THE UNIVERSITY OF BRITISH COLUMBIA
DEPARTMENT OF STATISTICS

STAT 447B Methods for Statistical Learning (2014/15 Term 1)
Assignment 1 Solution

1. (a) Since $\mathbb{E}(\epsilon|x) = 0$, we have

$$\mathbb{E}(Y|x) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \leq \kappa_1 \\ \beta_2 + \beta_3 x, & \text{if } x > \kappa_1. \end{cases} \tag{1}$$

The constraint that $\mathbb{E}(Y|x)$ is continuous at $x = \kappa_1$ implies that $\beta_0 + \beta_1\kappa_1 = \beta_2 + \beta_3\kappa_1$. Note that we can rewrite β_2 in terms of β_0, β_1 and β_3; there are thus only 3 free parameters in this regression model.

(b) Rearranging the terms of the constraint, we obtain $\beta_2 = \beta_0 + (\beta_1 - \beta_3)\,\kappa_1$. Put this back into (1) and the result is

$$\mathbb{E}(Y|x) = \begin{cases} \beta_0 + \beta_1 x, & \text{if } x \leq \kappa_1 \\ \beta_0 + \beta_1\kappa_1 + \beta_3\,(x - \kappa_1)\,, & \text{if } x > \kappa_1. \end{cases}$$

(c) Note that the matrix multiplication

$$\begin{pmatrix} 1 & 6 & 0 \\ 1 & 10 & 2 \\ 1 & 10 & 19 \\ 1 & 10 & 0 \\ 1 & 10 & 8 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_3 \end{pmatrix}$$

gives us the desired $\mathbb{E}(Y|x)$ for each observation. Hence the matrix passed to `lm` is

$$\begin{pmatrix} 6 & 0 \\ 10 & 2 \\ 10 & 19 \\ 10 & 0 \\ 10 & 8 \end{pmatrix}.$$

2. (*Please also refer to the R code for this assignment*)

(a) A simple linear regression results in a poor fit of the data. The variable `age` is highly significant, but the R^2 value of 0.047 clearly demonstrates the inadequacy of this fit.

(b) The fit is better than that of a simple linear regression (although still quite poor!), in the sense that the fitted value drops as `age` increases beyond the median, in response to the lack of large values of `logwage` beyond 70 years of age. Unlike simple linear regression, here changing `age` by 1 has different effects on the expected value of `logwage` depending on whether `age` is above or below the median. In particular, if `age` is below median, a unit increase in `age` results in an increase of 0.0198 in `logwage`; if `age` is above median, a unit increase in `age` results in a decrease of 0.0050 in `logwage` (i.e. the difference between the two estimated slope parameters).