9.1 Johdanto lukuun 9: lineaarinen regressio

- Havaintoaineisto koostuu kahden muuttujan x ja y arvoista

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

  Pari (x_i, y_i) on mitattu havaintoyksiköstä i.
- Emme aluksi aseta tilastollista mallia, vaan tarkastelemme tilannetta heuristisesti.
- Päämääränä 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.

Selittäjän ja vasteen yhteys

- Ajattelemme, että x vaikuttaa y:n arvoon osapuilleen kaavan

$$y = f(x)$$

  mukaisesti.
- Funktio f on ainakin osittain tuntematon; arvo $f(x)$ riippuu x:n lisäksi tuntemattomien parametrien arvoista.
- Pisteet (x_i, y_i) eivät 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) pitää määrittää havaintojen perusteella.

Lineaarinen rakennemalli

- 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 funktiolla.
- Parametri α on suoran vakiotermi (engl. *intercept*).
- Parametri β on suoran kulmakerroin (engl. *slope*).
- Tuntemattomien parametrien arvot pitää määrittää havaintojen perusteella.

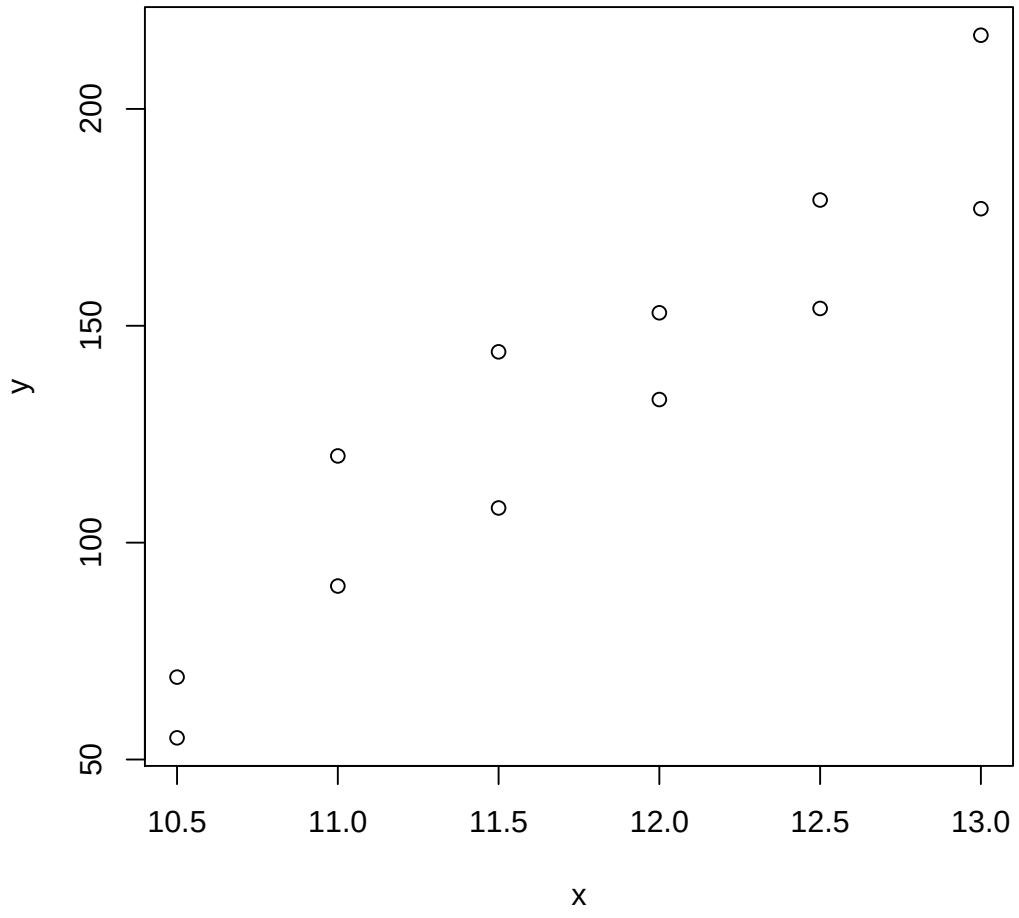
Esimerkki: kuminauha-aineisto

- 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.

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

Kuminauha-aineisto hajontakuvana



9.2 Suoran sovittaminen pienimmän neliösumman menetelmällä

- Emme vieläkään aseta tilastollista mallia, vaan tarkastelemme tehtävää heuristisesti.
- Parametrit eli suoran kertoimet α ja β voidaan määrittää sellaisella tavalla, jossa yritetään saada vasteet y_i sekä niitä vastaavat sovitteet tai ennusteet $\alpha + \beta x_i$ mahdollisimman samankaltaisiksi jonkin kriteerin mielessä.
- Samankaltaisuutta olisi mahdollista mitata erilaisilla tavoilla, ja erilaiset tavat johtaisivat erilaisin parametriestimaatteihin.

Pienimmän neliösumman suora

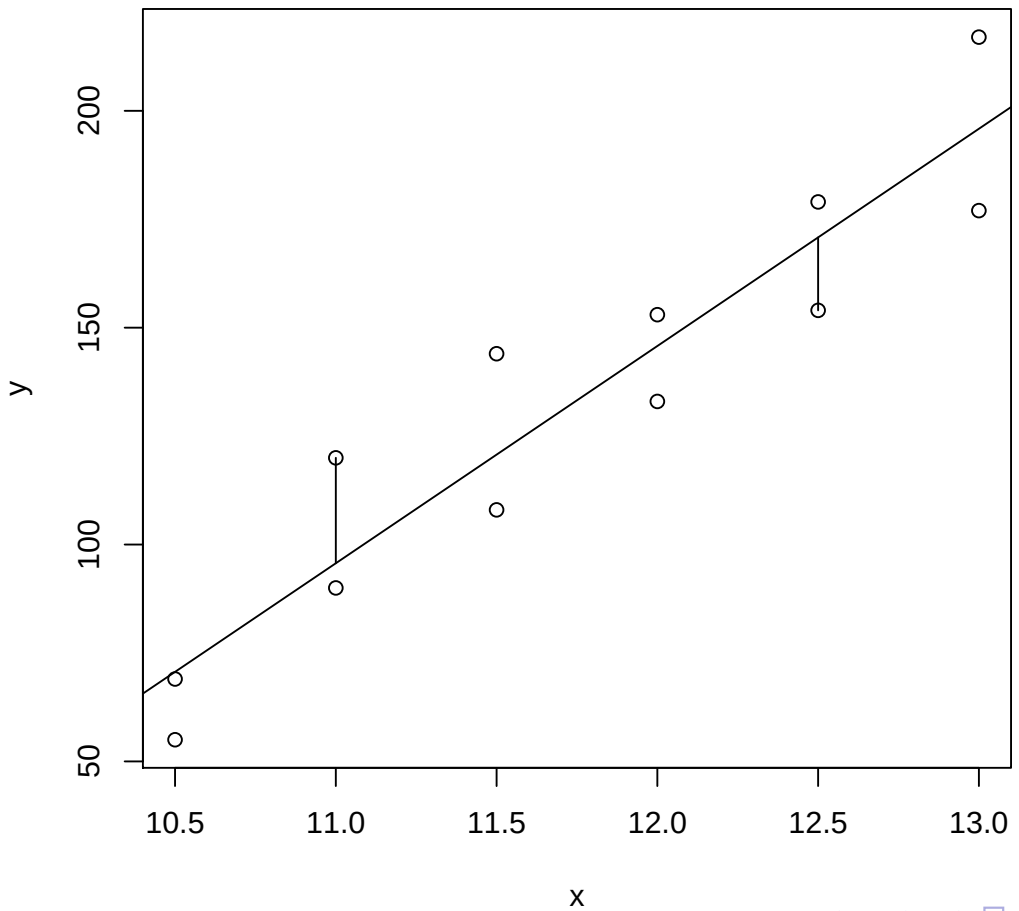
- Tärkein samankaltaisuuskriteeri on virheiden neliöiden summa

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

- 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.
- Pienimmän neliösumman suora on se suora, jota vastaavat kertoimet α ja β minimoivat virheiden neliöiden summan (1).
- Tällöin sovitamme suoran aineistoon käyttämällä pienimmän neliösumman menetelmää (engl. *method of least squares*) eli PNS-menetelmää.

PNS-suora kuminauha-aineistolle

Kahdelle (x, y)-pisteelle on piirretty sen y-akselin suuntaan mitattu etäisyys PNS-suorasta.



Miksi PNS-kriteeri on tärkeä

- Sen käyttö johtaa yksinkertaisiin kaavoihin:
  - ratkaisu on olemassa, se on yksikäsitteinen ja helposti laskettavissa.
- PNS-kriteeri voidaan perustellä myöhemmin esitettävällä yksinkertaisella tilastollisella mallilla sekä suurimman uskottavuuden periaattella.

PNS-suoran kulmakerroin

- PNS-menetelmä valitsee kulmakertoimelle β arvon

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

- $\bar{x}$ ja $\bar{y}$ ovat keskiarvot

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

- Jotta b olisi määritelty, täytyy olettaa, että kaavan (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.

PNS-suoran vakiotermi

- Kun PNS-suoran kulmakerroin b on laskettu, niin PNS-menetelmä antaa vakion α arvoksi

$$\hat{\alpha} = a = \bar{y} - b\bar{x}. \tag{3}$$

- Tuntemattomia parametreja (tässä kertoimia) merkitään kreikkalaisilla kirjaimilla. Niiden estimaatteja voidaan merkitä joko laittamalla hattu parametrin päälle tai sitten käyttämällä kreikkalaista kirjainta vastaavaa latinalaista kirjainta.

PNS-suoran kulmakerroin otoskovarianssin ja -varianssin 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

$$s(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \tag{4}$$

- Otoskovarianssissa käytetään jakajana lukua $n - 1$ sen takia, että näin saadaan populaatiokovarianssin harhaton estimaattori.

Satunnaismuuttujien kovarianssi

- Jos X ja Y ovat satunnaismuuttujia, niin niiden kovarianssi (eli populaatiokovarianssi) $\mathrm{cov}(X, Y)$ määritellään odotusarvona

$$\mathrm{cov}(X, Y) = E[(X - EX)(Y - EY)]. \tag{5}$$

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

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

missä $\mathbf{X} = (X_1, \ldots, X_n)$ ja $\mathbf{Y} = (Y_1, \ldots, Y_n)$ ja $s(\mathbf{X}, \mathbf{Y})$ on satunnaisotoksesta laskettu otoskovarianssi.

PNS-suoran kulmakerroin

- Otoskovarianssin (4) avulla lausuttuna

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

- Osoittaja on x- ja y-lukujen otoskovarianssi, ja nimittäjä on x-lukujen otosvarianssi.

PNS-suoran vakiotermi

- PNS-suoran yhtälö $y = a + bx$ voidaan esittää myös kaavalla

$$y - \bar{y} = b(x - \bar{x}), \tag{6}$$

- Tästä nähdään, että PNS-suora kulkee parien (x_i, y_i) keskiarvon $(\bar{x}, \bar{y})$ kautta.
- Tämä muotoilu on helpompi muistaa kuin aikaisempi kaava $a = \bar{y} - b\bar{x}$, ja tästä muotoilusta nähdään helposti oikea lauseke PNS-suoran vakiotermille.

PNS-suoran kertoimien laskeminen kuminauha-aineistolle

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)`).

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

## [1] 50.09

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

## [1] -455.3
```

PNS-suora kuminauha-aineistolle R:llä (jatkoa)

Otoskovarianssin saisi laskettua myös suoraan määritelmää käyttämällä.

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

## [1] 39.84
```

R:stä löytyy toki myös 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 PNS-suoran kertoimien kaavat (2) ja (3) suoralla laskulla.
- Todistus perustuu siihen tosiasiaan, että $\mathrm{SS}(\alpha, \beta)$ on kertoimien suhteen toisen asteen polynomi, jota voidaan analysoida neliöksi täydentämällä.
- Määritelmä oli

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

- Käytämme hyväksi esitystä

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

Laskun vaiheita

- Kerrotaan trinomi auki summan sisällä:

$$(a+b+c)^2 = a^2 + b^2 + c^2 + 2ab + 2ac + 2bc$$

- Hajotetaan alkuperäinen summa kuudeksi summaksi, ja otetaan vakiot pois summamerkin alta.
- Käytetään sitä havaintoa, että

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

Välivaihe

Näiden vaiheiden jälkeen saamme esityksen

$$\mathrm{SS}(\alpha, \beta) = q_{yy} + n(\alpha - \bar{y} + \beta\bar{x})^2 + \beta^2 q_{xx} - 2\beta q_{xy},$$

missä merkittiin

$$q_{xx} = \sum (x_i - \bar{x})^2, \quad q_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}), \quad q_{yy} = \sum (y_i - \bar{y})^2. \tag{7}$$

Täydennetään neliöksi

$$
\begin{aligned}
\mathrm{SS}(\alpha,\beta) &= q_{yy} + n(\alpha - \bar{y} + \beta\bar{x})^2 + \beta^2 q_{xx} - 2\beta q_{xy} \\
&= q_{yy} + n(\alpha - \bar{y} + \beta\bar{x})^2 \\
&\qquad q_{xx}\left(\beta^2 - 2\beta\frac{q_{xy}}{q_{xx}} + \frac{q_{xy}^2}{q_{xx}^2}\right) - \frac{q_{xy}^2}{q_{xx}} \\
&= q_{yy} - \frac{q_{xy}^2}{q_{xx}} + \underbrace{q_{xx}\left(\beta - \frac{q_{xy}}{q_{xx}}\right)^2}_{\geq 0} + \underbrace{n(\alpha - \bar{y} + \beta\bar{x})^2}_{\geq 0}.
\end{aligned}
$$

Johtopäätös

- Edellisen perusteella kaikilla (α, β) pätee

$$\mathrm{SS}(\alpha, \beta) \geq q_{yy} - \frac{q_{xy}^2}{q_{xx}} = \mathrm{SS}(a, b). \tag{8}$$

- Tässä alaraja saavutetaan valitsemalla $\alpha = a$ ja $\beta = b$, eli aikaisemmin väitetyt arvot PNS-suoran kertoimille:

$$b = \frac{q_{xy}}{q_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

  ja tämän avulla

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

- PNS-suoran kertoimien kaavat on saatu todistettua.

9.3 Lineaarinen malli

- Asetetaan lineaarinen malli, jossa virheet ovat normaalijakautuneita.
- Havaintoja $y_1, \ldots, y_n$ vastaa mallissa satunnaismuuttujat $Y_1, \ldots, Y_n$, joiden yhteisjakauma voidaan esittää kaavalla

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n. \tag{9}$$

- Luvut $x_1, \ldots, x_n$ ovat tunnettuja vakioita ja virheet ϵ_i ovat riippumattomia satunnaismuuttujia, jotka 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 $\sigma^2 > 0$.

Toinen muotoilu lineaariselle mallille

- Lineaarinen malli

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

  voidaan muotoilla toisella tavalla.
- Voidaan sanoa yhtäpitävästi, että satunnaismuuttujat Y_i ovat riippumattomia siten, että muuttujan Y_i jakauma on

$$Y_i \sim N(\alpha + \beta x_i, \sigma^2), \qquad i = 1, \ldots, n. \tag{10}$$

- Tämän huomion jälkeen uskottavuusfunktio on helppo kirjoittaa.

Lineaarisen mallin uskottavuusfunktio

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

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

Logaritminen uskottavuusfunktio on

$$\ell(\alpha, \beta, \sigma^2) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \sum_{i=1}^{n} (y_i - \alpha - \beta\, x_i)^2 \qquad (11)$$
$$= -\frac{n}{2} \log(\sigma^2) - \frac{1}{2\,\sigma^2} \mathrm{SS}(\alpha, \beta). \qquad (12)$$

Logaritmisen uskottavuusfunktion analysointia

- Tulos oli

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

- Oli virhevarianssin $\sigma^2 > 0$ arvo mikä tahansa, niin funktion

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

maksimipiste on sama kuin funktion $\operatorname{SS}(\alpha, \beta)$ minimipiste, joka on (a, b).

- 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, ja se on

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

Lineaarisen mallin SU-estimaattien käsittelyä

- Kertoimien SU-estimaatit ovat PNS-suoran kertoimet a ja b.
- Virhevarianssin SU-estimaatiksi saatiin

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

mutta tätä kaavaa ei käytetä. Syy selviää, kun analysoidaan vastaavan estimaattorin otantajakaumaa: estimaattori on harhainen.

- Virhevarianssin σ^2 estimaattina käytetään estimaattia

$$s^2 = \frac{1}{n-2}\,\mathrm{SS}(a,b) = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - a - b\,x_i)^2. \qquad (13)$$

Jakajana on otoskoon n sekä mallin kertoimien lukumäärän 2 erotus.

Lineaarisen mallin SU-estimaattorien otantajakauma

- Merkitään estimaatteja a, b ja s^2 vastaavia estimaattoreita symboleilla $a(\mathbf{Y})$, $b(\mathbf{Y})$ ja $s^2(\mathbf{Y})$.
- Kertoimien PNS-estimaattorit ovat harhattomia ja normaalijakautuneita.
- Kulmakertoimen estimaattorilla on otantajakauma

$$b(\mathbf{Y}) \sim N(\beta, \frac{\sigma^2}{q_{xx}}), \quad \text{jossa} \quad q_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \qquad (14)$$

- Tämän takia PNS-kulmakertoimen b keskivirhe lasketaan kaavalla

$$\mathrm{se}(b) = \frac{s}{\sqrt{q_{xx}}}. \qquad (15)$$

Estimaattorien otantajakauma (jatkoa)

Virhevarianssin estimaattorin $s^2(\mathbf{Y})$ otantajakauma:

- $s^2(\mathbf{Y})$ on harhaton
- $s^2(\mathbf{Y}) \perp\!\!\!\perp a(\mathbf{Y})$ ja $s^2(\mathbf{Y}) \perp\!\!\!\perp b(\mathbf{Y})$.
- Sopivasti skaalattuna $s^2(\mathbf{Y})$ noudattaa khiin neliön jakaumaa vapausasteluvulla $n-2$:

$$\frac{n-2}{\sigma^2}\, s^2(\mathbf{Y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (Y_i - a(\mathbf{Y}) - b(\mathbf{Y})\, x_i)^2 \sim \chi^2_{n-2}. \quad (16)$$

Taas päästään t-jakaumaan

- Käytetään hyväksi tuloksia

$$b(\mathbf{Y}) \sim N(\beta, \frac{\sigma^2}{q_{xx}}), \qquad \frac{n-2}{\sigma^2} s^2(\mathbf{Y}) \sim \chi^2_{n-2},$$

  jossa lisäksi estimaattorit ovat riippumattomia.

- Tällöin nähdään että

$$\frac{b(\mathbf{Y}) - \beta}{s(\mathbf{Y})/\sqrt{q_{xx}}} \sim t_{n-2}. \tag{17}$$

- Tämän perusteella mallin kulmakertoimelle voidaan laskea luottamusvälejä sekä voidaan testata kulmakerrointa koskevia hypoteeseja.

Kulmakertoimen luottamusväli ja sen testaus

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

$$b - t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}} \leq \beta \leq b + t_{n-2}(\alpha/2)\frac{s}{\sqrt{q_{xx}}}. \qquad (18)$$

- On mielekästä testata hypoteesiparia

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

sillä jos $\beta = 0$, niin tällöin mallissa ei tarvita selittäjää x.

- Tässä testissä aineistosta lasketaan testisuure

$$t = \frac{b-0}{s/\sqrt{q_{xx}}}. \qquad (19)$$

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

Kaksisuuntainen testi $\beta = 0$ kuminauha-aineistolle

```
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
```

Tulostuksesta poimittuja asioita

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

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

- Nollahypoteesi hylätään kaikilla tavanomaisilla luottamustasoilla.
- Toisin sanoen venytys on tarpeellinen selittäjä tässä lineaarisessa mallissa.

Kulmakertoimen luottamusväli kuminauha-aineistolle

```
confint(m)
```

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

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

Lineaarinen regressio, kun selittäjät ovat satunnaismuuttujia

- Edellä esitetty lineaarinen malli, jossa selittäjiä x_i pidetään vakioina on mielekäs lähinnä silloin, kun selittäjän arvot voidaan kokeessa itse asettaa.
- 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.

Lineaarinen regressio, kun selittäjät ovat satunnaisia

- Jos oletetaan, että satunnaismuuttujat

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

muodostavat satunnaisotoksen jostakin kaksiulotteisesta normaalijakaumasta, niin tällöin edellisen jakson päättelyn kannalta keskeiset tulokset kuten kaava (16) ja (17) pitävät paikkansa.

- Tämä väite perustellaan 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.

Satunnaiset selittäjät (jatkoa)

- Vektorin $(Y_1, \ldots, Y_n)$ ehdollinen yhteisjakauma ehdolla $\mathbf{X} = \mathbf{x}$ täyttää lineaarisen mallin oletukset.
- Tämän takia kaikki edellisessä jaksossa mainitut jakaumatulokset pätevät ehdolla $\mathbf{X} = \mathbf{x}$.
- Koska varianssiestimaattorin otantajakauma (16) tai t-saranasuureen (17) otantajakauma ei riipu arvosta $\mathbf{x}$, niin tästä seuraa, että kaavojen (16) ja (17) jakaumatulokset pätevät myös ei-ehdollisesti.

9.5 Muita lineaarisia malleja

- Lineaarinen malli (9) voidaan kirjoittaa matriisimerkinnöillä kaavalla

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{20}$$

- Tässä asetelmamatriisi (tai mallimatriisi) $\mathbf{X}$ on kiinteä ja tunnettu, ja kerroinvektori $\boldsymbol{\beta}$ kiinteä mutta tuntematon.
- Mallissa (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 $N(0, \sigma^2)$.
- Sekä kerroinvektori että virhevarianssi ovat tuntemattomia parametreja.

Olemme käsitelleet erikoistapauksia yleisestä mallista aikaisemmin

Usea tällä kurssilla käsitelty malli voidaan kirjoittaa lineaarisena mallina

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

Näitä ovat

- tämän luvun lineaarinen regressiomalli;
- 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.

Yleinen lineaarinen malli

- Lineaaristen mallien teoria antaa yhtenäisen teorian kaikille niille malleille, jotka voidaan ilmaista kaavalla

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

- Aiheeseen voi perehtyä esim. lineaaristen mallien kurssilla.