# Homework Assignment 2

DS4043, Spring 2023

**Due on March 29, 2023 at 11:59 pm**

1. Consider the multivariate normal distribution vector $\mathrm{X} = (\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3)^{\mathrm{T}}$ having mean vector $\boldsymbol{\mu} = (0, 1, 2, )^{\mathrm{T}}$ and covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 1 \end{bmatrix}$$

a) Generate 100 random observations from the multivariate normal distribution given above with *set.seed(12)*. (Hint: see *?mvrnorm* ) You may need to use the package MASS.

```
library(MASS) # you may need to use this package
set.seed(12)
```

b) Construct a scatterplot matrix for **X** and add a fitted smooth density curve on the diagonal panels for each $\mathrm{X}_1, \mathrm{X}_2, \mathrm{X}_3$ to verify that the location and correlation for each plot agrees with the parameters of the corresponding bivariate distributions.

c) Obtain the correlation plot for the generated sample **X**, where coefficients are added to the plot whose magnitude are presented by different colors. Let the visualization method of correlation matrix to be ellipse.

```
library(corrplot) # you may need to use this package
```

d) Given the covariance matrix $\Sigma$, find $\sigma_{x_1}$, $\sigma_{x_2}$ and $\rho_{x_1 x_2}$. Consider the joint PDF of bivariate normal distribution

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$
$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]\right\},$$

sketch a surface plot for $X_1$ and $X_2$, based on their bivariate probability density function. (Hint: if you want to use *curve3d*, please install and use the package *emdbook*)

```
library(emdbook) # you may need to use this package
```

e) Sketch 3-D scatter plots for each of $X_1, X_2$ and $X_3$ as a $z$ axis and rest two variables as $x$ and $y$ axes. Put these 3 plots in one picture.

```r
library(lattice) # you may need to use this package
```

2. A continuous random variable $X$ has the probability density function

$$f_X(t) = \begin{cases} at + bt^2 & 0 < t < 1 \\ 0 & \text{otherwise} \end{cases}.$$

If $E[X] = 1/2$, find (a) $a$ and $b$; (b) $P(X < 1/2)$; (c) $\text{Var}(X)$; (d) Generate the density plot of $X$

3. Consider a nonparametric regression model

$$y_i = g(x_i) + \epsilon_i, \quad 1 \leq i \leq n,$$

where $y_i$'s are observations, $g$ is an unknown function, and $\epsilon_i$'s are independent and identically distributed random errors with zero mean and variance $\sigma^2$. $n$ is the number of observations. Usually one fits the mean function $g$ first and then estimates the variance $\sigma^2$ from residual sum of squares $\hat{\sigma}^2 = \sum_{i=1}^n \hat{\epsilon}_i^2/(n-1)$ where $\hat{\epsilon}_i = y_i - \hat{g}(x_i)$. However this method requires an estimate of the unknown function $g$. Then some researchers proposed some difference-based estimators which does not require the estimation of $g$. Assume that $x$ is univariate and $0 \leq x_1 \leq \cdots \leq x_n \leq 1$. Rice (1984) proposed the first order difference-based estimator

$$\hat{\sigma}_R^2 = \frac{1}{2(n-1)} \sum_{i=2}^n (y_i - y_{i-1})^2.$$

Gasser, Sroka and Jennen-Steinmetz (1986) proposed the second order difference based estimator and for equidistant design points (i.e. $x_i$ and $x_{i+1}$ have the same distance for all $i = 1, 2, \ldots, n$), $\hat{\sigma}_{GSJ}^2$ reduces to

$$\hat{\sigma}_{GSJ}^2 = \frac{2}{3(n-2)} \sum_{i=2}^{n-1} \left( \frac{1}{2} y_{i-1} - y_i + \frac{1}{2} y_{i+1} \right)^2.$$

Consider the temperature anomaly dataset. Temperature anomalies in degrees Celsius are based on the new version HadCRUT4 land-sea dataset (Morice et al., 2012). We focus on the global median annual temperature anomalies from 1850 to 2019 relative to the 1961-1990 average. We try to build up the model between time and global median temperature $y_i$ and year $x_i$.

(a) Use *read.csv* to read the temperature anomaly dataset. Let $x$ be the vector of years from 1850-2019, $y$ be the vector of corresponding global median annual temperature anomalies, and $n$ be the number of observations

(b) Display a scatter plot between global median annual temperature anomalies and years with caption "Global median land-sea temperature anomaly relative to the 1961-1990 average temperature", $x-$label years and $y-$label temperature anomalies.

(c) Change the years $x$ to a new vector $x$ such that $x_i = i/n$. Compute the first order difference-based estimator. (Note: the change of $x$ or not will not affect the computation of the estimator)

(d) Compute the second order difference-based estimator.