Statistical Inference

Project Phase-2

Introduction

In this project, we intend to study and analyze a series of real datasets with what you learned in this course. The first step to begin analyzing a dataset is to get familiar with it. In the first step, this acquaintance can be made by observing the dataset features 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.

To answer each question, you have to fully explain the meaning of your analysis and interpret the generated plot and what you observe, even when it is not explicitly stated in the question. The more reasonable your analysis is, the more positive effect it has on the acquired grade of the corresponding question.

Note that there is no one way to solve each question correctly. Furthermore, whenever you need to do hypothesis testing, you must check all of the pre-requisite conditions (such as sample size, skewness, etc.). Finally, the validity of your results should be discussed.

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 how you created the categorical variable from the numerical variable.

- In most questions, you should use the ggplot2 library to visualize and produce the desired charts. You can find more about this elegant visualization package in references [1], [2].

- 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 for which at least one of them has more than two levels. Using these, follow these steps:

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: (Note: To answer the following questions, first implement them by yourself in R and then use R functions to solve them.)

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. (be sure to check the requirements for the test and write them down in your answer.)

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. Check the Linearity, Nearly Normal Residuals, and Constant Variability conditions in R.
    b. Compute the least squares regression.
    c. Write the predictive equation for the response variable and interpret its parameters.
    d. 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. How well do you think your model fits the data?

D. Develop the “best” possible multiple linear regression model for the response variable using different approaches and metrics.

E. Use 5-fold cross-validation and compare the model’s RMSE (part B and C). How do you interpret these values?

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

G. What percent of the variation in the response variable is explained by the model (part B and C)?

Question 6

Choose a binary categorical variable from your dataset as a response variable and choose several categorical and numerical variables which you think can best explain the response variable.

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

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

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

D. Calculate a 95% confidence interval for the odds ratio

Question 7

Please answer the following questions in respect to the model in the previous question.

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

B. Select another categorical variable except the response variable from the model, draw the OR (odd ratio) curve for this categorical variable and interpret the plot.

C. Select explanatory variables with the most meaningful roles in the model prediction, and construct the new Logistic Regression model, and then interpret the result.

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

References

[1] Intro to Data Visualization with R & ggplot2 (Link)

[2] Data visualization with R and ggplot2: the R Graph Gallery (Link)