Stat 414 Project

Overview:
In this project, your group will select or collect a multilevel data set of interest to you, thoroughly analyze the data using methods from Stat 414 (or new methods that expand upon ideas from Stat 414), and present your results in both a written report and a brief presentation. The data set can come from research you have conducted, friends, professors who have collected scientific data, reputable internet sites, etc.

Groups: 1-3 people

The dates below are tentative, let me know if you need more time/want to meet to discuss your project etc. Finding a good dataset is worth the most time in this project!

Grading:		Points	Tentative Due Date
Part I:	Proposal and Data assembly	5	Oct. 14
Part II:	Exploratory Data Analysis	5	Oct. 28
Part III:	Modeling Results	20	Nov. 18
Part IV:	Final Report	50	Dec. 11
	Final Presentation	20	
		100	

Data sources:
Be sure that your data is rich enough so that there are opportunities for model fitting choices, controlling for covariates, discovering interesting interactions, and generally providing interesting answers to real, compelling research questions (using or expanding upon methods from Stat 414). **Your data should have a multilevel structure**, meaning there are "level 1 units" and "level 2 units" **and** that you have variables for the Level 1 and the Level 2 observations (e.g., patient response and # of weeks since surgery, percentage Black in a county and percentage Black in the state, location of VRBO listing and average income in the county). If your response variable is categorical, we can work with that too, but may be towards the end of the course. It is quite common and you are encouraged to merge data from different sources (e.g., county level observations)

Here is a link to possible data sources in Canvas that you may use as a launching pad if you're searching for data. I'll also post some previous project topics. You may also want to find some analyses with citations to relevant papers as a starting point, but I will ask that your project is original in terms of the analyses you perform and references you find.

Part I: Proposal and Data Assembly

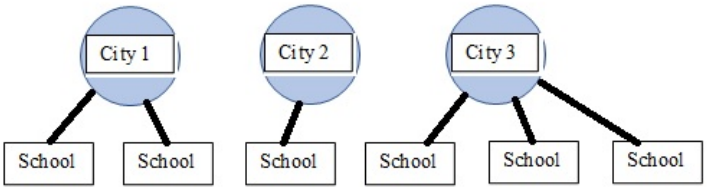
1. Identify the important **research question(s)** which will guide your project (e.g. Do youth who participate in physical exercise class have lower BMI? Does this relationship vary by neighborhood?) – and describe why your chosen project is interesting to you. Provide rationale for each variable included in your proposed data set (e.g., "We may subgroup by gender." "We need to control for diet.")
2. Describe the data source you will be using. If available, include a link to the raw data.
3. Identify the proposed Level 1 and Level 2 units. Ideally your Level 2 units can be conceptualized as a random sample from a larger population of such units (e.g., schools) and there are quite a few of them (so treating them as fixed effects could be a bit annoying) and you have several Level 1 units within each Level 2 unit (does not have to be balanced).

Example: Observational units = secondary school students
Level 1 = students (100 students from each neighborhood), Level 2 = neighborhood (30 neighborhoods)

Or

Level 1 = school (random sample of schools from each city), Level 2 = random sample of cities

Consider including a schematic of the multilevel structure.



4. Complete a *variable chart* (similar to the one below) for your anticipated variables. A typical list will include 6-10 variables. List the variable, whether the variable is quantitative or categorical, the units of measurement for each variable or the categories, and whether or not it is a Level 1 or Level 2 variable. For any variable whose definition is unclear, provide a short definition. As an example, if I were to measure my body mass index (BMI) weekly for a sample of individuals, variables could include whether the individual attended a physical exercise class that week and that individual's age at the start of the study (especially if I want to have age-adjusted effects of exercise class), the first few lines of the chart might read:

Name	Variable Role	Type	Values (units)
BMI	Response (weekly)	Quantitative	> 0 (kg/m)
attended PE class	Level 1 Predictor	Categorical	yes, no
age	Lever 2 Predictor	quantitative	12 to 18 years

Definitions: BMI = body mass index = weight / height squared. A measure of body fat.

Note: If you don't have natural Level 2 variables (or even if you do), an important option is aggregating a level 1 variable across the level 2 units (e.g., average GPA of students in each class, a person's average amount of sleep over the course of the study). Please consider including at least one of these "contextual" variables as well.

Part II: Exploratory Data Analysis

1. Remind the reader of the response variable, Level 1 and Level 2 units (and maybe even the graphic of the multilevel structure of your data). Note any major changes from our Part I report, including group membership.
2. Discuss any initial data cleaning steps based on the data collection method (e.g., redefining variables, creating any new variables).
3. Explore the relationships between your response variable and key predictor variables both graphically and numerically. Depending on the types of your response and explanatory variables, you may consider graphs such as boxplots, scatterplots, spaghetti plots, and segmented bar charts, and you may consider summary statistics (like mean, median, standard deviation) by group, correlations, regression equations, and two-way tables with proportions. At this stage of the project, the graphs can be loose with titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels! Of interest here is whether any variables should be transformed to help with linearity and equal variance assumptions, and whether any categorical variables should be regrouped.
    - I would grand-mean center any quantitative variables.

- If you do plan to include any aggregated variables, also consider creating "deviation" variables for Level 1.

4. Make note of any problematic data and observations that you are justified in removing. Discuss implications about any decisions you make about missing data (e.g., narrowing of the population to which you can generalize).
5. Based on your exploratory analysis, are there any variables that you are now skeptical of including in final models or any that seem like they will be especially important? Has this exploration expanded any of your primary research aims?

Below are guidelines for the final report that you may want to start working on:

- Begin with a short paragraph introducing your project and primary research question(s). (This introduction will be expanded into several paragraphs for the final paper.)
- Use your graphical and numerical summaries to tell a story, supporting your conclusions with summary statistics. Weave numerical summaries seamlessly into your text, and refer to graphs where appropriate. Integrate your output with your discussion. You don't have to discuss every variable or pair of variables, focus on the most interesting pieces of the story.
- Write well! Complete sentences, good flow, proper grammar, the works…
    - Aim your report at audience familiar with 313/324/334-level statistics, but may be a little rusty. Also, they have no specific knowledge on your research topic, but they have the ability to catch on quickly. Explain your terms clearly.
    - Give concise but precise statements interpreting summary statistics, etc. – in the context of your data set and research questions you pose. Avoid vague terms like "this data," "these results," etc. Also avoid cryptic variable names that you may have used in your statistical software.
    - Stat Majors – remember lessons you learned in Stat 365!
- Commented, reproducible output so that I can trace how you constructed your final data set, what the results of your exploratory data analyses were, and what plots and analyses you generated. I should be able to take your data file, go through your data organization and variable creation steps, and ultimately generate your same models.

Part III: Modeling Results

Include the following modeling steps. This may not find the best model, but will be an opportunity for you to build a multilevel model in a coherent fashion. You should be using your cleaned data set *with quantitative variables grand-mean centered.*

1. Include a graph exploring the variability in the response variable across the Level-2 units. Fit an ANOVA using OLS for your response variable and the Level 2 grouping variable (the Level 2 units). Does the variation in the response across the Level 2 units appear to be statistically significant?
2. Fit the "random intercepts only" (null) model. Interpret each of the estimated parameters in context. Interpret the intraclass correlation coefficient in context. Does the value of the ICC seem "substantial" to you? Report the likelihood, deviance, and AIC values for later comparison.
3. Add 1-3 Level 1 variables. Carry out a likelihood ratio test to compare this model to the model in step 2 (using ML, clearly explain how you find the chi-square value and df). Include details. Also report/compare the AIC values to the intercepts only model. Calculate a "proportion of variation explained" for this set of variables and interpret the results in context (be clear variation in what). Did the Level 2 variance decrease? What does the tell you? Remove (one at a time) any insignificant variables.
4. Add 1-3 Level 2 variables. Carry out a likelihood ratio test to compare the models (using ML). Include details. Also report/compare the AIC values. Calculate a "proportion of variation explained"

for each level and interpret the results in context. Remove (one at a time) any insignificant variables.

5. Consider random slopes for one Level 1 variable. (This could involve putting back in one of the variables that was removed earlier…) Include a graph illustrating variability in the estimated random slopes and discuss what you learn in context. Interpret the amount of group-to-group variation in these slopes in context. Once you have a model with at least one set of random slopes, compare this model to the model in step 4, is adding random slopes a significant improvement (REML, be clear how you are determining degrees of freedom)?
6. Add and interpret a cross-level interaction (you may have to use insignificant variables, focus on interpreting the interaction). Are you able to explain much of the slope variation you found in step 5? Is this a significantly better model?

Keep in mind: Doing what I tell you to do is ~ B work. Doing more or less will move your grade up or down. Possible Extras: Enhanced graphs; More than 2 levels; Compare model in step 3 to a random effects ANCOVA model (using OLS); Testing additional random slopes; Cross validation (or at least consider possible multiple comparison issues); Including and interpreting confidence intervals

For Discussion in Final Report

- Briefly resummarize your model building process (what was done at each step? To learn what?).
- Describe your favorite model from those above. Tell which features you like, and *provide interpretations for key parameters*. Make it an interesting multilevel model (e.g., include interactions?). Goal here is to impress me with what you have learned about multi-level modelling more than finding the best model for your research question.
- Residual analysis for this favorite model, including identifying any potentially problematic observations that you are not justified in removing.
- Concerns and future plans. Describe concerns you have with your current models and additional data analysis plans you have.

Don't go overboard on model building for this project, we don't have enough time to "find the best model," just want to get a sense of some reasonable models and how you are comparing them.

Part IV: Final Presentation and Final Report

Your Part IV score will be based on (a) the quality of your team's Final Report, (b) the quality of your team's Final Presentation, as determined by your classmates and me, and (c) your individual contribution to your team, as assessed by all group members. The audience for the Final Report and Final Presentation is anyone who has taken Stat 414 but might not be informed about your particular project.

Final Report

Your report should be a thoughtful, concise, polished, document, no longer than 8 pages. (Lessons learned in Stat 365.) Relevant tables and/or figures should be formatted neatly into your report (but they don't count as part of your 8 page maximum). Be sure to label and reference your graphs and tables so they are interpretable on their own. Any less relevant figures and tables along with important documentation and output should be attached as an annotated appendix to the end of your report (see below for more details). **Upload a copy of your final data file.**

1. **Introduction** A few paragraphs that contain background information, motivation for your research, and a statement of your research goals. Be sure to incorporate any supporting references into the text. The purpose of the background is to place your work in the greater context of the literature in the area you are investigating. Then you should explicitly identify a hypothesis that you will investigate with your analysis. *Don't assume the reader remembers any of your earlier project reports.*

2. **Data Source/Methods** A paragraph or two that…

    - Briefly describe your data, where your data came from (source), definitions of important variables, and how the data were collected. You should include a variables table and/or diagram representing your study.
    - Indicate any modifications made to the data, recoding, or decisions about missing data
    - Briefly describe the methods you used (e.g., multilevel regression) in your analysis
    - Do <u>not</u> report results in this section!

    <u>Note:</u> If you had a more involved data collection process or are using a method not covered in Stat 414, you may choose to expand Data Source/Methods a bit to describe your process/statistical method a bit more.

3. **Results** The meat of your report, with discussion/interpretation of your final *model* (Include the model building as part of the annotated appendix).

    - <u>Summarize</u> the results of your exploratory data analysis (e.g., include a matrix plot?). Which Level 1 and Level 2 variables appeared to be most promising before you built your final model?
    - <u>Summarize</u> the results of your initial anova exploring the significance of the Level 2 grouping variable.
    - Summarize, including (with a useful picture), your null model and what you learn, including interpreting the ICC value in context.
    - Brief discussion of how you arrived at your final model. Interpret all of the estimated parameters in your final model in context. Include at least one confidence interval or interval of random slopes and interpret in context. (But there shouldn't really be much "commentary" in this section.) How much of the unexplained Level 1 and/or Level 2 variation in the null model is explained by your final model?
    - Integrate the most relevant figures to help interpret the results with your discussion. This includes graphs illustrating fixed effects, random effects, and interactions.
        - These tables and figures should be well-labeled, numbered (e.g., Figure 1), and have a good, descriptive caption.
        - Especially effective graphs compare your data to the model and discuss how the model does and does not capture important features of the data. At least consider showing the "effects plots" in addition to the raw data graphs.
        - Each report should have a minimum of two plots (could be several more, just keep them well-sized and well-integrated into the discussion).
    - Diagnostic analysis: Discussion and inclusion of residual plots (with commentary) for your final model

4. **Discussion** A few paragraphs that:

- Describe how the results help answer your research questions and what was most interesting from your analysis. In fact, the first paragraph of the Discussion is very important – in professional journals, it is often the first and sometimes the only paragraph that is read in a paper. After the first sentence highlights primary results, the remainder of the first paragraph might compare your results to others in the literature or include interesting secondary results.
- Discuss possible implications of the results in the context of the research question.
- Identify any limitations of your study. Discuss the potential impact of such limitations on the conclusions. (e.g., potential confounding variables, generalizability - Don't give generic statements of possible causation and generalizability, but thoughtfully discuss relevant issues – confounding variables, representativeness of the sample, etc.)
- Identify strengths and weaknesses of your analysis.
- Make suggestions for future research. Identify important next steps that a researcher could take to build on your work.

5. **Annotated Appendix**

- Make sure you submit your raw data and make sure I know definitions of important variables and the source of the data.
- Your reproducible model building details.
- Tables and figures that are informative but were not referenced specifically in the main report. Include a short annotation – one or two sentences on what they show.
- Description of statistical modeling steps that were not included in the main body of your report. Possible entries here include:
    - o Evaluation of assumptions.
    - o Outlier analysis and how you decided to deal with any outliers along with rationale for your decision.
    - o Describe hypotheses testing you performed during model building and how you decided on the explanatory variables you ultimately included in your final model.
    - o Assessment of the final model.
- If applicable: How you went from the model output in R to interpretations in your report (e.g. exponentiate coefficients, then take inverse)
- Anticipate questions someone might have after reading your report, and make sure those questions can be answered with information in the appendix.
- A citation for any reference article(s) you included in your proposal. Also include a link, if appropriate.

Final Presentation

Each group will prepare 5-minute Powerpoint (or equivalent) presentation describing your findings and any new methods used.

- You can use your Final Report as an outline for your talk; the Results section should comprise the biggest chunk of your presentation, and you can probably skip the Methods

section (unless you're doing something unusual) but do provide enough context for the audience to understand your primary research question(s).

- Use plots to tell your story as much as possible. When you present a graph, be sure to orient listeners to what variables are on each axis, and what the main point of the graph is, etc.
- Avoid big chunks of output.
- You will be assessed based on: organization, verbal presentation, ability to use statistical terminology correctly and confidently, use of graphs to tell your story, success in anticipating and answering audience questions, and ability to hold the audience's interest.

Also consider submitting your project to the competition (next deadline Dec. 2024). I'm very happy to help with this process…(their format is a bit different…)
https://www.causeweb.org/usproc/usresp