Statistical Inference

Project Phase 2

University of Tehran

Spring 2019

Introduction

In the second phase of the project, we focus on topics discussed during the second half of the course. The questions are related to inference for categorical variables, linear and logistic regression. You should use the same dataset assigned to you in phase I.

For each question you answer, you have to fully explain the meaning of your analysis and interpret any generated plots and what you observe even when it is not explicitly stated in the question.

You must check the conditions for any test performed even when it is not explicitly stated in the question. You should also discuss whether the results of the test are valid or not.

You are free to manipulate your data, especially for questions 6 and 8. For example, you may convert numerical variables into categorical variables or even split the dataset based on a category and create separate models for each one. However, any changes to the dataset must be intuitive, and your reasoning should be explained properly.

Question 1 - Categorical Hypothesis Testing

Consider two categorical variables in your dataset such that at least one of them has more than 2 levels. Design and perform a hypothesis test to check if the variables are independent.

Question 2 - Sample Proportions

Choose a binary categorical variable (small sample size, e.g., n ≤ 15) and perform a hypothesis test for its success rate by means of the Simulation method.

Question 3 - Goodness of Fit

Choose a categorical variable that has more than two levels, calculate its probability distribution. Then choose two samples of size 100 from your dataset:

1. Randomly and without any bias
2. Bias the sampling method

Compare each sample with the real distribution using $\chi 2$ (goodness of fit) and interpret your results.

Question 4 - Bivariate Regression

Choose a numerical response variable that predicting its future value is meaningful within the context of your dataset. Select one explanatory variable which you believe is the best predictor for this response variable.

1. Compute the least squares regression.
2. Plot the relation between these two variables using a scatter plot overlaid with the least-squares fit as a dashed line.
3. What is the best predictive equation for the response variable? Interpret this formula using plain language.
4. How much of the total variability in the response variable is explained by the explanatory variable?

5. Compute the least squares regression of the explanatory variable on the response variable. What is the predictive equation? Are the two predictive equations equivalent? If not, what is the reason of the observed difference.
6. The least squares fitted line is a parametric method. The line is fitted according to an optimality criterion. However, various non-parametric methods exist for visualizing the relation between two variables. One such method is Locally Weighted Regression and Smoothing Scatterplots (LOWESS or LOESS).
    a. Briefly explain how LOESS works and compare it with ordinary least squares fit and state any benefits or shortcomings it may have for both bivariate and multiple regression.
    b. Draw a scatterplot of your two variables overlaid with the LOESS fit. Experiment with different span values.

Question 5 - Multiple Regression Part 1

1. Select another explanatory variable you think will be useful for predicting the response variable that you chose in previous question. Compute the least squares regression for the response variable using the two explanatory variables.
2. Compare this model with the model from the previous question using adjusted R2.
3. Compare the two models using an ANOVA table with the more complex model listed first.
    a. Calculate the proportional reduction in the residual sum of squares (RSS).
    b. What is the probability the two estimates of residual variance for the two models are equal?

Question 6 - Multiple Regression Part 2

1. In this question, you are to develop the best possible multiple linear regression model for the chosen response variable. You should explain what methods you used to derive this model. You can use as many explanatory variables as you deem necessary.
2. Use 5-fold cross-validation and report the final models RMSE.
3. Check the conditions for this regression model. Use relevant plots for each condition. Explain anything that may violate the conditions and whether you think they are important.
4. How much of the variability in the response variable is explained by the model? Do you think the model is suitable for predicting the response variable, i.e., can it be used in a real-world enterprise application?
5. Plot a correlogram for the chosen explanatory variables. Discuss any relations between these variables. Could correlations between explanatory variables be used for pruning them?

Question 7 - Parsimonious Models

Choose a numerical response variable and several categorical and numerical explanatory variables from your dataset (not less than 5 variables) and build a linear model for them.

1. Use the backward elimination method with regard to $adjusted\ R^2$ to attain the parsimonious model for the selected variables.
2. Use the forward selection method concerning p-value to attain the parsimonious model.

Question 8- Logistic Regression

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.

1. Construct a logistic regression model and interpret the intercept and the slopes in terms of log odds and log odds ratio.
2. Draw the ROC curve for the model.
3. Discuss the goodness of your obtained model in terms of AUC.
4. Which explanatory variable in the mode plays a more meaningful role in prediction?
5. Is there an instance of multicollinearity issue in your model? How do you explain that?
6. Calculate a 95% confidence interval for the odds ratio.