Math 189: Final Project

Instructor: Zhuosong Zhang

May, 2022

1. As part of a study of its sheet metal assembly process, a major automobile manufacturer uses sensors that record the deviation from the nominal thickness (millimeters) at six locations on a car. The first four are measured when the car body is complete and the last two are measured on the underbody at an earlier stage of assembly. Data on 50 cars are given in the file `assembly.csv`.

   a) Report the sample mean vector and the sample covariance matrix of the six variables $(X_1, \ldots, X_6)$.

   b) Provide a 95% t-type confidence interval for $\theta = 3 \times E(Y_1) + 2 \times E(Y_2)$.

   c) Let μ_1 be the population mean of X_1. Test $H_0 : \mu_1 = 0$ versus $H_1 : \mu_1 < 0$. You need to provide the p-value or rejection region, and make your conclusion.

   d) Choose arbitrary two variables, say, (Y_1, Y_2), from the original six variables $(X_1, \ldots, X_6)$. Draw a scatter plot between Y_1 and Y_2 for these 50 observations.

   e) Let $\boldsymbol{\mu} = E(Y_1, Y_2)^T$. Define $\boldsymbol{\mu}_0 = (0, 0)^T$. Test whether $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ withe the significance level 0.05.

   f) Based on the data set as well as what you learned in the course, ask a related question other than a)–e). Please also try to solve your question.

2. Let $\boldsymbol{X} = (X_1, X_2, X_3)^T$ be $N_3(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with $\boldsymbol{\mu}^T = (2, -3, 1)$ and

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

   a) Find the distribution of $3X_1 - 2X_2 + X_3$.

   b) Find a number a such that X_2 and $X_1 + aX_3$ are independent.

   c) Let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{10}$ be a random sample of size 10 from the population $\boldsymbol{X}$. Specify the following distribution:

      i) the distribution of $\bar{\boldsymbol{X}}$, where $\bar{\boldsymbol{X}}$ is the sample mean vector;

      ii) the distribution of $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$;

      iii) the distribution of $n(\bar{\boldsymbol{X}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1} (\bar{\boldsymbol{X}} - \boldsymbol{\mu})$ where $\boldsymbol{S}$ is the sample covariance matrix.

3. In this question, we continue considering the `Carseats` data set in the `ISLR2` library.

   a) Fit a linear regression model for (`Sales ~ .`) and report the estimators of the intercept and coefficients of the variables.

   b) Analyze the residuals. Discuss whether the linear model is appropriate.

   c) Let p be the number of predictors in the model in a), and let $\beta_1, \ldots, \beta_p$ be the corresponding coefficients. Test whether $\beta_1 = \cdots = \beta_p = 0$ with significance level 0.05.

   d) Provide a 95% confidence intervals of β_1 and β_2 separately.

   e) Give the 95% Bonferroni-type confidence intervals for $\beta_1, \ldots, \beta_p$.

   f) Based on the full model in a), test whether $\beta_1 = \beta_2 = 0$ with significance 0.05.

   g) Randomly split the data into a training set and a test set by using your own random seed.

      i. Fit a ridge regression model and LASSO model on the training set, and choose λ by cross-validation.

      ii. Report the coefficients obtained in the final model, and compare the results in the Ridge regression and LASSO.

   h) Based on the data set as well as what you learned in the course, ask a related question other than a)–g). Please also answer your question.

4. In this question, we consider some theories in linear regression.

   Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a sample from (X, Y). **Assume that** $\sum_{i=1}^n x_i = 0$. Suppose that we want to fit the following linear model:

   $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

   where ϵ_i is the error term.

   a) Suppose the loss function is defined by

   $$L(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

   Find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimizes the loss function. (It suffices to write the mathematical expressions using $(x_i, y_i), 1 \leq i \leq n$.)

   b) Let $\lambda \geq 0$ be a fixed number. Define the loss function as

   $$L(\beta_0, \beta_1 | \lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda \beta_1^2.$$

   Find $\hat{\beta}_0^\lambda$ and $\hat{\beta}_1^\lambda$ that minimizes the loss function $L(\beta_0, \beta_1 | \lambda)$. (It suffices to write the mathematical expressions using $(x_i, y_i), 1 \leq i \leq n$.)

5. In this question, we will use a simulated data to perform nonlinear relation between X and Y.

   a) Use the `rt()` function to generate a predictor X of length $n = 200$, as well as a noise vector ϵ. You may choose your own `df` in `rt()`.

   b) Generate a response vector Y of length n according to the model

$$Y = a_0 + a_1 \sin(X) + a_2 \times \frac{\exp(2 \times \cos(X))}{1 + \exp(2 \times \cos(X))} + \epsilon.$$

   c) Perform a multiple linear model using Y and $X, X^2, \ldots, X^5$. Choose the best model containing the predictors $X, X^2, \ldots, X^5$ using forward stepwise selection.

   d) Plot the polynomial fits for a range of different polynomial degrees (say, from 1 to 20), and report the associated residual sum of squares. Perform cross-validation or another approach to select the optimal degree for the polynomial, and explain your results.

   e) Fit a natural spline model to predict Y using X using a range of degrees of freedom, and plot the resulting fits and report the resulting RSS. Describe the results obtained.

   f) Perform cross-validation in order to select the best degrees of freedom for a natural spline model on this data. Describe your results. Please avoid using `cv.glm`, but use your own codes to calculate the cross-validation error.

   g) Based on the data set as well as what you learned in the course, ask a related question other than a)–f). Please also answer your question.

6. In this question, we consider the bootstraping method by a simulated data set.

   a) Use `mvrnorm()` to generate a sequence of 2-dimensional normal random vectors, $(x_1, y_1), \ldots, (x_{20}, y_{20})$, where the mean vector is

$$\mu = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

   and the covariance matrix is

$$\Sigma = \begin{pmatrix} 2 & 0.4 \\ 0.4 & 1 \end{pmatrix}.$$

   b) Let ρ be the theoretical correlation between X and Y. Report ρ, and estimate the sample correlation r between X and Y using the data in a).

   c) Repeat a) and b) for $B = 1000$ times using different random seeds, and give the average of $r_1, \ldots, r_B$. Denote the average as $\bar{r}$. Find the difference between $\bar{r}$ and ρ.

   d) Provide a Bootstrap-type bias of $\hat{r}$ using $R = 1000$ bootstrap estimators. Compare the results in c) and d).

   e) Provide a Bootstrap-type variance of $\hat{r}$ using $R = 1000$ bootstrap estimators.

f) Based on the data set as well as what you learnt in the course, ask a related question other than a)–e). Explain why you ask this question. You do not have to answer the question.

7. In this problem, we consider the classification methods using the `College` data set in the `ISLR2` library.

a) Randomly split the data into a training set and a test set.

b) Perform logistic regression on the training data to predict `Private` using all other variables. Interpret the meaning of the coefficients matrix in the `summary` result.

c) What is the test error of the model obtained?

d) Repeat b) by LDA and QDA, separately.

e) Fit a SVM using a nonlinear kernel to the data. What is the test error of the model?

f) Comment on the results in b), c) and d).