# DASC 5420: Assignment 1

**Due Date: February 19, 2024**

---

Make an **R Markdown** file and produce a **pdf** file by solving the following problems. In the title of your file make sure you include your name and student ID number.

1. (5 marks)

   Use **R** for this question and attach your source codes together with the outputs you obtain.

   The dataset `USArrests` (available in **R**) contains violent crime rates per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. You are also given the percent of the population living in urban areas.

   To see this dataset, simply type `USArrests` in **R**, and for the details use `?USArrests`.

DONE (a) Construct box plots for each of the four variables and comment on the univariate characteristics of each briefly.

DONE (b) Obtain the sample mean vector $\overline{\mathbf{X}}$, the sample covariance matrix $\mathbf{S}$ and the sample correlation matrix $\mathbf{R}$. What can you say about the relationship between the four variables?

DONE (c) Let

$$d_S(\mathbf{x}, \mathbf{q}) = \sqrt{(\mathbf{x}-\mathbf{q})' \mathbf{S}^{-1} (\mathbf{x}-\mathbf{q})},$$

where $\mathbf{S}$ is the sample covariance matrix, be the statistical distance between an observed point $\mathbf{x}$ and a fixed point $\mathbf{q}$.

Find the state with the highest murder rate and compute the statistical distance of each observation from that state.

For this, you can use the *stat.dist* function, which calculates the statistical distance between each point of a data matrix Y from a given point p:

```
stat.dist = function(Y,p){
S.inv= solve(var(Y))
dist= function(y){sqrt(t(y) %*% S.inv %*% y)}
return(apply(t(apply(Y,1, function(y) y-p)),1,dist))
}
```

Note that, `USArrests` is defined as a data frame object, with row names being the 50 states. To calculate the statistical distance, you need to define a matrix that contains only the numeric entries of `USArrests`. You can do this as follows,

```
X= as.matrix(USArrests)
rownames(X) = NULL
```

Now, you will be able to calculate the statistical distance of each observation from the $i^{th}$ one using

```
stat.dist(X, X[i, ])
```

DONE (d) Using these computed distances determine which six states are "closest" to the state with the highest murder rate. Can you provide an explanation why these seven states are statistically similar in terms of their violent crime rates and percentage of urban population in 1973?

DONE (e) Obtain pairwise scatterplots (also called matrixplot) for the four variables in R and identify (using a different color or symbol) the points for the seven states you found in part (d). Do what you see on the scatterplots and what you concluded in part (d) agree with each other?

2. (5 marks)

This question involves the use of Linear Regression on the *WHO Life Expectancy* data set. Use **R** to complete these tasks. Make sure you included all the **R** codes.

*WHO Life Expectancy* data related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from the United Nations website.

- Description: https://www.kaggle.com/kumarajarshi/life-expectancy-who

The data (`Life-Expectancy-Data.csv`) is available in Moodle.

DONE (a) Design a hypothesis for something you can predict from this data set using Linear Regression. Write your hypothesis and provide justification. Include an explanation of why you think you will be able to predict something and list the inputs and output.

- It is okay if you are predicting a new value as long as it is computed using existing values in the data set.
- It must be a regression task, i.e., predicting a scalar value.
- Include at least one visualization of the variables you are using in this hypothesis.

DONE (b) Process any necessary data transformation. Explain why you are using that transformation. This could include:

- Feature scaling such as standardizing or normalizing the data.
- Selecting or removing certain values (such as outliers or missing values).

DONE (c) Perform a Linear Regression and comment on the output. For instance:

- Is there a relationship between the inputs and the output?
- Which inputs appear to have a statistically significant relationship to the output?

DONE (d) Print out your algorithm performance. Choose the right metric(s) for judging the effectiveness of your prediction.

DONE (e) Iterate and improve your algorithm performance. Write down the options you tried and the methods you used to increase the performance of your learning algorithm. Key things to mention are how you determined algorithm parameters, which variables and feature scaling options you explored. Visualize the process if possible (e.g., performance for different parameter values).

DONE 3. (5 marks)

Let $Y$ be a $n \times 1$ vector for the response variable, $\mathbf{X}$ be a $n \times (p+1)$ design matrix where the first column contains 1's, $\boldsymbol{\beta}$ be the $(p+1) \times 1$ vector of regression coefficients, and $\boldsymbol{\epsilon}$ is a $n \times 1$ vector of error terms. Then the multiple linear regression model can be written as

$$Y = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim \text{MVN}(0, \sigma^2 I)$, where $I$ is an identity matrix.

(a) Find the least square (LS) estimate of $\boldsymbol{\beta}$. Show that LS estimator $\hat{\boldsymbol{\beta}}_{LS}$ is an unbiased estimate of $\boldsymbol{\beta}$ [*i.e.*, $\text{E}(\hat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{\beta}$] and

$$\text{Cov}(\hat{\boldsymbol{\beta}}_{LS}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}.$$

(b) Write down the likelihood function (LF) for $\boldsymbol{\beta}$ and $\sigma^2$. Find the maximum likelihood estimate (MLE) of $\boldsymbol{\beta}$ and $\sigma^2$. Compare the ML estimates with LS estimates from (a).

(c) Derive $\text{Var}(\widehat{\boldsymbol{\beta}}_{MLE})$, where $\widehat{\boldsymbol{\beta}}_{MLE}$ is the maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$.

(d) Given $x_k$, a data value from the training data, derive $\widehat{Y}_k$ and find $\text{Var}(\widehat{Y}_k)$.

(e) Given $x_{new}$, a new vector of feature variables from the test data, derive $\widehat{Y}_{new}$ and find $\text{Var}(\widehat{Y}_{new})$. Also find $\text{Var}(Y_{new} - \widehat{Y}_{new})$.

DONE 4. (4 marks)

Consider the regression model:

$$\text{E}(Y_j) = \beta_0 + \beta_1 x_j + \beta_2(3x_j^2 - 2), \quad j = 1, 2, 3$$

where $x_1 = -1, x_2 = 0$, and $x_3 = 1$.

(a) Find least squares estimate (LSE) of $\beta_0, \beta_1$ and $\beta_2$.

(b) Show that the LSE of $\beta_0, \beta_1$ are unchanged if $\beta_2 = 0$.

DONE 5. (2 marks)

Let $Y_i \in \{0, 1\}$ and $x_i$ be the output and input variables in a logistic regression problem where the probability of success is defined as

$$\pi_i = \Pr(Y_i = 1) = \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}},$$

where $\beta_0$ and $\beta_1$ are the intercept and slope parameters, respectively. To estimate the parameters $\beta_0$ and $\beta_1$, we define the likelihood function as

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i},$$

where n is the sample size. Write down the explicit form of the likelihood function to get the maximum likelihood estimate (MLE) of $\beta_0$ and $\beta_1$. How can we obtain the MLE of $\beta_0$ and $\beta_1$?

(Here, you don't need to find out the MLE of $\beta_0$ and $\beta_1$. You just need to write the equations which are used to get the MLE of $\beta_0$ and $\beta_1$.)

DONE 6. (4 marks)

Use **R** for this question. The file `twins.Rdata` (available on Moodle) contains a subset of data on the National Merit Twin Study. The variables on the data file, recorded for each twin pair, are:

| | |
|---|---|
| `Pairnum` | `Twin pair number` |
| `Zygosity` | `1= identical, 2 = fraternal` |
| `English1` | `English score of the first-born twin` |
| `Math1` | `Mathematics score of the first-born twin` |
| `SocSci1` | `Social Science score of the first-born twin` |
| `NatSci1` | `Natural Science score of the first-born twin` |
| `Vocab1` | `Vocabulary score of the first-born twin` |
| `English2` | `English score of the second-born twin` |
| `Math2` | `Mathematics score of the second-born twin` |
| `SocSci2` | `Social Science score of the second-born twin` |
| `NatSci2` | `Natural Science score of the second-born twin` |
| `Vocab2` | `Vocabulary score of the second-born twin` |

Double click on the data fille `twins.Rdata` to import it into R, or use: `load(file="twins.Rdata")` You can check the first few rows of the dataset using: `head(twins)`

Report (i) the number of twin pairs, (ii) the number of variables, (iii) the number of identical twins and (iv) the number of fraternal twins.

Upload the final **pdf** file in Assignment 1 in Moodle. Make sure you have included all the **R** codes in the output file. Also make sure your **pdf** file is well formatted and contains your name.

Good Luck!