Problem Set

STAT 380
UAEU

October 23, 2023

This is a Student's Activity Task containing a few multiple type questions and a number of descriptive type problems. The activity is not a graded component of the course. Its objective is to encourage students learning while solving problems and also to prepare for the upcoming exam.

Part I: Short answer type questions.

Must Review: Review All the Quiz Problems.

1. If we consider a Dataset that has 150 observations and 190 covariates (independent variables) to model a continuous response variable. Identify whether the following statement is True or False.
Statement: There will be no problem fitting a simple linear regression that can be used to obtain a predicted value of the response when the covariate values are available.

Ans: ☐ TURE ☐ FALSE

2. **Statement**: If it is possible to fit a Standard Linear Regression and a Ridge regression with L_2 penalty on the regression coefficients, then the absolute vakue of the estimated regression coefficients obtained from the Ridge Regression is always smaller than that of the Standard Linear Regression.

Ans: ☐ TURE ☐ FALSE

3. **Statement:** In the case of a LASSO regression that obtains the regression coefficients by minimizing the following objective function:

$$\|\underset{\sim}{\mathbf{y}} - \mathbf{X}\underset{\sim}{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \text{where } \lambda > 0.$$

The parameter $\lambda > 0$ is a tuning parameter that is often called a model selection parameter as well.
Statement: The optimal value of the tuning parameter is typically obtained using the AIC (or BIC, or Mallow's C_p) criteria.

Ans: TURE FALSE

4. Consider a dataset with 7 covariates (explanatory variables) to model a continuous response variable. Six among the seven covariates are numerical in nature while one of the covariate, 'Highest Education level' is categorical variable that can take either of the following four values:

$$\{\text{ High-School}, \text{Bachelor's degree}, \text{Master's Degree}, \text{PHD or Higher}\}$$

Note that, it is customary to introduce appropriate (multiple) 'dummy variables' as covariates to incorporate categorical. Categorical Covariates. How many total regression coefficients, including the intercept, are there in the constructed model ?

Ans: 7 11 10 9

5. When performing the Standard Linear Regression on a dataset to model a continuous response, a statistical analyst wants to verify whether the assumption of the Normality of the model errors are satisfied or not. Which one from the following statistical hypothesis tests are the most appropriate for the purpose.

Shapiro Wilk's Test χ^2-Test Durbin-Watson Test

6. One of the primary assumptions of the standard linear regression is the Normality of the corresponding errors. Based on the the visual inspection of the Normal-QQ plot (Theoretical Quantile vs Estimated Quantile) of the model residuals, a practitioner suspects that the Normality assumption may not be satisfied for the dataset. Therefore, he/she conducts the standard Shapiro-Wilks's Test for Normality of the corresponding residuals and obtained a p-value of 0.0017. What can be concluded from the above information:

Ans:

The Model Residuals Can Be Normally Distributed

There is a Strong Evidence to Reject the Fact that the Model Residuals are Normally Distributed.

7. When performing the Standard Linear Regression on a dataset to model a continuous response, a statistical analyst wants to verify whether the assumption of the independence of the model errors are satisfied or not. Which one from the following statistical hypothesis tests are the most appropriate for the purpose.

- Shapiro Wilk's Test
- χ^2-Test
- Durbin-Watson Test

8. One of the primary assumptions of the standard linear regression is the Normality of the corresponding errors. Based on the the visual inspection of the standard residual plot, a practitioner suspects that the assumption on the independence of the model errors may not be satisfied for the dataset. Therefore, he/she conducts the standard Durbin-Watson Test of the corresponding residuals and obtained a p-value of 0.37. What can be concluded from the above information:

Ans:

- The model residuals may not be correlated.
- There is a Strong Statistical Evidence that the residuals are correlated.
- There is a some Statistical Evidence that the residuals are correlated.

9. In the context of Standard Linear Regression to model a continuous response variable, how do we check if there is 'Influential' point in the dataset.

Ans:

- It is a Influential point if the corresponding Cook's Distance is more than 0.5.
- It is a Influential point if the corresponding Cook's Distance is less than 0.5.
- It is a Influential point if the corresponding Cook's Distance is less than 0.05.

10. If we consider a 3 degree regression splines with 7 knot points, then how many regression coefficient parameters are there in the model? (i.e. what is the dimension of the model?)

Ans:

- 20
- 11
- 9
- 10

11. If we consider a 3 degree regression splines (cubic splines) with 20 knot points, then how many regression coefficient parameters are there in the model? (i.e. what is the dimension of the model?)

Ans:

- 24
- 16
- 30
- 76

12. Let $\pi \in (0,1)$ such that $\text{Logit}(\pi) = 1.12$. What is the most appropriate value of π from the list below.

- 0.543
- 0.97
- 0.754
- 3.065

13. Let x be a real number such that $\text{Expit}(x) = 0.65$. What is the most appropriate value of x from the list below.

- 0.268
- -0.187
- 1.916
- 3.065

14. Let $\pi \in (0,1)$ such that $\text{Logit}(\pi) = 2.5$. What is the most appropriate value of $\text{Odds}(\pi)$ from the list below.

- 0.924
- 12.18
- 0.397
- 9.11

15. What is the value of the following function $\text{Logit}(0.8)$?

- 12.183
- 0.924
- 0.397
- 1.12

16. What is the value of the following function $\text{Expit}(2.5)$?

- 1.87
- 0.602
- 0.097
- 1.99

17. In a game of rolling a fair dice twice. A player wins if both throws result in sixes. What is the odds of winning the game?

Ans:
- $\frac{1}{36}$
- $\frac{1}{6}$
- $\frac{1}{35}$
- $\frac{35}{36}$

18. Imagine we have a sample of 150 points, 70 from 'Class 0' (For example: healthy) and the others from 'Class 1 (For Example: Not Healthy). Let us assume that a classification technique (Medical Testing) can correctly identify 65 out of the 70 belongs to the 'Class 0' (positive for all Diseased) while it can correctly detect 77 of objects that belongs to the 'Class1'(Not Healthy). Calculate the **Sensitivity** for the corresponding Classification Technique.

- 96.25%
- 91.31%
- 92.92%
- 93.90%

19. Imagine we have a sample of 150 points, 70 from 'Class 0' (For example: healthy) and the others from 'Class 1 (For Example: Not Healthy). Let us assume that a classification technique (Medical Testing) can correctly identify 65 out of the 70 belongs to the 'Class 0' (positive for all Diseased) while it can correctly detect 77 of objects that belongs to the 'Class1'(Not Healthy). Calculate the **Specificity** for the corresponding Classification Technique.

- 96.25%
- 95.59%
- 92.85%
- 93.90%

20. Imagine we have a sample of 150 points, 70 from ‘Class 0’ (For example: healthy) and the others from ‘Class 1 (For Example: Not Healthy). Let us assume that a classification technique (Medical Testing) can correctly identify 65 out of the 70 belongs to the ‘Class 0’ (positive for all Diseased) while it can correctly detect 77 of objects that belongs to the ’Class1’(Not Healthy). Calculate the **Yuden Index** for the corresponding Classification table.

- [ ] 0.891
- [ ] 0.938
- [ ] 0.929
- [ ] 1.86

21. Imagine we have a sample of 150 points, 70 from ‘Class 0’ (For example: healthy) and the others from ‘Class 1 (For Example: Not Healthy). Let us assume that a classification technique (Medical Testing) can correctly identify 65 out of the 70 belongs to the ‘Class 0’ (positive for all Diseased) while it can correctly detect 77 of objects that belongs to the ’Class1’(Not Healthy). Calculate the Yuden Index for the corresponding Classification table.

- [ ] 0.891
- [ ] 0.938
- [ ] 0.929
- [ ] 1.86

22. Consider analyzing a data-set that has a binary categorical response while all the covariates are numerical and continuous in nature. We know that a logistic regression and also a quadratic discriminant analysis is applied for modeling the response variable. To compare the performance of both the method a ROC curve is constructed and the corresponding Area Under the ROC Curve (AUC) is calculated based on their performance in a Testing set. The AUC for the logistic regression is 0.91 while the AUC for the QDA is 0.85. Identify whether the following statement is TRUE or FALSE. **Statement:** The Logistic regression is performing bettern then that of the QDA for this data set that is evaluated based on the testing set.

- [ ] TRUE
- [ ] FALSE

Part II: Descriptive Problems.

1. # Must Review: Review All the Quiz Problems.

2. Let us consider a Data-set that has a continuous response, Y and a numerical continuous covariate X. The observed data is provided as $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, where n is the number of observations. Assume that $n \geq 100$. Based on a preliminary analysis, it is evident that a nonlinear polynomial of the covariate would be a preferred choice for modeling the response. However, a major task is to identify the appropriate degree of the fitted polynomial.
   (a) Write a polynomial regression model of degree K for the response variable Y and the covariate X.
   (b) Provide a algorithm based on Cross-validation to select the optimal value for K.

Ans: The Dataset that has a continuous response, Y and a numerical continuous covariate X. The observed data is given as $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, where n is the number of observations. We are assuming that $n \geq 100$.

(a) The polynomial regression model for degree K, $K \geq 2$ is provided below:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_K X^K + \varepsilon$$

where the response is Y and the covariate is X. Note that, according to the model assumption, each data-point in the dataset satisfies the above equation. That is, If $\mathscr{D} = \{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$ denotes all the data point, then

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_K x_i^K + \varepsilon_i \text{ for } i = 1,2,\ldots,n,$$

where ε_i is assumed to be Normally distributed with mean 0 and constant variance σ^2. $\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n$ are assumed to be statistically independent to each other.

(b) A major problem in the polynomial regression is to select the degree of the polynomial equation based on minimizing the Cross Validated Errors. The basic idea of evaluating/calculation the model error for any Cross validation procedure is is the following:
- The Cross-validation (CV) is build upon the idea on partitioning the data randomly into folds, or non-overlapping sub-samples.
- Each time, one of the folds is used as the validation set and the remaining folds serve as the training set.
- The overall performance is computed by appropriately, combining the model's prediction error on each of the validation sets.

In the current context we apply the general principles of Cross Validation to compute the Cross-Validated Mean Square Error fro the polynomial model of each degree and choose the degree that corresponds to the minimum Cross Validated error. To describe the algorithm, let us use the following notation:

To perform a M Fold cross-validation, let $\mathscr{D}_1,\mathscr{D}_2,\ldots,\mathscr{D}_M$ is the random partition of the entire data. i.e. $\mathscr{D}_1 \cup \mathscr{D}_2 \cup \ldots \cup \mathscr{D}_M = \mathscr{D}$ and $\mathscr{D}_i \cap \mathscr{D}_j = \emptyset$ for $i \neq j$.

The following is the details of Algorithm:

Step 1: First partition the data in $\mathscr{D}_1,\mathscr{D}_2,\ldots,\mathscr{D}_M$. A typical procedure is to randomly permute the rows of the data and split it into M equal parts.

Step2: Select a value for k, (lets say $k = 3$ for example)

Step3: Calculate the Cross Validated MSE in the following way. For Each $m = 1, \ldots, M$.

A. Fit the k-degree polynomial model on all the data points **except** for the m^{th} validation set $\mathscr{D}_m$. If $k = 3$, fit the following model: $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$

B. Calculate the Predicted values for the response variable. Let $\widehat{y}_i$ denotes the corresponding predicted values for the response Y for all data-points that are in the m^{th} validation set $\mathscr{D}_m$.

C. Calculate the Cross-Validated Error for the m^{th} validation set as following

$$\mathrm{CV}_m(k) := \frac{1}{n_m} \sum_{i \in \mathscr{D}_m} (y_i - \widehat{y}_i)^2,$$

where n_m denotes the number of data points in the m^{th} validation set $\mathscr{D}_m$.

Calculate

$$CV(k) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{CV}_m(k).$$

Step4: Finally, when we calculate the $CV(k)$ for all possible candidate value of the degree k, lets say $k = 1, 2, 3, \ldots K$ then choose the optimal degree $k^\star$ that provides the minimum $CV(k^\star)$.

$$k^\star := \arg \min_{k \in 1,2,\ldots K} CV(k).$$

3. Let us consider a Dataset that has a continuous response, Y and a numerical continuous covariate X. The observed data is provided as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where n is the number of observations. Assume that $n \geq 100$. Based on a preliminary analysis, it is evident that a regression spline of degree 4. However, a major task is to identify the appropriate knots points for the fitted model.

   (a) Provide a algorithm based on Cross-validation to select the optimal knot-points.

4. Let us consider a Dataset that has a continuous response, Y and the numerical continuous covariates $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. The observed data is provided as $\{(\underset{\sim}{\mathbf{x}}_1, y_1), (\underset{\sim}{\mathbf{x}}_2, y_2), \ldots, (\underset{\sim}{\mathbf{x}}_n, y_n)\}$, where n is the number of observations. Assume that $n \geq 100$.
   (a) Write down the objective function of the LASSO regression model.
   (b) Write down the procedure to select the tuning parameter $\lambda > 0$.

Ans: The Dataset that has a continuous response, Y and a numerical continuous covariate X. The observed data is given as $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where n is the number of observations. We are assuming that $n \geq 100$.

(a) The Objective function for the lasso regression is:

LASSO Regression

Let $(y_i, \underset{\sim}{\mathbf{x}}_i) \in \mathbb{R} \times \mathbb{R}^p$ for $i = 1, \ldots n$ are observed data. A LASSO Regression estimator $\hat{\beta}_{\text{lasso},\lambda}$ is via following minimization problem:

$$\hat{\beta}_{\text{lasso},\lambda} = \text{Argmin}_{\underset{\sim}{\beta}} \|\mathbf{y} - \mathbf{X}\underset{\sim}{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| ,$$

$\lambda > 0$.

(b) A major issue in the LASSO regression is to select the tuning parameter $\lambda > 0$ by minimizing the corresponding Cross Validated Errors. The basic idea of evaluating/calculation the model error for any Cross validation procedure is is the following:
- The Cross-validation (CV) is build upon the idea on partitioning the data randomly into folds, or non-overlapping sub-samples.
- Each time, one of the folds is used as the validation set and the remaining folds serve as the training set.
- The overall performance is computed by appropriately, combining the model's prediction error on each of the validation sets.

In the current context we apply the general principles of Cross Validation to compute the Crosss-Validated Mean Square Error for the LASSO regression in choosing the optimal value of the tuning parameter $\lambda > 0$. To describe the algorithm, let us use the following notation:
To perform a M Fold crossvalidation, let $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_M$ is the random partition of the entire data. i.e. $\mathscr{D}_1 \cup \mathscr{D}_2 \cup \ldots \cup \mathscr{D}_M = \mathscr{D}$ and $\mathscr{D}_i \cap \mathscr{D}_j = \emptyset$ for $i \neq j$.
The following is the details of Algorithm:

Step 1: First partition the data in $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_M$. A typical procedure is to randomly permute the rows of the data and split it into M equal parts.

Step2: Select a value for λ, (lets say $k = 30.5$ for example)

Step3: Calculate the Cross Validated MSE in the following way. For Each $m = 1, \ldots, M$.

A. Fit the LASSO regression with the specific choice for Λ on all the data points **except** for the m^{th} validation set $\mathscr{D}_m$.
If $\lambda = 30.5$, we consider minimizing the following objective function:

$$\|\mathbf{y} - \mathbf{X}\underset{\sim}{\beta}\|^2 + 30.5 \sum_{j=1}^{p} |\beta_j| .$$

B. Calculate the Predicted values for the respons. Let $\widehat{y_i}$ denotes the corresponding predicted

values for the response Y for all data-points that are in the m^{th} validation set $\mathscr{D}_m$.

C. Calculatee the Cross-Validated Error for the m^{th} validation set as following

$$\mathrm{CV}_m(\lambda) := \frac{1}{n_m} \sum_{i \in \mathscr{D}_m} (y_i - \widehat{y}_i)^2,$$

where n_m denotes the number of data points in the m^{th} validation set $\mathscr{D}_m$.

Calculate

$$CV(\lambda) = \frac{1}{M} \sum_{m=1}^{M} \mathrm{CV}_m(\lambda).$$

Step4: Finally, when we calculate the $CV(\lambda)$ for all possible candidate values of λ , then choose the optimal $\lambda^\star$ that provides the minimum $CV(\lambda^\star)$.

$$\lambda^\star := \arg \min_{k \in 1,2,\ldots K} CV(\lambda).$$

5. Let us consider a Dataset that has a continuous response, Y and a numerical continuous covariate $\mathbf{X} = (X_1, X_2, \ldots, X_n)$. The observed data is provided as $\{(\underset{\sim}{\mathbf{x}}_1, y_1), (\underset{\sim}{\mathbf{x}}_2, y_2), \ldots, (\underset{\sim}{\mathbf{x}}_n, y_n)\}$, where n is the number of observations. Assume that $n \geq 100$.

   (a) Write down the objective function of a Ridge regression model.

   (b) For a given value of $\lambda > 0$, what is the estimated regression coefficients for the regression parameter $\underset{\sim}{\beta}$?

   (c) Write down the procedure to select the optimal value for the tuning parameter $\lambda > 0$.