Statistical Inference - Spring 2018

Project Phase 2

University of Tehran

ECE Department

Khordad 1396

INTRODUCTION

Welcome back. In this phase of the project, we investigate several topics discussed during the second half of the course. Specifically, inference for categorical variables, linear regression, and logistic regression. You have to answer the questions by analyzing the same dataset selected for you in phase one.

P.S. 1: We know that some of the datasets are not quite straightforward to work with. So do not panic and go as far as you possibly can. The course staff will consider some coefficient of difficulty to make it up for you.

P.S. 2: In some of the questions you need to work with categorical variables having a reasonable number of distinct values (levels). In any case, if you see yourself in trouble finding one that suits your needs, feel free to build categorical variables out of other (maybe numerical) variables that meet the requirements.

P.S. 3: Do not forget to check the test conditions wherever needed.

Question 1.

Choose a categorical variable out of your dataset. Do the followings:

1. Design and perform a hypothesis test for this categorical variable.
2. Calculate the effect size in both default and alternative cases, followed by interpretation of this effect size. What could you say about practical significance and statistical significance here?

Question 2.

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:

1. Derive a 95% confidence interval for the difference of these two variables and interpret it.
2. By hypothesis testing, determine if the two variables are independent or not.

Question 3.

Choose a binary categorical variable with a small sample size ($n \leq 15$) and perform a hypothesis test for its success rate by means of **Simulation Method**.

Question 4.

a) In this question you are asked to perform a $\chi 2$ goodness-of-fit test on a categorical variable with more than 2 levels on two different samples of size 500 from the dataset. One of the samples should be randomly obtained and the other should be biased on purpose. Analyze the results and determine if each sample distribution on the chosen variable is different than that of the original dataset.

b) Pick up one more categorical variable and compare it to the one you chose in part (a). Using the $\chi 2$ test, check if the two variables are independent or not.

Question 5.

Come up with a research question that can be answered with a hypothesis test or a confidence interval, e.g. "Is there a difference in mean scores between categories A and B?" or "What is the average difference in scores between entities that do and do not have some specific characteristic?" This question could be used to shed some light on your choice of the "best" linear model (in question 7). Carry out the appropriate inference task to answer your question.

Question 6.

Choose a preferably interesting numerical response variable (meaning that, good prediction of that variable is of great value in the context of your dataset) and 5 (or more) other mixed (categorical and numerical) explanatory variables.

a) Without building a model yet, which explanatory variables do you guess are significant predictors and why?

b) Derive the **parsimonious** linear model once using Backwards Elimination and the next time by Forward Selection. Was your impression in part (a) precise?

Question 7.

Consider the response variable you selected in Question 6.

a) Develop the "best" multiple linear regression model to explain your chosen response variable.
b) What were your metrics for selecting the "best" model?
c) Analyze the residuals, do they follow a normal distribution as assumed earlier? (Hint: Use a Q-Q plot)
d) Is your model reliable? (Check three conditions: 1. linearity, 2. nearly normal residuals, and 3. constant variability)
e) Interpret the model's intercept and coefficients?
f) Discuss the correlation between explanatory variables.
g) Which explanatory variable plays a more significant role in prediction?
   i) Design and perform a hypothesis test for this variable to see if it is really a suitable predictor for the response variable.
   ii) Build up a 90% confidence interval for the coefficient of this variable.
h) Using 5-fold cross-validation, report final model's test RMSE (Root Mean Squared Error). How do you interpret this value?

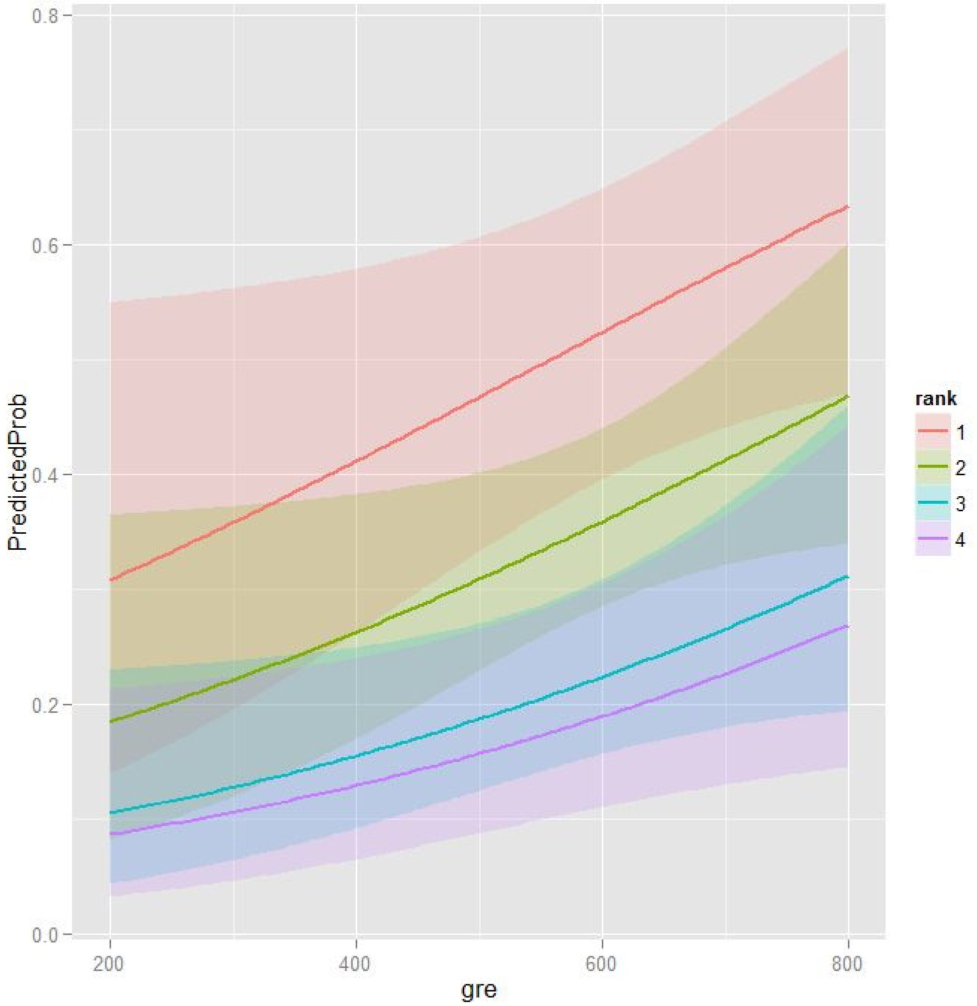
   (If you are not familiar with cross-validation, please refer to this OpenML webpage.)

i) What percent of variation in response variable is explained by the model?
j) How well do you think your model fits the data?
k) Is it the same model as in Question 6? How would you justify the difference?

Question 8.

Select a binary categorical variable (also an interesting one!) with 5 (or more) mixed explanatory variables.

a) Construct a Logistic Regression model.
b) Interpret the coefficients as odds ratios. How would you explain odds ratio for the intercept?
c) Obtain 98% confidence intervals for odds ratios.
d) Take a look at the following plot (taken from ref [2]):



In this part we are going to observe the effect of changing the value of explanatory variables on the predicted probability (of response variable). From explanatory variables, select a numerical (let's call it **N**) and a categorical variable (let's say **C**). Build a new dataset containing **N** and **C** (their values will be generated as described later) and other variables that their values are set to their mean (or mode). For each level of **C** generate a range of values for **N** (for example 1000 different values in range of [1, 100], note that these values totally depend on your numerical variable and it's range. Make sure you have generated enough different values). Afterwards, add another variable to the dataset using the **predict** method in R with the logistic regression model you made in part (a) passed as input. This variable is going to represent the predicted probability (regarding the response variable). Use ggplot2 to plot a graph for each level of **C** showing the value of **N** on x-axis and the value of predicted probability y-axis. Each line should be covered by a 95% CI band around it. Your final output will be much like the figure provided above.

e) Draw the ROC curve of the model of part (a) and compute the AUC. Interpret the AUC to justify if this model is considered a good model or not.

Question 9.

Considering the analysis you've done in the two phases of the project, what **story** do you think your data says? Does it have the potential to answer any interesting research questions? Did you answer any? Can your data be used to tackle a real world problem? Do you think of any further analysis that could help understanding data even more?
Tell us what it felt like analyzing the dataset you worked with.

REFERENCES

1. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.lm.html
2. https://stats.idre.ucla.edu/r/dae/logit-regression/