Tilastollinen mallintaminen, syksy 2000
Harjoituksia

1.1. Oletetaan, että satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n$ ovat riippumattomia ja noudattavat samaa jakaumaa (otos jakaumasta). Kirjoita jakauman tiheysfunktio (tf) ja logaritmoitu uskottavuusfunktio (uf) eksponentiaalisen jakuman standardimuodossa (ks. Altham, s. 15), kun

(a) $Y_i \sim N(\mu, \sigma^2)$

(b) $Y_i \sim Po(\mu)$

(c) $Y_i \sim Bi(r, \pi),\ i = 1, 2, \ldots, n.$

1.2. Sama kuin 1. tehtävä, mutta nyt oletetaan, että

(a) $Y_i \sim N(\mu_i, \sigma^2)$

(b) $Y_i \sim Po(\mu_i)$, missä $\mu_i = \beta_0 + \beta_1 x_i,\ i = 1, 2, \ldots, n$

(c) $Y_i \sim Bi(r_i, \pi_i),\ i = 1, 2, \ldots, n$, missä $log(\frac{\pi_i}{1-\pi_i}),\ i = 1, 2, \ldots, n.$

1.3. (a) Määritä 1. tehtävän jakaumien odotusarvot ja varianssit Apulauseen 1 avulla (Altham, s. 15, Lemma 1).

(b) 2. tehtävän jakaumien odotusarvot ja varianssit Apulauseen 1 avulla .

1.4. Althamin esimerkki sivulta 6. Tee pyydetyt täydennykset.

1.5. Satunnaismuuttujat $Y_1, Y_2, \ldots, Y_n$ ovat riippumattomia ja noudattavat Poissonin jakaumaa $Y_i \sim Po(\mu),\ i = 1, 2, \ldots, n.$

(a) Osoita, että parametrin μ suurimman uskottavuuden estimaatti (sue) $\hat{\mu} = \sum Y_i/n$ ja $\mathrm{var}(\hat{\mu}) = \mu/n$.

(b) Mitä on $\mathrm{E}(-\frac{\partial^2 L}{\partial \mu^2})$?

(c) Satunnaismuuttujan $n\hat{\mu}$ täsmällinenjakauma?

(d) Satunnaismuuttujan $\hat{\mu}$ asymptoottinen jakauma?

(Ks. verkosta Example Sheet 1, tehtävä 1 ja Outline solutions to Example Sheet 1.)

1.6. Tarkastellaan uusien kuukausittain (36 kk) rekisteröityjen AIDS-tapausten määriä Englannissa marraskuuhun 1985 asti (Arkki 10, Altham s. 36). Oletetaan, että uusien tapausten lukumäärät $Y_i \sim Po(\mu_i),\ \mu_i \geq 0, i = 1, \ldots, 36$ ovat riippumattomia ja noudattavat Poissonin jakaumaa.

(a) Estimoi malli, jossa odotusarvojen logaritmilla on lineaarinen trendi (selittäjänä kuukauden järjestysluku).

(b) Laske mallin antamat odotusarvojen estimaatit ja piirrä odotusarvokäyrä pisteparveen.

(c) Testaa lineaarisen trendin hypoteesi ω_c (a-kohta) täyttä mallia ω_f vastaan

1.7. Todenna edellisessä tehtävässä, että devianssi on

(a)

$$2\sum_{i=1}^{36} y_i log\frac{y_i}{e_i},$$

missä e_i on mallista estimoitu i. kuukauden odotusarvo.

(b) Laske devianssi suoraan edellä mainitulla kaavalla ja totea, että saadaan sama tulos kuin `glm()`:n tulostuksessa.

(c) Mahtaako devianssin jakauma-aproksimaatio olla hyvä tässä esimerkissä? Tee analyysi, jossa aproksimaatio on oletettavasti parempi.

1.8. Jatkoa edelliseen. Laske (numeerisesti) vaihtoehtoisten mallien (ω_c ja ω_f vallitessa uskottavuusfunktion maksimi ja niiden avulla uskottavuussuhde R_n. Totea, että $2logR_n$: arvo on sama kuin devianssi `glm()`:n tulostuksessa.

1.9. Olkoot $X \sim Po(\lambda)$ ja $Y \sim Po(\mu)$ toisistaan riippumattomia. Silloin $X + Y \sim Po(\lambda + \mu)$. Osoita, että X :n ehdollinen jakauma ehdolla $X+Y = N$ noudattaa binomijakaumaa $Bi(N, \pi)$, missä $\pi = \lambda/(\lambda+\mu)$.

1.10. Olkoot $X_i \sim Po(\lambda_i),\ i = 1, \ldots, k$ toisistaan riippumattomia. Silloin $X_1+X_2+\ldots+X_k \sim Po(\lambda_1+\lambda_2+\ldots+\lambda_k)$. Osoita, että satunnaismuuttujien $X_1, X_2, \ldots, X_k$ ehdollinen yhteisjakauma ehdolla $X_1+X_2+\ldots+X_k = N$ noudattaa multinomijakaumaa (ks. Althan s. 42) parametrein N ja $\pi_1, \pi_2, \ldots, \pi_k$, missä $\pi_i = \lambda_i/(\lambda_1 + \lambda_2 + \ldots + \lambda_k),\ i = 1, 2, \ldots, k$.

1.11. Oletetaan AIDS-tehtävässä identiteettilinkki ja Poissonin jakauma (ts. "tavallinen" lineaarinen malli, mutta Poissonin jakauma. Johda Apulausetta 2 ja sen jälkeen esitettyjä lausekkeita (Altham s. 17-18) soveltaen kertoimien estimaattorin asymptoottinen kovarianssimatriisi. Laske numeerisesti kovarianssimatriisin estimaatti.

1.12. Sama kuin edellinen tehtävä, mutta log-linkkiä käyttäen.

1.13. Tarkastellaan työarkilla WS_lm2 esitettyä aineistoa, jossa tutkitaan synteettisten kuitujen vetolujuuden y riippuvuutta kuituihin sekoitetun puuvillan prosenttiosudesta. Selittävän faktorin a tasot (5 tasoa) ovat puuvillan %-osuudet (15%,20%,25%, 30% ja 35%) kuidussa. Estimoi faktorin a ("käsittelyn") vaikutukset käyttäen seuraavia kontrasteja: `contr.helmert, contr.sum, contr.treatment`. Tulkitse tulokset.

1.14. Tee edellisen tehtävän estimoinnit identifiointirajoitteita käyttäen. Ts. määritä jokaista kontrastia vastaava identifiointirajoite, sijoita luennoissa esitettyyn kaavaan (RM) ja estimoi parametrit.

1.15. Tarkastellaan aineistoa, jossa on annettu tieliikenneonnettomuudet Cambridgessa 1978-81 kahdelta tieosuudelta kolmena eri vuorokauden aikana (ks. Altham s. 39). Aineisto on annettu myös arkilla WS10. Kuinka onnettomuudet riippuvat liikenteen määrästä, tiestä ja vuorokaudenajasta? Oletetaan, että onnettomuuksien lukumäärät $Y_{ij} \sim Po(\mu_{ij})$, $i = 1, 2; j = 1, 2, 3$ ovat toisistaan riippumattomia, missä i on tieindeksi ja j aikaindeksi. Tutki jokaisen selittäjän merkitsevyys. Voidaanko mallia jotenkin yksinkertaistaa?

1.16. The Daily Telegraph (28.10.1988) julkaisi artikkelin "Johtajat alttiita ajamaan humalassa". Artikkelissa esitettiin Royal Ascotin ja Henley Regatan yhteydessä järjestettyjen puhallustestien tuloksia (Royal Ascot ja Henley Regatta ovat tunnetusti "kosteita" englantilaisia urheilutapahtumia niinkuin hevos- ja soutukilpailutkin).

(a) Estimoi mallit `saturated` ja `independence` (ks. WS11).

(b) Vertaile mallien antamia tuloksia ja tulkitse.

(c) Jääkö RA:ssa todennäköisemmin kiinni kuin HR:ssä (jos puhallutetaan)?

RA:ssa puhallutettiin kaikkiaan 24+2210 ja HR:ssä 5+680 henkilä. RA:ssa oli siis 24 positiivista ja HR:ssa 5(ks. Altham s. 42 ja WS11).

1.17. Tarkastellaan edellisen tehtävän aineistoa binomijakauman avulla.

(a) Estimoidaan mallit `sat` ja `indep` (ks. WS11).

(b) Vertaile mallien antamia tuloksia ja tulkitse.

1.18. Otetaan sukupuoli mukaan edellisen tehtävän tarkasteluun. Oletetaan, että havainnot ovat muotoa $24 = 23 + 1,\ 2210 = 2 + 2208,\ 5 = 3 + 2$ ja $680 = 340 + 340$, missä jokaisen summan 1. luku on miesten lkm ja 2. luku naisten lkm.

(a) Miten naisten pidätysaste vaihtelee Ascotin ja Henleyn välillä?

(b) Onko miesten pidätysasteessa Ascotin ja Henleyn välillä?

(c) Miten siis riippuvuus muuttuu, kun sukupuoli otetaan mukaan analyysiin?

Tämä on esimerkki ns. Simpsonin paradoksista, jota Altham kutsuu "Yulen paradoksiksi". G.U.Yule kirjoitti asiasta jo 1900-luvun alussa ja Simpson vasta n. 50 vuotta myöhemmin vuonna 1951 (ks. Altham s. 49).

1.19. Tarkastellaan anoreksia aineistoa (`data anorexia, library mass`), jossa on 72 nuorta naispotilasta. Tutki hoitojen vaikutusta. Selitettävänä on painon muutos ja selittäjänä muuttuja `Treat` (Factor of three levels: Cont (Control), CBT (Cognitive Behavioural Treatment) and FT (Family treatment)). Ks. lisää WS_glm1 ja kysy aineistosta `help(anorexia)`.

   (a) Estimoi mallit `ax.1` ja `ax.0` (ks. WS_glm1).
   (b) Muodosta erotusmuuttuja `Postwt - Prewt` ja selitä sitä muuttujilla `Treat` ja Prewt sekä tulkitse.
   (c) Vertaile edellisten kohtien tuloksia. Mikä on offset-muuttujan vaikutus?

   Huomaa, että `glm` määrityksellä `family=gaussian` on lineaarinen malli (vrt. `lm`).

1.20. Tarkastellaan Hosmerin ja Lemeshowin (1989, s. 91) aineistoa, jossa selitämme alhaista syntymäpainoa logistisella regressiolla (`library(mass), data(birthwt), help("birthwt")`). Katso WS_glm2, jossa tehdään uusia muuttujia ja analysoitava aineisto `bwt`.

   (a) Estimoi malli `birthwt.glm`, jossa on mukana kaikki selittäjät (ks. WS_glm2). Mikä on paras selittäjä?
   (b) Pudotetaan mallista `birthwt.glm` yksi selittäjä. Tutki tulostuksen ja `update`-funktion avulla, mikä näistä pienemmistä malleista olisi paras.
   (c) Miten AIC-käyttäytyy valinnassa?

1.21. Tutkitaan edelleen `bwt`-aineistoa. Vertaillaan malleja niihin liittyvien testisuureiden avulla.

   (a) Tutki funktion `dropterm` avulla (`dropterm(birthwt.glm, test= "Chisq")`), mikä mallia `birthwt.glm` pienempi malli olisi paras.
   (b) Estimoi paras yhden selittäjän malli ja hae parasta mallia täydentämällä yhden selittäjän mallia funktion `addterm` avulla (`addterm(object, test= "Chisq")`).
   (c) Vertaile kohtien (a) ja (b) tuloksia. Tulkitse.

1.22. Tee edellisen tehtävän tarkastelut, mutta jätä funktioista `dropterm` ja `addterm` määre `(, test= "Chisq")` pois.

(a) Hae paras malli pelkästään AIC:n arvon perusteella.

(b) Hajota AIC uskottavuusfunktiosta tulevaan osuuteen sekä mallin koosta ja havaintojen lukumäärästä tulevaan komponenttiin. Miten AIC käyttäytyy muuttujia lisättäessä tai poistettaessa?

1.23. Estimoidaan nyt AIC:n mielessä paras malli `bwt`-aineistosta.

(a) Estimoi paras malli `birthwt.step` käyttäen funktiota `stepAIC, library(mass)`(ks. WS_glm2).

(b) Estimoi paras malli birthwt.step2 `stepAIC`:n avulla lähtien mallista, jossa on mukana myös kahden muuttujan yhdysvaikutukset ja muuttujien `age` ja `lwt` 2. potenssit.

1.24. Vertaile kahden mielestäsi parhaan mallin ennustuskykyä jonkin valitsemasi ennustekriteerin suhteen.

1.25. Selitetään edelleen alhaista syntymäpainoa `low bwt`-aineiston muuttujilla. Edellisessä tehtävässä haettiin parasta mallia `AIC`-criteerin avulla. Hae nyt kriteerin

$$BIC = -2 \times log(uf : n\ maksimi) + k \times log(n)$$

mielessä paras malli. Vertaile `AIC`:llä ja `BIC`:llä saatuja malleja toisiinsa. (Ks. WS_glm2 ja `library(mass) ja help("stepAIC")`).

1.26. (Jatketaan edellisen tehtävän tarkastelua.) Vertaile `AIC`:n ja `BIC`:n mielessaä parhaiden mallien ennustekykyä mallien antamien sovitteiden (todennäköisyyksiä) avulla.

1.27. Analysoi `anorexia`-aineistoa (vrt. tehtävä 1.19). Tarkastele graafisesti painoa ennen ja jälkeen hoidon jokaisessa ryhmässä. Piirrä jokaisessa ryhmässä oma regressiosuora ja vertaa sitä koko aineiston regressiosuoraan. Sopiiko suora aineistoon vai voisiko ehdottaa jotain muuta mallia?

1.28. Oletetaan, että havainnot noudattavat normaalijakaumaa

$$Y_i \sim N(\beta_0 + \beta_1 x_i + \beta_2 z_i, \sigma^2),\ i = 1, 2, \ldots, n$$

ja ovat toisistaan riippumattomia (Ks. Altham s.25 - 27).

(a) Kirjoita otoksen $Y_1, Y_2, \ldots, Y_n$ uskottavuusfunktio.

(b) Uskottavuusfunktion maksimi, kun σ^2 tunnetaan.

(c) Uskottavuusfunktion maksimi, kun σ^2 tuntematon (estimoidaan).

(d) `AIC`?

1.29. Oletetaan, että satunnaismuuttujat Y_i ovat riippumattomia ja noudattavat Poissonin jakaumaa $Po(\mu_i),\ i = 1, 2, \ldots, n$. Malli on

$$H:\ log(\mu_i) = \alpha + \beta x_i$$

missä $x_1, x_2, \ldots, x_n$ ovat annettuja (Ks. Altham s. 38-39). Kirjoita

(a) logaritmoidun uskottavuusfunktion $log f(\mathbf{y}; \alpha, \beta)$ lauseke ($\mathbf{y} = (y_1, y_2, \ldots, y_n)$ on havaintovektori).

(b) yhtälö, josta suurimman uskottavuuden estimaatit $(\hat{\alpha}, \hat{\beta})$ voidaan ratkaista.

(c) ilmaisun

$$\max_{\beta=0} f(\mathbf{y}; \alpha, \beta)$$

lauseke.

(d) Kuinka edellisen kohdan tulosta käytetään yhdessä Wilksin lauseen (Altham s. 4 ja s. 9) kanssa hypoteesin "$H_0:\ \beta = 0$" testaamiseen?

1.30. Tarkastellaan entsyymireaktion alkukiihtyvyyden (lkm/min^2) riippuvuutta substraatin (pohja-aineen) entsyymipitoisuudesta (miljoonasosia). Reaktiossa syntyvän radioaktiivisen tuotteen antamien merkkien lukumäärä mitattiin entsyymipitoisuuden (`conc`) funktiona ja laskettiin reaktion alkukiihtyvyys (`rate`). Kokeessa vertailtiin *Puromysiinillä* käsitellyllä entsyymillä saatuja tuloksia vertailukokeeseen, jossa entsyymiä ei oltu käsitelty. Aineisto on kirjastossa `nls` (ks. `library(nls)` ja `help(Puromycin)`). Michaelisin ja Mentenin mallissa alkukiihtyvyyden Y riippuvuus pitoisuudesta ilmaistaan yhtälöllä

$$f(x, \beta) = \frac{\beta_1 x}{\beta_2 + x},$$

missä $\beta = (\beta_1, \beta_2)'$.

(a) Derivoi f parametrien β_1 ja β_2 suhteen. Kommentoi mallin lineaarisuutta/epälineaarisuutta.

(b) Muodosta funktio $\frac{1}{f}$ ja muodosta uudet parametrit $\theta_1 = \frac{1}{\beta_1}$ ja $\theta_2 = \frac{\beta_2}{\beta_1}$. Onko funktio $\frac{1}{f}$ parametrien θ_1 ja θ_2 suhteen lineaarinen?

1.31. Sovita Michaelisin ja Mentenin malli Puromycinillä käsiteltyyn aineiston osaan (aineisto `Puromycin`). Tarkastele mallin sopivuutta aineistoon.

1.32. Sovita Puromycinillä käsiteltyyn aineiston osaan (aineisto `Puromycin`) lineneaarinen regressiomalli, kun selitettävä muuttuja on $\frac{1}{y}$ ja selittäjä $\frac{1}{x}$ (x on pitoisuus (`conc`) ja y kiihtyvyys (`rate`)). Tarkastele mallin sopivuutta aineistoon ja vertaile edellisen tehtävän tuloksiin.

1.33. Tutki, onko Puromycin-käsittelyllä vaikutusta. Tarkastele erojen tilastollista merkitsevyyttä.

1.34. Sovita aineistoihin `Puromycin` sekä `wtloss` jokin mielestäsi hyvä polynomimalli ja vertaa niiden sopivuutta tarkasteltuihin epälineaarisiin malleihin.

1.35. Tarkastellaan 1. asteen B-spliniä käsittelevää esimerkkiä (apumateriaalia työarkilla `WS_sr`). Oletetaan, että x saa kokonaislukuarvot väliltä $[0, 9]$ (`x <- 0:9`) ja solmut ovat pisteissä 3 ja 6 (ts. väli on jaettu kolmeen yhtä suureen osaväliin). Funktio `bs(x,knots=c(3,6),degree=1)` tulostaa kantafunktioiden $B_1(x), B_2(x)$ $B_3(x)$ arvot `x`:n arvoilla $0, 1, \ldots, 9$. Kantafunktiot ovat muotoa $a_i x + b_i,\ i = 1, 2, 3$ olevia paloittain lineaarisia funktioita. Tulosta `bs(x,knots=c(3,6),degree=1)` ja päättele tulostuksesta kantafunktioiden lausekkeet eli parametrien $a_i, b_i,\ i = 1, 2, 3$ arvot.

1.36. Jatketaan edellisen tehtävän tarkastelua. Olkoot x:n arvoja vastaavat y:n arvot $1, 2.5, 5, 4, 2, 2.1, 5, 7, 11, 11$. Sovitetaan aineistoon edellisessä tehtävässä määritelty B-splini eli estimoidaan muotoa `lm(y ~ bs())` oleva lineaarinen malli.

(a) Tulkitse tulos.

(b) Selitä, miten mallin sovite lasketaan.

(c) Kokeile yhden solmun B-spliniä. Koeta valita solmu siten, että yhteensopivuus dataan olisi mahdollisimman hyvä.

1.37. Työarkilla `WS_sr` on annettu erään kuusen halkaisijat (paksuus) $30cm$:n välein tyvestä lähtien. Mallinna kuusen runkokäyrä ($d = f(x)$, missä x on etäisyys tyvestä ja d halkaisija)

(a) jollain sopivalla 1. asteen splinillä

(b) 3. asteen splinillä

(c) 3. asteen polynomilla.

1.38. Aineistossa `mcycle` (`data(mcycle), library(MASS)`) on havaintosarja simuloiduista moottoripyöräonnettomuuksista, joissa on mitattu pään kiihtyvyys törmäyksen jälkeen eri ajanhetkinä. Sovita aineistoon tasoitussplini (`library(modreg) , smooth.spline()`). Tutki tasoitussplinin (`smooth.spline()`) käyttäytymistä tasoitusparametrin λ funktiona.

1.39. Muodosta luonnollisen kuutiosplinin (`ns(), library(splines)`) avulla `mcycle`-aineistoon lineaarinen malli (`lm()`) ja vertaile tulosta tasoitussplinin antamaan tulokseen.