Study sheet 2, Exercise 2

Kai Husmann

10 November 2016

In some applications of simple linear regression a model without an intercept is required (when the data are such that the line must go through the origin), that is, a model of the form

$$y_i = \beta x_i + \varepsilon_i, i = 1, ..., n$$

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

a)

Derive the least squares estimator for β. Book page 128, lecture 26.10.

$$\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta x_i)^2$$

$$\begin{aligned}
\frac{\partial \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\partial \beta} &= -2 \sum_{i=1}^{n} (y_i - \beta x_i) x_i = 0 \\
&\propto \sum_{i=1}^{n} (y_i - \beta x_i) x_i = 0 \\
&= \sum_{i=1}^{n} x_i y_i - \beta \sum_{i=1}^{n} x_i^2 \\
\beta \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} x_i y_i \\
\beta &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}
\end{aligned}$$

b)

Let $\tilde{\beta} = \frac{1}{\sum_{i=1}^{n} x_i} \sum_{i=1}^{n} y_i$, $\breve{\beta} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$ with $x_i \neq 0$ be two estimators of β.

Is $\tilde{\beta}$ unbiased? Book p. 131, lecture 26.10.

$$
\begin{aligned}
\mathrm{E}(\tilde{\beta}) &= \mathrm{E}(\frac{1}{\sum_{i=1}^{n} x_i}\sum_{i=1}^{n} y_i) \\
&= \mathrm{E}(\frac{1}{\sum_{i=1}^{n} x_i})\mathrm{E}(\sum_{i=1}^{n} y_i) \\
&= \frac{1}{\sum_{i=1}^{n} x_i}\sum_{i=1}^{n}\mathrm{E}(y_i) \\
&= \frac{\mathrm{E}(y_i)}{\sum_{i=1}^{n} x_i} \\
&= \frac{\beta\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} \\
&= \beta\frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i} \\
&= \beta
\end{aligned}
$$

Is $\breve{\beta}$ unbiased?

$$
\begin{aligned}
\mathrm{E}(\breve{\beta}) &= \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathrm{E}(y_i)}{x_i} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{x_i\beta}{x_i} \\
&= \frac{1}{n}n\beta = \beta
\end{aligned}
$$

c)

The **R** package `gamair`, `data(hubble)` contains data from the Hubble space telescope on distances and velocities of 24 galaxies (see also `?hubble`). Fit the following simple linear regression model without intercept to the data:

$$velocity = \beta_1 distance + \varepsilon$$

This is essentially what astronomers call Hubble's Law and β_1 is known as Hubble's constant. We can use the estimated value of β_1 to find an approximate value for the age of the universe.

Data structure

```
library(gamair)
data(hubble)
summary(hubble)
```

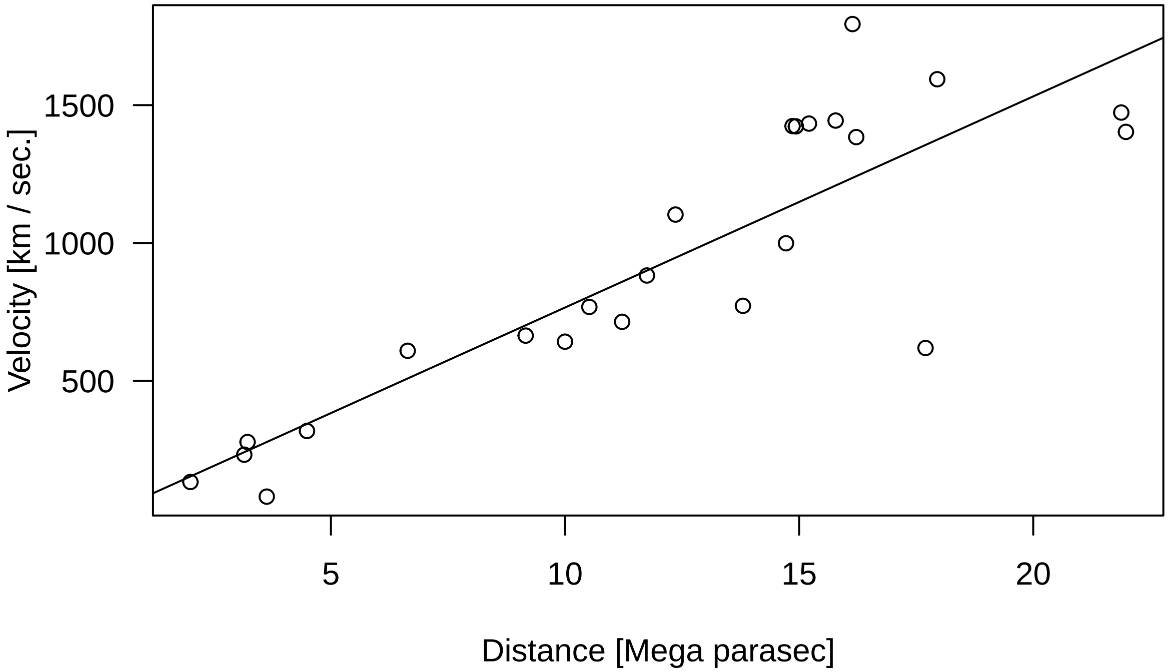
```
##       Galaxy          y                x
##  IC4182  : 1   Min.   :  80.0   Min.   : 2.00
##  NGC0300 : 1   1st Qu.: 616.5   1st Qu.: 8.53
```

```
##  NGC0925 : 1   Median : 827.0   Median :13.08
##  NGC1326A: 1   Mean   : 924.4   Mean   :12.05
##  NGC1365 : 1   3rd Qu.:1423.2   3rd Qu.:15.87
##  NGC1425 : 1   Max.   :1794.0   Max.   :21.98
##  (Other) :18
```

Linear regression

```
linreg <- lm(y ~ x -1, data = hubble)
summary(linreg)
```

```
##
## Call:
## lm(formula = y ~ x - 1, data = hubble)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -736.5 -132.5  -19.0  172.2  558.0
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   76.581      3.965   19.32 1.03e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 258.9 on 23 degrees of freedom
## Multiple R-squared:  0.9419, Adjusted R-squared:  0.9394
## F-statistic: 373.1 on 1 and 23 DF,  p-value: 1.032e-15
```



Age of the universe

The estimated coefficient for the velocity of the galaxy as a function of it´s distance β_1 is 76.58 [km/(sMpc)]. As 1 pc is 3.09e13 km, the age of the universe t_H is calculated as

$$t_H[s] = \frac{1}{\beta_1} = \frac{1}{76.68[km/sMpc]} = 4.03 * 10^{17}[s] = 12.8 \text{ billion years}$$