PSTAT 126 Final Project Instruction

Final Project Report Due Date: Saturday, September 15, 2018, 23:55 p.m. on GauchoSpace.

About the Project

The project is an opportunity to conduct a complete regression analysis of a real data set with what you have learned in this course. This will require all of the skills we have learned over the quarter: from plotting the data, to translating scientific questions into statistical regression terminology, answering those questions, and presenting the results in a concise manner. You can finish this project alone or find one or two partner(s). The project comprises 20% of the total course grade.

Again, there will be more than one viable model for a data set. The final model will be acceptable as long as you can justify it with what we've learned over the quarter. No matter what regression methods you will try to find a final model, it will be helpful to keep in mind that we prefer parsimonious models to overly complicated models. Before we use a linear regression model to answer questions of interest, the four conditions of a linear regression model must be met. Transformations are necessary when any of the four conditions is not met.

Choose Data Sets and Questions of Interest

First, you need to choose a data set from a public source. For example, you can choose one from the UC Irvine Machine Learning Repository. On this site, you can isolate data sets which are suitable for regression by clicking on the "Regression" link in the "Default Task" menu on the left hand side of the page. Once you have made your choice, you need to formulate one or more scientific questions which you will address in your regression analysis. Think about which variable(s) would be interesting as responses in a regression, and which other variables in the data set may be useful as predictors. Your report must include a thorough and appropriate regression analysis of two such scientifc questions.

Report Structure

The report should be prepared using 12pt font size with 1-inch margins, on U.S. letter paper, not including the appendix. The maximum page limit is 6 pages, not including the title page and appendix.

1. Title Page

Include the title of your project, and the names and section information of all group members.

2. Introduction (1 paragraph)

Briefly describe the data set of your project, state your research questions in terms of the data set and findings.

3. Questions of Interest

State clearly and concisely your questions of interest.

4. Regression Method

For each question of interest, state how you will attempt to answer that question using regression. There is no need to describe the mathematical or statistical details of this methods.

5. Regression Analysis, Results and Interpretation

In separate subsections, you should answer each of your questions of interest. Your narrative should include:

- Important Details of the Analysis: Depending on the questions you want to answer, this will include some from the following list: computing coefficient estimates, coefficient of determination R^2, p-values or test statistics, diagnostics, confidence intervals, prediction intervals, model selection procedures, etc. If you did a hypothesis test, then state your null and alternative hypothesis, the value of your test-statistic, p-value and your decision (you do not need to show all of your working – that should be included in the Appendix). Provide similar details for confidence intervals, sub-model tests, ANOVA tables, etc. Do not simply use every single method we've discussed in class. You will need to convince the reader that you have used the appropriate tools for answering the question of interest.
- Diagnostic Checks: Were your assumptions plausible? Why? How did you check them?
- Interpretation: What do your results mean for the questions you were trying to answer? You should give accurate and complete interpretations of your results in terms of the variables in a specific data set. The interpretations should involve a mix of statistical terminology, variable names, and appropriate units.

Relevant plots with proper title, variable names, legends, etc must be included within the body of the text. Useful plots include a scatterplot between two variables, a scatterplot matrix showing the pairwise relationship between variables, diagnostic plots of residuals, etc.

6. Conclusion (1 - 2 paragraphs)

In the conclusion section, you need to summarize your findings based on your final model in clearly understandable, non-statistical terms. What is the main message produced by your analysis. You may include any final comments and thoughts about your analysis. For example, is there any other possible way to improve the model such as to find predictors not in the data set? How general are your results, to what situations do they apply? Any other comments.

7. Appendix

In appendix, you need to include all the **R** code for your analysis. Optional contents include some analysis results or plots that you found interesting, but not of primary importance to your final analysis.

Grading Criteria

1. Well-Organized Writing

The format of the report should follow the **report structure**, starting from a title page to conclusion paragraph(s). Each section must have a clear title. Make sure that there is some margin between two consecutive sections. Title, variable names, legends, etc in a plot must be clear and easy to read.

2. Coherent Thought Process

To finish the project, you need to perform a complete analysis of the data including initial data analysis, such as making scatterplots with preliminary comments on the relationship between variables, regression modeling and diagnostics, a search for possible transformations and a consideration of model selection. You should include enough information for the steps leading to your selection of model. Your analysis should have a clear statement of the conclusion of your analysis.

3. Diversity of Methods

Instead of simply reporting estimates, standard errors, or p-values, etc, your analysis should demonstrate a wide understanding of the regression methodologies presented throughout the quarter.