**Jarkko Isotalo**

# Johdatus yleistettyihin lineaarisiin malleihin


INFORMAATIOTIETEIDEN YKSIKKÖ
TAMPEREEN YLIOPISTO

INFORMAATIOTIETEIDEN YKSIKÖN RAPORTTEJA 8/2012

TAMPERE 2012

TAMPEREEN YLIOPISTO
INFORMAATIOTIETEIDEN YKSIKKÖ
INFORMAATIOTIETEIDEN YKSIKÖN RAPORTTEJA 8/2012
TAMMIKUU 2012

**Jarkko Isotalo**

# Johdatus yleistettyihin lineaarisiin malleihin

INFORMAATIOTIETEIDEN YKSIKKÖ
33014 TAMPEREEN YLIOPISTO

ISBN 978-951-44-8734-7

ISSN-L 1799-8158
ISSN 1799-8158

# Esipuhe

Tätä luentomonistetta on käytetty oppimateriaalina Tampereen yliopistossa yleistettyjen lineaaristen mallien kursseilla. Lähdemateriaalina on käytetty seuraavia yleistettyjen lineaaristen mallien oppikirjoja.


Agresti, A. (2007). *An Introduction to Categorial Data Analysis.* Second Edition, Wiley.

Christensen, R. (1997). *Log-Linear Models and Logistic Regression.* Springer.

Dobson, A. (2002). *An Introduction to Generalized Linear Models.* Second Edition, Chapman & Hall.

Faraway, J.J. (2005). *Extending the Linear Model with R.* Chapman & Hall.

McCullagh, P. & Nelder, J.A. (1983). *Generalized Linear Models.* Chapman & Hall.


Tampere, tammikuu 2012 Jarkko Isotalo

# Sisältö

# Luku 1

# Johdatus tilastolliseen päättelyyn ja jakaumiin

## 1.1 Suurimman uskottavuuden estimaattori

Olkoon $f_Y(y;\beta)$ satunnaismuuttuja $Y$:n tiheysfunktio, mikä riippuu tuntemattomasta parametrista $\beta$. Olkoon $y_1, y_2, \ldots, y_n$ havaittu satunnaisotos $Y$:n jakaumasta. Tuntemattoman parametrin $\beta$ arvoa voidaan estimoida suurimman uskottavuuden menetelmällä. Parametrin $\beta$ suurimman uskottavuuden estimaatti $\hat{\beta}$ on ratkaisu seuraavaan maksimointiongelmaan:

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} f_Y(y_i;\beta). \tag{1.1}$$

Usein suurimman uskottavuuden estimaatti $\hat{\beta}$ on helpompi muodostaa ratkaisuna logaritmoidun yhteistiheysfunktion maksimointina:

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \log(f_Y(y_i;\beta)). \tag{1.2}$$

Suurimman uskottavuuden menetelmässä yhteistiheysfunktiota kutsutaan uskottavuusfunktioksi

$$L(\beta, \mathbf{y}) = \prod_{i=1}^{n} f_Y(y_i;\beta) \tag{1.3}$$

ja logaritmoitua yhteistiheysfunktiota logaritmoiduksi uskottavuusfunktioksi

$$l(\beta, \mathbf{y}) = \sum_{i=1}^{n} \log(f_Y(y_i;\beta)), \tag{1.4}$$

missä $\mathbf{y}$ on satunnaisvektori

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Koska $\hat{\beta}$ on satunnaisotoksen $\mathbf{y}$ funktio, se on itsestään satunnaismuuttuja tilanteessa, missä havaittuja arvoja $y_i$ kohdellaan satunnaismuuttujina. Tällöin $\hat{\beta}$:a kutsutaan suurimman uskottavuuden estimaattoriksi. Voidaan osoittaa, että suurimman uskottavuuden estimaattori $\hat{\beta}$ noudattaa asymptoottisesti (kun $n \to \infty$) normaalijakaumaa

$$\hat{\beta} \sim N\left(\beta, \frac{1}{I(\beta)}\right) = N\left(\beta, \sigma^2_{\hat{\beta}}(\beta)\right), \tag{1.5}$$

missä $I(\beta)$ on parametriin $\beta$ liittyvä informaatioluku

$$I(\beta) = \mathrm{E}\left[\left(\frac{\partial l(\beta, \mathbf{y})}{\partial \beta}\right)^2\right] = -\mathrm{E}\left(\frac{\partial^2 l(\beta, \mathbf{y})}{\partial \beta \partial \beta}\right). \tag{1.6}$$

## 1.2 Luottamusväliestimaatti, Waldin ja Score testit

Suurimman uskottavuuden estimaattorin $\hat{\beta}$ varianssi $\mathrm{Var}(\hat{\beta}) = \sigma^2_{\hat{\beta}}(\beta)$ voi riippua tuntemattomasta parametrista $\beta$ ja siten esimerkiksi keskihajontaa $\sigma_{\hat{\beta}}(\beta) = \sqrt{\mathrm{Var}(\hat{\beta})}$ ei mahdollisesti voida tarkasti laskea. Estimaattorin $\hat{\beta}$ varianssia ja siten keskihajontaa voidaan estimoida siten, että varianssin tuntemattoman parametrin arvo korvataan suurimman uskottavuuden estimaatin $\hat{\beta}$ arvolla:

$$\widehat{\mathrm{Var}}(\hat{\beta}) = \hat{\sigma}^2_{\hat{\beta}}(\hat{\beta}). \tag{1.7}$$

Estimoidun keskihajonnan avulla $\hat{\sigma}_{\hat{\beta}}(\hat{\beta})$ avulla voidaan nyt tuntemattomalle parametrille $\beta$ muodostaa normaalijakaumaan perustuva $100(1-\alpha)\%$ asymptoottinen luottamusväliestimaatti

$$\left(\hat{\beta} - z_{\alpha/2}\hat{\sigma}_{\hat{\beta}}(\hat{\beta}), \hat{\beta} + z_{\alpha/2}\hat{\sigma}_{\hat{\beta}}(\hat{\beta})\right), \tag{1.8}$$

missä $z_{\alpha/2}$ on luku, jolle on voimassa todennäköisyys $P(Z > z_{\alpha/2}) = 1 - \alpha/2$, missä $Z$ noudattaa standardoitua normaalijakaumaa $Z \sim N(0,1)$ .

Tarkastellaan seuraavaksi hypoteeseja

$$\begin{aligned} H_0 &: \beta = \beta_0, \\ H_a &: \beta \neq \beta_0, \end{aligned} \tag{1.9}$$

missä $\beta_0$ on jokin annettu arvo. Kun $H_0$ on tosi, otossuure

$$Z = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_{\hat{\beta}}(\hat{\beta})} \tag{1.10}$$

noudattaa asymptoottisesti standardoitua normaalijakaumaa $Z \sim N(0,1)$. Tämän tyylistä testisuuretta kutsutaan Waldin testiksi, missä suurimman uskottavuuden estimaattorin keskihajontaa $\sigma_{\hat{\beta}}(\beta)$ on estimoitu suurimman uskottavuuden estimaatin avulla.

Score testissä testataan edellä olevaa hypoteesia samalla testisuureella kuin Waldin testissä paitsi että keskihajonnan $\sigma_{\hat{\beta}}(\beta)$ estimaatti korvataan keskihajonnalla, mikä olisi $H_0$ hypoteesin vallitessa voimassa:

$$Z = \frac{\hat{\beta} - \beta_0}{\sigma_{\hat{\beta}}(\beta_0)}. \tag{1.11}$$

## 1.3 Uskottavuussuhdetesti

Tarkastellaan edelleen hypoteeseja

$$\begin{aligned} H_0 &: \beta = \beta_0, \\ H_a &: \beta \neq \beta_0, \end{aligned} \tag{1.12}$$

missä $\beta_0$ on jokin annettu arvo. Suurimman uskottavuuden menetelmän mukaisesti uskottavuusfunktio $L(\beta, \mathbf{y})$ saa suurimman arvonsa suurimman uskottavuuden estimaatin arvolla $L(\hat{\beta}, \mathbf{y})$. Toisaalta uskottavuusfunktion arvo voidaan laskea myös $H_0$ hypoteesin ollessa voimassa. Tällöin uskottavuusfunktio saa arvon $L(\beta_0, \mathbf{y})$. Suhdetta

$$\begin{aligned} \Delta = -2\log\left(\frac{L(\beta_0, \mathbf{y})}{L(\hat{\beta}, \mathbf{y})}\right) &= -2\left(l(\beta_0, \mathbf{y}) - l(\hat{\beta}, \mathbf{y})\right) \\ &= 2\left(l(\hat{\beta}, \mathbf{y}) - l(\beta_0, \mathbf{y})\right) \end{aligned} \tag{1.13}$$

kutsutaan uskottavuussuhteeksi. Hypoteesin $H_0 : \beta = \beta_0$ vallitessa uskottavuussuhde $\Delta$ noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = 1$.

## 1.4 Eksponentiaalinen jakaumaperhe

Mikäli satunnaismuuttujan $Y$ jakauma riippuu vain yhdestä tuntemattomasta parametrista $\beta$, satunnaismuuttuja $Y$:n todennäköisyysjakauma kuuluu eksponentiaaliseen jakaumaperheeseen, jos $Y$:n tiheysfunktio $f_Y(y, \beta)$ voidaan kirjoittaa muodossa

$$f_Y(y; \beta) = a(\beta)b(y)e^{yQ(\beta)}, \tag{1.14}$$

missä $a, b$ ja $Q$ merkitsevät joitakin funktioita.

Yleisemmin eksponentiaaliseen jakaumaperheeseen kuuluvat jatkuvista jakaumista muun muassa normaalijakauma, gamma jakauma, käänteinen normaalijakauma ja eksponenttijakauma. Diskreeteistä jakaumista eksponentiaaliseen jakaumaperheeseen kuuluvat muun muassa Bernoullin jakauma, binomijakauma, käänteinen binomijakauma, Poissonin jakauma ja multinomijakauma.

## 1.5 Normaalijakauma

Satunnaismuuttuja $Y$ noudattaa normaalijakaumaa $Y \sim N(\mu, \sigma^2)$, jos $Y$:n tiheysfunktio on muotoa

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(y-\mu)^2}{\sigma^2}}. \tag{1.15}$$

Normaalijakauma on tilastotieteen eniten käytetty jakauma. Normaalijakaumalla on tärkeä lineaarinen ominaisuus, eli jos $Y \sim N(\mu, \sigma^2)$, niin silloin lineaarinen muunnos

$$X = aY + b \tag{1.16}$$

noudattaa normaalijakaumaa $X \sim N(a\mu + b, a^2\sigma^2)$. Normaalijakaumaa $Z \sim N(0, 1)$ kutsutaan standardoiduksi normaalijakaumaksi.

## 1.6 Bernoullin jakauma

Bernoullin koe on satunnaiskoe, jolla on täsmälleen kaksi toisensa poissulkevaa tulosvaihtoehtoa. Bernoullin kokeen tulosvaihtoehdot voidaan koodata luvuilla 0 ja 1. Satunnaismuuttuja $Y$ noudattaa Bernoullin jakaumaa $Y \sim Ber(\pi)$, kun

$$P(Y = 1) = \pi, \qquad P(Y = 0) = 1 - \pi, \tag{1.17}$$

missä $0 \leq \pi \leq 1$. Bernoullin jakaumaa noudattavan satunnaismuuttujan $Y$:n odotusarvo ja varianssi ovat

$$\mathrm{E}(Y) = \pi, \qquad \mathrm{Var}(Y) = \pi(1 - \pi). \tag{1.18}$$

## 1.7 Binomijakauma

Olkoon $X_1, X_2 \ldots, X_n$ riippumattomia Bernoullin jakaumaa noudattavia satunnaismuuttujia $X_i \sim Ber(\pi)$. Tällöin satunnaismuuttuja $Y = X_1 + X_2 + \cdots + X_n$ noudattaa binomijakaumaa parametrein $n$ ja $\pi$. Satunnaismuuttuja $Y$:n jakaumaa merkitään $Y \sim Bin(n, \pi)$ ja pistetodennäköisyysfunktio on muotoa

$$P(Y = y) = f_Y(y, \pi) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}, \quad y = 0, 1, 2, \ldots, n. \tag{1.19}$$

Binomijakaumaa noudattavan satunnaismuuttuja $X$:n odotusarvo ja varianssi ovat

$$\mathrm{E}(Y) = n\pi, \quad \mathrm{Var}(Y) = n\pi(1 - \pi). \tag{1.20}$$

Jos $Y \sim Bin(n, \pi)$, niin silloin $X = n - Y$ noudattaa $X \sim Bin(n, 1 - \pi)$. Tällöin satunnaismuuttujien $Y$ ja $X$ välillä on täydellinen riippuvuus – kun $Y$ saa suuren arvon,

$X$ saa pienen. Satunnaismuuttujien $Y$ ja $X$ yhteisjakauma noudattaa kaksiulotteista multinomijakaumaa $(Y, X) \sim Mult\,(n, (\pi, 1-\pi))$.

Binomijakauman tilanteessa suurimman uskottavuuden estimaattori $\pi$:lle on muotoa

$$\hat{\pi} = \frac{Y}{n}. \tag{1.21}$$

Suurimman uskottavuuden estimaattorin $\hat{\pi}$ odotusarvo ja varianssi ovat

$$\mathrm{E}(\hat{\pi}) = \pi, \qquad \mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{n}. \tag{1.22}$$

## 1.8 Multinomijakauma

Multinomijakauma on binomijakauman yleistys. Multinomijakauma liittyy satunnaiskokeisiin, joissa on useampia kuin kaksi toisensa poissulkevaa tulosvaihtoehtoa. Toistettaessa tällaisia moniulotteisia riippumattomia satunnaiskokeita $n$ kappaletta, saatujen tulosten frekvenssijakauma voidaan kuvata multinomijakauman avulla. Tarkastellaan tilannetta, missä satunnaiskokeella on $k$ kappaletta toisensa poissulkevaa tulosvaihtoehtoa. Merkitään tulosvaihtoehtoja luvuilla $1, 2, \ldots, k$ ja olkoon $\pi_i$ tulosvaihtoehdon $i$ todennäköisyys. Toistetaan $k$-ulotteista satunnaiskoetta $n$ kappaletta ja merkitään $Y_i$:llä tuloksen $i$ lukumäärää $n$:n kokeen sarjassa. Tällöin satunnaisvektori $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ noudattaa $k$-ulotteista multinomijakaumaa parametrein $n$ ja $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$, $\mathbf{Y} \sim Mult(n, \boldsymbol{\pi})$. Multinomijakauman pistetodennäköisyysfunktio on muotoa

$$f_{\mathbf{Y}}(y_1, y_2, \ldots, y_k; \boldsymbol{\pi}) = \binom{n}{y_1\, y_2\, \ldots\, y_k} \pi_2^{y_1} \pi_2^{y_2} \cdots \pi_k^{y_k}, \tag{1.23}$$

missä $y_1 + y_2 + \cdots + y_k = n$, $\pi_1 + \pi_2 + \cdots + \pi_k = 1$ ja $\binom{n}{y_1\, y_2 \ldots y_k} = \frac{n!}{y_1!\, y_2! \ldots y_k!}$.

Multinomijakaumalle on voimassa seuraavat ominaisuudet:

$$Y_i \sim Bin(n, \pi_i), \quad \mathrm{E}(Y_i) = n\pi_i, \quad \mathrm{Var}(Y_i) = n\pi_i(1-\pi_i), \quad \mathrm{Cov}(Y_i, Y_j) = -n\pi_i\pi_j.$$

Suurimman uskottavuuden estimaattorit ovat muotoa

$$\hat{\pi}_i = \frac{Y_i}{n}. \tag{1.24}$$

## 1.9 Poissonin jakauma

Toisinaan frekvenssidata ei synny ehdolla, että jotain toistokoetta toistetaan tietyn $n$ kertaa. Usein on tilanteita, että jonkin ajan tai tilan aikana vain havainnoidaan jonkin satunnaisilmiön toteutuminen $y$ frekvenssin kerran. Poissonin jakauma sopii hyvin tällaisten frekvenssidatojen mallintamiseen. Satunnaismuuttuja $Y$ noudattaa Poissonin jakaumaa parametrilla $\lambda > 0$, jos $Y$ :n pistetodennäköisyysfunktio on muotoa

$$P(Y = y) = f_Y(y) = \frac{e^{-\lambda}\lambda^y}{y!}, \quad y = 0, 1, 2, \ldots. \tag{1.25}$$

Jos $Y \sim Poi(\lambda)$, niin silloin

$$\mathrm{E}(Y) = \lambda, \qquad \mathrm{Var}(Y) = \lambda. \tag{1.26}$$

# Luku 2

# Ristiintaulukot

## 2.1 Ristiintaulukoiden merkinnät

Olkoon $X$ ja $Y$ satunnaismuuttujia joilla kummallakin on kaksi toisensa poissulkevaa tulosvaihtoehtoa. Jos tulosvaihtoehtoja merkitään 0:lla ja 1:llä, niin satunnaismuuttujien yhteistodennäköisyysjakauma voidaan esittää $2 \times 2$-ristiintaulukon avulla:

| $P(X = x_i, Y = y_j)$ : | | $y_j=$ 1 | 0 | Yhteensä |
|---|---|---|---|---|
| $x_i=$ | 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
| | 0 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

Ristiintaulukossa rivi- ja sarakesummat

$$\pi_{1+} = \pi_{11} + \pi_{12}, \quad \pi_{2+} = \pi_{21} + \pi_{22},$$
$$\pi_{+1} = \pi_{11} + \pi_{21}, \quad \pi_{+2} = \pi_{12} + \pi_{22},$$

ovat muuttujien $X$ ja $Y$ marginaalijakaumia.

Ristiintaulukolla voidaan esittää myös ehdollisten todennäköisyyksien $P(Y = y_j|X = x_i)$ jakauma. Tällöin taulukon rivit tulkitaan riippumattomiksi binomijakaumiksi parametrein $\pi_1$ ja $\pi_2$:

| $P(Y = y_j\|X = x_i)$ : | | $y_j=$ 1 | 0 | Yhteensä |
|---|---|---|---|---|
| $x_i=$ | 1 | $\pi_1$ | $1 - \pi_1$ | 1 |
| | 0 | $\pi_2$ | $1 - \pi_2$ | 1 |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

Vastaavasti jos satunnaismuuttujilla $X$ ja $Y$ on $I$ ja $J$ toisensa poissulkevaa tulosvaihtoehtoa, satunnaismuuttujien $X$ ja $Y$ yhteistodennäköisyysjakauma voidaan esittää $I \times J$-ristiintaulukon avulla:

| $P(X = x_i, Y = y_j)$ : | | $y_j$= 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i$= | 1 | $\pi_{11}$ | $\pi_{12}$ | ... | $\pi_{1J}$ | $\pi_{1+}$ |
| | 2 | $\pi_{21}$ | $\pi_{22}$ | ... | $\pi_{2J}$ | $\pi_{2+}$ |
| | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| | I | $\pi_{I1}$ | $\pi_{I2}$ | ... | $\pi_{IJ}$ | $\pi_{I+}$ |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | ... | $\pi_{+J}$ | 1 |

Ehdollisten todennäköisyyksien $P(Y = y_j|X = x_i)$ jakauma yleisemmässä tilanteessa on muotoa:

| $P(Y = y_j\|X = x_i)$ : | | $y_j$= 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i$= | 1 | $\pi_{11}$ | $\pi_{12}$ | ... | $\pi_{1J}$ | 1 |
| | 2 | $\pi_{21}$ | $\pi_{22}$ | ... | $\pi_{2J}$ | 1 |
| | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| | I | $\pi_{I1}$ | $\pi_{I2}$ | ... | $\pi_{IJ}$ | 1 |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | ... | $\pi_{+J}$ | 1 |

Toistettaessa satunnaismuuttujien $X$ ja $Y$ muodostamaa 2-ulotteista satunnaiskoetta $n_{++}$ kertaa, voidaan tulosvaihtoehtojen $(X = x_i, Y = y_j)$ frekvenssijakauma kuvata ristiintaulukolla:

| $Freq(X = x_i, Y = y_j)$ : | | $y_j$= 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i$= | 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | ... | $n_{2J}$ | $n_{2+}$ |
| | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| | I | $n_{I1}$ | $n_{I2}$ | ... | $n_{IJ}$ | $n_{I+}$ |
| | Yhteensä | $n_{+1}$ | $n_{+2}$ | ... | $n_{+J}$ | $n_{++}$ |

## 2.2 Päättelyasetelmat $2 \times 2$-ristiintaulukossa

Ristiintaulukoissa havaittujen frekvenssien $n_{ij}$ avulla tehdään päättelyitä tuntemattomista todennäköisyyksistä $\pi_{ij}$. Riippuen koeasetelmasta ja päättelyn tavoitteista voidaan erotella seuraavia päättelyasetelmia:

- Jos $X$ on selittävä muuttuja ja $Y$ selitettävä muuttuja, silloin ollaan yleensä kiinnostuneita ehdollisten todennäköisyyksien $P(Y = y_j|X = 1)$ ja $P(Y =$

$y_j|X = 0$) eroavuuksista. Tällöin $2 \times 2$-frekvenssitaulukon rivien oletetaan olevan toteutuneita arvoja riippumattomista binomijakaumista $Bin(n_{1+}, \pi_1)$ ja $Bin(n_{2+}, \pi_2)$. Mikäli $2 \times 2$-frekvenssitaulukossa rivisummat $n_{1+}$ ja $n_{2+}$ ovat ennalta kiinnitettyjä, taulukon havaittujen frekvenssien tulkitaan myös olevan toteutuneita riippumattomista binomijakaumista $Bin(n_{1+}, \pi_1)$ ja $Bin(n_{2+}, \pi_2)$.

| | 1 | 0 | Yhteensä |
|---|---|---|---|
| 1 | $\pi_1$ | $1-\pi_1$ | 1 |
| 0 | $\pi_2$ | $1-\pi_2$ | 1 |
| Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

| | 1 | 0 | Yhteensä |
|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| 0 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| Yhteensä | $n_{+1}$ | $n_{+2}$ | $n_{++}$ |

- Jos kumpikin $X$ ja $Y$ ovat selitettäviä muuttujia, silloin ollaan yleensä kiinnostuneita ovatko muuttujat $X$ ja $Y$ riippumattomia toisistaan, eli onko voimassa $P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$. Tällöin $2 \times 2$-frekvenssitaulukon havaintojen $n_{ij}$ oletetaan olevan toteutuneita arvoja joko multinomijakaumasta $Mult\,(n_{++}, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$ tai siten niin, että $n_{ij} \sim Poi(\lambda_{ij})$.

| | 1 | 0 | Yhteensä |
|---|---|---|---|
| 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
| 0 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
| Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

| | 1 | 0 | Yhteensä |
|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| 0 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| Yhteensä | $n_{+1}$ | $n_{+2}$ | $n_{++}$ |

## 2.3 Kaksi riippumatonta binomijakaumaa

Oletetaan, että $2 \times 2$-frekvenssitaulukon havainnot $n_{ij}$ ovat toteutuneita arvoja riippumattomista binomijakaumista $Bin(n_{1+}, \pi_1)$ ja $Bin(n_{2+}, \pi_2)$. Testataan hypoteesia

$$\begin{aligned} H_0 :\ P(Y = y_j|X = 1) &= P(Y = y_j|X = 0) \\ \pi_1 &= \pi_2 \\ \pi_1 - \pi_2 &= 0. \end{aligned}$$

Testattaessa hypoteesia $H_0 :\ \pi_1 - \pi_2 = 0$, Waldin testisuure

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1)/n_{1+} + \hat{\pi}_2(1-\hat{\pi}_2)/n_{2+}}} \tag{2.1}$$

noudattaa asymptoottisesti standardoitua normaalijakaumaa $Z \sim N(0, 1)$.

Toisinaan erotuksen $\pi_1 - \pi_2$ sijaan voi olla järkevä tutkia todennäköisyyksien $\pi_1$ ja $\pi_2$ suhdetta. Suhteellinen riski $\delta$ on suhde

$$\delta = \frac{\pi_1}{\pi_2}, \tag{2.2}$$

ja sen estimaatti on $\hat{\delta} = \hat{\pi}_1/\hat{\pi}_2$.

## 2.4 Vedonlyöntisuhde

Oletetaan edelleen, että $2 \times 2$-frekvenssitaulukon havainnot $n_{ij}$ ovat toteutuneita arvoja riippumattomista binomijakaumista $Bin(n_{1+}, \pi_1)$ ja $Bin(n_{2+}, \pi_2)$. Todennäköisyyksistä $\pi_1$ ja $\pi_2$ voidaan muodostaa vedonlyöntikertoimet $\gamma_1$ ja $\gamma_2$:

$$\gamma_1 = \frac{\pi_1}{1-\pi_1}, \qquad \gamma_2 = \frac{\pi_2}{1-\pi_2}. \tag{2.3}$$

Vedonlyöntikertoimien $\gamma_1$ ja $\gamma_2$ suhdetta

$$\theta = \theta_{Y|X} = \frac{\gamma_1}{\gamma_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{2.4}$$

kutsutaan vedonlyöntisuhteeksi. Kun muuttujat $X$ ja $Y$ ovat riippumattomia, eli $\pi_1 = \pi_2$, vedonlyöntisuhde saa arvon $\theta = 1$. Vedonlyöntisuhteen estimaatti on muotoa

$$\hat{\theta} = \hat{\theta}_{Y|X} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.5}$$

Vedonlyöntisuhteella on sellainen hyödyllinen ominaisuus, että vedonlyöntisuhteen estimaatin arvo pysyy samana tilanteessa, missä $Y$:llä selitetään $X$:n arvoja. Tarkastellaan ehdollisia todennäköisyyksiä $P(X = x_i|Y = y_j)$:

| $P(X = x_i\|Y = y_j)$ : | | $y_j$= 1 | 0 | Yhteensä |
|---|---|---|---|---|
| $x_i$= | 1 | $\pi_1$ | $\pi_2$ | $\pi_{1+}$ |
| | 0 | $1-\pi_1$ | $1-\pi_2$ | $\pi_{2+}$ |
| | Yhteensä | 1 | 1 | 1 |

Tällöin vedonlyöntisuhteen estimaatti saa myös arvon

$$\hat{\theta} = \hat{\theta}_{X|Y} = \frac{\hat{\pi}_1/(1-\hat{\pi}_1)}{\hat{\pi}_2/(1-\hat{\pi}_2)} = \frac{n_{11}n_{22}}{n_{12}n_{21}}. \tag{2.6}$$

Täten sama estimaatti $\hat{\theta}$ estimoi vedonlyöntisuhteita $\theta_{Y|X}$ ja $\theta_{X|Y}$. Tämä ominaisuus tekee vedonlyöntisuhteesta erityisen hyödyllisen parametrin tilanteissa, missä muodostettu frekvenssidata kuvaa toteutuneita arvoja todennäköisyysjakauman $P(X = x_i|Y = y_j)$ tapauksessa ja silti varsinainen kiinnostuksen kohde on tutkia ehdollisen jakauman $P(Y = y_j|X = x_i)$ ominaisuuksia.

Silloin kun molemmat muuttujat $X$ ja $Y$ ovat selitettäviä muuttujia, vedonlyöntisuhde voidaan määritellä suhteena

$$\theta = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}}. \tag{2.7}$$

Estimaattina toimii edelleen $\hat{\theta}$.

Koska estimaatin $\hat{\theta}$ jakauma on hyvin vino, on hyödyllistä perustaa vedonlyöntisuhteen päättely logaritmoituun vedonlyöntisuhteeseen. Kun $\theta = 1$, niin $\log(\theta) = 0$. Logaritmoitu vedonlyöntisuhteen estimaatti $\log(\hat{\theta})$ noudattaa asymptoottisesti normaalijakaumaa parametrein

$$\mathrm{E}\left(\log(\hat{\theta})\right) = \log(\theta), \qquad \sigma\left(\log(\hat{\theta})\right) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}. \tag{2.8}$$

Logaritmoidulle vedonlyöntisuhteelle $\log(\theta)$ saadaan muodostettua $100(1-\alpha)\%$ luottamusväli kaavalla

$$\log(\hat{\theta})\pi \pm z_{\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}, \tag{2.9}$$

missä $z_{\alpha/2}$ on luku, jolle voimassa $P(Z > z_{\alpha/2}) = \alpha/2$ kun $Z \sim N(0,1)$. Korottamalla eksponenttiin logaritmoidun vedonlyöntisuhteen luottamusvälin raja-arvot, saadaan muodostettua luottamusväli itse vedonlyöntisuhteelle $\theta$.

Jos jokin $n_{ij} = 0$, niin $\hat{\theta}$ on 0 tai $\infty$. Tällöin voidaan käyttää muunneltua estimaattia

$$\tilde{\theta} = \frac{(n_{11} + 0.5)(n_{22} + 0.5)}{(n_{12} + 0.5)(n_{21} + 0.5)} \tag{2.10}$$

estimoimaan vedonlyöntisuhdetta $\theta$. Logaritmoidun estimaatin $\tilde{\theta}$ keskihajonta on muotoa

$$\sigma\left(\log(\tilde{\theta})\right) = \sqrt{\frac{1}{n_{11}+0.5} + \frac{1}{n_{12}+0.5} + \frac{1}{n_{21}+0.5} + \frac{1}{n_{22}+0.5}}. \tag{2.11}$$

## 2.5 Riippumattomuustestit $2 \times 2$-ristiintaulukossa

Oletetaan, että muuttujat $X$ ja $Y$ ovat molemmat selitettäviä muuttujia, ja että $2\times2$-frekvenssitaulukon havainnot $n_{ij}$ ovat toteutuneita arvoja joko multinomijakaumasta $Mult\,(n_{++}, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))$ tai siten niin, että $n_{ij} \sim Poi(\lambda_{ij})$.

Kun kokonaisfrekvenssisumma $n_{++}$ on tiedossa, voidaan jokaiselle ristiintaulukon solulle laskea odotetut frekvenssit

$$\mu_{ij} = n_{++}\pi_{ij}. \tag{2.12}$$

| $E(n_{ij})$ : | | $y_j$= 1 | 0 | Yhteensä |
|---|---|---|---|---|
| $x_i$= | 1 | $\mu_{11}$ | $\mu_{12}$ | $\mu_{1+}$ |
| | 0 | $\mu_{21}$ | $\mu_{22}$ | $\mu_{2+}$ |
| | Yhteensä | $\mu_{+1}$ | $\mu_{+2}$ | $\mu_{++}$ |

Tarkastellaan $X$:n ja $Y$:n riippumattomuutta. Testataan hypoteesia

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{kaikille } i \text{ ja } j$$

Hypoteesin $H_0$ vallitessa odotetut frekvenssit $\mu_{ij}$ ovat muotoa $\mu_{ij} = n_{++}\pi_{i+}\pi_{+j}$. Koska $\pi_{i+}$ ja $\pi_{+j}$ ovat tuntemattomia, pitää ne estimoida ja siten myös saadaan estimoidut odotetut frekvenssit

$$\hat{\mu}_{ij} = n_{++}\frac{n_{i+}}{n_{++}} \cdot \frac{n_{+j}}{n_{++}} = \frac{n_{i+}n_{+j}}{n_{++}}. \tag{2.13}$$

$H_0$ hypoteesin voimassaoloa voidaan nyt testata Pearsonin $X^2$-testisuureella

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}. \tag{2.14}$$

Testisuure $X^2$ noudattaa $H_0$ hypoteesin vallitessa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = 1$.

Vaihtoehtoisesti $H_0$ hypoteesin voimassaoloa voidaan testata uskottavuussuhteen avulla. Multinomijakauman tilanteessa uskottavuussuhde on muotoa

$$G^2 = 2\sum_{i=1}^{2}\sum_{j=1}^{2} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right). \tag{2.15}$$

Uskottavuussuhde $G^2$ noudattaa $H_0$ hypoteesin vallitessa asymptoottisesti myös $\chi^2$-jakaumaa vapausastein $df = 1$.

## 2.6 Riippumattomuustestit $I \times J$-ristiintaulukossa

Tarkastellaan tilannetta, jossa satunnaismuuttujilla $X$ ja $Y$ on $I$ ja $J$ toisensa poissulkevaa tulosvaihtoehtoa. $I \times J$-ristiintaulukon tilanteessa vedonlyöntisuhde $\theta$ voidaan määritellä lukuna

$$\theta = \frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}}. \tag{2.16}$$

Tilanteessa, jossa $I \times J$-ristiintaulukko kuvaa ehdollisten todennäköisyyksien $P(Y = y_j|X = x_i)$ jakaumaa, hypoteesi

$$H_0 : \pi_{1j} = \pi_{2j} = \cdots = \pi_{Ij} \qquad \text{kaikille } j = 1, 2, \ldots, J$$

on voimassa jos ja vain jos

$$\frac{\pi_{ij}\pi_{i'j'}}{\pi_{ij'}\pi_{i'j}} = 1$$

kaikille $i, i' = 1, 2, \ldots, I$ ja $j, j' = 1, 2, \ldots, J$.

Pearsonin $X^2$-testisuuren ja uskottavuussuhteen $G^2$ avulla voidaan testata yleistä muuttujien $X$ ja $Y$ välistä riippumattomuutta

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{kaikille } i = 1, 2, \ldots, I,\ j = 1, 2, \ldots, J.$$

Pearsonin $X^2$-testisuure ja uskottavuussuhde $G^2$ ovat $I \times J$-ristiintaulukon tilanteessa muotoa

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right), \tag{2.17}$$

missä estimoidut odotetut frekvenssit ovat muotoa

$$\hat{\mu}_{ij} = \frac{n_{i+}n_{+j}}{n_{++}}. \tag{2.18}$$

Testisuureet $X^2$ ja $G^2$ noudattavat $H_0$ hypoteesin vallitessa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = (I-1)(J-1)$.

Pearsonin $X^2$-testisuure ja uskottavuussuhde $G^2$ testaavat siis onko muuttujien $X$ ja $Y$ välillä riippuvuutta. Standardoitujen soluresiduaalien

$$\hat{\epsilon}_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}} \tag{2.19}$$

avulla voidaan tutkia minkä suuntainen muuttujien $X$ ja $Y$ välinen riippuvuus on. $H_0$ hypoteesin ollessa voimassa, standardoitu soluresiduaali noudattaa asymptoottisesti standardoitua normaalijakaumaa $\hat{\epsilon}_{ij} \sim N(0,1)$.

## 2.7 Trenditesti

Olkoon muuttujat $X$ ja $Y$ järjestysasteikollisia. Tällöin $X$:n ja $Y$:n tulosvaihtoehdot $i = 1, 2, \ldots, I$ ja $j = 1, 2, \ldots, J$ voidaan järjestää esimerkiksi nousevaan järjestykseen. Olkoon nyt $u_1 \leq u_2 \leq \cdots \leq u_I$ muuttujan $X$ tulosvaihtoehdoille $i = 1, 2, \ldots, I$ määriteltyjä lukuarvoja, ja vastaavasti olkoon $v_1 \leq v_2 \leq \cdots \leq v_J$ muuttujan $Y$ tulosvaihtoehdoille $j = 1, 2, \ldots, J$ määriteltyjä lukuarvoja. Korrelaatiokertoimen $r$ avulla voidaan testata onko muuttujien $X$ ja $Y$ välillä lineaarista riippuvuutta. Korrelaatiokerroin lasketaan kaavalla

$$r = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(u_i - \bar{u})(v_j - \bar{v})\hat{\pi}_{ij}}{\sqrt{\left[\sum_{i=1}^{I}(u_i - \bar{u})^2\hat{\pi}_{i+}\right]\left[\sum_{j=1}^{J}(v_j - \bar{v})^2\hat{\pi}_{+j}\right]}}, \tag{2.20}$$

missä $\hat{\pi}_{ij} = \frac{n_{ij}}{n_{++}}$, $\hat{\pi}_{i+} = \frac{n_{i+}}{n_{++}}$ ja $\hat{\pi}_{+j} = \frac{n_{+j}}{n_{++}}$, sekä $\bar{u} = \sum_{i=1}^{I} u_i\hat{\pi}_{i+}$ ja $\bar{v} = \sum_{j=1}^{J} v_i\hat{\pi}_{+j}$.

Merkitään muuttujien $X$ ja $Y$ välistä populaatiokorrelaatiokerrointa $\rho$:lla. Testataan lineaarista riippumattomuutta, eli hypoteesia

$$H_0 : \rho = 0.$$

Testisuure

$$M^2 = (n_{++} - 1)r^2 \tag{2.21}$$

noudattaa $H_0$ hypoteesin ollessa voimassa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = 1$.

# Luku 3

# Lineaaristen mallien perusteita

## 3.1 Parametrien estimoinnista

Olkoon $y_1, y_2, \ldots, y_n$ satunnaisotos normaalijakaumasta $Y_i \sim N(\mu_i, \sigma^2)$. Lineaarisessa mallissa oletetaan, että odotusarvo $\mu_i$ riippuu lineaarisesti selittävistä muuttujista

$$\mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathbf{x}_i' \boldsymbol{\beta}, \tag{3.1}$$

missä

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \qquad \mathbf{x}_i = \begin{pmatrix} 1 \\ x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}.$$

Tuntemattoman parametrivektorin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaattori saadaan ratkaisemalla logaritmoituun uskottavuusfunktioon liittyvä maksimointiongelma

$$\arg\max_{\boldsymbol{\beta}} l(\boldsymbol{\beta}, \sigma^2; y_1, y_2, \ldots, y_n) = \arg\max_{\boldsymbol{\beta}} \left( \log\left( (2\pi\sigma^2)^{-n/2} \right) - \frac{1}{2\sigma^2} (\sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 \right). \tag{3.2}$$

Suurimman uskottavuuden estimaattorin $\hat{\boldsymbol{\beta}}$ avulla saadaan laskettua jokaisen havainnon $i$ odotusarvon $\mu_i$ suurimman uskottavuuden estimaatti:

$$\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \mathbf{x}_i' \hat{\boldsymbol{\beta}}. \tag{3.3}$$

Lineaarisessa mallissa odotusarvon estimaatteja $\hat{\mu}_i$ kutsutaan usein sovitearvoiksi ja niistä käytetään merkintää $\hat{\mu}_i = \hat{y}_i$.

Odotusarvon estimaattien avulla voidaan laskea jokaiselle havainnolle residuaalit

$$e_i = y_i - \hat{\mu}_i. \tag{3.4}$$

Mallin varianssiin liittyvän tuntemattoman parametrin $\sigma^2$ estimaattorina yleisesti käytetään kaavaa

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-(k+1)}. \tag{3.5}$$

Tämä itse asiassa ei ole parametrin $\sigma^2$ suurimman uskottavuuden estimaattori, vaan niin sanottu rajoitettu suurimman uskottavuuden estimaattori.

Kun merkitään

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix},$$

niin estimaattorin $\hat{\boldsymbol{\beta}}$ kovarianssimatriisi on muotoa

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.6}$$

ja täten yksittäisen estimaattorin varianssi on muotoa

$$\mathrm{Var}(\hat{\beta}_j) = \sigma^2 t^{jj}, \tag{3.7}$$

missä $t^{jj}$ on matriisin $(\mathbf{X}'\mathbf{X})^{-1}$ $j$:nes diagonaalielementti.

Estimaattorin kovarianssimatriisin estimaattori on puolestaan muotoa

$$\widehat{\mathrm{Cov}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}, \tag{3.8}$$

ja siten siis

$$\widehat{\mathrm{Var}}(\hat{\beta}_j) = \hat{\sigma}^2 t^{jj}. \tag{3.9}$$

Yksittäiselle parametrille $\beta_j$ saadaan muodostettua $100(1-\alpha)\%$ luottamusväliestimaatti välin

$$\left(\hat{\beta} - t_{\alpha/2}\hat{\sigma}\sqrt{t^{jj}}, \hat{\beta} + t_{\alpha/2}\hat{\sigma}\sqrt{t^{jj}}\right) \tag{3.10}$$

avulla, missä $t_{\alpha/2}$ on luku, jolle on voimassa $P(t_{n-(k+1)} > t_{\alpha/2}) = \alpha/2$, kun $t_{n-(k+1)}$ noudattaa Studentin $t$-jakaumaa vapausastein $df = n-(k+1)$.

Luottamusväliestimaatti annetuilla $\mathbf{x}_*$:n arvoilla odotusarvolle $\mu_* = \mathbf{x}_*'\boldsymbol{\beta}$ saadaan puolestaan kaavan

$$\left(\mathbf{x}_*'\hat{\boldsymbol{\beta}} - t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}, \mathbf{x}_*'\hat{\boldsymbol{\beta}} + t_{\alpha/2}\hat{\sigma}\sqrt{\mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}\right) \tag{3.11}$$

avulla. Luottamusväliennuste uudelle havainnolle $y_*$ saadaan taasen välillä

$$\left(\mathbf{x}_*'\hat{\boldsymbol{\beta}} - t_{\alpha/2}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*)}, \mathbf{x}_*'\hat{\boldsymbol{\beta}} + t_{\alpha/2}\sqrt{\hat{\sigma}^2(1+\mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*)}\right). \tag{3.12}$$

## 3.2 Mallin selitysaste

Merkitään

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2, \qquad SSR = \sum_{i=1}^{n}(\hat{\mu}_i - \bar{y})^2, \quad SSE = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2. \tag{3.13}$$

Tällöin on voimassa $SST = SSR + SSE$.

Tarkastellaan nyt malleja

$$\begin{aligned} \mathcal{M}_0 &: \ \mu_i = \beta_0, \\ \mathcal{M} &: \ \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \end{aligned}$$

Mallin $\mathcal{M}$ selitysaste $R^2(\mathcal{M})$ määritellään nyt suhteena

$$\begin{aligned} R^2(\mathcal{M}) &= \frac{SSR(\mathcal{M})}{SST(\mathcal{M})} = 1 - \frac{SSE(\mathcal{M})}{SST(\mathcal{M})} \\ &= 1 - \frac{SSE(\mathcal{M})}{SSE(\mathcal{M}_0)}. \end{aligned} \tag{3.14}$$

## 3.3 Mallin devianssi

Tarkastellaan logaritmoitua uskottavuusfunktiota odotusarvojen $\mu_i$ suhteen. Normaalijakauman tilanteessa logaritmoitu uskottavuusfunktio on muotoa

$$l(\mu_i, \sigma^2; \mathbf{y}) = \log\left((2\pi\sigma^2)^{-n/2}\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \mu_i)^2\right). \tag{3.15}$$

Odotusarvojen $\mu_i$ estimaateiksi voidaan valitaan havaitut arvot $y_i$. Tällöin odotusarvojen $\mu_i$ estimaatteja kutsutaan kyllästetyiksi estimaateiksi

$$\hat{\mu}_{i,K} = y_i. \tag{3.16}$$

Kyllästettyjen estimaattien $\hat{\mu}_{i,K}$ arvoilla logaritmoidun uskottavuusfunktion arvo supistuu muotoon

$$l(\hat{\mu}_{i,K}, \sigma^2; \mathbf{y}) = \log\left((2\pi\sigma^2)^{-n/2}\right). \tag{3.17}$$

Mallin

$$\mathcal{M} : \ \mu_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \mathbf{x}_i'\boldsymbol{\beta}$$

tilanteessa suurimman uskottavuuden estimaattien $\hat{\mu}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ arvoilla logaritmoitu uskottavuusfunktio saa arvon

$$l(\hat{\mu}_i, \sigma^2; \mathbf{y}) = \log\left((2\pi\sigma^2)^{-n/2}\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right). \tag{3.18}$$

Mallin $\mathcal{M}$ devianssi $D(\mathcal{M})$ määritellään kyllästettyjen estimaattien ja suurimman uskottavuuden estimaattien arvoilla laskettujen logaritmoitujen uskottavuusfunktion erotuksena

$$\begin{aligned} D(\mathcal{M}) &= 2\left(l(\hat{\mu}_{i,K}, \sigma^2; \mathbf{y}) - l(\hat{\mu}_i, \sigma^2; \mathbf{y})\right) \\ &= 2\left(\log\left((2\pi\sigma^2)^{-n/2}\right) - \left(\log\left((2\pi\sigma^2)^{-n/2}\right) - \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right)\right)\right) \\ &= 2\left(\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2\right)\right) = \frac{\sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{\sigma^2}. \end{aligned} \tag{3.19}$$

Voidaan osoittaa, että devianssi $D(\mathcal{M})$ noudattaa $\chi^2$-jakaumaa vapausastein $df = n-(k+1)$. Normaalijakauman tapauksessa devianssi $D(\mathcal{M})$ riippuu tuntemattomasta parametrista $\sigma^2$. Täten devianssin käyttäminen hypoteesin testaamiseen tai mallin sopivuuden mittaamiseen ei ole normaalijakauman tilanteessa suoraan mahdollista.

Edellä määritelty devianssi on itse asiassa oikealta termiltään skaalattu devianssi. Normaalijakauman tilanteessa voidaan määritellä myös niin sanottu ei-skaalattu devianssi (engl. unscaled deviance)

$$D_u(\mathcal{M}) = \sigma^2 D(\mathcal{M}) = \sum_{i=1}^{n}(y_i - \hat{\mu}_i)^2 = \sum_{i=1}^{n} e_i^2. \tag{3.20}$$

Ei-skaalattu devianssi $D_u(\mathcal{M})$ ei kuitenkaan koskaan tarkasti ottaen noudata $\chi^2$-jakaumaa, joten on parempi käyttää muita otossuureita hypoteesin testaamiseen ja mallin sopivuuden tarkasteluun.

## 3.4 Hypoteesin testaus

Yksittäisiin parametreihin $\beta_j$ liittyviä hypoteeseja

$$\begin{aligned} H_0 &: \beta_j = b_j, \\ H_a &: \beta_j \neq b_j, \end{aligned} \tag{3.21}$$

voidaan testata $t$-testillä

$$t = \frac{\hat{\beta}_j - b_j}{\hat{\sigma}\sqrt{t^{jj}}}, \tag{3.22}$$

missä testisuure $t$ noudattaa Studentin $t$-jakaumaa vapausastein $df = n-(k+1)$ kun $H_0$ hypoteesi on tosi.

Ositetaan seuraavaksi selittävät muuttujat ja parametrit kahteen osaan

$$\mathbf{x}_i = \begin{pmatrix} \mathbf{x}_{i1} \\ \mathbf{x}_{i2} \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}.$$

Testataan hypoteeseja

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{b}_2,$$
$$H_a : \boldsymbol{\beta}_2 \neq \mathbf{b}_2. \tag{3.23}$$

Testattavat hypoteesit vastaavat mallien muodossa hypoteeseja

$$H_0 : \text{Malli } \mathcal{M}_1 : \mu_i = \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \mathbf{x}'_{i2}\mathbf{b}_2 \text{ on voimassa,}$$
$$H_a : \text{Malli } \mathcal{M}_2 : \mu_i = \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + \mathbf{x}'_{i2}\boldsymbol{\beta}_2 \text{ on voimassa.}$$

Mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ devianssit $D(\mathcal{M}_1)$ ja $D(\mathcal{M}_2)$ voidaan laskea ja samoin devianssien erotus

$$\Delta D = D(\mathcal{M}_1) - D(\mathcal{M}_2). \tag{3.24}$$

Hypoteeseja saadaan testattua $F$-testin avulla

$$F = \frac{\Delta D / \dim(\mathbf{x}_2)}{D(\mathcal{M}_2)/n - (k+1)}, \tag{3.25}$$

missä testisuure $F$ noudattaa $H_0$ hypoteesin ollessa voimassa $F$-jakaumaa vapausastein $df_1 = \dim(\mathbf{x}_2), df_2 = n - (k+1)$. Merkintä $\dim(\mathbf{x}_2)$ tarkoittaa vektoreiden $\mathbf{x}_{i2}$ pituutta.

# Luku 4

# Yleistettyjen lineaaristen mallien teoriaa

## 4.1 Mallin rakenne

Kaikissa yleistetyissä lineaarisissa malleissa on seuraavat kolme komponenttia.

- Satunnaiskomponentti – Määrittää mallin selitettävän muuttujan $Y$ ja $Y$:n jakauman.
- Systemaattinen komponentti – Määrittää mallin selittävät muuttujat $x_1, x_2, \ldots, x_k$ joidenka katsotaan vaikuttavan selitettävän muuttujan $Y$ odotusarvon $\mu$ arvoon.
- Linkkifunktio – Määrittää sen funktion $g(\cdot)$ rakenteen, minkä kautta muuttujan $Y$ odotusarvo $\mu$ riippuu lineaarisesti selittävistä muuttujista $x_1, x_2, \ldots, x_k$.

Satunnaiskomponentti identifioi yleistetyn lineaarisen mallin selitettävän muuttujan $Y$ ja $Y$:n jakauman. Olkoon $Y_1, Y_2, \ldots, Y_n$ satunnaisotos $Y$:n jakaumasta, eli oletetaan, että jokainen $Y_i$ noudattaa muuttujan $Y$ jakaumaa. Oletetaan myös, että $Y_i$:t ovat toisistaan riippumattomia. Yleistetyissä lineaarisissa malleissa oletetaan, että $Y_i$:n jakauma kuuluu eksponentiaaliseen jakaumaperheeseen, eli että $Y_i$:n tiheysfunktio on muoto

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)e^{y_i Q(\theta_i)}, \tag{4.1}$$

kun satunnaismuuttujan $Y_i$ jakauma riippuu vain yhdestä tuntemattomasta parametrista $\theta_i$. Termiä $Q(\theta_i)$ kutsutaan luonnolliseksi parametriksi.

Useissa tilanteissa satunnaismuuttujalla $Y_i$ on kaksi toisensa poissulkevaa tulosvaihtoehtoa. Tällöin satunnaismuuttuja $Y_i$ on binaarinen muuttuja ja tulosvaihtoehdot voidaan koodata 0:lla ja 1:llä. Yleisemmin $Y_i$ voi olla satunnaismuuttuja, mikä kuvaa binaarisen satunnaismuuttujan 1:s tulosvaihtoehtojen (onnistumisten) lukumäärää tilanteessa, missä havainnoidaan binaarisen satunnaismuuttujan toteutunut arvo $n$ kertaa. Kummassakin tilanteessa oletetaan, että $Y_i$:t noudattavat binomijakaumaa.

Toisinaan selitettävä muuttuja voi saada positiivisia lukumääräarvoja. Esimerkiksi ristiintaulukoissa solufrekvenssit ovat ei-negatiivisia kokonaislukuja. Kun selitettävän muuttujan $Y_i$ tulosvaihtoehdot ovat ei-negatiivisia kokonaislukuja, voidaan olettaa, että $Y_i$:t noudattavat Poissonin jakaumaa.

Mikäli selitettävä muuttuja voidaan määritellä suhde- tai intervalliasteikolliseksi muuttujaksi, voidaan olettaa, että $Y_i$:t noudattavat normaalijakaumaa.

Yleistetyssä lineaarisessa mallissa mallinnetaan selitettävän muuttujan $Y$ odotusarvosta $\mathrm{E}(Y) = \mu$ riippuvan linkkifunktion $g(\mu)$ arvoa selittävien muuttujien $x_1, x_2, \ldots, x_k$ avulla lineaarisen yhtälön

$$g(\mu) = \eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \tag{4.2}$$

kautta. Linkkifunktio $g(\mu)$ yhdistää selittävät muuttujat $x_1, x_2, \ldots, x_k$ (systemaattisen komponentin) selitettävän muuttujan $Y$ arvoihin (satunnaiskomponenttiin).

Yksinkertaisin linkkifunktio on identtilinkki $g(\mu) = \mu$. Tällöin satunnaismuuttujan $Y_i$ odotusarvon $\mu_i$ odotetaan olevan lineaarisesti riippuvainen selittävien muuttujien $x_1, x_2, \ldots, x_k$ arvoista

$$\mu_i = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{4.3}$$

Tavallisen lineaarisen regressiomallin tilanteessa oletetaan nimenomaan, että normaalisti jakautuneen $Y_i$:n odotusarvo $\mu_i$ riippuu identtilinkin kautta lineaarisesti selittävistä muuttujista.

Toisenlaiset linkkifunktiot mahdollistavat odotusarvon $\mu$ olevan epälineaarisesti riippuvainen selittävistä muuttujista $x_1, x_2, \ldots, x_k$. Hyödyllinen linkkifunktio on log-linkki $g(\mu) = \log(\mu)$ mikä sopii tilanteisiin, missä odotusarvo $\mu$ ei voi olla negatiivinen kuten frekvenssidatan tilanteessa. Yleistettyä lineaarista mallia kutsutaan log-lineaariseksi malliksi, mikäli linkkifunktio on log-linkki muotoa. Log-lineaarinen malli on muotoa

$$\log(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{4.4}$$

Jos odotusarvo on välillä $0 \leq \mu \leq 1$, kuten todennäköisyyksien tilanteessa, käyttökelpoinen linkkifunktio on logit-linkki

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right). \tag{4.5}$$

Yleistettyä lineaarista mallia kutsutaan logistiseksi regressiomalliksi, mikäli linkkifunktio on logit-linkki muotoa. Logistinen regressiomalli on muotoa

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}. \tag{4.6}$$

Jos satunnaismuuttujan $Y_i$ jakauman ainoa tuntematon parametri on sen odotusarvo $\mu_i$, ja jos satunnaismuuttuja $Y_i$ kuuluu eksponentiaaliseen jakaumaperheeseen, niin

funktio $Q(\mu_i)$ on satunnaismuuttujan $Y_i$ luonnollinen parametri. Linkkifunktiota $g(\mu_i)$ kutsutaan kanooniseksi linkiksi, mikäli linkkifunktio on muotoa

$$g(\mu_i) = Q(\mu_i). \tag{4.7}$$

Käytännössä usein linkkifunktioksi $g(\mu_i)$ valitaan satunnaismuuttujan $Y_i$ kanooninen linkkifunktio $Q(\mu_i)$.

## 4.2 Hypoteesin testaus yleistetyssä lineaarisessa mallissa

Oletetaan, että selitettävän muuttujan $Y$ jakauma kuuluu eksponentiaaliseen jakaumaperheeseen tuntemattomana parametrina $Y$:n odotusarvo $\mu$, ja että mallin linkkifunktio $g(\mu)$ on kanooninen linkkifunktio $g(\mu) = \eta = Q(\mu)$. Oletetaan lisäksi, että kanooniselle linkkifunktiolle on olemassa käänteisfunktio $\mu = g^{-1}(\eta)$, missä

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{4.8}$$

Tällöin havaittuun satunnaisotokseen $\mathbf{y} = (y_1, y_2, \ldots y_n)'$ perustuva logaritmoitu uskottavuusfunktio on muotoa

$$l(\boldsymbol{\beta}, \mathbf{y}) = \sum_{i=1}^{n} \left[\log(a(g^{-1}(\eta_i))) + \log(b(y_i)) + y_i \eta_i\right], \tag{4.9}$$

missä $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$. Logaritmoidun uskottavuusfunktion avulla parametreille $\boldsymbol{\beta}$ voidaan laskea suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\beta}}$. Yleensä estimaateille $\hat{\boldsymbol{\beta}}$ ei löydy suljetun muodon ratkaisua ja siten estimaattien arvot joudutaan numeerisesti ratkaisemaan käyttämällä esim. Newton-Raphson algoritmia.

Yleistetyssä lineaarisessa mallissa suurimman uskottavuuden estimaatit $\hat{\boldsymbol{\beta}}$ noudattavat asymptoottisesti normaalijakaumaa. Esimerkiksi testattaessa hypoteesia

$$\begin{aligned} H_0 &: \beta_j = b_j, \\ H_a &: \beta_j \neq b_j, \end{aligned} \tag{4.10}$$

Waldin testisuure

$$Z = \frac{\hat{\beta}_j - b_j}{\hat{\sigma}_{\hat{\beta}_j}(\hat{\beta}_j)} \tag{4.11}$$

noudattaa asymptoottisesti standardoitua normaalijakaumaa $H_0$ hypoteesin ollessa voimassa.

Vastaavasti uskottavuussuhde

$$\Delta = -2 \log \left( \frac{L(b_j, \mathbf{y})}{L(\hat{\beta}_j, \mathbf{y})} \right) \tag{4.12}$$

noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = 1$ kun $H_0$ hypoteesi on voimassa.

## 4.3 Mallin devianssi

Merkitään satunnaisotoksen $Y_1, Y_2, \ldots, Y_n$ toteutuneita arvoja satunnaisvektorilla $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ ja satunnaisotoksen odotusarvoja vektorilla $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$. Uskottavuusfunktio $L$ voidaan kirjoittaa odotusarvovektorin $\boldsymbol{\mu}$ funktiona $L(\boldsymbol{\mu}, \mathbf{y})$. Jos nyt odotusarvovektorin $\boldsymbol{\mu}$ estimaattina käytetään toteutuneita arvoja $\mathbf{y}$, eli $\hat{\boldsymbol{\mu}}_{\mathcal{K}} = \mathbf{y}$, saadaan täydelliset soviteарvot tarkasteltavan datan tilanteessa. Estimaatteja $\hat{\boldsymbol{\mu}}_{\mathcal{K}} = \mathbf{y}$ kutsutaan kyllästetyksi estimaateiksi. Kyllästetyt estimaatit eivät ole käytännössä hyödyllinen, koska ne ei tiivistä informaatiota mitenkään alkuperäisistä toteutuneista havainnoista. Merkitään kyllästettyjen estimaattien tilanteessa uskottavuusfunktiota $L(\hat{\boldsymbol{\mu}}_{\mathcal{K}}, \mathbf{y})$:llä.

Tarkastellaan yleistettyä lineaarista mallia

$$\mathcal{M}: \qquad g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},$$

missä parametrit $\beta_0, \beta_1, \ldots, \beta_k$ estimoidaan suurimman uskottavuuden menetelmällä. Odotusarvojen estimaatit $\hat{\mu}_i$ saadaan mallin $\mathcal{M}$ tilanteessa linkkifunktion $g$ käänteisfunktion $g^{-1}$ avulla. Merkitään mallin $\mathcal{M}$ tilanteessa uskottavuusfunktiota $L(\hat{\boldsymbol{\mu}}, \mathbf{y})$:llä. Yleistettyjen lineaaristen mallien devianssi määritellään uskottavuussuhteena

$$D(\mathcal{M}) = -2 \log \left( \frac{L(\hat{\boldsymbol{\mu}}, \mathbf{y})}{L(\hat{\boldsymbol{\mu}}_{\mathcal{K}}, \mathbf{y})} \right) = 2 \left( l(\hat{\boldsymbol{\mu}}_{\mathcal{K}}, \mathbf{y}) - l(\hat{\boldsymbol{\mu}}, \mathbf{y}) \right). \tag{4.13}$$

Tarkastellaan seuraavaksi kahta hierarkista mallia $\mathcal{M}_0$ ja $\mathcal{M}_1$:

$$\begin{aligned} \mathcal{M}_1: &\qquad g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \\ \mathcal{M}_2: &\qquad g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}, \end{aligned}$$

missä $p < k$. Tällöin mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ devianssien erotus

$$D(\mathcal{M}_1) - D(\mathcal{M}_2) \tag{4.14}$$

noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = k - p$. Devianssien erotus saa suuria arvoja tilanteessa, jossa malli $\mathcal{M}_1$ sopii aineistoon huonommin verrattuna malliin $\mathcal{M}_2$. Devianssien erotuksella voidaan siis vertailla mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ sopivuutta dataan.

## 4.4 Yleistetty lineaarinen malli binaaridatan tilanteessa

Oletetaan, että selitettävä muuttuja $Y$ on binaarinen ja noudattaa Bernoullin jakaumaa parametrilla $\pi$, $Y \sim Ber(\pi)$, eli $P(Y = 1) = \pi$ ja $P(Y = 0) = 1 - \pi$. Tällöin $Y$:n odotusarvo $\mu = \pi$.

Ajatellaan, että todennäköisyys $\pi$ riippuu jostain selittävästä muuttujasta $x$. Jos $\pi$:n ja $x$:n välinen riippuvuus havainnon $i$ tilanteessa noudattaa lineaarista yhtälöä

$$\pi(x_i) = \beta_0 + \beta_1 x_i, \tag{4.15}$$

niin mallia kutsutaan lineaariseksi todennäköisyysmalliksi. Lineaarinen todennäköisyysmalli on yleistetty lineaarinen malli, missä satunnaiskomponentti on binomijakautunut ($Ber(\pi) = Bin(1, \pi)$) ja linkkifunktio $g(\mu_i)$ on identtilinkki.

Lineaarisen todennäköisyysmalli ongelma on se, että sovitemalli saattaa antaa isoilla tai pienillä $x$:n arvoilla todennäköisyyden sovitearvoiksi $\hat{\pi}(x)$ arvoja, jotka eivät kuulu välille (0,1). Täten lineaarinen todennäköisyysmalli saattaa olla käyttökelpoinen vain tietyillä $x$:n arvoilla.

Logistinen regressiomalli on erittäin käyttökelpoinen malli kun oletetaan, että Bernoullin jakaumaa noudattavan selitettävän muuttujan $Y$ tulosvaihtoehdon 1 todennäköisyys $\pi(x)$ riippuu epälineaarisesti selittävän muuttujan $x$ arvoista. Logistisessa regressiomallissa $\pi$:n ja $x$:n välinen riippuvuus havainnon $i$ tilanteessa noudattaa yhtälöä

$$\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1 + e^{\beta_0+\beta_1 x_i}}. \tag{4.16}$$

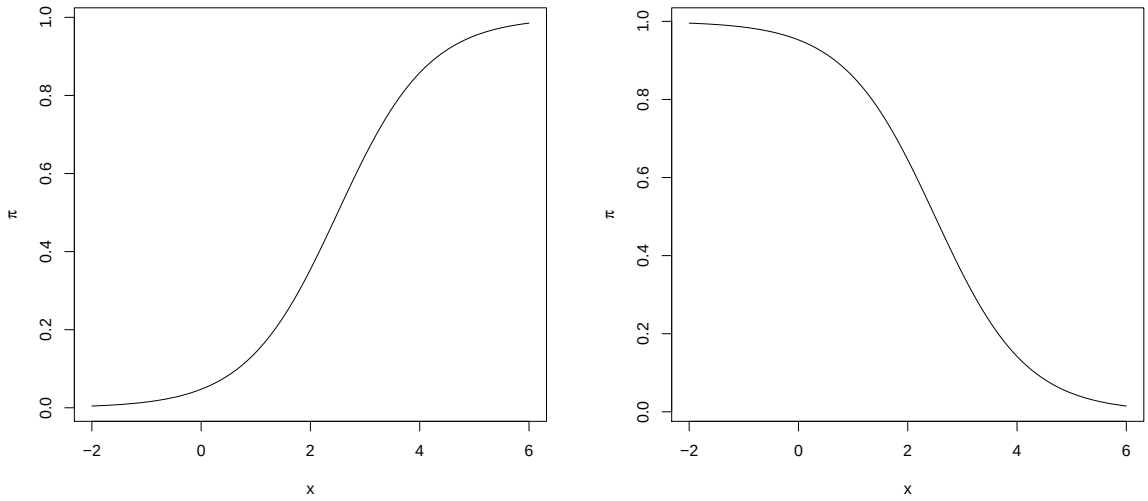
Logistisen regressiomallin tilanteessa vedonlyöntikerroin $\gamma(x_i)$ on muotoa

$$\gamma(x_i) = \frac{\pi(x_i)}{1 - \pi(x_i)} = e^{\beta_0+\beta_1 x_i}, \tag{4.17}$$

ja täten logaritmoitu vedonlyöntikerroin noudattaa lineaarista yhtälöä

$$\log(\gamma(x_i)) = \log\left(\frac{\pi(x_i)}{1 - \pi(x_i)}\right) = \text{logit}(\pi(x_i)) = \beta_0 + \beta_1 x_i. \tag{4.18}$$

Logistinen regressiomalli on yleistetty lineaarinen malli, missä satunnaiskomponentti on binomijakautunut ja linkkifunktio $g(\mu_i)$ on logit-linkki. Logit-linkki on binomijakauman tilanteessa kanooninen linkkifunktio. Logistista regressiomallia kutsutaan myös logit malliksi. Alla olevassa kuvassa on esitetty miltä $\pi$:n ja $x$:n välinen riippuvuus näyttää logistisen regressiomallin tilanteessa kun parametri $\beta_1 > 0$ ja $\beta_1 < 0$.



## 4.5 Mallintaminen $2 \times 2$-ristiintaulukossa

Tarkastellaan seuraavaa muuttujien $X$ ja $Y$ välistä $2 \times 2$-ristiintaulukkoa:

| | | $y_j =$ 1 | 0 | Yhteensä |
|---|---|---|---|---|
| $x_i =$ | 1 | $\pi_1$ | $1 - \pi_1$ | 1 |
| | 0 | $\pi_2$ | $1 - \pi_2$ | 1 |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

Oletetaan nyt, että $Y$ noudattaa Bernoullin jakaumaa $Y \sim Ber(\pi(x))$, missä todennäköisyys $\pi(x)$ riippuu $X$:n havaitusta arvosta $x$. Mikäli todennäköisyys $\pi$ riippuu havaitusta arvosta $x_i$ lineaarisen todennäköisyysmallin mukaan

$$\pi(x_i) = \beta_0 + \beta_1 x_i, \tag{4.19}$$

niin silloin

$$\beta_1 = \pi(x_i = 1) - \pi(x_i = 0). \tag{4.20}$$

Mikäli taas logistisessa regressiomalli kuvaa $\pi$:n ja $x$:n välistä riippuvuutta, niin silloin

$$\begin{aligned}\beta_1 &= \text{logit}[\pi(x_i = 1)] - \text{logit}[\pi(x_i = 0)] = \log\left(\frac{\pi(1)}{1 - \pi(1)}\right) - \log\left(\frac{\pi(0)}{1 - \pi(0)}\right) \\ &= \log\left(\frac{\pi(1)/1 - \pi(1)}{\pi(0)/1 - \pi(0)}\right).\end{aligned} \tag{4.21}$$

Logistisen regressiomallin tilanteessa siis parametri $\beta_1$ on logaritmoitu vedonlyöntisuhde $\log(\theta)$.

## 4.6 Yleistetty lineaarinen malli frekvenssidatan tilanteessa

Oletetaan, että selitettävä muuttuja $Y$ saa ei-negatiivisia kokonaislukuarvoja ja noudattaa Poissonin jakaumaa parametrilla $\lambda$, $Y \sim Poi(\lambda)$. Tällöin $Y$:n odotusarvo ja varianssi ovat $\mu = \lambda$ ja $\sigma^2 = \lambda$.

Poissonin jakaumaa noudattavan selitettävän muuttujan $Y$ odotusarvoa $\mu$ voidaan mallintaa yleistetyllä lineaarisella mallilla, missä linkkifunktio on identtilinkki. Tällöin odotusarvon $\mu$ ja selittävän muuttujan $x$ välinen riippuvuus havainnon $i$ tilanteessa noudattaa Poissonin lineaarista regressiomallia

$$\mu_i = \beta_0 + \beta_1 x_i. \tag{4.22}$$

Useimmin Poissonin jakauman tilanteessa kuitenkin mallinnetaan logaritmoitua odotusarvoa $\log(\mu)$. Poissonin log-lineaarinen malli on yleistetty lineaarinen malli, missä siis linkkifunktio on log-linkki:

$$\log(\mu_i) = \beta_0 + \beta_1 x_i. \tag{4.23}$$

Poissonin log-lineaarisen mallin tilanteessa havainnon $i$ odotusarvo $\mu_i$ riippuu epälineaarisesti selittävästä muuttujasta $x_i$:

$$\mu_i = e^{\beta_0+\beta_1 x_i} = e^{\beta_0}\left(e^{\beta_1}\right)^{x_i}. \tag{4.24}$$

Poissonin log-lineaariselle mallille tyypillinen ongelma on se, että datassa selitettävän muuttujan varianssi on suurempi verrattuna mallin antamaan varianssiin. Poissonin jakauman tilanteessa odotusarvon ja varianssin pitäisi olla yhtä suuret. Usein kuitenkin käytännön aineistoissa selitettävän muuttujan varianssi on suurempi kuin mitä Poissonin log-lineaarinen mallin mukaan varianssin pitäisi olla annetulla selittävän muuttujan $x_i$ arvolla. Tälläistä ilmiötä kutsutaan ylihajonnaksi.

## 4.7 Poissonin log-lineaarinen malli $I \times J$-ristiintaulukossa

Poissonin log-lineaarista mallia voidaan käyttää mallintamaan ristiintaulukon solufrekvenssejä. Olkoon $Y_{ij}$:t $I \times J$-ristiintaulukon solufrekvenssejä, jotka noudattavat Poissonin jakaumaa $Y_{ij} \sim Poi(\mu_{ij})$. Oletetaan, että ristiintaulukon rivi-ja sarakemuuttujat ovat riippumattomia, eli solutodennäköisyyksille on voimassa

$$\pi_{ij} = \pi_{i+}\pi_{+j}. \tag{4.25}$$

Tällöin odotetut frekvenssit $\mu_{ij}$ ovat muotoa

$$\mu_{ij} = n_{++}\pi_{ij} = n_{++}\pi_{i+}\pi_{+j}, \tag{4.26}$$

ja siten logaritmoidut odotusarvot muotoa

$$\begin{aligned}\log(\mu_{ij}) &= \log(n_{++}) + \log(\pi_{i+}) + \log(\pi_{+j}) \\ &= \alpha + \beta_i + \gamma_j.\end{aligned} \tag{4.27}$$

Eli jos ristiintaulukon rivi-ja sarakemuuttujat ovat riippumattomia, log-lineaarisessa mallissa on rivi-ja sarakemuuttujien päävaikutukset muttei niiden yhdysvaikutuksia.

# Luku 5

# Logistinen regressio

## 5.1 Mallin perusteet

Logistinen regressiomalli on tärkein binaarisen selitettävän muuttujan malli. Olkoon $Y$ Bernoullin jakaumaa noudattava selitettävä muuttuja ja olkoon $X$ selittävä muuttuja. Merkitään

$$\pi(x) = P(Y = 1|X = x). \tag{5.1}$$

Logistisessa regressiomallissa oletetaan, että

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}, \tag{5.2}$$

eli, että logaritmoitu vedonlyönti kerroin $\gamma(x)$ on lineaarisesti riippuvainen selittävästä muuttujasta $X$:

$$\log(\gamma(x)) = \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \beta_0 + \beta_1 x. \tag{5.3}$$

Jos selittäviä muuttujia on useita $\mathbf{X} = (X_1, X_2, \ldots, X_k)$ ja

$$\pi(\mathbf{x}) = P(Y = 1|\mathbf{X} = \mathbf{x}), \tag{5.4}$$

niin silloin logistinen regressiomalli on muotoa

$$g(\pi(\mathbf{x})) = \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) = \text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k. \tag{5.5}$$

Logistinen regressiomalli on yleistetty lineaarinen malli, missä linkkifunktio $g$ on logit-linkki.

Tarkastellaan logistista regressiomallia

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \beta_0 + \beta_1 x. \tag{5.6}$$

Jos $\beta_1 > 0$, niin $\pi(x)$ kasvaa kun $x$ kasvaa. Jos $\beta_1 = 0$, niin todennäköisyys $\pi(x)$ ei riipu selittävästä muuttujasta ja siten $Y$ on riippumaton $X$:stä.

Tarkastellaan logistisen regressiomallin arvoja $X$:n arvoilla $X = x$ ja $X = x + 1$. Tällöin logaritmoitu vedonlyöntisuhde $\theta_{x+1|x}$ on muotoa

$$\begin{aligned}\log(\theta_{x+1|x}) &= \log\left(\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x+1)}{1-\pi(x+1)}}\right) = \log\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1-\pi(x)}\right) \\ &= \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1(x)) = \beta_1, \end{aligned} \tag{5.7}$$

ja siten $\theta_{x+1|x} = e^{\beta_1}$ ja $\hat{\theta}_{x+1|x} = e^{\hat{\beta}_1}$.

Tarkastellaan logistisen regressiomallin tilanteessa hypoteesia

$$H_0 : \beta_1 = 0. \tag{5.8}$$

Tällöin Waldin testisuure

$$Z = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \tag{5.9}$$

noudattaa asymptoottisesti standardoitua normaalijakaumaa $H_0$ hypoteesin ollessa voimassa.

$H_0$ hypoteesi voidaan testata myös devianssien avulla. Tarkastellaan malleja $\mathcal{M}_0$ ja $\mathcal{M}_1$:

$$\begin{aligned}\mathcal{M}_0 : &\quad \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \beta_0, \\ \mathcal{M}_1 : &\quad \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \beta_0 + \beta_1 x. \end{aligned}$$

Tällöin mallien $\mathcal{M}_0$ ja $\mathcal{M}_1$ devianssien erotus

$$D(\mathcal{M}_0) - D(\mathcal{M}_1) \tag{5.10}$$

noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = 1$ kun $H_0$ hypoteesi on voimassa.

Waldin testisuureen avulla voidaan muodostaa parametrille $\beta_1$ $100(1-\alpha)\%$ asymptoottinen luottamusväli käyttäen kaavaa

$$\hat{\beta}_1 \pm z_{\alpha/2}\hat{\sigma}(\hat{\beta}_1), \tag{5.11}$$

missä $P(Z > z_{\alpha/2}) = \alpha/2$ kun $Z \sim N(0,1)$.

Todennäköisyydelle $\pi(x)$ voidaan luoda luottamusestimaatti logistisen regressiomallin kautta. Logit-linkin suurimman uskottavuuden estimaatti on muotoa

$$\text{logit}(\hat{\pi}(x)) = \hat{\beta}_0 + \hat{\beta}_1 x, \tag{5.12}$$

ja estimoidun logit-linkin varianssi on muotoa

$$\begin{aligned}\sigma^2\left(\text{logit}(\hat{\pi}(x))\right) &= \text{Var}\left(\text{logit}(\hat{\pi}(x))\right) = \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x\right) \\ &= \text{Var}(\hat{\beta}_0) + x^2\,\text{Var}(\hat{\beta}_1) + 2x\,\text{Cov}(\hat{\beta}_0, \hat{\beta}_1). \end{aligned} \tag{5.13}$$

Estimoidun logit-linkin estimoitu varianssi $\hat{\sigma}^2\left(\text{logit}(\hat{\pi}(x))\right)$ saadaan sitten laskettua korvaamalla varianssin kaavassa tuntemattomat varianssit ja kovarianssit niiden estimaateilla. Estimoidun varianssin avulla voidaan logit-linkille muodostaa $100(1-\alpha)\%$ asymptoottinen luottamusväli käyttäen kaavaa

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm z_{\alpha/2}\hat{\sigma}\left(\text{logit}(\hat{\pi}(x))\right). \tag{5.14}$$

Täten todennäköisyydelle $\pi(x)$ voidaan muodostaa luottamusväliestimaatti laskemalla logit-linkin käänteisfunktion arvot logit-linkin luottamusväliestimaatin päätepisteiden arvoilla

$$\left(\frac{e^{\hat{\beta}_0+\hat{\beta}_1 x - z_{\alpha/2}\hat{\sigma}(\text{logit}(\hat{\pi}(x)))}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x - z_{\alpha/2}\hat{\sigma}(\text{logit}(\hat{\pi}(x)))}}, \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x + z_{\alpha/2}\hat{\sigma}(\text{logit}(\hat{\pi}(x)))}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x + z_{\alpha/2}\hat{\sigma}(\text{logit}(\hat{\pi}(x)))}}\right). \tag{5.15}$$

## 5.2 Mallin arvioiminen

Tarkastellaan seuraavaksi kahta hierarkista logistista regressiomallia $\mathcal{M}_1$ ja $\mathcal{M}_2$:

$$\begin{aligned}\mathcal{M}_1: &\quad \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \text{logit}(\pi(\mathbf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}, \\ \mathcal{M}_2: &\quad \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \text{logit}(\pi(\mathbf{x}_i)) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik},\end{aligned}$$

missä $p < k$. Tällöin mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ devianssien erotus

$$D(\mathcal{M}_1) - D(\mathcal{M}_2) \tag{5.16}$$

noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = k - p$ mikäli mallin $\mathcal{M}_1$ osalta tietyt jakaumaoletukset ovat kunnossa ($n_i(\mathbf{x}_i)$ riittävän suuri). Devianssien erotus saa suuria arvoja tilanteessa, jossa malli $\mathcal{M}_1$ sopii aineistoon huonommin verrattuna malliin $\mathcal{M}_2$. Devianssien erotuksella voidaan siis vertailla mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ sopivuutta dataan.

Yksittäisen mallin $\mathcal{M}_1$ riittävyyttä verrattuna kyllästettyyn malliin $\mathcal{K}$ voidaan testata devianssilla $D(\mathcal{M}_1)$, mikäli jokaisella $\mathbf{x}_i$ arvolla $n_i(\mathbf{x}_i) \geq 5$. Tällöin devianssi $D(\mathcal{M}_1)$ noudattaa asymptoottisesti $\chi^2$-jakaumaa vapausastein $df = N_1 - p$, missä $N_1$ on eri $\mathbf{x}_i$ vektoreiden lukumäärä.

Vaihtoehtoinen tapa tarkastella mallien $\mathcal{M}_1$ ja $\mathcal{M}_2$ paremmuutta on laskea malleista Akaikenin informaatio kriteerit $AIC$:

$$AIC(\mathcal{M}_1) = -2[\log(L_{\mathcal{M}_1}(\hat{\boldsymbol{\beta}})) - p], \tag{5.17}$$

$$AIC(\mathcal{M}_2) = -2[\log(L_{\mathcal{M}_2}(\hat{\boldsymbol{\beta}})) - k]. \tag{5.18}$$

Malli, millä on pienempi $AIC$ arvo, on Akaikenin informaatio kriteerin mukaan parempi.

Malleista $\mathcal{M}_1$ ja $\mathcal{M}_2$ voidaan laskea myös lineaarisen mallin selitysastetta vastaava Naglekerken arvo:

$$R^2(\mathcal{M}_1) = \frac{1 - e^{(D(\mathcal{M}_1) - D(\mathcal{M}_0))/n_{++}}}{1 - e^{-D(\mathcal{M}_0)/n_{++}}}, \tag{5.19}$$

$$R^2(\mathcal{M}_2) = \frac{1 - e^{(D(\mathcal{M}_2) - D(\mathcal{M}_0))/n_{++}}}{1 - e^{-D(\mathcal{M}_0)/n_{++}}}, \tag{5.20}$$

missä $D(\mathcal{M}_0)$ on devianssi mallista

$$\mathcal{M}_0 : \qquad \log\left(\frac{\pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}\right) = \text{logit}(\pi(\mathbf{x}_i)) = \beta_0.$$

Naglekerken selitysaste saa arvoja väliltä $0 \leq R^2 \leq 1$.

## 5.3 Residuaalit logistisessa regressiomallissa

Olkoon $\hat{\pi}(\mathbf{x}_i)$ logistisen regressiomallin antama sovite todennäköisyydelle $\pi(\mathbf{x}_i)$ selittävien muuttujien $\mathbf{x}_i$ arvoilla. Jos $\mathbf{x}_i$ arvoilla on toistettu Bernoullin koetta $n_i(\mathbf{x}_i)$ kertaa, niin Pearsonin residuaali määritellään suhteena

$$e_i = \frac{y_i(\mathbf{x}_i) - n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)}{\sqrt{n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i))}}, \tag{5.21}$$

missä $y_i(\mathbf{x}_i)$ on onnistumisten lukumäärä arvoilla $\mathbf{x}_i$.

Standardoitu residuaali on puolestaan määritelty suhteena

$$r_i = \frac{y_i(\mathbf{x}_i) - n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)}{\sqrt{n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)(1 - \hat{\pi}(\mathbf{x}_i))(1 - \hat{h}_i(\mathbf{x}_i))}}, \tag{5.22}$$

missä $\hat{h}_i(\mathbf{x}_i)$ on selittävien muuttujien arvoista riippuva leverage arvo.

Logistisesta regressiomallista voidaan laskea myös devianssiresiduaalit

$$d_i = \sqrt{q_i} \times sign(y_i(\mathbf{x}_i) - n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)), \tag{5.23}$$

missä

$$q_i = 2\left(y_i(\mathbf{x}_i)\log\left(\frac{y_i(\mathbf{x}_i)}{n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)}\right) + (n_i(\mathbf{x}_i) - y_i(\mathbf{x}_i))\log\left(\frac{n_i(\mathbf{x}_i) - y_i(\mathbf{x}_i)}{n_i(\mathbf{x}_i) - n_i(\mathbf{x}_i)\hat{\pi}(\mathbf{x}_i)}\right)\right).$$

Muodostamalla pisteparvikuvioita residuaaleista ja selittävistä muuttujista tai $\text{logit}(\hat{\pi}(\mathbf{x}_i))$ sovitearvoista, voidaan kuvien avulla yrittää löytää mahdollisia syitä miksei malli mahdollisesti sovi tarpeeksi hyvin dataan. Kun $n_i(\mathbf{x}_i) = 1$ residuaalien käyttökelpoisuus on kuitenkin hyvin rajallinen.

## 5.4 Luokitteluasteikolliset selittävät muuttujat

Olkoon $Y$ Bernoullin jakaumaa noudattava selitettävä muuttuja ja olkoon $X_1$ ja $X_2$ binaarisia selittäviä muuttujia, jotka siten voivat saada kaksi toisensa poissulkevaa tulosvaihtoehtoa. Tässä tilanteessa satunnaisotoksen $Y_1, Y_2, \ldots, Y_n$ tulokset voidaan siten esittää $2 \times 2 \times 2$-ristiintaulukon avulla, missä todennäköisyyksillä on voimassa ristiintaulukko

| | | | | $y_k =$ 1 | 0 | Yhteensä |
|---|---|---|---|---|---|---|
| $x_{i1} =$ | 1 | $x_{j2} =$ | 1 | $\pi(1,1)$ | $1-\pi(1,1)$ | 1 |
| | | | 0 | $\pi(1,0)$ | $1-\pi(1,0)$ | 1 |
| $x_{i1} =$ | 0 | $x_{j2} =$ | 1 | $\pi(0,1)$ | $1-\pi(0,1)$ | 1 |
| | | | 0 | $\pi(0,0)$ | $1-\pi(0,0)$ | 1 |
| | | | Yhteensä | $\pi_{++1}$ | $\pi_{++2}$ | 1 |

Mallinnetaan todennäköisyyttä $P(Y = 1|X_1 = x_1, X_2 = x_2) = \pi(x_1, x_2)$ logistisella regressiomallilla

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2. \tag{5.24}$$

Mallia kutsutaan päävaikutusmalliksi. Muuttujat $x_1$ ja $x_2$ ovat indikaattorimuuttujia, jotka voivat saada arvoja 0 tai 1.

Jos logistiseen regressiomalliin lisätään muuttujien $x_1$ ja $x_2$ yhteisvaikutus, on malli muotoa

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2. \tag{5.25}$$

Yllä olevien mallien avulla voidaan tutkia, että selittävätkö kummatkin muuttujat $X_1$ ja $X_2$ selitettävän muuttujan $Y$ arvoja, ja onko muuttujilla $X_1$ ja $X_2$ lisäksi vielä yhdysvaikutusta muuttujan $Y$ arvoihin.

Mikäli muuttujilla $X_1$ ja $X_2$ olisi $I$ ja $J$ eri toisensa poissulkevaa tulosvaihtoehtoa, voidaan päävaikutusmalli kuvata parametrein

$$\text{logit}(\pi(\mathbf{x})) = \beta_0 + \beta_i^{x_1} + \beta_j^{x_2}, \tag{5.26}$$

missä tuntemattomia parametreja $\beta_i$ on $I - 1$ kappaletta ja parametreja $\beta_j$ $J - 1$ kappaletta. Eli $\beta_i^{x_1}$ tarkoittaa samaa kuin

$$\beta_i^{x_1} = \beta_{11} x_{11} + \beta_{12} x_{12} + \cdots + \beta_{1(I-1)} x_{1(I-1)}, \tag{5.27}$$

missä $x_{11}, x_{12}, \ldots, x_{1(I-1)}$ ovat kaikki indikaattorimuuttujia saaden arvoja 0 tai 1 riippuen muuttujan $X_1$ tulosvaihtoehdon toteutumisesta.

## 5.5 Moniluokkaiset logit mallit

Logistisen regressiomallin tilanteessa olettiin, että selitettävä muuttuja $Y$ on binaarinen Bernoullin jakaumaa $Y \sim Ber(\pi)$ noudattava satunnaismuuttuja, missä logit-linkkifunktio riippuu selittävästä muuttujasta $X$ lineaarisesti

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \text{logit}(\pi(x)) = \beta_0 + \beta_1 x. \tag{5.28}$$

Moniluokkaiset logit mallit ovat logistisen regressiomallin yleistyksiä tilanteeseen, missä selitettävällä muuttujalla $Y$ on $J$ toisensa poissulkevaa tulosvaihtoehtoa. Merkitään tulosvaihtoehtojen $J$ todennäköisyyksiä vektorilla

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_J). \tag{5.29}$$

Moniluokkaisissa logit malleissa valitaan jokin todennäköisyyksistä $\pi_1, \pi_2, \ldots, \pi_J$ vertailukohdaksi, esim. todennäköisyys $\pi_1$, ja sen jälkeen mallinnetaan logaritmoituja vedonlyöntikertoimia

$$\log\left(\frac{\pi_j}{\pi_1}\right), \qquad j = 2, 3, \ldots, J. \tag{5.30}$$

Moniluokkaisissa logit malleissa logaritmoidut vedonlyöntikertoimet riippuvat sitten lineaarisesti selittävästä muuttujasta $X$:

$$\log\left(\frac{\pi_j}{\pi_1}\right) = \alpha_j + \beta_j x, \qquad j = 2, 3, \ldots, J. \tag{5.31}$$

Täten moniluokkaisten logit mallien tilanteessa todennäköisyydet $\pi_j$ ovat muotoa

$$\pi_j = \frac{e^{\alpha_j + \beta_j x}}{1 + \sum_{h=2}^{J} e^{\alpha_h + \beta_h x}}. \tag{5.32}$$

## 5.6 Kumulatiiviset logit mallit

Jos selitettävä muuttuja $Y$ on järjestysasteikollinen muuttuja, voidaan luokkien $J$ luonnollinen järjestys ottaa mukaan analyysiin mallintamalla kumulatiivisia todennäköisyyksiä logit malleilla. Kumulatiiviset todennäköisyydet määritellään seuraavasti:

$$P(Y \leq j|x) = \pi_1 + \pi_2 + \cdots + \pi_j, \qquad j = 1, \ldots, J. \tag{5.33}$$

Yksi mahdollinen tapa mallintaa kumulatiivisia todennäköisyyksiä on käyttää suhteellisten vedonlyöntikertoimien kumulatiivista logit mallia

$$\log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \alpha_j + \beta x, \qquad j = 1, \ldots, J-1. \tag{5.34}$$

Suhteellisten vedonlyöntikertoimien kumulatiivisessa logit mallissa parametri $\beta$ on sama jokaisessa $J-1$ yhtälössä.

# Luku 6

# Poissonin log-lineaarinen malli

## 6.1 Log-lineaariset mallit kaksiulotteisissa ristiintaulukoissa

Poissonin log-lineaarisella mallilla voidaan mallintaa frekvenssityyppistä dataa. Siten Poissonin log-lineaarisia malleja voidaan käyttää mallintamaan ristiintaulukoiden solufrekvenssejä.

Olkoon satunnaismuuttujilla $X$ ja $Y$ $I$ ja $J$ toisensa poissulkevaa tulosvaihtoehtoa. Satunnaismuuttujien $X$ ja $Y$ yhteistodennäköisyysjakauma voidaan esittää $I \times J$-ristiintaulukon avulla:

| $P(X = x_i, Y = y_j)$ : | | $y_j=$ 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i=$ | 1 | $\pi_{11}$ | $\pi_{12}$ | ... | $\pi_{1J}$ | $\pi_{1+}$ |
| | 2 | $\pi_{21}$ | $\pi_{22}$ | ... | $\pi_{2J}$ | $\pi_{2+}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| | I | $\pi_{I1}$ | $\pi_{I2}$ | ... | $\pi_{IJ}$ | $\pi_{I+}$ |
| | Yhteensä | $\pi_{+1}$ | $\pi_{+2}$ | ... | $\pi_{+J}$ | 1 |

Toistettaessa satunnaismuuttujien $X$ ja $Y$ muodostamaa 2-ulotteista satunnaiskoetta $n_{++}$ kertaa, voidaan tulosvaihtoehtojen $(X = x_i, Y = y_j)$ frekvenssijakauma kuvata ristiintaulukolla:

| $Freq(X = x_i, Y = y_j)$ : | | $y_j=$ 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i=$ | 1 | $n_{11}$ | $n_{12}$ | ... | $n_{1J}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | ... | $n_{2J}$ | $n_{2+}$ |
| | $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| | I | $n_{I1}$ | $n_{I2}$ | ... | $n_{IJ}$ | $n_{I+}$ |
| | Yhteensä | $n_{+1}$ | $n_{+2}$ | ... | $n_{+J}$ | $n_{++}$ |

Havaittuja solufrekvenssejä $n_{ij}$ voidaan sitten verrata odotettuihin solufrekvensseihin $\mu_{ij} = n_{++}\pi_{ij}$:

| $\mathrm{E}(n_{ij})$ : | | $y_j$= 1 | 2 | ... | J | Yhteensä |
|---|---|---|---|---|---|---|
| $x_i$= | 1 | $\mu_{11}$ | $\mu_{12}$ | ... | $\mu_{1J}$ | $\mu_{1+}$ |
| | 2 | $\mu_{21}$ | $\mu_{22}$ | ... | $\mu_{2J}$ | $\mu_{2+}$ |
| | ⋮ | ⋮ | ⋮ | | ⋮ | ⋮ |
| | I | $\mu_{I1}$ | $\mu_{I2}$ | ... | $\mu_{IJ}$ | $\mu_{I+}$ |
| | Yhteensä | $\mu_{+1}$ | $\mu_{+2}$ | ... | $\mu_{+J}$ | 1 |

Tarkastellaan muuttujien $X$:n ja $Y$:n välistä riippuvuutta. Testataan hypoteesia, että $X$ ja $Y$ ovat riippumattomia:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \qquad \text{kaikille } i \text{ ja } j. \tag{6.1}$$

$H_0$ hypoteesin vallitessa frekvenssien $n_{ij}$ odotetut frekvenssit $\mu_{ij}$ ovat muotoa

$$\mu_{ij} = n_{++}\pi_{i+}\pi_{+j}. \tag{6.2}$$

Täten ottamalla odotetuista frekvensseistä $\mu_{ij}$ logaritmit, saadaan

$$\begin{aligned} \log(\mu_{ij}) &= \log(n_{++}) + \log(\pi_{i+}) + \log(\pi_{+j}) \\ &= \beta_0 + \beta_i^x + \beta_j^y. \end{aligned} \tag{6.3}$$

Yllä olevassa yhtälössä $x$ ja $y$ ovat yläindeksejä, eivät potenssiin korotuksia.

Mikäli nyt oletetaan, että solufrekvenssit $n_{ij}$ noudattavat Poissonin jakaumaa $n_{ij} \sim Poi(\mu_{ij})$ ja odotetuille frekvensseille on voimassa log-linkkifunktio

$$\mathcal{M}(X, Y) : \qquad \log(\mu_{ij}) = \beta_0 + \beta_i^x + \beta_j^y,$$

niin mallia kutsutaan Poissonin log-lineaariseksi päävaikutusmalliksi. Päävaikutusmalli liittyy hypoteesiin, että muuttujat $X$ ja $Y$ ovat riippumattomia. Hypoteesi hyväksytään, mikäli päävaikutusmallin katsotaan sopivan aineistoon riittävän hyvin.

Mallin sopivuutta dataan voidaan testata Pearsonin $X^2$-testisuureen ja uskottavuussuhteen $G^2$ avulla. Olkoon $\hat{\mu}_{ij}$ log-lineaarisen mallin antamat estimaatit odotetuille frekvensseille. Tällöin Pearsonin $X^2$-testisuure ja uskottavuussuhde $G^2$ ovat $I \times J$-ristiintaulukon tilanteessa muotoa

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right). \tag{6.4}$$

Uskottavuussuhde $G^2$ vastaa tarkasteltavan mallin $\mathcal{M}$ devianssia $D(\mathcal{M})$. Pearsonin $X^2$-testisuure ja uskottavuussuhde $G^2$ noudattavat asymptoottisesti $\chi^2$-jakaumaa. $\chi^2$-jakauman vapausasteet ovat ristiintaulukon solujen lukumäärä miinus log-lineaarisen mallin parametrien lukumäärä.

Mikäli $X$:n ja $Y$:n välillä on riippuvuutta, log-lineaarinen malli on muotoa

$$\mathcal{M}(XY): \qquad \log(\mu_{ij}) = \beta_0 + \beta_i^x + \beta_j^y + \beta_{ij}^{xy}.$$

Mallia $\mathcal{M}(XY)$ kutsutaan log-lineaariseksi yhdysvaikutusmalliksi. Parametrit $\beta_{ij}^{xy}$ ovat yhdysvaikutustermejä, jotka kuvaavat mallin poikkeavuutta päävaikutusmallista. Yhdysvaikutusmallissa $\mathcal{M}(XY)$ on itse asiassa yhtä paljon parametrejä kuin havaintoja $n_{ij}$, joten yhdysvaikutusmalli on kyllästetty malli, mikä sopii täydellisesti aineistoon.

Tarkasteltaessa onko muuttujien $X$:n ja $Y$:n välillä on riippuvuutta, tarkastellaan siis sopiiko päävaikutusmalli $\mathcal{M}(X,Y)$ riittävän hyvin aineistoon vai tarvitseeko ristiintaulukko mallintaa yhdysvaikutusmallin $\mathcal{M}(XY)$ avulla.

## 6.2 Log-lineaarinen malli ja logistinen regressio

Log-lineaarisessa mallissa ajatellaan, että kumpikin muuttujista $X$ ja $Y$ ovat selitettäviä muuttujia, joiden riippuvuutta tarkastellaan. Mikäli $X$:n katsotaan selittävän $Y$:n arvoja ja $Y$:llä on kaksi eri tulosvaihtoehtoa, niin $I \times 2$-ristiintaulukon voidaan ajatella kuvaavan $Y$:n ehdollista todennäköisyysjakaumaa eri $X$:n arvoilla.

Mallinnetaan $I \times 2$-ristiintaulukon tilanteessa $X$:n ja $Y$:n välistä riippuvuutta Poissonin log-lineaarisella päävaikutusmallilla

$$\mathcal{M}(X,Y): \qquad \log(\mu_{ij}) = \beta_0 + \beta_i^x + \beta_j^y.$$

Tällöin

$$\begin{aligned} \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) &= \beta_0 + \beta_i^x + \beta_1^y - \beta_0 - \beta_i^x - \beta_2^y \\ &= \beta_1^y - \beta_2^y. \end{aligned} \tag{6.5}$$

Koska

$$\log\left(\frac{\pi_{x_i}}{1-\pi_{x_i}}\right) = \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) \tag{6.6}$$

ja $\alpha = \beta_1^y - \beta_2^y$ ei riipu $x$:stä, niin log-lineaarinen päävaikutusmalli $\mathcal{M}(X,Y)$ vastaa logistista regressiomallia

$$\mathcal{M}_0: \qquad \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \operatorname{logit}(\pi(x_i)) = \beta_0.$$

Vastaavasti voidaan osoittaa, että log-lineaarinen yhdysvaikutusmalli $\mathcal{M}(XY)$ vastaan $I \times 2$-ristiintaulukon tilanteessa logistista regressiomallia

$$\mathcal{M}_1: \qquad \log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = \operatorname{logit}(\pi(x_i)) = \beta_0 + \beta_i^x.$$

## 6.3 Log-lineaariset mallit kolmeulotteisissa ristiintaulukoissa

Olkoon $X$,$Z$ ja $Y$ järjestys- tai luokitteluasteikkollisia muuttujia, joilla on $I$, $J$ ja $K$ toisensa poissulkevaa tulosvaihtoehtoa. Merkitään todennäköisyyksillä

$$P(X = x_i, Z = z_j, Y = y_k) = \pi_{ijk} \tag{6.7}$$

satunnaismuuttujien $X$, $Z$ ja $Y$ yhteistodennäköisyysjakaumaa. Olkoon $n_{ijk}$ solufrekvenssit muuttujien $X$,$Z$ ja $Y$ muodostamassa $I \times J \times K$-ristiintaulukossa, missä kolmeulotteista satunnaiskoetta on toistettu $n_{+++}$ kertaa.

Oletetaan, että solufrekvenssit $n_{ijk}$ noudattavat Poissonin jakaumaa $n_{ijk} \sim Poi(\mu_{ijk})$ ja mallinnetaan odotettuja frekvenssejä $\mu_{ijk}$ Poissonin log-lineaarisella mallilla.

Tarkastellaan ensiksi tilannetta, että muuttujat $X$, $Z$ ja $Y$ ovat keskenään riippumattomia. Eli testataan hypoteesia:

$$H_0 : \pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k} \qquad \text{kaikille } i, j \text{ ja } k. \tag{6.8}$$

$H_0$ hypoteesin vallitessa frekvenssien $n_{ijk}$ odotetut frekvenssit $\mu_{ijk}$ ovat muotoa

$$\mu_{ijk} = n_{+++}\pi_{i++}\pi_{+j+}\pi_{++k}. \tag{6.9}$$

Täten ottamalla odotetuista frekvensseistä $\mu_{ijk}$ logaritmit, saadaan muodostettua Poissonin log-lineaarinen päävaikutusmalli

$$\begin{aligned} \mathcal{M}(X, Z, Y) : \qquad \log(\mu_{ijk}) &= \log(n_{+++}) + \log(\pi_{i++}) + \log(\pi_{+j+}) + \log(\pi_{++k}) \\ &= \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y. \end{aligned}$$

Mikäli muuttujien $X$, $Z$ ja $Y$ välillä on jokaisen muuttujan suhteen riippuvuutta, voidaan muuttujien välistä riippuvuutta mallintaa Poissonin log-lineaarisella yhdysvaikutusmallilla

$$\mathcal{M}(XZY) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz} + \beta_{ik}^{xy} + \beta_{jk}^{zy} + \beta_{ijk}^{xzy}.$$

Päävaikutusmallin $\mathcal{M}(X, Z, Y)$ ja yhdysvaikutusmallin välille voidaan luoda monenlaisia $X$:n, $Z$:n ja $Y$:n välisiä riippuvuuksia kuvaavia malleja. Esimerkiksi tarkastellaan hypoteesia, että $Y$ on riippumaton $X$:stä, $Z$:sta, eli

$$H_0 : \pi_{ijk} = \pi_{ij+}\pi_{++k} \qquad \text{kaikille } i, j \text{ ja } k. \tag{6.10}$$

Tällöin Poissonin log-lineaarinen malli solufrekvenssien odotusarvoille $\mu_{ijk}$ on muotoa

$$\mathcal{M}(XZ, Y) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz}.$$

Mallissa $\mathcal{M}(XZ, Y)$ parametrit $\beta_{ij}^{xz}$ kuvaavat, että muuttujien $X$ ja $Z$ välillä voi olla riippuvuutta, mutta muita yhdysvaikutustermejä mallissa ei ole.

Voidaan myös ajatella, että $X$ ja $Y$ ovat keskenään ehdollisesti riippumattomia annetulla $Z = z_j$ arvolla. Merkitään ehdollisia todennäköisyyksiä seuraavasti:

$$P(X = x_i, Y = y_k | Z = z_j) = \pi_{ik|j}. \tag{6.11}$$

Jos $X$ ja $Y$ ovat keskenään ehdollisesti riippumattomia annetulla $Z = z_j$ arvolla, niin tällöin on voimassa hypoteesi

$$H_0 : \pi_{ik|j} = \pi_{i+|j}\pi_{+k|j} \qquad \text{kaikille } i, \ j \text{ ja } k. \tag{6.12}$$

$H_0$ hypoteesi on yhtä kuin hypoteesi

$$H_0 : \pi_{ijk} = \frac{\pi_{ij+}\pi_{+jk}}{\pi_{+j+}} \qquad \text{kaikille } i, \ j \text{ ja } k. \tag{6.13}$$

$H_0$ hypoteesia vastaava log-lineaarinen malli on nyt muotoa

$$\mathcal{M}(XZ, ZY) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz} + \beta_{jk}^{zy}.$$

Mikäli muuttujat $X$, $Z$ ja $Y$ voivat kaikki olla pareittain keskenään riippuvaisia, log-lineaarinen malli on silloin muotoa

$$\mathcal{M}(XZ, XY, ZY) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz} + \beta_{ik}^{xy} + \beta_{jk}^{zy}.$$

Mallia $\mathcal{M}(XZ, XY, ZY)$ kutsutaan homogeenisesti riippuvaiseksi malliksi.

Testataan hypoteesia, että $Y$ on riippumaton $X$:stä, $Z$:sta, eli

$$H_0 : \pi_{ijk} = \pi_{ij+}\pi_{++k} \qquad \text{kaikille } i, \ j \text{ ja } k. \tag{6.14}$$

$H_0$ hypoteesin vallitessa Poissonin log-lineaarinen malli solufrekvenssien odotusarvoille $\mu_{ijk}$ on muotoa

$$\mathcal{M}(XZ, Y) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz}.$$

Mikäli $H_0$ hypoteesi ei ole voimassa, voidaan ajatella, että solufrekvenssien odotusarvot $\mu_{ijk}$ noudattavat Poissonin log-lineaarista yhdysvaikutusmallia

$$\mathcal{M}(XZY) : \qquad \log(\mu_{ijk}) = \beta_0 + \beta_i^x + \beta_j^z + \beta_k^y + \beta_{ij}^{xz} + \beta_{ik}^{xy} + \beta_{jk}^{zy} + \beta_{ijk}^{xzy}.$$

Tällöin $H_0$ hypoteesi vastaa hypoteesia

$$H_0 : \beta_{ik}^{xy} = 0, \ \beta_{jk}^{zy} = 0, \ \beta_{ijk}^{xzy} = 0. \tag{6.15}$$

$H_0$ hypoteesia voidaan tällöin testata laskemalla mallien $\mathcal{M}(XZ, Y)$ ja $\mathcal{M}(XZY)$ devianssien erotus

$$D(\mathcal{M}(XZ, Y)) - D(\mathcal{M}(XZY)), \tag{6.16}$$

ja tutkimalla voisiko devianssien erotus noudattaa $\chi^2$-jakaumaa, jonka vapausasteet ovat mallien $\mathcal{M}(XZ, Y)$ ja $\mathcal{M}(XZY)$ parametrien lukumäärien erotus.

Jos kuitenkin tiedetään jo etukäteen, ettei muuttujien $X$, $Z$ ja $Y$ välillä ei ole voimassa kolmen muuttujan yhteinen yhdysvaikutus, eli mallissa $\mathcal{M}(XZY)$ parametrit $\beta_{ijk}^{xzy}$ ovat nollia, niin silloin mallia $\mathcal{M}(XZ, Y)$ voidaan verrata homogeenisesti riippuvaan malliin $\mathcal{M}(XZ, XY, ZY)$. Eli hypoteesi

$$H_0 : \pi_{ijk} = \pi_{ij+}\pi_{++k} \qquad \text{kaikille } i, j \text{ ja } k. \tag{6.17}$$

vastaa hypoteesia

$$H_0 : \beta_{ik}^{xy} = 0, \beta_{jk}^{zy} = 0, \tag{6.18}$$

ja täten $H_0$ hypoteesi voidaan testata laskemalla mallien $\mathcal{M}(XZ, Y)$ ja $\mathcal{M}(XZ, XY, ZY)$ devianssien erotus

$$D(\mathcal{M}(XZ, Y)) - D(\mathcal{M}(XZ, XY, ZY)). \tag{6.19}$$

## 6.4 Järjestysasteikolliset muuttujat

Olkoon muuttujat $X$ ja $Y$ järjestysasteikollisia. Tällöin $X$:n ja $Y$:n tulosvaihtoehdot $i = 1, 2, \ldots, I$ ja $j = 1, 2, \ldots, J$ voidaan järjestää esimerkiksi nousevaan järjestykseen. Olkoon nyt $u_1 \leq u_2 \leq \cdots \leq u_I$ muuttujan $X$ tulosvaihtoehdoille $i = 1, 2, \ldots, I$ määriteltyjä lukuarvoja, ja vastaavasti olkoon $v_1 \leq v_2 \leq \cdots \leq v_J$ muuttujan $Y$ tulosvaihtoehdoille $j = 1, 2, \ldots, J$ määriteltyjä lukuarvoja.

Tarkastellaan muuttujien $X$ ja $Y$ välistä lineaarista riippuvuutta. Mikäli muuttujat ovat riippumattomia, $I \times J$-ristiintaulukon solufrekvenssien odotusarvoja $\mu_{ij}$ voidaan tällöin mallintaa Poissonin log-lineaarisella päävaikutusmallilla

$$\mathcal{M}(X, Y) : \qquad \log(\mu_{ij}) = \beta_0 + \beta_i^x + \beta_j^y.$$

Mikäli muuttujien $X$ ja $Y$ välillä ajatellaan olevan lineaarista riippuvuutta, voidaan lukuarvojen $u_i$ ja $v_j$ avulla mallintaa muuttujien välistä mahdollista lineaarista riippuvuutta. Lineaarisen riippuvuuden tilanteessa oletetaan, että $I \times J$-ristiintaulukon solufrekvenssien odotusarvot $\mu_{ij}$ muodostuvat Poissonin log-lineaarisesta mallista

$$\mathcal{M}(X_L Y_L) : \qquad \log(\mu_{ij}) = \beta_0 + \beta_1 u_i v_j + \beta_i^x + \beta_j^y.$$

Yllä olevaa malli kutsutaan lineaarisen riippuvuuden malliksi.

Muuttujien $X$ ja $Y$ välistä lineaarista riippuvuutta voidaan nyt testata mallin $\mathcal{M}(X_L Y_L)$ avulla tarkastelemalla hypoteesia

$$\begin{aligned} H_0 &: \beta_1 = 0, \\ H_a &: \beta_1 \neq 0 \end{aligned} \tag{6.20}$$

$H_0$ hypoteesin testaus voidaan suorittaa joko Waldin testin $\hat{\beta}_1 / \hat{\sigma}(\hat{\beta}_1)$ avulla tai laskemalla devianssien erotus malleista $\mathcal{M}(X, Y)$ ja $\mathcal{M}(X_L Y_L)$:

$$D(\mathcal{M}(X, Y)) - D(\mathcal{M}(X_L Y_L)). \tag{6.21}$$