STAT 330 Final Group Project Details

Project Report Due Date: Friday, December 15, 11:59 pm

About the Final Project

The final project is an opportunity to put all that you have learned in STAT 330 to the test. Your task is to conduct a start-to-finish regression analysis of a real dataset. This will require all of the skills we have learned over the semester: from plotting the data, to translating scientific questions into statistical regression terminology, answering those questions, and presenting the results in a clear, concise and engaging manner.

You should conduct your analysis and write your report in a Quarto document using the template provided. The next section describes how to work your way through the project step-by-step, and details of the final project write-up are given in the final section of this document.

The Final Project Workflow

1. **Create a Group:**

   You may create your own groups, but the size should be 2 or 3 people. If you want to be in a solo group, or a group of 4, you need to get permission from Dr. Sandholtz.

2. **Choose Dataset and Questions of Interest:**

   Your first task is to choose a dataset for which regression analysis would be useful. A great source for such data is 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 “Task” menu on the left hand side of the page. I suggest you find a few that look interesting and browse the variables they contain in order to make a final selection. Once you have made your choice, you need to formulate some scientific questions that you think would be interesting to explore through regression analysis. Think about which variable(s) would be interesting as a response in a regression model, 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 scientific questions**.

   If you are unsure whether a particular data set is suitable, particularly if the data do not come from a public source, please discuss your concerns with Dr. Sandholtz as soon as possible during office hours.

3. **Exploratory Analysis:**

   Exploratory plots of the data and numerical summaries are essential in beginning any analysis. Determine which techniques that you have learned in STAT 330 will best allow you to answer your scientific questions, and conduct appropriate exploratory analyses. At this stage, scatterplots, added variable plots, boxplots, etc. can give you a sense of relationships that exist between relevant variables. As necessary, explore potential transformations, identify outliers and influential points, etc.

4. **Conduct the Analysis:**

   Next, perform the appropriate analysis in `R`. Depending on the questions you want to answer, this will include various items from the following list: computing coefficient estimates, R^2 values, p-values or test statistics, confidence intervals, prediction intervals, model selection procedures and results, diagnostics, 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.

   REGARDING DIAGNOSTICS: The diagnostics shown in your report are used to show that the analyses are valid. In other words, there should be no blatant violation of the linear regression assumptions. Of course, it may take a while to arrive at a model with good diagnostics, requiring transformations for example. In this case, you may wish to include plots from preliminary models that showed assumption violations in the Appendix (see below), rather than in the main body of the paper.

5. **Interpret the Results:**

   After conducting the analysis, you should formulate concrete (i.e. data-specific), accurate and complete interpretations of your results. These interpretations should involve a mix of statistical terminology, variable names and appropriate scientific units. If you are using hypothesis tests, do not focus too much on $p < 0.05$ or any other significance level, but rather on how strongly (or weakly) the data serve as evidence against the null hypothesis.

6. **Final Concusions:**

   In your conclusions you should boil down your interpretations from the previous section into clearly understandable, nonstatistical terms. What is the main message produced by your analysis? There may also be additional questions that arise, problems you encounter, or possible extensions of your analysis that could be addressed here.

Grading Criteria

Your grade on the project will be based on the following criteria:

1. **Compatibility of Scientific Question and Regression** - Is linear regression a suitable tool for the scientific questions of interest? If so, did the choice of response and predictor variables match this goal?

2. **Coherent Thought Process** - Does the analysis indicate a sound understanding of regression methods? This is often best judged by the preliminary comments on the questions of interest, judgement of diagnostics, as well as the conclusions made after the analysis.

3. **Diversity of Methods** - Did the analysis demonstrate a wide understanding of the regression methodologies presented throughout the semester? Or did you simply report estimates and standard errors for the regression coefficients?

The Final Project Report

Each group will submit one report. The maximum page limit (excluding the title page and possibly an appendix) for the project is **8 pages**, though fewer may suffice. The report will have the following sections that will be included in the project template.

1. **Title Page:** Must include the title of your project, and the names of all group members.
2. **Abstract (100 words):** Brief summary of your project, questions of interest and findings.
3. **Problem and Motivation:** In this section you should describe (i) the relevant background of your topic, and, (ii) the motivation for your project i.e., why your readers should be interested. It should include subsections for each of the following:
   - **Data:** Briefly state the source of your data and describe the relevant variables that **you will use** in your project.
   - **Questions of Interest:** State clearly and concisely your questions of interest, written in scientific terms and not statistical ones. These should be stated in relation to the variables in the data set.
   - **Regression Methods:** 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 the methods.
4. **Analyses, Results and Interpretation:** In separate subsections, you should answer each of your questions of interest. Your narrative should include:
   - Analysis Details: This includes defining the model and the components of the method you are applying, as well as detailed, coherent code. For example, for a hypothesis test, state your null and alternative hypothesis and show the necessary R output to support the conclusions you will make. Provide similar detail for confidence intervals, reduced model tests, ANOVA tables, model selection procedures, etc.
   - Diagnostic Checks: What were your assumptions, and were they plausible? Why? How did you check them?
   - Interpretation: What do your results mean for the questions you were trying to answer?

   Relevant plots and figures must be included within the body of the text.
   Plots should not be too small to read the axis labes, title, etc.!
5. **Conclusions:** Brief summary of your findings. Also, you may include any final comments and thoughts about your project. For example, do you trust your results? How general are your results, to what situations do they apply? Any other comments.
6. **Contributions:** For each member of the group, write a short paragraph describing their contribution to this project and the report. It is up to you to decide how to divide up the work, so long as everyone is contributing approximately equally.
7. **Appendix (Optional):** Any exploratory data analysis from the project, or figures and plots that you found interesting, but not of primary importance to your final analysis. For example, this appendix is appropriate to show diagnostic plots, Box-Cox plots, etc., for preliminary regression models which were not used as they showed violations of model assumptions. This may not be looked at in detail when grading but could be useful for your own future reference.