SU-estimointi

- Havaintovektori $\mathbf{Y} = (y_1, ..., y_n)$ toteuttaa $\mathbf{Y} \sim \mathrm{N}\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n\right)$.
- MN-jakauman tiheysfunktion kaavasta saadaan

$$f_{\mathbf{Y}}\left(\mathbf{y}; \boldsymbol{\beta}, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)\right\}$$

- Siis, parametrien $\boldsymbol{\beta}$ ja σ^2 log-uskottavuusfunktio on

$$l\left(\boldsymbol{\beta}, \sigma^2; \mathbf{y}\right) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathrm{S}\left(\boldsymbol{\beta}\right),$$

  jossa

$$\mathrm{S}\left(\boldsymbol{\beta}\right) = \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)'\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right) = \sum_{i=1}^{n}\left(y_i - \mathbf{x}_i'\boldsymbol{\beta}\right)^2$$

  on ns. *jäännösneliösummafunktio* ($\mathbf{x}_i'$ matriisin $\mathbf{X}$ i. rivi).
- SU-estimaatit $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}\left(\mathbf{y}\right)$ ja $\hat{\sigma}^2 = \hat{\sigma}^2\left(\mathbf{y}\right)$:

$$S(\hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\beta}} S(\boldsymbol{\beta}) \qquad \text{ja} \qquad \hat{\sigma}^2 = \frac{1}{n}S(\hat{\boldsymbol{\beta}})$$

SU-estimointi

- SU-estimaatit $\hat{\beta} = \hat{\beta}(\mathbf{y})$ ja $\hat{\sigma}^2 = \hat{\sigma}^2(\mathbf{y})$:

$$S(\hat{\beta}) = \min_{\beta} S(\beta) \qquad \text{ja} \qquad \hat{\sigma}^2 = \frac{1}{n} S(\hat{\beta})$$

- Minimointitehtävän ratkaisu johtaa *normaaliyhtälöihin*

$$\mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{y}.$$

- Oletuksesta $\mathrm{r}(\mathbf{X}) = p$ seuraa $\mathrm{r}(\mathbf{X}'\mathbf{X}) = p$ ja siten yksikäsitteinen ratkaisu (*pienimmän neliösumman (PNS) estimaatti*)

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

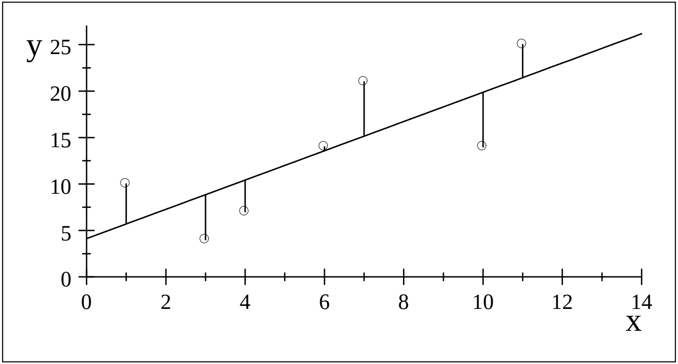
- Tämän jälkeen

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\hat{\beta})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

SU-estimointi

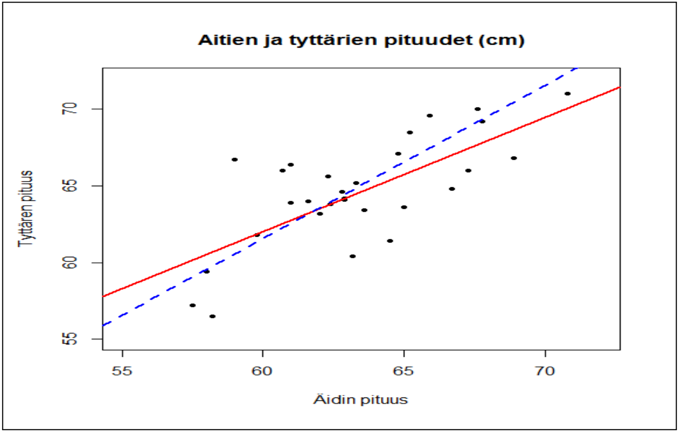
Malli $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i,\ \ i = 1, \ldots, n,\ \varepsilon_1, \ldots, \varepsilon_n \perp\!\!\!\perp,\ \ \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$

PNS-estimaatti $\hat{\boldsymbol{\beta}} = \left(\hat{\beta}_1, \hat{\beta}_2\right)$ minimoi funktion $\sum_{i=1}^{n} \left(y_i - \beta_1 - \beta_2 x_i\right)^2$ eli kuvan pystysuorien janojen pituuksien neliösumman.



SU-estimointi

Äitien ja tyttärien pituudet (tuumissa): n=30, punainen PNS-suora $y = 17.29 + 0.75x$, sininen katkoviiva $y = 1.53 + x$



- Tarkastellaan mallia

$$Y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_1, \ldots, \varepsilon_n \perp\!\!\!\perp, \; \varepsilon_i \sim \mathrm{N}\left(0, \sigma^2\right).$$

- PNS-sovite ja -residuaali ovat $\hat{y}_i = y_i - \hat{\beta}_1 - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}$ ja $\hat{\varepsilon}_i = y_i - \hat{y}_i$.
- Koska matriisin $\mathbf{X}$ ensimmäinen sarake on nyt ykkösvektori $\mathbf{1}_n$, seuraa normaaliyhtälöistä $\mathbf{X}'\hat{\boldsymbol{\varepsilon}} = \mathbf{0}$ ja edelleen $\mathbf{1}_n'\hat{\boldsymbol{\varepsilon}} = \hat{\varepsilon}_1 + \cdots + \hat{\varepsilon}_n = 0$ ja

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i + \hat{\varepsilon}_i) = \frac{1}{n}\sum_{i=1}^{n}\hat{y}_i.$$

- Määritellään käsitteet:
  - Kokonaisneliösumma $\mathrm{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - Regressioneliösumma $\mathrm{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - Residuaalineliösumma $\mathrm{SSE} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$

- Määritellään käsitteet:
  - Kokonaisneliösumma $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
  - Regressioneliösumma $\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
  - Residuaalineliösumma $\text{SSE} = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$
- *Kun mallissa on vakio*, näiden välillä on yhteys

$$\text{SST} = \text{SSR} + \text{SSE}.$$

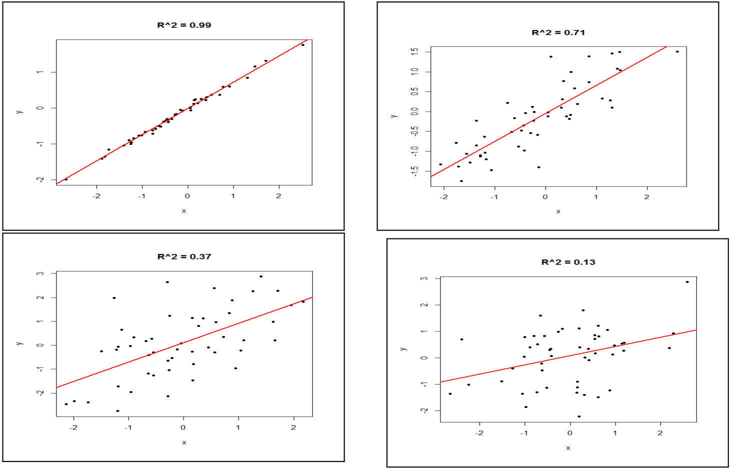
- Mallin *selitysaste* on

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = \frac{\text{SSR}}{\text{SST}},$$

  jossa jälkimmäinen yhtälö perustuu edellä todettuun identiteettiin.
- Selitysasteella ominaisuus

$$0 \leq R^2 \leq 1.$$

Simuloituihin aineistoihin sovitettuja PNS-suoria eri selitysastein, $n = 50$



- Selitysasteelle pätee

$$R^2 = r^2_{y\hat{y}},$$

jossa $r_{y\hat{y}}$ on selitettävän muuttujan havaintojen y_i ja sovitteiden $\hat{y}_i$ $(1, ..., n)$ välinen otoskorrelaatiokerroin.

- Koska

$$R = \sqrt{1 - \text{ SSE/SST}} \quad \text{ja} \quad SSE = \min_{\beta} \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right)' \left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right),$$

valitsee PNS-menetelmä sovitteeksi sen selittäjien lineaarikombinaation, jonka otoskorrelaatio selitettävän muuttujan kanssa maksimoituu (huomaa tulkinnan laskennallinen luonne).