Question 2.1:

This practical question aims to analyse a high dimensional microarray data set called `eyedata`, available on Ultra, which contains gene expression data of 200 genes for 120 samples, along with the TRIM32 gene of 120 samples as the response variable of interest. The data originates from microarray experiments of mammalian eye tissue samples. The first variable (column) in data is the TRIM32 gene as the response variable, and the other variables (columns) are the 200 genes as the covariates, quantified for 120 samples (observations in rows).

The main objectives of this study were to find out which genes are associated with the TRIM32 and to predict the expression levels of TRIM32 from the expression levels of the 200 genes measured in the microarray experiment.

(a) Load the `eyedata` into R using the command `load(file="eyedata.rda")`. Note that you may need to paste the right directory of the data file in your computer. For example, `load(file="C:/Users/bvft24/Desktop/eyedata.rda")`

(b) In R, create the response variable vector $\boldsymbol{Y}$ being the first column of `eyedata`. Also, create the covariates design matrix $\boldsymbol{X}$ being all columns of `eyedata` except the first column. Do centre both $\boldsymbol{Y}$ and $\boldsymbol{X}$ using the R function `scale` so that all the covariates and the response variable have zero mean (no need to scale them having variance 1).

(c) Consider a linear regression model with $\boldsymbol{Y}$ and $\boldsymbol{X}$. Show that the OLS method fails to work here. For this, first show that $\boldsymbol{X}$ does not have full column rank by computing the rank of $\boldsymbol{X}$ using the R function `qr`. Then calculate the inverse of matrix $\boldsymbol{X}^T\boldsymbol{X}$ using the R function `solve` and observe that the inverse does not exist due to singularity as $n < p$. Despite the OLS does not work properly, apply it using the R function `lm` and interpret what goes wrong in the results.

(d) Apply the ridge regression model with $\boldsymbol{Y}$ and $\boldsymbol{X}$ using the R function `glmnet` with the regularisation parameter set to $\lambda = 2$. You may need to first install the R package `glmnet` and load it in R using the following commands:

```
install.packages("glmnet")
library(glmnet)
```

For this case of $\lambda = 2$, extract the ridge estimates of regression coefficients $\boldsymbol{\beta}$ using the R function `coef`.

Now, choose an optimal value of λ for ridge regression by applying the cross validation (CV) using the R function `cv.glmnet` which by default minimises the deviance. Plot the CV output using the R function `plot`. Fit the ridge regression model to the data with this optimal value of λ and extract the estimates of regression coefficients $\boldsymbol{\beta}$. What do you observe when comparing the estimates with the previous case of $\lambda = 2$?

(e) Apply the lasso regression with $\boldsymbol{Y}$ and $\boldsymbol{X}$ using the R function `glmnet` with an optimal value of λ being chosen using the cross validation with the R function `cv.glmnet`. Plot the CV output using the R function `plot`. Extract the lasso estimates of parameters $\boldsymbol{\beta}$ and compare with the ridge estimates in terms of the sparsity of solution.

(f) Randomly partition the data to training data (70%, say $\boldsymbol{Y}_{\text{train}}$ and $\boldsymbol{X}_{\text{train}}$) and test data (30%, say $\boldsymbol{Y}_{\text{test}}$ and $\boldsymbol{X}_{\text{test}}$). Apply ridge and lasso regression to the training data, both with λ being chosen by CV, and then use the fitted models to predict the response values in the test data. Calculate the mean squared error of predictions for the two methods and comment which method provides better predictions and discuss why.

Question 2.2:

Consider the high dimensional data set called `riboflavin`, available on Ultra, which is obtained from a high-throughput genomic study concerning the riboflavin (vitamin B_2) production by bacillus subtilis. The main objective of the study was to find out which genes are associated with the production rate of vitamin B_2. The data set contains 71 samples and 4088 covariates corresponding to 4088 genes, along with the logarithm of riboflavin production rate as response variable.

(a) First, load the `riboflavin` data into R environment using the command `load(file="riboflavin.rda")`. Again you might need to specify the right directory of the data file in your computer. Then, create the response variable vector $\boldsymbol{Y}$ from this data set using the command `Y <- riboflavin$y`. Also, create the covariates design matrix $\boldsymbol{X}$ using the command `X <- riboflavin$x`. Center the covariates $\boldsymbol{X}$ to have zero mean but not the response variable $\boldsymbol{Y}$.

(b) Calculate the pairwise correlations between covariates (or genes) to see if they are correlated. For this, use the R function `cor`. Create a histogram of the pairwise correlations using the R function `hist`. What do you observe here?

(c) Apply the ridge and lasso regression to the `riboflavin` data, both with λ being chosen by CV. Interpret the results.

(d) Randomly partition the data to training data (70%, say $\boldsymbol{Y}_{\text{train}}$ and $\boldsymbol{X}_{\text{train}}$) and test data (30%, say $\boldsymbol{Y}_{\text{test}}$ and $\boldsymbol{X}_{\text{test}}$). Apply ridge and lasso regression to the training data and then use the fitted models to predict the response values in the test data. Calculate the mean squared error of predictions for the two methods and discuss which method provides better predictions.

(e) Note that this and the next parts of question shall be tried later as they concern "elastic net regularisation" for correlated variables which will be covered at the end of Chapter 2. In R package `glmnet`, for computational convenience, the elastic net estimator is defined as

$$\hat{\boldsymbol{\beta}}_{\text{EN}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ 1/n \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \alpha\|\boldsymbol{\beta}\|_1^1 + (1-\alpha)/2\|\boldsymbol{\beta}\|_2^2 \right\},$$

in which $0 \le \alpha \le 1$ is a single regularisation parameter. Note that $\alpha = 1$ results in the lasso penalty and $\alpha = 0$ leads to the ridge penalty. Note also that the factor $1/n$ is just a normalising constant.

Setting $\alpha = 0.5$ in `glmnet`, apply the elastic net to the `riboflavin` data and extract the parameter estimates. What are the estimates with $\alpha = 0.8$?

(f) Use cross validation to find an optimal value of α for the elastic net. For this, consider a range of values for α, say $\alpha = i/20$ for $i = 0, 1, \ldots, 20$, and investigate which one of these α values results in smallest prediction error when applied to randomly selected training (70%) and test (30%) data. Can this be related to the correlations among genes investigated in Part (b)?