STAT3215Q. Applied Linear Regression in Data Science (Fall 2023)
Course Project Description

Project Overview

This course project consists of writing a report about a published, peer-reviewed research paper, reproducing the results, and extending or enhancing the results in some way. Specifically, this includes

- identifying a published paper in the literature that (1) has employed either the linear regression methods we have discussed in this class *or* closely related methods and (2) has a publically available dataset,
- obtaining the exact dataset(s) described in the published paper,
- reproducing the results described in the published paper,
- extending or enhancing the results of the published paper using an approach not discussed in the paper,
- and writing a report detailing the linear regression models and methods used.

You may work alone or in pairs ("groups" of 1 or 2). After approval of your project proposal (worth 20% of your project grade), each group will ultimately submit to me the original published research paper, the project report written in R Markdown (with text and code for reproducibility), and the dataset for reproducing the results. More details on each of these items are provided below.

Important Dates and Deadlines

By 11:59pm, Friday October 13, 2023: Identify a partner if you are working in pairs and inform me by email, copying your partner. Just one member of the pair should email me. If I do not receive an email from you and you were also not copied on a partner's email to me by this date, I will assume you are working alone.

By 11:59pm, Friday October 27, 2023: Identify paper and dataset. Submit an approximately one-page proposal (submit both .Rmd and .pdf file), along with the original paper and dataset (as a .csv file). This is worth 20% of your project grade. **Late submissions will be penalized.** A late submission within a 24 hour grace period will be graded for only 50% of the proposal credit. Late submissions beyond 24 hours of the due date and time will receive no proposal credit, but still need to be completed and approved by me. Following discussion with me, you may need to submit revised dataset and/or proposal. Await my approval before proceeding.

By 11:59pm, Friday December 8, 2023: Project report due date; submit all material (original paper, project report [both .Rmd and .pdf], data, code). The actual report is worth the remaining 80% of your project grade. **Late submissions will be penalized.** A late submission within a 48-hour grace period will be penalized 20% of the possible score for each 24 hour period after the due date and time. Late submissions beyond 48 hours of the due date and time will receive no credit.

Details

About File Submission: If you are working in pairs, select one group member to be the corresponding group member, who will be responsible for submitting all project-related files (proposal(s), data, final report, etc.). That is, only one person in each group of two will be submitting files to me via email: elizabeth.schifano@uconn.edu . If you are working alone, you are the corresponding group member in a group of size one.

All submitted files should be named using the following format: LastName-FirstName-Description, where LastName, FirstName, and Description (Proposal, Data, Revised Proposal, Report, etc) are filled in appropriately for the corresponding group member (e.g., Schifano-Elizabeth-Proposal.pdf).

About Paper Selection and Proposal (20% of project grade): When selecting a research paper, avoid articles that are heavily mathematical. If you see a large number of formulas, chances are the paper uses some fairly advanced statistical techniques. Our goal is to find papers with applications of linear regression (or simple experimental design) rather than papers developing new methods or theories.

How to find a research paper?

- Search PubMed [www.ncbi.nlm.nih.gov/pubmed/]. Enter keywords from class and/or interest areas.
- Search Google Scholar for some of the same keywords.
- Look through specific online journals, e.g., Nutrition Journal, Journal of Sports Science and Medicine, Data in Brief.
- Verify that the paper you select provides all the data required for you to perform the described analyses.

I can help point you in the right direction if you are having trouble and you can ask me if you aren't sure if the paper that you have selected is appropriate for this project.

After selection of a research paper, the corresponding group member will email me an electronic copy of the paper with an approximately one-page proposal describing the analysis to be replicated, how you will recreate the dataset used in the paper, if needed (e.g., any merging of files, any filtering of data to include subjects of particular age range or gender, any creation of new variables derived from the initial dataset), and the extension/enhancement you plan to implement.

What is an extension or enhancement? These are meant to be additional subanalyses that do not exist (or at least are not reported) in the published paper. Here are some possible examples (you are not limited to these):

- Many times checkable statistical assumptions are not checked, or if checked, are not reported in the published paper. The extension or enhancement could be a thorough diagnostic check of the assumptions, and if a violation exists, the implementation of a remedy.
- Likewise, checking for influential data points and collinearity may also not be reported; the extension or enhancement could examine these and include a remedy if needed/available.
- Suppose the authors of the selected paper did not implement a particular analysis correctly (e.g., they treated a numerically coded *qualitative* predictor as a *quantitative* predictor). The enhancement could be the re-analysis using the correct implementation.
- Suppose the authors wish to use model selection for prediction, but the model was not (properly) validated. The enhancement could be the inclusion of a proper validation method (e.g. cross-validation).

The proposal must be written in R Markdown and submitted as a pdf and Rmd file, named appropriately, and written in grammatically correct English. Be sure to include the name and student ID number of all group members on top of submitted proposal. *Each group must find a unique paper and data set; the same paper/data may not be used by different groups.* **I must approve your data set and proposal before your group can proceed with completing the project.**

About Project Report (80% of project grade): The project report must be written in R Markdown and submitted as a pdf and Rmd file, named appropriately, and written in grammatically correct English. The report must describe *in detail* the statistical model(s) being employed, the assumptions required for the statistical analysis, any additional steps required for replicating the analyses and results, the replicated results, and the details of the extension/enhancement implemented. Simply writing a summary of the research paper's methods section is not sufficient. If you cannot replicate the results exactly, you must discuss in detail the potential sources of discrepancy. Note that you do not need to use the same software as used in the original paper, but this could contribute to differences in reported results and this must be described in the report.

The project report must include the following components in this order:

1. Title (Sub)page: This includes title of the report, and name(s) and email address(es) of group member(s).
2. Section 1. Introduction: This section should include the statement of the problem investigated in the original paper, a brief discussion why this problem is important, and an outline of the rest of the report.
3. Section 2. Description of Data: Provide a description of the data, providing details of the study design and methods of data collection. You should also replicate various graphical and numerical descriptions of the variables that were included in the original paper. (If they do not exist in the original paper, you must include them in your report anyway.) The data set must also be submitted as a .csv file.
4. Section 3. Methods (or Models) and Analysis: Clearly write out the statistical methods, model(s), and formulas needed for reproducing the analytic results in the paper, defining all terms, and listing all assumptions. Analyze the data to (attempt to) replicate the published results, and include your reproduced results. Your reproducible code must be included within the R Markdown (Rmd) file. Give conclusions associated with the results.
5. Section 4. Extension/Enhancement: Describe in detail the additional subanalyses, why these additional subanalyses will extend or enhance the results of the original paper, and include results of the additional subanalyses.
6. Section 5. Discussion: Describe in detail discrepancies of the results, if any, and discuss whether the listed assumptions in Section 3 are reasonably satisfied. Think critically about any conclusions drawn in the paper and discuss whether they are adequately supported.
7. References (list of references cited in the report in alphabetical order based on the last names of the first author of the articles or books; APA or MLA format); this must include at a minimum the citation of the original published paper.

Note that points will be deducted if any of the above components are not included, or if they are included in a different order.

Do NOT directly copy text, figures or tables from the original paper; summarize text in your own words and create your own figures and tables.

About Reproducibility and Dataset: The dataset must be submitted as a .csv file, and the code within the Rmd file must be ``fully reproducible'' in the sense that one should be able to reproduce all of the results in your report by simply running the relevant portions of the code. In other words, additional code should not need to be written in order to run your analyses. Since you are submitting both the Rmd and pdf report file, you may use echo=FALSE to suppress code appearing in the final report. NOTE: if the file size of the dataset is too big to be sent directly through email, share it with me via OneDrive.