikkeuksellisen hyvin. Sitä, mistä vastaanottajan vao(4) poikkeuksellinen vaikutus johtui, ei tässä pystytä selittämään. Oppilaathan eivät voi valita ajokokeen vastaanottajaa ja vastaanottajat arvostelevat samoin kriteerein kaikkien oppilaiden ajosuoritukset. Vastaanottajan koulutuksella ei ollut vaikutusta ajokokeen lopputulokseen ($p = 0.955$). Tietoja vastaanottajien iästä ja sukupuolesta ei ollut käytettävissä, joten niiden merkitystä ei voitu selvittää.

Muuttuja *ajo-opetus yhteensä* oli ongelmallinen siksi, että autokoululaisten ja opetuslupalaisten ajokertojen määrät jakautuivat aivan eri tavoin (Taulukko(4), s. 13). Tavoite saman kokoisista luokista toteutui hyvin vain autokoululaisten osalta. Kolmiluokkainen luokittelu kertoo myös paljon tarkemmin autokoululaisten kuin opetuslupalaisten ajo-opetusmääristä. Autokouluissa pyritään siihen, että 30 ajokertaa riittää. Sen sijaan opetuslupalaiset ajavat enemmän ja heidän ajokertojensa määrä vaihtelee paljon enemmän kuin autokoululaisilla. Muuttuja *ajo-opetus yhteensä* myös vaikutti voimakkaasti muiden riippumattomien muuttujien kertoimiin eli se on sekoittava muuttuja.

Kirjallisuutta

[1] Hosmer, David W. & Lemeshow, Stanley. (1989) *Applied logistic regression*. New York: John Wiley & Sons, Inc.

[2] Paakkanen, Mikko. *Inssiajojen reputusmäärät vaihtelevat maakuntien ja sukupuolten välillä*. Helsingin Sanomat 14.1.2007