University of Tehran
ECE Department

Statistical Inference

Spring 2020

INTRODUCTION

In this project, we intend to study and analyze a series of real datasets with what you learned in this course. To begin analyzing a dataset, the first step is to get familiar with it. In the first step, this acquaintance can be made by observing the features of the dataset and distribution of the values and visualizing the data to make initial guesses about it. In the next step, by performing statistical tests, we make sure our guesses are correct and make our claims with certainty.

Datasets Description

Dataset name	Filename	Description
Healthcare	health.csv	This dataset includes some information regarding the health situations of around 5000 individuals as well as how much they yearly spend on their health bills.
University admissions	admission.csv	This dataset contains some information about some students applying for university admissions.
Students’ performance	student.csv	This dataset includes the information about a sample of students studying in two different institutes as well as their grades in three different exams.

IMPORTANT NOTICES

- Use the R language in answering questions. Submit your codes in a separate file next to your report. Reports without R codes are pointless.
- In some datasets, you need to clean the data and convert the format and data type to more appropriate formats. So do this before answering the questions and explain the steps at the beginning of your report.
- If you need more categorical variables, you can add a new one to the dataset using some of your numerical variables. In this case, you need to describe the way you created the categorical variable from the numerical variable.
- In most of the questions, you must use the ggplot2 library to visualize and produce the desired charts.
- For each question, you need to fully explain your answer. An important part of the score will be attributed to your description. Drawing charts and performing calculations without sufficient explanations will result in losing the score. These descriptions show how much you understand the dataset. If you see interesting things in the diagrams, don't forget to mention them.
- When performing statistical tests, be sure to check the requirements for that test and write it down in your answer.

Question 1

Consider two categorical variables in your dataset such that at least one of them has more than 2 levels. Having these at hand, do the followings:

A. Derive a 95% confidence interval for the difference of these two variables and interpret it.

B. By hypothesis testing, determine if the two variables are independent or not.

Question 2

Choose a binary categorical variable and randomly select a small sample of your data (small sample size, e.g., $n \leq 15$). Then, perform a hypothesis test for the variable's success rate by means of the Simulation method.

Question 3

Answer the following questions:

A. Choose a categorical variable that has more than two levels, calculate its probability distribution. Then choose two samples of size 100 from your dataset. One of the samples should be randomly selected and the other should be biased on purpose. Compare each sample with the real distribution using χ^2 (goodness of fit) and interpret your results.

B. Pick up another categorical variable and compare it to the one you chose in part (a). Using the χ^2 test, check if the two variables are independent or not.

Question 4

From your dataset choose a numerical variable that predicts its future value is meaningful within the context of your dataset. next, choose two explanatory variables which you believe are the best predictors for your response variable:

A. Without building a model yet, which explanatory variable do you guess is the more significant predictor and why? (use your knowledge from phase 1)

B. for each explanatory variable:

   a. Compute the least squares regression.

   b. Write the predictive equation for the response variable and interpret its parameters.

   c. Draw a scatter plot of the relation between these two variables overlaid with this least-squares fit as a dashed line.

C. By using the previous part results, try to explain which explanatory variable is the more significant predictor.

D. Now, Compare your models, once using adjusted R2 and another time by ANOVA table. Explain results.

E. According to the results that you found in the previous parts, list the features of a good predictor.

F. Choose a random sample of 100 data points from the dataset.
    a. By 90 percent of data, Build two Linear Regression models and design hypothesis tests to see if these explanatory variables are a significant predictor of the response variable or not.

    b. Calculate the 95% confidence interval for the slope of the relationship between response variable and explanatory variables. Interpret these CIs.

    c. Use your models to predict the values of the response variable for the remaining percent of samples.

    d. Compare the predicted values with actuals. Report success rate.

Question 5

Consider the response variable you selected in the previous question. You can use as many explanatory variables as you deem necessary:

A. Plot a correlogram for explanatory variables and discuss the correlation between them. Could you find which explanatory variable plays a more significant role in prediction?

B. Develop a multiple linear regression model for the response variable using explanatory variables you found in part A.

C. What percent of the variation in the response variable is explained by the model?

D. How well do you think your model fits the data?

E. Now, Develop the "best" possible multiple linear regression model for the response variable using different approaches and metrics.

F. Check diagnostics for your model in part E (Three conditions: 1. Linearity, 2. Nearly normal residuals, and 3. Constant variability) and explain if this is a reliable model or not.

G. Use 5-fold cross-validation and compare the models' RMSE (part B and E). How do you interpret these values?

Question 6

Choose a binary categorical variable from your dataset as a response variable and choose various categorical and numerical variables that you guess might explain the response variable accurately.

A. Construct a logistic regression model and interpret the intercept and the slopes in terms of log odds and log odds ratio.

B. Choose a categorical variable in your model among the explanatory variables and plot the odds ratio curve for that variable. Interpret the plot.

C. Draw the ROC curve for the model. What does this diagram signify? Discuss the goodness of the model based on the AUC.

D. Which explanatory variable in the model plays the most significant role in the prediction? Why?

E. Select the explanatory variables with the most significant contribution to the model. Then, construct a new logistic regression model using these variables. Interpret the results.

F. Draw the utility curve for the model you've created in part E (define the utility of different outcomes yourself). What is the best threshold for this model?

Question 7

Healthcare Dataset

Create a new boolean variable called "high medical costs". For each individual person in the dataset, this variable will either be 1 or 0 depending on whether the medical costs for that specific person are higher than a certain threshold. This threshold can be something like the median of the health costs or anything you think might be reasonable. Create a logistic regression model to predict whether or not that person will incur high medical costs. Which variable has the most impact on the prediction?

University Admissions Dataset

Create a linear regression model that predicts the chance of admission for each applicant based on some other variables. In your model, which variable has the biggest effect on the chance of admission?

Students' Performance Dataset

Create a new boolean variable called "academic probation". For each student, this variable is 1 if the total score (G1+G2+G3) is below 25 and is 0 otherwise. Create a logistic regression model that, based on each student's characteristics, predicts whether or not that student is going to be on academic probation. Based on your model, which variable has the most effect on it?