BST-STA 223 W21
Class Projects
Due dates: Project 1: Wednesday, February 2, 3pm
Project 2, Wednesday, March 11, 3pm

1. Select a complex data set, research project, or topic for a new wikipedia page of your choice that involves topics from the lecture notes and is suitable to apply, review, demonstrate or extend any of the methods we are discussing in class or that are covered in the lecture notes. If you cannot find a project on your own, please ask for advice how to find one. More on this below. The two class projects must be different, with the exception that a research project (which can be data-oriented, computational, methodological or theoretical) can count for two projects and then Project 1 is about the initial results of the research. In all other cases, one of the projects should be an advanced data analysis. If both projects are an advanced data analysis, the data and methodology must be different between the two projects.

   You are not allowed to repeat an analysis of a data set that has already been done in published or online work or otherwise is accessible to you. Also each data should be used only for one project, and research projects should be unique. You may select data that have previously been analyzed as long as your analysis differs substantially from the previous one, in case you are aware of the previous analysis, and you then must properly refer to and disclose such previous analysis in your project report, and quote all relevant references.

   Specifically, your 223 projects cannot overlap with any other projects you have done previously or are currently doing for another class and must be new work for you; they must be based on your independent work; and you must report all sources (papers, reports, books) that you have used when completing it. Failure to adhere to these standards constitutes a violation of the Academic Code of Conduct.

   A high quality project will go beyond a straightforward application of GLM and will address some challenges in the data or include a comparison of different approaches. Important is an initial check for outliers and missing data and appropriate preprocessing. You need to make it clear whether your project is about prediction (find the best predictor) or modeling (find the relation between an outcome and predictors). In the former case you need to ascertain the misclassification rate by selecting training and test sets and preferrably compare several predictors. In the latter case you need to interpret your findings and try to connect the results to the underlying question.

   All projects need to be approved, and you should discuss the project with me in person, during office hours or by special appointment, before you seriously start on it. You are requested to settle on a project very soon, as you likely will need about 3-4 weeks to complete it. If you change the scope of your project substantially, i.e., change the type of response variable you are considering, you also need to obtain approval for such a change. When you present a data analysis project for approval, you need to state the number of independent units/subjects, nature of the data and variables available per subject, predictors/responses, and what is the overall question to be addressed by the analysis.

To get initial approval, please send a two paragraph project outline by e-mail. It is highly recommended that you also talk with me about your plans. The first paragraph of your e-mail should include the name and the source for the data, as well as details about predictors, responses and their type (sample size, no. of predictors, type of predictors, and type of response need to be included). The second paragraph should contain a brief description of the approaches, models and methods that you anticipate to use when you start the project, and should also include the overall question you aim to address. For non-data analysis projects, one paragraph will suffice.

2. If you select a data analysis project, in case you don't have your own data from a prior collaboration, you can obtain data from popular data bases, such as:

   (a) UCI machine learning repository: http://archive.ics.uci.edu/ml/

   (b) Princeton datasets for GLM:

   `http://data.princeton.edu/wws509/datasets/#phd`

   (c) Kaggle datasets: https://www.kaggle.com/datasets

   (d) Datasets for the book Applied Linear Analysis and Generalized Linear Models, 2nd Edition, can be found at:
   http://socserv.mcmaster.ca/jfox/Books/Applied-Regression-2E/datasets/

   (e) datasets in R package "datasets".

   (f) You can also easily find lots of other datasets that are suitable for GLM by simply googling "datasets for GLM".

   (g) Since all of you need to choose different datasets, each dataset will be considered/approved on a "first-come, first-serve" basis (determined by the order I receive the e-mails with the requested projects).

   (h) If all else fails, I can provide you with a dataset as a last resort.

   (i) To ascertain whether a dataset is suitable for the class project, you should carefully read through data descriptions. Are you aiming at applying smoothing, at the GLM for modeling a regression relationship, or at predicting/classifying an outcome?

   (j) You need to identify the overall goal of your class project and motivate it. Also consider the difficulty level associated with the analysis, as you want to choose data that are neither too challenging nor too trivial. Basic data features such as sample size, number of predictors, and whether responses can be assumed to be independent should be considered, as these impact the level of difficulty.

   (k) Also do a search of papers that have used the data you are considering since you are not allowed to repeat an analysis that has been published or posted before. The existence of a paper on the data you select does not mean that you cannot use the data, but your analysis has to differ in substantial aspects from any previous posted/published analysis, as mentioned above.

Once you have settled on a data analysis or research question of interest, or the creation of a wikipedia page, you need to develop a plan for the analysis of your data, and select an appropriate model and implementation (software of your choice). If you apply a GLM, your emphasis should be on modeling and interpretation, where we aim to select only key predictors, while you can also do prediction after you completed the modeling. For prediction we typically select more predictors and may opt for somewhat less interpretable models. You always need to start with descriptive data analysis, checking for outliers and missing data and generally assess the state of the data (box plots, matrix plots). You are welcome to discuss your data analysis plan and preliminary results with me or the TA.

3. Once you have finished the project, you will report in two ways, with both parts to be submitted online at the time of the due date:

   - A poster that presents your major findings, limited to the size of one large poster sheet (can be black and white). The poster will be presented by you on zoom in less than 5 minutes. Be prepared to defend your analysis in the discussion of your poster. The poster should briefly describe the data or research project, research question, statistical model, fitting, diagnostics, and conclusions, and should provide a meaningful and appealing summary of your work. Graphics are a good way to efficiently convey complex information and for most projects should be included.
   - A written report, limited to a maximum of 5 pages of main text (typed, with 11 sized font and 1.5 spacing as well as top, bottom, left and right margin at one inch) which includes your major figures, tables (they will go to the main text) and references. All sources that you use in any way must be attributed and included in the list of references (which does not count towards the page limit). If you choose a data analysis, the data and your analysis code should also be submitted in a suitable form, alternatively you can provide a link where the data can be obtained. For proprietary data, exceptions can be made. You also need to include a description of the format and of the variables involved. Your report can also have an additional Appendix (not subject to the page limit), where you can list additional details that exceed the page limit, such as additional figures or derivations, outputs and tables. It is mandatory that you refer to these materials from the main part of the report.

4. Your written report should include a brief summary and introduction, in which you describe the data or research project, including the main question to be addressed and a very brief description of the data if it is a data analysis. The variables (predictors and responses) need to be clearly described including their number and units. This is followed by methods, results and discussion. Describe the models you use, why and how they were chosen, any goodness-of-fit analysis in the methods section, and the fitted values, test outcomes, selected predictors etc. in the results section. The discussion should contain interpretations and a critical assessment of your analysis. For a research project, you need to run a simulation or extend an existing method

or theory, where the extension should be the main part of your report and you refer to existing and previous approaches in the appendix. In exceptional cases, a research project can also consist of a summary of recently published research.

5. The decisions that you make for your data analysis or research project should be clearly stated, justified and well-reasoned, and the strengths and weaknesses of your analysis and possible alternative approaches described in the Discussion. All material in the Appendix must be referenced from the main text (for example, say ... see Appendix A,... Appendix B,..., i.e., structure your Appendix in sections A, B,..., as needed).

6. You are responsible for following the rules: You are not allowed to exchange or collaborate on your actual code, output, models, analyses or write-ups with other students, and not allowed to copy any online or otherwise existing analysis or project from others. You are allowed to discuss general issues of your project with other students and are welcome and expected to ask the TA or me for advice. You are not allowed to reuse any material or code that you have used previously or are using concurrently for any project in another class that you are taking or have taken. You need to clearly reference any material and publication that you consulted for your analysis, and list the data source. It is highly recommended that you discuss your research project, wikipedia page or plan for a data analysis with me and the TA. If you use notions or methods that were not discussed in the 223 class, you need to define/describe them and provide references.

7. Projects will be graded according to intellectual originality and innovative ideas in the data analysis or research project, clear justifications for the steps of your analysis, as well as appropriateness and correctness of your analysis in the light of your stated goals and the data. Do not overlook hidden dependencies of subjects/cases. Describe explicitly, clearly and correctly all models you use. Make sure you write and describe your models correctly, give reasons why you use them, clearly and correctly identify your estimates and hypotheses and test statistics and how the statistical results address the (scientific or subject-matter) question you aim to answer. Give reasons why you do what you do and interpret the results, both statistically and in the larger context of the problem you study.