Tampereen yliopisto

Taru Hakamäki

ZERO-INFLATED-MALLIT RATKAISUNA VASTEMUUTTUJAN NOLLA-ARVOJEN YLIEDUSTUKSEEN

Informaatioteknologian ja viestinnän tiedekunta
Kandidaattitutkielma
Huhtikuu 2020

Tiivistelmä

Taru Hakamäki: Zero-inflated-mallit ratkaisuna vastemuuttujan nolla-arvojen yliedustukseen
Kandidaattitutkielma
Tampereen yliopisto
Matematiikan ja tilastotieteen tutkinto-ohjelma
Huhtikuu 2020

Tämän tutkielman tarkoituksena on esitellä zero-inflated-Poissonin jakauma ja zero-inflated-negatiivinen binomijakauma. Näillä jakaumilla voidaan mallintaa lukumäärävasteisia aineistoja, joissa vastemuuttujassa on suuri määrä nolla-arvon saavia havaintoja. Vastemuuttujan nolla-arvojen yliedustuksen ymmärtäminen ehkäisee virheellisiä tulkintoja ja auttaa ymmärtämään havaintojen muodostumista.

Perinteisiä tapoja mallintaa lukumäärävasteita ovat Poissonin jakauma ja negatiivinen binomijakauma. Poissonin jakauma olettaa vastemuuttujan odotusarvon ja varianssin yhtä suuriksi. Jos tämä oletus ei toteudu, aineistossa sanotaan olevan ylihajontaa. Tällöin negatiivinen binomijakauma on parempi valinta mallintamaan aineistoa. Zero-inflated-malleja voidaan soveltaa perinteisten menetelmien sijaan, kun vastemuuttuja sisältää ylihajonnan lisäksi runsaasti nolla-arvoja. Nämä mallit jakavat vastemuuttujan arvot kahteen kuvitteelliseen ryhmään. Toinen ryhmä muodostuu lukumääristä, jotka noudattavat Poissonin jakaumaa tai negatiivista binomijakaumaa. Tässä ryhmässä mahdollisia arvoja ovat kaikki luonnolliset luvut. Toinen ryhmä muodostuu ainoastaan rakenteellisista nolla-arvoista, jotka syntyvät usein aineistonkeruun puutteista. Tällaiset nollat eivät ole seurausta satunnaisvaihtelusta, vaan ne saavat aina arvon nolla. Rakenteellisten nollien todennäköisyyttä mallinnetaan logistisella regressiolla.

Edellä esiteltyjä jakaumia sovelletaan esimerkkitutkimuksessa työntekijöiden sairauspoissaolokertojen lukumääriä kuvaavaan aineistoon. Yleisen periaatteen mukaan parhaaksi malliksi valitaan mahdollisimman yksinkertainen malli, joka selittää vastemuuttujan arvoja riittävän hyvin. Tässä tutkimuksessa ei voida valita yksiselitteisesti yhtä parasta mallia kuvaamaan tutkittua aineistoa. Sovitetuista malleista zero-inflated-negatiivinen binomijakauma näyttää olevan paras, kun vertaillaan Akaiken informaatiokriteerejä. Ero ei kuitenkaan ole suuri verrattuna negatiiviseen binomijakaumaan, joka on yksinkertaisempi ja helpommin tulkittava.

Avainsanat: lukumäärävaste, Poissonin jakauma, ylihajonta, negatiivinen binomijakauma, zero-inflated-mallit

Tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck –ohjelmalla

Sisältö

1 Johdanto

Tilastollinen päättely aineistosta, jossa vastemuuttujan nolla-arvot ovat yliedustettuna, on usein tehoton tai jopa virheellinen, jollei nolla-arvojen alkuperää ja syitä niiden muodostumiselle pohdita tarkoin. Zero-inflated-mallit (ZI) ovat yksi tapa käsitellä nolla-arvojen yliedustusta. Tässä tutkielmassa esitellään Zero-inflated-Poisson- (ZINP) ja zero-inflated-negatiivinen binomijakauma (ZINB) ratkaisuna mallintaa aineistoja, joissa on runsaasti nolla-arvoja. ZI-mallien lisäksi esitellään lyhyesti perinteisemmät lukumäärävasteita mallintavat Poisson- ja negatiivinen binomimalli (NB). Myös näiden perinteisempien jakaumien perusidea tulee ymmärtää, sillä ZIP-jakauma hyödyntää Poissonin jakaumaa ja ZINB-jakauma hyödyntää negatiivista binomijakaumaa. Näillä neljällä mallilla saadaan mallinnettua suuri osa lukumäärävasteisista aineistoista.

ZI-malleja voidaan soveltaa aineistoihin, joiden vastemuuttuja mittaa lukumääriä, joista suuri osa on nolla-arvoja. Lukumäärämuotoisista muuttujista puhutaan, kun muuttuja voi saada vain arvoja, jotka kuuluvat luonnollisten lukujen joukkoon. Beaujean ja Grant (2016) listaavat lukumäärämuotoisille muuttujille kolme ominai-suutta: ne ovat aina kokonaislukuja, pienin mahdollinen arvo on nolla eivätkä arvot koskaan ole negatiivisia ja niiden jakauma on usein oikealle vinoutunut niin, että suuri osa arvoista on pieniä ja suuremmat arvot ovat harvinaisempia.

Tämä tutkielma jakautuu Menetelmät-osioon ja Esimerkkitutkimus-osioon. Menetelmät-osio on jaettu kolmeen osaan. Ensimmäinen osa esittelee perinteisinä lukumäärämuotoisia aineistoja mallintavina jakaumina lyhyesti Poissonin jakauman ja negatiivisen binomijakauman. Toisessa osassa esitellään ZI-mallit yleisemmin ja tarkemmin ZIP- ja ZINB-mallit. Kolmannessa osassa esitellään menetelmiä, joilla voidaan valita aineistoon parhaiten sopiva malli. Esiteltyjä menetelmiä sovelletaan esimerkkitutkimuksessa R-ohjelmistoa hyödyntäen työntekijöiden sairauspoissaoloja kuvaavaan aineistoon. Tutkimuksessa valitaan aineistoon parhaiten sopiva malli vertailemalla Poisson-, NB-, ZIP-, ja ZINB-malleja sekä tulkitaan mallin antamat tulokset.

2 Menetelmät

2.1 Perinteisiä jakaumia lukumäärämuotoisille aineistoille

2.1.1 Poissonin jakauma

Poissonin jakauma on yksinkertaisin lukumääräaineistoja kuvaava jakauma. Sillä on vain yksi parametri, $\mu \geq 0$, joka kuvaa sekä odotusarvoa että varianssia

$$E(Y) = \text{var}(Y) = \mu. \tag{2.1}$$

Kun μ on lähellä nollaa, jakauma on oikealle vinoutunut, mutta lähestyy normaalijakaumaa, kun μ kasvaa. (Beaujean & Grant, 2016.) Poissonin jakauman pistetodennäköisyysfunktio on muotoa:

$$f(y;\mu) = \frac{\mu^y \cdot e^{-\mu}}{y!}, y = 0, 1, 2, \ldots \text{ ja } \mu \geq 0. \tag{2.2}$$

Funktiosta (2.2) saadaan todennäköisyys vastemuuttujan arvolle y_i, kun odotusarvo μ tunnetaan. (Zuur et al. 2009.)

Poissonin regressiomalli olettaa, että selittävät muuttujat ovat linkittyneitä vastemuuttujaan log-muunnoksella. Yksinkertaiselle Poissonin regressiomallille, jossa on vain yksi selittävä muuttuja, voidaan kirjoittaa malli

$$log(\mu_i) = \alpha + \beta \cdot x_i, \tag{2.3}$$

missä μ_i on odotusarvo vastemuuttujalle, α on mallin vakiotermi, β on estimoitu regressiokerroin ja x_i kuvaa ryhmää selittävän muuttujan arvoista, jotka saavat saman arvon x. Mallin vastemuuttujan palauttaminen lukumääräasteikolle vaati käänteisen linkkifunktion käyttöä, mikä tarkoittaa lukumäärämuotoisten muuttujien ja log-linkkifunktion kohdalla eksponenttimuunnosta. Eksponenttimuunnos yhtälöstä (2.3) on

$$\mu_i = e^{\alpha+\beta \cdot x_i}. \tag{2.4}$$

(Beaujean & Grant, 2016.)

2.1.2 Ylihajonta

Poissonin jakauman keskeisin oletus on odotusarvon ja varianssin yhtäsuuruus. Todellisissa aineistoissa varianssi on kuitenkin usein odotusarvoa suurempi, jolloin puhutaan ylihajonnasta. Joissakin tapauksissa varianssi voi olla odotusarvoa pienempi, jolloin puhutaan alihajonnasta, mutta tässä työssä keskitytään ylihajonnan vaikutuksiin. Kun aineistossa on runsaasti ylihajontaa, Poissonin jakauma ei useinkaan ole yksin riittävä tapa mallintaa aineistoa, sillä sen oletus odotusarvon ja varianssin yhtäsuuruudesta ei toteudu. (Zuur et al. 2009.)

Hilbe (2011) esittää, että ylihajonta voi olla näennäistä, mallin puutteista johtuvaa tai aineistossa todella esiintyvää ylihajontaa. Näennäisen ylihajonnan taustalla saattaa olla mallista puuttuva tärkeä selittävä tekijä tai interaktio, poikkeavat havainnot, vastemuuttujan virheellinen skaalaus tai väärän linkkifunktion käyttö. Jos ylihajonnan ei voida katsoa johtuvan edellä mainituista tekijöistä, sen voidaan olettaa olevan seurausta aineistosta, esimerkiksi nollien yliedustuksesta. Ylihajonta voi johtaa estimaattien keskivirheiden pienenemiseen, jolloin muuttuja saatetaan katsoa tilastollisesti merkitseväksi, vaikkei se sitä olisi (Hilbe 2011).

Zuur et al. (2009) esittävät ylihajonnan selvittämiseksi laskukaavan:

$$\hat{\phi} = \frac{D}{n-p}. \tag{2.5}$$

Kaavassa (2.5) ylihajontaa kuvaava estimaatti $\hat{\phi}$ saadaan jakamalla devianssi D vapausasteilla n-p. Devianssi saadaan Poisson- ja binomimalleille kaavasta

$$D = -2[L(\hat{\boldsymbol{\mu}}; \mathbf{y}) - L(\mathbf{y}; \mathbf{y})], \tag{2.6}$$

missä L($\hat{\boldsymbol{\mu}}$;**y**) on tutkitun mallin uskottavuusfunktio ja L(**y**;**y**)] on saturoidun mallin uskottavuusfunktio. Saturoidussa mallissa on yhtä monta parametria kuin aineistossa on havaintoja, joten sen sovite on täydellinen kopio aineistosta. Devianssi kertoo kuinka paljon malli eroaa saturoidusta mallista, joten sillä voidaan tutkia mallin sopivuutta aineistoon. (Agresti 2003.)

Jos kaavasta (2.5) saatu ylihajontaa kuvaava estimaatti $\hat{\phi}$ on suurempi kuin 1, aineistossa voidaan katsoa olevan ylihajontaa eikä Poissonin jakauman oletus varianssin ja odotusarvon yhtäsuuruudesta toteudu. Hilbe (2011) muistuttaa kuitenkin, että pieni määrä ylihajontaa ei ole vaarallista. Se, kuinka paljon yli yhden ylihajontaa kuvaava estimaatti voi olla ilman, että siihen tarvitsee puuttua, riippuu havaintojen lukumäärästä. Hilben esimerkin mukaan ylihajontaa kuvaavan estimaatin arvoon 1.1 ei tarvitse puuttua, jos havaintojen lukumäärä on 100, mutta jos havaintojen lukumäärä on 100 000, ylihajontaan on reagoitava. Kun aineistoissa on ylihajontaa, Hilbe suosittelee käyttämään negatiivista binomijakaumaa tai, jos vastemuuttujassa on lisäksi runsaasti nollia, zero-inflated-mallia tai zero-altered-mallia.

2.1.3 Negatiivinen binomijakauma

Negatiivisella binomijakaumalla (NB) voidaan mallintaa luonnollisista luvuista muodostuvia vasteita. Alunperin negatiivinen binomijakauma kehitettiin mallintamaan epäonnistumisten lukumäärää ennen r:ttä onnistumista riippumattomissa Bernoullin kokeissa. Negatiivinen binomijakauma voidaan kuitenkin nähdä myös Poissonin jakauman ja gammajakauman sekoitteena. Tällöin sillä voidaan mallintaa diskreettejä vasteita, joita ei voi mallintaa Poissonin jakaumalla ylihajonnan vuoksi. (Hilbe 2011.)

NB-jakauman pistetodennäköisyysfunktio on muotoa:

$$f(y;k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\cdot\Gamma(y+1)}\cdot(\frac{k}{\mu+k})^k\cdot((1-\frac{k}{\mu+k})^y), \tag{2.7}$$

missä y = 0, 1, 2, ..., $\mu \geq 0$ on odotusarvo, k > 0 on hajontaparametri ja Γ(y+1)=y!. NB-jakauman odotusarvo ja varianssi ovat:

$$E(Y) = \mu, \tag{2.8}$$

$$\text{var}(Y) = \mu + \frac{\mu^2}{k}. \tag{2.9}$$

(Zuur et al. 2009.)

NB-jakauman varianssin ensimmäinen termi μ on Poissonin jakauman varianssi, ja toinen termi μ^2/k on gammajakauman varianssi (Hilbe 2011). NB-jakaumalla on kaksi parametria μ ja k, missä μ on odotusarvo ja k kuvaa epäsuorasti ylihajonnan määrää aineistossa. Mitä pienemmän arvon k saa, sitä enemmän aineistossa on ylihajontaa. Kun k kasvaa suureksi suhteessa odotusarvoon, varianssin toinen termi lähestyy nollaa ja NB-jakauma lähestyy Poissonin jakaumaa. (Zuur et al. 2009.)

Myös negatiivisen binomijakauman regressiomalli käyttää log-linkkifunktiota Poissonin regressiomallin tapaan. Log-linkki varmistaa, että ennustetut arvot ovat aina ei-negatiivisia. (Zuur et al. 2009.)

2.2 Nolla-arvojen yliedustus vastemuuttujassa

Toisinaan aineiston ylihajonta johtuu vastemuuttujan nolla-arvojen yliedustuksesta. Nolla-arvojen yliedustuksen voi yleensä havaita jo vastemuuttujan frekvenssijakaumasta, missä nolla-arvojen kohdalla on muihin arvoihin verrattuna selkeä piikki. Negatiivinen binomijakauma ei pysty käsittelemään ylihajontaa, joka johtuu nollien yliedustuksesta. Tällöin on käytettävä zero-inflated-malleja (ZI) tai zero-altered-malleja (ZA), jotka voivat käsitellä nollien yliedustuksesta johtuvaa ylihajontaa. (Zuur et al. 2009.)

Ylimääräisten nollien ja niiden lähteen huomioiminen on tärkeää, sillä huomiotta jättäminen saattaa johtaa estimointivirheisiin tai lisätä aineiston ylihajontaa (Zuur et al. 2009). Esimerkiksi virheelliseen päättelyyn johtavasta tilanteesta voisi olla tutkimus alkoholin käytön riskeistä. Jos tutkitaan juotujen alkoholiannosten lukumäärää edellisellä viikolla, saadaan nolla-arvoja henkilöiltä, jotka juovat alkoholia, mutta eivät ole juoneet sitä juuri edellisen viikon aikana sekä henkilöiltä, jotka eivät koskaan juo alkoholia. Nämä kaksi ryhmää ovat kuitenkin todennäköisesti täysin eri riskiryhmissä alkoholin käytön riskejä tutkittaessa, vaikka molemmat tuottavat nolla-arvon.

Nolla-arvojen tunteminen auttaa myös sopivan mallin valinnassa. ZI-malleja käytetään, kun ajatellaan, että nollia muodostuu kahdesta eri prosessista: satunnaisvaihtelusta johtuvat nollat ja aineistonkeruun heikkouksista johtuvat nollat (Zuur et al. 2009). Edelliseen esimerkkiin, jossa tutkittiin alkoholinkäyttöä, olisi hyvä soveltaa ZI-mallia. Siinä ylimääräisiä nolla-arvoja tuottaa lyhyt tutkimusperiodi sekä henkilöt, jotka eivät koskaan juo alkoholia.

ZA-malleja sovelletaan, jos nolla-arvojen yliedustuksen voidaan ajatella johtuvan jostakin kynnyksestä, joka ilmiön tulee ylittää ennen kuin se saa nollaa suuremman arvon. (Zuur et al. 2009.) Esimerkki tutkimuksesta, jossa olisi hyvä käyttää ZA-mallia, on liikennevahinkojen lukumäärien tutkiminen. Vakuutusyhtiöiden tietoon tulee usein vain suuremmat vahingot, sillä vakuutuksenottaja saattaa jättää ilmoittamatta vahingosta, jos vahinko ei ylitä omavastuurajaa tai vakuutuksenottaja ei halua menettää vakuutusmaksuihin liittyviä bonusalennuksia. Tällöin vakuutusyhtiö tilastoi vakuutuksenottajalle nolla vahinkoa, vaikka vahinko olisikin todellisuudessa tapahtunut.

ZA-mallit jakavat vastemuuttujan arvot nolliin ja positiivisiin arvoihin. Binomiprosessi mallintaa todennäköisyyttä, että havainto saa arvon nolla. Positiiviset arvot mallinnetaan perinteisesti käyttäen Poissonin jakaumaa tai negatiivista binomijakaumaa. (Zuur et al. 2009.) Tässä työssä keskitytään pääasiassa zero-inflated-malleihin, joita esitellään tarkemmin seuraavissa alaluvuissa.

2.2.1 Zero-inflated-mallit

Zero-inflated-malleja (ZI) voidaan soveltaa, kun lukumääräarvoisessa vastemuuttujassa esiintyy suuri määrä nollia ja nollien voidaan olettaa muodostuvan tosista ja rakenteellisista nollista. Todet nolla-arvot voivat saada myös positiivisen arvon, mutta satunnaisvaihtelusta johtuen ne ovat saaneet arvon nolla. Myös rakenteelliset nolla-arvot ovat yhtä lailla todellisia havaintoja, mutta nämä nollat eivät synny vain satunnaisvaihtelun seurauksena, vaan ovat usein seurausta puutteellisesta tutkimusasetelmasta tai ei-riskiryhmistä.

Martin et al. (2005) esittää kaksi eri tyyppistä todellista nolla-arvoa ja kaksi eri tyyppistä rakenteellista nolla-arvoa. Ensimmäinen tyyppi todellisesta nolla-arvosta ilmenee, kun havaintopaikassa ei esiinny tutkittavaa ilmiötä. Esimerkiksi sairauspoissaolokertoja tutkittaessa tällaisen nollan voisi tuottaa henkilö, joka ei ole työelämässä. Toinen tyyppi todellisesta nolla-arvosta on tilanne, jossa ympäristö on sopiva ilmiön esiintymiselle, mutta satunnaisvaihtelun vuoksi havaitaan silti nolla. Esimerkiksi työntekijä on voinut pysyä terveenä eikä kertaakaan ole tarvinnut sairauslomaa. Ensimmäinen tyyppi rakenteellisista nolla-arvoista on tilanne, jossa ilmiötä ei havaita, vaikka tutkitaan oikeassa ympäristössä, sillä tutkittava aikaperiodi on liian lyhyt. Esimerkiksi sairauspoissaoloja tutkittaessa saadaan sitä enemmän nolla-arvoja, mitä lyhyempää periodia tarkastellaan. Toinen rakenteellisia nollia tuottava tilanne esiintyy, kun ilmiö tapahtuu, mutta sitä ei havaita. Esimerkiksi henkilö sairastuu niin, että hänen tulisi jäädä sairauslomalle, mutta jostakin syystä hän kuitenkin työskentelee eikä sairauspoissaoloa hänen kohdaltaan näin tilastoida. Tällöin työntekijä kuuluu niin kutsuttuun ei-riskiryhmään, sillä tutkittu ilmiö ei koskaan tapahdu hänen kohdallaan.

ZI-malleissa aineisto jaetaan kuvitteellisesti kahteen osaan. Toinen osa mallintaa rakenteellisia nolla-arvoja binaarisesti käyttäen logistista regressiota ja toinen mallintaa Poisson- tai NB-jakaumalla tosia nolla-arvoja ja nollaa suurempia arvoja. Todellisuudessa aineistoa ei jaeta näihin osiin eikä tiedetä, kumpaan ryhmään tietty nolla kuuluu, vaan ainoastaan oletetaan, että nämä kaksi joukkoa ovat olemassa. Jos malli käyttää todellisten nollien ja positiivisten havaintojen mallintamiseen Poissonin jakaumaa, mallia kutsutaan zero-inflated-Poisson-jakaumaksi (ZIP). Jos malli hyödyntää taas negatiivista binomijakaumaa, mallia kutsutaan zero-inflated-negatiiviseksi binomijakaumaksi (ZINB). Valinta ZIP- ja ZINB-mallin välillä tulisi perustua vastemuuttujan positiivisten havaintojen ja tosien nolla-arvojen jakaumaan. ZIP-mallia voidaan käyttää, jos aineiston ylihajonta johtuu ainoastaan nollien yliedustuksesta ja todet nollat ja positiiviset havainnot noudattavat Poissonin jakaumaa. ZINB-malli on tarpeellinen, jos ylihajonta ei korjaannu ainoastaan huomioimalla rakenteelliset nollat käyttäen ZIP-jakaumaa, vaan ylihajontaa on myös tosien nollien ja positiivisten havaintojen joukossa. ZINB-mallin soveltami-

nen on siis hyödyllistä, kun vastemuuttujan todet nollat ja positiiviset havainnot noudattavat NB-jakaumaa, mutta lisäksi vasteen nolla-arvot ovat yliedustettuna johtuen rakenteellisista nollista. (Zuur et al. 2009.)

2.2.2 ZIP- ja ZINB-mallien matematiikka

Tässä luvussa on hyödynnetty Zuur et al. (2009, 274 – 278) esittämää tapaa johtaa ZIP- ja ZINB-mallien todennäköisyysfunktiot.

ZIP-jakauma

ZIP-mallin todennäköisyysjakauma on kaksiosainen. Toinen osa mallintaa todennäköisyyttä, että satunnaismuuttuja Y_i saa arvon nolla ja toinen todennäköisyyksiä, kun Y_i saa nollaa suuremman arvon. Muodostetaan ensin todennäköisyys, että Y_i saa arvon nolla.

$$P(Y_i = 0) = P(\text{rakenteellinen nolla}) + (1 - P(\text{rakenteellinen nolla})) \cdot P(\text{nolla Poisson-jakaumasta})$$

Todennäköisyys, että Y_i saa arvon nolla, koostuu kahdesta termistä. Ensimmäinen laskee todennäköisyyden rakenteellisille nolla-arvoille. Toisesta termistä saadaan todennäköisyys tosille nolla-arvoille kertomalla todennäköisyys, että nolla ei ole rakenteellinen nolla ja todennäköisyys, että Poissonin jakauma tuottaa nollan. Todennäköisyys P(rakenteellinen nolla) on binomijakautunut. Merkitään todennäköisyydeksi, että Y_i saa arvon rakenteellinen nolla π_i ja merkitään myös, että Y_i ei ole rakenteellinen nolla merkinnällä 1 - π_i. Näillä merkinnöillä voidaan kirjoittaa aiempi yhtälö muodossa

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \cdot P(\text{nolla Poisson-jakaumasta}).$$

Todennäköisyys P(nolla Poisson-jakaumasta) saadaan Poissonin jakauman pistetodennäköisyysfunktiosta (2.2), kun asetetaan $y_i = 0$.

$$P(y_i = 0; \mu) = \frac{\mu^0 \cdot e^{-\mu}}{0!} = e^{-\mu}$$

Kun edellinen tulos sijoitetaan todennäköisyyteen P($Y_i = 0$), ZIP-jakautunut aineisto tuottaa nolla-arvon todennäköisyydellä

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \cdot e^{-\mu}.$$

Seuraavaksi selvitetään todennäköisyys, että ZIP-jakautunut aineisto tuottaa nollaa suuremman arvon. Tämä saadaan kertomalla todennäköisyys, että arvo ei

ole rakenteellinen nolla ja todennäköisyys, että Poissonin jakauma tuottaa nollaa suuremman arvon.

$P(Y_i = y_i|y_i > 0) = (1-P(\text{rakent. nolla}))\cdot P(\text{nollaa suurempi arvo Poisson-jakaumasta})$

Koska oletetaan binomijakauma rakenteellisille nollille todennäköisyydellä π_i ja lukumääräarvoille Poissonin jakauma, voidaan kirjoittaa

$$P(Y_i = y_i|y_i > 0) = (1 - \pi_i) \cdot \frac{\mu^y \cdot e^{-\mu}}{y!}.$$

Saamme todennäköisyysfunktion ZIP-jakaumalle yhdistämällä todennäköisyys, että Y_i saa arvon nolla ja todennäköisyys, että Y_i saa positiivisen arvon

$$f(y_i) = \begin{cases} \pi_i + (1 - \pi_i) \cdot e^{-\mu} & , y_i = 0 \\ (1 - \pi_i) \cdot \frac{\mu^y \cdot e^{-\mu}}{y!} & , y_i > 0 \end{cases}. \tag{2.10}$$

Kaavassa 2.10 ylempi osa antaa todennäköisyyden, että vastemuuttujan havainto saa arvon nolla. Alempi osa kertoo todennäköisyyden arvolle y_i, kun $y_i>0$. ZIP-mallin odotusarvo ja varianssi ovat

$$E(Y_i) = \mu_i \cdot (1 - \pi_i) \tag{2.11}$$

$$\text{var}(Y_i) = (1 - \pi_i) \cdot (\mu_i + \pi_i \cdot \mu_i^2). \tag{2.12}$$

ZINB-jakauma

ZIP- ja ZINB-mallien todennäköisyysjakaumien ainoa ero on, että lukumääriä mallintava Poissonin jakauma vaihdetaan negatiiviseksi binomijakaumaksi. ZINB-mallin johtaminen seuraa samaa kaavaa kuin ZIP-mallin johtaminen. Aluksi selvitetään todennäköisyys, että malli tuottaa arvon nolla ja sitten todennäköisyysjakauma positiivisille arvoille.

Nolla-arvojen todennäköisyys ZINB-jakaumassa on muotoa

$$P(Y_i = 0) = \pi_i + (1 - \pi_i) \cdot (\frac{k}{\mu + k})^k,$$

missä π_i on todennäköisyys saada rakenteellinen nolla ja $(\frac{k}{\mu+k})^k$ on todennäköisyys saada nolla-arvo NB-jakaumasta, mikä saadaan sijoittamalla $y_i = 0$ NB-jakauman pistetodennäköisyysfunktioon (2.7). Todennäköisyys, että ZINB-jakauma tuottaa nollaa suuremman arvon on

$$P(Y_i = y_i|y_i > 0) = (1 - \pi_i) \cdot \frac{\Gamma(y + k)}{\Gamma(k) \cdot \Gamma(y + 1)} \cdot (\frac{k}{\mu + k})^k \cdot ((1 - \frac{k}{\mu + k})^y),$$

missä on kerrottu todennäköisyys, että havainto ei ole rakenteellinen nolla ja todennäköisyys positiivisille arvoille NB-jakaumasta. ZINB-mallin tiheysfunktio (2.13) saadaan yhdistämällä kaksi edellä esitettyä todennäköisyyttä.

$$f(y_i) = \begin{cases} \pi_i + (1-\pi_i) \cdot \frac{k}{\mu+k}^k & , y_i = 0 \\ (1-\pi_i) \cdot \frac{\Gamma(y+k)}{\Gamma(k)\cdot\Gamma(y+1)} \cdot (\frac{k}{\mu+k})^k \cdot ((1-\frac{k}{\mu+k})^y) & , y_i > 0 \end{cases} . \tag{2.13}$$

ZINB-jakauman tiheysfunktion ylempi osa antaa todennäköisyyden ZINB-jakauman nolla-arvoille ja alempi osa antaa todennäköisyydet ZINB-jakauman positiivisille arvoille. ZINB-jakauman odotusarvo ja varianssi ovat

$$E(Y_i) = \mu_i \cdot (1-\pi_i), \tag{2.14}$$

$$\text{var}(Y_i) = (1-\pi_i) \cdot (\mu_i + \frac{\mu_i^2}{k}) + \mu_i^2 \cdot (\pi_i^2 + \pi_i). \tag{2.15}$$

2.3 Menetelmiä mallin valitsemiseksi

Yksi tärkeä osa regressioanalyysiä on valita aineistoon parhaiten sopiva malli ja tutkia, kuinka hyvin sovitettu malli kuvaa havaittua aineistoa.

Ensimmäisenä sopivan mallin valintaa tulee pohtia pohjautuen tietoon aineiston käyttäytymisestä sekä tutkittavasta ilmiöstä. Jos aineistossa ei näytä olevan erityisen suurta edustusta nolla-arvoille eikä siinä ole ylihajontaa, voidaan mallinnukseen soveltaa Poissonin jakaumaa. Jos ylihajontaa on vain vähän, voidaan soveltaa Quasi-Poisson-menetelmää, mutta suuremman ylihajonnan tilanteessa tulee käyttää NB-jakaumaa. Jos aineistossa on runsaasti nollia, on pohdittava, tulisiko siihen soveltaa ZI- tai ZA-mallia. ZI-mallia käytetään, jos nollien yliedustuksen voidaan katsoa johtuvan rakenteellisista nollista. ZA-mallia sovelletaan, jos positiivisten arvojen voidaan ajatella muodostuvan vasta, kun ilmiö on ylittänyt jonkin kynnyksen. ZI- ja ZA- mallien kohdalla on vielä pohdittava, mitä mallia käytetään positiivisten arvojen ja ZI-malleissa myös todellisten nollien mallintamiseen. Ensisijaisesti halutaan käyttää yksinkertaisempaa Poissonin jakaumaa hyödyntävää ZIP- tai ZAP-mallia, mutta jos malliin jää ylihajonta ZIP- tai ZAP-mallista huolimatta, tarvitaan mallintamiseen NB-jakaumaa, ja on hyödynnettävä ZINB- tai ZANB-mallia. (Zuur et al. 2009.)

Yksinkertainen tapa vertailla mallien hyvyyttä on hyödyntää informaatiokriteerejä, joista perinteisimmin käytettyjä ovat Akaiken informaatiokriteeri (AIC) ja Bayesianin informaatiokriteeri (BIC). Malli, joka saa pienimmän AIC- tai BIC-arvon, valitaan parhaiten aineistoon sopivaksi malliksi (Zuur et al. 2009). AIC- ja BIC-arvot saadaan kaavoista

$$AIC = -2\ell + 2\mathrm{p} \tag{2.16}$$

$$BIC = -2\ell + 2\mathrm{p} \cdot \log \mathrm{n}, \tag{2.17}$$

missä p on mallin parametrien lukumäärä, ℓ on logaritmi suurimman uskottavuuden estimaatista ja n on havaintojen lukumäärä (Posada & Buckley 2004). Informaatiokriteerit pyrkivät valitsemaan parhaaksi malliksi mahdollisimman yksinkertaisen mallin, joka selittää vastemuuttujan arvoja mahdollisimman hyvin. Yksittäiset informaatiokriteerien arvot eivät ole tulkittavissa, sillä niihin vaikuttaa muun muassa otoskoko. Informaatiokriteereillä voidaan verrata vain mallien hyvyyttä keskenään, mutta ne eivät kerro, sopiiko malli tutkittuun aineistoon. Jos kaikki tutkitut mallit ovat huonoja, pienin informaatiokriteeri kertoo vain vähiten huonon mallin. (Beaujean & Grant 2016.)

Mallin hyvyyden tarkastelemiseksi voidaan verrata esimerkiksi aineiston alkuperäisiä havaintoja ja sovitettuja arvoja, joiden plottauksesta toivotaan muodostuvan suora viiva. Yksi vaihtoehto on myös verrata Pearsonin residuaaleja vastemuuttujan ennustettuihin arvoihin sekä selittävien muuttujien arvoihin. Tässä vertailussa toivotaan, että plotatut residuaalit eivät muodosta mitään kuviota, vaan näyttävät satunnaisilta. (Zuur et al. 2009.) Pearsonin residuaalit saadaan kaavasta

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\mathrm{var}(Y_i)}}. \tag{2.18}$$

Kaavassa (2.18) niin kutsutut raa'at residuaalit y_i - $\hat{\mu}_i$ on jaettu varianssin neliöjuurella. Mitä suuremman Pearsonin residuaalin havainto saa, sitä huonommin alkuperäinen aineisto ja sovitettu malli kohtaavat kyseisen havainnon kohdalla. (Dunteman & Ho 2006.)

3 Esimerkkitutkimus sairauspoissaoloista

Havainnoidaan mallin valintaa ja tulkintaa tutkimalla aineistoa erään yrityksen työntekijöiden sairauspoissaolokertojen lukumääristä. Aineiston vastemuuttuja mittaa lukumääriä ja siinä on runsaasti nolla-arvoja. Mallin valinta aloitetaan Poissonin jakaumasta. Seuraavaksi tutkitaan aineiston ylihajontaa ja tarvetta NB-jakaumalle sekä nollien yliedustusta ja tarvetta ZIP- tai ZINB-mallille. ZA-mallit jätetään tässä tutkimuksessa tarkastelun ulkopuolelle. Lopuksi tulkitaan parhaaksi valitun mallin tuottamat tulokset.

Esimerkkitutkimus toteutetaan käyttäen R-ohjelmistoa. Tärkein tässä työssä käytetty R-ohjelmiston kirjasto on pscl-kirjasto, jolla pystytään luomaan zero-inflated-malleja. Mallin valinta tehdään perustuen Akaiken informaatiokriteereihin (AIC) ja tietoon aineiston ylihajonnasta ja mahdollisesta nollien yliedustuksesta. Kaikkien mallien R-tulosteet on esitetty liitteissä.

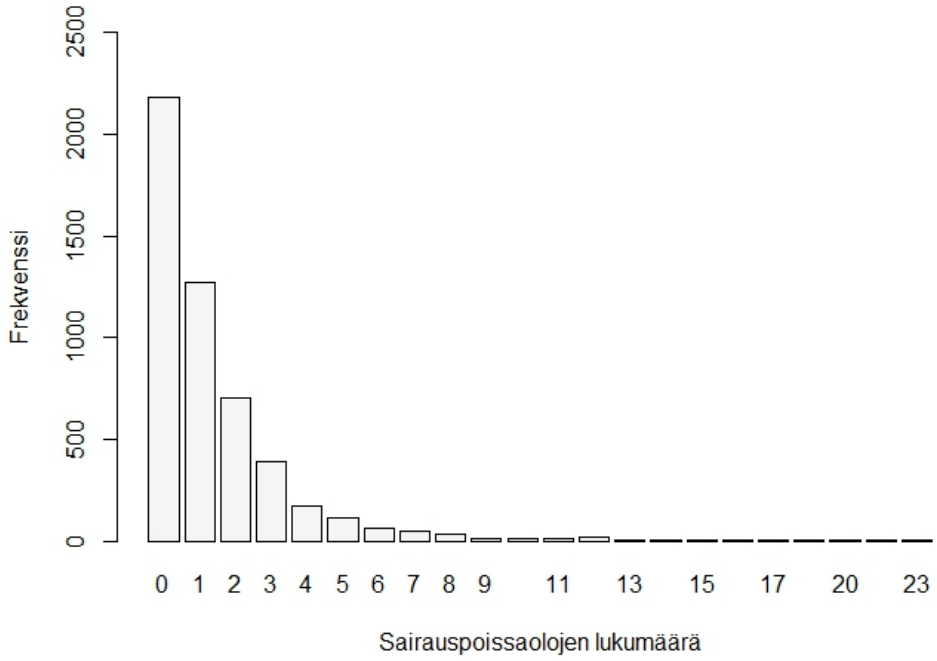
3.1 Aineisto

Aineistossa on 5049 tilastoyksikköä, jotka vastaavat yksittäisiä työntekijöitä erässä yrityksessä. Aineistossa on tieto kunkin työntekijän sairauspoissaolojen lukumääristä vuonna 2014. Sairauspoissaolokertojen lukumäärä valitaan tutkimuksen vastemuuttujaksi. Muuttujat, joilla vastetta pyritään selittämään ovat työntekijän ikä, sukupuoli ja tieto siitä, onko työntekijä ollut kuntoutuksessa. Kuvassa 1 on frekvenssijakauma sairauspoissaolokertojen lukumääristä. Kuvaajasta nähdään, että aineiston vastemuuttujassa on runsaasti nollia suhteessa muihin arvoihin, joten mallinnuksessa saatetaan tarvita ZI-jakaumaa.

Tauluun 1 on koottu aineiston muuttujat. Jatkuvista muuttujista on kerrottu niiden keskiarvo, keskihajonta sekä minimi- ja maksimiarvo. Kategorisista muuttujista on esitetty niiden lukumäärä sekä suhteellinen osuus aineistossa. Taulussa 2 on esitetty selittävien muuttujien (ikä, sukupuoli ja kuntoutus) suhde selitettävään muuttujaan (sairauspoissaolokertojen lukumäärä). Muuttuja ikä on jaettu vertailussa kolmeen ikäryhmään.

Yrityksen työntekijät olivat olleet vuonna 2014 keskimäärin sairauslomalla 1.4 kertaa keskihajonnalla 2.1. Sairauspoissaolokerrat vaihtelevat aineistossa välillä 0 - 23. Havaituista lukumääristä 43 prosenttia on nollia. Naisia aineistossa on 1680 ja miehiä 3369. Miehet olivat olleet keskimäärin sairauslomalla 1.7 kertaa ja naiset 1.2 kertaa. 272 työntekijää oli ollut kuntoutuksessa ja 4777 työntekijää eivät olleet käyneet kuntoutuksessa. Kuntoutuksessa olleet olivat sairauslomalla keskimäärin 1.9 kertaa ja muut 1.4 kertaa. Työntekijöiden keski-ikä aineistossa on 46 vuotta yhdeksän vuoden keskihajonnalla. Työntekijöiden iät vaihtelevat välillä 19 - 67 vuotta. Iän ja sairauspoissaolokertojen lukumäärän korrelaatio on negatiivinen,

mutta se on todella pieni (-0.05).



Kuva 1. Frekvenssijakauma sairauspoissaolojen lukumääristä.

Taulu 1. Aineiston muuttujat.

Jatkuvat			
Muuttuja	Keskiarvo	Keskihajonta	Min – Max
sairauspoissaolojen lukumäärä	1.4	2.1	0 – 23
ikä	46.2	9.0	19 – 67
Kategoriset			
Muuttuja	Kategoria	Lukumäärä	Suhteellinen osuus
sukupuoli	nainen	1680	0.33
	mies	3369	0.67
kuntoutus	ei kuntoutuksessa	4777	0.94
	kuntoutuksessa	272	0.05

Taulu 2. Selittävien muuttujien suhde sairauspoissaolokertojen lukumäärään.

Muuttuja	Kategoria	Keskiarvo	Keskihajonta	Korrelaatio
ikä				-0.05
	<30	1.7	1.7	
	30 - 49	1.4	2.0	
	>49	1.3	2.1	
sukupuoli	nainen	1.2	1.9	
	mies	1.7	2.4	
kuntoutus	ei kuntoutuksessa	1.4	2.0	
	kuntoutuksessa	1.9	2.7	

3.2 Mallien vertailu

Mallinnetaan aineistoa Poisson-, NB-, ZIP- ja ZINB-malleilla. Mallien muuttujien estimaatit ja niiden merkitsevyydet on koottu Tauluun 3. Kaikki mallit mallintavat lukumääriä käyttäen log-linkkifunktiota, mutta ZIP- ja ZINB-mallit mallintavat lisäksi rakenteellisia nolla-arvoja logistisella regressiolla. Poisson- ja NB-malli antavat hyvin samanlaiset estimaatit kaikille muuttujille ja kaikki muuttujat ovat myös tilastollisesti erittäin merkitseviä. ZIP- ja ZINB-mallien lukumääriä mallintava osa on myös hyvin samankaltainen, kuin Poisson- ja NB-mallit muuttujaa ikä lukuun ottamatta. Ikä ei ole merkitsevä ZIP- eikä ZINB-mallien lukumääriä mallintavassa osassa. Toisaalta ikä on merkitsevä muuttuja ZIP- ja ZINB-mallien zero-inflation-osassa. ZIP- ja ZINB-mallien zero-inflation-osat eroavat niin, että ZIP-mallissa myös muuttuja kuntoutus on merkitsevä ja kaikki estimaatit poikkeavat toisistaan.

Mallien informaatiokriteerejä, ennustettuja arvoja ja residuaaleja tarkastelemalla voidaan tutkia, miten hyvin kukin malli sopii havaittuun aineistoon. Myös tietoa vastemuuttujan ylihajonnasta ja nolla-arvojen yliedustuksesta voidaan käyttää hyödyksi sopivaa mallia valittaessa.

Poissonin jakaumalla on oletus odotusarvon ja varianssin yhtä suuruudesta. Sairauspoissaolokertojen odotusarvo on 1.4 ja varianssi on 4.3. Aineistossa on siis ylihajontaa. Ylihajontaa voidaan tutkia kaavalla (2.5)

$$\hat{\phi} = \frac{D}{n-p} = \frac{11753}{5045} = 2.33.$$

Ylihajontaa kuvaava estimaatti on selkeästi suurempi kuin yksi, joten Poissonin jakauman oletus ei toteudu eikä se sovi mallintamaan aineistoa. Jos aineistosta kuitenkin halutaan tehdä Poisson-regressiomalli, mallin AIC on 19122.

Taulu 3. Poisson-, NB-, ZIP- ja ZINB-mallien muuttujien estimaattien vertailu.

Lukumääräarvojen mallinnus (log-linkki)				
	Poisson	NB	ZIP	ZINB
(Vakiotermi)	0.735 ***	0.752***	0.591 ***	0.471 ***
ikä	-0.012***	-0.012***	0.001	-0.005
sukupuoli	0.292 ***	0.294***	0.280 ***	0.300 ***
kuntoutus	0.329 ***	0.334***	0.220 ***	0.328 ***
Zero-inflation-malli (logit-linkki)				
			ZIP	ZINB
(Vakiotermi)			-2.326 ***	-16.222 ***
ikä			0.037 ***	0.245 ***
sukupuoli			-0.037	0.162
kuntoutus			-0.313 *	-0.037

[1]*** Tilastollisesti erittäin merkitsevä (p-arvo < 0.001).

[2]** Tilastollisesti merkitsevä (p-arvo < 0.01).

[3]* Tilastollisesti melkein merkitsevä (p-arvo <0.05).

Koska aineistossa on ylihajontaa, sovitetaan seuraavaksi NB-malli. Mallin AIC-arvo on 16243, joka on selkeästi parempi kuin Poissonin mallin vastaava arvo.

Kuvan 1 frekvenssijakaumasta nähdään, että vastemuuttujassa on runsaasti nollia. Osa nollista on mahdollisesti rakenteellisia nollia, sillä osa työntekijöistä saattaa jättää pitämättä sairauslomaa, vaikka olisikin sairastunut. Tosia nolla-arvoja taas tuottavat työntekijät, jotka eivät ole sairastuneet tutkittavana aikana ja eivät siksi ole olleet sairauslomalla. Aineistoon saattaa siis sopia myös ZI-malli. Jos aineiston ylihajonta johtuu vain rakenteellisista nollista, ZIP-malli saattaa sopia aineistoon. ZIP-mallin AIC on 17560.

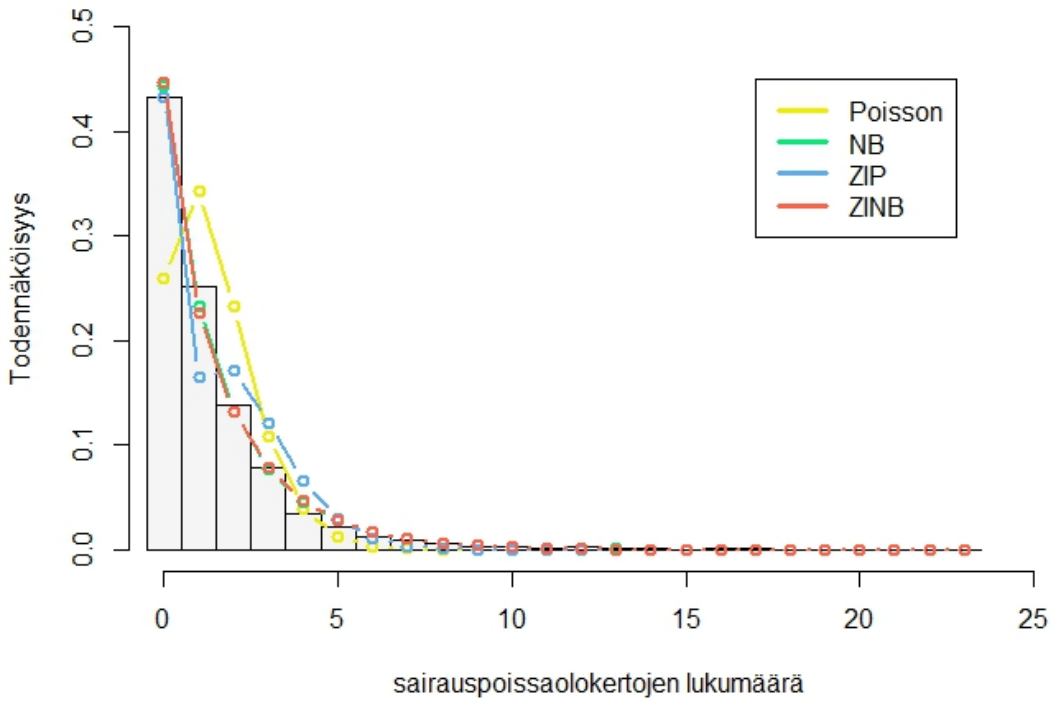
ZIP-malli sai huonomman AIC-arvon kuin NB-malli, joten todennäköisesti ylihajontaa on myös rakenteellisista nollista huolimatta. ZINB-malli huomioi sekä rakenteelliset nollat että jäljelle jääneen aineiston ylihajonnan. ZINB-mallin AIC-arvoksi saadaan 16217, joka on pienin testatuista malleista. Tutkittujen mallien AIC-arvot on vielä koottu Tauluun 4. Myös tieto aineiston ylihajonnasta ja vastemuuttujan nolla-arvojen yliedustuksesta puoltaa ZINB-mallin käyttöä.

Kuvassa 2 vertaillaan, miten hyvin kunkin tutkitun mallin ennustetut arvot vastaavat todellisia havaintoja. Tässä kuvaajassa lukumääriä tutkitaan niiden suhteellisina osuuksina. Poissonin regressiomallin ennusteet näyttävät poikkeavan eniten

Taulu 4. Akaiken informaatiokriteerit.

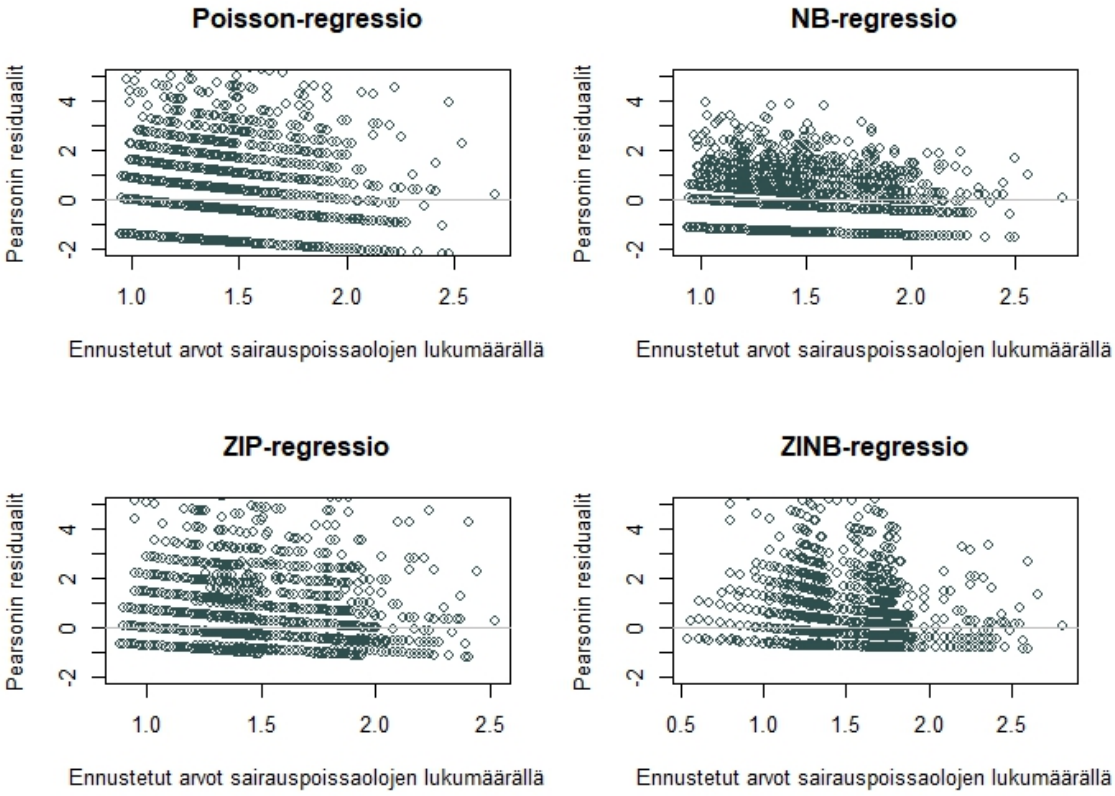
	Poisson	NB	ZIP	ZINB
AIC	19122	16243	17560	16217

todellisista havainnoista. Verrattuna todellisiin arvoihin se ennustaa selkeästi pienemmän määrän nolla-arvoja ja suuremman määrän muita pieniä arvoja. Muut mallit näyttävät ennustavan nollien osuuden lähes havaittujen nolla-arvojen osuuden mukaisesti. ZIP-mallin muut ennustetut osuudet poikkeavat jonkin verran havaituista osuuksista. NB- ja ZINB-mallien ennustetut osuudet lukumääräarvoille ovat hyvin lähellä toisiaan ja vastaavat melko hyvin todellisia havaintoja.



Kuva 2. Havaittujen arvojen todennäköisyydet verrattuna ennustettujen arvojen todennäköisyyksiin.

Kuvassa 3 on plotattu Poisson-, NB-, ZIP- ja ZINB-mallien Pearsonin residuaalit ja ennustetut havainnot. Ihanteellisessa kuvaajassa residuaalit eivät muodosta minkäänlaista kuviota, vaan näyttävät muodostuvan satunnaisesti.



Kuva 3. Ennustetut arvot sairauspoissaolojen lukumäärille verrattuna Pearsonin residuaaleihin.

NB- ja ZINB-malli näyttävät molemmat kuvaavan hyvin tutkittua aineistoa. ZINB-mallin AIC-arvo on pienin, sillä on hyvä ennustavuus sekä se pystyy käsittelemään ylihajontaan ja mahdollisia rakenteellisia nolla-arvoja. Myös yksinkertaisempi ja helpommin tulkittava NB-malli pystyy käsittelemään ylihajontaa, sen ennustavuus on hyvä ja AIC-arvo on lähes yhtä hyvät kuin ZINB-mallin. Ei siis voida yksiselitteisesti valita yhtä parasta mallia. Valitaan kuitenkin tarkempaan tarkasteluun ZINB-malli, jonka tuloksia tarkastellaan seuraavassa alaluvussa.

3.3 ZINB-mallin tulkinta

Valitsimme tarkempaan tarkasteluun ZINB-mallin perustuen AIC-arvoon ja tietoon vastemuuttujan nolla-arvojen yliedustuksesta ja ylihajonnasta. Mallin muuttujien estimaatit ja merkitsevyydet on esitetty taulussa 3. Malli koostuu kahdesta osasta. Ensimmäinen osa mallintaa NB-mallilla tosia nolla-arvoja ja positiivisia havaintoja käyttäen log-linkkifunktiota eli se mallintaa lukumääriä. Zero-inflation-osa mallintaa puolestaan rakenteellisia nolla-arvoja käyttäen logit-linkkifunktiota eli sillä on binaarinen tulos: tuottaako kyseinen muuttuja rakenteellisia nollia.

Lukumääriä mallintavan osan tulkinta

Mallin lukumääriä mallintava osa tulkitaan kuten perinteinen NB-malli tulkittaisiin. Malli estimoiduilla regressiokertoimilla on:

$$log(\hat{\mu}) = 0.471 - 0.005 \cdot \text{ikä} + 0.300 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus},$$

missä $\hat{\mu}$ on sairauspoissaolokertojen lukumäärän estimoitu odotusarvo tutkituilla regressiotermeillä. Ikä ei tämän mallin mukaan ole tilastollisesti merkitsevä tekijä selittämään sairauspoissaolokertojen lukumäärää (p-arvo > 0.05).

Mallin vakion eksponenttimuunnos antaa estimoidun odotusarvon sairauspoissaolokertojen lukumäärälle, kun muut muuttujat saavat arvon nolla, eli tilanteessa, jossa työntekijän ikä on nolla, sukupuoli on nainen eikä työntekijä ole ollut kuntoutuksessa. Tämä ei ole relevantti tulkinta, sillä mallia tuskin tarvitsee soveltaa 0-vuotiaisiin. Vakiolle saadaan tulkinta, kun muuttuja ikä keskitetään vähentämällä siitä sen keskiarvo 46. Sovite kirjoitetaan uudelleen muotoon

$$\begin{aligned} log(\hat{\mu}) =& 0.471 - 0.005 \cdot \text{ikä} + 0.300 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus} \\ =& 0.471 - 0.005 \cdot 46 - 0.005(\text{ikä} - 46) + 0.300 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus} \\ =& 0.241 - 0.005(\text{ikä} - 46) + 0.300 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus}. \end{aligned}$$

Saadun vakion 0.241 eksponenttimuunnos kertoo sairauspoissaolokertojen lukumäärän ennusteen, kun työntekijän ikä on 46 ja muut muuttujat saavat arvon nolla, eli työntekijä on nainen, joka ei ole ollut kuntoutuksessa. Ennuste on $e^{0.241}$ =1.3 sairauspoissaolokertaa.

Muuttujan sukupuoli estimaatti on positiivinen, joten sairauspoissaolokertojen lukumäärän voidaan ennustaa olevan keskimäärin suurempi miehillä kuin naisilla. Myös muuttujan kuntoutus estimaatti on positiivinen, joten voidaan ennustaa, että kuntoutuksessa ollut työntekijä on keskimäärin useammin sairauslomalla kuin työntekijä, joka ei ole ollut kuntoutuksessa. Muuttujan sukupuoli estimaatti 0.300

kertoo, että sairauspoissaolokertojen lukumäärän ennustetaan kasvavan keskimäärin $e^{0.300} = 1.35$ kertaiseksi, kun työntekijän tarkastelu vaihdetaan naisista miehiin muiden muuttujien pysyessä samana. Vastaavasti muuttujan kuntoutus estimaatti 0.328 kertoo, että sairauspoissaolokertojen lukumäärän ennustetaan kasvavan keskimäärin $e^{0.328} = 1.39$ kertaiseksi, kun työntekijän tarkastelu vaihdetaan ei kuntoutuksessa olleesta kuntoutuksessa olleeseen muiden muuttujien pysyessä samana.

Lasketaan sairauspoissaolokertojen lukumäärien odotusarvot neljälle eri ryhmälle: kuntoutuksessa olleille miehille, kuntoutuksessa olleille naisille, ei kuntoutuksessa olleille miehille ja ei kuntoutuksessa olleille naisille. Koska tilanne, jossa ikä on nolla, ei ole relevantti tulkinta, kerrotaan iän estimaatti 0.005 muuttujan ikä keskiarvolla 46. Tällöin sairauspoissaolokertojen odotusarvot saadaan 46-vuotiaalle työntekijälle.

46-vuotiaiden miesten, jotka ovat olleet kuntoutuksessa, estimoitu odotusarvo sairauspoissaolojen lukumäärälle saadaan laskemalla

$$\begin{aligned} log(\hat{\mu}) &= 0.471 - 0.005 \cdot 46 + 0.300 + 0.328 \\ \hat{\mu} &= e^{0.471 - 0.005 \cdot 46 + 0.300 + 0.328} \\ \hat{\mu} &= 2.4. \end{aligned}$$

Kun lasketaan odotusarvo saman ikäiselle naisille, joka on ollut kuntoutuksessa, muuttujan sukupuoli estimaatin kerroin asetetaan nollaksi

$$\begin{aligned} log(\hat{\mu}) &= 0.471 - 0.005 \cdot 46 + 0.328 \\ \hat{\mu} &= e^{0.471 - 0.005 \cdot 46 + 0.328} \\ \hat{\mu} &= 1.8. \end{aligned}$$

Kun lasketaan odotusarvo 46-vuotiaalle miehelle, joka ei ole ollut kuntoutuksessa, muuttujan kuntoutus estimaatin kerroin asetetaan nollaksi ja odotusarvoksi saadaan

$$\begin{aligned} log(\hat{\mu}) &= 0.471 - 0.005 \cdot 46 + 0.300 \\ \hat{\mu} &= e^{0.471 - 0.005 \cdot 46 + 0.300} \\ \hat{\mu} &= 1.7. \end{aligned}$$

Sairauspoissaolokertojen lukumäärän ennuste 46-vuotiaille naisille, jotka eivät ole olleet kuntoutuksessa saadaan asettamalla sekä muuttujan sukupuoli että kuntoutus estimaattien kertoimet nolliksi

$$\begin{aligned} log(\hat{\mu}) &= 0.471 - 0.005 \cdot 46 \\ \hat{\mu} &= e^{0.471 - 0.005 \cdot 46} \\ \hat{\mu} &= 1.3. \end{aligned}$$

Odotusarvoksi saatiin 1.3, joka saatiin aiemmin myös tulkittaessa mallin vakiotermiä.

Rakenteellisia nollia mallintavan osan tulkinta

ZINB-mallin zero-inflation-osa ennustaa logistisella regressiolla, tuottaako havainto rakenteellisia nolla-arvoja. Malli estimoiduilla regressiokertoimilla on

$$logit(\hat{\pi}) = -16.222 + 0.254 \cdot \text{ikä} + 0.162 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus},$$

missä $\hat{\pi}$ on estimoitu todennäköisyys tuottaa rakenteellisia nolla-arvoja ja muuttuja ikä on mallin ainoa merkittävästi rakenteellisten nollien syntymiseen vaikuttava muuttuja (p-arvo < 0.05). Mallin vakiolle ei tässäkään saada suoraan relevanttia tulkintaa, mutta tulkinta saadaan, kun muuttuja ikä keskitetään vähentämällä siitä sen keskiarvo 46. Sovite kirjoitetaan uudelleen muotoon

$$\begin{aligned} logit(\hat{\pi}) &= -16.222 + 0.254 \cdot \text{ikä} + 0.162 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus} \\ &= -16.222 + 0.254 \cdot 46 + 0.245(\text{ikä} - 46) + 0.162 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus} \\ &= -4.538 - 0.254(\text{ikä} - 46) + 0.162 \cdot \text{sukupuoli} + 0.328 \cdot \text{kuntoutus}. \end{aligned}$$

Vakion tulkinta saadaan nyt sen logit-muunnoksesta 46-vuotiaalle työntekijälle, kun muut muuttujat saavat arvon nolla. Vakio kertoo siis 46-vuotiaan naistyöntekijän, joka ei ole ollut kuntoutuksessa, todennäköisyyden tuottaa rakenteellinen nolla-arvo. Todennäköisyydeksi saadaan

$$\begin{aligned} logit(\hat{\pi}) &= -4.538 \\ \hat{\pi} &= \frac{e^{-4.538}}{1 + e^{-4.538}} \\ \hat{\pi} &= 0.01. \end{aligned}$$

Ainoa tilastollisesti merkitsevä muuttuja on ikä, jonka estimaatti 0.245 on etumerkiltään positiivinen. Sen voidaan siis katsoa olevan ainoa muuttuja, joka tilastollisesti merkitsevästi saattaa tuottaa rakenteellisia nollia niin, että iän lisääntyessä rakenteellisten nollien todennäköisyys lisääntyy. Koska muuttujat sukupuoli ja kuntoutus eivät ole tilastollisesti merkitseviä, niiden suhteen luotujen neljän eri ryhmän vertaaminen on turhaa, sillä ryhmien todennäköisyydet tuottaa rakenteellisia nollia ovat hyvin lähellä toisiaan. Jos kuitenkin halutaan laskea todennäköisyys, että 46-vuotias mies, joka on ollut kuntoutuksessa, tuottaa rakenteellisen nolla-arvon, se saadaan seuraavasti

$$
\begin{aligned}
logit(\hat{\pi}) &= -16.222 + 0.245 \cdot 46 + 0.162 + 0.036 \\
\hat{\pi} &= \frac{e^{-16.222+0.245\cdot 46+0.162+0.036}}{1 + e^{-16.222+0.245\cdot 46+0.162+0.036}} \\
\hat{\pi} &= 0.01.
\end{aligned}
$$

Muiden ryhmien todennäköisyydet saadaan asettamalla haluttua ryhmää kuvaavat kertoimet muuttujille. Kun halutaan tutkia naisia, muuttujan sukupuoli estimaatin kerroin asetetaan nollaksi ja miesten kohdalla se saa arvon yksi. Kuntoutuksessa olleita tutkittaessa muuttujan kuntoutus estimaatin kerroin saa arvon yksi ja muulloin se saa arvon nolla. Muuttujan ikä estimaatin kerroin on edellä iän keskiarvo 46, mutta sen tilalle voi asettaa haluamansa arvon.

4 Yhteenveto

Tässä tutkielmassa esiteltiin rakenteellisista nolla-arvoista johtuva nollien yliedustus vastemuuttujassa sekä, miten käsitellä tätä yliedustusta käyttäen zero-inflated-malleja (ZI). Zero-inflated-Poisson-mallia (ZIP) voidaan käyttää, kun ylihajonta ja nollien yliedustus johtuvat ainoastaan rakenteellisista nollista. Zero-inflated-negatiivinen binomijakauma (ZINB) on hyödyllinen, kun rakenteellisten nollien mallintaminen ei riitä poistamaan ylihajontaa. ZI-mallit eivät ainoastaan korjaa ylihajontaa vaan auttavat ymmärtämään erilaisia riski- ja ei-riskiryhmiä. Nolla-havaintojen suuri osuus vastemuuttujassa ei aina kerro tarpeesta ZI-mallille. Toisinaan yliedustus saattaa johtua ainoastaan ylihajonnasta ilman rakenteellisia nollia. Tällöin aineistoon saattaa sopia paremmin ylihajontaa käsittelevä NB-jakauma, joka ei kuitenkaan ota huomioon rakenteellisia nollia. Jos vasteessa ei ole ylihajontaa eikä rakenteellisia nolla-arvoja, voidaan käyttää Poissonin jakaumaa.

Sopivan mallin valinta aineistoon on tärkeä vaihe, sillä eri mallit saattavat tuottaa erilaisia tulkintoja aineistosta. Poisson-, NB-, ZIP-, ja ZINB-malleja sovellettiin työntekijöiden sairauspoissaoloja kuvaavaan aineistoon. Tutkimuksessa pyrittiin valitsemaan paras malli kyseiseen aineistoon käyttäen tietoa aineiston ylihajonnasta ja mahdollisista rakenteellisista nollista, Akaiken informaatiokriteeriä (AIC) ja ennustettujen arvojen suhdetta havaittuihin arvoihin. AIC-arvojen mukaan ZINB-malli olisi paras sovite aineistoon, mutta ero ei ollut suuri NB-mallin AIC-arvoon. Kuva 2 ennustetuista ja havaituista arvoista näyttää, että ZINB- ja NB-mallien ennusteet ovat hyvin lähellä toisiaan ja vastaavat melko hyvin todellisia havaintoja. Yleinen periaate on, että parhaaksi malliksi valitaan mahdollisimman yksinkertainen malli, joka mallintaa aineistoa riittävän hyvin. Aineistolle ei voida valita yhtä selkeää parasta mallia, sillä ZINB- malli on hiukan parempi perustuen AIC-arvoon, mutta NB-malli on yksinkertaisempi ja helpommin tulkittava.

Tarkempi tarkastelu toteutettiin kuitenkin ZINB-mallille, jolla on kaksiosainen tulkinta. Malli jaetaan NB-mallilla lukumääriä mallintavaan osaan ja logistisella regressiolla rakenteellisia nollia mallintavaan osaan. Sovitetun ZINB-mallin mukaan sairauspoissaolokertojen lukumääriä selittävät muuttujat sukupuoli ja kuntoutus. Miesten ennustettiin olevan naisia keskimäärin useammin sairauslomalla ja kuntoutuksessa olleiden työntekijöiden enemmän kuin ei kuntoutuksessa olleiden. Logistisen regression mukaan rakenteellisia nollia selittää muuttuja ikä niin, että iän kasvaessa, myös todennäköisyys rakenteellisille nolla-arvoille kasvaa.

Tutkielmaa olisi voinut laajentaa koskemaan myös zero-altered-malleja (ZA), joita voidaan myös käyttää mallintamaan vastemuuttujan nolla-arvojen yliedustusta. Mielenkiintoista voisi olla esimerkiksi verrata ZI- ja ZA-malleja keskenään.

5 Lähdeluettelo

Agresti, A. (2003). *Categorical data analysis*. Volume 482. John Wiley & Sons.

Beaujean, A., A. & Grant, M., B. (2016). *Tutorial on Using Regression Models with Count Outcomes using R*. Practical Assessment, Research, and Evaluation: Volume 21 , Article 2.

Dunteman, G. H., & Ho., M. R. (2006). *Quantitative Applications in the Social Sciences: An introduction to generalized linear models*. Thousand Oaks, CA: SAGE Publications, Inc. doi: 10.4135/9781412983273.

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Martin T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Trey, A. J., and Possingham H. P. (2005). *Zero tolerance ecology: improving ecological inference by modelling the source of zero observations*. Ecology letters, 8(11):1235-1246.

Posada D., Buckley, T. R. (2004), *Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests*, Systematic Biology, Volume 53, Issue 5, Pages 793–808.

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A., and Smith, G. M. (2009). *Mixed effects models and extension in ecology with R.* Springer Science & Business Media.

A Liite: R-tulosteet

Liite 1. Regressiomalli Poissonin jakaumasta

```
>model.p = glm(sairauskrt ~ ika + factor(sukupuoli) + factor(kuntoutus), poisson)
>summary(model.p)

Call:
glm(formula = sairauskrt ~ ika + factor(sukupuoli) + factor(kuntoutus),
family = poisson)

Deviance Residuals:
Min       1Q   Median       3Q      Max
-2.2263  -1.5845  -0.4202   0.5093   8.9526
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.735388   0.062560  11.755  < 2e-16 ***
ika                -0.011815   0.001365  -8.654  < 2e-16 ***
factor(sukupuoli)   0.291795   0.024842  11.746  < 2e-16 ***
factor(kuntoutus)   0.329325   0.047648   6.912 4.79e-12 ***
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 11995  on 5048  degrees of freedom
Residual deviance: 11753  on 5045  degrees of freedom
AIC: 19122

Number of Fisher Scoring iterations: 6
```

Liite 2. Regressiomalli negatiivisesta binomijakaumasta

```
Call:
glm.nb(formula = sairauskrt ~ ika + factor(sukupuoli) + factor(kuntoutus),
    init.theta = 0.8624177457, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5320  -1.2448  -0.2688   0.3020   3.9074

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.752163   0.101863   7.384 1.54e-13 ***
ika                -0.012203   0.002203  -5.540 3.03e-08 ***
factor(sukupuoli)   0.293650   0.041053   7.153 8.49e-13 ***
factor(kuntoutus)   0.334288   0.083919   3.983 6.79e-05 ***
---
(Dispersion parameter for Negative Binomial(0.8624) family taken to be 1)

    Null deviance: 5181.5  on 5048  degrees of freedom
Residual deviance: 5089.1  on 5045  degrees of freedom
AIC: 16243

Number of Fisher Scoring iterations: 1


              Theta:  0.8624
          Std. Err.:  0.0331

 2 x log-likelihood:  -16232.6920
```

Liite 3. Regressiomalli ZIP-jakaumasta

```
f = formula(sairauskrt ~ ika + factor(sukupuoli) + factor(kuntoutus))
model.zip = zeroinfl(f, dist = "poisson", link = "logit", data = tdata)
summary(model.zip)

Call:
zeroinfl(formula = f, data = tdata, dist = "poisson", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-1.1958 -0.8572 -0.2991  0.4355 13.0030

Count model coefficients (poisson with log link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.591079   0.074425   7.942 1.99e-15 ***
ika                 0.001025   0.001628   0.630    0.529
factor(sukupuoli)   0.280234   0.028973   9.672  < 2e-16 ***
factor(kuntoutus)   0.220063   0.052556   4.187 2.82e-05 ***

Zero-inflation model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -2.326321   0.207681 -11.201   <2e-16 ***
ika                 0.037370   0.004306   8.678   <2e-16 ***
factor(sukupuoli)  -0.037459   0.077396  -0.484   0.6284
factor(kuntoutus)  -0.313133   0.151319  -2.069   0.0385 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 13
Log-likelihood: -8772 on 8 Df

> AIC(m.zip)

[1] 17557.97
```

Liite 4. Regressiomalli ZINB-jakaumasta

```
f = formula(sairauskrt ~ ika + factor(sukupuoli) + factor(kuntoutus))
model.zinb = zeroinfl(f, dist = "negbin", link = "logit", data = tdata)
summary(model.zinb)

Call:
zeroinfl(formula = f1, data = tdata, dist = "negbin", link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-0.8257 -0.7304 -0.2479  0.3747 10.5275

Count model coefficients (negbin with log link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.471089   0.117235   4.018 5.86e-05 ***
ika                -0.005157   0.002710  -1.903 0.056992 .
factor(sukupuoli)   0.300030   0.042662   7.033 2.03e-12 ***
factor(kuntoutus)   0.327652   0.091112   3.596 0.000323 ***
Log(theta)         -0.056719   0.048800  -1.162 0.245129

Zero-inflation model coefficients (binomial with logit link):
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -16.22210    4.28483  -3.786 0.000153 ***
ika                  0.24504    0.06838   3.584 0.000339 ***
factor(sukupuoli)    0.16221    0.38831   0.418 0.676135
factor(kuntoutus)   -0.03673    0.57054  -0.064 0.948673
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 0.9449
Number of iterations in BFGS optimization: 34
Log-likelihood: -8100 on 9 Df

> AIC(model.zinb)

[1] 16217.19
```