Project Description
Biostatistics 755, Spring 2024

1 Overview

Learning is best done by example. I don't know who first came up with this idea, but it's something that has been passed down since the beginning of time. As a parent of three kids, I've seen the difference in the power of something learned through my voice versus something learned through experience. You can tell your kids to eat neatly thousands of times with little impact. Kids at a lunch table can pick on them once, laugh, and instantly change their behavior for a lifetime.

In this course, the overall goal is to learn how to do a complete analysis:

1. decide on the correct statistical model (e.g., linear mixed model, logistic GEE/GLMM, etc.),
2. develop a covariate strategy that will answer the question of interest (i.e., the hypothesis test for β_3 tests my question of interest),
3. know how to implement the model in statistical software (e.g., SAS, R),
4. use data-driven techniques to fine-tune the model and arrive at the final analysis(es) (i.e., coming up with appropriate adjustment variables, appropriately modeling covariance, checking model diagnostics),
5. and clearly interpret the results in general scientific language (i.e., interpreting the value of β_3 so someone with basic scientific knowledge can understand).

Since we learn best by example, you will learn these steps by implementing them on two projects of your design: a midterm and a final project.

1.1 Midterm versus Final Project

The final project will be more rigorous than the midterm project. **The final project requires a final report and a presentation, while the midterm project only requires a presentation. The outcome for the midterm project is required to be continuous since that's all we will have completed by then. The final project's outcome can be continuous, binary (yes/no), categorical (yes/no/maybe), or count. You can use the same study for both projects, but the study questions should be notably different.** For the midterm project, we will have not discussed missing data issues and you will use all data that are available (even if they have only 1 observation). As we'll see later in the course, this is an acceptable (if inefficient) way of analyzing data (especially for linear models).

You will work in teams of two or three (no solo projects). You may select with whom you want to work with; working with the same team for both projects is not required. For both projects, we can accomidate at most 10 teams due to time limits on the in class presentations. For the midterm project, the first 5 pairs that tell me they are going to work together can do so. The remaining people in the class must be in teams of 3. For the final project, those that didn't get to work in a team of 2 but wanted to do so will be given first chance to work as a pair.

1.2 Topic Selection

The project aims are to explore some aspects of the methods for longitudinal or clustered data covered in the course. The topic of the project will be selected by the team. Learning "what questions to ask" is not a goal of this course, but you will have to do some of that here. As far as what data you will use, some possible places to look for data are your advisor, ADD Health Study, Wisconsin Longitudinal Study, or NHLBI. Finding a dataset is part of these projects. This project should not require any data collection.

Some tips on choosing a topic:

- It is best to pick a subject that interests you since it's more likely that it will interest the rest of us.
- Choose a new problem and not something that you have already done.
- The data do not have to be related to public health.
- The goal of this project is to implement the methods from class. Adding topics outside this course's scope usually will not be a good place to put your effort (at least in terms of your grade on the project). You want to spend the majority of your time ensuring that you do and interpret the things we learned in here the correct way. For example, I will have no problem ignoring sampling weights.
- If your data are longitudinal the question *should* utilize their longitudinality (if that's a word). That is, are you looking for within-subject change over time? Consider asking yourself "could this be answered reasonably well with cross-sectional data?" If so, then your question is not using the longitudinality of the data.
- **Depth over breadth.** Whenever you have questions about what you should be looking at in your project, always choose to investigate one question in-depth over looking at many associations. It's better to say something definitive about one relationship than to say nothing about three relationships.
- **You don't have to win.** When doing an analysis where there is no relationship between Y and X the best answer is to find no relationship (even when you know there "has to be one"). Do not repeatedly apply different methods looking for a "better" result. You can make sure your model is correctly specified, but stick to your analyis plan (even when things don't turn out as planned). Again, depth over breadth.

1.3 Types of Projects

The main differentiator in the types of projects is the data used for analysis.

- *Static type:* these projects will use data that require straightforward cleaning or transforming of the variables. These are commonly "well-worn" datasets.

- *Dynamic type:* these projects the data require some meaningful cleaning, transformations, or exploration before use.

All projects require exploratory data analyses to determine how variables should be included in models (i.e., should I use X, X^2, $\log(X)$ or categorize X, same for Y).

For static projects, the team must test multiple model dependence structures (i.e., random effect specifications, covariance types, etc.) and adjustment strategies.

For dynamic projects, a single general dependence structure (i.e., baseline random intercepts) can be used and the investigation into how to adjust doesn't have to be as rigorous. If you have a dynamic project, you must demonstrate this in your presentation and paper. That is, give a background of the data and explain the difficulties you had (and your solutions).

My goal in making this differentiation is to respect those projects that require significant time on the data cleaning side. For all projects, I would love to see a rigorous determination of the best dependence structure and adjustment strategy. However, if you spend 12 hours cleaning your data, you deserve a break on the analyses side. This is my attempt to do so.

2 Midterm Project

Project abstract (due on 2/9)

One-page proposals are due in class by the deadline stated above. These proposals should list your team members and spell out briefly the main goal of the study, your basic approach, and each team member's role. The purpose of the abstract is to make sure:

- the dataset is appropriate for this class (i.e., longitudinal or clustered data),
- the goals of the project are well defined and
- the outcome and exposure of interest (if any) are feasible.

I don't really grade the abstracts (if you hand it in, you get 10/10). So, just do your best to describe what analysis approach you'll use. If you feel that you don't know what approach to use, just do your best, and we'll work through it together.

If you don't think through your analysis before handing in your abstract, the worst thing that will happen is that you won't get meaningful feedback, which could hurt you down the road. So, think it through, do your best, and make sure you understand the feedback I give you so that the analysis is clear.

Oral Presentation (2/27 and 2/29)

Each team will give an oral presentation of 15 minutes. The presentation should include all group members. The course website will contain examples of presentations from past years. Your slides should contain all relevant information for the analysis (see the description of the final report for an example outline). The slides and what your group says during the presentation are the majority of your report. Your slides **must** include an interpretation of the main coefficient of interest. This is usually a β coefficient, but could also be a variance term or a least squared means estimate. Statistical code should be emailed separately. Your code is very important as it tells me what you specifically ran.

Grading

Team average scores for these projects will be assigned based on the following rubric:

- Quality of the hypothesis generation 10%
- Need for longitudinal/clustered methods 10%
- Analysis designed to address aims 10%
- Quality of exploratory analysis (figures and tables) 10%
- Model Selection (covariates, covariance, random effects) 10%
- Interpretation of coefficients (even if they're not significant) 10%
- Profession appearance of slides (don't just paste output) 10%
- Strength and limitations of the study 5%
- Seminar presentation:
  - Keeping the interest of the class 10%
  - Handling questions 5%
- Implementation in statistical software 10%

Implementation in statistical software is judged based on your statistical code (sent separately and not discussed in the presentation).

3 Final Project

Project abstract (due on 3/22)

See the description from the midterm project.

Final report (due on 4/12)

You should thoroughly but concisely report your entire investigation. Include at least:

- Introduction
  - Background/Motivation
  - Current literature on the topic
  - Specific Aims of the project
- Methods
  - description of data variables
  - description of how the final analytic sample was obtained

    - description of why data are missing and how they were handled

- Statistical methods section
    - Write-up of the analysis plan (exploratory analysis, model section and inferential models)
    - Statement of the statistical model
- Results
    - discussion of exploratory analysis (table/figures)
    - discussion of model section
    - discussion of final results
- Conclusion
    - Relate final results back to the aims of the study
    - Interpretation of coefficients (even if they are not significant).
    - A statement of the implications of your study
    - A discussion of further questions raised by your study
- Meaningful figures and tables that show preliminary patterns in the data.

The main body of the report should include only the end products of any statistical calculations. Statistical jargon will not be well received (i.e., include your findings' interpretation). To quote Richard Feynman "*Being able to explain or teach a concept shows mastery, but being able to do so simply shows that you truly understand the subject at hand.*"

The paper is to be no more than 5 pages long with normal margins (it may be single-spaced). The 5 pages include all text, but not tables, figures, or bibliography. The order of the report should be main text, bibliography, then the tables and figures at the end (statistical code should be emailed separately).

Oral Presentation (4/16 and 4/18)

Each team will give an oral presentation of 15 minutes. The order of the presentations will be randomized. The course website will contain examples of completed projects and presentations from past years.

Grading

Team average scores for these projects will be assigned based on the following rubric:

- Quality of the hypothesis generation 10%
- Need for longitudinal/clustered methods 10%
- Analysis designed to address aims 10%
- Quality of exploratory analysis (figures and tables) 10%

- Model Selection (covariates, covariance, random effects) 10%
- Interpretation of coefficients (even if they're not significant) 10%
- Readability of the report 10%
- Strength and limitations of the study 5%
- Viability of study conclusions 10%
- Seminar presentation:
    - Keeping the interest of the class 10%
    - Handling questions 5%