STA302H1 Fall 2025 Final Project

Part 1: Research Proposal

Overview

Due Date: Thursday, October 23, 11:59pm

- No Questions Asked Extension until: Sunday, October 26, 11:59pm (Please note the separate dropbox on Quercus). No additional extensions will be granted for any reasons, including corrupted file submissions.
- Please note that if you intend to use the NQA extension time, you should not submit any documents prior to the original assigned deadline. If you submit to both dropboxes, the latest submission submitted before the assigned deadline will be graded.

Goal of the Assessment:

- To have the opportunity to work on a topic of interest and to be creative about this topic.
- To experience the process of conducting a small literature review and incorporating knowledge gained into analysis.
- To think about whether a research question and/or a dataset is appropriate for use with linear regression.
- To create a draft of the components to be included in an introduction section of a report, as well as summary figures and/or tables for results section.

Learning Outcomes being Assessed:

- Apply multiple linear models on various datasets using R statistical software.
- Differentiate the relationships modelled using qualitative predictors, interactions between predictors, and continuous predictors.
- Create appropriate residuals plots to evaluate model assumptions for a given data set using software.
- Recognize distinct patterns in appropriate residual plots and correctly conclude which assumption is violated.
- Report the results of a residual plot analysis and recommend a course of action.

Summary of Instructions

1. Locate open-source data in an area of interest to the group that meets the data requirements listed below. Some examples could be (but are certainly not limited to) sports, medicine, public health, economics, video games, literature, etc. Groups will also need to discuss why their dataset is suitable to be used with a linear regression model.
2. Define an explicit research question using the information in that dataset. Note that groups will need to discuss why linear regression is appropriate to answer this question with this dataset (as opposed to another simpler statistical technique).
3. Locate three **peer-reviewed and relevant** academic papers related to the specific research question or topic of interest. Students/groups will need to describe how each article relates back to their proposed research question.
4. Select at least 9 predictor variables from the dataset to be included in a preliminary multiple linear regression model, with at least one being categorical in nature, and one being numerical. Include an interaction term using a categorical variable with another variable of your choice. The model will then be fit and a complete residual analysis to assess model assumptions will be done.
5. Provide a table that numerically summarizes each variable used in their preliminary model, with an informative caption that highlights any interesting features of the variables (e.g., skews, possible outliers or non-sensical observations, high spread, missing values).

Dataset Requirements

- Dataset must be open-source and the website where it was found/downloaded from must be provided.
- MUST contain at least 1000 observations (i.e., rows).
- MUST contain 1 response variable suitable for linear regression and at least 9 predictor variables, at least one of which must be categorical and at least one of which must be numerical. Categorical variables with multiple levels count as 1 variable here.
    - Since at least one predictor will need to be categorical, you may convert one of your numerical variables to categorical if no such variable is available in your downloaded dataset. However, you will need to justify your choice of variable and categorization in the proposal. For example, you may cite literature (is there precedent for what you are doing?), or use properties of the dataset itself (are there clusters or natural break points in the data?) to

justify your decision. As a rule of thumb, you do not want to lose any important information by converting your numerical variable to categorical.

- Should NOT be from an educational resource, such as a textbook dataset. If you're not sure, please ask the instructor or one of the TAs.
- Should NOT be one of the following datasets: Boston Housing dataset or Red Wine Quality dataset.
- If the dataset was found in a data repository (e.g., Kaggle, UCI Repository, etc.), you MUST ensure that your research question is novel and different from the original usage of the data.

Proposal Format

Your group will create a written proposal that should introduce your research question and data, summarize existing knowledge in that area, fit a preliminary model based on the existing knowledge, and conduct a residual analysis of the model. The proposal **must include** the following sections and **must not exceed** the word count in each case:

- **Contributions:** each group member's name is listed and a description of their contribution to the proposal is outlined (this does not count towards the word limit).
- **Introduction (400 words):**
  - State your proposed research question and discuss why fitting a linear model to answer your research question is preferred over a simpler statistical technique. Consider what aspect(s) of the linear model provides the answer to your research question (e.g., estimated coefficients, predictions, confidence intervals on specific quantities, p-values of certain hypothesis tests, etc.).
  - Summarize the results of three peer-reviewed research papers with a focus on what these papers tell you about the relationship between your response and predictors.
  - Finally, discuss who would benefit most from your linear model and the answer to the research question and why.
- **Data description (300 words):**
  - State where the data was found, explain how and why the data was originally collected (not how you found the data but how the original curator of the data collected it and why)
    - if you cannot locate this information, you should change datasets
  - Provide a statistical summary of the response variable and discuss whether it is suitable for use in a linear model (e.g., can you consider the observations independent, is your response reasonably continuous)

    - Summarize numerically or graphically (in a single figure/table) each predictor in your dataset that will be used in the preliminary model and identify important data characteristics, such as skews, high variance, missing values, outliers, etc.
    - Justify your choice of interacted predictors (1 categorical with another predictor). This could involve citing results from your literature review and/or provided a figure/table demonstrating the need for an interaction term in your model.
        - NOTE: if you had to convert a numerical predictor to a categorical predictor to meet the data requirements, you must also justify your choice and the chosen categories in this section.
- **Ethics discussion (300 words):** Answer the prompts, referring to the content of the ethics module during Module 3.
    - Would you consider your dataset to be trustworthy, given the criteria discussed in the ethics module? Justify briefly using material and terminology from the module.
    - Was your dataset collected ethically, and are you making ethically appropriate use of it, given the issues raised in the ethics module?
- **Preliminary results (300 words):**
    - Fit a preliminary model using at least 9 predictors from your dataset
    - Assess the assumptions of the linear model, noting any violations and what led to your conclusions.
        - NOTE: Place residual plots into the document in a grid (i.e., 2-3 plots placed horizontally in a single figure) so that multiple plots will display in a single figure for improved readability (see Resources below).
        - DO NOT correct the assumptions yet – you will do this in the next part of the project.
    - Display your preliminary model in a table or a figure, ensuring that standard errors and/or confidence intervals are included.
    - Discuss the results of your preliminary model (recognizing it is not your final model). Consider whether there are any surprises in the results (e.g. magnitude or direction of coefficients, significance, etc.) by comparing to what was summarized from the literature.
- **Plan (300 words):** Based on the preliminary results you have, outline the steps you will take to arrive at a final model that answers your research question, based on the content covered thus far. **You will not need to write code (or do any of what you outline) for this part of the assessment.**

    - Discuss how you will reduce or increase the number of predictors in your model. What statistical techniques will you use? Will you consider context and the information from your literature review in your decisions? Are there predictors that you must keep in your model, and why?
    - When will model assumptions be checked and how often? Based on your preliminary model, will you need to consider transforming any of the variables in your model? If so, what type(s) of transformation(s) will you consider and why?
    - Consider what a reasonable schedule would be to complete the project. Outline a schedule that shows your group has a good plan to accomplish your goals for the final project.
    - As part of your plan, consider including visuals such as a Gantt chart or flow chart to summarize this section
- **Bibliography:** an appropriately formatted list of resources and literature cited in the proposal (not included in work count). Please use APA citation format.

What to Submit

Only ONE member of the group should submit ALL required submission components. A complete submission to Quercus will include:

- Your group's completed Group Teamwork Agreement, saved as a PDF.
- The completed proposal, saved as a PDF.
- The Rmd file containing the code used to subset and clean the data, fit the model, produce a summary table, and conduct the residual analysis for checking assumptions. The grader should be able to run this code with the data you provide to obtain the same results and/or figures as what appears in your proposal.
- The original and cleaned (where appropriate) datasets as CSV files, uploaded to your UofT OneDrive, with the **shareable link** included as a **submission comment** on Quercus. Select the sharing permission "anyone in UofT with the link can access".

Failure to meet these submission requirements, including incorrect format of components, missing components, and OneDrive links that do not allow shared access will result in a **one-mark deduction** on the grade of the proposal if we can access the files, or a grade of 0 on any components we are unable to access.

Resources

Should your group have difficulty locating a suitable dataset that meets the group's interest and the dataset requirements, your group can consider using one of the datasets below.

- Ames Housing dataset
- NHANES survey dataset
- AirBnB dataset (needs you to create a free account)
- Million Song dataset
- NBA player dataset
- Spotify Dataset

You may also consider consulting the library resources below for help performing your literature search and citing the results.

- How to search for academic articles
- Using search operators to find articles
- Limiting search to peer-reviewed articles
- Why and how to cite your references
- Help getting the correct citation format
- Exporting a citation

Should your group use R Markdown to produce the proposal, the R Markdown resources below will help you format your document and make it more presentable.

- Settings for displaying or not displaying R code in knitted document
- Adding captions and other plotting features
- Including multiple plots in a grid using patchwork or base R plot commands
- Creating tables in RMarkdown using Kabble or manually
- Exporting plots in RStudio

For some advice in formulating a research question and searching the academic literature, see our Tip Sheet for Creating a Research Question, designed by Dory Abelman, a previous TA for the course.

Rubric

Criteria of Assessment	Excellent (2 points)	Satisfactory (1 point)	Needs Revision (0 points)
Introduction Section			
Proposed research question: • The response variable has been well-defined/explained so that no subject-matter experience with the data/topic is needed to understand the research question. • The research question avoids ambiguous language like "factors" and instead lists predictors of interest explicitly or lists groups of predictors sharing characteristics (e.g. demographic information). • Who would benefit most from the linear model and the answer to the research question and why is explicitly stated and realistic.	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.
Literature summary: • At least three legitimate peer-reviewed articles are summarized. • The main result of each article is summarized concisely and in the context of the original study population. • What each article tells you about the relationship between your response and predictors is explicitly article provided and is consistent with the summary.	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.
Suitability of linear regression: • Provides a reasonable justification for why a linear model is preferred over a non-modelling statistical approach. • Aspects of the linear model that can be used to answer the research question are explicitly stated and are consistent with the research question. • Uses appropriate terminology from the course materials.	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.

Data Description Section			
Description of data source: • Where the data was sourced/downloaded from is explicitly mentioned with a corresponding citation in the bibliography. • The original usage or purpose of the dataset is described (i.e. why was the data collected by the curator). • How the data were originally collected by the curator of the dataset is described (e.g. survey, census, web-scraped, etc.).	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.
Variable summaries/data descriptions: • The dataset meets the requirements set out in the assignment instructions (i.e., at least 1000 observations/rows, 1 response variable and at least 9 reasonable predictor variables (at least 1 numerical and 1 categorical)). • A justification for why the response variable is appropriate for linear regression has been provided and is supported by the variable summaries. • All variables to be included in the preliminary model have been summarized using appropriate plots or numerical statistics, and important variable characteristics have been identified correctly. • The choice of variables making up the interaction term has been justified (includes choice of variable and categories created if categorical variable not already part of the dataset) and is supported by context, the summarized literature, and/or an appropriate figure/table.	All four criteria are met.	Only two or three criteria are met.	One or fewer criteria are met.

Ethics Discussion Section			
• Answer correctly references some of the criteria discussed in the first ethics module. • Response makes a reasonable and clear attempt to argue for its conclusion.	Both criteria are met.	One criterion is met.	No criteria are met.
Preliminary Model Results Section			
Residual analysis of preliminary model: • All plots needed for a complete residual analysis have been presented, are correct, and are easily readable with appropriate axes and labels. • Each assumption and condition are assessed and a conclusion for each is provided. • Correct details are provided, with reference to the appropriate plot, to describe how such a conclusion was made for each assumption and condition.	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.
Preliminary model discussion: • Model estimates (with measure of error) from preliminary model are presented in an easily readable, understandable, and professional way. • A comparison is made between the preliminary model results and those summarized from the literature, and it is discussed whether the model yields surprising or inconsistent results. • The discussion of preliminary model results makes explicit reference to supporting evidence from the model and is explained using appropriate course terminology.	All three criteria are met.	Only two criteria are met.	One or fewer criteria are met.

Analysis and Team Plan			
Analysis plan: • Statistical methods intended to help reduce/increase the number of predictors in the model are described with sufficient detail to understand how decisions about predictor inclusion will be made and are correct. • A discussion on whether context and/or the summarized literature will be used to help decide inclusion/exclusion of predictors and why is provided and is consistent with the literature summaries and research question. • The plan correctly outlines when and how often model assumptions will be assessed as adjustments to the model are made. • A discussion about potential variable transformations based on the assumptions of the preliminary model, including how you plan to select a transformation to address specific violations, is provided, and is correct and consistent with the preliminary model results.	All four criteria are met.	Only two or three criteria are met.	No criteria are met.
Team plan: • A team plan is provided that includes dates of completion for conducting the analysis, creating the poster, and recording the presentation and indicates who will be working on each. • The team plan is sufficiently detailed for the whole group to follow, is presented in an organized way, and allows sufficient time to complete the final project.	Both criteria are met.	One criteria is met.	No criteria are met.
Overall Proposal Formatting			
• The bibliography and in-text citations are formatted correctly using a consistent style. • Word counts for each section are met or are no more than 15 words in excess. • Headers and paragraphs are used effectively to increase readability and separate ideas for increased comprehension. • No R code or R output (other than plots) is displayed in the written proposal.	All four criteria are met.	Only three criteria are met.	Two or fewer criteria are met.
			Total Points: /22