Luku 9

Lineaarinen regressio

9.1 Johdanto

Oletamme, että havaintoaineisto koostuu kahden muuttujan x ja y arvoista

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n),$$

siten että pari (x_i, y_i) on mitattu havaintoyksiköstä i. Emme aluksi lainkaan aseta tilastollista mallia, vaan tarkastelemme tilannetta heuristisesti.

Päämääränä regressiomalleissa on kuvailla, kuinka *selittävä muuttuja* x (engl. *explanatory variable*, *independent variable*) vaikuttaa *selitettävään muuttujaan* eli *vasteeseen* y (engl. *dependent variable*, *response*). Yleisemmässä tilanteessa selittäviä muuttujia voisi olla usampia, mutta tässä kappaleessa selittäjiä on vain yksi.

Ajattelemme, että x vaikuttaa y:n arvoon osapuilleen kaavan

$$y = f(x)$$

mukaisesti, jossa f on funktio, jonka muoto on ainakin osittain tuntematon. Tarkemmin sanoen funktion arvo $f(x)$ riippuu x:n arvon lisäksi tuntemattomien parametrien arvoista vaikka tätä ei olla merkinnöissä huomioitu. Erilaisten virhelähteiden takia pisteet (x_i, y_i) eivät kuitenkaan tarkalleen asetu funktion f kuvaajalle millään parametrin arvolla, minkä takia tyydymme malliin

$$y = f(x) + \epsilon,$$

jossa muuttuja ϵ tarkoittaa virhettä. Funktion f muoto (ts. parametrien arvot) pitäisi määrittää havaintojen perusteella.

Tämän yleisen rakennemallin $f(x)$ sijasta tarkastelemme lineaarista lauseketta

$$\alpha + \beta\, x,$$

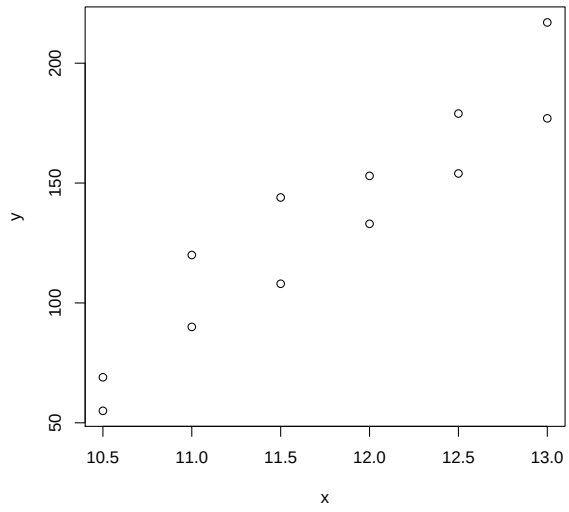
jonka kuvaaja on suora. Jos x-arvojen vaihteluväli on pieni, niin melkein mitä tahansa funktiota $f(x)$ voidaan approksimoida tällaisella lineaarisella funktiolla.

Lineaarisessa rakennemallissa $\alpha + \beta\, x$ parametri α on suoran vakiotermi (engl. *intercept*) ja parametri β on suoran kulmakerroin (engl. *slope*). Ne ovat tuntemattomia parametreja, joiden arvot pitää määrittää havaintojen perusteella. Lähdemme seuraavaksi etsimään tälle tehtävälle toimivaa ratkaisua.

Taulukko 9.1 Kuminauhan lentomatkoja y eri venytyksen x arvoilla.

x	y	x	y
11	120	10.5	69
12	153	10.5	55
13	217	11.5	108
13	177	11.5	144
12	133	12.5	154
11	90	12.5	179

Kuva 9.1 Kuminauhan lentomatkoja y eri venytyksen x arvoilla.

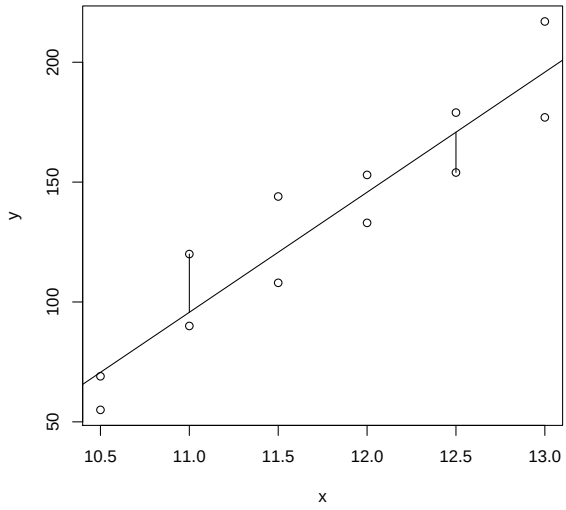


Analysoimme tässä kappaleessa kurssin luennoitsijan kotonaan mittaamaa aineistoa, jossa tutkittiin, miten pitkälle kuminauha lentää, jos sitä ensin venytetään ja sitten sen toinen pää päästetään vapaaksi. Kuminauhaa venytettiin siten, että sen toista päätä pingoitettiin viivottimen nollapisteen puoleiseen päähän ja venytystä kontrolloitiin asettamalla toinen pää tiettyyn kohtaan x (cm) viivoittimen asteikolla. Viivoitin ja kuminauha pidettiin vaakasuorassa 50 cm:n korkeudessa lattialta (pöydän avulla), ja lentomatka y (cm) mitattiin katsomalla, miten kauas kuminauhan kauimmainen kohta päätyi siitä lattian pisteestä, jonka yläpuolella viivottimen nollapisteen puoleinen pää oli. Mittaukset ovat taulukossa 9.1, ja kuvassa 9.1 ne esitetään hajontakuvana.

9.2 Suoran sovittaminen pienimmän neliösumman menetelmällä

Emme vieläkään aseta tilastollista mallia, vaan tarkastelemme edelleen tehtävää heuristisesti. Parametrit eli kertoimet α ja β voidaan määrittää sellaisella taval-

Kuva 9.2 Kuminauha-aineistoon sovitettu pienimmän neliösumman suora $a + bx$, jossa $a = -455.3$ ja $b = 50.09$.



la, jossa yritetään saada vasteet y_i sekä niitä vastaavat sovitteet tai ennusteet $\alpha+\beta x_i$ mahdollisimman samankaltaisiksi jonkin kriteerin mielessä. Tässä yhteydessä samankaltaisuutta olisi mahdollista mitata erilaisilla tavoilla, ja erilaiset tavat johtaisivat erilaisin parametriestimaatteihin.

Käytännössä tärkein kriteeri on virheiden neliöiden summa

$$\mathrm{SS}(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2. \tag{9.1}$$

Tässä $|y_i - \alpha - \beta x_i|$ on pisteen (x_i, y_i) y-akselin suunnassa mitattu etäisyys suorasta $\alpha + \beta x$. Kriteerissä $\mathrm{SS}(\alpha, \beta)$ nämä y-akselin suuntaiset etäisyydet neliöidään ja sitten ne summataan yhteen. Tämän kriteerin mielessä paras suora eli pienimmän neliösumman suora on se suora, jota vastaavat kertoimet α ja β minimoivat virheiden neliöiden summan (9.1). Voimme myös sanoa, että sovitamme suoran aineistoon käyttämällä pienimmän neliösumman menetelmää (engl. *method of least squares*) eli PNS-menetelmää. Kuvassa 9.2 on piirretty kuminauha-aineisto sekä siihen sovitettu PNS-suora. Lisäksi kahdelle (x, y)-pisteelle on piirretty sen y-akselin suuntaan mitattu etäisyys PNS-suorasta.

Kriteeri (9.1) on tärkeä toisaalta sen takia, että se johtaa yksinkertaisiin kaavoihin ja toisaalta sen takia, että sen käyttö voidaan perustellä myöhemmin esitettävällä yksinkertaisella tilastollisella mallilla sekä suurimman uskottavuuden periaattella.

Todistamme jakson lopussa, että PNS-menetelmä valitsee kulmakertoimelle β arvon

$$\hat{\beta} = b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}. \tag{9.2}$$

Tässä tuttuun tapaan $\bar{x}$ ja $\bar{y}$ tarkoittavat keskiarvoja

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Jotta b olisi hyvin määritelty, täytyy olettaa, että kaavan (9.2) nimittäjä on aidosti positiivinen. Tämä on voimassa, mikäli jonossa $x_1, \ldots, x_n$ on vähintään kaksi erisuurta arvoa, ja tämän ehdon oletamme jatkossa. Kun kulmakerroin on laskettu, niin PNS-menetelmä antaa vakion α arvoksi

$$\hat{\alpha} = a = \bar{y} - b\bar{x}. \tag{9.3}$$

Huomaa että tuntemattomia parametreja (tai kertoimia) merkitään kreikkalaisilla kirjaimilla, ja niiden estimaatteja voidaan merkitä joko laittamalla hattu parametrin päälle tai sitten käyttämällä kreikkalaista kirjainta vastaavaa latinalaista kirjainta.

PNS-suoran kulmakertoimelle voidaan antaa tulkinta otoskovarianssin ja otosvarianssin avulla. Määrittelemme vektoreista

$$\mathbf{x} = (x_1, \ldots, x_n), \qquad \mathbf{y} = (y_1, \ldots, y_n).$$

lasketun otoskovarianssin $s(\mathbf{x}, \mathbf{y})$ kaavalla

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}). \tag{9.4}$$

Otoskovarianssissa käytetään jakajana lukua $n-1$ sen takia, että näin saadaan populaatiokovarianssin harhaton estimaattori. Jos X ja Y ovat satunnaismuuttujia, niin niiden kovarianssi (eli populaatiokovarianssi) $\operatorname{cov}(X, Y)$ määritellään odotusarvona

$$\operatorname{cov}(X, Y) = E[(X - EX)(Y - EY)]. \tag{9.5}$$

Jos $(X_1, Y_1), \ldots, (X_n, Y_n)$ on satunnaisotos satunnaismuuttujaparin (X, Y) jakaumasta, niin helpohkoilla laskuilla nähdään, että

$$Es(\mathbf{X}, \mathbf{Y}) = \operatorname{cov}(X, Y),$$

missä tietenkin $\mathbf{X} = (X_1, \ldots, X_n)$ ja $\mathbf{Y} = (Y_1, \ldots, Y_n)$.

PNS-suoran kulmakerroin (9.2) saadaan lausuttua otoskovarianssin (9.4) avulla kaavalla

$$b = \frac{s(\mathbf{x}, \mathbf{y})}{s(\mathbf{x}, \mathbf{x})}.$$

Osoittaja on x- ja y-lukujen otoskovarianssi, ja nimittäjä on x-lukujen otosvarianssi. PNS-suoran yhtälö $y = a + bx$ voidaan esittää myös kaavalla

$$y - \bar{y} = b(x - \bar{x}), \tag{9.6}$$

mistä nähdään, että se kulkee parien (x_i, y_i) keskiarvon $(\bar{x}, \bar{y})$ kautta. Tämä muotoilu on helpompi muistaa kuin PNS-suoran vakiotermin kaava (9.3), ja tästä muotoilusta nähdään helposti oikea lauseke PNS-suoran vakiotermille.

Esimerkki 9.1 R:ssä vektorien `x` ja `y` otoskovarianssi saadaan laskettua kutsulla `cov(x, y)`, ja vektorin otoskovarianssi saadaan laskettua joko kutsulla `var(x)` (tai kutsulla `cov(x, x)`). Tällä keinolla PNS-suoran kertoimet saadaan laskettua itse helposti. Otoskovarianssin saa toki laskettua myös suoraan määritelmää käyttämällä.

```
x <- c(11, 12, 13, 13, 12, 11, 10.5, 10.5, 11.5, 11.5, 12.5,
    12.5)
y <- c(120, 153, 217, 177, 133, 90, 69, 55, 108, 144, 154, 179)
print(c(mean(x), mean(y), cov(x, y), var(x)))

## [1]  11.7500 133.2500  39.8409   0.7955

print(omaSxy <- 1/(length(x) - 1) * sum((x - mean(x)) * (y -
    mean(y))))

## [1] 39.84

print(b <- cov(x, y)/var(x))

## [1] 50.09

print(a <- mean(y) - b * mean(x))

## [1] -455.3
```

Toki R:stä löyty myös valmis funktio `lm`, jolla PNS-suora saadaan helposti laskettua.

```
print(lm(y ~ x))

##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      -455.3         50.1
```

△

PNS-suoran kertoimien kaavojen todistus

Todistamme kaavat (9.2) ja (9.3) suoralla laskulla. Todistus perustuu siihen tosiasiaan, että $\mathrm{SS}(\alpha, \beta)$ on kertoimien suhteen toisen asteen polynomi, jota voidaan analysoida neliöksi täydentämällä.

Lähdemme liikkeelle esityksestä

$$y_i - \alpha - \beta x_i = (y_i - \bar{y}) - (\alpha - \bar{y} + \beta\bar{x}) - \beta(x_i - \bar{x}),$$

jonka avulla saamme

$$\begin{aligned}\mathrm{SS}(\alpha,\beta) &= \sum_{i=1}^{n}\left[(y_i-\bar{y})-(\alpha-\bar{y}+\beta\bar{x})-\beta(x_i-\bar{x})\right]^2\\ &= \sum(y_i-\bar{y})^2 + n(\alpha-\bar{y}+\beta\bar{x})^2 + \beta^2\sum(x_i-\bar{x})^2\\ &\quad -2(\alpha-\bar{y}+\beta\bar{x})\sum(y_i-\bar{y})\\ &\quad -2\beta\sum(x_i-\bar{x})(y_i-\bar{y})\\ &\quad +2(\alpha-\bar{y}+\beta\bar{x})\beta\sum(x_i-\bar{x}).\end{aligned}$$

Tässä kerrottiin trinomin neliö auki, ja summaoperaattorin alta otetiin pois vakiona pysyvät termit. Koska

$$\sum(x_i-\bar{x}) = \sum x_i - n\bar{x} = 0, \quad \text{ja} \quad \sum(y_i-\bar{y}) = \sum y_i - n\bar{y} = 0,$$

niin kaksi funktion $\mathrm{SS}(\alpha,\beta)$ esityksen kuudesta termistä häviää. Otamme käyttöön merkinnät

$$q_{xx} = \sum(x_i-\bar{x})^2, \quad q_{xy} = \sum(x_i-\bar{x})(y_i-\bar{y}), \quad q_{yy} = \sum(y_i-\bar{y})^2, \tag{9.7}$$

minkä jälkeen $\mathrm{SS}(\alpha,\beta)$ voidaan esittää kaavalla

$$\begin{aligned}\mathrm{SS}(\alpha,\beta) &= q_{yy} + n(\alpha-\bar{y}+\beta\bar{x})^2\\ &\qquad q_{xx}\left(\beta^2 - 2\beta\frac{q_{xy}}{q_{xx}} + \frac{q_{xy}^2}{q_{xx}^2}\right) - \frac{q_{xy}^2}{q_{xx}}\\ &= q_{yy} - \frac{q_{xy}^2}{q_{xx}} + q_{xx}\left(\beta - \frac{q_{xy}}{q_{xx}}\right)^2 + n(\alpha-\bar{y}+\beta\bar{x})^2.\end{aligned}$$

Koska oletusten mukaan $q_{xx} > 0$ ja $n > 0$, niin kaikilla (α,β) pätee

$$\mathrm{SS}(\alpha,\beta) \geq q_{yy} - \frac{q_{xy}^2}{q_{xx}} = \mathrm{SS}(a,b) \tag{9.8}$$

jossa alaraja saavutetaan valitsemalla β:lle kaavan (9.2) mukainen arvo b ja sen jälkeen α:lle kaavan (9.3) mukainen arvo a. Nyt on PNS-suoran kertoimien kaavat saatu todistettua.

9.3 Lineaarinen malli

Pienimmän neliösumman periaate voidaan johtaa suurimman uskottavuuden periaatteesta, kun ensin asetetaan lineaarinen malli, jossa virheet ovat normaalijakautuneita. Lineaarisessa mallissa havaintoja $y_1,\ldots,y_n$ vastaa satunnaismuuttujat $Y_1,\ldots,Y_n$, joiden yhteisjakauma voidaan esittää kaavalla

$$Y_i = \alpha + \beta\, x_i + \epsilon_i, \qquad i = 1,\ldots,n. \tag{9.9}$$

Tässä ajatellaan, että luvut $x_1,\ldots,x_n$ ovat tunnettuja vakioita ja että virheet ϵ_i ovat riippumattomia satunnaismuuttujia, jotka kaikki noudattavat normaalijakaumaa

$$\epsilon_i \sim N(0,\sigma^2).$$

Virhesatunnaismuuttujien odotusarvo on nolla ja varianssi σ^2 ei riipu indeksistä i. Tuntemattomia parametreja ovat kertoimet α ja β sekä virhevarianssi $\sigma^2 > 0$.

Lineaarinen malli (9.9) voidaan yhtäpitävästi muotoilla siten, että satunnaismuuttujat Y_i ovat riippumattomia siten, että muuttujan Y_i jakauma on

$$Y_i \sim N(\alpha + \beta\, x_i, \sigma^2), \qquad i = 1, \ldots, n. \tag{9.10}$$

Tämän huomion jälkeen uskottavuusfunktio voidaan kirjoittaa.

Kun 2π:n potenssit jätetään pois, niin uskottavuusfunktioksi saadaan

$$\begin{aligned} L(\alpha, \beta, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{\sqrt{\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i - \alpha - \beta\, x_i)^2}{\sigma^2}\right) \\ &= (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2\right). \end{aligned}$$

Logaritminen uskottavuusfunktio on

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 \tag{9.11}$$

$$= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \operatorname{SS}(\alpha, \beta), \tag{9.12}$$

jossa $\operatorname{SS}(\alpha, \beta)$ on virheiden neliösumma (9.1). Oli varianssiparametrin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion

$$(\alpha, \beta) \mapsto \ell(\alpha, \beta, \sigma^2)$$

maksimipiste on sama kuin funktion $\operatorname{SS}(\alpha, \beta)$ minimipiste, joka puolestaan on (a, b), jossa a ja b ovat PNS-suoran vakiotermi ja kulmakerroin. Tämän ansiosta SU-estimaatin haku saadaan palautettua yhden muuttujan maksimointitehtäväksi, jossa pitää etsiä funktion

$$u(\sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \operatorname{SS}(a, b), \qquad \sigma^2 > 0$$

maksimipiste. Tämän funktion maksimi löytyy pisteestä

$$\hat{\sigma}^2 = \frac{1}{n} \operatorname{SS}(a, b).$$

Kertoimien SU-estimaateiksi saatiin PNS-suoran kertoimet, ja varianssiparametrin SU-estimaatiksi saatiin

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - a - b\, x_i)^2.$$

Lineaarisessa mallissa (9.9) kertoimien estimaatteina käytetään PNS-estimaatteeja a ja b, mutta virhevarianssin σ^2 estimaattina ei käytetä SU-estimaattia, vaan estimaattia

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - a - b\, x_i)^2, \tag{9.13}$$

jossa jakajana on otoskoon n sekä mallin kertoimien lukumäärän 2 erotus.

Vastaavien estimaattorien $a(\mathbf{Y})$, $b(\mathbf{Y})$ ja $s^2(\mathbf{Y})$ otantajakauma tunnetaan. On suhteellisen yksinkertaista nähdä, että kertoimien PNS-estimaattorit ovat harhattomia ja normaalijakautuneita. Lyhyehkö lasku näyttää, että

$$b(\mathbf{Y}) \sim N(\beta, \frac{\sigma^2}{q_{xx}}). \tag{9.14}$$

Tämän takia PNS-kulmakertoimen b keskivirhe lasketaan kaavalla

$$\mathrm{se}(b) = \frac{s}{\sqrt{q_{xx}}}, \tag{9.15}$$

jossa käytetään kaavan (9.7) merkintää neliösummalle

$$q_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2,$$

jota ei jaeta $(n-1)$:llä.

On paljon vaativampaa näyttää, että $s^2(\mathbf{Y})$ harhaton sekä riippumaton sekä estimaattorista $a(\mathbf{Y})$ että estimaattorista $b(\mathbf{Y})$. Osoittautuu, että sopivasti skaalattuna $s^2(\mathbf{Y})$ noudattaa khiin neliön jakaumaa vapausasteluvulla $n-2$, eli

$$\frac{n-2}{\sigma^2}\, s^2(\mathbf{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_i - a - b\, x_i)^2 \sim \chi^2_{n-2}. \tag{9.16}$$

Kun otantajakaumat (9.14) ja (9.16) yhdistetään sen tiedon kanssa, että nämä estimaattorit ovat riippumattomia, niin nähdään että

$$\frac{b(\mathbf{Y}) - \beta}{s(\mathbf{Y})/\sqrt{q_{xx}}} \sim t_{n-2}. \tag{9.17}$$

Näihin tuloksiin perustuen voidaan mallin kulmakertoimelle laskea luottamusvälejä sekä voidaan testata sitä koskevia hypoteeseja.

Esimerkiksi β:n kaksisuuntainen luottamusväli luottamustasolla $(1-\alpha)$ lasketaan kaavalla

$$b - t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}} \leq \beta \leq b + t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}}. \tag{9.18}$$

On mielekästä testata hypoteesiparia

$$H_0 : \beta = 0, \qquad H_0 \neq 0,$$

sillä jos todellinen kulmakerroin on nolla, niin tällöin selittäjää x ei lainkaan tarvita mallissa. Tässä testissä aineistosta lasketaan testisuure

$$t = \frac{b - 0}{s/\sqrt{q_{xx}}}, \tag{9.19}$$

jota verrataan t_{n-2}-jakauman yläkvantiiliin $t_{n-2}(\alpha/2)$. Nollahypoteesi hylätään, jos $|t| > t_{n-2}(\alpha/2)$.

Esimerkiksi kuminauha-aineistolle tämän testin tulos nähdään tutkimalla seuraavien komentojen tulostusta.

```
m <- lm(y ~ x)
summary(m)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -18.86 -13.49  -3.66  11.43  24.31
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -455.3       68.3   -6.66  5.6e-05 ***
## x               50.1        5.8    8.64  6.0e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.2 on 10 degrees of freedom
## Multiple R-squared: 0.882,Adjusted R-squared: 0.87
## F-statistic: 74.6 on 1 and 10 DF,  p-value: 5.98e-06
```

Nollahypoteesia $H_0 : \beta = 0$ vastaava t-arvo ja p-arvo ovat

$$t = 8.64, \qquad p = 5.98 \times 10^{-6}.$$

Tämä hypoteesi hylätään kaikilla tavanomaisilla luottamustasoilla. Toisin sanoen venytys on tarpeellinen selittäjä tässä lineaarisessa mallissa.

Kulmakertoimen 95 % luottamusvälin saa helpoimmin laskettua komennolla

```
confint(m)

##               2.5 %  97.5 %
## (Intercept) -607.47 -303.04
## x             37.17   63.01
```

Kulmakertoimen 95 % luottamusväli on $[37.17, 63.01]$.

9.4 Lineaarinen regressio, kun selittäjät ovat satunnaismuuttujia

Edellisen jakson tilastollinen malli on mielekäs lähinnä silloin, kun selittäjän arvot voidaan kokeessa itse asettaa, koska ne oletetaan tunnetuiksi vakioiksi. Lineaarista regressiota käytetään kuitenkin myös sellaisissa tilanteissa, joissa (x, y)-arvot mitataan yksilöistä, jotka on poimittu satunnaisesti populaatiosta. Tällöin myös x-arvot pitää mallintaa satunnaismuuttujina.

Jos tehdään mallissa sellainen oletus, että satunnaismuuttujat

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

muodostavat satunnaisotoksen jostakin kaksiulotteisesta normaalijakaumasta, niin tällöin edellisen jakson päättelyn kannalta keskeiset jakaumatulokset pitävät edelleen paikkansa. Erityisesti varianssiparametrin estimaattorin otantajakaumatulos (9.16) pitää edelleen paikkansa ja t-tunnusluvun (9.17) jakauma on edelleen t_{n-2}.

Tämä väite voidaan perustella tarkastelemalla sitä ehdollista jakaumaa, joka syntyy, kun selittäjävektorin $\mathbf{X} = (X_1, \ldots, X_n)$ arvoksi kiinnitetään havaitut arvot $\mathbf{x} = (x_1, \ldots, x_n)$. Satunnaismuuttujien $Y_1, \ldots, Y_n$ ehdollisessa yhteisjakaumassa ne ovat riippumattomia ja noudattavat normaalijakaumia

$$Y_i \mid (\mathbf{X} = \mathbf{x}) \sim N(\alpha + \beta\, x_i, \sigma^2),$$

jossa kertoimet α ja β sekä virhevarianssi σ^2 voidaan ilmaista kaksiulotteisen normaalijakauman parametrien avulla. Tämä seuraa kaksiulotteisen normaalijakauman erityisominaisuuksista.

Siis ehdollinen yhteisjakauma täyttää lineaarisen mallin oletukset, joten kaikki edellisessä jaksossa mainitut jakaumatulokset pätevät ehdolla $\mathbf{X} = \mathbf{x}$. Tästä seuraa edelleen, että sellaiset jakaumatulokset, jotka eivät riipu arvosta $\mathbf{x}$ (erityisesti (9.16) sekä (9.17)) pätevät myös ei-ehdollisesti.

9.5 Muita lineaarisia malleja

Jakson 9.3 malli (9.9) voidaan kirjoittaa matriisimerkinnöillä kaavalla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{9.20}$$

Tässä asetelmamatriisi (tai mallimatriisi) $\mathbf{X}$ on kiinteä ja tunnettu, ja kerroinvektori $\boldsymbol{\beta}$ kiinteä mutta tuntematon. Mallille (9.9) ne ovat muotoa

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

Virhevektorin $\boldsymbol{\epsilon}$ komponentit ϵ_i ovat riippumattomia ja noudattavat kukin normaalijakaumaa $N(0, \sigma^2)$. Tässä mallissa sekä kerroinvektori että virhevarianssi ovat tuntemattomia parametreja.

Usea tällä kurssilla käsitelty malli voidaan lineaarisena mallina kaavalla (9.20). Näitä ovat

- lineaarinen regressiomalli (9.9)
- satunnaisotos normaalijakaumasta $N(\mu, \sigma^2)$, jossa μ ja σ^2 ovat tuntemattomia;
- kaksi riippumatonta otosta kahdesta eri normaalijakaumasta $N(\mu_1, \sigma^2)$ ja $N(\mu_2, \sigma^2)$, (varianssianalyysimalli), kun jakaumilla on sama tuntematon varianssiparametri.

Lineaaristen mallien teoria antaa yhtenäisen teorian kaikille näille malleille sekä kaikille niiden muotoa (9.20) oleville yleistyksille. Aiheeseen voi perehtyä esim. lineaaristen mallien kurssilla.

Luku 10

Bayes-päättelyn alkeita

Bayesiläisessä päättelyssä havaintosatunnaisvektorin jakauma mallinnetaan täysin samalla tavalla kuin frekventistisessä lähestymistavassa silloin, kun parametrin arvo on kiinnitetty. Frekventistisessä lähestymistavassa parametri on kiinteä (ts. ei-satunnainen) mutta tuntematon. Bayesiläisessä lähestymistavassa parametria käsitellään satunnaisena suureena.

Tämä ehkä vähäiseltä tuntuva ero johtaa suuriin eroihin laskutekniikoissa ja tulosten tulkinnoissa. *Bayesiläisessä päättelyssä kaikki on toisin* kuin frekventistisessä.

10.1 Todennäköisyyslaskentaa

Tämän jakson kaavoissa oletetaan hiljaisesti, että osamäärien nimittäjät ovat erisuuria kuin nolla.

Olkoot A ja B tapahtumia. Jos tiedämme (täsmälleen sen), että B on sattunut, niin tapahtuman A todennäköisyys lasketaan *ehdollisen todennäköisyyden* kaavalla

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}. \tag{10.1}$$

Ehdollisen todennäköisyyden kaavasta saadaan todennäköisyyksien *kertolaskukaava*

$$P(A \cap B) = P(B)\, P(A \mid B) = P(A)\, P(B \mid A). \tag{10.2}$$

Tästä nähdään kaava

$$P(A \mid B) = \frac{P(A)\, P(B \mid A)}{P(B)}. \tag{10.3}$$

Jos tapahtumat $A_1, \ldots, A_M$ ovat jokin perusjoukon ositus ts.

- joukot A_i ovat erillisiä: jos $i \neq j$, niin $A_i \cap A_j = \emptyset$.
- niiden yhdiste on koko perusjoukko,

niin $P(B)$ voidaan laskea (todennäköisyyden additiivisuuden perusteella) seuraavasti,

$$P(B) = P(\cup_{i=1}^{M}(B \cap A_i)) = \sum_{i=1}^{M} P(B \cap A_i)$$
$$= \sum_{i=1}^{M} P(A_i)\, P(B \mid A_i)$$

Kun tätä ns. *kokonaistodennäköisyyden* kaavaa käytetään kaavassa (10.3) saadaan *Bayesin kaava*

$$P(A_k \mid B) = \frac{P(A_k)\, P(B \mid A_k)}{\sum_{i=1}^{M} P(A_i)\, P(B \mid A_i)} \tag{10.4}$$

johon bayesiläinen päättely perustuu diskreetin parametrin ja diskreetin havaintovektorin tapauksessa.

Kirjoitetaan edelliset kaavat vielä siinä tapauksessa, jossa käsitellään kahden diskreetin satunnaismuuttujan (tai satunnaisvektorin) $\tilde{\theta}$ ja $\mathbf{Y}$ yhteisjakaumaa, jonka määrää niiden yhteispistetodennäköisyysfunktio

$$f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}) = P(\tilde{\theta} = \theta, \mathbf{Y} = \mathbf{y}) = P(\{\tilde{\theta} = \theta\} \cap \{\mathbf{Y} = \mathbf{y}\}). \tag{10.5}$$

Satunnaismuuttujan (tai satunnaisvektorin) $\tilde{\theta}$ mahdolliset arvot ovat $\theta_1, \theta_2, \ldots$ ja satunnaismuuttujan (tai satunnaisvektorin) $\mathbf{Y}$ mahdolliset arvot ovat $\mathbf{y}_1, \mathbf{y}_2, \ldots$. Yhteispistetodennäköisyysfunktion arvoja ja muiden pistetodennäköisyysfunktioiden arvoja lasketaan jatkossa sellaisissa pisteissä $(\theta, \mathbf{y})$, joiden θ-koordinaatti on jokin $\tilde{\theta}$ mahdollisista arvoista ja $\mathbf{y}$-koordinaatti on jokin $\mathbf{Y}$:n mahdollisista arvoista.

Käytämme seuraavia merkintöjä.

- $p(\theta)$ on satunnaismuuttujan $\tilde{\theta}$ reunajakauman ptnf, eli
$$p(\theta) = P(\tilde{\theta} = \theta)$$
- $f(\mathbf{y} \mid \theta)$ on satunnaisvektorin $\mathbf{Y}$ pntf, kun $\tilde{\theta} = \theta$, eli
$$f(\mathbf{y} \mid \theta) = P(\mathbf{Y} = \mathbf{y} \mid \tilde{\theta} = \theta).$$
- $p(\theta \mid \mathbf{y})$ on satunnaismuuttujan $\tilde{\theta}$ ptnf, kun $\mathbf{Y} = \mathbf{y}$, eli
$$p(\theta \mid \mathbf{y}) = P(\tilde{\theta} = \theta \mid \mathbf{Y} = \mathbf{y})$$
- $f(\mathbf{y})$ on satunnaisvektorin $\mathbf{Y}$ reunajakauman ptnf, eli
$$f(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y})$$

Satunnaismuuttujien $\tilde{\theta}$ ja $\mathbf{Y}$ reunapistetodennäköisyysfunktiot saadaan niiden yhteispistetodennäköisyysfunktiosta kokonaistodennäköisyyden kaavalla, nimittäin

$$p(\theta) = P(\tilde{\theta} = \theta) = \sum_{\mathbf{y}} P(\tilde{\theta} = \theta, \mathbf{Y} = \mathbf{y}) = \sum_{\mathbf{y}} f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}) \tag{10.6}$$

$$f(\mathbf{y}) = P(\mathbf{Y} = \mathbf{y}) = \sum_{\theta} f_{\tilde{\theta},\mathbf{Y}}(\theta, \mathbf{y}). \tag{10.7}$$

Todennäköisyyksien kertolaskukaavan (10.2) mukaan yhteispistetodennäköisyysfunktio voidaan jakaa tekijöihin molemmilla seuraavista tavoista

$$f_{\tilde{\theta},\mathbf{Y}}(\theta,\mathbf{y}) = p(\theta)\, f(\mathbf{y} \mid \theta) = f(\mathbf{y})\, p(\theta \mid \mathbf{y}). \tag{10.8}$$

Tapahtuman $\tilde{\theta} = \theta$ todennäköisyys ehdolla $\mathbf{Y} = \mathbf{y}$ saadaan ratkaistua edellisestä identiteetistä, nimittäin

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})}. \tag{10.9}$$

Tämä on se Bayesin kaavan versio, johon bayesiläinen päättely perustuu silloin kuin aineiston otantajakauma on diskreetti ja parametriavaruus on diskreetti. Tässä parametria pidetään satunnaismuuttujana $\tilde{\theta}$ joka saa jonkin arvon parametriavaruudessa Θ.

Ennen (lat. *a priori*) havaintojen tekoa parametrilla on ns. *priorijakauma* (engl. *prior distribution*) eli jakauma, jonka ptnf on $p(\theta)$. Seuraavaksi tehdään havainto $\mathbf{Y} = \mathbf{y}$. Bayesiläinen päättely tarkoittaa sitä, että havaintojen jälkeen (lat. *a posteriori*) priorikäsitys päivitetään siirtymällä havaintoja vastaavaan ehdolliseen jakaumaan. Ennakkokäsitys päivitetään Bayesin kaavalla havaintojen jälkeiseksi käsitykseksi käyttämällä hyväksi priorijakaumaa ja havaintoja vastaavaa uskottavuusfunktiota $f(\mathbf{y} \mid \theta)$. Tulos on parametrin *posteriorijakauma* (engl. *posterior distribution*), eli parametrin ehdollinen jakauma $p(\theta \mid \mathbf{y})$ ehdolla $\mathbf{Y} = \mathbf{y}$.

Bayesin kaava kannattaa pitää mielessä muodossa

$$\text{posteriori} \propto \text{priori} \times \text{uskottavuus} \tag{10.10}$$

Tässä merkintä $\propto$ tarkoittaa verrannollisuutta, ts. edellä väitetään, että posteriori on vakio kertaa priorin ja uskottavuusfunktion tulo. Tässä posterioria $p(\theta \mid \mathbf{y})$ ajatellaan muuttujan θ funktiona kuten myös priorin ja uskottavuusfunktion tuloa. Ts. edellä väitetään, että

$$p(\theta \mid \mathbf{y}) = C\, p(\theta)\, f(\mathbf{y} \mid \theta), \qquad \text{kaikilla } \theta, \tag{10.11}$$

ja tämän on totta, sillä

$$C = \frac{1}{f(\mathbf{y})} = \frac{1}{\sum_{\theta=0}^{N} p(\theta)\, f(\mathbf{y} \mid \theta)}$$

Verrannollisuusvakio C toki riippuu havainnoista $\mathbf{y}$, mutta muuttujan θ funktiona ajateltuna se on vakio.

Bayesiläisessä analyysissä havaintoja vastaavalle satunnaisvektorille kiinnitetään sen havaittu arvo, ja sitten pohditaan eri parametrinarvojen todennäköisyyksiä. Frekventistisessä analyysissä parametri on kiinteä, ja todennäköisyyslaskentaa käytetään aineistoa vastaavan satunnaisvektorin jakauman ja siitä johdettujen tunnuslukujen jakaumien johtamiseen erilaisilla hypoteettisilla parametrinarvoilla. Useimmat frekventistisen tilastotieteen käsitteet perustuvat sellaisten aineistojen ominaisuuksien pohtimiseen, joita ei kokeessa havaittu.

Voimme ajatella, että bayesiläisessä analyysissä diskreetin parametrin tapauksessa lasketaan priorin ja uskottavuuden tulo kaikilla mahdollisilla parametrin arvoilla, jonka jälkeen tulos normalisoidaan pistetodennäköisyysfunktioksi jakamalla laskettujen arvojen summalla. Tämän algoritmin vaiheet ovat

1. Laske
$$s = \sum_{\theta} p(\theta)\, f(\mathbf{y} \mid \theta). \tag{10.12}$$

2. Laske
$$p(\theta \mid \mathbf{y}) = \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{s}, \qquad \theta \in \Theta. \tag{10.13}$$

Näemme sovelluksia seuraavissa jaksossa, joissa palaamme käsittelemään tilastollista päättelyä luvun 2 esimerkeissä.

10.2 Pallot kulhossa: diskreetti parametri

Kulhossa on N palloa ja niistä θ kpl on mustia ja $N - \theta$ valkoisia. Lukumäärä $0 \leq \theta \leq N$ on tuntematon. Palloja nostetaan satunnaisesti ja palauttaen n kertaa.

Oletetaan nyt, että luontoäiti on laittanut pallot kulhoon sillä tavalla, että hän ensin arpoi valkoisten pallojen lukumääräksi θ yhden luvuista $0, 1, \ldots, N$ siten, että kaikki vaihtoehdot ovat yhtä todennäköisiä. Sen jälkeen hän laittoi kulhoon θ valkoista ja $N - \theta$ mustaa palloja.

Pallojen poiminta tuottaa jonkin jonon $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ onnistumisia (valkoisen pallon nosto koodataan arvolla $y_i = 1$) tai epäonnistumisia (mustan pallon nosto koodataan arvolla $y_i = 0$). Vastaavien satunnaismuuttujien $Y_1, \ldots, Y_n$ yhteispistetodennäköisyysfunktio tunnetaan, jos θ tunnetaan, sillä se on
$$f(\mathbf{y} \mid \theta) = \left(\frac{\theta}{N}\right)^{t(\mathbf{y})} \left(1 - \frac{\theta}{N}\right)^{n - t(\mathbf{y})}, \tag{10.14}$$
jossa $t(\mathbf{y}) = \sum_i y_i$ on onnistumisten lukumäärä.

Kun on havaittu tietty jono $\mathbf{y}$, niin sitten voidaan kysyä, millä todennäköisyydellä kulhossa on θ kappaletta valkoisia palloja. Kun tähän kysymykseen vastataan kaikilla mahdollisilla parametrin θ arvoilla $0, 1, \ldots, N$, saadaan tuloksena posteriorijakauma.

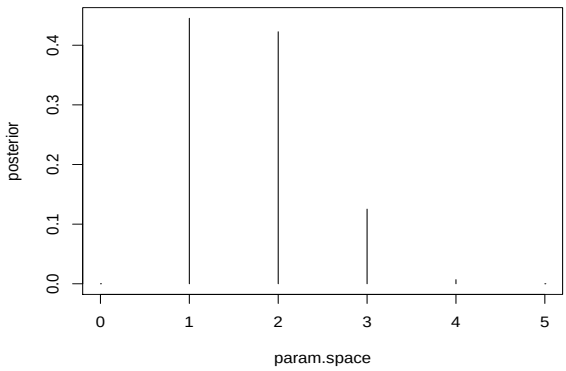
Ennen havaintoja satunnaismuuttujan $\tilde{\theta}$ jakauma eli priorijakauma on tilanteen kuvauksen perusteella diskreetti tasajakauma, jonka pistetodennäköisyysfunktio on
$$p(\theta) = \frac{1}{N+1}, \qquad \theta = 0, 1, \ldots, N.$$

Posteriorijakauma lasketaan seuraavassa esimerkissä soveltamalla kaavoja (10.12)–(10.13).

Esimerkki 10.1 Kulhossa on $N = 5$ palloa ja nostoja tehdään $n = 7$ ja tulokset ovat $\mathbf{y} = (1, 0, 0, 0, 1, 0, 0)$, jolloin onnistumisia on $x = 2$ kappaletta. Lasketaan R:llä priorin ja uskottavuusfunktion tulo, ja normalisoidaan se posteriorijakaumaksi.

```
N <- 5
n <- 7
x <- 2
param.space <- 0:N
print(prior <- rep(1, N + 1)/(N + 1))
```

```
## [1] 0.1667 0.1667 0.1667 0.1667 0.1667 0.1667

print(likelihood <- (param.space/N)^x * (1 - param.space/N)^(n -
    x))

## [1] 0.0000000 0.0131072 0.0124416 0.0036864 0.0002048 0.0000000

h <- prior * likelihood
posterior <- h/sum(h)
names(posterior) <- as.character(param.space)
print(posterior)

##        0        1        2        3        4        5
## 0.000000 0.445217 0.422609 0.125217 0.006957 0.000000

plot(param.space, posterior, "h")
```



Tarkkaavainen lukija huomasi, että tasaisesti priorista seurasi se, että posteriorijakauma oli yhtä kuin uskottavuusfunktio normalisoituna todennäköisyysjakaumaksi. Havaintojen jälkeen parametrin todennäköisin arvo on 1, mutta arvo 2 on lähes yhtä todennäköinen. Tässä tilanteessa on on paikallaan puhua parametrin todennäköisyydestä (ennen ja jälkeen havaintojen teon); tässä yhteydessä ei tarvitse puhua esim. uskottavuudesta. △

10.3 Priorin ja posteriorin tulkitseminen epävarmuuden kuvauksina

Edellisen jakson laskuissa ei ole mitään kiistanalaista, vaan ne seuraavat suoraan todennäköisyyslaskennan sääntöjen avulla tilanteen kuvauksesta. Kuitenkin bayesiläistä päättelyä pidettiin tilastotieteilijöiden piirissä yleisesti ainakin 1920–1960 luvuilla vanhentuneena ja suorastaan tuomittavana tapana lähestyä tilastollisen päättelyn ongelmia. Nykypäivänä tilastotieteen ammattilehdissä suurin osa artikkeleistä käyttää tavalla tai toisella bayesiläistä lähestymistapaa. Yritän tässä jaksossa valottaa, mikä asia bayesiläisessä päättelyssä aikanaan aiheutti tämän voimakkaan vastustuksen.

Ongelmaksi koettiin se, että bayesiläistä lähestymistapaa sovellettiin tilanteissa, joissa parametrin arvoa ei määrännyt mikään satunnaismekanismi. Tällöin priorijakauma ei kuvaa todellista arpomista, vaan se pitää ymmärtää kvantitatiivisena kuvauksena soveltajan epävarmuudesta parametrin todellisesta arvosta ennen havaintojen tekemistä. Posteriorijakauman tulkinnaksi tulee puolestaan se, että se on kvantitatiivinen kuvaus menetelmän soveltajan epävarmuudesta parametrin todellisesta arvosta, kun havainnot on otettu huomioon. Kiistan ydin oli siinä, että frekventistisen tilastotieteen perustajat eivät hyväksyneet sitä ajatusta, että todennäköisyysjakauma saataisiin tulkita subjektiivisena epävarmuuden kuvauksena. Heidän mielestään todennäköisyyden käsite oli objektiivinen, ja sitä saadaan käyttää ainoastaan sellaisissa tilanteessa, jossa jotakin koetta (jossakin mielessä) toistetaan useita kertoja.

Nykyaikana bayesiläisessä päättelyssä lähes aina käytetään todennäköisyyden subjektiivista tulkintaa epävarmuuden kvantitatiivisena kuvauksena silloin, kun puhutaan parametrin todennäköisyysjakaumasta. Tätä ajatusta ei nykyään enää koeta ongelmallisena.

Posteriorijakauma on tällöin tietenkin subjektiivinen, sillä se riippuu siitä, minkälaista priorijakaumaa kyseinen subjekti pitää hyvänä kuvauksena omasta epävarmuudestaan. Priorijakaumalla on voimakas vaikutus posteriorijakaumaan, mikäli otoskoko on pieni. Jos otoskoko on suuri, niin tällöin erilaisilla järkevillä prioreilla saavutetaan lähes samanlainen posteriorijakauma. Otoskoon kasvaessa järkevien soveltajien subjektiiviset posteriorijakaumat alkavat siis muistuttaa yhä enenenevässä määrin toisiaan.

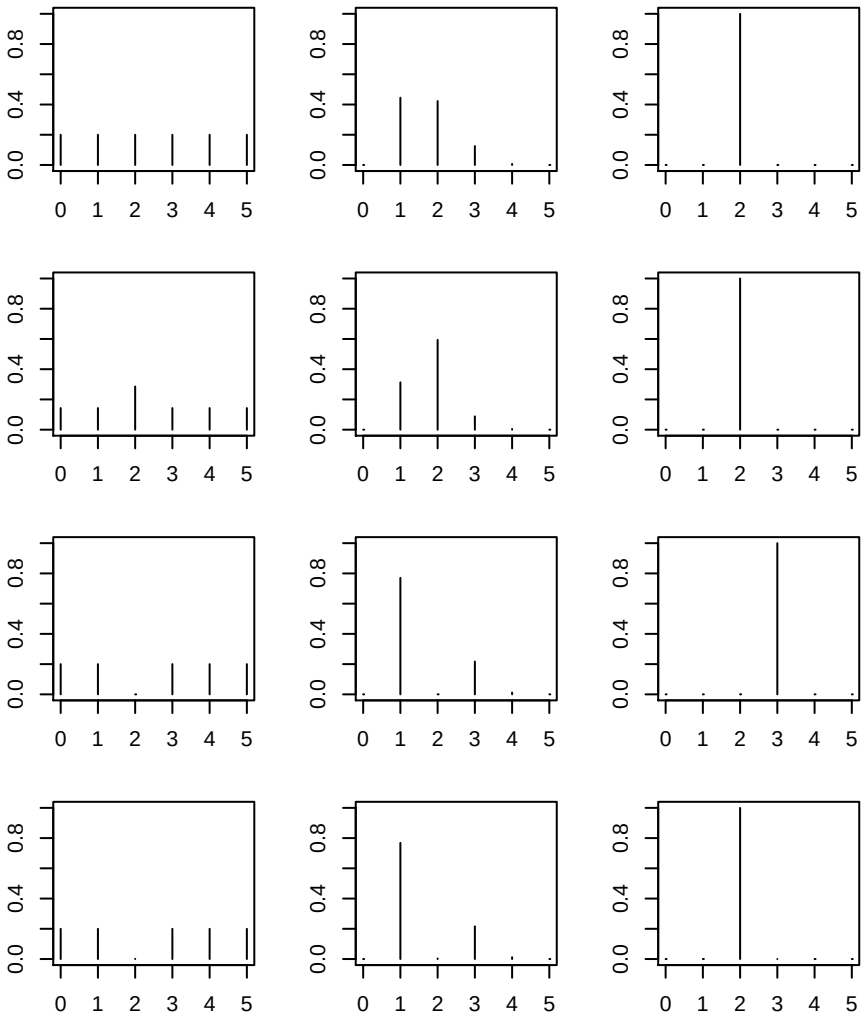
Lasketaan esimerkin vuoksi neljän eri henkilön A, B, C ja D posteriorijakaumat pallot kulhossa -tilanteessa: ensin seitsemän noston (2 onnistumista) ja sitten 300 noston (133 onnistumista) jälkeen.

- A:n priorijakauma on tasajakauma arvoilla $0, 1, \ldots, 5$.
- B suosii sitä vaihtoehto, että kulhossa on kaksi valkoista palloa. B:n priorijakauman mukaan valkoisten pallojen lukumäärä 2 on kaksi kertaa todennäköisempi kuin muut, jotka puolestaan ovat keskenään yhtä todennäköisiä.
- C on dogmaattisesti sitä mieltä, että kulhossa ei voi olla kahta valkoista palloa. C:n ennakkokäsityksen mukaan arvo 2 on mahdoton, ja kaikki muut mahdollisuudet ovat yhtä todennäköisiä.
- Myös D vastustaa kiivaasti sitä ajatusta, että kulhossa voisi olla kaksi valkoista palloa. Hän kuitenkin muotoilee käsityksensä vähemmän dogmaattisesti kuin C. D:n ennakkokäsityksen mukaan arvolla 2 on todennäköisyys 1/1000, ja kaikki muut arvot ovat keskenään yhtä todennäköisiä.

Eri henkilöiden priorijakaumat ja posteriorijakaumat on esitetty kuvassa 10.3

Otoskoolla $n = 7$ nähdään, että posteriorijakaumat riippuvat vahvasti kunkin henkilön priorikäsityksistä, mutta henkilöt C ja D ovat käytännössä yhtä mieltä eri valkoisten pallojen lukumäärän todennäköisyyksistä. Sen sijaan otoskoolla $n = 300$ kaikkien muiden henkilöiden paitsi C:n posteriorikäsitykset ovat käytännössä yhtenevät. C sulkee omalla priorin valinnallaan kokonaan pois sen mahdollisuuden, että valkoisia palloja voisi kulhossa olla kaksi. Tämän takia C:n posterioritodennäköisyys kahdelle valkoiselle pallolle on aina nolla riippumatta lainkaan siitä, mitä havaintoja saadaan. Tällä kertaa aineisto todistaa otoskoolla $n = 300$ vahvasti sen puolesta, että valkoisia palloja olisi kulhossa

Kuva 10.1 Vaakariveillä on henkilöiden A, B, C ja D priorijakaumat, posteriorijakaumat $n = 7$ toiston ja 2 onnistumisen sekä $n = 300$ toiston ja 133 onnistumisen jälkeen.



todellisuudessa kaksi kappaletta, ja henkilöt A, B ja D ovat käytännössä täysin vakuuttuneita siitä, että valkoisia palloja on kulhossa kaksi.

Sellaista dogmaattista priorin valintaa (kuten C:n priori), joka sulkee kokonaan pois tarkastelusta jonkin järkevän osan parametriavaruudesta, voidaan pitää bayesiläisen päättelyn pelisääntöjen vastaisena.

10.4 Nasta purkissa: jatkuva parametri

Nyt binomikokeen onnistumistodennäköisyydellä on jokin arvo avoimella välillä $(0, 1)$. Jos tämä arvo θ tunnetaan, niin havaintosatunnaisvektorin pistetodennäköisyysfunktio on

$$f(\mathbf{y} \mid \theta) = \theta^{t(\mathbf{y})} \, (1 - \theta)^{n - t(\mathbf{y})},$$

kun $\mathbf{y} = (y_1, \ldots, y_n)$ on jono nollia tai ykkösiä, ja $t(\mathbf{y}) = \sum_i y_i$ on onnistumisten lukumäärä n toistossa. Nyt parametriavaruus on jatkuva, ja tämä tuo mukanaan uusia teknisiä ongelmia.

Parametri $\tilde{\theta}$ on nyt jatkuvasti jakautunut satunnaismuuttuja, jonka arvot kuuluvat parametriavaruuteen $\Theta = (0, 1)$. Ennen havaintojen tekoa soveltajan pitää onnistua kuvaamaan oma epävarmuutensa parametrin arvoista priorijakaumalla, jonka tiheysfunktiota merkitsemme

$$p(\theta), \qquad \theta \in \Theta.$$

Havaintojen jälkeen siirrytään tarkastelemaan parametrin ehdollista tiheysfunktiota eli sen posteriorijakaumaa.

Suuri osa jakson 10.1 kaavoista pitää sellaisenaan paikkansa myös tässä uudessa tilanteessa, mutta summaus pitää korvata integroinnilla. Yhteisjakauma voidaan esittää funktion

$$f_{\tilde{\theta}, \mathbf{Y}}(\theta, \mathbf{y}) = p(\theta) \, f(\mathbf{y} \mid \theta)$$

avulla. Se on muuttujan θ suhteen tiheysfunktio ja muuttujan $\mathbf{y}$ suhteen pistetodennäköisyysfunktio, ts. todennäköisyyksiä lasketaan integroimalla muuttujan θ suhteen ja summaamalla muuttujan $\mathbf{y}$ suhteen.

Bayesin kaava saa tutun muodon

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta) \, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})},$$

mutta nyt normalisointivakio $f(\mathbf{y})$ eli havaintosatunnaisvektorin reunajakauman pistetodennäköisyysfunktion arvo pitää laskea integroimalla,

$$f(\mathbf{y}) = \int_0^1 p(\theta) \, f(\mathbf{y} \mid \theta) \, \mathrm{d}\theta.$$

Tätä integraalia ei yleisesti ottaen osata laskea analyyttisesti. Sitä voi toki yrittää approksimoida jollakin numeerisella menetelmällä.

Posteriorijakauman tiheysfunktion kuva voidaan piirtää suoraan käyttämällä verrannollisuustulosta

$$p(\theta \mid \mathbf{y}) \propto p(\theta) \, f(\mathbf{y} \mid \theta),$$

mikäli tyydytään siihen, että y-akselin skaala jää selvittämättä.

10.5 Liittojakauma eli konjugaattijakauma

Joillekin uskottavuusfunktioille on mahdollista löytää sellainen parametrinen perhe jakaumia, joilla on seuraava miellyttävä ominaisuus. Mikäli priorijakauma valitaan kyseisestä perheestä, niin myös posteriorijakauma kuuluu samaan perheeseen. Näemme kohta, että binomikokeen uskottavuusfunktiolle beeta-jakaumat muodostavat tällaisen perheen. Tämä voidaan ilmaista sanomalla, että beeta-jakauma on binomiuskottavuuden *liittopriori* (engl. *conjugate prior*) tai että havaintosatunnaisvektorin otantajakauma ja priorijakauma ovat toistensa liittojakaumia.

Mikäli tutkijan ennakkokäsitys parametrinarvosta voidaan esittää jollakin liittoperheen jakaumalla, niin tällöin posteriorijakauma saadaan laskettua johtamalla päivityskaavat, joilla priorijakauman parametrit päivitetään posteriorijakauman parametreiksi. Kutsutaan näitä liittojakaumaperheen parametreja selvyyden vuoksi hyperparametreiksi, jotta ne saadaan erotettua tilastollisen mallin parametrista θ.

Beeta-jakauma on jatkuva jakauma, jonka tiheysfunktio on muotoa

$$g(x) \propto x^{\alpha-1}\,(1-x)^{\beta-1}, \quad \text{kun } 0 < x < 1, \tag{10.15}$$

missä $\alpha > 0$ ja $\beta > 0$ ovat jakaumaperheen parametrit. Tämä lauseke määrittelee yksikäsiteisesti tietyn todennäköisyysjakauman, jonka tiheysfunktio saadaan selville jakamalla lauseke sen välin $(0,1)$ yli lasketulla integraalilla, sillä tiheysfunktion integraalin koko satunnaismuuttujan arvoalueen yli täytyy olla yksi. Kaavassa (10.15) merkitsemättä jätetty normalisointivakio saadaan ns. Eulerin beetafunktion

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}\,(1-x)^{\beta-1}\,\mathrm{d}x$$

avulla (jossa B on iso beeta-kirjain). Beeta-jakauman tiheysfunktion täydellinen kaava on

$$g(x) = \begin{cases} \dfrac{1}{B(\alpha,\beta)}\,x^{\alpha-1}\,(1-x)^{\beta-1}, & \text{kun } 0 < x < 1, \\ 0 & \text{muuten.} \end{cases} \tag{10.16}$$

Erityisesti valinnoilla $\alpha = 1$ ja $\beta = 1$ saadaan välin $(0,1)$ tasajakauma.

Beeta-jakauman $\mathrm{Beta}(\alpha,\beta)$ ominaisuudet tunnetaan. Sitä noudattavan satunnaismuuttujan X arvot ovat välillä $(0,1)$, ja esim. sen odotusarvo ja varianssi saadaan laskettua tunnetuilla kaavoilla

$$EX = \frac{\alpha}{\alpha+\beta}, \qquad \operatorname{var} X = \frac{\alpha\,\beta}{(\alpha+\beta)^2\,(\alpha+\beta+1)}. \tag{10.17}$$

Jakauman moodi (eli tiheysfunktion maksimipiste) on

$$\frac{\alpha-1}{\alpha+\beta-2},$$

mikäli $\alpha > 1$ ja $\beta > 1$ (muilla parametrinarvoilla jakaumalla ei ole hyvin määriteltyä moodia).

Tarkistetaan nyt, että beeta-jakauma on binomiuskottavuuden liittopriori. Priorin hyperparametrit ovat $\alpha, \beta > 0$ ja onnistumisia havaitaan k kappaletta. Tarkistuksen voi tehdä kahdella erilaisella tavalla.

Tapa 1: normalisointivakion laskeminen integroimalla

Integroidaan tulo priori kertaa uskottavuus. Huomaa, että priorijakauman tiheysfunktiossa argumenttina pitää käyttää integrointimuuttujaa, joka on θ.

$$\begin{aligned} f(\mathbf{y}) &= \int_0^1 p(\theta)\, f(\mathbf{y} \mid \theta)\, \mathrm{d}\theta \\ &= \int_0^1 \frac{1}{B(\alpha,\beta)}\, \theta^{\alpha-1}\, (1-\theta)^{\beta-1}\, \theta^k\, \theta^{n-k}\, \mathrm{d}\theta \\ &= \frac{1}{B(\alpha,\beta)} \int_0^1 \theta^{\alpha+k-1}\, (1-\theta)^{\beta+n-k-1}\, \mathrm{d}\theta = \frac{B(\alpha+k, \beta+n-k)}{B(\alpha,\beta)}. \end{aligned}$$

Integraali osattiin laskea, koska huomattiin käyttää Eulerin beeta-funktion määritelmää.

Bayesin kaavan mukaan

$$\begin{aligned} p(\theta \mid \mathbf{y}) &= \frac{p(\theta)\, f(\mathbf{y} \mid \theta)}{f(\mathbf{y})} \\ &= \frac{\frac{1}{B(\alpha,\beta)}\, \theta^{\alpha-1}\, (1-\theta)^{\beta-1}\, \theta^k\, \theta^{n-k}}{B(\alpha+k, \beta+n-k)/B(\alpha,\beta)} \\ &= \frac{1}{B(\alpha+k, \beta+n-k)}\, \theta^{\alpha+k-1}\, (1-\theta)^{\beta+n-k-1}, \qquad 0 < \theta < 1 \end{aligned}$$

Tästä nähdään, että posteriorijakauma on beeta-jakauma

$$\mathrm{Beta}(\alpha+k, \beta+n-k).$$

Tapa 2: verrannollisuustarkastelu

Edellisessä tavassa tulee matkan varrella helposti virheitä. Tulos on paljon helpompi johtaa seuraavalla tekniikalla. Muuttujan θ funktiona posteriorijakauman tiheys on verrannollinen priorijakauman tiheyden ja uskottavuusfunktion tuloon, joten välillä $0 < \theta < 1$ pätee verrannollisuus

$$\begin{aligned} p(\theta \mid \mathbf{y}) &\propto p(\theta)\, f(\mathbf{y} \mid \theta) \\ &= \theta^{\alpha-1}\, (1-\theta)^{\beta-1}\, \theta^k\, \theta^{n-k} \\ &= \theta^{\alpha+k-1}\, (1-\theta)^{\beta+n-k-1}. \end{aligned}$$

Vertaamalla tulosta kaavaan (10.15) nähdään, että posteriorijakauma on beeta-jakauma

$$\mathrm{Beta}(\alpha+k, \beta+n-k),$$

sillä ainoa todennäköisyysjakauma, jonka kantaja on $(0,1)$ ja jonka tiheysfunktio on verrannollinen johdettuun funktioon on tämä beeta-jakauma.

10.6 Posteriorijakauman yhteenvetoja

Kun posteriorijakauma on selvitetty, niin lopuksi voidaan yrittää laskea siitä tiettyjä yhteenvetoja. Jos posteriorijakauma on yleisesti tunnettu yksinkertainen jakauma, niin tämä vaihe voidaan sivuuttaa.

Posteriorijakauman keskikohtaa voidaan luonnehtia parametrin posteriori-odotusarvolla. Binomikokeen tapauksessa tämä on luku

$$E[\tilde{\theta} \mid \mathbf{y}] = \int_0^1 \theta\, p(\theta \mid \mathbf{y})\, \mathrm{d}\theta.$$

Jos priori on Beta(α, β), niin posterioriodotusarvo voidaan lukea suoraan beetajakauman odotusarvon kaavasta (10.17), kun siihen sijoitetaan posteriorijakauman hyperparametrit, jotka ovat binomikokeessa

$$\alpha_1 = \alpha + k, \qquad \beta_1 = \beta + n - k.$$

Vastaavasti voidaan laskea posteriorimoodi eli posteriorijakauman moodi. Posterioriodotusarvoa ja posteriorimoodia voidaan ajatella bayesiläisinä piste-estimaatteina. Posteriorijakauman keskittyneisyyttä voidaan kuvailla esim. laskemalla parametrin posteriorivarianssi eli posteriorijakauman varianssi.

Mikäli posteriorijakauman kvantiilifunktion $q(u)$ arvoja osataan laskea tavalla tai toisella, niin sen jälkeen parametriavaruudesta löytyy helposti väli $[L, U]$, jolle

$$P(L \leq \tilde{\theta} \leq U \mid \mathbf{Y} = \mathbf{y}) = 1 - \alpha$$

millä tahansa annetulla $0 < \alpha < 1$. Eräs tällainen väli saadaan jakamalla virhetodennäköisyys α tasan alemman ja ylemmän jakauman hännän kesken valitsemalla

$$L = q(\alpha/2), \qquad U = q(\alpha/2).$$

Havainnon jälkeen parametri kuuluu tälle välille todennäköisyydellä $1 - \alpha$. Tälläistä väliä voidaan kutsua tason $1 - \alpha$ todennäköisyysväliksi tai bayesiläiseksi luottamusväliksi.

Jos priorijakauma on beetajakauma, niin binomikokeessa posteriorijakauma on eräs toinen beetajakauma. Tilastollisista ohjelmistoista löytyy välineet beetajakauman kvantiilifunktion laskemiseksi: esim. R-ohjelmistossa kvantiilifunktion saa laskettua funktiolla `qbeta`. Tällaisessa tilanteessa todennäköisyysvälin päätepisteet saa laskettua käden käänteessä tietokoneella.

Mikäli ollaan kiinnostuneita muotoa

$$\tilde{\theta} \in A$$

olevasta hypoteesista, niin sen todennäköisyys havainnon jälkeen voidaan laskea integroimalla posteriorijakauman tiheysfunktiota

$$P(\tilde{\theta} \in A \mid \mathbf{Y} = \mathbf{y}) = \int_A p(\theta \mid \mathbf{y})\, \mathrm{d}\theta.$$

Jos posteriorijakauman kertymäfunktio osataan laskea ja jos A on väli, niin välin A posterioritodennäköisyys saadaan laskemalla kerymäfunktion arvot välin päätepisteissä sekä laskemalla suuremman arvon ja pienemmän arvon erotus. Vastahypoteesin $\tilde{\theta} \notin A$ todennäköisyys havainnon jälkeen saadaan sitten vähentämällä luvusta yksi tarkasteltavan hypoteesin todennäköisyys.

Jos priori on beetajakauma, niin posteriori on myös beetajakauma, ja tilastoohjelmista löytyy valmiudet laskea beetajakauman kertymäfunktion arvoja. Tässä tilanteessa välien posterioritodennäköisyydet saadaan laskettua käden käänteessä.

Koska parametria käsitellään satunnaismuuttujana, niin päästään puhumaan parametrin todennäköisyyksistä tai parametrin todennäköisyysjakaumasta tai kyseisen jakauman ominaisuuksista havainnon tekemisen jälkeen. Johdot ovat periaatteessa täysin suoraviivaisia ja niissä tarvitaan ainoastaan todennäköisyyslaskentaa, eikä mitään sen ulkopuolisia periaatteita. Monimutkaisemmille tilastollisille malleille vaadittavat laskut ovat kuitenkin liian hankalia analyyttisesti suoritettaviksi.

10.7 Bayesiläisen päättelyn laskentamenetelmiä

Posteriorijakauma saadaan selvitettyä kaavojen avulla seuraavissa tilanteissa.

1. Jos parametri on diskreetti ja sillä on äärellinen määrä mahdollisia arvoja.
2. Jos käytetään priorijakaumaa, joka kuuluu uskottavuusfunktion liittoperheeseen.

Muissa tapauksissa ei saada johdettua käyttökelpoisia analyyttisiä kaavoja.

Nykyään bayesiläinen analyysi tehdään tietokoneen avulla. Laskentamenetelmät perustuvat tyypillisesti satunnaisuuden simulointiin tietokoneen avulla eli ns. Monte Carlo -menetelmiin.

Perustana on se ajatus, että koska posteriorijakauma on todennäköisyysjakauma, niin sitä voidaan yrittää simuloida tietokoneella. On kehitetty menetelmiä, joissa lasketaan suuri määrä arvoja $t_1, \ldots, t_N$, joita voidaan pitää sellaisten satunnaismuuttujien $\tilde{\theta}_1, \ldots, \tilde{\theta}_N$ havaittuina arvoina, joista kukin on jakautunut posteriorijakauman mukaisesti.

Tämän jälkeen otosta $t_1, \ldots, t_N$ käsitellään data-analyysin keinoin.

- Posterioriodotusarvoa voidaan arvioida vastaavilla otoskeskiarvoilla,

$$E[\tilde{\theta} \mid \mathbf{y}] \approx \frac{1}{N} \sum_{i=1} t_i$$

- Parametrille voidaan muodostaa bayesiläisiä luottamusvälejä otoksesta laskettujen kvantiilipisteiden avulla.
- Posteriorijakauman pistetodennäköisyysfunktiota voidaan arvioida vastaavilla suhteellisilla frekvensseillä (diskreetin parametrin tapauksessa).
- Posteriorijakauman tiheysfunktiota voidaan arvioida esim. histogrammilla (jatkuvan parametrin tapauksessa).
- Mielivaltaisen parametrin funktion $g(\tilde{\theta})$ posteriorijakauman ominaisuuksia (esim. posterioriodotusarvoa tai posteriorijakaumaa) voidaan selvittää otoksesta $g(t_1), \ldots, g(t_N)$ aivan samoilla menetelmillä.

Tällaiset menetelmät ovat suhteellisen uusia: intensiivinen kehitysvaihe alkoi 1980-luvun lopulla. Niiden ansiosta nykyään on mahdollista analysoida lähes mielivaltaisen monimutkaisia bayesiläisiä tilastollisia malleja.