JYVÄSKYLÄN YLIOPISTO
Matematiikan ja tilastotieteen laitos
Johdatus tilastolliseen mallintamiseen

Harjoitukset 2
20.9.2011

1. Aineisto `airpol`. Aineisto sisältää tietoja 50 USA:n suurimmasta kaupungista:

   `city` = kaupungin nimi,

   `tmr` = kuolevuus 100000 asukasta kohti,

   `smean` = sulfaattipitoisuus, 10 mikrogrammaa kuutiometrissä,

   `pmean` = hiukkaspitoisuus, 10 mikrogrammaa kuutiometrissä,

   `perwh` = valkoisten osuus (%),

   `nonpoor` = köyhyysrajan yläpuolisten osuus (%),

   `ge65` = yli 65 vuotiaiden osuus 1000 asukasta kohti.

   a) Aja regressio, jossa kuolevuutta selitetään vain ilman saasteita mittaavilla muuttujilla `smean` ja `pmean`. b) Aja regressio, jossa kuolevuutta selitetään vain kaikilla muilla numeerisilla muuttujilla. c) Miksi kertoimien tulkinta käytännön kannalta on jälkimmäisessä analyysissä on järkevämpi?

2. Osoita, että a) $\widehat{\mathbf{Y}}'\widehat{\mathbf{e}} = \mathbf{0}$ ja b) $\widehat{\mathbf{e}}'\mathbf{X} = \mathbf{0}$

3. Osoita, että luentojen $\widetilde{\sigma}^2$ antaa globaalin maksimin funktiolle

$$\log L(\widehat{\boldsymbol{\beta}}, \sigma^2) = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log \sigma^2 - \frac{S(\widehat{\boldsymbol{\beta}})}{2\sigma^2}.$$

4. Osoita, että yhden selittäjän mallissa

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ \widehat{\beta}_0 &= \bar{y} - \widehat{\beta}_1 \bar{x}. \end{aligned}$$

5. Jatkoa edelliseen. a) Osoita, että regressiokerroin voidaan kirjoittaa myös muotoon

$$\widehat{\beta}_1 = r\frac{s_y}{s_x},$$

   missä r on otoskorrelaatiokerroin sekä s_x ja s_y ovat vastaavasti x- ja y-havaintojen otoskeskihajonnat. b) Päättele, että jos regressiokerroin on itseisarvoltaa yli ykkösen, niin $s_y > s_x$. c) Päättele myös, että jos vaste ja selittäjä standardoidaan, niin regressiokerroin on sama kuin korrelaatiokerroin ("regressio kohti keskiarvoa").

6. Oletetaan, että $x \sim N(175, 8.3^2)$ on isän pituus ja $y \sim N(175 + 0.8 \times (x - 175), 8.3^2 \times (1 - 0.8^2))$ on pojan pituus. Generoi 1000 kpl isä-poika -pareja (funktio `rnorm()`). a) Sovita regressiomalli

$$y = \beta_0 + \beta_1 x + \varepsilon$$

koko aineistoon. b) Sovita regressiomalli aineistoon, jossa isän pituus rajoitetaan välille 170–190. c) Vertaile estimoituja regressiokertoimia, jäännöskeskihajontoja ja selitysasteita. Mitä huomaat?