Lineaarisen mallin määrittely

Yksinkertainen esimerkki

- Peltotilkut $i = 1, ..., n$
- Lannoitemäärät $x_1, ..., x_n$
- Satomäärät $y_1, ..., y_n$
- Mikä on lannoituksen vaikutus odotettavissa olevaan satomäärään?
- Oletetaan satomääriä vastaavista sm:sta
  - $Y_1, ..., Y_n$ $\underline{\parallel}$ eli riippumattomuus
  - $\mu_i := \mathrm{E}(Y_i) = \beta_1 + \beta_2 x_i$ odotusarvon lineaarisuus
  - $\mathrm{Var}(Y_i) = \sigma^2 \ \forall i$ varianssin koko ei riipu lannoitemäärästä
- Tilanne voidaan kuvata yhtälöllä

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \ \ i = 1, ..., n,$$

jossa sm:t $\varepsilon_1, ..., \varepsilon_n$ ovat ei-havaittavia, $\underline{\parallel}$ ja toteuttavat

$$\mathrm{E}(\varepsilon_i) = 0, \ \mathrm{Var}(\varepsilon_i) = \sigma^2 \ \forall i$$

Lineaarisen mallin määrittely

Yksinkertainen esimerkki

- Kuvattaessa lannoituksen vaikutusta odotettavissa olevaan satomäärään päädyttiin siis yhtälöön

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \;\; i = 1, ..., n,$$

jossa sm:t $\varepsilon_1, ..., \varepsilon_n$ ovat ei-havaittavia, $\underline{\parallel}$ ja toteuttavat

$$\mathrm{E}(\varepsilon_i) = 0, \;\; \mathrm{Var}(\varepsilon_i) = \sigma^2 \;\; \forall i$$

- Sm:t $\varepsilon_1, ..., \varepsilon_n$ voidaan tulkita havaituissa satomäärissä ilmeneväksi 'puhtaaksi satunnaisvaihteluksi' tai selitysvirheeksi, joka ei selity lannoitemäärällä.
- Edellä esitettyä yhtälöä kutsutaan *yhden selittävän muuttujan lineaariseksi regressiomalliksi*.
- Tiukasti tilastollisen päättelyn näkökulmasta, ei kysymyksessä ole vielä kuitenkaan tilastollinen malli, jollainen vaatii havaintojen yhteistodennäköisyysjakauman ja parametriavaruuden spesifioinnin.

Lineaarisen mallin määrittely

Yksinkertainen esimerkki

Klassinen lineaarinen malli perustuu olettaa normaalijakaumaan ja tilastollinen malli saadaan olettamalla

$$Y_1, ..., Y_n \underline{\parallel} \quad Y_i \sim \mathrm{N}\left(\beta_1 + \beta_2 x_i, \sigma^2\right), \quad \beta_1, \beta_2 \in \mathbb{R}, \quad \sigma^2 > 0$$

tai yhtäpitävästi

$$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i, \ i = 1, ..., n, \quad \beta_1, \beta_2 \in \mathbb{R}, \quad \sigma^2 > 0$$

$$\varepsilon_1, ..., \varepsilon_n \underline{\parallel}, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$$

Tarkasteltavassa esimerkkitapauksessa oletus $\beta_1 > 0$ tuntuisi järkevältä, mutta johtaa mutkikkaampaa malliin, joka ei sisälly klassiseen lineaariseen malliin.

Lineaarisen mallin määrittely

Yksinkertainen esimerkki

- Yhden selittävän muuttujan lineaarinen regressiomallissa

$$Y_1, ..., Y_n \;\underline{\parallel}\; Y_i \sim \mathrm{N}\left(\beta_1 + \beta_2 x_i, \sigma^2\right), \quad \beta_1, \beta_2 \in \mathbb{R}, \quad \sigma^2 > 0.$$

  selittävien muuttujien havaintoja $x_1, ..., x_n$ ei tulkita satunnaisiksi (loogista tarkasteltavassa tapauksessa).

- Jos selittävät muuttujat olisivat satunnaisia, täytyisi tilastollinen malli periaatteessa laajentaa koskemaan sv:ta $\left[Y_i \;\; X_i\right]'$, $i = 1, ..., n$.

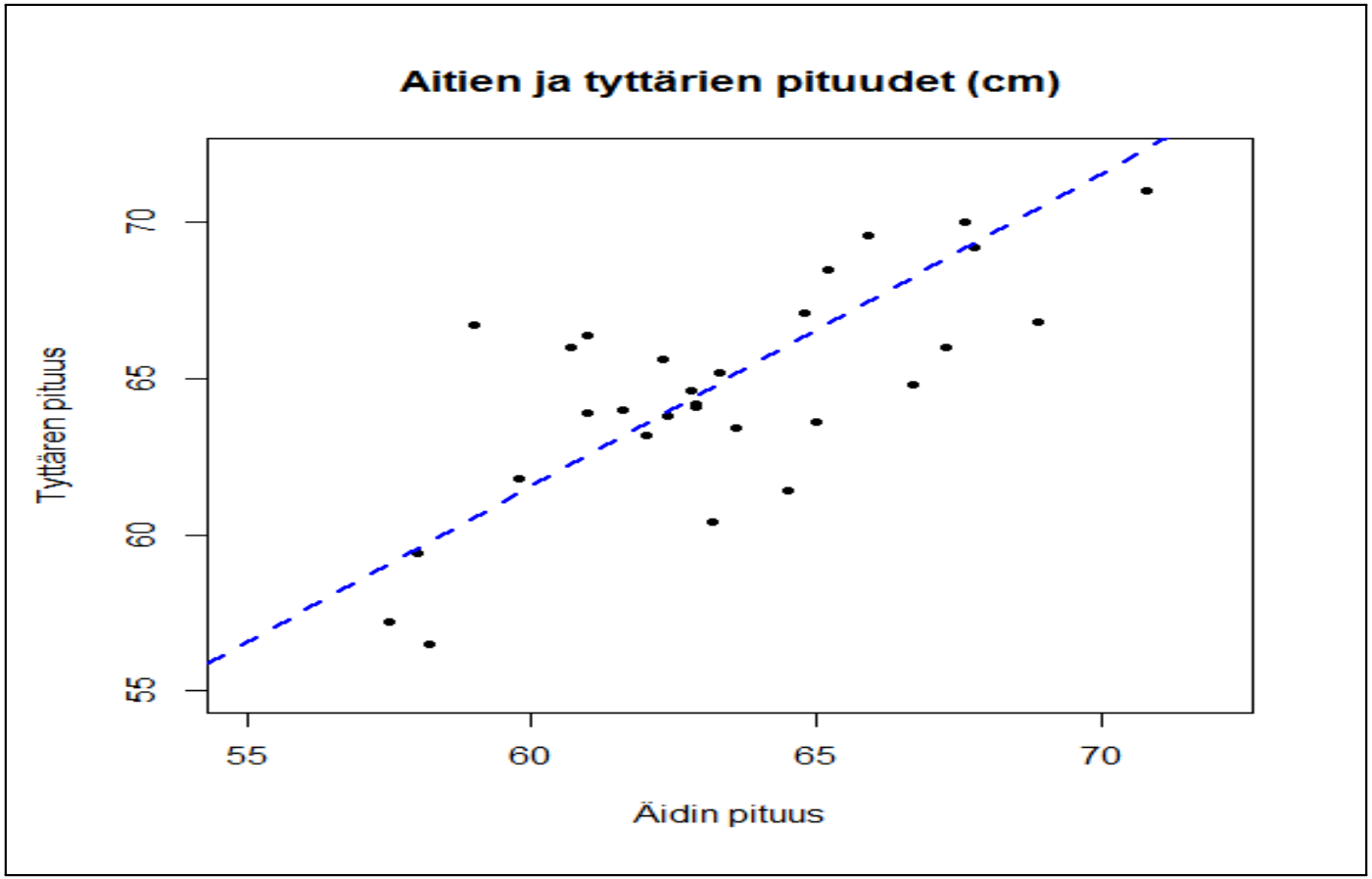
- Usein selitettäviä muuttujia $Y_1, ..., Y_n$ voidaan tarkastella ehdollisesti ehdolla $(X_1, ..., X_n) = (x_1, ..., x_n)$, jolloin edellä esitetty malli soveltuu.

- Usein tarvitaan useita selittäviä muuttujia

Lineaarisen mallin määrittely

Yksinkertainen esimerkki

Äitien ja tyttärien pituudet (tuumissa): n=30, katkoviiva 45 asteen suora



Yleinen lineaarinen malli

Tausta: Aineiston muuttujista yksi on luonteeltaan *selitettävä* ja loput sen vaihtelua *selittäviä* muuttujia eli kaaviona

Havainto yksikkö	Selitettävä m:t; y	Selittävät m:t; $x_1, \ldots, x_p$
1	y_1	$x_{11}, \ldots, x_{1p}$
$\vdots$	$\vdots$	$\vdots$
n	y_n	$x_{n1}, \ldots, x_{np}$

Lineaarisessa mallissa selittävän muuttujan vaikutus selitettävään muuttujaan oletetaan (tietyssä mielessä) *lineaariseksi*.

Yleinen lineaarinen malli

- Lineaarisen mallin määritelmä voidaan perustaa yhtälöön

$$Y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n,$$

  jossa Y_i on havaittava sm, $x_{i1}, ..., x_{ip}$ ovat havaittavia ja ei-satunnaisia (eli kiinteitä), ε_i on ei-havaittava sm ja $\beta_1, ..., \beta_p$ ovat tuntemattomia parametreja.
- ε_i on ns. *virhe(termi)* eli se osa Y_i:stä, jota mallin systemaattinen osa tai rakenne $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ ei kykyne selittämään
- Lineaarisuus oletetaan parametrien $\beta_1, ..., \beta_p$ suhteen ja virhetermi lisätään systemaattiseen osaan additiivisesti.
  - Esimerkiksi

$$Y_i = \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, ..., n,$$

    on vaaditulla tavalla lineaarinen

Yleinen lineaarinen malli

- Lineaariseen malliin voidaan päätyä lähtemällä oletuksesta

$$Y_1, ..., Y_n \perp\!\!\!\perp , \; Y_i \sim \mathrm{N}\left(\mu_i, \sigma^2\right)$$

ja olettamalla odotusarvoille μ_i (esim. taustatiedon perusteella) rakenne $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$, jossa $\mathbf{x}_i' = [x_{i1} \; \cdots \; x_{ip}]$ ja $\boldsymbol{\beta} = [\beta_1 \; \cdots \; \beta_p]'$.

- Miksi yhtälö

$$Y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i, \quad (i = 1, ..., n)$$

ja virhetermi ε_i?

- Yksi syy on, että tämä on toisinaan luonteva muotoilu.
- Toinen, että käsitteet *residuaali* $\hat{\boldsymbol{\varepsilon}}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ ja *sovite* $\hat{\mu}_i = \mathbf{x}_i'\hat{\boldsymbol{\beta}} = \hat{y}_i$ ovat hyödyllisiä ($\hat{\boldsymbol{\beta}}$ on parametrin $\boldsymbol{\beta}$ estimaatti).
- Residuaali on teoreettisen virheen $\varepsilon_i = Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ empiirinen vastine ja sovite on mallin systemaattisen osan $\mu_i = \mathbf{x}_i'\boldsymbol{\beta}$ empiirinen vastine.

Yleinen lineaarinen malli

- Residuaalit $\hat{\varepsilon}_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ sisältävät ilmeisestikin informaatiota virheiden $\varepsilon_i = Y_i - \mathbf{x}_i'\boldsymbol{\beta}$ varianssista eli parametrista $\sigma^2 = \mathrm{Var}(\varepsilon_i) = \mathrm{Var}(Y_i)$.
- Residuaalien avulla voidaan myös tutkia mallin oletusten paikkansapitävyyttä yleensä helpommin kuin alkuperäisteen havaintojen $y_1, ..., y_n$ avulla.
- Tutkittavia oletuksia:
  - Onko $\mathrm{Var}(Y_i) = \mathrm{Var}(\varepsilon_i)$ vakio kuten oletetaan?
  - Päteekö oletettu riippumattomuus $Y_1, ..., Y_n \perp\!\!\!\perp \Leftrightarrow \varepsilon_1, ..., \varepsilon_n \perp\!\!\!\perp$?
  - Päteekö oletettu normaalisuus $Y_i \sim \mathrm{N}(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2) \Leftrightarrow \varepsilon_i \sim \mathrm{N}(0, \sigma^2)$?

Yleinen lineaarinen malli

Lineaarisen mallin erikoistapauksia

- Riippumaton otos normaalijakaumasta (vrt. tilastollinen päättely).

$$Y_1, ..., Y_n \quad \underline{\parallel} \quad , \quad Y_i \sim \mathrm{N}\left(\mu, \sigma^2\right)$$

-

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

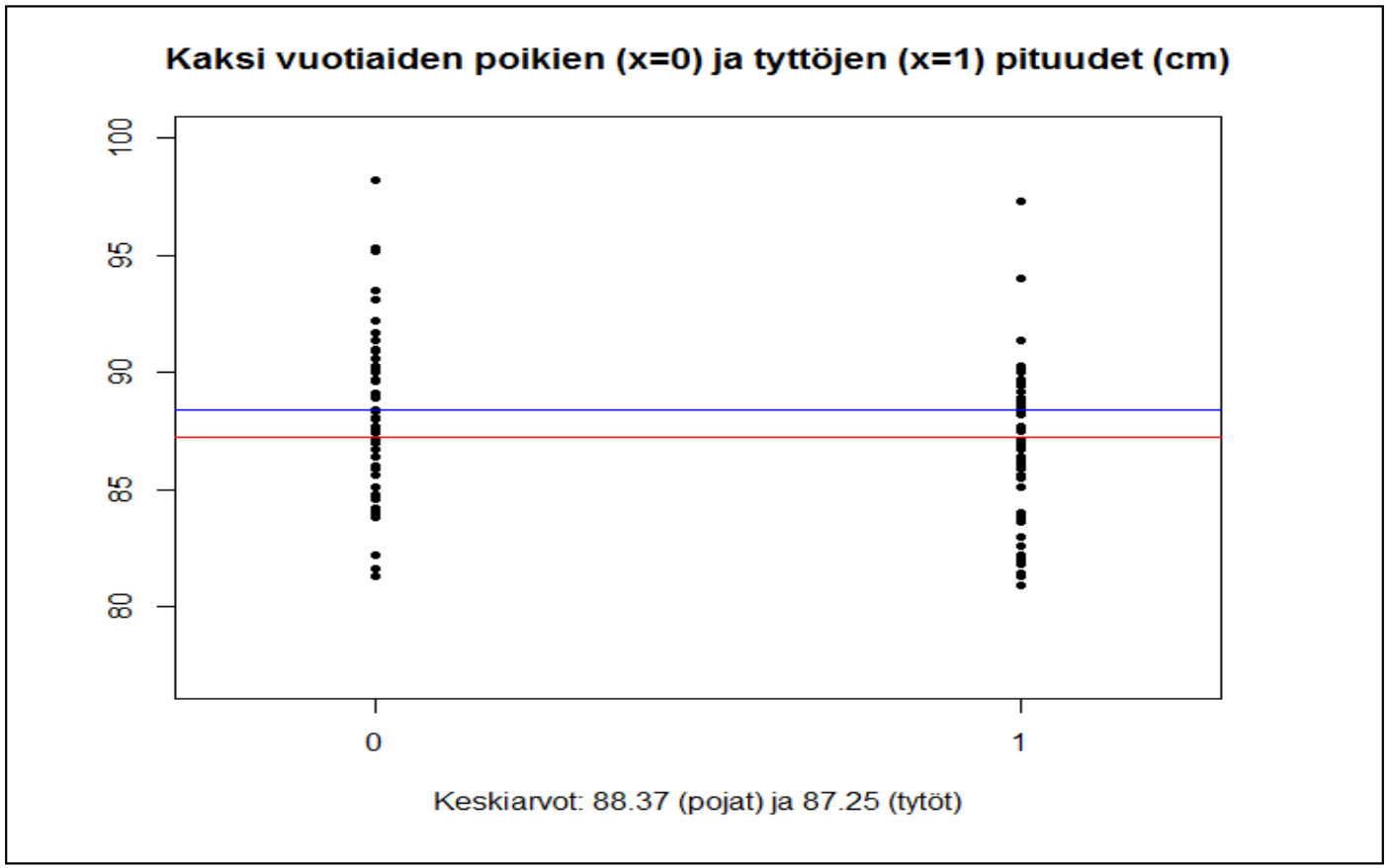
eli

$$\mathbf{Y} = \mathbf{1}_n \mu + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right) \quad \Leftrightarrow \quad \mathbf{Y} \sim \mathrm{N}\left(\mathbf{1}_n \mu, \sigma^2 \mathbf{I}_n\right)$$

Matriiseiksi $\mathbf{X}$ tulee siten vektori $\mathbf{1}_n = \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}' \quad (n \times 1)$.

- Kiinnostava hypoteesi $H_0 : \mu = \mu_0$

Kaksivuotiaiden poikien (x=0) ja tyttöjen (x=1) pituudet (cm)
$n_1 = 66, \quad n_2 = 70$



- Kahden odotusarvoltaan (mahdollisesti) poikkeavan riippumattoman normaalisen otoksen malli eli

$$Y_1, ..., Y_n \quad \underline{\parallel} \quad , \quad Y_i \sim \begin{cases} \mathrm{N}\left(\mu_1, \sigma^2\right), & \text{kun } i = 1, ..., n_1 \\ \mathrm{N}\left(\mu_2, \sigma^2\right), & \text{kun } i = n_1 + 1, ..., n_1 + n_2 = n. \end{cases}$$

-

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

eli

$$\mathbf{Y} = \begin{bmatrix} \mathbf{1}_{n_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{n_2} \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}_n\right); \quad H_0 : \mu_1 = \mu_2$$

Yleinen lineaarinen malli

Lineaarisen mallin erikoistapauksia

- *Yksisuuntainen varianssianalyysimalli*. Havaintoina on p riippumatonta otosta jakaumista $\mathrm{N}(\mu_j, \sigma^2)$ $(j = 1, ..., p)$.
- Esimerkiksi vehnälajikkeiden $A_1, ..., A_p$ satoisuuden tutkiminen, kun niitä viljellään samoissa olosuhteissa.
- Kiinnostuksen kohteena on odotusarvoissa $\mu_1, .., \mu_p$ mahdollisesti ilmenevät erot; esim. $H_0 : \mu_1 = \cdots = \mu_p$
- *Kaksisuuntainen varianssianalyysimalli*. Esimerkiksi vehnälajikkeita A_1 ja A_2 lannoitetaan kahta eri lannoitetta B_1 ja B_2 käyttäen, jolloin havainnot peräisin neljästä ryhmästä.
- Mallin avulla voidaan tutkia onko satomäärissä eroja eri ryhmien välillä ja johtuvatko mahdolliset erot vehnälajikkeesta, lannoitteesta vai niiden yhteisvaikutuksesta.
- Näissä malleissa selittävät muuttujat ovat ryhmää osoittavia indikaattoreita.

- *Usean selittäjän lineaarinen regressiomalli*

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, ..., n, \quad \boldsymbol{\beta} \in \mathbb{R}^p, \sigma^2 > 0.$$

β_1 ns. *(regressio)vakio* ja β_j $(j = 2, .., p)$ ns. *regressiokerroin*, joka kuvaa paljonko selitettävä muuttuja (tai sen odotusarvo) muuttuu, kun j:nnen selittäjän arvo muuttuu yhden yksikön ja muiden selittäjien arvot pysyvät ennallaan.

- Saadaan yleisestä mallista valitsemalla $x_{i1} = 1$, joten

$$\mathbf{X} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1p} \\ & & \vdots & \\ 1 & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

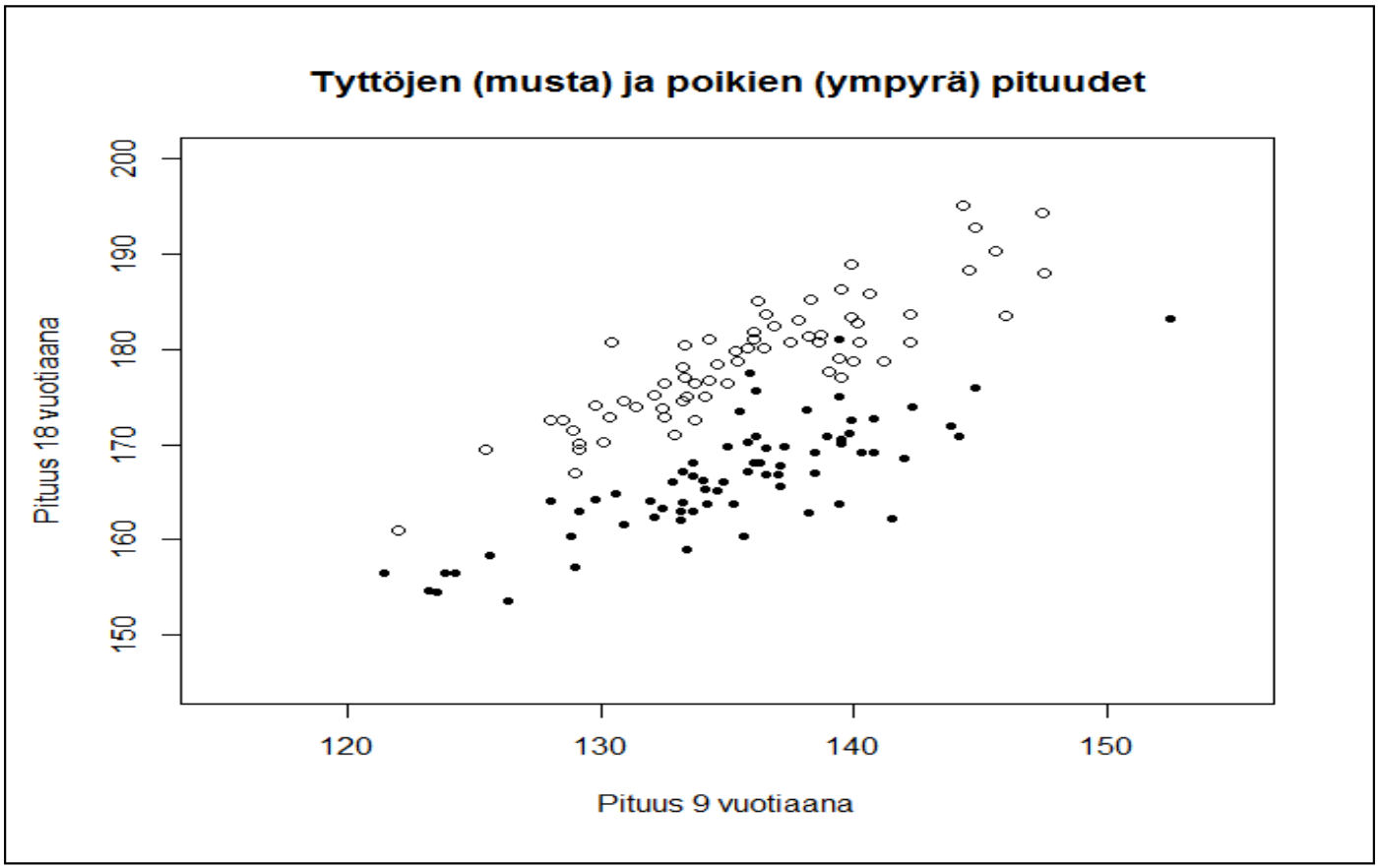
- Samassa mallissa voi olla sekä kvantitatiivisia selittäjiä että kvalitatiivisia ryhmää osoittavia indikaattoreita.
- Tutkitaan esimerkiksi vehnälajikkeiden A_1 ja A_2 satoisuutta, kun molempia lannoitetaan samaa lannoitetta käyttäen:

$$Y_1, ..., Y_n \quad \underline{\perp\!\!\!\perp} \quad , \quad Y_i \sim \begin{cases} \mathrm{N}\left(\beta_1 + \beta_3 x_i, \sigma^2\right), & \text{kun } i = 1, ..., n_1 \\ \mathrm{N}\left(\beta_2 + \beta_3 x_i, \sigma^2\right), & \text{kun } i = n_1 + 1, ..., n_1 + n_2 \end{cases}$$

eli

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_{n_1} \\ Y_{n_1+1} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & x_{n_1} \\ 0 & 1 & x_{n_1+1} \\ \vdots & \vdots & \vdots \\ 0 & 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{n_1} \\ \varepsilon_{n_1+1} \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Tyttöjen (musta) ja poikien (ympyrä) pituudet 9 ja 18 v. ikäisinä (cm)
$n_1 = 66, \quad n_2 = 70$



- Joskus epälineaarinen malli voi olla *linearisoituva*:

$$Y_i = e^{\beta_1} x_{i2}^{\beta_2} \cdots x_{ip}^{\beta_p} e^{\varepsilon_i}, \quad i = 1, \ldots, n,$$

jossa muuttujat oletetaan positiivisiksi.

- Ottamalla (luonnollinen) logaritmi puolittain päädytään yhtälöön

$$\log Y_i = \beta_1 + \beta_2 \log x_{i2} + \cdots + \beta_p \log x_{ip} + \varepsilon_i, \quad i = 1, \ldots, n,$$

eli (merkintöjä vaille) usean selittäjän lineaarinen regressiomalli.

- Tällaiset ns. multiplikatiiviset mallit ovat tavallisia taloudellisissa sovelluksissa mallinnettaessa esimerkiksi jonkin tuotteen kysyntää.

Yleinen lineaarinen malli

- Lineaarisen mallin määritelmä voidaan perustaa yhtälöön

$$\begin{aligned} Y_i &= \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i \\ &= \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, ..., n, \end{aligned}$$

  jossa Y_i on havaittava sm, $\mathbf{x}_i = [x_{i1} \ \cdots \ x_{ip}]'$ on havaittava ja ei-satunnainen (eli kiinteitä), ε_i on ei-havaittava sm ja $\beta_1, ..., \beta_p$ ovat tuntemattomia parametreja.

- ε_i on ns. *virhe(termi)* eli se osa Y_i stä, jota mallin systemaattinen osa tai rakenne $\beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ ei kykyne selittämään.

- Tilastollista mallia varten oletetaan

$$\varepsilon_1, ..., \varepsilon_n \perp\!\!\!\perp, \quad \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right) \Leftrightarrow Y_1, ..., Y_n \perp\!\!\!\perp Y_i \sim \mathrm{N}\left(\mathbf{x}_i' \boldsymbol{\beta}, \sigma^2\right)$$

Yleinen lineaarinen malli

- Mallin matriisiesitys

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ & \vdots & \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

eli

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

jossa $\varepsilon_1, ..., \varepsilon_n \underline{\parallel}$, $\varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right) \Leftrightarrow \boldsymbol{\varepsilon} \sim \mathrm{N}\left(0, \sigma^2 \mathbf{I}_n\right)$.

- Nyt malli voidaan määritellä lyhyesti kirjoittamalla

$$\mathbf{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0.$$

- Lisäoletus: Matriisi $\mathbf{X}$ $(n \times p)$ täyttä sarakeastetta eli $\mathrm{r}\left(\mathbf{X}\right) = p$.

Yleinen lineaarinen malli

- Lineaarinen malli voidaan siis määritellä kahdella yhtäpitävällä tavalla:

$$(1) \qquad \mathbf{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0.$$

tai

$$(2) \qquad Y_1, ..., Y_n \underline{\parallel} Y_i \sim \mathrm{N}\left(\mathbf{x}_i'\boldsymbol{\beta}, \sigma^2\right), \quad \boldsymbol{\beta} \in \mathbb{R}^p, \ \sigma^2 > 0$$

- Uskottavuusfunktio voidaan johtaa kumpaa tahansa käyttäen.
- Lisäoletus, että matriisi $\mathbf{X}$ $(n \times p)$ on täyttä sarakeastetta eli $\mathrm{r}(\mathbf{X}) = p$ ei vaikuta uskottavuusfunktion johtoon.