STAT 400 Statistical Modeling II – Project Information

(50 Points)

This is an opportunity for you to apply methods and tools discussed in STAT 400 to work on a problem or data set that interests you. You will need to find (at least one) data set to address your questions of interest as described below. If you are not sure where to look, some good sources are <https://www.kaggle.com/datasets>, <http://www.statsci.org/datasets.html>, <https://data.world>, and <https://data.gov>. If you have data from other classes or your own research that you think would be appropriate for this class, you may use those too. You may ***not*** use any data from the textbook authors’ (Roback and Legler) site AND you may not use data from the textbook you used in the prerequisite course (STAT 300/STAT 462).

The project is broken into 2 parts: GLM’s and Multilevel Modeling. In each part, you will ask research questions, perform an explanatory data analysis, and thoroughly analyze your model to answer the research question. The goal of this project is to demonstrate that you understand the steps for analyzing the dataset/model and that you can effectively communicate your decisions and analysis. While I do not tell you the exact steps that should be taken for the analysis, your goal is to demonstrate that you understand the process and analysis.

For all parts of the project, students are to work in groups of 3-4. Once the groups are finalized, you may not change partners. If you do not pick a partner, you will be randomly assigned to a group.

A breakdown of the grading is as follows:

* Part 1 (GLMs) – 25 Points:
  + Meet Preliminary Milestone Deadlines (10 Points)
  + Write a final report that summarizes the complete analysis (15 Points)
* Part 2 (Multilevel Models) – 25 Points
  + Meet Preliminary Milestone Deadline (10 Points)
  + Write a final report that summarizes the complete analysis (15 Points)
* Review of Group Members
  + All team members will be asked to evaluate the performance of their teammates.
  + If your teammates indicate that you have not participated, do not understand the material, or did not contribute to the project, the instructor reserves the right to lower your project grade. This could result in a minor or significant reduction to your grade depending on the situation.

# Preliminary Milestones

## Part 1 (GLM’s) Preliminary Milestones

* Due Friday, March 29 – One group member should send an email to the instructor (yvy5509@psu.edu). The email should mention the names of all group members, and the summarization of project goal.
* Due Friday, April 12 – Each group should submit a data set suitable for building a generalized linear model (GLM). Describe the variable to be used as the response, the predictor variables (including whether each predictor is quantitative/categorical (with how many levels)), and your research question. The variable used as the response should be either Poisson, binomial, or multinomial. There should also be at least three variables remaining to be used as predictors/input variables. Describe where you have found the dataset (provide a link if appropriate) and submit the dataset as a .csv file in Canvas. Also, submit your description of the variables and research question in Canvas. Each group member should submit, but please submit identical documents.
* Due Friday, April 19 – Read the dataset into R, perform any data cleaning (such as recoding categorical variables and describe the encoding (1 = yes, 0 = no, etc.), handling missing values, etc.) and summarize these results with the head() , glimpse(), or str() functions. Then, being to perform your exploratory data analysis (EDA). The R code required to complete these steps should be saved in an R script or Rmd file and included as part of your submission. It is fine to do you write using R markdown. If so, submit both the .Rmd and .html files. NOTE: For this deadline, you do NOT have to complete the full analysis of the model. You simply have to complete the data wrangling (and explain any data cleaning choices you make), summarize the results, and make some progress with your EDA. Each group member should submit, but please submit identical documents.

## Part 2 (Multilevel Modeling) Preliminary Milestone

* Due Friday, April 26 – Mention which dataset you have chosen and state your intended research question (or research questions if you have more than one). The research question should be involved enough that you will consider at least one level 1 covariate and at least one level 2 covariate. Identify the response, the level 1 and 2 observational units, the level 1 and 2 covariates, the fixed effect(s), and the random effect(s).

# Project Reports and Presentations

After completing the preliminary milestones listed above for each part listed above, you will perform a complete analysis related to the research question. A complete analysis includes a thorough EDA, model analysis, thoughtful discussion, and answers to your research questions. At the completion of the project, there are three project deliverables:

1. Presentation
2. Final Report
3. Review of Group Members

# **Besides preliminary milestones, you are highly encouraged to use more advanced models and tools in final project:**

1. Advanced machine learning/deep learning methods outside this course
2. Visualization/Demo
3. In-depth analysis using domain knowledge

## The grade of final project can be significantly increase based on the quality of the advanced analysis (also bonus points). If the final project is promising and you are interested in publishing the results in journal/conference, I am happy to discuss with team further after this course.

## Final Report (due May 3rd)

This is essentially the main analysis and model-building steps, but you should also include the work you’ve done in the earlier parts. (This means I should be able to read your final report and understand all aspects of your project. If I must refer to earlier submissions to understand what you have done, the report will be viewed as missing important information.) The format should be either html, doc, of pdf and should be organized in two parts: one chapter for the GLM and one chapter for the multilevel model. Add an appendix if you are including any other relevant items. You may include R code in the document itself (helpful if you’re working from an Rmd file), but you should not include any raw data or other lengthy output that distracts from the main content; attach those separately if applicable. (If you are not using a markdown document, you are still expected to submit your code in the appendix or as a separate file.) The grading will be based on the appropriateness of your analysis and your explanation of what you try, not necessarily on finding the "correct” model.

For the GLM modeling chapter, the main components are:

1. Some EDA that describe the variables of interest. If you’re considering a lot, you may choose only a few representatives for this.

* Every plot you include should be labeled or explained clearly. What characteristics are you trying to illustrate?
* Be aware of how your plots appear in the final document. If too many plots are forced in one panel, the detail may be lost.
* Plots should be near the narrative text that describes them. Avoid putting all plots in the appendix and forcing your audience to endlessly flip to connect your explanations to you visualizations.
* Avoid including plots/output that is not discussed in your narrative.

1. Statistical analysis and conclusions. This includes variable selection methods and justification for your final model, interpretation of model parameters, and conclusions for your research questions of interest.

* Keep answers clear and concise. Making two conflicting statements may lead to a deduction, even if one of them is correct.
* Refer to specific parts of the output to support your comments. For example, if the output has multiple p-values displayed, it should be clear which one you’re referring to.

For the Multilevel modeling chapters, the main components are:

1. Some EDA that describe the variables of interest. The EDA should reflect the response, level 1, and level 2 variables of interest. If you’re considering a lot, you may choose only a few representative plots for this.

* Every plot you include should be labeled or explained clearly. What characteristics are you trying to illustrate?
* Be aware of how your plots appear in the final document. If too many plots are forced in one panel, the detail may be lost.
* Plots should be near the narrative text that describes them. Avoid putting all plots in the appendix and forcing your audience to endlessly flip to connect your explanations to you visualizations.
* Avoid including plots/output that is not discussed in your narrative.

1. Statistical analysis and conclusions. This includes demonstrating an understanding of model representations (i.e., write out your models), the model building approach for multilevel data, interpretation of model parameters, comparisons of models, and conclusions for your research question(s) of interest. A big part of this task is demonstrating that you understand the process of multilevel moding

* Keep answers clear and concise. Making two conflicting statements may lead to a deduction, even if one of them is correct.
* Refer to specific parts of the output to support your comments. For example, if the output has multiple p-values displayed, it should be clear which one you’re referring to.

NOTE: Part of the grades (for both EDA and the modeling in both parts of the project) will be based on the quality/professionalism of your report. One sentence answers are of a lower quality than thoughtful paragraphs. Poorly labeled plots and typos are of a lower quality than a carefully proofread document written with one coherent voice. Doing a minimal analysis is of lower quality than a thorough analysis. Etc. Show off what you have learned!!!

## Final Presentations (Lecture time in week before Final Week: April 22 - April 26)

This part involves a short 11-13 minutes summary presentation of your project **in class**. Be sure to describe the dataset, variables of interest, research question(s), relevant EDA, the analysis of your model, and your conclusions. Each group member must participate in the presentation.

Some notes:

* I am expecting your group to create a professional looking presentation (likely using some type of slides) to discuss your project.
  + A presentation is different than a video of you reading your paper to me.
* All group members must participate equally during the presentation.
  + I expect all group members to present during the presentation.
  + Participation should be balanced.
  + In other words, avoid having one person speak for 8 minutes and the others speak for 30 seconds.
* When presenting project, please consider the following questions:
  + Did we describe the dataset and explain the variables of interest clearly?
  + Did we state and answer our research question(s)?
  + Is the analysis appropriate, complete, and correct?
  + Did all group members participate equally?
  + Did the presentation look and sound professional? Was the presentation of a high quality?

# Review of Group Members (due May 3rd)

At the conclusion of the project, each group member will be asked to rate the contributions of the other group members. Specifically, you should evaluate your teammates in terms of participation (did they come to meetings, were the responsive to group discussions, etc.) and contribution (did they understand the material, offer insights, add to the quality of the overall product, etc.). Your group evaluation should also discuss the contributions of each team member. Finally, I am expecting a thoughtful reflection of the group experience. The reflection should be much more detailed than just saying “All group members were great.” or “10/10 for everyone.” I want you to take time to actually reflect on the experience and contributions.

Each person in a group should submit their own group evaluation of their teammates. If you have not participated or contributed to the group, I reserve the right to lower your project grade.