Project Phase 2

Statistical Inference
Spring 2020

University of Tehran
ECE Department

INTRODUCTION

In phase 1, you got familiar with your dataset by performing statistical tests as well as plotting various diagrams. In this phase of the project, you'll study categorical variables and perform statistical tests on this type of data. Also you will work with different regression models and examine predictability of the data by these models.

IMPORTANT NOTICES

- Use the R language in answering questions. Submit your codes in a separate file next to your report. **Reports without R codes will not be graded**.
- 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 should 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 observe interesting patterns in the diagrams, don't forget to mention them.
- When performing statistical tests, be sure to check the requirements for that test and write them down in your report.

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 (n<15) and perform a hypothesis test for the variable's success rate by means of 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

Choose a numerical variable as a response variable, which predicting its future value is meaningful within the context of your dataset, and an explanatory variable which you believe is the best predictor for this response variable.

a. Compute the least squares regression.
b. Write the predictive equation for the response variable and interpret its parameters (including slope and intercept).
c. Draw a scatter plot of the relation between these two variables overlaid with this least-squares fit as a dashed line.
d. Choose a random sample of 27 data points from the dataset.
    1. Build a Linear Regression model for this sample and design a hypothesis test to see if this explanatory variable is a significant predictor of the response variable or not.
    2. Calculate the 95% confidence interval for the slope of the relationship between response variable and explanatory variable. Interpret this CI.
    3. Choose another explanatory variable you think will be useful and build the Linear Regression model for the response variable and these two explanatory variables. Compare this model with the previous model once using adjusted R^2 and another time by ANOVA table.

QUESTION 5

Consider the response variable you selected in the previous question.

a. Develop the “best” possible multiple linear regression model for the response variable once using Backward Elimination and the next time by Forward Selection.
b. Use 5-fold cross-validation and report the final model’s RMSE (Root Mean Squared Error). How do you interpret this value?
c. Check diagnostics for your model (Three conditions: 1. Linearity, 2. Nearly normal residuals, and 3. Constant variability) and explain is this a reliable model?
d. Plot a correlogram for explanatory variables and discuss the correlation between them. Which explanatory variable plays a more significant role in prediction?
e. What percent of variation in response variable is explained by the model?
f. How well do you think your model fits the data?

QUESTION 6

Select a binary categorical variable as response variable, and choose several explanatory variables (numerical and categorical), and then answer the following questions.

a. Construct a Logistic Regression model for these variables. Report the results.
b. Interpret and discuss about the intercept and the slopes in terms of log odds and log odds ratio.
c. Draw the ROC curve for the model. And compute the AUC of the model and then discuss the goodness of the model in terms of AUC.
d. Calculate 98% confidence intervals for the odd ratios.

QUESTION 7

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

a. Which explanatory variable in the model plays a more meaningful role in prediction? Explain your reason(s).
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. Select an appropriate threshold for the new model.
e. Draw utility curve, and then compare between the best threshold in this section and the threshold that is calculated in section “d”.

QUESTION 8

Answer this question based on the dataset assigned to you.

- IMDB Movies:

  Create a Boolean variable called "Award Winner" that is based on the number of awards won by that film, and if a film has received at least one award, this value is true. Then, using logistic regression, estimate the chance that a film will win a prize based solely on its genre. Which genre has the greatest impact on the chance of winning an award?

- Used Car:

  Using linear regression, create a model that measures a vehicle's fuel consumption based on vehicle technical variables such as gearbox type, engine size, year of manufacture, and more. Make your choice only from the variables that are related to the technical specifications of the car. For example, the price or color of a car cannot have a significant effect on its consumption. After creating the model, which factor has the most impact on car fuel consumption?

- Insurance:

  Create a new Boolean variable called "high medical costs". For medical expenses, consider a specific limit such as the mean or median, and the amount of this new variable is determined based on the boundaries. Using logistic regression, create a model that, based on each person's characteristics, predicts whether or not that person will incur high medical costs. Discuss which variable has the most impact on this prediction.

- Spotify:

Is it possible to predict the duration of a music track based on its characteristics? Using linear regression, tell if this is possible. Also, if possible, check which variable has the most impact on the duration of a music track?