STAT 439 Group Data Analysis Project

Spring 2022

1 Learning Objectives

The learning goals of the STAT 439 data analysis project are

- Formulate clear scientific research questions;
- Explore public data sources and find data that will answer specific questions;
- Recognize limitations of the available data and adapt research questions to the data at hand;
- Conduct a full data analysis with the goal of answering specific research questions;
- Summarize the data analysis process in a concise, organized, clearly written scientific report;
- Present report findings in an oral presentation;
- Develop communication and teamwork skills which will be useful in future careers.

2 General Instructions

Your assignment is to analyze a data set of your choice to best address your scientific questions of interest. Your final analysis will be presented in the form of a report, as well as a 10-15 minute in-class oral presentation. The required components, deadlines, and grading for the data analysis project are as follows:

Component	**Due Date (by 11pm)**	**Percent of project grade**
Data analysis proposal	Fri. Apr. 8	8%
Draft report	Fri. Apr. 22	5%
Peer assessments	Fri. Apr. 29	5%
Final report	Fri. May 6	50%
Oral presentations	Thur. May 12 8:00–9:50am	30%
Group evaluations	Thur. May 12	2%

You will be working in self-selected groups of 2–4 students. Group members are expected to contribute to the project equally and will evaluate each other at the end of the project. Each group member should contribute to all portions of the data analysis process. All group members will receive the same grade on the project. If you experience problems with your group members, talk to Prof. Hancock immediately so we can remedy the situation.

You may use any written or online references for this project that you wish (be sure to cite any references used!). Feel free to discuss your project with other students in the course, but do not ask for help or provide help on the data analysis of another group. If there are questions that you have about the analysis, feel free to contact Prof. Hancock.

3 Project Component Details

All project components will be submitted via Gradescope. Detailed instructions for each component of the project are below.

1. **Data analysis proposal (Due in Gradescope by 11:00pm 4/8)**. You are required to select your own data according to the following rules:

   - the data set cannot be part of your current or past research;
   - the data set cannot have been analyzed by other available sources (e.g., textbook, journal article, research groups);
   - the primary response variable must be either binary, binomial or Poisson;
   - the data set should include at least three explanatory variables of interest;
   - the data set should include at least 50 cases ($n \geq 50$).

   My data sources webpage may be a good starting place:
   http://www.math.montana.edu/shancock/data.html

   Your data analysis proposal should be $3 - 4$ pages, double-spaced, with 12 point font and one inch margins. Sections of the proposal should include: 1. Background, 2. Data set description, 3. Scientific goals and primary questions of interest, 4. Preliminary data exploration (i.e., summary statistics and plots), and 5. Analysis plan and modeling. Cite any references used. The goal of the proposal is to give me an idea of the direction of your analysis, and give me an opportunity to give you feedback on your analysis before the draft report is due.

2. **Data analysis draft report (Due in Gradescope by 11:00pm 4/22)**. You will submit an anonymized draft report (see report guidelines below). Do not include your names on the draft report. Drafts will be assessed by your peers.

3. **Peer assessments (Due in Gradescope by 11:00pm 4/22**. Each individual will receive one draft report to assess and provide feedback. You may work on these together as a group, or individually. An assessment/feedback form will be provided.

4. **Final report (Due in Gradescope by 11:00pm 5/1)**. Your data analysis report may NOT exceed **ten pages**, including any plots and tables. Your report should be double-spaced with 12 point font and one inch margins. References and appendices (if needed) do not count towards the page limit. You will also turn in a separate R Markdown file with your analysis. Detailed instructions for the format of the report are below.

5. **Presentation**. Your group will present your findings in a 10-15 minute class presentation during our final examination time, 8:00–9:50am on Thursday, May 12 in our usual classroom.

6. **Group evaluation (Due in Gradescope by 11:00pm 5/12)**. You will evaluate the level of participation of your group members through a Gradescope Quiz.

4 Data Analysis Strategies

- Perform adequate exploratory analysis of the data and provide a complete, yet succinct, presentation of the results including both summary statistics and plots that are relevant to the research questions.
- Clearly state the model building/selection/validation criteria used to address the scientific question(s) of interest.
- Clearly state the statistical model equation before presenting model estimates.
- Perform adequate model diagnostics.
- Provide precise interpretations of the relevant estimated parameters and their interval estimates in the context of the scientific problem, including scope of inference, and assessing statistical and practical significance.

5 Data Analysis Report Guidelines

Your data analysis report should describe the results of your analysis and the conclusions you would reach from those results. This report should look like a formal report to a statistically naïve researcher or an interested lay person. Because a statistical analysis aims to answer a scientific question, you should organize your report in the manner which is customarily used in science:

1. **Abstract**: Provide a concise description of the question, the data used to try to answer it, and the conclusions of your analysis. Only give the most pertinent estimates, confidence intervals, and p-values, if needed. Don't give too much detail here, but do note any significant problems that were encountered.
2. **Background/Introduction**: Provide a description of the scientific motivation for the analysis and relevant background literature. You don't have to go into great detail here, but do give all the facts that entered into your decision process during the analysis. List your specific questions research questions of interest as well as the questions that you were able to answer with available data (if different).
3. **Methods**:
   (a) **Source of the Data**: Describe the source and sampling methods for the data, if known. Describe the variables that are available and their meaning for the analysis. Highlight patterns of missing data as well as possible confounding by measured or unmeasured variables. This should not be a detailed presentation of descriptive statistics, however. That will come under Results.
   (b) **Statistical Methods**: Describe the methods used for the analysis at two levels. 1) Give a low-level technical description of the analysis for a potential reader. Include references for non-standard techniques. You may want to describe the software used, and certainly want to describe the methods used for assessing the appropriateness of your models. 2) Explain the basic philosophy behind the analysis techniques in layman's terms. Explain why you didn't use more common techniques if applicable.

4. **Results**: Provide the pertinent results of your analyses. Do not include all the dead-end analyses you might have done unless they provide insight into the question. Do lead the reader up to the analyses gradually.

   (a) Start off with descriptive statistics. The goal is to describe the basic characteristics of the sample used to address the question, as well as to present simple descriptive statistics (non-model based) that address the questions. Tables and plots are the key tools. If there are any characteristics of the data that present technical problems that needed to be addressed in the modeling, try to present descriptive statistics illustrating those issues. The basic idea is to presage all the issues you will talk about when presenting the models used in statistical inference, insofar as possible with simple descriptive statistics.

   (b) Then go to the major models used to answer the primary questions. Present summaries of the statistical inference obtained from these models (point estimates, confidence intervals, p-values). Highlight any particular issues that materially affected the models used to answer the question (confounding, interactions, nonlinearities, etc.) Tables can often be used to good effect here. Provide interpretations for all parameter estimates of interest.

   (c) Leave exploratory analyses (if any) for last and highlight the exploratory nature of those analyses. Present the results of your analyses in tables and publishing quality figures.

   Do not include output from statistical programs in the body of your report. (Such means little to me and nothing to a reader). When possible, use words instead of cryptic variable names. Use forms of estimates that have some meaning to a statistically naïve researcher. For example, present odds ratios rather than logistic regression parameters. Present confidence intervals rather than the values of Z, t, F, or χ^2 statistics.

5. **Discussion**: Discuss the conclusions which you feel can be drawn from the analyses, including scope of inference, and assessing statistical and practical significance. Suggest directions for future studies and analyses. Highlight the limitations of the data and your analyses.

6. **Appendix**: (*if needed*) Anything of an overly technical nature, or additional plots and tables may be included in an Appendix. The Appendix does not count towards the page limit.

7. **R Code**: A separate R Markdown file should be submitted with your report that outlines the details of your analysis process and includes relevant code used to generate plots, summary statistics, and model estimates.

The major theme of the above is to write to the scientific community rather than to a statistician. If you cannot explain your findings in a straightforward manner, then the analysis is of little value to anyone.

Lead your reader to all the proper results. You spent a long time analyzing the data. Now provide a brief tour through the high points of your work. Statistical diagnostics, which take a lot of our time, can most often be summarized in a single sentence (e.g., "We found no evidence to suggest that the final model did not fit the data adequately.") with relevant diagnostics relegated to an appendix. You are reporting your major results and impressions of the data. If the reader wanted to see every detail, he/she would have to do the analysis himself/herself.

6 Additional Resources

Your report should be well written, organized, and succinct. Dan Gillen's "Organizing Your Approach to a Data Analysis" as well as Givens' and Hoeting's "Communicating Statistical Results", both posted in D2L, will be helpful.