STAT 437 - Paper Review/Critique
Due: Monday, April 25 on Learn (Dropbox Provided)

The goal of this assignment is to have you demonstrate course knowledge and engage with scientific research directly. My intention is to give you **as much** flexibility with this project as possible. If there is a direction that you would like to take with this project, that is not listed here, please email me for us to discuss. Please read these directions carefully, and ask any questions you may have.

For the final project in this course you will be asked to produce either applied or theoretical research based on the methods we have covered in the course. You will present your research in a form of your choosing (for instance, a written research report, a research poster, a video explainer, a blog post, etc.), in such a way so as to convey your understanding of the course content *as well as* your engagement with the selected research problem. In what follows I will outline two different ways of approaching this project, **however**, you are free to select any topic and any form of presentation which you think will convey your understanding. If you have any questions about whether or not a particular approach is acceptable, please ask. For this project, **you may work independently or in pairs (of your choosing). I would encourage you to use the course Teams channel to post possible project ideas, and arrange yourselves into pairs, if you desire.** Please indicate to me what your intention is (via email, dylan.spicker@gmail.com). If you choose to work in pairs, please note that the project will receive a common grade, which will be assigned to both members.

The assignment is due on the last day of exams (though earlier submissions are welcome, if applicable!) and is worth 30% of the grade. The full mark will come from the submission (and will not, as the syllabus had initially stated, involve an interview portion). You will be graded on your grasp of the course material, the novelty and appropriateness of the methods that you use, and on your ability to communicate and exercise a suitable approach to problem solving. I have provided guidance for both **applied projects** and **theoretical projects**, though it is **entirely acceptable** if your project fits between these two (some of each) or does not neatly fit into these categories.

Before you begin into your project, please submit a brief description of what you wish to do (including any particular references) to me over email, and indicate if you will be working in pairs or individually. There is no specific due date on proposing your project to me, but note that the earlier you do it (1) the more time you will have to make required changes if there are any concerns, (2) the less likely that another group will have taken exactly the same project as you, and (3) the more I will be able to provide helpful tips of guidance to the approach.

The form of the submission can be anything that effectively presents the results of your research. This may be a report (with the outlined sections), a research poster (covering the required material), a research talk (in a video format), an interactive web application (using, for instance, R Shiny), a blog post, or whatever other form you find interesting, useful, or appropriate. As always, if there is anyway that this project can serve as useful to you, beyond the scope of this course, please reach out and discuss options with me! I am very open to accommodating projects that will help you, if we can have them showcase the knowledge you learned in this course!

Directions for an Applied Project

If you work on an applied research problem, the idea is to **find** and **analyze** a real-world dataset, with a clear scientific objective, and communicate your results. You can work on **any** real-world dataset, so long as the methods we have covered in this course (either relating to longitudinal data analysis, survival analysis, or both) are applicable. For your project you should:

- Find and gain access to a real-world data set, that is **of interest** to you.
- Motivated from these data, **pose a scientific question** that can (in theory) be answered from these data.
- Download the data and transform them to a suitable format, and then apply modelling procedures discussed in this course in an attempt to answer the question.
- Conduct relevant inference, hypothesis testing, model fitting, etc. to ensure that the models that are used are suitable both to answer your question of interest **and** are suitable based on the observed data.
- Communicate the findings in a report, poster, presentation, etc.

While the specific goals of your analysis will need to be guided by the questions which are of interest to you, there are some general pieces of guidance that will likely be applicable. First, note that there will always be choices for you to make in terms of modelling (including nonlinear trends, interactions, individual heterogeneity, etc.): make sure that you are making these choices based both on what is of scientific interest, and what the data suggests is correct. Second, if you follow an automated process for variable selection in your modelling you will likely miss factors that are interesting or important. A variable which does not seem to be related to the outcome should still be included in the model **if** you were interested in asking "is this variable related to the outcome?", for instance. Third, it is important to focus on the scientific plausibility of the models in addition to the statistical validity of them: if a factor should influence an outcome, there's reason to include it, even if it doesn't seem relevant, and vice versa.

Your project should contain **background information** and a description of the problem that you are setting out to solve. You should indicate if other research has looked at this question before, and if so, what did they find? How does your approach differ from any existing research? The background should make clear what your question is, why it is interesting, and how you intend to answer it. You should discuss **exploratory analyses** where the data are plotted and explored, and where any concerns or special considerations are noted. Is there missingness? If so, how do you plan on handling it? Do you need to consider a subset of the observations for one reason or another? How might these choices impact your question of interest? You should discuss the **methods**, explaining the statistical techniques you will employ, and justifying why these are selected. You should indicate your approach to model building, including a discussion of what criteria are used to select models and why some models were ruled out. You should discuss the **results** of the analysis, including relevant estimates and summary statistics, as well as any plots or code that are relevant. The interpretation of the specific models and terms should always be in reference to the underlying subject matter, and should answer your question of interest (and other relevant questions). **The results section will likely be the largest section.** Finally, you should include a **discussion**. This will discuss any issues/limitations for the analysis you performed, and an indication of how these might be able to be overcome. You should draw conclusions regarding the underlying questions, and illustrate points that may be interesting to look in the future.

Example Project Topics

The following provide some examples of (possible) types of projects that you could consider. Note, you do not need to select one of these, but they may serve as inspiration into possible types of questions!

- Using data from the Framingham Heart Study (FHS) to determine the impact of age, BMI, and smoking quantity on serum cholesterol levels. In particular, investigating whether (when controlling for relevant factors) there is a notable impact (cross-sectional or longitudinal) of the quantity that someone smokes on their cholesterol levels, and if so what form does this effect take (is it linear in the amount smoked, based on any smoking at all, etc.).
- Determine the impact that exposure to ultraviolet radiation has on mortality rates for malignant melanoma. In particular, test the hypothesis that exposure to UVB has a significant and sizable impact on the rates of mortality, and if there are any confounding relationships in these data.
- Determine the efficacy of particular public health measures on preventing the spread of COVID-19 infections in a region. In particular, using international data, determine whether the presence of indoor mask mandates reduces the probability of a mass outbreak (defined, for instance, relative to the capacity of the hospitals in the region).
- Investigate the impacts of aging on the performance of professional athletes. In particular, considering publicly available data from the national hockey league, consider the impact that aging has on a player's offensive performance, and what factors influence these aging curves.
- Using the podcast data posted on the course website, determine a model that predicts a podcasts popularity in the future. In particular, determine the impact that podcast age, category, and tone have on the popularity of podcasts, and use this to forecast podcast popularity into the future.
- By scraping relevant fan websites, build a dataset from the reality TV show *Survivor*, and using this data conduct an analysis on the factors that are relevant to predicting survival rates for individuals on the show. What baseline factors (either in terms of personal characteristics, or production decisions) seem to impact the longevity of a contestant's gameplay?

Directions for a Theoretical Project

If you work on a theoretical research problem, the idea is **expand on** and **answer** theoretical questions related to the subject matter in the course, with clearly presented results. The theoretical question that is being asked can be one that has been previously answered (but which we did not explicitly cover), or one that you cannot find an existing answer to, but which interests you nonetheless. If your question has been previously answered, you should find a way to contribute further to the finding, for instance by exploring simulation experiments related to the question, using a different method to prove the result, or relaxing unnecessary assumptions that were made. For your project you should:

- Pose a theoretical question, that is **of interest** to you, and related to a topic in this course.
- Using either rigorous proof, simulation experiments, example cases, or a combination of these techniques, explore the question with an attempt of providing an answer to (part of) it.
- Demonstrate the theoretical merits of the approach, on the basis of mathematical properties or through thorough simulated results.
- Communicate the findings in a report, poster, presentation, etc.

The question of interest is ultimately up to you. It does not need to be an *important* question, nor does it need to be a particularly challenging question to answer. However, you need to ensure that your approach to answering the question demonstrates a keen awareness of the methods discussed throughout the course on longitudinal data, and provides non-trivial insight into the problem. You do not need to prove something novel. If you pose a question which you cannot answer in general, but which you thoroughly explore through simulation and examples, you can receive full marks for that theoretical investigation. The question can have been approached elsewhere by others, however, in this case you need to provide additional value beyond explaining their theoretical result (for instance, by testing the limits of the method through simulation).

Your project should contain **background information** and a description of the problem that you are setting out to solve. You should indicate if other research has looked at this question before, and if so, what did they find? What are you planning to add compared to their previous work? The background should make clear what your question is, why it is interesting, and how you intend to answer it. You should discuss **methods and notation** that are required to understand your problem, and if relevant, provide a useful example to serve as a guide through the problem. You should discuss the **results** of your investigation, including any relevant proofs, simulation results, code, or derivations. You should make clear how your work answers the question that you set out to answer. **The results section will likely be the largest section.** Finally, you should include a **discussion**. This will discuss any issues/limitations for the analysis you performed, and an indication of how these might be able to be overcome. You should draw conclusions regarding the underlying questions, and illustrate points that may be interesting to look in the future.

Example Project Topics

The following provide some examples of (possible) types of projects that you could consider. Note, you do not need to select one of these, but they may serve as inspiration into possible types of questions!

- Why does the structure of a generalized linear marginal model imply (conditional) independence between outcomes and covariates, and why does this independence mean that stochastic time-varying covariates are inadmissible in these models? The proof for this already exists, so instead, this theoretical investigation will clearly present the results of the proof and then thoroughly investigate (using well-planned simulations) the impacts of ignoring this implied independence on the consistent estimation of model parameters.

- Why does the restricted maximum likelihood (REML) procedure produce estimators with reduced small-sample bias, as compared to standard ML procedures? This is a theoretical question which has been answered, so this project will explore these results both through deriving the small sample bias in a simple example and demonstrating the degree of bias based both on sample size and the specific model parameters in a series of simulations.

- What are the impacts on longitudinal methods if within-subject correlation is negative? Positive correlation leads to increased efficiency in estimation and negative correlation should (theoretically) harm efficiency by the same token. Are there settings where negative correlations may genuinely arise in data, and if so, how do existing methods (which either explicitly or implicitly assume positive correlation) handle these types of data?

- What is the impact of informative censoring on an analysis conducted under the assumption of non-informative censoring? In survival analysis we often make the assumption that any censoring is non-informative. What happens when there is added information in the censoring process? Are there certain informative mechanisms that can be easily adapted into the likelihood framework we have considered? This analysis could proceed both on the basis of theory (looking at specific cases which have not been meaningfully looked at in the literature) in addition to simulation-based results, and may build on previous literature or not.

- In addition to the topics covered in the course, we could have used **multi-state models** to analyze time-to-event data. These models are related to the transition models we discussed for categorical data. How might these models be formulated, estimated, and implemented in practice? This would take existing research, place it into the context of our course, and then demonstrate the utility through example analyses and simulated analysis.