MATH 3333 3.0 - Winter 2019-20
Assignment 1
(Due Date: January 21, 2020)

Question 1: Please download the bike sharing data set from the following website: https://archive.ics.uci.edu/ml/datasets/bike+sharing+dataset. This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information. There are two data sets and we will use the day.csv. Import the data into R and perform the following exploratory analysis.
a) The variable "registered" records the number of registered users used the bike sharing service on a particular day. Please provide the mean value of the variable "registered" for each day of the week.

b) Plot the conditional density plot of the variable "registered" conditional on each day of the weak.

c) Produce a two-dimensional levelplot of the variable "registered" against the combination of temperature (variable "temp") and humidity (variable "hum").

Question 2: Perform linear regression model on the bike sharing data set from Question 2.

a. Provide the summary result of the regression model with "registered" as the response variable and "temp", "hum" as the predictors. You can copy and paste the regression result from the R output.

b. What other predictors do you think might be important for the modelling of the variable "registered"? Please construct another linear model including more predictors and provide the summary result of the second model.

c. Use adjusted R-square to determine which model is a better model.

Question 3: In the following marketing set, we have 9 years with the sales in 10 million euro and the advertising expenditure in million euro.

Year	Sales	Advertisement
1	65	23
2	76	26
3	85	30
4	106	34
5	119	43
6	129	48
7	142	52
8	144	57
9	151	58

a) Formulate the response vector Y, which has nine entries.

b) Formulate the data matrix of X, the first column should be all ones corresponding to the intercept, and the second column should be the predictors. The dimension of X should be 9×2.

c) Write R code to compute X^tX.

d) Compute X^tX by hand using calculator and compare with part c.

e) Write R code to compute $(X^tX)^-$.

f) Compute part d by hand using calculator.

g) Write R code to compute $\theta = (X^tX)^-X^tY$. This is the estimated linear regression coefficient of the linear model with Y as the response and X as the data matrix.

f)Compute part (g) by hand using calculator.

Question 4:

a. In this question, we will learn to program cross validation. Using R to partition the bike sharing data into two equal parts. Run the regression model using "registered" as the response and "temp" and "hum" only using the first half of the data. Based on the output model, build a predictive function with "temp" and "hum" as input and the "registered" as output. The predictive function is registered=θ_0+θ_1 temp+ θ_2 hum. The regression coefficients are obtained from the R output.

b. Apply the predictive function on each of the observations in the second half of the data. For each observation, calculate the error between the observed "registered" value and the value using the predictive function from part a. Use a loop to add up all the squared errors. Report the total of sum of squared errors for the second half of the data.

c. Repeat part (a) and part (b) using the regression model you proposed in Question 2.b.

d Compare the two models based on the cross validation errors.

Question 5: Consider a data set with the response vector $Y = (y_1, \ldots, y_n)^t$ and the data matrix

$$\begin{pmatrix} 1 & x_{11} & x_{12} \\ . & . & . \\ . & . & . \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

We model the relationship between X and Y using the linear regression model: $y_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2} + \epsilon_i,\ i = 1, \ldots, n$, where $\epsilon \sim N(0, \sigma^2)$. Let the parameter vector be denoted as $\theta = (\theta_0, \theta_1, \theta_2)^t$. We wish to minimize the sum of squared residuals: $SSE = \sum_{i=1}^{n}(y_i - (\theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}))^2$. Let the fitted value be denoted as $\hat{y}_i = \theta_0 + \theta_1 x_{i1} + \theta_2 x_{i2}$, and let the fitted value vector be denoted as $\hat{Y}$.

a) Show that $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

b) Show that $SSE = (Y - \hat{Y})^t(Y - \hat{Y})$.

c) Show that $\hat{Y} = X\theta$.

d) Simplify the derivative equation $\frac{\partial SSE}{\partial \theta} = 0$.

e) Find the solution of θ which solves the equation in part d.