Homework 1

DSA1101
Introduction to Data Science

September 7, 2018

Problem 1 (10 points). Suppose we have two data vectors $x = c(x_1, x_2, ..., x_n)$ and $y = c(y_1, y_2, ..., y_n)$, both of length n. Remember in lecture that their means are given by $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ respectively.

(a) Show that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) &= \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{n}n\bar{x} \\ &= \bar{x} - \bar{x} = 0.\end{aligned}$$

(b) Show that $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$.

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}n\bar{x}^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2.\end{aligned}$$

Problem 2 (15 points). Suppose we have two data vectors $x = c(x_1, x_2, x_3) = c(1, 2.5, 4)$ and $y = c(y_1, y_2, y_3) = (0, 3, 3)$. We postulate the following simple linear relationship between y and x:

$$y \approx \beta_0 + \beta_1 x.$$

(a) Complete the following table based on the 3 data points, leaving your answers in terms of β_0 and β_1:

i	x_i	y_i	$\beta_0 + \beta_1 x_i$	residual: $e_i = y_i - (\beta_0 + \beta_1 x_i)$
1	1	0	$\beta_0 + \beta_1$	$-(\beta_0 + \beta_1)$
2	2.5	3	$\beta_0 + 2.5\beta_1$	$3 - (\beta_0 + 2.5\beta_1)$
3	4	3	$\beta_0 + 4\beta_1$	$3 - (\beta_0 + 4\beta_1)$

(b) Write down an expression for the Residual Sum of Squares, $RSS = e_1^2 + e_2^2 + e_3^2$, leaving your answer in terms of β_0 and β_1:

$$RSS = [\beta_0 + \beta_1]^2 + [3 - (\beta_0 + 2.5\beta_1)]^2 + [3 - (\beta_0 + 4\beta_1)]^2$$

(c) Based on the RSS given in (b), derive and write down the *least squares* estimates $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$\begin{aligned} \frac{\partial RSS}{\partial \beta_0} &= 2\left[\beta_0 + \beta_1\right] + 2\left[3 - (\beta_0 + 2.5\beta_1)\right](-1) + 2\left[3 - (\beta_0 + 4\beta_1)\right](-1) \\ &= -12 + 6\beta_0 + 15\beta_1 = 0. \end{aligned}$$

$$\begin{aligned} \frac{\partial RSS}{\partial \beta_1} &= 2\left[\beta_0 + \beta_1\right] + 2\left[3 - (\beta_0 + 2.5\beta_1)\right](-2.5) + 2\left[3 - (\beta_0 + 4\beta_1)\right](-4) \\ &= -39 + 15\beta_0 + 46.5\beta_1 = 0. \end{aligned}$$

$\rightarrow$

$$\begin{aligned} &-39 + 15\left(2 - \frac{15}{6}\beta_1\right) + 46.5\beta_1 = 0 \\ &-9 + 9\beta_1 = 0 \end{aligned}$$

$\rightarrow$

$$\hat{\beta}_1 = 1$$
$$\hat{\beta}_0 = -0.5$$