Stat 316 – Fall 2019:

The Research Project Overview

**Overview**:

In this project, you will select or collect a data set of interest to you, thoroughly analyze the data using methods from Stat 316 (or new methods that expand upon ideas from Stat 316), and present your results in both oral and written formats. The data set can come from research you have conducted, friends or professors who have collected scientific data, reputable internet sites, etc.

**Groups**: comprised of 2-3 people – assigned by me with your input (complete Project Partner Feedback on Moodle by 5:00 PM on Thursday Oct 17). I will (hopefully) announce groups on Friday Oct 18.

**Grading**: Points Tentative Due Date

## Stage I: Proposal and Data Assembly 10 Wed, Oct 30

Stage II: Exploratory Data Analysis 10 Tues, Nov 12

Stage III: Initial Modeling Results 10 Tues, Nov 26

# Stage IV: Final Report 100 Mon, Dec 16

# Final Presentation 20 Mon, Dec 16

150

**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 316). Your data should either have a multilevel structure or non-normal responses, or both.

I have posted several data repositories on Moodle that you may use as a launching pad if you’re searching for data. Some of these data sets have sample analyses and citations for relevant papers; although those can be helpful starting points, I will ask that your project is original in terms of the analyses you perform and references you find.

**Stage I: Proposal**

1. Identify the important **research question(s)** which will guide your project (e.g. What factors are associated with lower BMI?) – and describe why your chosen project is interesting to you. Provide rationale for each variable included in your proposed data set (e.g. Does taking a PE class help? We may subgroup by gender. We need to control for diet.)

2. Find **references** for *at least* two articles in the refereed literature that are relevant to your question of interest. You should avoid articles that are too technical to be relevant to the project or to be informative for the non-specialist (e.g. “Beta 2-adrenergic receptor polymorphisms and haplotypes are associated with airways hyperresponsiveness among nonsmoking men,” Chest, 2004). Articles that appear in the popular press (such as *The New York Times* or *Washington Post*) or news articles in journals are **not** acceptable as refereed references, although they may help motivate ideas for your project. Be sure you obtain the entire paper and not just an abstract!

You will eventually use these references in the introduction of your paper. Pay close attention to the figures, tables, and methods sections of the papers you select as they can give you an idea of what I'll be expecting from you in your final write up.

For the project proposal, include the following information:

* Give the citation for each reference (in APA format or similar) and a link, if appropriate.
* In 1-2 paragraphs, summarize the primary findings and how they relate to your proposal.

3. Complete a **variable chart** (similar to the one that follows) for your anticipated variables. A typical list will include 6-10 variables. List the variable name, variable role if known (response, explanatory, potential confounder), an indication of whether or not the variable is quantitative or categorical, the range of values for each variable, and the units of measurement for each variable (if appropriate). For any variable whose definition is unclear, provide a **short definition**. Also mention your **observational units**; if your data has multilevel structure, *provide the level for each variable as well*. As an example, if body mass index (BMI) were the response variable, attending a physical exercise class daily is the explanatory variable, and age is a potential confounder, the first few lines of the chart might read:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Name | Variable Role | Type | Values | Units |
| BMI | response | quantitative | > 0 | kg/m2 |
| attend PE class daily | explanatory | categorical | yes, no | NA |
| age | potential confounder | quantitative | 12 to 18 | years |

Observational units = secondary school students

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

4. Outline how you plan to address your research question(s) with the data you have listed in (3) (e.g. We plan to run multiple linear regression models with BMI as response and daily attendance and age as explanatory variables, possibly examining interactions between the two).

5. Describe how you obtained your data, providing a link if appropriate. **Store your data, properly labeled, as a .csv file in the Project folder of our class folder on the RStudio server.**

**Stage II: Exploratory Data Analysis**

1. **Exploratory data analysis**. Calculate or prepare appropriate numeric and graphical summaries for all relevant variables. Summarize the data using methods that are appropriate to the data type.

(a) Most likely you will need to “clean” your dataset first. Make note of any problematic data and observations that need to be removed. Consider implications about any decisions you make about missing data.

(b) Descriptive statistics (5 number summaries for continuous variables; tables of counts and proportions for categorical variables) for all relevant variables in your data set. Plot continuous variables using histograms.

(c) Explore the relationships between important *pairs* of variables both graphically and numerically. Depending on the type of your response and explanatory variables, you may consider graphs such as boxplots, scatterplots, cdplots, 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. Exploratory plots can be loose with titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels!

2. **EDA Report**. In no more than 3 pages, summarize the main findings of your exploratory analysis, referring to specific plots and summary statistics where necessary. In addition, describe your plans for building models to address important research questions, including which variables will be important to consider in light of your exploratory analyses.

A few extra details on the EDA Report:

(a) The **Main Body of your report** should follow these guidelines:

* No more than 3 pages
* Begin with a short paragraph introducing your project and primary research questions. (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.
* Include at least 2 interesting plots (hopefully more!). Name each plot (e.g. Figure 1) so they are easily referred to in your report, and format the figures neatly within your report (without taking up too much space). Exploratory plots can be loose with captions and axis labels, but for your final paper it is essential that your figures have meaningful captions and axis labels!
* Preview directions you plan to go with modeling. What models will you begin by fitting, and what variables will be involved.
* Write well! Complete sentences, good flow, proper grammar, the works…

(b) Your **Annotated Appendix and References** section should include these elements:

* Definitions of important variables and the source of the data.
* R scripts and (commented) 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 you generated.
* A short annotation – one or two sentences – on what each analysis shows.
* 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.
* A citation for each reference article (in APA format or something similar) you included in your proposal. Also include a link, if appropriate. Remember that you must have the entire paper and not just an abstract, and at least two must be from peer-reviewed journals.

(c) General hints for your Stage II report:

* Aim your report at audience familiar with 272-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.
* Show your work and output in an Appendix so your analyses can be reconstructed, if necessary.
* 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 R.

**Stage III: Initial Modeling Results**

Based on the results of your exploratory data analysis in Stage II, you will begin to fit a taxonomy of models, trying to ultimately settle on a single model (or models) to address your research questions. Then you will prepare an Initial Modeling Report with Annotated Appendix as described below:

(a) The **Main Body of your report** should follow these guidelines:

* 2-3 pages, organized into the following sections:
  + Model building process. Describe the steps you took, and the reasons for those steps.
  + Current model. Describe your tentative final model at this point – tell which features you like, and *provide interpretations for key parameters*.
  + Concerns and future plans. Describe concerns you have with your current models and additional data analysis plans you have.
* Write well! Complete sentences, good flow, proper grammar, active verbs, the works. 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”, “looking at”, etc.

(b) Your **Annotated Appendix** section should follow these guidelines:

* Include R scripts and (commented) output so that I could take your R code, read in your data, go through your data organization and variable creation steps, and ultimately generate your taxonomy of models.
* Include preliminary models that you considered along the way.
* Any plot or table referenced in the main body should be labeled (e.g. Figure 1). All other plots and tables can just appear as is.

**Stage IV: Final Presentation and Final Report**

On the date and time of the final you will present your project results. Your Stage 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 316 but might not be informed about your particular project. Additional details on the format of the Final Report and the Final Presentation follow.

**Final Report**

Your report should be a thoughtful, concise, polished, document, no longer than 8 pages. Relevant tables and/or figures should be formatted neatly into your report (since they 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. An annotated appendix containing less relevant figures and tables along with important R scripts and output should be attached to the end of your report (see below for more details).

1. **Introduction** A few paragraphs that contain background information, motivation for your research, and a statement of your research goals. Be sure to incorporate your 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.

2. **Materials and Methods** Three to five paragraphs (or fewer) that…

* Briefly describe your data, where it came from (source), definitions of important variables, and how it was collected
* Indicate any modifications made to the data, recoding, or decisions about missing data
* Briefly but thoroughly describe the statistical inference methods used to quantify the association between your outcome and predictor variables (at least one method must have been introduced in Stat 316). What summary statistics were calculated? What statistical tests were performed?
* Specify strategies employed when building your models
* Do not report results in the Materials and Methods section!

Note: If you are using a method not covered in Stat 316, you may choose to expand Materials and Methods a bit to describe your statistical method.

3. **Results** The meat of your report, which should include…

* A general description of your data (completed via your exploratory data analysis)
* A description of the results from your analyses, including interpretations of parameter estimates, tests, and confidence intervals in context.
* Tables that summarize results and figures that illustrate results. These tables and figures should be well-labeled, numbered (e.g. Figure 1), and have a good, descriptive caption. Each report should have a minimum of two plots; rarely are residual plots part of the main body of the report unless they are an integral part of the story.
* While you should interpret tests, confidence intervals, and coefficients in this section, you should not editorialize here! Save that for the Discussion.

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

* Begin with an accurate summary statement; 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.
* Make a statement regarding potential confounding variables in your study
* Make a statement about the generalizability of your results. Don’t give generic statements of possible causation and generalizability, but thoughtfully discuss relevant issues – confounding variables, representativeness of the sample, etc.
* Identify any limitations of your study. Discuss the potential impact of such limitations on the conclusions.
* 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.
* Do not include test statistics or p-values in this section.

5. **Annotated Appendix**

* 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.
* R scripts and output (annotated) so that I can trace how you constructed your final data set, what models you ran to produce the results quoted in your report, and what intermediate models you also considered.
* Description of statistical modeling steps that were not included in the main body of your report. Possible entries here include:

o How you handled missing data

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.

* 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 each reference article (in APA format or something similar) you included in your proposal. Also include a link, if appropriate. Remember that you must have the entire paper and not just an abstract, and at least two must be from peer-reviewed journals.

**Final Presentation**

During our scheduled final exam time, each group will make an 8-minute Powerpoint presentation describing your findings and any new methods used.

* The time each team member speaks must be approximately equal.
* You MAY NOT exceed 8 minutes! I WILL cut you off in mid-sentence at 8:00!
* 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).
* 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.
* Avoid big chunks of R 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.