Homework 1

DSA1101
Introduction to Data Science

August 21, 2018

Name:

Matriculation card number:

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

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

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				
2				
3				

(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:

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