School of Mathematics and Statistics
MAST90083: Computational Statistics and Data Science
Assignment 1
Due date: No later than 08:00am on Monday 9th September 2024
Weight: 20%

Question 1 Linear Regression

In this question, we will apply linear regression, ridge regression and Lasso to the R built-in dataset `Hitters` and compare their performance. Type `?Hitters` to see the description of this data. Our goal is to predict the response variable `salary` using these three regression models. To complete this task, you'll need to install two packages: `ISLR` and `glmnet`. (Note: When using random functions in this exercise, please set the random number generator with `set.seed(10)` to ensure reproducibility.)

1. **Data preparation:** Load the Hitters dataset from `library(ISLR)`. Remove all those rows from Hitters dataset that have entry NA in the `salary` column. For a design matrix construction, include all variables in Hitters dataset except the `salary` and store them in variable $\mathbf{X}$ (matrix of covariates). Also, read the `salary` variable and store it in variable $\mathbf{y}$ (response).

2. **Linear regression model:** First, we consider the full model that includes all the potential variables and do the following tasks: (1) compute the least square estimates $\hat{\boldsymbol{\beta}}_{LS}$ of $\boldsymbol{\beta}$ and the corresponding p-values manually using the sampling distribution property of the least squares estimator and check with the R built-in function `lm()`. From these p values, which variable (variables) can you consider significant (or meaningful) ? (2) Now we consider a smaller model using those variables that have p values smaller than 0.05 in (1). Compute the $\hat{\boldsymbol{\beta}}_{LS1}$ and the corresponding p-values for this smaller model. (3) Compare these two models using F-statistics to test whether the smaller model provides a better characterization of the linear relationship between $\mathbf{X}$ and $\mathbf{y}$. First compute manually the F statistics and associated p-value and then check with the F statistics obtained from R built-in function `anova()`. What conclusions can you make from this comparison? Can you reject the hypothesis that the smaller model is correct?

3. **Ridge regression:** Generate a sequence of 100 values of λ between 10^5 and 10^{-2} and use the `glmnet` function. For ridge regression, we set $\alpha = 0$ and do the following tasks: (1) compute estimates of the ridge regression coefficients for the 100 different λ values. Then, observe the set of coefficients for two extreme values of λ i.e. 10^5 and also for 10^{-2}. For which value of λ among these two, the ridge coefficient values are closer to zero? (2) Draw a plot of $l2$-norm of the ridge coefficient values (excluding the intercept's coefficient value) against the logarithm of the λ values. From this plot, can you decide the optimal λ value between 10^5 and 10^{-2}, or is it better to use the

mean square error (MSE) plot against the λ values? justify your response. (3) Use the `cv.glmnet` function to perform ten fold cross validation (CV). First, we randomly pick half samples from $\mathbf{x}$ for all variables and also the corresponding samples from $\mathbf{y}$ to construct a training dataset (use `set.seed(10)` for random number generator). The rest of the samples are saved for the testing dataset. Using this training dataset, plot the cross validation results, and find the best λ (the one that results in smallest CV error) value and its corresponding MSE_{RR} value obtained using the testing dataset. You can use `predict()` function here. Now refit the ridge regression model on the full data set using the λ chosen by CV. Examine the ridge regression coefficients, are they all present, similar to the linear regression case?

4. **Lasso:** This time we set $\alpha = 1$ and again (1) plot the cross validation results, and find the best λ value using the same training set used in ridge regression above and its corresponding $\text{MSE}_{\text{Lasso}}$ value using the same testing set used in ridge regression model above. (2) Now obtain the coefficients again using the best λ that we just selected using CV. Are all the coefficients selected again?

5. Here we consider a regularization technique that combines both the ridge regression penalty and lasso penalty, which can be written as $\frac{1-\alpha}{2}||\boldsymbol{\beta}||_2^2 + \alpha|\boldsymbol{\beta}|_1, 0 \leq \alpha \leq 1$ (you can see that when $\alpha = 1$, it corresponds to the Lasso penalty and when $\alpha = 0$, it corresponds to the ridge regression penalty).
   Now, use this penalty to combine the benefits of ridge regression and Lasso by finding a best α. This still can be achieved using `glmnet()`. Type `?glmnet()` to see the details of each argument in this function. Let $\boldsymbol{\alpha} \in [0, 1]$ be a set of 11 different values obtained using `seq(0,1,by=0.1)`. Then, (1) for each of this α_i, i=1,...,11. apply the ten fold cross-validation to select the best λ values and compute the associated MSE_i using the same testing set used in ridge regression and Lasso above. (2) Select the best α as the one that results in the smallest MSE and estimate the regression coefficients again using the best α and the corresponding best λ that where selected. (3) Are all the coefficients selected again? Compare this results with the coefficients obtained from ridge regression in 3 and Lasso in 4, what do you observe? (Hint: You can check whether your code is correct or not by checking $\text{MSE}_1 = \text{MSE}_{\text{RR}}$ and $\text{MSE}_{11} = \text{MSE}_{\text{Lasso}}$)

Question 2

In this question, we will analyze the `weather_stations.csv` dataset, which contains daily average pressure readings (measured in inches) from 17 weather stations located in the central part of Oklahoma State, USA. The data covers the period from May 2, 2017, to September 30, 2017, covering a total of 152 days. In other words, we have 17 time series, each representing daily average pressure over 152 days in 17 different weather stations. In addition to the pressure data, we also have geographic information for these 17 weather stations stored in the file `geoinfo_stations.csv`. We are interested in three spatial variables: `NLAT` (North Latitude),`ELON` (East Longitude) and `ELEV` (Elevation in meters) for each weather station in this file.

1. At weather station $j, j = 1, \cdots, 17$, the daily average pressure data at time $t, t =$

$1, \cdots, 152$ is $y_{j,t}$. For each time series, we consider the following model of order p

$$y_{j,t} = \sum_{i=1}^{p} \phi_{ij} y_{j,t-i} + \beta_{1j}\texttt{NLAT}_j + \beta_{2j}\texttt{ELON}_j + \beta_{3j}\texttt{ELEV}_j + \epsilon_{j,t}, t = p+1, \cdots, 152. \quad (1)$$

Provide a linear matrix form for model (1) where the parameters are presented in a $p \times 1$ vector $\boldsymbol{\phi} = (\phi_1, \cdots, \phi_p)^T$ and a 3×1 vector $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3)^T$, derive the expressions of least square estimators of $\boldsymbol{\phi}_j$ and $\boldsymbol{\beta}_j$. (Note a linear matrix form for model includes the expression for observed time series $\boldsymbol{y}_j$, design matrices and the error term in vector form.)

2. First, only consider the centered daily pressure data and don't include the geographic variables in the model. In this case the time series can be expressed using the following model

$$y_{j,t} = \sum_{i=1}^{p} \phi_{ij} y_{j,t-i} + \epsilon_{j,t}, t = p+1, \cdots, 152. \quad (2)$$

This also known as autoregression (AR) model with order p. Now we aim to use AIC and BIC to find the best order p in model (2) for each time series $j, j = 1, \cdots, 17$. This means, for time series j, use the R built-in function `AIC()` and `BIC()` to compute the AIC and BIC values for different order p, say `p=seq(1,15,by=1)` then select the best order p using AIC and BIC. Repeat these steps for all the 17 time series. Draw two plots (one for AIC and the other for BIC) and each of them should have 17 curves. Each curve represent the AIC (or BIC) values against the order p for the time series j and provide your best order p for the 17 time series. (Hint: You can use `arima()` function in R to fit AR(p) model and get the log-likelihood for computing AIC and BIC)

3. Now we use the same best order $p^\star$ obtained from 2 for each of the 17 time series in order to perform the following analysis. From 2, we can get 17 pairs of coefficients $(\widehat{\boldsymbol{\phi}}_j, \hat{\eta}_j)$, for the 17 time series, or 17 sets of parameters in total so that each time series is modeled as

$$y_{j,t} = \sum_{i=1}^{i=p^\star} \phi_{ij} y_{j,t-i} + \hat{\eta}_j + \epsilon_{j,t}, t = p+1, \cdots, 152. \quad (3)$$

It is aimed to have a single model for all stations time series. For this purpose, it is assumed that $\hat{\eta}_j$ depends on the spatial covariates, $\texttt{NLAT}_j$, $\texttt{ELON}_j$ and $\texttt{ELEV}_j$. This assumption can be checked if it is reasonable by fitting the following linear regression model

$$\hat{\boldsymbol{\eta}} = \gamma_0 + \gamma_1 \texttt{NLAT} + \gamma_2 \texttt{ELON} + \gamma_3 \texttt{ELEV}.$$

where $\hat{\boldsymbol{\eta}}, \texttt{NLAT}, \texttt{ELON}, \texttt{ELEV}$ are a 17×1 vector and $\hat{\boldsymbol{\eta}}$ obtained from 2. Perform a t-test to determine whether all three spatial covariates should be included. Then, remove any insignificant covariate and use the F-test to assess whether the smaller model is better. Compare the results of the t-test and F-test. What conclusions can you make ?