Poisson-regressiomalli

Juuso Tuominen

LuK-tutkielma
Elokuu 2025

Tarkastajat:
Jouko Katajisto

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Juuso Tuominen: Poisson-regressiomalli
LuK-tutkielma, 17 s.
Tilastotiede
Elokuu 2025

Tässä tutkielmassa käsitellään yleistettyjen lineaaristen mallien erikoistapausta Poisson-regressiota. Ensimmäisenä tutkielmassa esitetään yleistettyjen lineaaristen mallien oletuksia matriisilaskennan avulla. Tämän jälkeen esitetään Poisson-regressio erikoistapauksena eksponenttiperheen muodossa. Lopuksi näytetään mallin käyttö konkreettisen aineiston avulla sekä kokonaismallin ja yksittäisten parametrien hyvyyden testaaminen.

Asiasanat: yleistetty lineaarinen malli, Poisson-regressio, tilastollinen testaaminen.

Sisällys

1 Johdanto

Poisson-satunnaismuuttujaa käytetään usein mallintamaan määrällisesti tai ajallisesti laskettavia muuttujia. Koska Poisson-satunnaismuuttuja on laskettu tapahtumamäärä, sen pienin mahdollinen arvo on nolla, ja teoreettisesti sen suurin arvo on ääretön [1]. Mallinnettavana on pääparametri λ, joka edustaa keskimääräistä lukumäärää aika- tai tilayksikköä kohti. Mallinnukseen voidaan käyttää yhtä tai useampaa selittäjää.

Lineaarisessa pienimmän neliösumman regressiomallissa kiinnostuksen kohteena oleva parametri on keskimääräinen vaste μ_i tutkittavalle yksikölle i. Tämä μ_i mallinnetaan suorana, kun selittäjiä on yksi. Intuitiivisesti Poisson-parametrin λ_i mallinnuus lineaarisena funktiona selittävästä muuttujasta voi vaikuttaa järkevältä, mutta tämä lähestymistapa tuottaa ongelmia. Itse asiassa malli

$$\lambda_i = \beta_0 + \beta_1 x_i \tag{1}$$

ei toimi hyvin Poisson-jakautuneelle aineistolle. Suora voi tuottaa negatiivisia arvoja tietyille x_i:n arvoille, mutta λ_i voi saada vain arvoja väliltä $[0, \infty)$. Lisäksi lineaarisen regression yhtälössä oletettu varianssin vakioisuus rikkoutuu, koska Poisson-muuttujan odotusarvon kasvaessa myös sen varianssi kasvaa.

Yksi tapa välttää edellä mainitut ongelmat on mallintaa $\log(\lambda_i)$ muuttujana, eikä suoraan λ_i:ta. Logaritmi $\log(\lambda_i)$ voi saada arvoja välillä $(-\infty, \infty)$, joten tämä ratkaisee negatiivisten arvojen aiheuttaman ongelman. Lisäksi tämä lähestymistapa huomioi varianssin kasvun odotusarvon kasvaessa. Näin ollen tarkastellaan yleistettyjen lineaaristen mallien erikoistapauksena Poisson-regressiomallia

$$\log(\lambda_i) = \beta_0 + \beta_1 x_i, \tag{2}$$

missä havaittu selitettävän muuttujan arvo Y_i on Poisson-jakautunut parametriin λ_i nähden, kun on annettuna selittävä muuttuja x_i.

2 Poisson-jakauma

Esitetään aluksi Poisson-jakauman määritelmä. Tilastolliset jakaumat voidaan jakaa diskreetteihin ja jatkuviin jakaumiin. Jatkuvista jakaumista voidaan muodostaa tiheysfunktio, kun taas diskreeteistä pistetodennäköisyysfunktio. Tutkitaan Poisson-jakaumaa.

Määritelmä 1. Poisson-jakauman pistetodennäköisyysfunktio on

$$P(Y_i = y_i \mid \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \quad y_i = 0, 1, 2, \ldots \tag{3}$$

Jokaisella diskreetillä tilastollisella mallilla on pistetodennäköisyysfunktiosta johdettu uskottavuusfunktio.

Määritelmä 2. Poisson-jakauman uskottavuusfunktio on

$$L(\mu_1, \mu_2, \ldots, \mu_n \mid y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}. \tag{4}$$

Uskottavuusfunktio on usein laskennallisesti epäkäytännöllinen sen haastavuuden takia. Siksi se esitetään yleensä logaritmoidussa muodossa.

Määritelmä 3. Poisson-jakauman logaritminen uskottavuusfunktio on

$$\log L(\mu_1, \mu_2, \ldots, \mu_n \mid y_1, y_2, \ldots, y_n) = \sum_{i=1}^{n} \left(y_i \log(\mu_i) - \mu_i - \log(y_i!)\right). \tag{5}$$

Poisson-regressio on yleistettyjen lineaaristen mallien erikoistapaus. Tarkastellaan seuraavaksi yleistettyjen lineaaristen mallien käsitteistöä sekä mallin rakennetta.

3 Yleistetty lineaarinen malli

Tässä luvussa esitetään yleistetyn lineaarisen mallin rakenne ja johdetaan siihen liittyviä tärkeimpiä keskeisiä tuloksia.

3.1 Mallin yleinen rakenne

Yleistetty lineaarinen malli esitetään yleensä yhtälönä, jossa selitettävää muuttujaa y yritetään mallintaa selittävien muuttujien x_1, x_2... lineaarisena yhtälönä. Yleensä malli ilmaistaan muodossa

$$y_i = \beta_0 + \beta_1 x_1 i + \beta_2 x_2 i + ... + \beta_p x_p i + e_i, \quad i = 1, 2, ..., n, \tag{6}$$

jossa β_0 on vakiotermi, β_i on selitettävän muuttujan regressiokerroin ja x_i on selitettävä muuttuja [2].

Yleistetyn lineaarisen mallin muodostamisen ja käytön ymmärtämiseksi tarkastellaan sen perusoletuksia sekä ominaisuuksia. Käsitellään aluksi eksponenttisen hajontaperheen suhdetta yleistettyihin lineaarisiin malleihin.

3.2 Eksponenttinen hajontaperhe

Riippumattomien satunnaismuuttujien tiheysfunktioiden oletetaan olevan eksponenttisen hajontaperheen muotoa [2]. Tällöin tilastollisen mallin satunnaisen osan perusrakenne on muotoa

$$f_{Y_i}(y_i; \theta_i, \phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi)\right), \tag{7}$$

jossa ϕ on hajontaparametri ja θ_i on tuntematon luonnollinen parametri.

3.2.1 Varianssi ja odotusarvo

Eksponenttiperheen mallikehikosta saadaan suoraan logaritmiseksi uskottavuusfunktioksi

$$l_i = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi). \tag{8}$$

Suoraan derivoimalla saadaan ensimmäiseksi derivaataksi

$$\frac{\partial l_i}{\partial \theta_i} = \frac{y_i - b'(\theta_i)}{a(\phi)} \tag{9}$$

ja toiseksi derivaataksi

$$\frac{\partial^2 l_i}{\partial \theta_i^2} = -\frac{b''(\theta_i)}{a(\phi)}. \tag{10}$$

Nyt voidaan käyttää yleisiä uskottavuusyhtälöitä koskevia tuloksia

$$E\left(\frac{\partial l}{\partial \theta}\right) = 0 \quad \text{ja} \quad -E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right], \tag{11}$$

jotka pätevät eksponenttijakauman säännöllisyysehtojen ollessa voimassa. Nyt satunnaismuuttujan Y_i odotusarvoksi saadaan

$$\mu_i = \mathrm{E}(Y_i) = b'(\theta_i). \tag{12}$$

Varianssi on tällöin

$$\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi). \tag{13}$$

3.3 Linkkifunktio

Linkkifunktio tarvitaan käsiteltäessä yleistettyjä lineaarisia malleja yhdistämään mallin satunnaisosaa ja lineaarista prediktoria η_i. Linkkifunktio $g(\cdot)$ on tällöin muotoa

$$g(\mu_i) = \eta_i = \sum_{j=1}^{p} \beta_j x_{ij}. \tag{14}$$

Eksponenttiperheen jakaumille linkkifunktiona toimivat identiteettilinkki (normaalijakauma), logit-linkki (binomijakauma) ja log-linkki (Poisson-jakauma). Tarkastellaan log-linkkifunktiota sen soveltuvuuden vuoksi. Log-linkkifunktio määritellään

$$g(\mu_i) = \log(\mu_i). \tag{15}$$

Poisson-regressiota sekä siihen liittyvän linkkifunktion ominaisuuksia käsitellään tarkemmin luvussa 4.

3.4 Uskottavuusyhtälöt ja SU-estimointi

Yleistetyn lineaarisen mallin tapauksessa logaritminen uskottavuusfunktio määriteltiin luvussa 3.2.1. Tämä uskottavuusfunktio voidaan kirjoittaa n-havainnon tapauksessa summana muodossa

$$l(\beta) = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + \sum_{i=1}^{n} c(y_i, \phi). \tag{16}$$

Suurimman uskottavuuden estimaattorin muodostamiseen käytetään yleistettyjen lineaaristen mallien tapauksessa iteratiivisia numeerisia menetelmiä. Näistä eniten käytetty on Newtonin–Raphsonin menetelmä. Menetelmää käydään tarkemmin Poisson-regression yhteydessä luvussa 4.3.

4 Poisson-regressio

Poisson-regressiomalli on yksi yleistettyjen lineaaristen mallien erikoistapauksista. Sitä käytetään mallintamaan lukumäärävasteita. Poisson-jakaumassa käytettävä todennäköisyysmassa on ei-negatiivisten kokonaislukujen joukko eli väliltä $[0, \infty)$, ja se kuuluu eksponenttiperheeseen.

Poisson-regression avulla voi analysoida sekä lukumäärällisiä että suhteellisia (rate) aineistoja. Tarkoituksena on mallintaa, mitkä selittävät muuttujat $x_i, i = 1, ..., n$ vaikuttavat tiettyyn vastemuuttujaan. Satunnaismuuttuja on mallin muuttuja Y, joka voi olla tarkka lukumäärä tai suhteellinen arvo. Esimerkkinä lukumäärän ja vasteen erosta voidaan sanoa, että ”minut ohittaa keskimäärin kuusi autoa päivän aikana” tai että ”minut ohittaa keskimäärin noin 0,25 autoa tunnissa”.

4.1 Poisson-jakauman käyttö

Aloitetaan tarkastelemalla Poisson-jakauman käyttöä yleistetyn lineaarisen mallin mallikehikossa. Poisson-regression logaritminen uskottavuusfunktio voidaan esittää yleistetyn lineaarisen mallin eksponenttisen hajontaperheen muodossa. Kuten luvussa 3.2.1 määriteltiin, eksponenttiperheen tapauksessa on mielekästä käyttää logaritmista uskottavuusfunktiota osana uskottavuuspäättelyä.

Määritelmä 4. Poisson-regression logaritminen uskottavuusfunktion muoto määritellään seuraavasti:

$$l(\beta; y) = \sum_{i=1}^{n}(-\mu_i + y_i\log(\mu_i) - \log(y_i!)) = \sum_{i=1}^{n}(-\exp(\mathbf{x}_i^{\cdot}\beta) + y_i(\mathbf{x}_i^{\cdot}\beta) - \log(y_i!)). \tag{17}$$

Todistus. Poisson-jakaumaoletuksen perusteella saadaan mallioletus

$$\mu_i = \mathrm{E}(Y_i) = \mathrm{var}(Y_i) = \exp(x_i^{\cdot}\beta). \tag{18}$$

Nyt käyttämällä Poisson-jakauman pistetodennäköisyysfunktiota saadaan

$$l_i(y_i;\beta) = \log((\exp(-\mu_i)\mu_i^{y_i})/y_i!) = -\mu_i + y_i\log(\mu_i) - \log(y_i!). \tag{19}$$

Summan logaritminen uskottavuusfunktio on tällöin

$$l(\beta;y) = \sum_{i=1}^{n} l_i(\beta;y) = \sum_{i=1}^{n}(-\mu_i + y_i\log(\mu_i) - \log(y_i!)), \tag{20}$$

ja kun $\mu_i = \exp(\mathbf{x}_i^{\cdot}\beta)$, saadaan logaritminen uskottavuus tiivistettyä muotoon

$$\sum_{i=1}^{n}(-\exp(\mathbf{x}_i^{\cdot}\beta) + y_i(\mathbf{x}_i^{\cdot}\beta) - \log(y_i!)). \tag{21}$$

□

Edellisen määritelmän todistuksen perusteella logaritminen uskottavuusfunktio saadaan yleistetyn lineaarisen mallin mallikehikon erikoistapauksena

$$l(\beta;y) = \sum_{i=1}^{n} \frac{(y_i\theta_i - b(\theta))}{a(\phi)} + \sum_{i=1}^{n} c(y_i;\phi) \tag{22}$$

valinnoilla

$$\theta_i = \log(\mu_i), \quad b(\theta) = \exp(\theta_i) = \mathbf{x}_i^{\cdot}\beta, \quad a(\phi) = 1 \quad \text{ja} \quad c(y_i;\phi) = -\log(y_i!). \tag{23}$$

Kuten nyt selvästi ilmenee, odotusarvoksi ja varianssiksi saadaan

$$\mathrm{E}(Y_i) = b'(\theta_i) = \exp(\theta_i) = \mu_i, \tag{24}$$

$$\mathrm{Var}(Y_i) = b''(\theta_i)a(\phi) = \exp(\theta_i) = \mu_i. \tag{25}$$

Näin ollen odotusarvo ja varianssi ovat Poisson-mallin oletuksen tavoin yhtä suuret. Poisson-regression erikoisuutena on, että aina odotusarvon kasvaessa myös varianssi kasvaa.

4.2 Logaritminen linkkifunktio

Kuten luvussa 3.3 todettiin, yleistettyjä lineaarisia malleja käytettäessä mallin satunnaisosa ja lineaarinen prediktori η_i tulee yhdistää jollain linkkifunktiolla $g(\cdot)$. Poisson-regressiomallissa käytettävä linkki on log-linkkifunktio eli

$$g(\mu_i) = \log(\mu_i). \tag{26}$$

Log-linkki varmistaa Poisson-jakaumaa käytettäessä odotusarvon μ positiivisuuden. Se yksinkertaistaa mallin estimointia ja testaamista toimiessaan mallin kanonisena linkkufunktiona.

4.3 SU-estimointi

Poisson-regressiomallin β-parametrien estimointiin käytetään Newtonin–Raphsonin menetelmää. Jälleen lähdetään liikkeelle mallin logaritmoidusta uskottavuusfunktiosta

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i \log(\mu_i) - \mu_i - \log(y_i!) \right), \tag{27}$$

jossa $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$. Sen ensimmäinen derivaatta ($g(\boldsymbol{\beta})$) voidaan kirjoittaa

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}), \tag{28}$$

jossa $\mathbf{X}$ on selittävien muuttujien matriisi, $\mathbf{y}$ on havaintojen vektori ja $\boldsymbol{\mu}$ on odotetut arvot, jossa $\mu_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$.

Toinen derivaatta on Hessin matriisi ($H(\boldsymbol{\beta})$), joka voidaan kirjoittaa muodossa

$$\mathbf{H}(\boldsymbol{\beta}) = -\mathbf{X}^\top \mathbf{W} \mathbf{X}, \tag{29}$$

jossa $\mathbf{W}$ on painomatriisi, jonka diagonaalielementit ovat μ_i.

Poisson-regression Newtonin–Raphsonin päivitys on

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{y} - \boldsymbol{\mu}), \tag{30}$$

jossa $\mathbf{y} - \boldsymbol{\mu}$ on virheiden vektori ja $(\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1}$ on Fisherin informaatiomatriisin käänteismatriisi. Näin informaatiomatriisin ja virhevektorin yhdisteenä saadaan SU-estimaatit yksikäsitteisesti ratkaistua.

5 Mallin käyttö

Tässä luvussa esitellään Poisson-regressiomallin käyttö empiirisen aineiston avulla. Tarkoituksena on havainnollistaa, miten havainnoidaan aineiston kelpoisuus Poisson-regressiolle soveltuvaksi ja miten regressiomalli muodostetaan. Analysointiin käytetään tässä tutkielmassa R-ohjelmistoa.

5.1 Käsiteltävä aineisto

Tarkasteltavana on vuoden 2015 Filippiinien perheiden taloudellisten tulojen ja menojen kysely [5]. Alkuperäisessä aineistossa on 60 muuttujaa ja noin 40000 havaintoa.

Tutkimuskysymyksenä halutaan selvittää, mikä on kotitalouden johtajan ikä, kun kotitalous on henkilömäärältään suurimmillaan?

Muodostetaan alkuperäisen aineiston pohjalta 1500 havainnon kokoinen satunnaisotos. Tutkimuskysymyksen perusteella valitaan otokseen seuraavat muuttujat:

- location eli kotitalouden sijainti Filippiineillä
- age eli kotitalouden johtajan ikä
- total eli kotitalouden henkilölukumäärä
- numLT5 eli kotitalouden alle viisi vuotiaiden henkilöiden lukumäärä

Alla tulostettuna kyseisen otosaineiston kuusi ensimmäistä riviä.

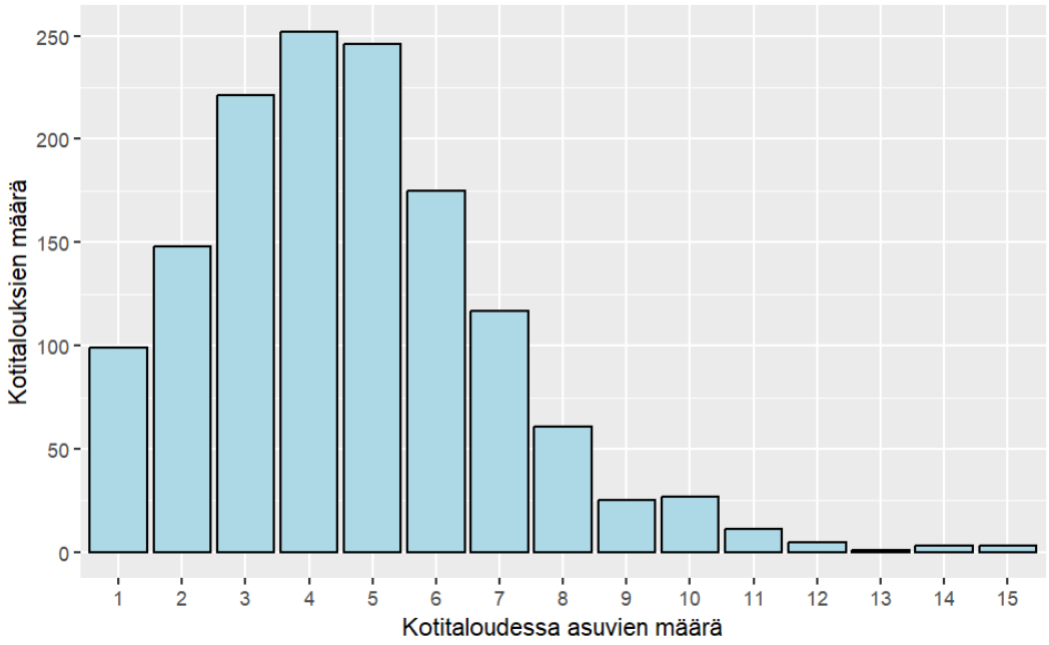
```
             location    age   total  numLT5
1     I - Ilocos Region  61     2      0
2                   NCR  33     4      0
3               Caraga  62     4      0
4               Caraga  51     7      1
5 X - Northern Mindanao  55     1      0
6      IVA - CALABARZON  38     4      1
```

Jo ensimmäisellä tarkastelulla voidaan havaita, että muuttujat ovat todennäköisesti diskreettejä ja ei-neagatiivisia eli Poisson-oletuksen mukaisia. Tutkitaan seuraavaksi tarkemmin aineiston sovelutuvuutta ja jakaumaa.

5.2 Aineiston ominaisuudet ja soveltuvuus

Poisson-regressiolle pätee Robackin [1] mukaan neljä oletusta:

1. Vastemuuttuja on lukumäärä aika- tai tilayksikköä kohti, joka kuvataan Poisson-jakaumalla.
2. Riippumattomuusoletus eli havaintojen on oltava toisistaan riippumattomia.



Kuva 1: Kotitaloudessa asuvien määrä suhteessa kotitalouksien määrään

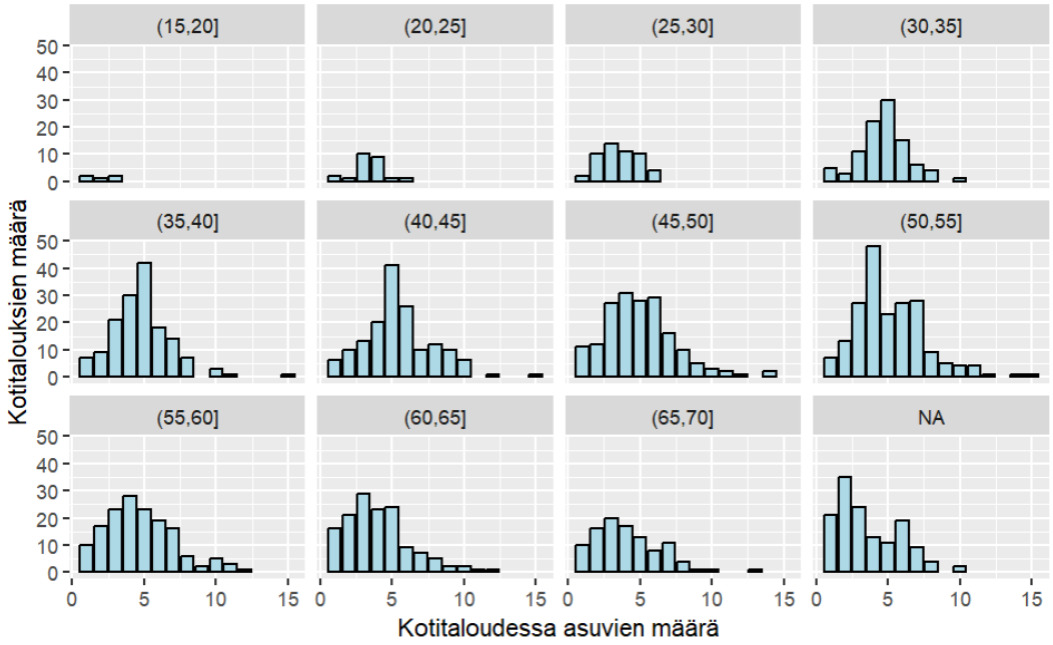
3. Määritelmän mukaan Poissonin satunnaismuuttujan keskiarvon on oltava yhtä suuri kuin sen varianssi.

4. Lineaarisuusoletus eli keskimääräisen nopeuden $\log(\lambda)$ on oltava x:n lineaarinen funktio.

Ensimmäisenä oletuksena Poisson-mallille tarkastellaan muuttujien diskreettisyyttä ja ei-negatiivisuutta. Kotitalouden johtajan ikä ja kotitaloudessa asuvien henkiöiden määrä ovat positiivisia kokonaislukuja. Ensimmäinen oletus Poisson-jakauma käytettävyydestä on oikea. Aineiston perusteella yhdessä kotitaloudessa asuu keskimäärin 4,11 ihmistä ja varianssi on noin 5,28. Yhdessä kotitaloudessa on ihmisiä yhden ja 15 henkilön väliltä.

Kuten kuvaajasta (kuva 1) nähdään, kotitalouksissa asuvien määrä on jakaumaltaan oikealle vino, ja jakaumalla on hyvin pitkälle oikealle jatkuva häntä. Tämä on hyvin yleistä Poisson-jakautuneelle aineistolle. Selvästi kotitaloudessa asuvien ihmisten lukumäärä ei noudata normaalijakaumaa. Myös syvemmän analyysin tarjoavan kuvaajan (Kuva 2) perusteella selittävät muuttujat voi mallintaa tarpeeksi hyvin Poisson-jakauman avulla. Täten oletus selittävien muuttujien Poisson-jakaumasta (oletus 1) on tyydyttävä, joten jatketaan aineiston analysointia Poisson-jakaumaa ajatellen.

Tutkitaan seuraavaksi oletusta selitettävän muuttujan varianssin eli keskihajonnan neliön ja odotusarvon yhtäsuuruudesta (oletus 3). Odotusarvon noustessa siis myös varianssin tulisi nousta. Tutkitaan arvoja aineiston kotitalouksien johtajien eri ikäryhmissä. Ikäryhmät on jaettu viiden vuoden välein. Nuorimpana ikäryhmänä on $(15, 20]$ vuotitaat ja vanhimpana ikäryhmänä on $(65, 70]$ vuotiaat. Kotitalouksien johtajat, jotka ovat iältään välin $(15, 70]$ ulkopuolella, eivät kuulu tarkastelussa



Kuva 2: Kotitaloudessa asuvien määrä kotitalouden johtajien ikäryhmittäin suhteessa kotitalouksien määrään

mihinkään ikäryhmään.

	AgeGroup	Mean	Variance	n
1	(15,20]	2.000000	1.000000	5
2	(20,25]	3.375000	1.201087	24
3	(25,30]	3.568627	1.770196	51
4	(30,35]	4.701031	2.753436	97
5	(35,40]	4.810458	4.273048	153
6	(40,45]	5.442308	5.551489	156
7	(45,50]	4.988701	5.931690	177
8	(50,55]	5.156566	5.858611	198
9	(55,60]	4.751634	5.753698	153
10	(60,65]	4.007143	4.999949	140
11	(65,70]	4.156863	5.163269	102

Taloudessa asuvien ihmisten keskiarvoa ja varianssia eri ikäryhmissä kuvaavan taulukon perusteella varianssi seuraa melko hyvin odotusarvon nousua ja laskua. Varianssi on kuitenkin joissain ikäryhmissä pienempi tai suurempi kuin kyseisen ikäryhmän odotusarvo. Tämä antaa näyttöä oletuksen rikkoutumisesta, mutta kuitenkin kyseiset rikkeet ovat suhteellisen pieniä. Varsinkin nuoremmissa ikäryhmissä otoskoko on hyvin pieni ja äärilukujen aiheuttamat heilahtelut saatavat näkyä vahvasti. Otoskoon ollessa yli 100 varianssi seuraa hyvin odotusarvon nousua ja laskua. Näiden tarkastelujen perusteella myös kolmas oletus pätee, ja Poisson-jakaumaa voidaan käyttää aineistoa analysoitaessa.

Poisson-regressiomalli tarkoittaa siis sitä, että logaritmiarvo $\log(\lambda_i)$, eikä keskimääräinen kotitalouden koko λ_i, on iän lineaarinen funktio, eli:

$$\log(\lambda_i) = \beta_0 + \beta_1 \text{age}_i. \tag{31}$$

Siksi Poisson-regression lineaarisuusoletuksen (oletus 4) tarkistamiseksi halutaan piirtää $\log(\lambda_i)$ iän funktiona. Valitettavasti λ_i on tässä tuntematon. Paras arvio λ_i:lle on kullekin ikäryhmälle havaittu kotitalouden keskimääräinen koko (eli muuttujan X tasoilla lasketut keskiarvot). Koska nämä keskiarvot on laskettu havaituista aineistotiedoista, niitä kutsutaan empiirisiksi keskiarvoiksi.

Empiirisistä keskiarvoista logaritmien ottaminen ja niiden esittäminen iän funktiona tarjoaa keinon arvioida lineaarisuusoletusta. Keskimääräistä kotitalouden logaritmoitua kokoa iän suhteen mallintavaan kuvaajaan (kuva 3) lisätty ”pehmennetty” käyrä viittaa siihen, että iän ja kotitalouden keskimääräisen koon logaritmin välillä on kaareva suhde. Tämä tarkoittaa sitä, että regressiomaalliin tulisi harkita neliöidyn termin lisäämistä.

Tämä havainto tukee hypoteesia, että on olemassa johtajan ikä, jolloin kotitalouden koko saavuttaa maksimiarvonsa. On tärkeää huomata, että nyt ei mallinneta empiiristen keskiarvojen logaritmia, vaan todellisen tapahtumanopeuden logaritmia. Empiiristen keskiarvojen tarkastelu tarjoaa kuitenkin käsityksen siitä, millainen suhde on $\log(\lambda)$:n ja x_i:n välillä.

Havaintojen riippumattomuusoletuksen (oletus 2) kannalta voi todeta, että tutkimuksen suorittanut taho on valinnut tutkimukseen satunnaiset perheet. Tästä ei ole tarkempaa tietoa, joten riippumattomuusoletuksen voi olettaa todeksi.

5.3 Poisson-regressiomallin käyttö

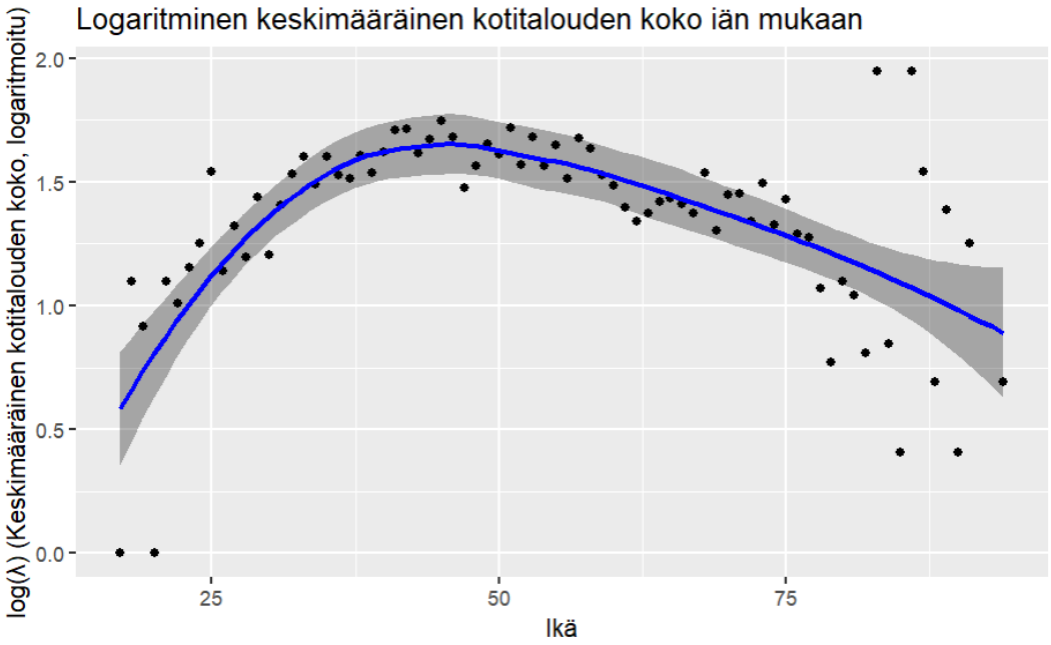
Aineiston soveltuvuuden toteamisen ja tunnuslukujen tarkastelun jälkeen voidaan muodostaa tutkimuskysymyksen perusteella haluttu regressiomalli, tässä tapauksessa Poisson-regressiomalli. Muodostetaan aineistosta muutamia erilaisia malleja. Näiden mallien käyttökelpoisuutta ja luotettavuutta vertaillaan myöhemmin luvussa 6.

5.3.1 Yhden muuttujan malli

Käytetään ensimmäisessä mallissa selittävänä muuttujana kotitalouden johtajan ikää. Haluttu malli on muotoa

$$\log(\text{total}) = \beta_0 + \beta_1 \cdot \text{age}. \tag{32}$$

R:n glm-funktioon eli R:n sisäänrakennettuun yleistetyn lineaarisen mallin työkaluun syötetään aluksi selitettävä muuttuja ja tämän jälkeen selittävät muuttujat. Lisäksi muuttujien perään kirjoitetaan haluttu jakauma eli Poisson-jakauma



Kuva 3: Lineaarisuus

sekä mahdollinen linkkifunktio. Jos linkkifunktion jättää syöttämättä, R tulkitsee Poisson-jakauman tapauksessa linkkifunktion olevan logaritminen. Summary-komennolla saadaan näkyviin ajetun glm-funktion tuloste.

```
model1 <- glm(total ~ age, family = poisson, data = House)
summary(model1)
```

```
Call:
glm(formula = total ~ age, family = poisson, data = House)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.7153218  0.0474590  36.143  < 2e-16 ***
age         -0.0036376  0.0008984  -4.049 5.15e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1588.2  on 1393  degrees of freedom
Residual deviance: 1571.7  on 1392  degrees of freedom
AIC: 6140.3

Number of Fisher Scoring iterations: 4
```

Funktion tulosteen perusteella halutun mallin kertomiksi saadaan

$$log(\hat{\lambda}) = 1,7153 - 0,0036\text{age}. \tag{33}$$

Miten regressiokertoimien estimaatteja tulkitaan Poisson-regressiomallissa? Kuten lineaarisessa pienimmän neliösumman regressiomallissa tulkittaessa kulmakertoimia halutaan nyt tarkastella, miten kotitalouden keskimääräinen henkilömäärä λ muuttuu, kun kotitalouden päämiehen ikä kasvaa vuodella. Poisson-regressiossa emme kuitenkaan tarkastele suoraan keskimääräisen henkilömäärän muutosta, vaan keskimääräisen henkilömäärän logaritmia, joka muunnetaan takaisin alkuperäisiksi yksiköiksi.

Esimerkiksi jos vertaillaan kahta mallia. Yhtä tietylle iälle ($\mathbf{X}$) ja toista sen jälkeen, kun ikää on kasvatettu yhdellä vuodella ($\mathbf{X}+1$), niin tällöin saadaan

$$\log(\lambda_{\mathbf{X}}) = \beta_0 + \beta_1\mathbf{X} \tag{34}$$

$$\log(\lambda_{\mathbf{X}+1}) = \beta_0 + \beta_1(\mathbf{X}+1). \tag{35}$$

Vähennetään yhtälöt toisistaan niin saadaan

$$\log(\lambda_{\mathbf{X}+1}) - \log(\lambda_{\mathbf{X}}) = \beta_1 \tag{36}$$

$$\log\left(\frac{\lambda_{\mathbf{X}+1}}{\lambda_{\mathbf{X}}}\right) = \beta_1 \tag{37}$$

$$\frac{\lambda_{\mathbf{X}+1}}{\lambda_{\mathbf{X}}} = e^{\beta_1}. \tag{38}$$

Tämän tuloksen perusteella ikäkertoimen eksponointi antaa halutun tuloksen. Tämän perusteella siis kotitaloudessa asuvien henkilöiden määrän keskiarvon muutos on

$$e^{-0,0036} = 0,9964. \tag{39}$$

Eli kotitalouden johtajan iän noustessa yhdellä vuodella kotitalouden henkilömäärä laskee 0,36%. Ja toisin päin, kotitalouden johtajan iän laskiessa vuodella kotitalouden henkilömäärä kasvaa noin 0,36%, sillä $1/0,9964 = 1,0036$.

5.3.2 Kahden muuttujan malli

Jotta tutkimuskysymykseen voidaan vastata, tulisi Poisson-regressiomallista muodostettavan polynomiyhtälön olla toisen asteen yhtälö. Tällöin muodostetaan aineistosta niin kutsuttu neliöity malli, jotta lineaarisuusehto täyttyy. Malliin lisätään toiseksi selittäväksi muuttujaksi ensimmäisen selittävän muuttujan neliö. Nyt siis annettuna regressiomallina on

$$\log(\text{total}) = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{age}^2. \tag{40}$$

Muodostetaan haluttu malli taas käyttäen R:n glm-funktiota. Määritellään aluksi age-muuttujan neliöity muoto aineistoon ja muodostetaan tämän jälkeen malli.

```
House <- House %>% mutate(age2 = age*age)
model2 <- glm(total ~ age + age2, family = poisson, data =
   House)
summary(model2)
```

```
Call:
glm(formula = total ~ age + age2, family = poisson, data = House)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.958e-01  1.575e-01   2.513    0.012 *
age          5.057e-02  6.187e-03   8.173 3.01e-16 ***
age2        -5.198e-04  5.876e-05  -8.847  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1588.2  on 1393  degrees of freedom
Residual deviance: 1486.7  on 1391  degrees of freedom
AIC: 6057.2

Number of Fisher Scoring iterations: 4
```

Tulosteen perusteella halutun kahden selittävän muuttujan regressiomallin kertoimiksi saadaan

$$log(\hat{\lambda}) = 0,3958 + 0,05057\text{age} - 0,0005198\text{age}^2. \tag{41}$$

Tämän regressiomallin pohjalta voidaan laskea vastaus tutkimuskysymykseen eli mikä on kotitalouden johtajan ikä kotitalouden henkilömäärän ollessa suurimmillaan? Regressiomallin toisen asteen termin kerroin on negatiivinen, jolloin regressiomallin paraabeli on alaspäin aukeava ja sen kuvaajalla on olemassa yksiselitteinen maksimikohta. Tämän globaalin maksimin selvittämiseksi lasketaan aluksi age-muuttujan suhteen ensimmäinen derivaatta

$$\frac{d}{d\text{age}} \log(\text{total}) = \beta_1 + 2\beta_2 \cdot \text{age} \tag{42}$$

ja asetetaan derivaatta nollaksi

$$0 = 0,05057 + 2 \cdot (-0,0005198) \cdot \text{age}. \tag{43}$$

Nyt voidaan ratkaista maksimikohta eli haluttu ikä

$$\text{age} = \frac{0,05057}{-2 \cdot (-0,0005198)} = 48,64. \tag{44}$$

Sijoitetaan age $= 48,64$ takaisin regressioyhtälöön

$$\log(\text{total}) = 0,3958 + 0,05057 \cdot 48,64 - 0,0005198 \cdot (48,64)^2. \tag{45}$$

Yhtälön ratkaisuksi ja maksimiarvoksi saadaan

$$\log(\text{total}) = 1,441. \tag{46}$$

Tämä tarkoittaa sitä, että talouden suurin henkilömäärä on noin $e^{1,441} = 4,225$ henkilöä silloin, kun talouden johtajan ikä on noin $48,64$ vuotta. Sama tulos on myös suoraan havaittavissa lineaarisuutta havainnollistavan kuvaajan (kuva 3) globaalista maksimista.

6 Mallien tilastollinen testaaminen

Mallin tilastollinen testaaminen on tärkeää, sillä se auttaa arvioimaan yleistetyn lineaarisen mallin pätevyyttä, selitysvoimaa ja yleistettävyyttä. Tämä prosessi varmistaa, että malli sopii riittävän hyvin dataan käytettäessä valittua luottamusastetta ja se tuottaa luotettavia tuloksia päätöksenteon tai ennusteiden tueksi. Mallia voidaan testata niin kokonaistasolla kuin myös yksittäisten parametrien osalta. Mallin estimaatteihin liittyvää epävarmuutta voidaan analysoida logaritmisen uskottavuusfunktion avulla.

6.1 Kokonaismallin testaaminen

Kokonaismallia testattaessa devianssi on yleinen ja tärkeä mittari käsiteltäessä yleistettyjä lineaarisia malleja. Sen avulla voidaan mitata, kuinka hyvin muodostettu malli selittää aineiston havaintoja suhteessa täydelliseen malliin. Poisson-jakautuneen yleistetyn lineaarisen mallin devianssin määritelmä on

$$D = 2 \sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right], \tag{47}$$

jossa y_i on havaittu arvo, $\hat{\mu}$ on mallin ennustama odotusarvo havaintopisteelle i ja n on havaintojen lukumäärä. Kun mallilla on Poisson-regression tapaan logaritminen linkkifunktio ja se sisältää vakiotermin, devianssi yksinkertaistuu muotoon

$$D = 2 \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right). \tag{48}$$

Devianssi saa arvoja väliltä $[0, \infty)$. Mitä pienempi sen arvo on, sitä paremmin malli sopii aineistoon. Poisson-regressiossa devianssi vastaa nollahypoteesin testaukseen tarvittavaa tilastollista perustaa.

Poisson-regressiomallissa devianssi on erityisen hyödyllinen, sillä Poisson-data ei noudata normaalijakaumaa ja devianssi antaa yleisesti tilastollisesti perustellun mittarin ei-normaaleille malleille. Lisäksi kuten aiemmin osoitettiin, Poisson-mallin erityispiirre on, että havaintojen varianssi riippuu odotusarvosta eli siis $\mathrm{Var}(Y_i) = \mathbb{E}(Y_i) = \mu_i$. Tämä tekee tavallisten residuaalimenetelmien eli jäännösten analysoinnin käytön haastavaksi. Devianssi tarjoaa robustin eli tilastollisesti luotettavan ja tarkan tavan arvioida sovituskykyä.

Verrataan aluksi alaluvussa 5.3.1 muodostettua yhden muuttujan mallia niin kutsuttuun nollamalliin, jossa ei ole yhtäkään selittävää muuttujaa. Käytetään R:stä löytyvää drop-in-deviance -funktiota. Siihen syöttämällä halutut yleistetyt lineaariset regressiomallit funktio laskee molempien mallien devianssit sekä niiden välisen erotuksen.

```
model0 <- glm(total ~ 1, family = poisson, data = House)
deviance_test <- anova(model0, model1, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: total ~ 1
Model 2: total ~ age
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1393     1588.2
2      1392     1571.7  1   16.444 5.011e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Kuten testin tuloksista voidaan havaita, yhden muuttujan mallin devianssi on noin $16,444$ yksikköä pienempi kuin nollamallin. Lisäksi p-arvo on selvästi merkitsevä, joten uusi muuttuja on selvästi tilastollisesti merkitsevä.

Vertaillaan seuravaaksi alaluvussa 5.3.2 muodostettua kahden muuttujan regressiomallia yhden muuttujan malliin. Toistetaan sama testi käyttäen nyt näitä malleja.

```
deviance_test2 <- anova(model1, model2, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: total ~ age
Model 2: total ~ age + age2
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      1392     1571.7
```

```
2      1391     1486.7  1   85.067 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Funktion tulosteen perusteella devianssi laskee noin 85 yksiköllä, kun malliin lisätään toinen muuttuja. Myös p-arvo on hyvin pieni, joten lisätty muuttuja on tilastollisesti merkitsevä ja se parantaa mallin sopivuutta. Tämä on odotettu tulos.

6.2 Yksittäisten parametrien testaaminen

Mallin yksittäisiä β-kertoimia voidaan testata käyttämällä Waldin testiä. Se on yksi yleisimmistä regressiomalleihin käytettävistä tilastollisista testeistä. Se perustuu mallin parametriestimaattien ja niiden keskivirheiden suhteeseen, ja sen tarkoituksena on testata nollahypoteesia. Waldin testi aloitetaan määrittämällä nollahypoteesi

$$H_0 : \beta_j = \beta_{j,0}, \tag{49}$$

johon normaalisti valitaan parametriksi $\beta_{j,0} = 0$. Testisuureena käytetään tällöin z-testisuuretta

$$z = \frac{\hat{\beta}_j - \beta_{j,0}}{\mathrm{SE}(\hat{\beta}_j)}, \tag{50}$$

jossa $\hat{\beta}_j$ on estimaatti regressiomallin parametrista β_j ja $\mathrm{SE}(\hat{\beta}_j)$ on estimaatin keskivirhe. Jos H_o oletetaan todeksi, testisuure z noudattaa standardoitua normaalijakaumaa $N(0, 1)$.

Waldin testin avulla arvioidaan, onko valittu selittävä muuttuja tilastollisesti merkitsevä. Nollahypoteesi H_0 hylätään, jos $|z|$ on suurempi kuin kriittinen arvo, joka saadaan normaalijakauman kvantiileista. Tätä tilastollista merkitsevyyttä kuvastava p-arvo saadaan laskemalla testisuureen z todennäköisyys standardin normaalijakauman avulla

$$p = 2 \cdot (1 - \Phi(|z|)), \tag{51}$$

missä Φ on standardin normaalijakauman kertymäfunktio. Yleensä p-arvoa laskettaessa käytetään 95 %:n luottamustasoa. Testisuuretta z käytettäessä tämä tarkoittaa, että jos $|z| > 1,96$, niin nollahypoteesi H_0 hylätään.

R-ohjelmisto laskee automaattisesti Waldin testin tulokset mallin jokaiselle parametrille käytettäessä glm-funktiota eli komentoa yleistetylle lineaariselle mallille. Tutkitaan alaluvussa 5.3.2 muodostettua kahden muuttujan mallia. Mallin muuttujien z-testisuureet ovat glm-tulosteen mukaan

$$\beta_1(|z|) = 8,173 \text{ ja } \beta_2(|z|) = 8,847. \tag{52}$$

Selvästi molempien muuttujien z-testisuureiden itseisarvot ovat korkeampia kuin yleisesti tunnettu merkitsevyysraja $|z| > 1,96$. Lisäksi taulukossa on z-testisuureiden jälkeen vielä listattuna parametrien tilastollista merkitsevyyttä kuvaavat p-arvot, ja ne ovat molemmat alle 0,05. Muuttujat ovat siis selvästi tilastollisesti merkitseviä ja nollahypoteesi H_0 voidaan molempien muuttujien osalta hylätä.

6.3 Mallin sopivuuden arviointi ja ongelmat

Poisson-jakaumaa käytettäessä mallin sopivuutta arvioitaessa tutkimuksen tekijä saattaa törmätä ylidispersioon. Se on tilanne, jossa tapahtumien varianssi ylittää tapahtumien odotusarvon. Malli saattaa esimerkiksi näyttää jakaumaltaan hyvin pitkälti Poisson-jakaumaa, mutta odotusarvoa tarkasteltaessa varianssi kasvaa huomattavasti nopeammin kuin odosotusarvo. Ylidispersion tilanteessa Poisson-mallin käyttö ei olisi perusteltua, vaan malliksi soveltuisi paremmin negatiivinen binomijakauma. Tässä tutkielmassa käytettävässä aineistossa ei ylidispersiota ollut havaittavissa, eikä sen ominaisuuksiin tämän tarkemmin perehdytä.

7 Johtopäätökset

Poisson-regressiomalli on tehokas menetelmä, kun analysoidaan laskettavia muuttujia, kuten tapahtumien tai yksiköiden lukumääriä. Tässä tutkielmassa tarkasteltiin kotitalouden henkilöiden lukumäärää suhteessa talouden johtajan ikään. Poisson-regressiomalli on käyttökelpoinen silloin, kun aineisto noudattaa Poisson-jakaumaa eikä esimerkiksi normaalijakaumaa. Lisäksi odotusarvon kasvaessa myös varianssi tulisi kasvaa, ja arvojen tulisi olla yhtä suuret. Tämä myös osoitettiin matemaattisesti tutkittaessa Poisson-jakauman käyttöä osana yleistetyn lineaarisen mallin mallikehikkoa.

Tutkittavassa aineistossa Poisson-regressiomallia käytettäessä tulokset osoittivat, että ikä vaikuttaa tilastollisesti merkitsevästi perheen kokoon, mutta vaikutus ei ole lineaarinen. Sen sijaan neliöllinen termi vaikuttaa olevan merkittävä. Tämä viittaa siihen, että perheen henkilöiden lukumäärä kasvaa talouden johtajan iän kasvaessa noin 48, 64 ikävuoteen asti ja alkaa tämän jälkeen laskea. Tämä voidaan yleisesti tulkita niin, että nuoremmilla ja vanhemmilla perheenpäillä on tyypillisesti pienempi perhekoko kuin keski-ikäisillä. Syvempiin syihin tämä tutkielma ei ota kantaa.

On kuitenkin hyvä ymmärtää, että perhekoko voi riippua myös muista tekijöistä, kuten tuloista, koulutustasosta ja alueellisista eroista. Tämä on hyvä tiedostaa, vaikka näihin aiheisiin liittyvät muuttujat eivät olleetkaan tutkimuksen kohteena. Lineaarisen ja neliöllisen termin lisäksi voisi olla hyödyllistä lisätä vuorovaikutustekijöitä tai muita luokittelumuuttujia. Poisson-jakauman oletukset eivät välttämättä täyty aivan täydellisesti, ja tämä voi osaltaan vaikuttaa tulosten tarkkuuteen. Jatkossa olisi hyödyllistä vertailla Poisson-mallia muihin tilastollisiin malleihin, kuten negatiiviseen binomimalliin, ja arvioida mallin soveltuvuutta eri aineistoihin.

Viitteet

[1] P. Roback, J. Legler: *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*, Francis & Taylor Group, 2021

[2] H. Nyberg: *Lineaariset ja yleistetyt lineaariset mallit*, Turun yliopisto, 2024

[3] A. Agresti: *Foundations of Linear and Generalized Linear Models*, John Wiley & Sons, 2015

[4] G. Casella & R. L. Berger: *Statistical Inference*, 2nd Edition, Duxbury Press, 2002

[5] Philippine Statistics Authority: *Family Income and Expenditure Survey*, 2015, https://www.kaggle.com/grosvenpaul/family-income-and-expenditure