Jyrki Möttönen

Yleistetyt lineaariset mallit

Helsingin yliopisto
Sosiaalitieteiden laitos
2014

Sisältö

Luku 1

Johdanto

1.1 Tilastollisista malleista

Miksi mallitetaan?

Halutaan tutkia **altisteen** x (vaaratekijä, selittäjä, riippumaton muuttuja, syy) vaikutusta **vasteeseen** y (selitettävä muuttuja, riippuva muuttuja, seuraus):

- tieto sinänsä, ymmärtäminen
- ennustaminen
- manipulointi, ehkäisy

Esimerkki 1.1.1
Halutaan tutkia äidin raskaudenaikaisen tupakoinnin vaikutusta lapsen syntymäpainoon.

Esimerkki 1.1.2
Halutaan mallittaa lapsen syntymäpainoa, kun selittävinä tekijöinä ovat äidin ikä, pariteetti (elossa syntyneiden lasten lukumäärä) ja lapsen sukupuoli.

Esimerkki 1.1.3
Halutaan tutkia onko vajaaperheisyydellä vaikutusta nuorten poikien säännölliseen alkoholin käyttöön.

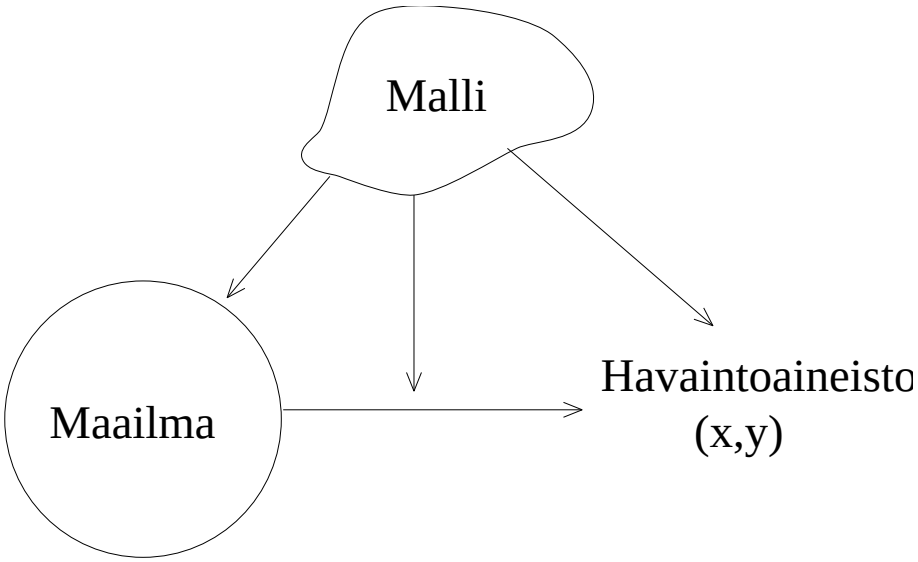
Esimerkki 1.1.4
Halutaan selittää tietyn tyyppisten loisten lukumäärää ahvenessa, kun selittäjinä on pyyntipaikka, pyyntiaika, kalan pituus ja kalan sukupuoli.

- koetulos ('vaste') $\boldsymbol{y}$
- koeolosuhteet ('selittäjävektori') $\boldsymbol{x}_i$
- satunnaisuus ('virhetermi') $\boldsymbol{\epsilon}$.

Usein:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} 1 & x_{12} & \cdots & x_{1p} \\ 1 & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix} \quad \text{ja} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

Tilastollinen malli kuvailee mahdollisimman uskottavasti sitä prosessia, joka tuottaa havaintoaineiston:



Kuva 1.1: Tilastollinen malli

Mallin valinta

"Where do probability models come from? To judge by the resounding silence over this question on the part of most statisticians, it seems highly embarassing. In general, the theoretician is happy to accept that his abstract probability triple (Ω, A, P) was found under a gooseberry bush, while the applied statistician's model "just growed"' A.P.Dawid (1982).

Ei ole olemassa mitään yleispätevää menetelmää tilastollisen mallin valintaan. Seuraavassa muutamia mallinnukseen kuuluvia perusperiaatteita:

- "A first, though at first sight, not a very helpful principle, is that all models are wrong; some, though, are more useful than others and we should seek those." McCullagh and Nelder (1989).

  Meidän on hyväksyttävä se väistämätön tosiasia, että me ei pystytä löytämään "ikuista totuutta". Mallit ovat aina todellisen maailman yksinkertaistuksia.

- "A second principle (which applies also to artists!) is not to fall in love with one model to the exclusion of alternatives." McCullagh and Nelder (1989).

  Aineisto voi antaa aiheen useiden erilaisten tilastollisten mallien kokeilemiseen. Ei pidä sulkea pois muita mahdollisia malleja.

- "A third principle recommends thorough checks on the fit of a model to the data, for example by using residuals and other statistics derived from the fit to look for outlying observations and so on." McCullagh and Nelder (1989).

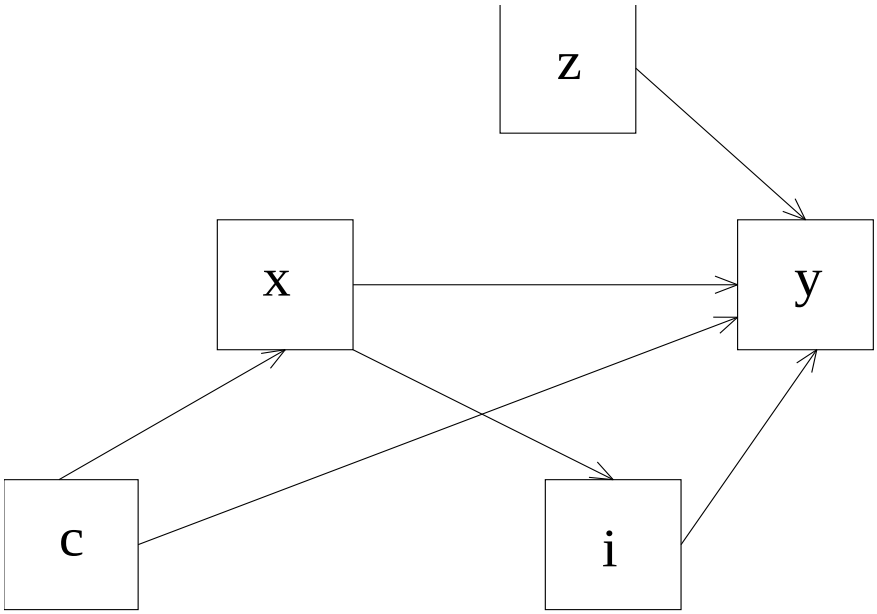
  On erittäin tärkeää tutkia kuinka hyvin malli sopii havintoaineistoon esimerkiksi residuaaliplottien ja erilaisten tilastollisten tunnuslukujen avulla.

- Valitse malli, jonka matemaattinen muoto sopii mallinnettavaan vasteeseen

  Esim. Monet tutkijat ovat käyttäneet tavallista lineaarista regressiomallia mallittamaan binäärisiä (0/1) vasteita. Sellaisilla malleilla estimoidut todennäköisyydet voivat olla välin $[0, 1]$ ulkopuolella!!

1.2 Selittävien tekijöiden valinta

Sekoittavat tekijät (c), väliintulevat (i) tekijät sekä riippumattomat syytekijät (z) tutkittaessa altisteen (x) vaikutusta vasteeseen (y):
Mitä tekijöitä on otettava malliin mukaan, tutkittaessa, onko x:llä vaikutusta y:hyn? Tutkija vai tietokone (askeltavat analyysit) valitsee?



Kuva 1.2: Selittävät tekijät

Muita usein esiintulevia kysymyksiä:

- selittävien tekijöiden laatu: jatkuvat, dikotomiset, luokitellut
- yhdysvaikutukset (vaikutus erilaista sekoittavien tekijöiden eri tasoilla; sekoittava tekijä 'muovaa' päätekijän vaikutusta)
- puuttuvien tietojen käsittely (poisjättäminen, korvaaminen ('imputointi'), ottaminen mallituksessa huomioon)

Eri päättelytilanteet

- Jatkuva vaste:
$$y_i = h(\boldsymbol{x}_i) + \epsilon_i, \quad i = 1, \ldots, n$$
jossa y_i:n odotusarvo riippuu "olosuhteista" $\boldsymbol{x}_i$, jatkuvat (virhe)muuttujat ϵ_i ovat riippumattomia ja samoin jakautuneita tiheysfunktiolla $f(e)$.
- Dikotominen vaste: y_i noudattaa Bernoullijakaumaa $Bin(1, p_i)$, jossa p_i riippuu "olosuhteista" $\boldsymbol{x}_i$ s.e.
$$p_i = h(\boldsymbol{x}_i), \quad i = 1, \ldots, n.$$
y_i:t ovat riippumattomia.

- Lukumäärävaste: y_i noudattaa Poissonjakaumaa $Poi(\lambda_i)$, jossa λ_i riippuu "olosuhteista" $\boldsymbol{x}_i$ s.e.

$$\lambda_i = h(\boldsymbol{x}_i), \quad i = 1, \ldots, n.$$

  y_i:t riippumattomia.

- Elinaikavaste: Elinaikaan y_i liittyvä vaarafunktio (hazard)

$$h_i(y) = f_i(y)/(1 - F_i(y))$$

  muotoa

$$h_i(y) = e^{h(\boldsymbol{x}_i)} h_0(y)$$

  (suhteellisen vaaran malli, proportional hazards model) tai

$$F_i(y) = F(e^{-h(\boldsymbol{x}_i)} y)$$

  (nopeutetun elämän malli, accelerated life model, accelerated failure time model). y_i:t riippumattomia. Elinaikavasteen tapausta ei käsitellä tällä kurssilla

1.3 Yleistetyt lineaariset mallit

Yhteisiä piirteitä yo. mallitustilanteissa:

1. *Satunnaisosa:* Kuvaa prosessiin liittyvää satunnaisuutta. Riippumattomat samaa jakaumaa noudattavat satunnaismuuttujat y_i odotusarvoilla μ_i sekä mahdollisesti tuntemattomalla hajontaparametrilla σ.

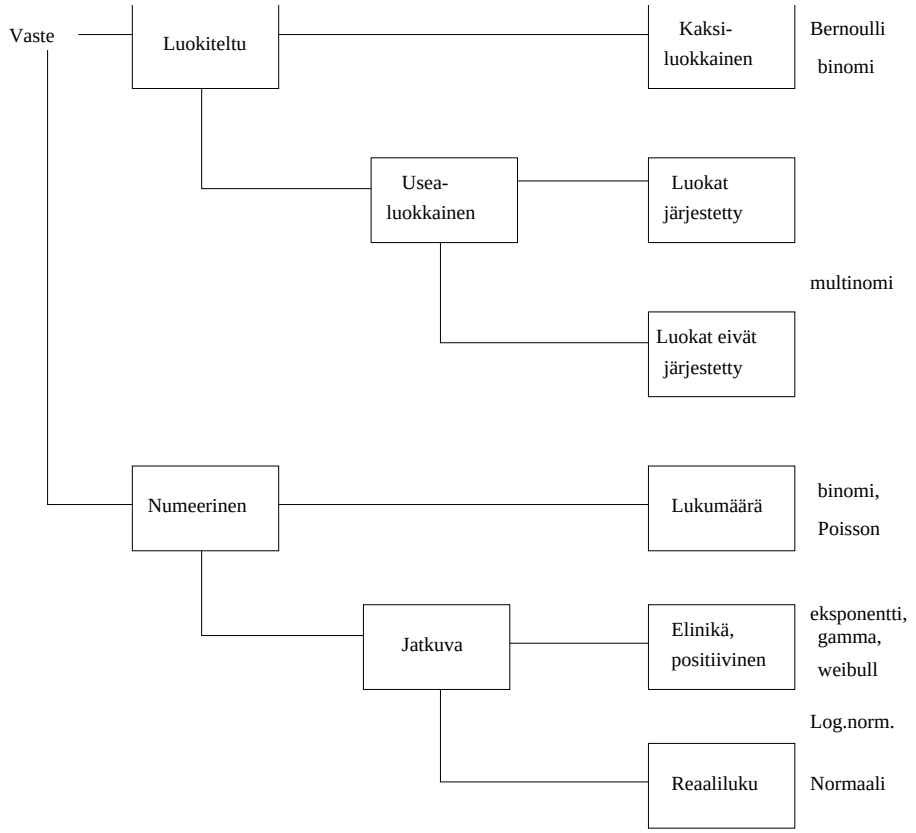
2. *Systemaattinen osa:* Kuvaa prosessin kiinteää osaa. Selittäjävektoreihin $\boldsymbol{x}_i$ liittyvät ns. **lineaariset ennusteet** (linear predictor)

$$\nu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$$

3. *Yhteys* (link) satunnais- ja systemaattisen osan välillä:

$$\nu_i = g(\mu_i) \quad (\text{tai } \mu_i = h(\nu_i) \ ),$$

   missä g on ns. **linkkifunktio**.



Kuva 1.3: Vasteen jakauma

HUOM. Useimmissa tarkastelluissa tapauksissa (normaalijakauma, Poisson-jakauma, binomijakauma, eksponenttijauma, gammajakauma, jne) vastemuuttujan jakauma kuuluu niin sanottuun **eksponentiaaliseen perheeseen**, ja tiheys- tai todennäköisyysfunktio on muotoa

$$f(y;\theta,\phi) = \exp\Big\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\Big\}$$

tietyillä funktioilla

$$a(\cdot), \quad b(\cdot) \quad \text{ja} \quad c(\cdot,\cdot).$$

Tällöin voidaan osoittaa, että

$$E(Y) = \mu = b'(\theta) \quad \text{ja} \quad Var(Y) = \sigma^2 = b''(\theta)a(\phi).$$

Tavallisimpia linkkifunktion valintoja:

- Identtinen linkki: $g(\mu) = \mu$ (normaalinen vaste)
- Logaritmilinkki $g(\mu) = \log(\mu)$ (Poissonvaste)
- Logitlinkki $g(\mu) = \text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$ (Binomivaste)
- Probitlinkki $g(\mu) = \Phi^{-1}(\mu)$ (Binomivaste)
- Complementary log-log-linkki $g(\mu) = \log\{-\log(1-\mu)\}$ (Binomivaste)

Linkkifunktion valinta: riippuvuuden laatu, selittäjien laatu (jatkuva, dikotominen, luokiteltu), laskettavuus, luonnollisuus. Linkkifunktion valinta vaikuttaa luonnollisesti parametrien tulkintaan.

Parametria $\boldsymbol{\beta}$ (ja ϕ) koskeva päättely: uskottavuuspäättely (suurimman uskottavuuden estimointi, uskottavuusosamäärätestit ja vastaavat luottamusvälit).

Luku 2

Uskottavuuspäättely

Halutaan tehdä päätelmiä, jotka koskevat tuntematonta **parametrivektoria** $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T$. Parametrin $\boldsymbol{\theta}$ mahdollisten arvojen muodostamaa joukkoa $\Omega \subset \mathbb{R}^k$ kutsutaan **parametriavaruudeksi** (parameter space).

Tehdään **koe**, jonka tulos $\boldsymbol{y}$ "sisältää parametria $\boldsymbol{\theta}$ koskevaa informaatiota". Koetulokseen $\boldsymbol{y}$ liittyvä todennäköisyys (tai tiheysfunktio)

$$f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta})$$

riippuu (tunnetulla tavalla) parametrista $\boldsymbol{\theta}$.

Havaittua koetulosta $\boldsymbol{y}$ vastaava **uskottavuusfunktio**

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{y}) = f_{\boldsymbol{Y}}(\boldsymbol{y}; \boldsymbol{\theta}), \ \boldsymbol{\theta} \in \Omega$$

kertoo (karkeasti tulkiten) "todennäköisyyden, jolla $\boldsymbol{\theta}$ tuottaa nyt saadun havainnon $\boldsymbol{y}$". **Logaritminen uskottavuusfunktio** on

$$l(\boldsymbol{\theta}) = l(\boldsymbol{\theta}; \boldsymbol{y}) = l(\theta_1, \ldots, \theta_k) = \log L(\boldsymbol{\theta}).$$

2.1 Suurimman uskottavuuden estimaatti

Suurimman uskottavuuden estimaatti (MLE)

$$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T \quad \in \Omega$$

on se parametriarvojen muodostama vektori, joka maksimoi (logaritmisen) uskottavuusfunktion.

Edelleen **pistemääräfunktio** (score function) on vektoriarvoinen funktio $\boldsymbol{s}(\boldsymbol{\theta})$, jonka j:s komponentti on

$$s_j(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_j} l(\boldsymbol{\theta}), \quad j = 1, \ldots, k,$$

ja **havaittu informaatio** $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})$ on matriisiarvoinen funktio, jonka (r, s)-komponentti on

$$j_{rs}(\boldsymbol{\theta}) = -\frac{\partial^2}{\partial \theta_r \partial \theta_s} l(\boldsymbol{\theta}).$$

Fisherin informaatio $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ saadaan ottamalla odotusarvot alkioittain matriisisista $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})$, ts. $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$ on matriisi, jonka (r, s)-komponentti on $i_{rs}(\boldsymbol{\theta}) = \mathrm{E}(j_{rs}(\boldsymbol{\theta}; \boldsymbol{y}))$

Suurimman uskottavuuden estimaatti (MLE) löytyy ratkaisemalla samanaikaisesti k yhtälöä

$$S_j(\boldsymbol{\theta}) = 0, \quad j = 1, \ldots, k$$

tai numeerisesti esimerkikiksi Newtonin-Raphsonin menetelmän avulla. Newtonin-Raphsonin menetelmä esitellään liitteessä A.1.

Esimerkki 2.1.1 Normaalijakauma, yksi otos
Olkoon $y_1, \ldots, y_n$ satunnaisotos normaalijakaumasta $N(\mu, \theta)$; μ ja $\theta > 0$ ovat tuntemattomia. Nyt

$$l(\mu, \theta) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\theta) - \frac{1}{2\theta} \sum_{i=1}^{n} (y_i - \mu)^2,$$

ja suurimman uskottavuuden estimaatti on

$$\hat{\mu} = \bar{y} \;\text{ ja }\; \hat{\theta} = s^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

Havaittu informaatiomatriisi on

$$\hat{\boldsymbol{\mathcal{J}}} = \mathrm{diag}\Big(\frac{n}{s^2}, \frac{n}{2(s^2)^2}\Big) = \begin{pmatrix} \frac{n}{s^2} & 0 \\ 0 & \frac{n}{2(s^2)^2} \end{pmatrix}.$$

Esimerkki 2.1.2 Normaalijakauma, kaksi otosta
Olkoon

- $x_1, \ldots, x_m$ satunnaisotos normaalijakaumasta $N(\mu_1, \theta)$
- $y_1, \ldots, y_n$ satunnaisotos normaalijakaumasta $N(\mu_2, \theta)$
- otokset riippumattomia

Silloin

$$l(\mu_1, \mu_2, \theta) = -\frac{m+n}{2}\log(2\pi) - \frac{m+n}{2}\log(\theta) - \frac{1}{2\theta}\sum_{i=1}^{m}(x_i - \mu_1)^2 - \frac{1}{2\theta}\sum_{j=1}^{n}(y_j - \mu_2)^2.$$

Suurimman uskottavuuden estimaateiksi saadaan

$$\hat{\mu}_x = \bar{x}, \quad \hat{\mu}_y = \bar{y}$$

ja

$$\hat{\theta} = s^2 = \frac{1}{m+n}\Big[\sum_{i=1}^{m}(x_i - \bar{x})^2 + \sum_{j=1}^{n}(y_j - \bar{y})^2\Big].$$

Havaittu informaatiomatriisi on

$$\hat{\boldsymbol{\mathcal{J}}} = \operatorname{diag}\Big(\frac{m}{s^2}, \frac{n}{s^2}, \frac{m+n}{2(s^2)^2}\Big) = \begin{pmatrix} \frac{m}{s^2} & 0 & 0 \\ 0 & \frac{n}{s^2} & 0 \\ 0 & 0 & \frac{m+n}{2(s^2)^2} \end{pmatrix}.$$

2.2 Jakaumatuloksia

Uskottavuusfunktio, logaritminen uskottavuusfunktio, jne., ovat **satunnaisia**; niiden arvot vaihtelevat kokeesta toiseen. Edelleen luonnollisesti näistä satunnaisfunktioista johdettavat suureet, kuten suurimman uskottavuuden estimaattivektori ja Fisherin informaatiomatriisi, ovat **satunnaismuuttujia** (satunnaisvektoreita, satunnaismatriiseja), joiden jakauman yksikäsitteisesti määrää koetuloksen $\boldsymbol{y}$ jakauma. Suureiden jakaumat siis riippuvat tuntemattomasta parametrista $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T \in \Omega$.

Tuloksia: Tarkastellaan satunnaisfunktioita $\boldsymbol{s}(\boldsymbol{\theta})$ ja $\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})$. Eräitten yleisten ehtojen vallitessa

$$E_{\boldsymbol{\theta}}(\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{Y})) = \boldsymbol{0} \quad \text{ja} \quad E_{\boldsymbol{\theta}}(\boldsymbol{s}(\boldsymbol{\theta}; \boldsymbol{Y})\boldsymbol{s}^T(\boldsymbol{\theta}; \boldsymbol{Y})) = E_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{J}}(\boldsymbol{\theta}; \boldsymbol{Y})) = \boldsymbol{\mathcal{I}}(\boldsymbol{\theta}).$$

Olkoon $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ satunnaisotos jakaumasta, jonka tiheysfuntio tai todennäköisyysfunktio riippuu tuntemattomasta k-ulotteisesta parametrista

$$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)^T.$$

Olkoon

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{y}) = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$$

koetulokseen $\boldsymbol{y}$ liittyvä suurimman uskottavuuden estimaatti ja $\hat{\boldsymbol{\mathcal{J}}}$ suurimman uskottavuuden ratkaisuun liittyvä havaittu informaatio. Suurimman uskottavuuden estimaatti ei välttämättä ole **harhaton** (jolloin $E_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ olipa $\boldsymbol{\theta}$ mikä tahansa), mutta eräiden yleisten ehtojen vallitessa se on **tarkentuva**, toisin sanoen,

$$P_{\boldsymbol{\theta}}(||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}|| > \epsilon) \to 0, \quad \forall \epsilon > 0, \quad \text{kun } n \to \infty$$

olipa $\boldsymbol{\theta} \in \Omega$ mikä tahansa ($\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^T\boldsymbol{x}}$ on vektorin $\boldsymbol{x}$ pituus). Edelleen eräiden yleisten ehtojen vallitessa suurilla otoskoilla likimain pätee:

$$\hat{\boldsymbol{\theta}} \sim N_k(\boldsymbol{\theta}, \hat{\boldsymbol{\mathcal{J}}}^{-1}).$$

Erityisesti likimain

$$\hat{\theta}_j \sim N(\theta_j, \hat{j}^{jj}), \quad j = 1, \ldots, k,$$

jossa $\hat{j}^{jj}$ on matriisin $\hat{\boldsymbol{\mathcal{J}}}^{-1}$ komponentti (j, j). Siis parametrin θ_j, $j = 1, \ldots, k$, likimääräinen $100(1-\alpha)$ %:n luottamusväli on muotoa

$$\hat{\theta}_j \ \pm \ z_{\alpha/2}\sqrt{\hat{j}^{jj}}.$$

Oletetaan seuraavaksi, että halutaan tehdä päätelmiä "uusista parametreista"

$$\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta}) \quad \text{tai} \quad \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_r \end{pmatrix} = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ \vdots \\ g_r(\boldsymbol{\theta}) \end{pmatrix}.$$

(HUOM: $r \leq k$; $\boldsymbol{\gamma}$:n dimensio voi siis olla aidosti pienempi kuin k.) Merkitään

$$q_{ij}(\boldsymbol{\theta}) = \frac{\partial g_i(\boldsymbol{\theta})}{\partial \theta_j} \quad \text{ja} \quad \boldsymbol{Q}(\boldsymbol{\theta}) = \Big(q_{ij}(\boldsymbol{\theta})\Big),$$

($r \times k$-matriisi). Jälleen (eräiden yleisten ehtojen vallitessa) isoilla otoskoilla n likimain pätee:

$$\hat{\boldsymbol{\gamma}} \sim N_r(\boldsymbol{\gamma}, \hat{\boldsymbol{Q}}\hat{\boldsymbol{\mathcal{J}}}^{-1}\hat{\boldsymbol{Q}}^T),$$

jossa

$$\hat{\boldsymbol{Q}} = \boldsymbol{Q}(\hat{\boldsymbol{\theta}}).$$

(**Delta-menetelmä**.)

2.3 Testaus

Tutkitaan seuraavaksi hypoteesia (väitettä)

$$H_0 : \quad \boldsymbol{\theta} = \boldsymbol{\theta}_0.$$

Niin sanotun **uskottavuusosamäärätestin** (likelihood ratio test) testisuure

$$r(\boldsymbol{y}) = 2 \log \frac{L(\hat{\boldsymbol{\theta}})}{L(\boldsymbol{\theta})} = 2[l(\hat{\boldsymbol{\theta}}) - l(\boldsymbol{\theta})] \sim \chi^2_k$$

likimain, kun H_0 tosi. Muita ('asymptoottisesti ekvivalentteja') testiversioita ovat (**Waldin testisuure**)

$$w(\boldsymbol{y}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \ \boldsymbol{\mathcal{J}}(\boldsymbol{\theta}_0) \ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \ \text{ tai } \ w(\boldsymbol{y}) = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \ \hat{\boldsymbol{\mathcal{J}}} \ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Niin sanottu **pistemäärätestisuure** (score test) on

$$u(\boldsymbol{y}) = \boldsymbol{s}^T(\boldsymbol{\theta}_0) \ \hat{\boldsymbol{\mathcal{J}}}^{-1} \ \boldsymbol{s}(\boldsymbol{\theta}_0).$$

Jokaisen neljän testisuureen rajajakauma on χ^2_k, kun $n \to \infty$ ja nollahypoteesi on tosi. Edelleen likimääräinen 95 %:n luottamusalue on

$$\{ \ \boldsymbol{\theta} \ : \ r(\boldsymbol{Y}) \leq \chi^2_k(0.95) \}$$

tai (kvadraattinen approksimaatio)

$$\{ \ \boldsymbol{\theta} \ : \ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T \ \hat{\boldsymbol{j}} \ (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \leq \chi^2_k(0.95) \}.$$

Merkitään seuraavaksi

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\theta}_1 \\ \boldsymbol{\theta}_2 \end{pmatrix},$$

missä $\boldsymbol{\theta}_1$ on r-vektori ja $\boldsymbol{\theta}_2$ on s-vektori ($r + s = k$). Oletaan, että halutaan tutkia väitteen (hypoteesin)

$$H_0 : \ \boldsymbol{\theta}_2 = \mathbf{0}$$

uskottavuutta. Väitteen suhteellinen uskottavuus on siis

$$\frac{L(\hat{\boldsymbol{\theta}}_1(\mathbf{0}), \mathbf{0})}{L(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2)},$$

missä $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^T, \hat{\boldsymbol{\theta}}_2^T)^T$ on $\boldsymbol{\theta}$:n suurimman uskottavuuden estimaatti ja $\hat{\boldsymbol{\theta}}_1(\mathbf{0})$ maksimoi funktion $L(\boldsymbol{\theta}_1, \mathbf{0})$. Jälleen eräiden yleisten ehtojen vallitessa ja kun väite $H_0 : \ \boldsymbol{\theta}_2 = \mathbf{0}$ on tosi, niin

$$r(\boldsymbol{y}) = 2[l(\hat{\boldsymbol{\theta}}_1, \hat{\boldsymbol{\theta}}_2) - l(\hat{\boldsymbol{\theta}}_1(\mathbf{0}), \mathbf{0})] \sim \chi^2_s$$

likimain, kun väite $H_0 : \ \boldsymbol{\theta}_2 = \mathbf{0}$ on tosi. Testisuurta $r(\boldsymbol{y})$ sanotaan **uskottavuusosamäärätestisuureeksi** testattaessa nollahypoteesia $H_0 : \ \boldsymbol{\theta}_2 = \mathbf{0}$.

2.4 Lisää uskottavuuspäättelystä

- Kokeiden yhdistäminen: Olkoot $L_1(\boldsymbol{\theta})$ ja $L_2(\boldsymbol{\theta})$ kahteen riippumattomaan kokeeseen liittyvät samaa tuntematonta parametria $\boldsymbol{\theta}$ koskevat uskottavuusfunktiot. Silloin 'yhdistetyssä' kokeessa

$$L(\boldsymbol{\theta}) = L_1(\boldsymbol{\theta}) \cdot L_2(\boldsymbol{\theta}) \ \text{ ja } \ l(\boldsymbol{\theta}) = l_1(\boldsymbol{\theta}) + l_2(\boldsymbol{\theta})$$

- Tunnusluku $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{y})$ on **tyhjentävä** (sufficient), jos

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \boldsymbol{y}) = L_1(\boldsymbol{\theta}; \boldsymbol{t}(\boldsymbol{y})) \cdot L_2(\boldsymbol{y}).$$

- **Tehokkuus:** Suurimman uskottavuuden testit ja estimaatit ovat asymptoottisesti optimaalisia.

- **Invarianttisuusominaisuus**: Parametrisointitapa ei vaikuta päättelytuloksiin.

2.5 Eksponentiaalinen perhe

Oletetaan, että y_i:n tiheys- tai todennäköisyysfunktio on

$$f(y_i;\theta_i,\phi) = \exp\Big\{\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\Big\}.$$

Oletetaan, että ϕ on tunnettu. Silloin havaintoon y_i liittyvä pisteluku- ja informaatiofunktio (θ_i tuntematon parametri) ovat

$$\frac{y_i - b'(\theta_i)}{a(\phi)} \quad \text{ja} \quad \frac{b''(\theta_i)}{a(\phi)}.$$

Koska

$$E_{\boldsymbol{\theta}}(\boldsymbol{s}(\boldsymbol{\theta})) = \mathbf{0} \quad \text{ja} \quad E_{\boldsymbol{\theta}}(\boldsymbol{s}(\boldsymbol{\theta})\boldsymbol{s}^T(\boldsymbol{\theta})) = E_{\boldsymbol{\theta}}(\boldsymbol{\mathcal{J}}(\boldsymbol{\theta})),$$

niin

$$E(y_i) = \mu_i = b'(\theta_i) \quad \text{ja} \quad Var(y_i) = \sigma_i^2 = b''(\theta_i)a(\phi).$$

Olkoon nyt

$$\nu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

lineaarinen prediktori, $i = 1,\ldots,n$. Havaintoaineistoon $y_1,\ldots,y_n$ liittyvä pistelukufunktio β_j:n suhteen on

$$\sum_{i=1}^{n}\Big\{\frac{y_i-\mu_i}{a(\phi)}\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\nu_i}x_{ij}\Big\}$$

Linkkiä, jonka vallitessa $\nu_i = \theta_i$ tai $g(\mu_i) = \theta_i$ kutsutaan **kanoniseksi linkiksi**. Tällöin $\boldsymbol{X}^T\boldsymbol{y}$ on tyhjentävä tunnusluku $\boldsymbol{\beta}$:lle ja

$$\frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\nu_i} = 1$$

Yleisesti: Jos

$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_i^2) \quad \text{ja} \quad \boldsymbol{D} = \mathrm{diag}\left(\frac{\partial\mu_i}{\partial\nu_i}\right),$$

niin pisteluku- ja informaatiofunktio ovat

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}(\boldsymbol{y}-\boldsymbol{\mu}) \quad \text{ja} \quad \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \mathrm{E}(\boldsymbol{\mathcal{J}}(\boldsymbol{\beta})) = \boldsymbol{X}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{D}^2\boldsymbol{X}$$

Luku 3

Jatkuva vaste

3.1 Lineaarinen regressioanalyysi

Aineisto:

$$\boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad \text{ja} \quad X = \begin{pmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}$$

Oletukset: y_i noudattaa normaalijakaumaa $N(\mu_i, \sigma^2)$ ja

$$\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}.$$

Muuttujat y_i riippumattomia.

Normaalijakauman $N(\mu, \sigma^2)$ tiheysfunktio:

$$f(y; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big\{ -\frac{(y-\mu)^2}{2\sigma^2} \Big\}$$

Siis havaintoaineistoon liittyvä logaritminen uskottavuusfunktio on

$$l(\boldsymbol{\beta}, \sigma^2) = \sum_{i=1}^{n} \log \; f(y_i; \mu_i, \sigma^2),$$

missä

$$\log f(y_i; \mu_i, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_i - \mu_i)^2}{2\sigma^2} = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

Parametrin $\boldsymbol{\beta}$ suurimman uskottavuuden estimaatti:

Suurimman uskottavuuden estimaatti (normaalijakaumatapauksessa pienimmän neliösumman estimaatti) saadaan minimoimalla lauseke

$$\sum (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 = (\boldsymbol{y} - X\boldsymbol{\beta})^T (\boldsymbol{y} - X\boldsymbol{\beta})$$

tai ratkaisemalla

$$X^T \boldsymbol{y} - (X^T X)\boldsymbol{\beta} = \mathbf{0}.$$

Ratkaisuna saadaan

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}.$$

Olkoon

$$H = X(X^T X)^{-1} X^T \quad \text{(ns. H-matriisi)}.$$

Silloin ennustevektori $\hat{\boldsymbol{\mu}}$ ja residuaalivektori $\hat{\boldsymbol{e}}$ voidaan kirjoittaa muodossa

$$\hat{\boldsymbol{\mu}} = X\hat{\boldsymbol{\beta}} = H\boldsymbol{y} \;\text{ ja }\; \hat{\boldsymbol{e}} = \boldsymbol{y} - X\hat{\boldsymbol{\beta}} = (I - H)\boldsymbol{y}.$$

Huomataan, että

$$\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2 (X^T X)^{-1})$$

ja

$$\hat{\boldsymbol{\mu}} \sim N_n(\boldsymbol{\mu}, \sigma^2 H) \;\text{ ja }\; \hat{\boldsymbol{e}} \sim N_n(\mathbf{0}, \sigma^2 (I - H)),$$

missä $\boldsymbol{\mu} = X\boldsymbol{\beta}$ ($\hat{\boldsymbol{\mu}}$ ja $\hat{\boldsymbol{e}}$ ovat riippumattomia). Parametrin σ^2 (harhaton) estimaattori on

$$s^2 = \frac{1}{n-p} \hat{\boldsymbol{e}}^T \hat{\boldsymbol{e}} = \frac{1}{n-p} \boldsymbol{y}^T (I - H)\boldsymbol{y}.$$

Varianssin σ^2 suurimman uskottavuuden estimaatti on (harhainen)

$$\hat{\sigma}^2 = \frac{1}{n} \boldsymbol{y}^T (I - H)\boldsymbol{y}.$$

Estimaatit, luottamusvälit, testit

Nyt

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2 (X^T X)^{-1}_{jj})$$

ja

$$\frac{\hat{\beta}_j - \beta_j}{s\sqrt{(X^T X)^{-1}_{jj}}} \sim t(n-p)$$

(likimain $N(0,1)$-jakautunut), ja likimääräinen 95 %:n luottamusväli on muotoa

$$\hat{\beta}_j \ \pm \ 1.96 \cdot s\sqrt{(X^T X)^{-1}_{jj}}.$$

Testattaessa nollahypoteesia $H_0 : \beta_j = 0$ "optimaalinen"testisuure on nollahypoteesin vallitessa likimain $N(0,1)$-jakautunut (tarkka jakauma $t(n-p)$)

$$t = \frac{\hat{\beta}_j}{s\sqrt{(X^T X)^{-1}_{jj}}}.$$

Parametrin β_j tulkinta (identtinen linkki): Oletetaan ensin, että selittäjä x_j on dikotominen (0/1)-muuttuja. Silloin β_j kuvaa muutosta vasteen odotusarvossa, kun siirrytään muuttujan x_j luokasta 0 luokkaan 1 (muiden selittävien muuttujien pysyessä arvoiltaan samoina).

Kun muuttuja x_j on jatkuva, β_j kertoo muutoksen vasteen odotusarvossa x_j:n kasvaessa yhden mittayksikkönsä verran (ja muiden selittäjien pysyessä arvoiltaan samoina.)

Huom. Käytettäessä logaritmilinkkiä

$$\mu_i = \exp(\beta_1 x_{i1} + \ldots \beta_p x_{ip}) = \exp(\beta_1 x_{i1}) \ldots \exp(\beta_p x_{ip})$$

ja siis nyt $\exp(\beta_j)$ kertoo, **monikokertaiseksi** vasteen odotusarvo muuttuu selittäjän x_j arvon muuttuessa yhden mittayksikön verran!

Sisäkkäisten mallien vertailu

Logaritmisen uskottavuusfunktion maksimiarvo on

$$l(\hat{\beta}, \hat{\sigma}^2) = -\frac{n}{2}\log(2\pi\hat{\sigma}^2) - \frac{n}{2}$$

ja se riippuu havainnoista vain residuaalien neliöiden summan

$$\boldsymbol{y}^T(I - H)\boldsymbol{y}$$

kautta. (GLIM-ohjelmistossa tästä residuaalien neliöiden summasta käytetään nimitystä **devianssi**).

Tarkastellaan kahta "sisäkkäistä"mallia M_1 ja M_2 ja oletetaan, mallissa M_1 on r selittäjää ja M_2 on saatu M_1:stä lisäämällä s uutta selittäjää. Jos H_1 ja H_2 ovat malleja vastaavat H-matriisit, niin devianssin muutos on $\boldsymbol{y}^T(H_1 - H_2)\boldsymbol{y}$, ja "suhteellinen muutos"

$$\frac{\boldsymbol{y}^T(H_2 - H_1)\boldsymbol{y}/s}{\boldsymbol{y}^T(I - H_2)\boldsymbol{y}/(n - r - s)}$$

noudattaa F-jakaumaa vapausasteilla s ja $n - r - s$ (tai likimain $\chi^2(s)$-jakaumaa), kun malli M_1 "oikea".

Normaalisuuden tutkiminen

Mallin "hyvyyttä" tai oikeellisuutta on mahdollista tarkastella tutkimalla estimoituja residuaaleja. Olkoon

$$e_{(1)} < \ldots < e_{(n)}$$

residuaalit $e_1, \ldots, e_n$ järjestettynä pienimmästä suurimpaan. Jos residuaalit tulevat $N(\mu, \sigma^2)$-jakaumasta, niin

$$\mathrm{E}(e_{(i)}) \approx \mu + \sigma\Phi^{-1}\left(\frac{i}{n+1}\right), \quad i = 1, \ldots, n.$$

jossa $\Phi^{-1}(x)$ on $N(0,1)$-jakauman kertymäfunktion $\Phi(x)$ käänteisfunktio, ts. $\Phi^{-1}(p) = y \Leftrightarrow p = \Phi(y)$. Jos residuaalit tulevat normaalijakaumasta, niin pisteet

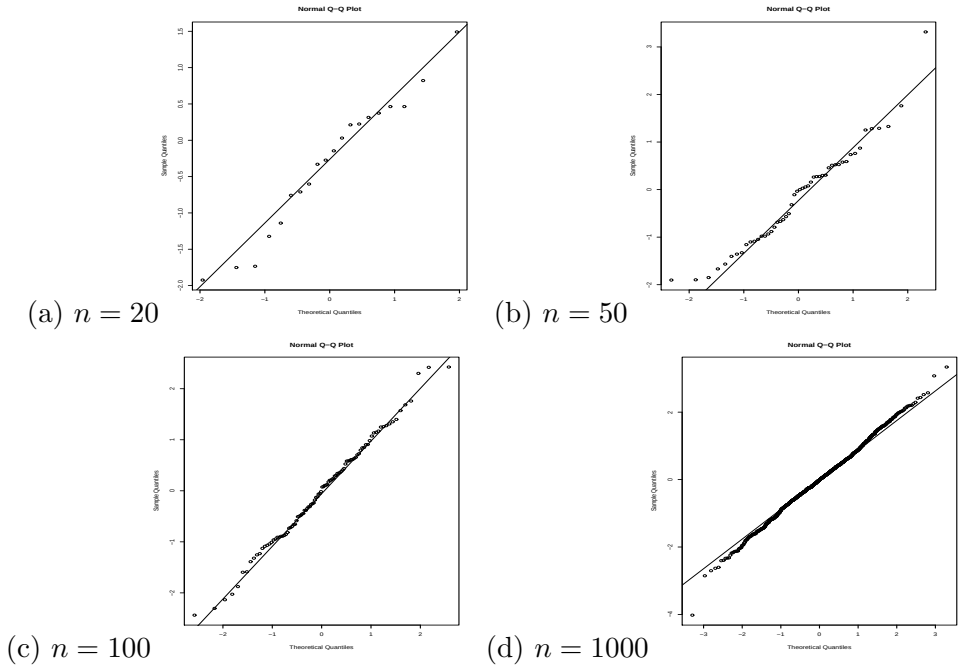
$$\left(\Phi^{-1}\left(\frac{i}{n+1}\right),\ e_{(i)}\right), \quad i = 1, \ldots, n$$

sijaitsevat likimain suoralla.

R-ohjelmiston komennolla `qqnorm(x)` tehdään "Q-Q plotti"eli kvantiilikuvio aineistolle `x`. Komennolla `qqline(x)` lisätään suora, joka kulkee alakavartiilien ja yläkvartiilien kautta.

Esim.

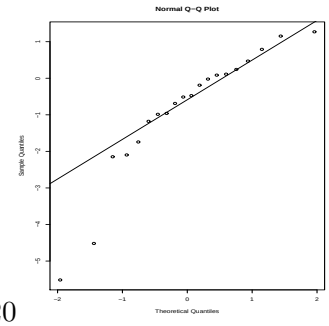
```
> x<-rnorm(20); qqnorm(x); qqline(x)
> x<-rt(20,df=2);qqnorm(x);qqline(x)
```



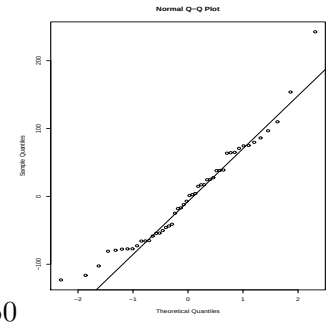
(a) $n = 20$ (b) $n = 50$

(c) $n = 100$ (d) $n = 1000$

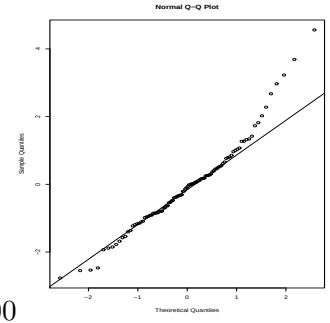
Kuva 3.1: Kvantiilikuvioita. Sat.otokset N(0,1)-jakaumasta.



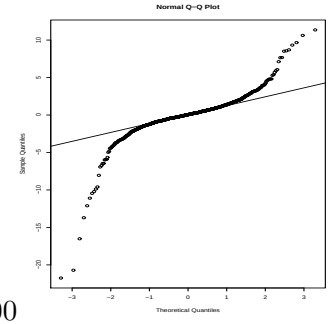
(a) $n = 20$



(b) $n = 50$

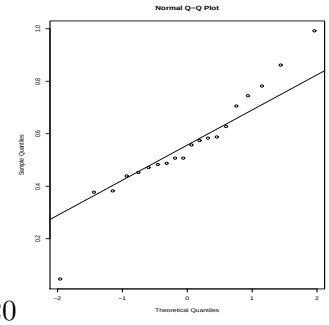


(c) $n = 100$

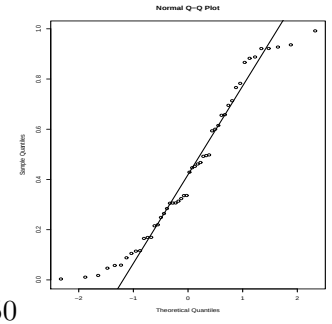


(d) $n = 1000$

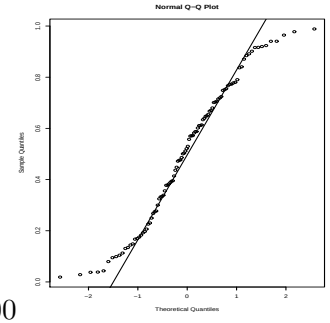
Kuva 3.2: Kvantiilikuvioita. Huipukkaampi kuin norm.jak. ($t(2)$).



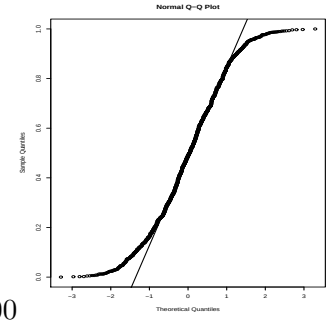
(a) $n = 20$



(b) $n = 50$

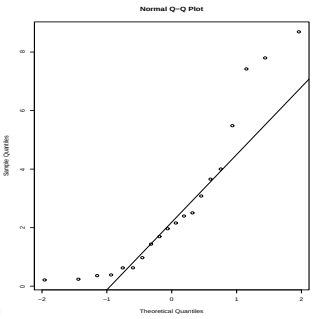


(c) $n = 100$

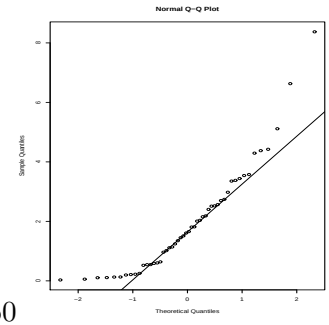


(d) $n = 1000$

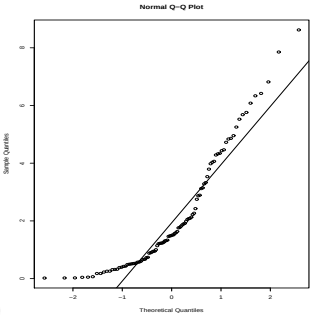
Kuva 3.3: Kvantiilikuvioita. Litteämpi kuin norm.jak. (tas(0,1)).



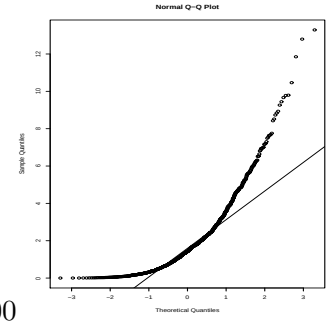
(a) $n = 20$



(b) $n = 50$

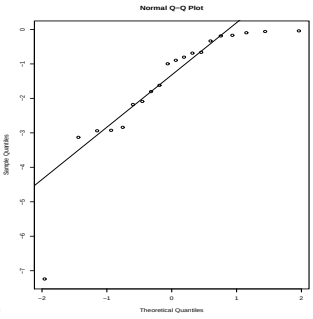


(c) $n = 100$

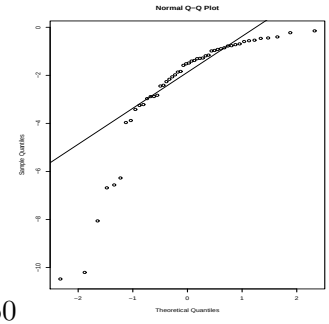


(d) $n = 1000$

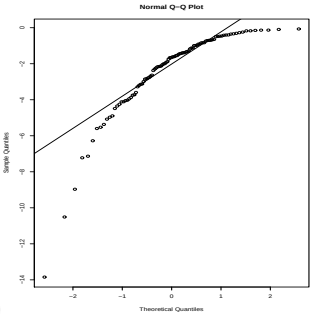
Kuva 3.4: Kvantiilikuvioita. Oikealle vino jakauma ($\chi^2(2)$).



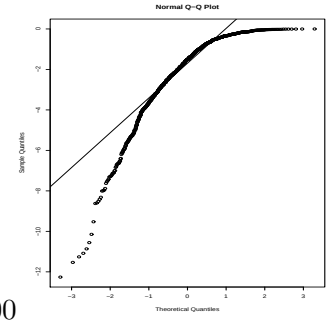
(a) $n = 20$



(b) $n = 50$



(c) $n = 100$



(d) $n = 1000$

Kuva 3.5: Kvantiilikuvioita. Vasemmalle vino jakauma ($(-1) \cdot \chi^2(2)$).

Sirontakuviot

Mallin "hyvyyttä" voidaan tarkastella myös sirontakuvioiden avulla. Sirontakuvioilla, joissa vastakkain ovat residuaalit ja ennusteet ($\hat{e}_i$ vs $\hat{y}_i$) tai residuaalit ja selittäjät ($\hat{e}_i$ vs x_{i2}, ..., $\hat{e}_i$ vs x_{ip}) tarkkaillaan mahdollista epälineaarista riippuvuutta, residuaalien varianssien käyttäytymistä $\boldsymbol{y}$:n, $\boldsymbol{x}_1$:n, ..., funktiona, jne.

Sirontakuviossa $\hat{y}_i$ vs. e_i pisteiden tulisi sijaita tasaisesti nollan molemmilla puolilla. Sirontakuviossa käytetään ennusteita $\hat{y}_i$ eikä alkuperäisiä havaintoja y_i. Tämä johtuu siitä, että e_i ja y_i ovat yleensä korreloituneita mutta e_i ja $\hat{y}_i$ ovat korreloimattomia.

Residuaalien e_i paikalla käytetään joskus *standardoituja residuaaleja* (standardized residual)

$$r_i = e_i / \sqrt{s^2(1 - h_i)}.$$

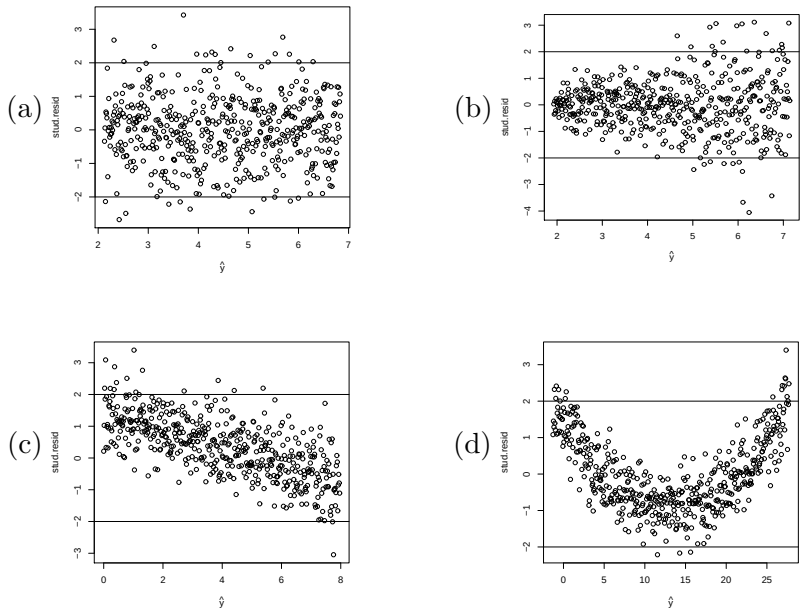
jossa $s^2 = \hat{\sigma^2}$ ja h_i on H-matriisin i:s diagonaalialkio. Huom! Nimittäjän valinta perustuu aiemmin annettuun tulokseen

$$\boldsymbol{e} \sim N_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H})).$$

Niin sanottujen *studentisoitujen residuaalien* (studentized residual, jacknife residual) avulla voidaan tutkia onko ainestossa poikkeavia havaintoja (outlier). Havaintoyksikköä i vastaava studentisoitu residuaali on

$$t_i = e_i / \sqrt{s(i)^2(1 - h_i)},$$

jossa $s(i)^2$ on σ^2:n estimaatti, kun i:s havainto on poistettu aineistosta. Jos e_i:t tulevat normaalijakaumasta, niin studentisoidut residuaalit ovat $t(n-p-2)$-jakautuneita. On odotettavissa, että noin 95% studentisoiduista residuaaleista on välillä $[-2, 2]$, kun malli on oikea.



Kuva 3.6: Residuaaliplotteja. (a) Malli kunnossa, (b) Varianssi ei ole vakio, (c) Mallista puuttuu vakiotermi, (d) Mahd. toisen asteen termi puuttuu.

Normaalijakaumasta

- Teoreettinen peruste käytölle: keskeinen raja-arvolause; käytännössä on huomattu (jo 1800-luvulla) empiiristen jakaumien usein noudattavan normaalijakaumaa
- Testien ja estimaattien (esim. uskottavuuspäättelyssä) rajajakauma otoskoon kasvaessa on usein normaalijakauma (keskeinen raja-arvolause)
- Muunnokset; ns. deltamenetelmä: Jos parametrin θ estimaatille $\hat{\theta}_n$ pätee jakauma-approksimaatio $\hat{\theta}_n \sim N(\theta, \tau^2/n)$ (likimain), niin muunnokselle $h(\hat{\theta}_n)$ pätee approksimaatio $h(\hat{\theta}_n) \sim N(h(\theta), (h'(\theta))^2\tau^2/n)$ (likimain).
- Summan normaalisuus: Jos $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \Sigma)$, niin $\boldsymbol{a}^T\boldsymbol{x} \sim N(\boldsymbol{a}^T\boldsymbol{\mu}, \boldsymbol{a}^T\Sigma\boldsymbol{a})$; odotusarvon ja varianssien summautuminen riippumattomassa tapauksessa
- Yhteys binomijakaumaan ja Poisson-jakaumaan
- Normaalijakaumasta johdetut jakaumat: χ^2-jakauma (keskinen ja epäkeskinen), t-jakauma (keskinen ja epäkeskinen) ja F-jakauma (keskinen ja epäkeskinen)

3.2 Muita estimointimenetelmiä

- Suurimman uskottavuuden estimointi muiden virhejakaumien tapauksessa
- M-estimointi: robusti vaihtoehto
- LAD (Least Absolute Deviation): minimoi objektifunktion $\sum_{i=1}^{n} |r_i|$, missä $r_i = y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}$, $i = 1, \ldots, n$.
- LMS (Least Median of Squares): minimoi objektifunktion $\text{Med}(r_i^2)$.
- RREG (Rank REGression): minimoi objektifunktion $\sum_{i<j} |r_i - r_j|$.

Luku 4

Dikotominen vaste

4.1 Yksi otos

Vastemuuttuja y on dikotominen (0/1)-arvoinen muuttuja. Olkoon tapauksen $y_i = 1$ todennäköisyys tai **vaara**

$$p_i = P(y_i = 1)$$

ja myöhemmin oletetaan, että $g(p_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$ jollakin tunnetulla linkkifunktiolla $g(p)$.

Tarkastellaan aluksi vaaran suhteen homogeenista populaatiota ja oletetaan, että vaara $p_i = p$, $i = 1, \ldots n$, on vakio. Silloin

$$y = \sum_{i=1}^{n} y_i \sim Bin(n, p).$$

Edelleen

$$P(y = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \ldots, n,$$

ja

$$E(y) = np \quad \text{ja} \quad Var(y) = np(1-p).$$

Tuntemattoman parametrin p suurimman uskottavuuden estimaattori (ja pienimmän varianssin harhaton estimaattori, MVUE) on

$$\hat{p} = \frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Estimaattorille pätee

$$E(\hat{p}) = p \quad \text{ja} \quad Var(\hat{p}) = \frac{p(1-p)}{n}$$

ja

$$\sqrt{n}\frac{\hat{p}-p}{\sqrt{p(1-p)}} \sim AN(0,1).$$

Fraasi '$\ldots\sim AN(0,1)$' luetaan '... noudattaa likimain standardoitua normaalijakaumaa, kun n suuri' tai '... rajajakauma n:n kasvaessa on standardoitu normaalijakauma.'. Likimääräinen 95 %:n luottamusväli on siis muotoa

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

HUOM. Deltamenetelmää käyttäen (koska $\hat{p} \sim N(p, p(1-p)/n)$ likimain) huomataan, että esimerkiksi

$$\log(\hat{p}) \sim N(\log(p), \frac{1-p}{np})$$

(likimain), mistä saadaan $g(p) = \log(p)$:lle luottamusväli

$$\log(\hat{p}) \pm 1.96 \times \sqrt{\frac{1-\hat{p}}{n\hat{p}}}$$

ja edelleen p:lle luottamusväli

$$\hat{p} \times \exp\left\{\pm 1.96 \times \sqrt{\frac{1-\hat{p}}{n\hat{p}}}\right\}.$$

(Hyvä vaihtoehto silloin, kun p pieni: Tällöin $\hat{p}$:n jakauma oikealle vino ja logaritmimuunnos johtaa parempaan normaalijakauma-approksimaatioon korjaamalla tämän vinouden.)

Vaihtoehtoinen tapa parametrisoida vaaraa on **vedonlyöntisuhde** (odds)

$$o = \frac{p}{1-p}$$

tai vedonlyöntisuhteen logaritmi $\log(o) = \log(p/(1-p))$ (log-odds). Muunnos

$$g(p) = \log\Big(\frac{p}{1-p}\Big)$$

on niin sanottu **logit-muunnos**. Vaara lausuttuna vedonlyöntisuhteen avulla on

$$p = \frac{o}{1+o} = \Big(1 + \frac{1}{o}\Big)^{-1}.$$

Vedonlyöntisuhteen o suurimman uskottavuuden estimaatti on $\hat{o} = \hat{p}/(1-\hat{p})$. Deltamenetelmää käyttäen pätee

$$\log(\hat{o}) \sim N\Big(\log(o), \frac{1}{np(1-p)}\Big)$$

(likimain). Siis $\log(o)$:n 95 %:n likimääräinen luottamusväli on

$$\log(\hat{o}) \pm 1.96 \times \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}},$$

mikä antaa o:lle luottamusvälin

$$\hat{o} \times \exp\Big\{\pm 1.96 \times \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}\Big\}$$

ja lopulta p:n luottamusvälin

$$\Big[1 + \frac{1-\hat{p}}{\hat{p}} \exp\Big\{\pm 1.96 \times \sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}\Big\}\Big]^{-1}.$$

4.2 Vertailuparametreista, kahden otoksen tapaus

Olkoot $y_1, \ldots, y_{n_1}$ ja $y_{n_1+1}, \ldots, y_{n_1+n_2}$ riippumattomat satunnaisotokset Bernoullijakaumista parametreilla p_1 ja p_2. Siis p_1 (p_2) on ensimmäiseen (toiseen) otokseen/populaatioon liittyvä vaara. Olkoot edelleen

$$o_1 = \frac{p_1}{1-p_1} \quad \text{ja} \quad o_2 = \frac{p_2}{1-p_2}$$

populaatioihin liittyvät vedonlyöntisuhteet.

Luonnollisesti suurimman uskottavuuden estimaatit parametreille p_1 ja p_2 ovat

$$\hat{p}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i \quad \text{ja} \quad \hat{p}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} y_i$$

ja luottamusvälit p_1:lle ja p_2:lle (ja o_1:lle ja o_2:lle) saadaan kuten edellä.

Vertailuparametreja.

Yleisimmät vertailuparametrit verrattaessa vaaraa kahdessa eri populaatiossa ovat

- Vaaraero (Risk Difference):

$$RD = p_2 - p_1$$

- Vaarasuhde (Risk Ratio):

$$RR = \frac{p_2}{p_1}$$

- Ristitulosuhde (Odds Ratio):

$$OR = \frac{o_2}{o_1} = \frac{p_2(1-p_1)}{(1-p_2)p_1}$$

Estimaatit ja luottamusvälit:

- Vaaraero, RD:

$$\widehat{RD} = \hat{p}_2 - \hat{p}_1$$

$$\widehat{RD} \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Vaarasuhde, RR:

$$\widehat{RR} = \frac{\hat{p}_2}{\hat{p}_1}$$

$$\widehat{RR} \times \exp\left\{\pm 1.96\sqrt{\frac{(1-\hat{p}_1)}{\hat{p}_1 n_1} + \frac{(1-\hat{p}_2)}{\hat{p}_2 n_2}}\right\}$$

- Ristitulosuhde, OR:

$$\widehat{OR} = \frac{\hat{o}_2}{\hat{o}_1}$$

$$\widehat{OR} \times \exp\left\{\pm 1.96\sqrt{\frac{1}{\hat{p}_1(1-\hat{p}_1)n_1} + \frac{1}{\hat{p}_2(1-\hat{p}_2)n_2}}\right\}$$

4.3 Kaksi dikotomista selittäjää

Tarkastellaan kahden dikotomisen (0/1-arvoisen) selittäjän (arvot matriisin X sarakkeissa 2 ja 3; muuttujat x_2 ja x_3) samanaikaista vaikutusta dikotomiseen vasteeseen (arvot vektorissa $\boldsymbol{y}$). Määritellään myös **yhdysvaikutusmuuttuja** (arvot sarakkeella 4; muuttuja x_4: $x_{i4} = x_{i2} \cdot x_{i3},\ i = 1, \ldots, n$.

Vaarat ja prediktorit ($\boldsymbol{\beta}^T\boldsymbol{x} = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$) eri tilanteissa (kun yhdysvaikutusmuuttuja on mukana prediktorissa) ovat taulukoissa 4.1 ja 4.2.

	x_3	
x_2	0	1
0	p_{00}	p_{01}
1	p_{10}	p_{11}

Taulukko 4.1: Vaarat eri tapauksissa.

	x_3	
x_2	0	1
0	β_1	$\beta_1 + \beta_3$
1	$\beta_1 + \beta_2$	$\beta_1 + \beta_2 + \beta_3 + \beta_4$

Taulukko 4.2: Lineaariset prediktorit eri tapauksissa.

y	n	x_1	x_2	x_3	x_4
y_{00}	n_{00}	1	0	0	0
y_{01}	n_{01}	1	0	1	0
y_{10}	n_{10}	1	1	0	0
y_{11}	n_{11}	1	1	1	1

Taulukko 4.3: Havaintoaineisto

Tulkinnat

Identtinen linkki: $p_i = \boldsymbol{\beta}^T \boldsymbol{x}_i$

- β_1 "perusvaara"eli vaara tilanteessa $\{x_2 = 0, x_3 = 0\}$.
- β_2 on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 1, x_3 = 0\}$ välinen vaaraero. Jos $\beta_2 > 0 (< 0)$, niin vaara on β_2:n verran suurempi (pienempi) tilanteessa $\{x_2 = 1, x_3 = 0\}$ kuin tilanteessa $\{x_2 = 0, x_3 = 0\}$.
- β_3 on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 0, x_3 = 1\}$ välinen vaaraero. Jos $\beta_3 > 0$ (< 0), niin vaara on β_3:n verran suurempi (pienempi) tilanteessa $\{x_2 = 0, x_3 = 1\}$ kuin tilanteessa $\{x_2 = 0, x_3 = 0\}$.
- β_4 on yhdysvaikutus. Tasolla $x_2 = 0$ tapausten $\{x_3 = 0\}$ ja $\{x_3 = 1\}$ välinen vaaraero on β_3 mutta tasolla $x_2 = 1$ vaaraero on $\beta_3 + \beta_4$. Jos siis $\beta_4 > 0$ (< 0), niin vaaraero on β_4:n verran suurempi (pienempi) tasolla $x_2 = 1$ kuin tasolla $x_2 = 0$.

Logaritmilinkki: $\log(p_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$ tai $p_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i)$

- $\exp(\beta_1)$ "perusvaara"eli vaara tilanteessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_2)$ on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 1, x_3 = 0\}$ välinen vaarasuhde. Jos $\beta_2 > 0$ eli $\exp(\beta_2) > 1$, niin vaara on $\exp(\beta_2)$ kertaa suurempi tapauksessa $\{x_2 = 1, x_3 = 0\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$. Jos $\beta_2 < 0$ eli $0 < \exp(\beta_2) < 1$, niin vaara on $100(1 - \exp(\beta_2))\%$ pienempi tapauksessa $\{x_2 = 1, x_3 = 0\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_3)$ on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 0, x_3 = 1\}$ välinen vaarasuhde. Jos $\beta_3 > 0$ eli $\exp(\beta_3) > 1$, niin vaara on $\exp(\beta_3)$ kertaa suurempi tapauksessa $\{x_2 = 0, x_3 = 1\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$. Jos $\beta_3 < 0$ eli $0 < \exp(\beta_3) < 1$, niin vaara on $100(1 - \exp(\beta_3))\%$ pienempi tapauksessa $\{x_2 = 0, x_3 = 1\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_4)$ on yhdysvaikutus. Tasolla $x_2 = 0$ tapausten $\{x_3 = 0\}$ ja $\{x_3 = 1\}$ välinen vaarasuhde on $\exp(\beta_3)$ mutta tasolla $x_2 = 1$ vaarasuhde on $\exp(\beta_3 + \beta_4)$. Jos $\beta_4 > 0$, niin vaarasuhde on $\exp(\beta_4)$ kertaa suurempi tasolla $x_2 = 1$ kuin tasolla $x_2 = 0$. Jos $\beta_4 < 0$, niin vaarasuhde on $100(1 - \exp(\beta_4))\%$ pienempi tasolla $x_2 = 1$ kuin tasolla $x_2 = 0$.

Logitlinkki: $\log(o_i) = \log(p_i/(1-p_i)) = \boldsymbol{\beta}^T\boldsymbol{x}_i$ tai $o_i = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)$ tai $p_i = \exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)/(1+\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i))$

- $\exp(\beta_1)$ "perusodds"eli vedonlyöntisuhde tilanteessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_2)$ on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 1, x_3 = 0\}$ välinen ristitulosuhde (OR). Jos $\beta_2 > 0$ eli $\exp(\beta_2) > 1$, niin ristitulosuhde on $\exp(\beta_2)$ kertaa suurempi tilanteessa $\{x_2 = 1, x_3 = 0\}$ kuin tilanteessa $\{x_2 = 0, x_3 = 0\}$. Jos $\beta_2 < 0$ eli $0 < \exp(\beta_2) < 1$, niin ristitulosuhde on $100(1 - \exp(\beta_2))\%$ pienempi tapauksessa $\{x_2 = 1, x_3 = 0\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_3)$ on tapausten $\{x_2 = 0, x_3 = 0\}$ ja $\{x_2 = 0, x_3 = 1\}$ välinen ristitulosuhde (OR). Jos $\beta_3 > 0$ eli $\exp(\beta_3) > 1$, niin ristitulosuhde on $\exp(\beta_3)$ kertaa suurempi tilanteessa $\{x_2 = 0, x_3 = 1\}$ kuin tilanteessa $\{x_2 = 0, x_3 = 0\}$. Jos $\beta_3 < 0$ eli $0 < \exp(\beta_3) < 1$, niin ristitulosuhde on $100(1 - \exp(\beta_3))\%$ pienempi tapauksessa $\{x_2 = 0, x_3 = 1\}$ kuin tapauksessa $\{x_2 = 0, x_3 = 0\}$.
- $\exp(\beta_4)$ on yhdysvaikutus. Tasolla $x_2 = 0$ tapausten $\{x_3 = 0\}$ ja $\{x_3 = 1\}$ välinen ristitulosuhde on $\exp(\beta_3)$ mutta tasolla $x_2 = 1$ ristitulosuhde on $\exp(\beta_3 + \beta_4)$. Jos $\beta_4 > 0$, niin ristitulosuhde on $\exp(\beta_4)$ kertaa suurempi tasolla $x_2 = 1$ kuin tasolla $x_2 = 0$. Jos $\beta_4 < 0$, niin ristitulosuhde on $100(1 - \exp(\beta_4))\%$ pienempi tasolla $x_2 = 1$ kuin tasolla $x_2 = 0$.

4.4 Kohorttitutkimus vs. tapaus-verrokkitutkimus

Kohorttitutkimuksessa tutkitaan altisteen vaikutusta valitussa väestössä. Havaintoyksiköt on valittu aineistoon altisteen perusteella, joten havaintoyksiköiden lukumäärät eri altistetasoilla ovat kiinteitä. Tällöin vaaraa ja vedonlyöntisuhdetta pystytään estimoimaan eri altistetasoilla. Taulukossa 4.4 on havainnollistava esimerkki kahdesta kohorttitutkimusaineistosta. Altistetasolla Altiste=1 molemmissa aineistoissa on samat solufrekvenssit. Altistetasolla Altiste=0 toisessa aineistossa on viisinkertaiset solufrekvessit ensimmäiseen verrattuna. Tässä oletetaan siis, että tapausten suhteellinen osuus pysyy täysin samana. Käytännön tutkimuksissa suhteelliset osuudet tietenkin hieman vaihtelevat mutta odotusarvo pysyy vakiona ($E(\hat{p}) = p$) otoskoosta riippumatta. Huomataan, että molemmissa tutkimuksissa vaaran estimaateiksi saadaan $\hat{p_1} = A/(A+B) = 50/70 = 5/7$ (Altiste=1) ja $\hat{p_0} = C/(C+D) = 10/30 = 50/150 = 1/3$ (Altiste=0). Vedonlyöntisuhteiksi saadaan näin ollen $\hat{o_1} = A/B = 5/2$ (Altiste=1) ja $\hat{o_0} = C/D = 1/2$ (Altiste=0). Ristitulosuhde on nyt $\hat{OR} = \hat{o_1}/\hat{o_0} = (AD)/(BC) = (50 \cdot 20)/(20 \cdot 10) = 5$.

	Tutkimus 1			Tutkimus 2		
	Vaste=1	Vaste=0	Yht.	Vaste=1	Vaste=0	Yht.
Altiste=1	50 (A)	20 (B)	70	50 (A)	20 (B)	70
Altiste=0	10 (C)	20 (D)	30	50 (C)	100 (D)	150

Taulukko 4.4: Kohorttitutkimus

Tapaus-verrokkitutkimuksessa tapausten ja verrokkien ("ei-tapausten") lukumäärät kiinnitetään ennalta, joten tutkija käytännössä määrää keinotekoisesti tapausten suhteelliset osuudet eri altisteryhmissä. Tällöin luonnollisestikaan vaaraa ja vedonlyöntisuhdetta **ei pystytä** estimoimaan eri altistetasoilla. Merkitään $y = Vaste$ ja $x = Altiste$ ja olkoon

$$\begin{aligned}
\theta_1 &= P(x=1|y=1), \quad 1-\theta_1 = P(x=0|y=1),\\
\theta_0 &= P(x=1|y=0), \quad 1-\theta_0 = P(x=0|y=0),\\
p_1 &= P(y=1|x=1), \quad 1-p_1 = P(y=0|x=1),\\
p_0 &= P(y=1|x=0), \quad 1-p_0 = P(y=0|x=0).
\end{aligned}$$

Tällöin ristitulosuhteeksi saadaan Bayesin kaavaa käyttämällä

$$\begin{aligned}
OR &= \frac{p_1/(1-p_1)}{p_0/(1-p_0)} \\
&= \frac{P(y=1|x=1)/P(y=0|x=1)}{P(y=1|x=0)/P(y=0|x=0)} \\
&= P(y=1|x=1)[P(y=0|x=1)]^{-1}P(y=0|x=0)[P(y=1|x=0)]^{-1} \\
&= \frac{P(y=1)P(x=1|y=1)}{P(x=1)}\left[\frac{P(y=0)P(x=1|y=0)}{P(x=1)}\right]^{-1} \\
&\quad \cdot\frac{P(y=0)P(x=0|y=0)}{P(x=0)}\left[\frac{P(y=1)P(x=0|y=1)}{P(x=0)}\right]^{-1} \\
&= \frac{P(y=1)\theta_1}{P(x=1)}\left[\frac{P(y=0)\theta_0}{P(x=1)}\right]^{-1}\frac{P(y=0)(1-\theta_0)}{P(x=0)}\left[\frac{P(y=1)(1-\theta_1)}{P(x=0)}\right]^{-1} \\
&= \frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}.
\end{aligned}$$

Parametrit θ_0 ja θ_1 pystytään estimoimaan aineistosta, joten ristitulosuhteelle saadaan estimaatti myös tapaus-verrokkitutkimuksessa!! Estimaatin kaava on täsmälleen sama kuin kohorttitutkimuksen tapauksessa:

$$\hat{OR} = \frac{AD}{BC}.$$

Taulukossa 4.5 on havainnollistava esimerkki kahdesta tapaus-verrokkiaineistosta. Tutkimuksessa 2 on otettu viisinkertainen otos verrokkiryhmästä (Vaste=0). Ristitulosuhteen estimaatiksi saadaan molemmissa tutkimuksissa

$$OR = \frac{50 \cdot 20}{20 \cdot 10} = \frac{50 \cdot 100}{100 \cdot 10} = 5.$$

	Tutkimus 1		Tutkimus 2	
	Vaste=1	Vaste=0	Vaste=1	Vaste=0
Altiste=1	50 (A)	20 (B)	50 (A)	100 (B)
Altiste=0	10 (C)	20 (D)	10 (C)	100 (D)
Yht.	60	40	60	200

Taulukko 4.5: Tapaus-verrokkitutkimus

4.5 Yleinen malli, suurimman uskottavuuden estimointi

Dikotomisen vasteen tapauksessa havaintomatriisia on usein mahdollista ”tiivistää” keräämällä yhteen tapaukset, joilla on samat selittäjien arvot (”profiilit”).

Alkuperäinen data:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}, \quad \text{missä} \;\; y_i \sim Bin(1, p_i), \;\; g(p_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

Tiivistetty data (r erilaista profiilia, $r \leq n$; n_i on profiilia i vastaavien tapausten lukumäärä):

$$\begin{pmatrix} y_1 \\ \vdots \\ y_r \end{pmatrix}, \begin{pmatrix} n_1 \\ \vdots \\ n_r \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_r^T \end{pmatrix}, \quad \text{missä} \;\; y_i \sim Bin(n_i, p_i), \;\; g(p_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

Tarkastellaan seuraavassa kolmikkoa $(\boldsymbol{y}, \boldsymbol{n}, \boldsymbol{x})$.

Olkoon $g(p)$ valittu linkkifunktio ja $h(x) = g^{-1}(x)$ sen käänteiskuvaus. Esimerkiksi:

- Identtinen linkki: $g(p) = p$, $h(x) = x$
- Logaritmilinkki: $g(p) = \log(p)$, $h(x) = \exp(x)$
- Logitlinkki: $g(p) = \text{logit}(p) = \log(p/(1-p))$, $h(x) = \exp(x)/(1 + \exp(x))$

Uskottavuusfunktio ja logaritminen uskottavuusfunktio ovat

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{r} \binom{n_i}{y_i} p_i^{y_i} (1-p_i)^{n_i - y_i}$$

ja

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \text{const} + \sum_{i=1}^{r} \Big\{ y_i \log(p_i) + (n_i - y_i) \log(1-p_i) \Big\},$$

jossa siis $p_i = h(\boldsymbol{\beta}^T \boldsymbol{x}_i)$.

Suurimman uskottavuuden ratkaisu löytyy (periaatteessa) $l(\boldsymbol{\beta})$:n (tai $L(\boldsymbol{\beta})$:n) ensimmäisten derivaattojen nollakohdasta: Merkitään

$$s_j(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}) = \sum \Big\{ \frac{y_i - n_i p_i}{p_i(1-p_i)} \; h'(\boldsymbol{\beta}^T \boldsymbol{x}_i) \; x_{ij} \Big\}$$

($l(\boldsymbol{\beta})$:n osittaisderivaatta β_j:n suhteen) ja

$$\boldsymbol{s}(\boldsymbol{\beta}) = \begin{pmatrix} s_1(\boldsymbol{\beta}) \\ \vdots \\ s_p(\boldsymbol{\beta}) \end{pmatrix}$$

(osittaisderivaattojan muodostama vektoriarvoinen funktio). Silloin $\hat{\beta}$ määräytyy ehdoista

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \mathbf{0} \quad \text{tai} \quad \begin{pmatrix} s_1(\boldsymbol{\beta}) \\ \vdots \\ s_p(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Valitettavasti ratkaisua ei voi antaa suljetussa muodossa (kuin aivan yksinkertaisissa tapauksissa).

4.6 Logistinen regressio

Tarkastellaan logit-linkkiä, jolloin siis

$$\text{logit}(p_i) = \log\Big(\frac{p_i}{1-p_i}\Big) = \boldsymbol{\beta}^T\boldsymbol{x}_i$$

eli

$$p_i = \frac{\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}{1+\exp(\boldsymbol{\beta}^T\boldsymbol{x}_i)}.$$

On helppo nähdä, että

$$s_j(\boldsymbol{\beta}) = \sum_{i=1}^{r}\Big[(y_i - n_ip_i)x_{ij}\Big] \;\text{ ja }\; i_{jk}(\boldsymbol{\beta}) = \sum_{i=1}^{r}\Big[n_ip_i(1-p_i)x_{ij}x_{ik}\Big]$$

eli

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu}) \;\text{ ja }\; \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X},$$

jossa $\mu_i = n_ip_i$, $i = 1,\ldots,r$, ja $\boldsymbol{W} = \text{diag}\Big(n_ip_i(1-p_i)\Big)$. Siis

$$E(\boldsymbol{y}) = \boldsymbol{\mu} \;\text{ ja }\; Cov(\boldsymbol{y}) = \boldsymbol{W}.$$

Olkoon nyt $\hat{\boldsymbol{\beta}}_{(k)}$ k:s arvaus ja olkoon $\boldsymbol{\mu}_{(k)}$ ja $\boldsymbol{W}_{(k)}$ tähän arvaukseen liittyvä odotusarvovektori ja kovarianssimatriisi. Silloin Newtonin-Raphsonin menetelmän mukainen $k+1$:s arvaus on

$$\hat{\boldsymbol{\beta}}_{(k+1)} = \hat{\boldsymbol{\beta}}_{(k)} + (\boldsymbol{X}^T\boldsymbol{W}_{(k)}\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{\mu}_{(k)}).$$

HUOM. $\boldsymbol{\mathcal{I}}(\hat{\boldsymbol{\beta}}) = \boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}$ on kyseiseen ongelmaan liittyvä ns. **Fisherin informaatiomatriisi**, ja sen käänteismatriisi estimoi $\hat{\boldsymbol{\beta}}$:n kovarianssimatriisia:

$$\widehat{Cov}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X})^{-1}$$

4.7 Mallien vertailu

Tarkastellaan "sisäkkäisiä" malleja M_1 (selittäjämatriisi $\boldsymbol{X}_1$) ja M_2 (selittäjämatriisi $\boldsymbol{X}_2$), ja oletetaan, että malli M_2 on saatu mallista M_1 lisäämällä siihen h uutta selittäjää. ($\boldsymbol{X}_2$ saadaan $\boldsymbol{X}_1$:stä lisäämällä siihen h uutta saraketta.)

Olkoot malliin M_1 (M_2) liittyvien vaaraestimaattien ja odotusarvojen vektorit $\hat{\boldsymbol{p}}_1$ ($\hat{\boldsymbol{p}}_2$) ja $\hat{\boldsymbol{\mu}}_1$ ($\hat{\boldsymbol{\mu}}_2$). Silloin näihin malleihin liittyvät logaritmisten uskottavuusfunktioiden maksimiarvot ovat

$$l_1(\hat{\boldsymbol{\beta}}_1) = c + \sum_{i=1}^{r} \Big\{ y_i \ \log(\hat{\mu}_{1i}) + (n_i - y_i) \ \log(n_i - \hat{\mu}_{1i}) \Big\}$$

ja

$$l_2(\hat{\boldsymbol{\beta}}_2) = c + \sum_{i=1}^{r} \Big\{ y_i \ \log(\hat{\mu}_{2i}) + (n_i - y_i) \ \log(n_i - \hat{\mu}_{2i}) \Big\}.$$

Mallien vertailuun sopiva uskottavuusosamäärätestisuureen logaritmi (kahdella kerrottuna)

$$\chi^2 = 2\Big(l_2(\hat{\boldsymbol{\beta}}_2) - l_1(\hat{\boldsymbol{\beta}}_1)\Big) = 2 \sum_{i=1}^{r} \left\{ y_i \ \log\left(\frac{\hat{\mu}_{2i}}{\hat{\mu}_{1i}}\right) + (n_i - y_i) \ \log\left(\frac{n_i - \hat{\mu}_{2i}}{n_i - \hat{\mu}_{1i}}\right) \right\}$$

on likimain $\chi^2(h)$-jakautunut, kun M_1 on tosi.

HUOM. GLIM-termistössä malliin M (ja sen antamaan estimaattiin $\hat{\mu}$) liittyvä **devianssi** on

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\ \sum_{i=1}^{r}\Big\{y_i\ \log\Big(\frac{y_i}{\hat{\mu}_i}\Big) + (n_i - y_i)\ \log\Big(\frac{n_i - y_i}{n_i - \hat{\mu}_i}\Big)\Big\}.$$

Devianssi siis vertaa mallia M "täydelliseen"tai "saturoituun"malliin, jossa kaikki r (=profiilien tai luokkien lukumäärä) parametria ovat käytössä.

Ylläoleva (uskottavuusosamäärä)testisuure on siis **devianssien muutos** siirryttäessä mallista M_1 malliin M_2. Devianssia tai **Pearsonin χ^2-testisuuretta**

$$\chi^2 = \sum_{i=1}^{r}\frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

käytetään usein lopullisen mallin M "hyvyyden"tutkimiseen. Eräiden yleisten ehtojen vallitessa molemmat suureet ovat (jos testattava malli on oikea) "asymptoottisesti ekvivalentteja"ja niiden rajajakauma on χ^2-jakauma vapausasteilla

$$r - \text{mallin } M \text{ parametrien lkm.}$$

4.8 Ylihajontaongelma

Mallin M (tai siihen liittyvän estimaatin $\hat{\mu}$) hyvyyttä voi siis karkeasti tutkia suureen

$$\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}})}{r-p}$$

avulla, missä r on profiilien lukumäärä ja p mallin M parametrien lukumäärä. Voidaan odottaa, että yo. suure on lähellä (likimääräistä) odotusarvoaan 1, jos malli M on oikea. Jos suure kuitenkin on "suuri", puhutaan ns. ylihajonnasta (over-dispersion). Ylihajonnan syy voi olla paitsi mallin virheellisyys, myös esimerkiksi havaintojen klusterisoituminen (jolloin riippumattomuusoletus ei pidä paikkaansa):
Muodostukoon (yhden otoksen) havaintoaineisto esimerkiksi m klusterista, ja olkoon jokaisen klusterin koko k; havaintojen lukumäärä on siis $n = mk$. Havaintoaineiston muodostavat dikotomiset havainnot

- 1. klusteri: $y_{11}, \ldots, y_{1k}$
- 2. klusteri: $y_{21}, \ldots, y_{2k}$
- ...
- m. klusteri: $y_{m1}, \ldots, y_{mk}$

Oletetaan, että klusterin sisällä on voimakasta riippuvuutta s.e. y-arvot ovat samat.

Virheellinen oletus: $y = \sum_i \sum_j y_{ij} \sim Bin(n,p)$, jolloin

$$E(y) = np \;\text{ ja }\; Var(y) = np(1-p)$$

Todellinen tilanne: $y = \sum_i k y_{i1} = k \sum y_{i1}$, missä y_{i1}:t riippumattomia ja $\sum y_{i1}$ siis noudattaa $Bin(m,p)$-jakaumaa. Nyt

$$E(y) = kmp = np \;\text{ ja }\; Var(y) = k^2 mp(1-p) = knp(1-p)$$

HUOM. Jos ylihajontaa esiintyy, estimaattien kovarianssimatriisiin ei voi luottaa ja tilanteen korjaamiseksi suositellaan yleensä $\hat{\boldsymbol{\beta}}$:n (väärään oletukseen perustuvan) kovarianssimatriisiestimaatin kertomista "ylihajontaparametrilla"

$$\frac{D(\boldsymbol{y};\hat{\boldsymbol{\mu}})}{r-p} \;\text{ tai }\; \frac{1}{r-p}\sum \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)} = \frac{\chi^2}{r-p}$$

4.9 Luokiteltu (polytominen) vaste

Yhden otoksen tapaus: Oletetaan, että vaste on luokiteltu ja että luokkien 1, 2, ..., k todennäköisyydet ovat $p_1, \ldots, p_k$. Olkoot $y_1, \ldots, y_k$ eri luokkiin sattuneiden havaintojen lukumäärät n havainnon satunnaisotoksessa. Merkitään

$$\boldsymbol{p} = \begin{pmatrix} p_1 \\ \vdots \\ p_k \end{pmatrix} \quad \text{ja} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix},$$

missä siis $\sum p_j = 1$ ja $\sum y_j = n$. Silloin $\boldsymbol{y}$ noudattaa **multinomijakaumaa** $Multin(n; (p_1, \ldots, p_k))$ ja

$$P(y_1 = n_1, \ldots y_k = n_k) = \frac{n!}{n_1! \ldots n_k!} p_1^{n_1} \ldots p_k^{n_k}, \quad n_1 + \ldots + n_k = n.$$

HUOM. Jos $\boldsymbol{y} \sim Multin(n; (p_1, \ldots, p_k))$, niin $y_j \sim Bin(n, p_j)$ ja

$$E(y_j) = np_j, \quad Var(y_j) = np_j(1 - p_j) \quad \text{ja} \quad Cov(y_j, y_l) = -np_j p_l.$$

Edelleen

$$\hat{\boldsymbol{p}} = (y_1/n, \ldots, y_k/n)^T \sim N_k\Big(\boldsymbol{p}, \frac{1}{n}[\text{diag}(p_1, \ldots, p_k) - \boldsymbol{p}\boldsymbol{p}^T]\Big)$$

likimain, kun n on suuri.

Luokat nominaaliasteikolla: Logistisen mallin yleistys saadaan, kun

$$p_j = \frac{\exp(\nu_j)}{\sum \exp(\nu_j)}, \quad j = 1, \ldots k,$$

missä $\nu_1 = 0$ ja

$$\nu_j = \boldsymbol{\beta}_j^T \boldsymbol{x}, \quad j = 2, \ldots, k.$$

Kun $k = 2$ (binäärinen vaste), ν_2 on luokkaan 2 liittyvä log-odds ja saadaan tavallinen logistinen malli.

Parametrivektori $\boldsymbol{\beta}_j$ kertoo, miten luokan j todennäköisyydet riippuvat selittäjistä (verrattuna luokkaan 1) ja luokkien j ja j' todennäköisyyksiä verrattaessa huomataan, että

$$\log\left(\frac{p_j}{p_{j'}}\right) = (\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j'})^T \boldsymbol{x}.$$

Luokat ordinaaliasteikolla:

Merkitään

$$P_1 = p_1,$$
$$P_2 = p_1 + p_2$$
$$\ldots$$
$$P_j = p_1 + \ldots + p_j$$
$$\ldots$$
$$P_k = 1$$

ja rakennetaan "logistinen malli"olettaen, että

$$\log \frac{P_j}{1 - P_j} = \theta_j - \boldsymbol{\beta}^T \boldsymbol{x}.$$

Sisäkkäiset tai hierarkiset mallit:

Luku 5

Lukumäärävaste

5.1 Poisson-jakauma

Satunnaismuuttuja y noudattaa **Poisson-jakaumaa** parametrilla (odotusarvolla) μ, jos sen mahdolliset arvot ovat $0, 1, 2, \ldots$ todennäköisyyksillä

$$P(y = j) = \frac{\mu^j}{j!} e^{-\mu}, \quad j = 0, 1, 2, \ldots$$

Silloin $E(y) = \mu$ ja $Var(y) = \mu$.

Tärkeä tulos: Jos $y_1 \sim Poi(\mu_1)$ ja $y_2 \sim Poi(\mu_2)$ ja y_1 ja y_2 ovat riippumattomia, niin $y_1 + y_2 \sim Poi(\mu_1 + \mu_2)$.

HUOM. Poisson-oletus on yleensä sopiva mallitettaessa harvinaisten tapahtumien sattumiskertoja (tietyllä alueella, tiettynä aikavälinä). Ns. Poisson-prosessi: Toistuvan tapahtuman "hetkellinen"sattumistodennäköisyys on ajasta riippumaton vakio. Silloin odotusaika peräkkäisten sattumiskertojen välillä noudattaa **eksponenttijakaumaa** ja sattumiskertojen lukumäärä tietyllä kiinteällä välillä noudattaa Poissonjakaumaa. Jos eksponenttijakauman odotusarvo on λ, niin t:n pituisen välin sattumiskertojen lukumäärän odotusarvo on t/λ.

HUOM. Yhteydet binomijakaumaan ja normaalijakaumaan:

(i). Jos $y \sim Bin(n, p)$ ja n on "suuri"ja p on "pieni", niin $y \sim Poi(\mu)$ likimain, missä $\mu = np$.

(ii) Jos $y \sim Poi(\mu)$ ja μ "suuri", niin $(y - \mu)/\sqrt{\mu} \sim N(0, 1)$ likimain.

Yksi otos

Oletetaan, että $y_1, \ldots, y_n$ on satunnaisotos jakaumasta $Poi(\mu)$. Merkitään $y = \sum y_i$. ($y \sim Poi(n\mu)$.)

Silloin

$$L(\mu) = \frac{\mu^y}{\prod_{i=1}^n y_i!} e^{-n\mu} = \frac{y!}{y_1! \ldots y_n!} \left(\frac{1}{n}\right)^{\sum_{i=1}^n y_i} \frac{(n\mu)^y}{y!} e^{-n\mu}$$

ja siis

$$l(\mu) = vakio + y \cdot \log(\mu) - n \cdot \mu,$$

joten μ:n suurimman uskottavuuden estimaatti on otoskeskiarvo $\hat{\mu} = y/n$. Nyt $E(\hat{\mu}) = \mu$ ja $Var(\hat{\mu}) = \mu/n$ ja μ:n likimääräinen 95 %:n luottamusväli on

$$\hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}}{n}} \quad \text{tai} \quad \hat{\mu} \times \exp\left\{\pm 1.96\sqrt{\frac{1}{\hat{\mu}n}}\right\}.$$

5.2 Kahden otoksen tapaus

Oletetaan, että $y_1, \ldots, y_m$ ja $y_{m+1}, \ldots, y_{m+n}$ ovat riippumattomat satunnaisotokset jakaumista $Poi(\mu_1)$ ja $Poi(\mu_2)$.

Merkitään

$$z_1 = y_1 + \ldots + y_m \quad \text{ja} \quad z_2 = y_{m+1} + \ldots + y_{m+n} \quad \text{sekä} \quad z = z_1 + z_2.$$

Silloin $z_1 \sim Poi(m\mu_1)$ ja $z_2 \sim Poi(n\mu_2)$ sekä $z \sim Poi(m\mu_1 + n\mu_2)$. Lukumäärät z_1 ja z_2 ovat riippumattomia.

Uskottavuusfunktio voidaan osittaa (multinomi-, binomi- ja Poisson-osiin) seuraavasti:

$$L(\mu_1,\mu_2) = \frac{z_1!}{\prod_{i=1}^{m} y_i!}\left(\frac{1}{m}\right)^{z_1} \cdot \frac{z_2!}{\prod_{i=m+1}^{m+n} y_j!}\left(\frac{1}{n}\right)^{z_2} \cdot$$
$$\cdot \frac{z!}{z_1!z_2!}\left(\frac{m\mu_1}{m\mu_1+n\mu_2}\right)^{z_1}\left(\frac{n\mu_2}{m\mu_1+n\mu_2}\right)^{z_2} \cdot$$
$$\cdot \frac{1}{z!}(m\mu_1+n\mu_2)^z e^{-m\mu_1-n\mu_2}.$$

Logaritmilinkkiä vastaava uudelleenparametrisointi:

$$\mu_1 = e^{\beta_1} \quad \text{ja} \quad \mu_2 = e^{\beta_1+\beta_2},$$

missä $\Delta = e^{\beta_2} = \mu_2/\mu_1$ on populaatioiden (käsittelyjen, tms.) eroa kuvaava parametri, odotusarvojen suhde. (Sillä on usein myös riskisuhdetulkinta). Toinen mahdollinen parametrisointitapa on määritellä

$$\tau = m\mu_1 + n\mu_2 \quad \text{ja} \quad \Delta = \mu_2/\mu_1,$$

jolloin uskottavuusfunktio on

$$L(\tau,\Delta) = vakio \cdot \tau^z e^{-\tau} \cdot \left(\frac{1}{1+\frac{n}{m}\Delta}\right)^{z_1}\left(\frac{\frac{n}{m}\Delta}{1+\frac{n}{m}\Delta}\right)^{z_2}$$

ja siis estimaatit $\hat{\tau}$ ja $\hat{\Delta}$ ovat (asymptoottisesti) riippumattomia.

Ylläolevasta seuraa, että jos $\hat{o} = z_2/z_1$ on kokeeseen $z_2 \sim Bin(z,p)$ liittyvä estimoitu vedonlyöntisuhde (odds), niin

$$\hat{\Delta} = \frac{n}{m}\hat{o} = \frac{\hat{\mu}_2}{\hat{\mu}_1}$$

ja luottamusvälit saadaan käyttäen o:n luottamusvälejä.

5.3 Yleinen malli, suurimman uskottavuuden estimointi

Aineisto:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \begin{pmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{pmatrix}, \quad \text{missä} \;\; y_i \sim Poi(\mu_i), \;\; g(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$$

Olkoon siis $g(p)$ valittu linkkifunktio, ja $h(x) = g^{-1}(x)$ sen käänteiskuvaus. Merkitään

$$\mu_i = h(\boldsymbol{\beta}^T \boldsymbol{x}_i), \quad i = 1, \ldots, n$$

Logaritminen uskottavuusfunktio on

$$l(\boldsymbol{\beta}) = \log(L(\boldsymbol{\beta})) = \text{const} + \sum_{i=1}^{n} \Big\{ y_i \log(\mu_i) - \mu_i \Big\}.$$

Edelleen

$$s_j(\boldsymbol{\beta}) = \frac{\partial}{\partial \beta_j} l(\beta) = \sum \Big\{ x_{ij} h'(\boldsymbol{\beta}^T \boldsymbol{x}_i) \Big(\frac{y_i}{\mu_i} - 1 \Big) \Big\}, \quad j = 1, \ldots p,$$

ja suurimman uskottavuuden estimaatti $\hat{\boldsymbol{\beta}}$ määräytyy ehdosta

$$\boldsymbol{s}(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} s_1(\boldsymbol{\beta}) \\ \vdots \\ s_p(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Ratkaisua ei voi yleisesti antaa suljetussa muodossa.

Logaritmilinkki: Tarkastellaan logaritmilinkkiä, jolloin siis

$$\log(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i \;\; \text{ja} \;\; \mu_i = \exp(\boldsymbol{\beta}^T \boldsymbol{x}_i).$$

On helppo nähdä, että nyt

$$s_j(\boldsymbol{\beta}) = \sum_i x_{ij}(y_i - \mu_i) \;\; \text{ja} \;\; i_{jk}(\boldsymbol{\beta}) = \sum_i x_{ij} x_{ik} \mu_i$$

tai

$$\boldsymbol{s}(\boldsymbol{\beta}) = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{\mu}) \;\; \text{ja} \;\; \boldsymbol{\mathcal{I}}(\boldsymbol{\beta}) = \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

jossa siis

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^T \;\; \text{ja} \;\; \boldsymbol{W} = \text{diag}(\mu_1, \ldots, \mu_n).$$

	1	2	...	k	Yht
	y_1	y_2	...	y_k	y

Taulukko 5.1: Yksi otos

	1	2	...	k	Yht
Käsittely 1	y_{11}	y_{12}	...	y_{1k}	y_1
Käsittely 2	y_{21}	y_{22}	...	y_{2k}	y_2

Taulukko 5.2: Kaksi otosta

5.4 Multinomivaste (revisited)

Palataan takaisin tilanteeseen, jossa vaste y on luokiteltu (k luokkaa); yhden satunnaisotoksen tapauksessa aineisto voidaan esittää taulukon 5.1 esittämässä muodossa. Jos nyt oletetaan, että havainnot $y_1, \ldots, y_k$ ovat riippumattomia ja noudattavat Poissonjakaumaa odotusarvoilla

$$e^{\nu_1},\ e^{\nu_1+\nu_2},\ \ e^{\nu_1+\nu_3},\ \ \ldots,\ \ e^{\nu_1+\nu_k},$$

niin $(y_1, \ldots, y_k)$:n ehdollinen jakauma ehdolla y on multinomijakauma parametreilla y ja

$$p_j = \frac{\exp(\nu_j)}{\sum_{i=1}^{k} \exp(\nu_i)},\quad j = 1, \ldots k$$

missä $\nu_1 = 0$. (Ko. todennäköisyydet eivät riipu lainkaan parametrista ν_1.) Koska

$$P(y_1, \ldots, y_k) = \ P(y) \cdot P(y_1, \ldots, y_k|y),$$

niin multinomijakaumaan liittyvät parametrit $\nu_2, \ldots, \nu_k$ voidaan estimoida Poissonmallin (logaritmilinkki) avulla, vasteina taulukon 5.1 y-arvot ja selittäjinä dikotomisoitu luokkamuuttuja (indikaattorit luokille $2, \ldots, k$).!

Tarkastellaan seuraavaksi kahden käsittelyn vertailua luokitellun vasteen tapauksessa (taulukko 5.2). Oletetaan jälleen (hetkeksi), että havainnot

$$y_{11}, \ldots y_{1k}, \quad y_{21}, \ldots, y_{2k}$$

ovat riippumattomia ja Poissonjakautuneita. Kuten yhden otoksen tapauksessa huomataan, että

$$P(y_{11}, \ldots, y_{1k}, y_{21}, \ldots, y_{2k}) = P(y_1, y_2) \cdot P(y_{11}, \ldots, y_{1k}|y_1) \cdot P(y_{21}, \ldots, y_{2k}|y_2),$$

ja multinomijakaumiin liittyvät parametrit voidaan jälleen estimoida Poissonmallitusta käyttäen: Vasteina taulukon 5.2 y-arvot ja selittäjinä dikotomisoitu käsittelymuuttuja (toista käsittelyä vastaava indikaattorimuuttuja) sekä dikotomisoitu luokkamuuttuja (indikaattorit luokille $2, \ldots, k$) sekä näiden yhdysvaikutukset. Yhdysvaikutusparametrien estimaatit antavat estimaatit käsittelyjen vaikutukselle multinomimallissa.

Yleinen tilanne: Vaste y on luokiteltu ja luokan j todennäköisyys riippuu olosuhteista $\boldsymbol{x}$ seuraavasti($\boldsymbol{\beta}_1$=$\mathbf{0}$)

$$P(y = j) = \frac{\exp(\boldsymbol{\beta}_j^T \boldsymbol{x})}{\sum_{i=1}^{k} \exp(\boldsymbol{\beta}_i^T \boldsymbol{x})}, \quad j = 2, \ldots, k.$$

Aineisto voidaan kirjoittaa muodossa (r profiilia):

$x_{11}, \ldots, x_{1p}, y_{11}, \ldots, y_{1k}$
$x_{21}, \ldots, x_{2p}, y_{21}, \ldots, y_{2k}$
$\ldots$
$x_{r1}, \ldots, x_{rp}, y_{r1}, \ldots, y_{rk}$

missä y_{ij} kertoo, montako kertaa olosuhteessa $\boldsymbol{x}_i$ havainto sattuu luokkaan j.

Oletetaan, että y_{ij}:t ovat riippumattomia ja Poisson-jakautuneita. Mallitetaan y-havainnot käyttäen log-linkkiä; selittäjinä x-arvot, dikotomisoitu luokkamuuttuja (indikaattorit luokille $2, \ldots, k$) sekä näiden yhdysvaikutukset. Yhdysvaikutusparametrien estimaatit antavat estimaatit käsittelyjen erotukselle multinomimallissa!

5.5 Log-lineaariset mallit

Tarkastellaan aluksi kahteen kaksiluokkaiseen muuttujaan A ja B liittyvää 2x2 kontingenssitaulua

$$\begin{pmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{pmatrix}$$

ja oletetaan, että aineisto on kerätty siten, että on realistista olettaa että lukumäärät $y_{11}, y_{12}, y_{21}, y_{22}$ ovat riipumattomia ja noudattavat Poisson-jakaumaa. Oletetaan, että solujen odotusarvojen logaritmit ovat (logaritmilinkki)

$$\begin{pmatrix} \log(\mu_{11}) = \beta_0 & \log(\mu_{12}) = \beta_0 + \beta_1 \\ \log(\mu_{21}) = \beta_0 + \beta_2 & \log(\mu_{22}) = \beta_0 + \beta_1 + \beta_2 + \beta_3. \end{pmatrix}$$

Siis kontingenssitaulun odotusarvo(matriisi)n logaritmi on

$$\log \begin{pmatrix} \mu_{11} & \mu_{12} \\ \mu_{21} & \mu_{22} \end{pmatrix} = \beta_0 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} + \beta_1 \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} + \beta_2 \begin{pmatrix} 0 & 0 \\ 1 & 1 \end{pmatrix} + \beta_3 \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$$

Muuttujat A ja B ovat riippumattomia, jos $\beta_3 = 0$, jota nyt siis voidaan testata Poisson-vastetta ja logaritmilinkkiä käyttäen.

HUOM. Jos kontingenssitaulu ehdollistetaan reunasummille $y_{1.} = y_{11} + y_{12}$ ja $y_{2.} = y_{21} + y_{22}$, niin y_{12} ja y_{22} ovat (ehdollisesti) riippumattomia ja noudattavat binomijakaumia $Bin(y_{1.}, p_1)$ja $Bin(y_{2.}, p_2)$, missä

$$o_1 = p_1/(1 - p_1) = \exp(\beta_1) \quad \text{ja} \quad o_2 = p_2/(1 - p_2) = \exp(\beta_1 + \beta_3).$$

Siis $\exp(\beta_3)$ on kyseiseen "binomikoevertailuun"liittyvä OR! (Jos $y_1 \sim Poi(\mu_1)$ ja $y_2 \sim Poi(\mu_2)$ ja y_1 ja y_2 riippumattomia $y = y_1 + y_2 \sim Poi(\mu)$, missä $\mu = \mu_1 + \mu_2$ ja $y_2|y \sim Bin(y, \mu_2/\mu)$).

			B			
A	1	...	j	...	b	
1						
$\vdots$						
i			π_{ij}			$\pi_{i.}$
$\vdots$						
a						
			$\pi_{.j}$			1

Taulukko 5.3: Populaatio

			B			
A	1	...	j	...	b	
1						
$\vdots$						
i			y_{ij}			$y_{i.}$
$\vdots$						
a						
			$y_{.j}$			y

Taulukko 5.4: Aineisto

Tarkastellaan samanaikaisesti kahta luokiteltua muuttujaa (A ja B), joiden kaksiulotteinen jakauma populaatiossa on annettu taulukossa 5.3. Oletetaan, että on kerätty y:n suuruinen empiirinen aineisto, johon liittyvä jakauma (kontingenssitaulukko) on taulukossa 5.4. Havaintoaineistoon (kontingenssitaulukkoon) liittyvä todennäköisyysjakauma riippuu luonnollisesti siitä, miten aineisto on kerätty.

Kontingenssitaulukkoihin liittyviä testejä:

- Yhteensopivuustestit: b riippumatonta otosta, otoskokoina $y_{\cdot 1}, \ldots, y_{\cdot b}$. Silloin

$$(y_{1j}, \ldots, y_{aj})|y_{\cdot j} \sim Multin\Big(y_{\cdot j}, \Big(\frac{\pi_{1j}}{\pi_{\cdot j}}, \ldots, \frac{\pi_{aj}}{\pi_{\cdot j}}\Big)\Big).$$

  Siis uskottavuusfunktio on

$$\prod_j P((y_{1j}, \ldots, y_{aj})|y_{\cdot j})$$

  ja vertailtavissa malleissa on $a-1$ ja $b(a-1)$ parametria.

- Riippumattomuustestit: Satunnaisotos kaksiulotteisesta luokitellusta jakaumasta:

$$(\ldots, y_{ij}, \ldots)|y \sim Multin(y, (\ldots, \pi_{ij}, \ldots))$$

  Uskottavuusfunktio on

$$P((y_{\cdot 1}, \ldots, y_{\cdot b})|y) \cdot \prod_j P((y_{1j}, \ldots, y_{aj})|y_{\cdot j})$$

  ja vertaltavissa malleissa on $(a-1)+(b-1)$ ja $ab-1$ parametria.

- Yhdysvaikutustestit: Loglineaarinen malli, missä $y_{ij} \sim Poi(\mu_{ij})$ ja y_{ij}:t riippumattomia. Uskottavuusfunktio on nyt

$$P(y) \cdot P(y_{\cdot 1}, \ldots, y_{\cdot b}|y) \cdot \prod_j P((y_{1j}, \ldots, y_{aj})|y_{\cdot j})$$

  ja vertailtavissa malleissa ("$A+B$" vs $A+B+A*B$; logaritmilinkki) on $a+b-1$ ja ab parametria. Kaikissa kolmessa tapauksessa uskottavuustestisuure on täsmälleen sama!!

Tarkastellaan nyt kolmeen muuttujaan (A, B ja C) liittyvää 3-ulotteista kontingenssitaulua ja merkitään solufrekvenssejä

$$y_{ijk}, \quad i = 1, \ldots, a;\ j = 1, \ldots, b;\ k = 1, \ldots, c.$$

Jos jälleen oletetaan solufrekvenssien riippumattomuus ja Poisson-jakautuneisuus, muuttujien A, B ja C riippumattomuus vastaa mallia

$$\mu_{ijk} = E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_k.$$

(Malli "$A + B + C$").

Malli "$A+B+C+A*B+A*C$", jossa on päävaikutusten lisäksi muuttujien A ja B sekä muuttujien A ja C väliset yhdysvaikutukset kuvailee tilanteen, jossa B ja C ovat ehdollisesti riippumattomia ehdolla A.

Mallien "$B + C$" ja "$B + C + B * C$" vertailua voidaan siis käyttää (vakiomattomaan tai standardoimattomaan) B:n ja C:n riippuvuuden tutkimiseen. Kuitenkin mahdollinen löydetty riippuvuus saattaa olla "sekoittavan tekijän"A aikaansaamaa ja sekoittavan tekijän eliminointi siis tapahtuu malleja

"$A+B+C+A*B+A*C$" ja "$A+B+C+A*B+A*C+B*C(+A*B*C)$"

vertaamalla.

5.6 Mallien vertailu

Kuten aikaisemmin, tarkastellaan "sisäkkäisiä"malleja M_1 (selittäjämatriisi $\boldsymbol{X}_1$) ja M_2 (selittäjämatriisi $\boldsymbol{X}_2$), ja oletetaan, että malli M_2 on saatu mallista M_1 lisäämällä siihen h uutta selittäjää. ($\boldsymbol{X}_2$ saadaan $\boldsymbol{X}_1$:stä lisäämällä siihen h uutta saraketta.)

Olkoot malleihin M_1 ja M_2 liittyvät ennustevektorit $\hat{\boldsymbol{\mu}}_1$ ja $\hat{\boldsymbol{\mu}}_2$. Silloin malleihin liittyvät logaritmisten uskottavuusfunktioiden maksimiarvot ovat

$$l(\hat{\boldsymbol{\mu}}_1) = c + \sum_{i=1}^{r} \{y_i \log(\hat{\mu}_{1i}) - \hat{\mu}_{1i}\}$$

ja

$$l(\hat{\boldsymbol{\mu}}_2) = c + \sum_{i=1}^{r} \{y_i \log(\hat{\mu}_{2i}) - \hat{\mu}_{2i}\}$$

Mallien vertailuun sopiva uskottavuusosamäärätestisuureen logaritmi (kahdella kerrottuna)

$$\chi^2 = 2(l(\hat{\boldsymbol{\mu}}_2) - l(\hat{\boldsymbol{\mu}}_1)) = 2 \sum_{i=1}^{r} \left\{ y_i \log\left(\frac{\hat{\mu}_{2i}}{\hat{\mu}_{1i}}\right) - (\hat{\mu}_{2i} - \hat{\mu}_{1i}) \right\}$$

on likimain $\chi^2(h)$-jakautunut, kun M_1 on tosi.

Malliin M (ja siihen liittyvään ennustevektoriin $\hat{\boldsymbol{\mu}}$) liittyvä **devianssi** on nyt

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = 2 \sum_{i=1}^{r} \left\{ y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i) \right\}.$$

Kuten aikaisemminkin, devianssi vertaa mallia M "täydelliseen"tai "saturoituun"malliin, jossa kaikki r (=profiilien tai luokkien lukumäärä) parametria ovat käytössä.

Ylläoleva (uskottavuusosamäärä)testisuure on siis **devianssien muutos** siirryttäessä mallista M_1 malliin M_2. Devianssia tai **Pearsonin χ^2-testisuuretta**

$$\chi^2 = \sum_{i=1}^{r} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

voidaan (tiettyjen ehtojen vallitessa) käyttää lopullisen mallin M "hyvyyden"tutkimiseen. Yleisten ehtojen vallitessa molemmat suureet ovat (jos testattava malli on oikea) "asymptoottisesti ekvivalentteja"ja niiden rajajakauma on χ^2-jakauma vapausasteilla

$$r \; - \; \text{mallin } M \text{ estimoitujen parametrien lkm.}$$

Ylihajontaongelma

Mallin M (tai siihen liittyvän estimaatin $\hat{\mu}$) hyvyyttä voi siis karkeasti tutkia suureen

$$\frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{r - p}$$

avulla, missä r on profiilien lukumäärä ja p mallin M parametrien lukumäärä. Jos malli M on oikea, voidaan odottaa, että yo. suure on lähellä (likimääräistä) odotusarvoaan 1. "Ylihajonnan"syy voi olla mallin virheellisyys (tärkeitä selittäjiä puuttuu) tai tapausten (joiden lukumäärä toimii vasteena) ajallinen tai paikallinen klusterisoituminen (vastoin Poissonprosessioletusta).

Jos ylihajontaa esiintyy, on suositeltavaa kertoa $\hat{\beta}$:n (väärään oletukseen perustuvan) kovarianssimatriisiestimaatti "ylihajontaa"estimoivalla kertoimella

$$\frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{r - p}$$

tai

$$\frac{1}{r - p} \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Kirjallisuutta

[1] Dobson, A. J. and Barnett, A. (2008). *An Introduction to Generalized Linear Models*, 3rd edn. London: Chapman & Hall.

[2] Dunteman, G. H. and Ho, M.-H. R. (2006). *An Introduction to Generalized Linear Models*. SAGE QASS Series.

[3] Gill, J. (2000). *Generalized Linear Models: A Unified Approach*. SAGE QASS Series.

[4] Hardin, J. W. and Hilbe, J. M. (2007). *Generalized Linear Models and Extensions*, 2nd edn. College Station, TX: Stata Press.

[5] Liao, T. F. (1994). *Interpreting probability models : logit, probit and other generalized linear models*. SAGE QASS Series.

[6] McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd edn. London: Chapman & Hall.

Liite A

Funktion numeerinen maksimointi

A.1 Newtonin-Raphsonin menetelmä

Olkoon ensin $l : \mathbb{R} \to \mathbb{R}$ ja halutaan etsiä $l(b)$:n maksimikohta. Merkitään $s(b) = l'(b)$ ja $j(b) = -l''(b)$ (oletetaan, että derivaatat olemassa). Yhtäpitävää on (usein) etsiä $s(b)$:n nollakohta. Olkoon $b_{(k)}$ funktion $s(b)$ nollakohtaan liittyvä k :s arvaus. Silloin seuraava "parempi"arvaus muodostetaan säännöllä

$$b_{(k+1)} = b_{(k)} + \frac{s(b_{(k)})}{j(b_{(k)})}, \quad k = 0, 1, 2, \ldots .$$

Iterointia jatketaan kunnes haluttu tarkkuus tai iteraatioiden maksimimäärä (ennalta annettu) saavutetaan. Alkuarvo $b_{(0)} =?$.

Maksimoidaan seuraavaksi funktiota $l : \mathbb{R}^p \to \mathbb{R}$ eli etsitään arvo $\hat{\boldsymbol{b}}$, joka maksimoi funktion $l(\boldsymbol{b})$. Olkoon $\boldsymbol{S}(\boldsymbol{b})$ ensimmäisten derivaattojen muodostama vektoriarvoinen funktio ja $(-1)\boldsymbol{\mathcal{J}}(\boldsymbol{b})$ toisten derivaattojen muodostama matriisiarvoinen funktio. Silloin Newtonin-Raphsonin menetelmän mukainen ratkaisua kohden konvergoiva jono saadaan säännöllä

$$\boldsymbol{b}_{(k+1)} = \boldsymbol{b}_{(k)} + \boldsymbol{\mathcal{J}}^{-1}(\boldsymbol{b}_{(k)})\boldsymbol{s}(\boldsymbol{b}_{(k)}), \quad k = 0, 1, 2, \ldots .$$

Alkuarvo $\boldsymbol{b}_{(0)} =?$.

A.2 IRLS-menetelmä