STA302 Summer 2025 Final Project Part 1
Research Proposal and Data Introduction
Due: May 21, 2025, by 11:00PM ET

Goal of the Assessment:	Learning Outcomes being Assessed:
• To have the opportunity to work on a topic of interest to them and to be creative about this topic. • 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.	• 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.

Instruction Summary:

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, ecology, finance, etc. Students/groups will also need to argue for 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 students/groups will need to argue for why linear regression is appropriate to answer this question with this dataset.

3. Select at least 5 variables from the dataset to be predictors in a preliminary multiple linear regression model, with at least one of these five being categorical in nature. The model will then be fit and a complete residual analysis to assess model assumptions will be done.

4. 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, one of which must be categorical. 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.
- 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, 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 (350 words):** introduce the relevance/importance of the topic, state the research question of interest, and describe why linear regression is a suitable statistical tool to answer the research question.
    - i.e., why should someone be interested in your project, what are you trying to answer, and why should you use linear regression.

- **Data description (300 words):** state where the data was found, explain how the data was originally collected (not how you found the data but how the original curator of the data collected it), describe the response variable (both statistically and with a written description of what it measures and why it meets the requirements for use in a linear model), summarize numerically or graphically (in a single figure/table) each predictor in your dataset that will be used in the preliminary model, and interpret the descriptive statistics in the context of what the predictors measure and how it relates to the research question.
    - NOTE: if you had to convert a numerical predictor to a categorical predictor to meet the data requirements, you must justify your choice and the chosen categories in this section.

- **Preliminary results (300 words):** fit a preliminary model using at least 5 predictors, conduct a full analysis of the linear regression assumptions noting any violations and what led to your conclusions. Discuss whether your preliminary model results are similar or different to results in the literature and why.
    - 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).

- **Bibliography:** an appropriately formatted list of resources and literature cited in the proposal (not included in work count). APA format is acceptable.

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 original and cleaned (where appropriate) datasets as CSV files, uploaded to a cloud-based storage service (e.g., OneDrive), with the **shareable link** included as a **submission comment** on Quercus.

Failure to meet these submission requirements, including incorrect format of components, missing components, and cloud links that do not allow shared access will result in a **one-mark deduction** on the grade of the proposal.

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 in the table below. You may also consider consulting the library resources for help citing the results. Should your group use R Markdown to produce the proposal, the R Markdown resources will help you format your document and make it more presentable.

Dataset Resources	Library Resources	R Markdown Resources
• Ames Housing dataset • NHANES survey dataset • AirBnB dataset (needs you to create a free account) • Million Song dataset • NBA player dataset	• Why and how to cite your references • Help getting the correct citation format • Exporting a citation	• 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

You may also wish to consider the writing resources posted on the Resources portion on Quercus.

Criteria of Assessment	Excellent (2 points)	Satisfactory (1 point)	Needs Revision (0 points)
Introduction Section			
Proposed research question: • The response variable of interest is clearly identifiable, and the predictors hypothesized to be related to the response are explicitly stated (or at minimum groups of common predictor characteristics are listed). • It is phrased using clear language and familiar terminology and makes a clear hypothesis about the population relationship. • It is directly connected to the stated importance/relevance of the project topic.	All three criteria are met.	Only two criteria are met.	Only one or fewer criteria are met.
Suitability of linear regression: • Uses appropriate terminology from the course materials. • Provides a reasonable justification for why and how estimating a linear trend will answer the research question proposed. • Provides a reasonable justification for whether the focus of the model will be on interpretability (description) or precision/accuracy (prediction).	All three criteria are met.	Only two criteria are met.	Only 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, and it is explicit how that usage differs from the current research proposal. • How the data were originally collected by the curator of the dataset is described and a corresponding reference is cited from the bibliography.	All three criteria are met.	Only two criteria are met.	Only one or fewer criteria are met.
Response variable summary: • An appropriate and suitably presentable numerical or graphical summary is used to statistically describe the response variable. • A written description of the response variable highlights important features of the response distribution, in the context of what is being measured/the research question. • A justification for why the chosen response variable is suitable to be used in a linear regression model is provided and is correct, based on the statistical summary presented.	All three criteria are met.	Only two criteria are met.	Only one or fewer criteria are met.

Predictor variable summaries: • An appropriate and suitably presentable numerical or graphical summary is used to statistically describe the chosen predictor variables. • Important/interesting variable characteristics (e.g. skews, abnormal values) or lack thereof are, in the context of what is being measured/the research question. • A justification for why the chosen predictor variables are relevant to answering the research question, making explicit reference to any modifications to variables that have been made.	All three criteria are met.	Only two criteria are met.	Only one or fewer 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.	Only one or fewer criteria are met.
Preliminary model discussion: • Model estimates from preliminary model are presented in an easily readable, understandable, and professional way. • A discussion on what these estimates tell the reader about a possible answer to the research question is provided in context, highlighting the effect of at least one numerical and one categorical predictor explicitly.	All two criteria are met.	Only one criterion is met.	None of the 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) are displayed in the written proposal.	All four criteria are met.	Only three criteria are met.	Only two or fewer criteria are met.
			Total Points: /16