STA302/1001HF: Methods of Data Analysis I **Summer 2017**

Assignment 3 : Linear Regression Project

Out: June 6, 2017 *Due: June 22, 2017*

Reminder : You MUST write your solution independently and turn in your own write-up.

This assignment is due 11 :00pm, June 22, 2017. Submit your project report as instructed by Crowdmark. Late assignments will be subject to a deduction of 10% of the total marks for the assignment for each day late.

The project assignment may be summarized as follows : find an interesting and challenging data to analyze, and conduct a multiple linear regression to analyze it.

More specifically, you should begin to find a data set with multiple variables (at least 5 variables includes the dependent variable with data size more than 50). You will then investigate the relationship by constructing necessary plots, computing correlation and etc until you attain a final MLR (Multiple Linear Regression) model for a dependent variable (Y) that you are interested in.

The main aim of this analysis could be inspired by a research paper that you read, or an application related to your own study field, or a topic of general interest to you. You could focus on anything from statistical inference to bioinformatics applications to artificial intelligence to game playing to astronomy - **try to be creative**. Your analysis procedure does not have to be completely original, i.e. it can be related to topics discussed elsewhere, as you long as you cite this in your project. (Your project should not directly repeat material from another course or project, though it could be related. If you do make any use of results or programs or ideas from other sources or other courses, then this should be clearly explained.)

Finally, you should state your conclusions, regarding how to pick up important variables and do the variable selection until you attain a final model, and what can we learn from the final model.

The main part of your project should be **a maximum of 4 single-spaced pages**(or the default spacing 1.08 lines in Word. If you use Rmarkdown, set font size at 12 pt). However, it can also include an **Appendix of any length**, which may or may not be read by the grader. **Full source code and program output from R** you write should be included (perhaps in the Appendix, with appropriate summaries in the main part), with programs well commented and easy to follow. Your topic, motivation, analysisi procedure, and results, should all be very clearly explained within the main part of the project, with supplementary materials and additional explanations in the Appendix.

Extra references on building a model and finding data :

- The assignment 2 and 3 solutions from fall 2016 (available under Assignments on portal) might be helpful and useful to this project.
- Where to find data ? See the data repositories I listed in FindData.pdf. (I expect each individual has his or her own interest, so the data as well as the analysis should be different)

Project task

Below is the overall sequence of tasks you will follow to complete the linear regression project.

1. Choose a research questions that can be addressed using the data set you found online (remember to give link or reference to the data you used).
2. Carry out your analysis :
   — Conduct linear regression analysis
   — Write your results in a report, using the outline given below
3. Upload your written report in a PDF file. Due Thursday, June 22, 2017 (by 11pm). (You don't have to use Rmarkdown to produce your report although I highly recommend it to you.)

Project Report Outline

Below is a detailed outline of the content that should be included in your project report. The components are listed in outline form so that they can be used as a checklist. However, your project report is expected to be a formal paper (not an outline). Your results should be stated in complete sentences, and your paper should be written in paragraph form. Although you may choose to use headings, you should not number your paragraphs.

Introduction (one page maximum)

Introduction to the data and motivation of analysis. Please state the main aim of your data analysis or why you are interested in a specific data. In this part, you should also define clearly the variable(s) that are analyzing (e.g. age, salary, price, income, etc.). This must be specific : "time spent watching TV" is too vague ; "number of hours spent watching TV in the last 3 days" would be specific enough. If your variable is a measurement (e.g. height) give units (e.g. inches). If your variable is score (e.g on an achievement test), give the range of possible scores. Be clear on which variable was selected as the explanatory variable, and which the response variable ? Why ? What type of correlation (linear or nonlinear) did you expect ? And so on.

Grade

5 - Excellent : Strong evidence of original thinking and a clear introduction to the data and motivation of analysis.

4 - Good : Grasped the basics of the data ; a good introduction to the data and motivation of analysis.

3 - Adequate : Understood the basics of data and motivation of analysis.

2 - Marginal : Some evidence of understanding of the data and motivation of analysis.

1 - Inadequate : Little evidence of a good understanding of the data. Little explanation about how or why the data was chosen.

Analysis of the data (two pages maximum)

Include preliminary data analysis (such as correlation analysis among variables and so on), appropriate plots, model diagnostics and steps to arrive a final model.

Grade

10 - Excellent : Strong evidence of data analysis skills. Probably used R to do the analysis, steps, calculations and plots.

08 - Good : Good evidence of data analysis skills. Appropriate analysis, steps, and calculations were done, and maybe appropriate plots were included.

06 - Adequate : Understood the basics of required data analysis and have made some mistakes.

04 - Marginal : Some evidence of understanding the basic data analysis required, but fail to carry out all the appropriate analysis procedure, calculation and plots.

02 - Inadequate : Little evidence of even a superficial understanding of the data analysis required to analyze the data.

Conclusion (one page maximum)

What conclusions can you make based on the results of your final model or analysis ? Write a paragraph or two outlining these conclusions.

Grade

5 - Excellent : Conclusions are highly appropriate given the data analyzed. Clearly written.

4 - Good : Conclusions are appropriate given the data analyzed. Writing is good.

3 - Adequate : Some conclusions are appropriate ; other obvious conclusions might be missing.

2 - Marginal : Some evidence that there was an understanding of the basic conclusions, but several obvious conclusions not stated.

1 - Inadequate : Little evidence of even a superficial understanding of the conclusions that can be drawn from the analysis.

Appendix (any length)

Grade

2 - Clearly documented, programs well commented and easy to follow.

1 - Code is provided but hard to follow.