**STAT 3220 - Spring 2023**

**Final Project (25% of grade)**

# Purpose

Often times in “the real world” when you are given a data set or are trying to answer a research question, you may not be equipped with the analysis skills to answer the question. You will have to research and learn a technique on your own. Additionally, you will need to communicate the understanding of the new technique and your analyses to any audience. Throughout the semester, you have had opportunities to develop the skills to be able to do just that. The purpose of the final project is to demonstrate your ability to learn a new analysis technique and present an analysis both verbally and in writing. Academic poster sessions are common-place events at most research fairs and conferences. So developing a poster, and presenting at these sessions are useful skills. To demonstrate your ability to communicate your findings orally and visually, you are to present your research and findings at our STAT 3220 poster presentation.

# Project Scope

You will work in a group of 3-4 students to research a question of interest to you and your group. You will be responsible for finding your own data set, establishing a research question, applying the techniques we’ve covered, while extending beyond the course, and presenting your findings. It is your responsibility to decide upon your group dynamic. You will perform an analysis using your data from project 1 or selecting new data that meets the project 1 expectations. You will choose to perform a MLR or logistic regression analysis. The following criteria are required for whichever option you choose.

You will include the following components in your poster presentation and analysis:

* **Introduction**: Brief summary of the research question you intend to answer. This section should include motivation behind your research question (why it is interesting and relevant) and what your question is. You will want to follow the structure of one general question and 2-3 specific research questions questions. You will include at least two relevant sources to support your claims and topic of interest. This may include aspects of part 1, if relevant.
* **Data Summary:** This section will include the relevant information about your data. And the important exploratory data analysis. This may include aspects of part 1, if relevant.
* **Analysis- see next section for specifics:** You will perform a multiple linear regression OR logistic regression for your data. You should summarize your model building and highlight important findings about your variables of interest. You should include relevant plots and limited output. Including the following components:
  + Include required analysis steps
  + Include the ”added techniques” that you selected
  + Assessing the model.
  + Selecting a final ”best” model.
* **Conclusion:** Your conclusion should consist of 2 parts: 1) interpreting your results of the analyses in context of the problem and 2) commenting on areas of future improvements. Your interpretation would include the actual prediction equation, comments on its usefulness for prediction, and an example of actually using the model for prediction or estimation. The second part you would comment on a limitations of this analysis and your data and how you would continue or improve on this research in the future.
* **Works Cited:** You should include any references you used in your project. This can be submitted directly on your poster, or an added pdf appendix submission to your presentation page.
* **Formatting:** Your poster should include summarized components from each of the following sections. Remember, there should not be paragraphs, instead using bulleted lists, tables, and relevant graphs to display and communicate your findings.

**Your written report will include the following:**

* The submitted report should be 3-pages and may include additional appendices (see below). You should use 1 inch margins, Times New Roman font 12 point, and single spacing, and page numbers. You should outline the following sections with headings.
  + **Introduction**: The research question(s) & Background/significance of the research.
  + **Methods/Analysis**: Describe the analysis process. This would include specific techniques and appropriate testing languages (appropriate order of analysis, assumptions, etc)
  + **Results**: The results of the final model. This should be in statistical terms and overall interpreting and assessing the statistical usefulness of the model. You should limit the output here (and maybe do not include anything). This would include your final model.
  + **Conclusions:** Your conclusion should consist of 2 parts: 1) interpreting your results of the analyses in context of the problem and 2) commenting on areas of future improvements. Your interpretation would include the actual prediction equation, comments on its usefulness for prediction, and an example of actually using the model for prediction or estimation. The second part you would comment on a limitations of this analysis and your data and how you would continue or improve on this research in the future.
* In addition to the 3-page paper, will have a **title page** which includes:
  + The title of the project and an image that best represents your project
* **References** should be listed at the end of the paper on a new page and do not count against the 3-page limit.
* (Optional) You may include up to 5 additional pages of information about your project in an appendix  This could include secondary analysis results, charts/tables, etc. The optional appendix **may** be reviewed if questions arise when reading the 3-page paper, but will not be considered primary content for scoring.  There is no guarantee of this additional review; thus no information deemed critical to the evaluation of your project should be included in the Appendix.

Analysis Choices + Requirements: You will select multiple linear regression or logistic regression as your analysis technique.

## Multiple Linear Regression Analysis Requirement

* Model building with significance testing (should be supported by EDA)
* Check for regression assumptions. Correct if violated (you only need to try up to three corrections, if the first doesn’t work)
* Identify and check for multicollinearity
* ADDED ANYWHERE IN YOUR ANALYSIS THAT IS RELEVANT:
  + Include at least THREE additional techniques from below to add to your analysis: Weighted least squares (Ch 9.4), External Model validation (Ch 5.11), variable screening (CH 6), influential observation analysis (Ch 8.6), box -cox transformation (supplemental), or another technique or aspect of MLR you learned on your own.
  + OR COMPARE your model to a model created from ridge or LASSO see Ch 9 and additional resources
  + OR COMPARE to Poisson regression (for “count” data)
* Final model selected should be assessed. It may not be great, but you will explain that in conclusion

## Logistic Regression Analysis Requirements

* Model building with significance testing (should be supported by EDA)
* Perform a minimum of goodness-of fit test
* Assess with metrics for association
* ADDED ANYWHERE IN YOUR ANALYSIS THAT IS RELEVANT:
  + Include at least THREE additional technique from: Check for multicollinearity, perform variable screening, check and correct for influential observations with diagnostic plots, assess for overdispersion or something else you learn in your research
  + OR COMPARE to a survival analysis.
* Final model selected should be assessed. It may not be great, but you will explain that in conclusion

# Logistics

## Due Thursday, April 27th at 11:00am (as a group, for all sections)

* **PDF of Poster:** Submit a PDF of your poster to gradescope- this is due for all groups.

## Poster Presenting April 27 + May 2 during class time:

* The purpose of the poster session is to present your results and findings of your research question, with less emphasis on “teaching” the analysis. You should be prepared to answer questions about the technique, but you should focus your presentation on answering your research question. Your audience is other students in this class. You should prepare a 10-12 minute presentation.
* You will print 5 copies of your poster on standard printer paper (in color or black and white). You can also use a laptop to zoom into plots and figures that are not as readable. You will submit one copy of your poster (regular printer paper size) to your reviewer.

## Due Tuesday, May 2nd at 11:59pm

* **Executive Summary Paper:** Submit 3 page written report to Gradescope. One per group.
* **Exit Survey (Quiz 10- submitted individually):** Submitted to Gradescope about your group work and course reflection.

## Scoring: 300 Points

|  |  |  |  |
| --- | --- | --- | --- |
| **What?** | **How?** | **Who?** | **How much?** |
| **Poster + Paper Includes Appropriate Summaries of Components** | Highlighted findings from Background, Data Summary, Analysis, and Conclusion all written coherently. The focus is on the research question, supported by the analysis. You do not need to include every single component of your analysis on the poster. | Scored by evaluators (Professor, TA, GTA)  100 Poster  50 Paper | **150 points**  **See** [**Instructor Rubric**](#_Project_2_Instructor_1) |
| **Presentation of poster:**  **Visual Aesthetic/ Formatting of Poster**  **Presentation** | There is not too much text or bulleted lists. Relevant graphs and related images are included. Overall, the poster is professional.  You demonstrate that you are knowledgeable about your topic and research question. Your presenting is clear and energetic | Average rating from your reviewers and me  \*If you are not present on your day to presentation, you will earn a 0 for this portion. | **30 points**  **70 points**  **See** [**Peer Review Rubric**](#_Peer_Review_Rubric:) |
| **Attending your presentation day** | You are able to answer questions by your reviewers | **You**  \*If you are not present on your day to presentation, you will earn a 0 for this portion. | **20 Points** |
| **Evaluation of other groups** | You will submit a formal evaluation of 2 other groups during the poster session | You submit thoughtful, appropriate feedback | **30 points** |

## Project 2 Instructor Rubric

**The following components should be addressed within your presentation, whether written on the poster or spoken. Remember the poster is an aide in the oral presentation.**

Components (Total = 150 points)

|  |  |  |
| --- | --- | --- |
|  | **Total Points Worth** | **Points Given** |
| **Introduction:**   * Acquaint the reader with the rationale behind the research question you intend to answer, with the intention of defending it. * Discussed questions like: Why is this question important, who might it impact, why do you have personal interest, etc * A clear benefit, or novelty to your research question is discussed * You should site at least two relevant works about the content topic * Briefly discuss the statistical analysis you intend to use to evaluate the question | 10 |  |
| **Data Summary:**  *Data Description):*   * Highlight the data collection process and an overview of the data set and the variables (include basic exploratory analyses such as outliers, missing values for important variables), describe data format, any cleaning/preprocessing steps taken   *EDA:*   * You have included your relevant EDA (limited to 3-4 of the most notable plots/summaries) * Plots are relevant and interpreted appropriately   + If you chose to not include certain variables because they were not interesting, you discuss why * It should be clear which variables you deem important and why you will use them for your model building process | 15  (broken into)  5  10 |  |
| **Analysis:**   * *Analysis Requirements*: The following are included and executed accurately   + Include required analysis steps and components as listed under your technique   + Include the appropriate added techniques as listed under your technique   + Your model is assessed with at least three criteria that are appropriate for your analysis   + Final Model is selected at the end of the process with a clear reason as to why it was selected over others. * *Analysis Presentation*: Appropriate tests are selected and justified * Discussed in a storytelling process of “first we did \_\_\_\_because \_\_\_ by doing \_\_\_\_ and came to \_\_\_\_ conclusion... then ….” * Be sure you are using the appropriate statistical language (ie model vs prediction equation) * There is limited output- only included if it aids in your presentation. | 50  Broken into:  40  10 |  |
| **Conclusion:**   * *Interpretation*: Appropriate interpretation of final prediction equation in context * Specific Example and interpretation of using for it * Overall answering- is this a good model? * *Future Research*: well-thought out areas for improvement * Include relevant citations * New variables to be collected * What you could do next, etc | 25  (broken into)  15  10 |  |
| **Paper Submission:** well written, contains all of the required components. | 50 |  |
| **TOTAL SCORE:** | 100 |  |
|  |  |  |

Each item in the presentation and written report will be scored as:

exceeds expectations, meets expectations, or does not meet expectations.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Group 1 Eval | Group 2 Eval | Group 3 Eval |
| **Visual Aesthetic/ Formatting** |  |  |  |
| Appropriate amount of text and readable (Visual Aesthetic) |  |  |  |
| Graphics and Visuals **added** to the presentation and were appropriate (Visual Aesthetic) |  |  |  |
| The poster is visually appealing (Visual Aesthetic) |  |  |  |
|  |  |  |  |
| **Presentation** |  |  |  |
| It was clear that the entire group understood the techniques used (Presentation) |  |  |  |
| It was clear that the group understood the content of the research topic (Presentation) |  |  |  |
| The group investigated an interesting research question(s) (Presentation) |  |  |  |
| The group used the poster as an effective aid in presenting their research (Presentation) |  |  |  |
| The organization of the presentation was logical, and the material covered was complete (Presentation) |  |  |  |
| The group adequately answered your questions. (Presentation) |  |  |  |
| The group had effective presenting skills (energetic, spoke loudly and clearly) (Presentation) |  |  |  |
| Comments: |  |  |  |

## Peer Review Rubric

This will be submitted through a Google Form.

Rate each group on the following:

1 (Strongly Disagree) 2 (Disagree) 3 (Neutral) 4 (Agree) 5 (Strongly Agree)

# Stat 3220 - Final Project FAQ

**Do resources/citations need to be on the poster?**

No. You can print a supplement document with your sources. You will want to minimally share your references for you background section.

**Where can I create my poster?**

Refer to the Project Tab there are some resources there. Previous semester students used Google Slides (or PowerPoint), Canva, or Adobe. If you google search “academic poster template” you should be able to find something. You will want to create it as if you were printing it to be 36”X48”. That way spacing and such will be easier to manage.

**How do we know which new “additional technique” might be appropriate?**

* **Weighted Least Squares (Ch 9.4).** In regular MLR each observation is equally weighted when building the model. Weighted least squares will “weight” some observations higher than others. This may be useful when you have data dealing with states or countries. (EX: is California “equal” to Rhode Island). This can be especially helpful when the observations are big outliers.
* **External Model validation (Ch 5.11).** The most common way you will do this is through data-splitting. Data-splitting should be used when you have a larger data set (>100). Evaluating Jackknife MSE or R^2 or PRESS can be used for smaller data sets. Just evaluating parameter estimates and prediction intervals is not sufficient for this criteria.
* **Box-Cox Transformation (supplemental).** This is used when you need to transform the response variable, but you are not certain what transformation to do. This procedure will give you the ideal power transformation, so that your data will meet the assumptions.
* **Ridge or LASSO**. Is used when there are a lot of variables and severe multicollinearity.
* **Poisson regression**. This is used when you have “count” data, particularly with a very unlikely response. Meaning you may have a lot of 0s.
* **The project tab has links to SAS documentation for each of these techniques that have some examples and sample code to follow**.

**Does our final model need to be “good”?**

No. The purpose of this assignment is that you have completed a thorough analysis based on what we have covered in our course. So long as you have attempted though several iterations to improve your model, that is ok. You can explain in your conclusion why your model is not good.

**What order should everything go in our analysis?**

In unit 3, I reviewed the MLR concept map. Refer to that as your guide for model building. Remember some steps happen simultaneously as others, and some steps need to be repeated.

**I have a conflict with one of the presentation dates.**

Speak to Professor Varanyak ASAP!