**Linear Regression Models**
**Statistics GR5205/GU4205 — Fall 2020**

**Homework 2**

**The following problems are due on Monday, October 5, 11:59pm.**
**Throughout this homework, we assume that the ground truth $\beta_0$, $\beta_1$ and $\sigma^2$ are all real-valued numbers or vectors!**

1. **Least Square Estimator in Simple Linear Regression** Given $x_1, \ldots, x_n \in \mathbb{R}$, and assume that not all of $x_i$ are the same. Suppose the simple linear regression model holds, where[1]:

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n$$

where $\mathbb{E}[\epsilon_i] = 0, \mathrm{Var}[\epsilon_i] = \sigma^2$, and for $i \neq j$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$. In class, we introduce the matrix form as

$$Y = \beta_0 + x\beta_1 + \epsilon,$$

where $\mathbb{E}[\epsilon] = 0, \mathrm{Var}[\epsilon] = \sigma^2 I_n$. The least square estimator for this model is

$$\widehat{\beta}_1 = \frac{(x - \bar{x}\mathbb{1}_n)^\top (Y - \bar{Y}\mathbb{1}_n)}{\|x - \bar{x}\mathbb{1}_n\|^2}, \qquad \widehat{\beta}_0 = \bar{Y} - \bar{x}\widehat{\beta}_1, \qquad \widehat{\sigma}^2 = \frac{1}{n-2}\left\|Y - \widehat{\beta}_0\mathbb{1}_n - x\widehat{\beta}_1\right\|^2.$$

(a) Show that, the least estimators are unbiased, i.e.

$$\mathbb{E}[\widehat{\beta}_1] = \beta_1, \qquad \mathbb{E}[\widehat{\beta}_0] = \beta_0, \qquad \mathbb{E}[\widehat{\sigma}^2] = \sigma^2.$$

(b) Verify that

$$\mathrm{Var}[\widehat{\beta}_1] = \frac{\sigma^2}{\|x - \bar{x}\mathbb{1}_n\|^2}, \qquad \mathrm{Var}[\widehat{\beta}_0] = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\|x - \bar{x}\mathbb{1}_n\|^2}\right).$$

**Hint: try the same argument introduced in Lecture 4.**

[1] In this expression, we treat $x_i$ as given and $Y_i$ are observable response random variables and $\epsilon_i$ are unobservable error term.

2. **Parameters Estimation in Multivariate Linear Regression**

Given $x_1, \ldots, x_n \in \mathbb{R}$, and assume that not all of $x_i$ are the same. Recall the simple linear regression model with Gaussian errors[2]:

$$Y_i = \beta_0 + x_i\beta_1 + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n$$

where $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$. Or equivalently, we can write

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}.$$

Define

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad x = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1, & x_n \end{bmatrix}, \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

then an expression of the normal simple linear regression model in matrix terms is

$$Y = x\beta + \epsilon, \qquad \text{where } \epsilon \sim \mathcal{N}\left(0, \sigma^2 I_n\right).$$

Let $\|\cdot\|$ be the $L^2$-norm of an $n$-dim vector, i.e. for $y \in \mathbb{R}^{n\times 1}$,

$$\|y\| = \sqrt{y_1^2 + y_2^2 + \cdots + y_n^2},$$

the mean squared error can be expressed as $Q(\beta) = \|Y - x\beta\|^2$.

(a) Derive the MSE minimizer $\widehat{\beta}$ in matrix from, using only $Y$ and $x$.

**Hint:** in matrix calculus, we have

$$\frac{\partial x\beta}{\partial \beta} = x^\top, \qquad \frac{\partial \beta^\top \Sigma \beta}{\partial \beta} = \left(\Sigma + \Sigma^\top\right)\beta.$$

(b) Now fix a $p \geq 2$. Suppose $x_i \in \mathbb{R}^{(p-1)\times 1}$ and $\beta_1 \in \mathbb{R}^{(p-1)\times 1}$ are both $(p-1)$-dim vector, i.e. we are using $p-1$ predictor variables to predict the 1-dim response variable $Y$. We write

$$x = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_1^{(p-1)} \\ 1 & x_2^{(1)} & \cdots & x_2^{(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & \cdots & x_n^{(p-1)} \end{bmatrix} = \begin{bmatrix} 1 & x_1^\top \\ 1 & x_2^\top \\ \vdots & \vdots \\ 1 & x_n^\top \end{bmatrix} \in \mathbb{R}^{n\times p},$$

and we claim that $\text{rank}(x) = p$. Will the MSE minimizer $\widehat{\beta}$ the same as in (a)?

[2]Again, in this expression, we treat $x_i$ as given and $Y_i$ are observable response random variables and $\epsilon_i$ are unobservable error term.

(c) Write the fitted values $\widehat{Y}$ in matrix form for fixed $p \geq 2$, using only $Y$ and $x$.

(d) (**linear algebra practice, will not assign credit.**) Try to show that the results from HW1 Q3 also hold for multivariate regression model ($p \geq 2$), using matrix representation.

(e) Let $\mathbb{1}_n$ denote the $n$-dim all one vector. For fixed $p \geq 2$, show that

$$\|\widehat{Y} - \bar{Y}\mathbb{1}_n\|^2 = \left(\widehat{Y} - \bar{Y}\mathbb{1}_n\right)^\top \left(Y - \bar{Y}\mathbb{1}_n\right).$$

Notice that the second term on the right-hand side is the response variable $Y$ instead of the prediction $\widehat{Y}$.

**Hint: The results from HW1 Q3 also hold in multivariate linear regression models as well, and you can directly use that. Or, notice that**

$$x(x^\top x)^{-1}x^\top x = x,$$

**which mean if choosing the first column of the matrices on both sides, we have**

$$x(x^\top x)^{-1}x^\top \mathbb{1}_n = \mathbb{1}_n.$$

3. **Problems 2.2 and 2.11–2.12 in KNN**

(a) In a test of $H_0 : \beta_1 \leq 0$ versus $H_a : \beta_1 > 0$ we fail to reject $H_0$, and an analyst concludes that there is no linear association between $X$ and $Y$. Do you agree? Explain.

(b) The same analyst later claims that "estimating the mean response at $x = x_0$" and "predicting the mean of $m$ new observations at $x = x_0$" are essentially the same problem. Do you agree? Explain.

(c) An expression for the variance of $\widehat{Y}_0 = b_0 + x_0 b_1$ is given by

$$\text{Var}(\widehat{Y}_0) = \sigma^2 \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] ,$$

and thus (the same analyst claims) we can be 95% confident that the next response observed at $x = x_h$ will fall within bounds given by

$$\widehat{Y}_0 \pm t(.975; n-2) \left\{ MSE \left[ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right] \right\}^{1/2} .$$

Do you agree? Explain.

4. **Problem 2.5 in KNN**

Continue with the *Copier maintenance* data introduced on the previous homework assignment; recall $X$ denotes the number of copiers serviced and $Y$ the total number of minutes spent on a service call. Assume the normal SLR model is appropriate.

(a) Conduct a $t$-test to determine whether or not there is a linear association between $X$ and $Y$. Clearly state the null and alternative hypotheses in terms of model parameters. Report and interpret the $p$-value from your test.

(b) Use a 95% confidence interval to estimate the change in mean service time when the number of copiers serviced increases by one. Interpret your confidence interval.

(c) The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Address this question in two ways:

i. By inspection of your confidence interval in part (b), and

ii. by conducting a formal significance test of the appropriate hypotheses, reporting and interpreting the $p$-value from the test.

Are your conclusions consistent?

(d) Does $b_0$ give any relevant information here about the “start-up” time on calls, i.e., about the time required before service work is begun on the copiers at a customer location?