Assignment 1 CSC/DSC 265 Spring 2017 Due February 14^{th} Fengyi Zhao 29352739

Q1: Given a simple linear regression model $Y = \beta_0 + \beta_1 X$.

(a) $\hat{\beta}_i$ is the least squares estimates of β_i, $i = 0, 1$. A constant c is added to each response. We want to know: the new least squares estimates of β_i.

$Y = f(X) + c \quad Y = \beta_0 + \beta_i X + c$ for $i = 0, 1$

So we define the residual sum of squares (RSS) as:

$$RSS = c^2 + c^2 + c^2 + \dots + c^2 = nc^2$$

$$= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

According to the note: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Since a constant c is added to each response:

$\hat{\beta}_i$, $i = 1$: $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i + c - (\bar{y} + c))}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

$i = 0$: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = (\bar{y} + c) - \left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)\bar{x}$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i + c - (\bar{y} + c))}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \hat{\beta}_1$$

$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad \hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i + c - \beta}{n}$

$\hat{\beta}_0 = \bar{y} + c - \beta_1 \bar{x}$

$= \beta_0 + c$

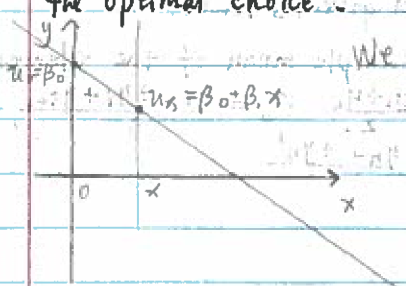
New equations:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = \beta_0 + c + \hat{\beta}_1 x$

where $\hat{\beta}_0 = \beta_0 + c$

$\hat{\beta}_1 = \beta_1$

(b) The coefficient β_0 is referred to as the intercept term. It can be interpreted as a summary of the vertical location of the response Y. Since the effect of changing the constant c of part (a) is directly observable in β_0. $\beta_0 = u_0$ when $u_x = \beta_0 + \beta_1 x$. So construct a new intercept u_x at any vertical line $X = x$. The least squares estimate will be $\hat{u}_x = \hat{\beta}_0 + \hat{\beta}_1 x$. What analytical criterion can be used to select x to get the optimal choice?



We want to choose a x that minimize the varience. So according to the formular (3.2) in the note:

$$\sigma^2_{\hat{u}x} = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

So when $x = \bar{x}$, the value of the varience $\sigma^2_{\hat{u}x}$ becomes the smallest. $\sigma^2_{\hat{u}x} = \sigma^2\left(\frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right) = \sigma\left(\frac{1}{n}\right) = \frac{\sigma^2}{n}$ So we want the line of best fit to pass through $(\bar{x}, \bar{y})$.

Q2: We are given a 'hat matrix' associated with multiple linear regression with q predictors. H is a linear transformation of an n-dimensional response vector y to the n-dimensional fitted vector $\hat{y} = Hy$. $\hat{y}$ is obtained by minimizing the SSE, the action taken by H on y is to project it onto the q-dimensional subspace $S_q \subset \mathbb{R}^n$ spanned by the q predictors. $\hat{y}$ is the point in S_q closest to y.