Luku 8

Johdanto tilastolliseen päättelyyn

8.1 Tilastollinen ongelma

Tilastollisessa päättely kasittelee johtopäätösten tekoa havainnoista, joihin sisältyy epävarmuutta ja satunnasivaihtelua. Tässä prosessissa käytetään apuna todennäköisyyteen perustuvaa tilastollista mallintamista. Ei ole itsestään selvää, että tuollaisesta aiheesta voidaan esittää mitään täsmällistä tai tieteellistä. Tilastotieteessäkin on omaksuttu erilaisia lähestymistapoja epävarmuuden käsittelyyn, jotka voidaan jaotella kahteen pääkoulukuntaan: Bayesilaiseen ja frekventistiseen. Tässä luvussa esitellään uskottavuuden (uskottavuusfunktio) käsite, joka on kummankin edellä mainitun lähestymistavan keskeinen peruskäsite.

Esimerkki 8.1 (Päättely vs. päätäntä) Oletetaan, että taudin T toteamiseksi on olemassa hyvä testi. Testin tilastollinen käyttäytyminen on kuvattu Taulukossa 8.1 Potilaalle tehtiin testi, jonka tulos oli $+$. Lääkäri saattaa

Taulukko 8.1. Testi taudin T toteamiseksi. Tuloksen $+/-$ todennäköisyydet.

	Testin tulos	
	$+$	$-$
Tauti on	0.95	0.05
Ei tautia	0.02	0.98

tehdä yhden seuraavista johtopäätöksistä (tai niiden negaation):

1. Potilaalla ei todennäköisesti ole tautia T.
2. Potilasta pitäisi hoitaa ikäänkuin hänellä olisi tauti T.
3. Testin tulos vahvistaa hypoteesia, että potilaalla on tauti T.

Johtopäätöstä 1 voidaan tarkastella laskemalla todennäkäisyys $P(T|+)$, jos testin tulos + (Vastaavasti $P(T|-)$). Todennäköisyys riippuu siitä, miten harvinaisesta taudista on kyse. Jos $P(T)$ on hyvin pieni, niin myös $P(T|+)$ on pieni. Johtopäätöksen 1 pätevyys riippuu siis ratkaisevasti taudin T prioritodennäköisyydestä $P(T)$. Lääkärillä on vaihtoehtoiset hypoteesit:

A: Potilaalla on tauti T.

B: Potilaalla ei ole tautia T.

Edellä esitetyt lääkärin johtopäätökset $1-3$ voidaan muotoilla myös seuraavasti:

1. Uskon hypoteesin B olevan tosi.
2. Minun pitäisi toimia ikäänkuin A olisi tosi.
3. Testitulos + on todiste hypoteesin A puolesta B:tä vastaan.

Lääkärin tekemät päätelmät ovat vastauksia kolmeen yleiseen kysymykseen:

1. Mitä minun pitäisi uskoa nyt, kun minulla on tämä havainto?
2. Mitä minun pitäisi tehdä nyt, kun minulla on tämä havainto?
3. Miten tekemäni havainto vaikuttaa hypoteesien A ja B uskottavuuteen ts. vahvistaako tai heikentääkö havainto A:n uskottavuutta ja B:n suhteen?

Tämän kurssin kannalta tyyppiä 3 oleva kysymys on keskeisin. Esimerkiksi raportoitaessa tieteellisen tutkimuksen tuloksia, pohditaan tavallisesti juuri tyypin 3 kysymyksiä. □

Esimerkki 8.2 (Aspiriinitutkimus) Taulukon 8.2 tulokset ovat peräisin tutkimuksesta, jossa selvitettiin, ehkäiseekö aspiriini aivohalvauksia ja sydäninfarkteja (infarctus myocardii acutus) (Steering Committee of the Physicians' Health Study Research Group 1989). Tutkimuksessa satunnaistettiin 22071 tervettä henkilöä aspiriiniryhmään ja lumeryhmään. Aspiriiniryhmään kuuluvat saivat päivittäin pienen annoksen aspiriinia ja lumeryhmään kuuluvat vastaavasti lumetabletin. Henkilöiden terveydentilaa seurattiin keskimäärin 5 vuotta. Asetelma on satunnaistettu kliininen koe, jossa tutkittiin aspiriinin käytön vaikutusta sydäninfarktikuolleisuuteen. Tutkimukseen osallistuneet eivät tienneet, kumpaan ryhmään he kuuluivat.

Pääkysymys on siis tämä: Onko aspiriinista hyötyä sydäninfarktin ehkäisyssä? Aspiriiniryhmässä on vähemmän sydäninfarkteja kuin lumeryhmässä, 139 vastaan 239. Mitä tämä todistaa, mitä luvuista voidaan päätellä? Onko todistusaineisto kyllin vahva, jotta voimme vastata kysymykseen? Aivohalvauksia oli kuitenkin enemmän aspiriiniryhmässä: 119 vastaan 98. Onko tuo

Taulukko 8.2. Aspiriinin käytön mahdollinen aivohalvauksia ja sydäninfarkteja ehkäisevä vaikutus.

Ryhmä	Sydänkohtaus	Aivohalvaus	Yhteensä
Aspiriini	139	119	11037
Lume	239	98	11034
Yhteensä	378	217	22071

ero merkitsevä? Tällaisiin kysymyksiin vastaaminen edellyttää *tilastollista* eli *stokastista mallia*, joka kuvaa havintojen stokastista käyttäytymistä.

Suhteellinen riski

$$\frac{139/11037}{239/11034} = 0.58$$

on eräs vakiintunut keino vertailla kahta suhteellista osuutta. Aspiriinin hyöty suhteellisena riskinä ilmaistuna on 0.58. Jos suhteellinen riski olisi 1, se tarkoittaisi, että aspiriinilla ei ole vaikutusta. Ykköstä selvästi pienempi arvo osoittaa, että aspiriinista on hyötyä. Onko siis 0.58 tarpeeksi paljon ykköstä pienempi? Tässä esimerkissä voidaan olettaa binomimalli, jossa sydäninfarktiin sairastuneiden lukumäärä aspiriiniryhmässä noudattaa binomijakaumaa $\text{Bin}(\theta_A, n_A)$ ja lumeryhmässä binomijakaumaa $\text{Bin}(\theta_L, n_L)$, missä $n_A = 11037, n_L = 11034$ ja todennäköisyydet θ_A sekä θ_L ovat tuntemattomia parametreja. Tuntematon suhteellinen riski on $\theta_A/\theta_L \equiv \theta_{SR}$.

Olemme siis laskeneet suhteellisen riskin estimaatin $\hat{\theta}_{SR} = 0.58$. Siihen ei ole liitetty mitään arvon luotettavuudesta kertovaa mittaa, joten se ei yksin pysty vastaamaan alkuperäiseen kysymykseen. Antaako koe niin paljon informaatiota, että voimme estimaatin $\hat{\theta}_{SR}$ perusteella väittää θ_{SR}:n olevan paljon pienempi kuin 1. Ajatellaan, että olisi tehty 10 kertaa laajempi koe ja olisi havaittu 1390 vastaan 2390 sydänkohtausta. Silloin jälleen olisi $\hat{\theta}_{SR} = 0.58$, mutta intuitiivisesti tämän kokeen tulos tuntuu vakuuttavammalta. Pelkkä estimaatin arvo ei riitä, vaan siihen pitää liittää jokin estimaatin täsmällisyyttä kuvaava mitta, jonka avulla voimme arvioida estimaatin luotettavuutta. Tämä on tilastollisen päättelyn perusongelma: Miten havaintojen avulla voidaan tehdä tuntemattomia parametreja koskevia päteviä päätelmiä?

Jos esimerkiksi todennäköisyys sairastua johonkin tautiin muuttuu jossain väestössä vaikkapa todennäköisyydestä 0.0001 todennäköisyyteen 0.0010, on muutos suhteellisesti ottaen erittäin suuri. Jonkin tavallisen tapahtuman kohdalla yhtä suuri todennäköisyyden muutos, esimerkiksi todennäköisyydestä 0.2001 todennäköisyyteen 0.2010 ei ole merkittävä. Samoilla todennäköisyyksien muutoksilla lähellä ääripäitä 0 ja 1 on usein suurempi merkitys kuin vaihtelualueen keskivaiheilla. Todennäköisyyksien suhteen avulla voidaan tarkastella suhteellista muutosta.

Taulukosta 8.2 laskettu suhteellinen riski $\theta_{SR} = \theta_A/\theta_L$, voi teoreettisesti

saada minkä tahansa ei-negatiivisen arvon. Todennäköisyyksien arvoilla $\theta_A = 0.0010$ ja $\theta_L = 0.0001$ suhteellinen riski on $\theta_{SR} = 0.0010/0.0001 = 10$ ja arvoilla $\theta_A = 0.2010$ ja $\theta_L = 0.2001$ suhteellinen riski on $\theta_{SR} = 0.2010/0.2001 = 1.004$.

Taulukko 8.3. Sairastumistodennäköisyys aspiriiniyhmässä on θ_A ja lumeryhmässä θ_L. Indikaattori $Y = 1$, kun henkilöllä on infarkti ja muutoin $Y = 0$, joten $P(Y = 1|\text{Aspiriiniryhmä}) = \theta_A$ ja $P(Y = 1|\text{Lumeryhmä}) = \theta_L$.

	Y		
	1	0	Yhteensä
Aspiriini	θ_A	$1 - \theta_A$	1.0
Lume	θ_L	$1 - \theta_L$	1.0

Sairastumistodennäköisyyden estimaatti aspiriiniyhmässä on $\hat{\theta}_A = 139/11037 = 0.012594$ ja lumeryhmässä $\hat{\theta}_L = 239/11034 = 0.021660$, joten $\hat{\theta}_L = 1.719892\,\hat{\theta}_A$ ja $\hat{\theta}_L - \hat{\theta}_A = 0.009066$. □

8.2 Tilastolliset mallit

Tilastotiede kehittelee menetelmiä, joiden avulla voidaan oppia havainnoista. Tilastollisessa mallintamisessa havaintoja tarkastellaan satunnaiskokeen tuloksena, satunnaisilmiönä. Ajatelkaamme, että havainnot ovat "mustan laatikon"tuottamia. Syötemuuttujien arvovektori $\boldsymbol{x} = (x_1, \ldots, x_p)$ (selittävät muuttujat, riippumattomat muuttuja) työnnetään laatikkoon ja saadaan tuloksen vastemuuttujien (riippuvat muuttujat, selitettävät muuttujat) arvot $\boldsymbol{y} = (y_1, \ldots, y_m)$:

$$\boldsymbol{x} \longrightarrow \boxed{\text{luonto}} \longrightarrow \boldsymbol{y}$$

Laatikon sisällä luonto liittää jonkin funktion avulla selittävät muuttujat ja vasteet yhteen. Havaintojen analysoinnin tavoitteet voidaan karkeati jakaa kahteen ryhmään:

Ennustaminen. Mallilla halutaan ennustaa, mitä vastemuuttujan arvoja saadaan tulevaisuuden syötteillä.

Tieto riippuvuuksista. Halutaan saada selvyys siitä, millä tavoin luonto on liittänyt yhteen syötteet ja vastemuuttujat.

Tilastolliset tulokset ja näkemykset voidaan täsmällisimmin ilmaista matematiikan keinoin, mutta yhteys havaintoihin ja tieteelliseen päättelyyn on ominaista tilastotieteelliselle ajattelutavalle. Monet tilastolliset tulokset ovat syntyneet ja syntyvät vastauksena melko konkreettisiin kysymyksiin, joihin ei aina ole olemassa yleistä eleganttia ratkaisua. Tällaisten ongelmien valtava määrä tekee vaikeaksi kehittää kaiken kattavaa teoriaa, vaikka toisaalta yhteisiä yleisiä periaatteita voidaan esittää. Periaatteet voidaan kuitenkin kiteyttää tilastollisen mallin käsitteeseen.

Parametrinen mallintaminen

Parametrisessa lahestymistavassa oletetaan mustan laatikon sisälle jokin havaintoja gereroiva todennäköisyysmalli. Havaintojen malli on esimerkiksi muotoa

(8.2.1) $$\text{vaste} = f(\text{selittäjät, satunnaisvirhe, parametrit}),$$

missä funktion f yleinen matemaattinen muoto oletetaan tunnetuksi. Funktio riippuu kuitenkin yleensä tuntemattomista parametreista, jotka on estimoitava havainnoista. Sen jälkeen mallia voidaan käyttää riippuvuuksien tarkasteluun tai ennustamiseen.

Musta laatikko voisi näyttää esimerkiksi seuraavalta:

$$\boldsymbol{x} \longrightarrow \boxed{\text{lineaarinen regressio}} \longrightarrow y$$

Tässä siis ajatellaan, että lineaarinen regressiomalli kuvaa riittävän hyvin vasteen y riippuvuutta selittäjistä $\boldsymbol{x}$. Silloin siis f on 1. asteen polynomi (parametrien suhteen). Mallin pätevyys pyritään vahvistamaan havaintojen avulla tekemällä esimerkiksi yhteensopivuustestejä ja tarkastelemalla havaintoihin sovitetun mallin residuaaleja.

Esimerkki 8.3 Oletetaan, että riippumattomat havainnot $Y_1, Y_2, \ldots, Y_n$ noudattavat normaalijakaumaa $\mathrm{N}(\mu_i, \sigma^2)$,missä $E(Y_i) = \mu_i,\ i = 1, \ldots, n$. Oletetaan lisäksi, että $E(Y_i)$ riippuu lineaarisesti selittävästä muuttujasta x, joten

(8.2.2) $$\mu_i = \alpha + \beta x_i, \qquad 1 \leq i \leq n.$$

Malli voidaan kirjoittaa myös muodossa

$$Y_i = \mu_i + V_i,$$

missä $V_i = Y_i - E(Y_i)$. Virhetermi V_i noudattaa normaalijakaumaa $\mathrm{N}(0, \sigma^2)$. □

Esimerkki 8.4 Count Rumford Baierilainen oli ensimmäisiä, joka teki lämpöfysiikan kokeita. Vuonna 1798 hän teki kokeen, jossa kanuunan putki kuumennettiin 130°F lämpötilaan (°F = °$C \times 1.8 + 32$). Sitten putken annettiin jäähtyä ja lämpötila mitattiin tietyin väliajoin. Ulkolämpötila kokeen aikana oli 60°F. Newtonin jäähtymislaki sanoo, että $df/dt = -\theta(f - t_0)$, missä t_0 on ulkolämpötila. Silloin putken lämpötilan hetkellä t pitäisi olla

$$f(t, \theta) = 60 + 70\, e^{-\theta t}.$$

Kun mittauksia tehdään käytännössä, havainnot eivät aivan täsmällisesti toteuta lakia. Poikkeamat tulkitaan mittausvirheiksi. Tilastollinen malli, jossa mittausvirheet otetaan huomioon, on muotoa

(8.2.3) $$Y = f(t, \theta) + \epsilon,$$

missä $f(t,\theta) = 60 + 70\,e^{-\theta t}$. Mittausvirhe ϵ on satunnaismuuttuja, jonka odotusarvo $E(\epsilon)$ oletetaan nollaksi.

Mittausvirheet noudattavat yleensä erittäin hyvin normaalijakaumaa, siksi tilastollista päättelyä ajatellen oletetaankin tavallisesti, että ϵ noudattaa normaalijakaumaa. Näillä tilastollisilla oletuksilla malli (8.2.3) voidaan luonnehtia seuraavasti:

$$Y \sim \mathrm{N}(60 + 70\,e^{-\theta t}, \sigma^2).$$

Varsinainen fysikaalisesti kiinnostava estimoitava parametri on θ ja σ^2 koejärjestelyyn liittyvä virhevarianssi. □

Esimerkki 8.5 Tutkitaan virran voimakkuuden (ampeereina) vaikutusta hitsauksessa syntyvän hitsaussauman minimiläpimittaan (Aineisto: The Welding Institute, Abingdon, P.M.E.Altham). Virran voimakkuus (x) on selittävä muuttuja ja syntyvän hitsaussauman minimiläpimitta (y) on selitettävä muuttuja. Kokeilaan aineistoon lineaarista regressiomallia.

$$Y_i = \alpha + \beta x_i + V_i,\ i = 1, \ldots, n,$$

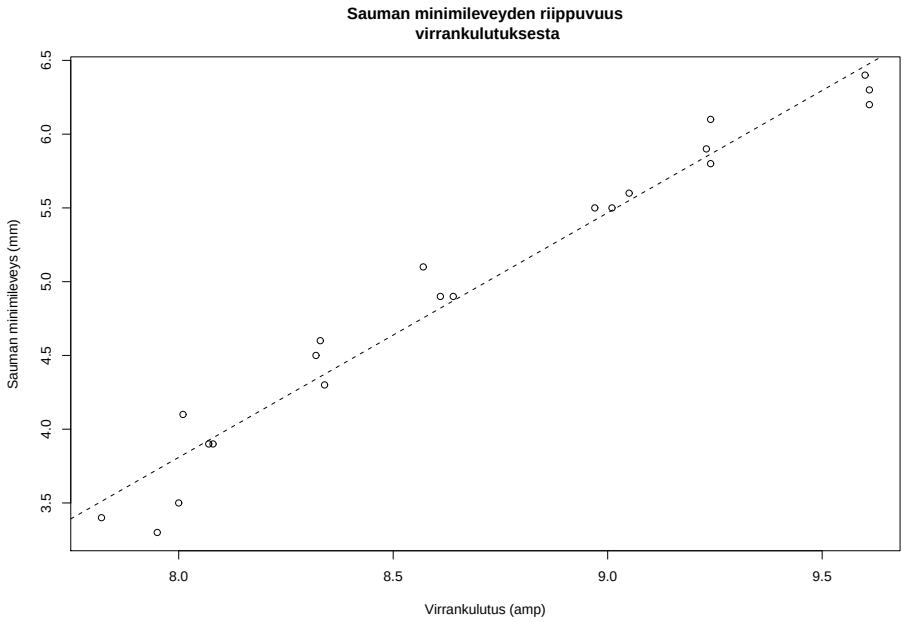
missä x_i:t ovat vakioita, $E(V_i) = 0$ ja $V_i \perp\!\!\!\perp V_j$, kun $i \neq j$. Aineistossa on 21 havaintoa, eli $n = 21$. Kuviossa 8.1 on aineistoon sovitettu pienimmän neliösumman menetelmällä suora.

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42623 -0.07282  0.01637  0.08269  0.34586

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -9.45427    0.65526  -14.43 1.09e-11 ***
x            1.65793    0.07531   22.01 5.53e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2012 on 19 degrees of freedom
Multiple R-squared: 0.9623,     Adjusted R-squared: 0.9603
F-statistic: 484.6 on 1 and 19 DF,  p-value: 5.529e-15
```

Tutkittaessa tarkemmin mallin residuaaleja voidaan havaita kvadraattinen trendi, mikä viittaa siihen, että kannattaa kokeilla kvadraattista mallia

$$Y_i = \alpha + \beta x_i + \gamma x_i^2 + V_i,\ i = 1, \ldots, n.$$

Myös 2. asteen termin kertoimen estimaatti $\hat{\gamma}$ on tilastollisesti merkitsevä ja mallin selitysaste on 0.9761. Kun lasketaan estimaattien $\hat{\beta}$ ja $\hat{\gamma}$ välinen korrelaatio, saadaan melko tarkkaan -1. Mitä tästä pitäisi päätellä? □



Kuvio 8.1. Hitsauksessa käytetyn virran voimakkuudella (ampeeria) selitetään syntyvän hitsaussauman minimiläpimittaa.

Esimerkissä 8.3 oletettiin, että havainnot noudattavat normaalijakaumaa. Esimerkissä 8.5 ei tällaista oletusta tehty. Kun aineistoon sovitetaan regressiomallia, on valittava parametrien estimointimenetelmä. Tavanomainen lähestymistapa on sovittaa pienimmän neliösumman suora aineistoon. Siinä minimoidaan neliösumma

$$g(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{8.2.4}$$

parametrien α ja α suhteen. Minimi saavutetaan pienimmän neliösumman ratkaisulla

$$\begin{aligned} \hat{\alpha} &= \bar{y} - \hat{\beta}\bar{x} \\ \hat{\beta} &= \frac{S_{xy}}{S_{xx}}, \end{aligned} \tag{8.2.5}$$

missä $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ ja $S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$. Matriisimuodossa lausekkeen (8.2.4) minimoiva ratkaisu (8.2.5) on

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}, \tag{8.2.6}$$

missä

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{pmatrix}^T \quad \text{ja} \quad \boldsymbol{y} = \begin{pmatrix} y_1 & y_2 & \cdots & y_n \end{pmatrix}^T.$$

Huomattakoon, että pienimmän neliösumman ratkaisu (8.2.5) on lineaarinen estimaattori. Selittäjän arvota $x_1, \ldots, x_n$ ovat vakioita ja havainnot $Y_1, \ldots, Y_n$ satunnaismuuttujia. Esimerkiksi $\hat{\beta}$ voidaan kirjoittaa

$$\hat{\beta} = \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{S_{xx}} Y_i.$$

Estimaattori $\hat{\beta}$ on siis havaintojen $Y_1, \ldots, Y_n$ lineaarinen yhdiste

$$\hat{\beta} = \sum_{i=1}^{n} b_i Y_i,$$

missä kertoimet $b_i = \frac{(x_i - \bar{x})}{S_{xx}}$, $1 \leq i \leq n$, ovat vakioita. Sama asia nähdään tietysti yleisemmin lausekkeesta (8.2.6), missä havaintovektori $\boldsymbol{y}$ kerrotaan vakiomatriisilla $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. Pienimmän neliösumman estimaattorit ovat myös harhattomia ja ne ovat *o*ptimaalisia lineaaristen harhattomien estimaattoreiden joukossa: niillä on pienin varianssi harhattomien lineaaristen estimaattoreiden luokassa (Gaussin ja Markovin lause). Sanomme, että sellainen estimaattori on paras lineaarinen harhaton estimaattori (Best Linear Unbiased Estimator, BLUE).

Pienimmän neliösumman (pns) estimointimenetelmässä minimoidaan kvadraattinen tappiofunktio (8.2.4). On tunnettua, että poikkeavilla havainnoilla on suuri vaikutus pns-estimaatteihin. Jos minimointikriteeriksi valitaan absoluuttipoikkeamien summa

$$\sum_{i=1}^{n} |y_i - \alpha - \beta x_i|,$$

saadaan pienimpien absoluuttipoikkeamien estimaatit. Nämä estimaatit eivät ole yhtä herkkiä poikkeavien havaintojen vaikutukselle. Sanomme, että pienimmän absoluuttipoikkeaman estimaattori on robustimpi kuin pns-estimaattori. Robustin regression käsite perustuu lineaaristen harhattomien estimaattoreiden luokkaa laajempaan estimaattoriluokkaan, jossa minimoidaan funktio

$$\sum_{i=1}^{n} \frac{\psi(y_i - \alpha - \beta x_i)}{\sigma},$$

missä ψ on tappiofunktio ja σ on skaalaustekijä. Kun $\psi(x) = x^2$, on kyseessä tavallinen pns-regressio ja pienimmän absoluuttipoikkeaman regressio saadaan, kun $\psi(x) = |x|$. Huber (1975) esitti yleisen tappiofunktion

$$\psi_c(x) = \begin{cases} x^2, & |x| \leq c \\ 2c|x| - c^2 & |x| > c. \end{cases}$$

Funktio $\psi_c(x)$ on siis kvadraattinen välillä $[-c, c]$ ja lineaarinen tämän välin ulkopuolella.

Esimerkki 8.6 Oletetaan, että havainnot $Y_1, Y_2, \ldots, Y_n$ noudattavat mallia

$$Y_i = \beta x_i + V_i, \qquad 1 \leq i \leq n, \tag{8.2.7}$$

missä $E(Y_i) = \beta x_i \Leftrightarrow E(V_i) = 0$, $\mathrm{Var}(Y_i) = \sigma^2$, $1 \leq i \leq$ ja $Y_i \perp\!\!\!\perp Y_j$, $i \neq j$. Mallissa ei ole siis vakiotermiä ja sovitesuorat kulkevat origon kautta. Parametrin β pienimmän neliösumman estimaattori on

$$\hat{\beta} = \sum_{i=1}^{n} \frac{x_i}{S_{xx}} Y_i, \tag{8.2.8}$$

missä nyt $S_{xx} = \sum_{i=1}^{n} x_i^2$.

Tarkastellaan nyt estimaattorin (8.2.8) optimaalisuutta. Estimaattori on harhaton, koska

$$E(\hat{\beta}) = \sum_{i=1}^{n} \frac{x_i}{S_{xx}} E(Y_i) = \frac{\beta}{S_{xx}} \sum_{i=1}^{n} x_i x_i = \beta.$$

Olkoon $\tilde{\beta} = \sum_{i=1}^{n} d_i Y_i$ jokin β:n lineaarinen harhaton estimaattori, missä siis $d_1, \ldots, d_n$ ovat vakioita. Harhattomuudesta seuraa, että

$$E(\tilde{\beta}) = \sum_{i=1}^{n} d_i E(Y_i) = (\sum_{i=1}^{n} d_i x_i)\beta = \beta$$

kaikilla β:n arvoilla, joten $\sum_{i=1}^{n} d_i x_i = 1$. Kirjoitetaan kertoimet d_i muodossa $d_i = g_i + e_i$, missä $g_i = x_i / S_{xx}$, $1 \leq i \leq n$, ovat pienimmän neliösumman estimaattorin (8.2.8) kertoimet. Silloin $\sum_{i=1}^{n} e_i x_i = 0$, koska $\hat{\beta}$:n ja $\tilde{\beta}$:n harhattomuuden perusteella $\sum_{i=1}^{n} d_i x_i = \sum_{i=1}^{n} g_i x_i = 1$. Nyt

$$\begin{aligned}
\mathrm{Var}(\tilde{\beta}) &= \sum_{i=1}^{n} d_i^2 \,\mathrm{Var}(Y_i) = \sum_{i=1}^{n} (g_i + e_i)^2 \sigma^2 \\
&= \sigma^2 (\sum_{i=1}^{n} g_i^2 + \sum_{i=1}^{n} e_i^2 + 2 \sum_{i=1}^{n} e_i g_i) \\
&= \sigma^2 (\sum_{i=1}^{n} g_i^2 + \sum_{i=1}^{n} e_i^2),
\end{aligned}$$

sillä $\sum_{i=1}^{n} e_i g_i = (\sum_{i=1}^{n} e_i x_i)/S_{xx} = 0$, koska $\sum_{i=1}^{n} e_i x_i = 0$. Olipa siis $\tilde{\beta}$ mikä tahansa β:n lineaarinen harhaton estimaattori mallissa (8.2.7), niin

$$\mathrm{Var}(\hat{\beta}) \leq \mathrm{Var}(\tilde{\beta}).$$

□

Algoritminen mallintaminen

Algoritminen lähestymistapa on saavuttanut suosiota ja sovellusmahdollisuuksia tietokoneiden laskentakapasiteetin kasvun myötä. Tässä ajattelutavassa mustan laatikon sisältö on monimutkainen ja tuntematon. Funktion matemaattista suljetun muodon lauseketta ei tunneta. Sen sijaan funktio $f(\boldsymbol{x})$ pyritään määrittämään algoritmisesti– algoritmi laskee $\boldsymbol{x}$:n perusteella ennusteita y:lle. Algoritmi siis pyritään muokkaamaan sellaiseksi, että se antaa hyviä ennusteita. Musta laatikko näyttäisi tältä:

$$\boldsymbol{x} \longrightarrow \boxed{\text{tuntematon}} \longrightarrow y$$

Esimerkiksi neuroverkot kuuluvat tähän kategoriaan. Mallin pätevyyttä arvioidaan ennustevirheen avulla.

Tavallisesti havaintojen vaihtelu jaetaan systemaattiseen osaan ja satunnaisosaan ja havintoarvojen ajatellaan muodostuvan näistä kahdesta komponentista additiivisesti:

$$\begin{aligned}\text{havainnot} &= f(\text{selittävät muuttujat, parametrit}) + \text{satunnaisosa} \\ &= \text{systemaattinen osa} + \text{satunnaisosa}.\end{aligned}$$

Esitys (8.2.1) on itse asiassa varsin yleinen, joka sallii monimutkaisetkin vaikutusmekanismit. Yksinkertaistavia oletuksia kuitenkin tarvitaan, jotta mallit pystytään ymmärtämään ja analysoimaan. Havaintojen oletetaan olevan peräisin jostain jakaumaperheestä, tavallisimmin ns. parametrisesta jakaumaperheestä. Systemaattinen osa on esimerkiksi havaintojen Y_1, Y_2, ..., Y_n odotusarvoja $E(Y_i)$, $1 \leq i \leq n$, koskeva oletus, joka lausutaan vaikkapa regressiofunktiona. Tavallisesti odotusarvo riippuu joistain selittävistä muuttujista (eli kovariaatista). Tilastollisen mallin voidaan sanoa olevan havaintojen yhteisjakaumaa ja systemaattista osaa koskevien oletusten joukko.

Ehdollinen odotusarvo

Kokeellisessa tilanteessa selittäjä x on koevakio. Tutkija päättää, millä x:n arvoilla hän tekee havaintoja riippuvasta muuttujasta Y. Esimerkiksi törmäystestissä valitaan törmäysnopeudet $x_1, \ldots, x_n$. Näillä selittäjän arvoilla mitataan vastemuuttujan (tai vastemuuttujien) arvot. Regressioanalyysia käytetään kuitenkin myös ei-kokeellisessa tilanteessa, jossa tutkija ei voi kontrolloida x:n arvoja. Silloin x on satunnaismuuttuja, jonka arvo havainnoidaan usein samanaikaisesti vastemuuttujan kanssa. On huomattava, että regressiomallissa (8.2.2) tarkastellaan ehdollista odotusarvoa

$$E(Y|x) = \mu(x),$$

missä Y:n ehdollisen odotusarvon oletetaan olevan x:n linearinen funktio $\mu(x) = \alpha + \beta x$. Suoran kertoimet α ja β ovat tuntemattomia parametreja, jotka estimoidaan havainnoista.

Parametrinen jakaumaperhe

Tilastotieteen oppikirjoissa lähdetään tavallisesti liikkeelle melko teknisesti. Sanotaan, että havainnot $Y_1, \ldots, Y_n$ ovat otos jostain tuntenmattomasta jakaumasta F, missä F on siis jakauman kertymäfunktio. Tavallisesti jakaumasta tehdään joitain oletuksia. Tilanne voi olla esimerkiksi sellainen, että jakauma voidaan oletettaa symmetriseksi. Tällä kurssilla käytetään useimmiten parametrista lähestymistapaa. Silloin jakauman ajatellaan kuuluvan johonkin parametriseen jakaumaperheeseen

$$\mathcal{F} = \{\, F(x;\theta),\ \theta \in \Theta \,\}$$

missä $F(x;\theta)$ on kertymäfunktio jokaisella kiinnitetyllä θ:n arvolla.

Käsittelemissämme päättelyongelmissa operoimme tavallisesti tiheysfunktioiden avulla, joten jakaumaperhe on silloin suoraviivaisempaa luonnehtia tiheysfunktioiden joukkona

$$\mathcal{F} = \{\, f(x;\theta),\ \theta \in \Theta \,\}.$$

Suure θ on siis parametri ja sen arvojoukko Θ on parametriavaruus. Valitsemalla yksi parametrin θ arvo saadaan täysin määrätty jakauma. Edellä olemme nähneet, että θ voi riippua selittävien muuttujie arvoista. Kun parametrin θ arvo valitaan havaintojen perusteella, saadaan θ:n *piste-estimaatti*. Parametrin (parametrien) arvon määrittämistä havaintojen perusteella sanotaan *piste-estimoinniksi*.

Esimerkki 8.7 Tarkastellaan auto-onnettomuuksien vakavuusastetta, kun selittäjänä on kuljettajan ikä. Usein väitetään, että nuoret kuljettajat aiheuttavat keskimääräistä enemmän vakavia onnettomuuksia.

Taulukko 8.4. Vakavien onnettomuuksien lukumäärä alueella A tammikuussa vuonna 2000.

Yli 21-vuotiaat		Alle 21-vuotiaat	
Kuolemaan johtaneet	Muut	Kuolemaan johtaneet	Muut
Y_1	Y_2	Y_3	Y_4
11	62	4	7

Oletetaan, että onnettomuuksien lukumäärä kuukaudessa noudattaa Poissonin jakaumaa $\mathrm{Poi}(\lambda)$. Tarkastellaan neljää onnettomuustyyppiä, jotka on määritelty kuljettajan iän ja onnettomuuden vakavuusasteen mukaan. Onnettomuuksien lukumäärien $Y_i, 1 \leq i \leq 4$, eri kategorioissa oletetaan noudattavan toisistaan riippumatta Poissonin jakaumaa $\mathrm{Poi}(\lambda_i)$. Oheisessa taulukossa on annettu eräs aineisto. Silloin esimerkiksi kuolemaan johtaneiden

onnettomuuksien lukumäärä Y_3 alle 21-vuotiaiden ryhmässä noudattaa Poissonin jakaumaa Poi(λ_3). Parametrit λ_1, λ_2, λ_3 ja λ_4 ovat satunnaismuuttujien Y_1, Y_2, Y_3 ja Y_4 odotusarvoja. Odotusarvo λ_i kertoo onnettomuusasteen i. kategoriassa. Vastaavasti esimerkiksi yli 21-vuotiaiden onnettomuusaste on $\lambda_1 + \lambda_2$ ja alle 21-vuotiaiden $\lambda_3 + \lambda_4$. Merkitään $\theta_1 = \lambda_1 + \lambda_2$ ja $\theta_L = \lambda_3 + \lambda_4$. Näin todennäköisyys, että yli 21-vuotias aiheuttaa kohtalokkaan onnettomuuden, on

$$\pi_1 = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

ja alle 21-vuotiaan todennäköisyys aiheuttaa kohtalokas onnettomuus on

$$\pi_2 = \frac{\lambda_3}{\lambda_3 + \lambda_4}.$$

Nelikko $(\theta_1, \pi_1, \theta_2, \pi_2)$ muodostaa uuden parametrisoinnin, joka saattaa olla tulkinnallisesti selkeämpi ja mielenkiintoisempi kuin alkuperäinen. □

Esimerkki 8.8 Tarkastellaan nyt logistista regressiomallia, kun havainnot noudattavat binomijakaumaa. Tarkastellaan lentokoneiden rakennuksessa käytettävien metallin kiinnittimien puristuskestävyyttä. Aineisto on kirjasta "Introduction to Linear Regression Analysisi" (Montgomery & Peck, 1982). Painekuormitus x on selittävä muuttuja, jonka arvo kasvaa 2500:sta 4300:aan 200:n yksikön (psi) välein. Yksikkö psi (Pounds per Square Inch) on paunaa (naulaa)/per neliötuuma) ja pauna = 425 g. Aineistossa

$$\begin{aligned} n &= \text{testattavien kiinnittimien lkm annetulla kuormituksella} \\ y &= \text{särkyvien kiinnittimien lkm annetulla kuormituksella.} \end{aligned}$$

Oletetaan, että särkyvien kiinnittimien lukumäärät noudattavat binomijakaumaa

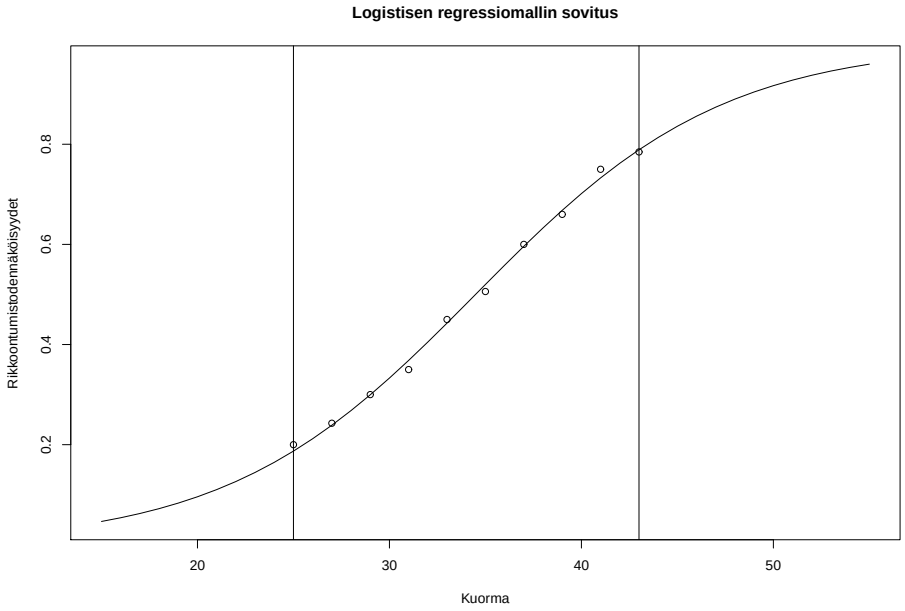
$$Y_i \sim \text{Bin}(n_i, \pi_i),\ i = 1, \ldots, 10$$

ja $Y_i \perp\!\!\!\perp Y_j$, kun $i \neq j$. Mallinnetaan parametrien π_i arvojen riippuvuutta kuormituksen määristä, joita on 10. Logistinen malli on muotoa

$$\log(\frac{\pi_i}{1-\pi_i}) = \alpha + \beta x_i, \quad 1 \leq i \leq 10,$$

missä $x_1, \ldots, x_{10}$ ovat kuormituksen määriä. Funktio $\log(\frac{\pi_i}{1-\pi_i})$ on yleistetyissä lineaarisissa malleissa ns. logit-linkki. Selitettävänä muuttujana on särkyvien kiinnittimien suhteellinen osuus $p = r/n$.

```
Call:
glm(formula = p ~ kuorma, family = binomial, weights = n)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.29475  -0.11129   0.04162   0.08847   0.35016
```



Kuvio 8.2. Metallinkiinnittimien rikkoontumistodennäköisyys kuormituksen funktiona.

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.33971    0.54569  -9.785   <2e-16 ***
kuorma       0.15484    0.01575   9.829   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 112.83207  on 9  degrees of freedom
Residual deviance:   0.37192  on 8  degrees of freedom
AIC: 49.088

Number of Fisher Scoring iterations: 3
```

□

8.3 Estimoinnista

Tarkastelemme nyt satunnaismuuttujia, joiden todennäköisyysfunktion (tai tiheysfunktion) funktionaalinen muoto tunnetaan, mutta jakauma riippuu jostain tuntemattomasta parametrista θ. Parametrin θ mahdolliset arvot

kuuluvat johonkin annettuun joukkoon Θ, jota kutsutaan *parametriavaruudeksi*. Tiedetään esimerkiksi, että jonkin tuotteen elinaika X noudattaa eksponenttijakaumaa

$$f(x;\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}, \qquad 0 < x < \infty,$$

missä $\theta \in \Theta = \{\, \theta \mid 0 < \theta < \infty \,\}$. Parametriavaruus Θ on siis positiivisten reaalilukujen joukko. Haluamme valita funktioperheestä

$$\mathcal{F} = \{\, f(x;\theta),\ \theta \in \Theta \,\}$$

yhden tiheysfunktion, joka esittää parhaiten tuotteen elinaikaa. Valitaan siis yksi parametrin θ arvo eli parametrin θ *piste-estimaatti*, joka määrittää jakauman.

Parametrin arvo arvioidaan eli estimoidaan havaintojen perusteella. Teemme jakaumasta havainnon $X = x$ ja estimoimme parametrin θ arvon havainnon x perusteella. Parametrin θ estimoimiseen käytettävää otosfunktiota $T(X)$ kutsutaan parametrin θ *estimaattoriksi* ja estimaattorin $T(X)$ arvoa $t = T(x)$ kutsutaan parametrin θ *estimaatiksi*. Estimaattori pyritään valitsemaan siten, että se antaa hyviä arvioita parametrista θ.

Esimerkki 8.9 Estimoidaan ehdokkaan A kannattajien suhteellinen osuus θ eräässä suuressa kaupungissa. Valitaan kaupungin äänioikeutetuista satunnaisesti n henkilöä, joilta tiedustellaan heidän kantaansa ehdokkaasta A. Olkoon X ehdokkaan A kannattajien lukumäärä otoksessa. Koska populaation koko on suuri verrattuna otoskokoon n, voidaan olettaa, että $X \sim \mathrm{Bin}(n,\theta)$, missä θ on todennäköisyys, että satunnaisesti valittu henkilö kannattaa A:ta. Binomijakaumaa noudattavan satunnaismuuttujan X todennäköisyysfunktio on muotoa

$$f(x;n,\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad x = 0,1,\ldots,n; \quad 0 \le \theta \le 1.$$

Binomijakauman parametriavaruus on $\Theta = \{\, \theta \mid 0 \le \theta \le 1 \,\}$. Tehtävänämme on määrittää θ:n estimaattori $T(X)$ siten, että havaitun arvon $X = x$ perusteella saadaan hyvä θ:n piste-estimaatti $T(x)$. Havainnon $X = x$ todennäköisyys on

$$P(X = x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}. \tag{8.3.1}$$

Eräs tapa määrittää θ:n estimaatti on tarkastella todennäköisyyttä $P(X = x;\theta)$ parametrin θ funktiona ja etsiä sellainen θ:n arvo, että havainnon x todennäköisyys saavuttaa maksiminsa. Voidaan osoittaa, että havainnon $X = x$ todennäköisyys maksimoituu, kun $\theta = x/n$. Tätä estimaattia kutsutaan θ:n *suurimman uskottavuuden estimaatiksi* ja sitä merkitään

$$\hat{\theta} = \frac{x}{n}.$$

□

8.4 Uskottavuussuhde

Todennäköisyyden (8.3.1) lausekkeessa tekijä $\binom{n}{x}$ ei riipu parametrista θ. *Uskottavuusfunktio* on parametrin θ funktio ja merkitsemme

$$L(\theta) = f(x;\theta), \tag{8.4.1}$$

missä $f(x;\theta)$ on todennäköisyysfunktio ja $f(x;\theta)$ on siis havainnon $X = x$ todennäköisyys. Uskottavuussuhteen

$$\frac{L(\theta_1)}{L(\theta_2)} = \frac{f(x;\theta_1)}{f(x;\theta_1)}, \tag{8.4.2}$$

avulla vertaillaan kahden parametrin arvon θ_1 ja θ_2 suhteellista uskottavuutta, kun on havaittu $X = x$. Uskottavuuspäättelyn perusta on uskottavuussuhde. Silloin uskottavuusfunktio

$$L(\theta) = c \cdot f(x;\theta), \tag{8.4.3}$$

antaa samat uskottavuussuhteet kuin (8.4.1), kun vakio c ei riipu parametrista θ. Sen sijaan c voi riippua havainnosta x. Monesti vakio c pyritään valitsemaan siten, että $L(\theta)$:lle saadaan yksinkertainen lauseke. Uskottavuusfunktioon perustuvat päätelmät eivät siis riipu vakion c valinnasta.

Tavallisesti uskottavuusfunktio tulee olemaan useiden tekijöiden tulo ja mm. siitä syystä on osoittautunut käteväksi työskennellä uskottavuusfunktion logaritmin avulla. *Logaritmoitu uskottavuusfunktio* $l(\theta)$ on uskottavuusfunktion luonnollinen logaritmi eli

$$l(\theta) = \log L(\theta). \tag{8.4.4}$$

Esityksestä (8.4.3) seuraa, että

$$l(\theta) = \log c + \log f(x;\theta),$$

missä vakio c ei siis riipu θ:sta. Jatkossakin kaikki logaritmit ovat luonnollisia logaritmeja, ellei toisin mainita.

Suurimman uskottavuuden estimaatti (SUE) $\hat{\theta}$ on se parametrin θ arvo, joka maksimoi havainnon x todennäköisyyden $f(x;\theta)$. Sama arvo $\hat{\theta}$ maksimoi myös funktiot $L(\theta)$ ja $l(\theta)$. Suurimman uskottavuuden estimaatti $\hat{\theta}$ on siis uskottavuusfunktion ja logaritmoidun uskottavuusfunktion maksimikohta. Tavallisesti tarkastellaan logaritmoitua uskottavuusfunktiota, koska se on usein matemaattisesti yksinkertaisempi kuin uskottavuusfunktio. Logaritmoidulla uskottavuusfunktiolla on myös teoreettisesti merkittävä tilastollinen tulkinta.

Esimerkki 8.10 Tarkastellaan edelleen Esimerkkiä 8.9, jossa havaintojen todennäköisyysfunktio on $f(x;\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$. Kun uskottavuusfunktiossa (8.4.1) valitaan vakion arvoksi $c = 1/\binom{n}{x}$, saadaan esitysmuoto

$$L(\theta) = \theta^x(1-\theta)^{n-x}, \qquad 0 \le \theta \le 1.$$

Tässä uskottavuusfunktion esityksessä ei ole ”turhia” vakiotekijöitä. Tätä uskottavuusfunktion esitysmuotoa kutsutaan myös *uskottavuusfunktion ytimeksi*. Esitämme uskottavuusfunktion usein tässä *ydinmuodossa*. Logaritmoitu uskottavuusfunktio on

$$l(\theta) = x \log \theta + (n - x) \log(1 - \theta), \qquad 0 < \theta < 1.$$

Parametrin θ suurimman uskottavuuden estimaatti on siis se θ:n arvo, joka maksimoi funktion $l(\theta)$. Huomattakoon, että $l(\theta)$ ei ole määritelty välin $[0, 1]$ päätepisteissä, mutta $L(\theta)$ on. □