# Applied Stats: The Final Research Project

## Overview:

In this project, you will select or collect a data set of interest to you, pose a research question(s) that you will attempt to answer using multiple linear and/or multiple logistic regression, and write a paper in a formal scientific style. The data set can come from reputable online sources, research you have conducted, friends or professors who have collected scientific data, etc.

**Data 231 students:** This is an *individual* project.

**Data 230 students:** For this project, you may work with 1-2 other students from this class (or you may work alone).

## Timeline & Grading:

Stage I: Proposal and Data Assembly... *Tuesday, Oct 19 (5%)*

Stage II: Exploratory Data Analysis... *Thursday, Nov 4 (5%)*

Stage III: Project Paper... *Thursday, Nov 23 (40%)*

Stage IV: Peer review...  *Thursday, Dec 2 (5%)*

Stage V: Revised Project Paper… *Friday, Dec 10 (45%)*

## 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. Many of you are majoring or minoring in something other than statistics – I encourage you to find a research topic in these or other fields in which you have knowledge and/or interest.

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.

A key aspect of this project is to use statistics to answer a research question *in context.* The dataset you choose must have enough documentation to give sufficient context for your project. You should be able to answer these questions from the data documentation:

* How were data collected? From whom, by whom, where and when.
* How is each variable defined and measured?

Because you’ll want to use several explanatory variables and explore interactions, it is best if your dataset contains *at least* 100-150 observations. (Rule of thumb says at least 10 observations per variable to avoid over-fitting). There is no maximum size, though R does start to slow down for *really* big data; if you have a truly humungous data set you may want to take a random sample to work with for your project.

## Stage I: Proposal and Data Assembly

Your proposal, to be turned in as a Word document on Moodle, will include the following:

1. Identify the *original* data source. Include a brief summary of how, from whom, and by whom, the data were collected. The data can come from one source or multiple sources. Students can collect their own data through experiment, survey, or study, but this is not required. Upload a copy of the data set: not a link to it, but an actual downloaded file in a format that R can open. (This is so I can look at it if I have questions about your proposal). If you plan to collect your own data, you must provide details on the design and implementation of the experiment or survey (in that case, you don’t actually need to have the data set “in hand” yet).
2. Identify the important research question(s) which will guide your project (e.g. “Do youth who participate in physical exercise class have lower BMI?”, “Are males more likely to drink and drive after adjusting for confounding variables?”) and describe why your chosen project is interesting to you.
3. Provide a list of variables of interest and their definitions (including units). You should also include rationale for inclusion for each variable and identify the variable type, and whether it may need recoding. A table is a good way to summarize this information; for example:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable name** | **Original definition** | **Units** | **Range or Levels** | **Possible recoding** | **Rationale** |
| bmi | Body mass index | Kg/m^2 | >0 |  | Response variable |
| pe | How many days per week attends Physical education class | Days/week | 0-5, integers | Currently categorical var. Recode to same values but numeric | Main explanatory var of interest |
| age | Age of student | years | 12-19, integers |  | Possible confounding var |
| lunch | Percent of students at the school receiving free/reduced lunch | % | 0-100 |  | Possible confounding var (socio-econ status) |

1. 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. Be sure you obtain the entire paper and not just an abstract! You will eventually use these references in the introduction of your paper. Your proposal should include:

• The citation for each reference (in a standard format) and a link, if appropriate.

• A few sentences that summarize the primary findings and how they relate to your proposal.

1. Briefly outline how you plan to address your research question(s) with your data. (e.g, “I 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. I will also adjust for socioeconomic status at the school level.)

## Stage II: Exploratory Data Analysis

The second stage of your project is to conduct Exploratory data analysis.

1. You may 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. This website shares some simple approaches to missing data (and the relative advantages/disadvantages of each approach):

<http://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html>

1. Produce descriptive statistics (5-number summaries and histograms or boxplots for continuous variables; tables of counts or proportions for categorical variables) for **all relevant variables** in your data set.
2. 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, or segmented bar charts. You may consider summary statistics (like mean, median, or standard deviation) by group, correlations, and two-way tables with proportions. At this stage plots can be loose in terms of titles and labels, but for your final paper it is essential that your figures have (meaningful) captions and axis labels!
3. *You should not be fitting any models at this point.*

### Your EDA report, to be turned in on Moodle, will meet these guidelines:

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. This report should be meaningful and readable to someone familiar with statistics but unfamiliar with your particular research topic and dataset (i.e. your professor). 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. A report like this might be something you’d share with collaborators or store as a reference as you proceed with your analysis.

(a) The Main Body of your EDA 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.

• You do not need to include EVERY plot or table you made in the report. Include at least 2 interesting plots/tables (if not more!). Name each figure (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). These exploratory plots/tables don’t have to be perfect in terms of titles and 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. This should be the last paragraph of the report.

• Write well! Complete sentences, good flow, proper grammar, etc.

(b) Your EDA report should also include an Annotated Appendix and References section (not included in the 3 page limit) which include these elements:

• Clear definitions of important variables and the (properly cited) source of the data.

• 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 a standard format) you included in your proposal. Also include a link, if appropriate.

(c) You should also upload to Moodle:

* An RMarkdown document with R code 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. Comments should be short, but clarify what you’ve done and why.
* The knitted Word document (not html) that was created from the RMarkdown above.
* The data file (as .txt, .csv, or .xls).

## Stage III: Project Report

Your report should be a thoughtful, concise, and polished document, **no longer than 8 pages** (Calibri 11 font, 1” margins, 1.5 line spacing). Relevant tables and/or figures should be **formatted neatly into your report** (since they count as part of your 8 page maximum, although they may have smaller font and single-line spacing). 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 code and output should be attached to the end of your report (see below for more details).

The report should include four LABELED sections plus bibliography and annotated appendix. (no page limit on bibliography and appendix). The rubric that I use to grade the reports is on Moodle.

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 the research question that you will investigate with your analysis.

2. MethodsA few paragraphs that…

* Briefly describe your data, where it came from (source), definitions of important variables, and **how it was collected** (random/representative sampling? From what population?)
* Indicate any modifications made to the data, recoding, or decisions about missing data. Be sure to report the **sample size** and the number of observations removed, if any.
* Briefly but thoroughly describe the statistical methods used to investigate the association between your outcome and predictor variables. What summary statistics were calculated? (Just list: e.g. proportions, and conditional proportions for categorical variables; means, medians, and standard deviation for quantitative variables). What statistical tests were performed? What type of modeling was done?
* Do not report results in the Methods section!

Note: If you are using a method not covered in this class (e.g., a nonparametric method, time series), you may choose to expand Methods a bit to describe your statistical method.

3. Results The densest (but not necessarily longest) section of your report, which should include…

* A general description of your data. (This is where you integrate your exploratory data analysis from Stage II.)
* The final model, presented in a table with variables, coefficients, standard errors, p-values, and confidence intervals for each coefficient.
* A description of the results from your analyses, including interpretations of parameter estimates and/or 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 good, descriptive captions. Each report should have a minimum of two tables/figures. Rarely are residual plots part of the main body of the report unless they are an integral part of the story.
* It may be helpful to refer to the articles you’ve cited as a guide for the writing style of the results section.
* While you should interpret confidence intervals and/or coefficients in this section, you should not editorialize here! Save that for the Discussion.
* R2 (for linear regression) or misclassification rate (for logistic regression) should be reported here or in the Discussion (or both).
* Keep in mind what you learned about this section from the Group Project!

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. You should demonstrate your knowledge of these issues and their importance **in context**; the research topic is the focus here, not statistics for statistics sake.
* Identify any limitations of your study. Discuss the potential impact of such limitations on the conclusions.
* Identify strengths and weaknesses of your analysis. **This is where you should (briefly) mention whether the conditions of the model are met.**
* 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, although you can of course reference Results when discussing your overall conclusions.
* See the end of this document for some additional advice about writing Discussions.

### Bibliography

Include full citations for any papers cited in your introduction or other sections of your report. Also cite the dataset you use as appropriate. You may use any bibliographic style you are familiar with (APA, MLA, Chicago, etc), but you MUST be consistent with this style (use the same style for both the bibliography and in-text citations).

### 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 code and output (commented and 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:
* How you handled missing data or recoded variables
* Evaluation of assumptions (including residual plots or empirical logit plots)
* How you went from the model output in R to interpretations in your report (e.g. exponentiate coefficients, then take inverse)
* Steps in your model building process -- how you decided on the explanatory variables you ultimately included in your final model.
* Two tips on constructing this section:
  + Anticipate questions someone might have after reading your report, and make sure those questions can be answered with information in the appendix.
  + *The easiest way to create the appendix is to use R Markdown*, then knit it into a Word document (just to create this section, not the entire report).
* The appendix should be uploaded as a separate document on Moodle and titled “Appendix”.

## Stage IV: Peer Review

An important part of any work is reflection and review. When you submit your paper to Moodle, it will be reviewed by another student, and you will be asked to review as well. More detailed instructions on this exercise will be given at the time of the review. You will be evaluated on your effort in the review, including how constructive and insightful your comments are.

## Stage V: Project Revision

After submitting your project report, you will receive feedback from your professor and 1-2 peers. You will have the opportunity to revise your project paper and resubmit it. A portion of your revision grade is based on the adequacy of your response to comments. (This means that if you just resubmit the first copy of your paper without changes, it will receive a lower grade.)

## Writing Tips

* Recall what you learned during the Group Project, including:
  + Results writing tips (see Group Project Guidelines)
  + *Never* include R code or output in the body of your paper
  + CI's should be reported as (#, #), including in Tables.
  + Everything should look professional and organized, including page numbers, numbered and captions tables/figures, etc.
  + Remember your audience: a scientist interested in your research question. This scientist has knowledge of statistical techniques and interpretation, but is NOT a statistician.
* Go back and re-read the documents by Jane Miller (on Moodle), especially her advice about the “Goldilocks Principle” when interpreting coefficients.
* In the caption of the Table with your model results, state the response variable, including units if applicable.
* Keep # of decimal places less than 4 (unless you have a very small value; e.g., 0.000000345).
* Don't use codes in your prose unless absolutely necessary.  For example, if you're using "Return on Investment" = ROI, this is fine because it's a well-known and easy-to-remember acronym.  But "sem\_trim\_15" is meaningless, so don't use it!  If you must use codes, *italicize*them for easy reading.
* **Some tips for writing the Discussion section**. As stated above, in the Discussion you should do the following (among other things):
  + Identify any limitations of your *study*. Discuss the potential impact of such limitations on the conclusions.
  + Identify strengths and weaknesses of your *analysis*.

It can often be difficult to differentiate between these two, but they are two separate issues.

*Limitations of the study* often have to do with study design and/or data collection... is it generalizable? If not, why not? Are there potential measurement errors or other biases? (e.g. students feel social pressure to lie about drug use, sex, etc when asked in survey). Are you trying to measure something that isn’t easily defined (e.g. “Very happy”, “Somewhat happy”, etc)?

*Strengths and weaknesses of the analysis* are statistical decisions that you made which may strengthen or weaken your potential conclusions. This is also where we mention that model conditions are met (hopefully!), or discuss possible problems with those conditions.

Example of a strength: Very large sample size means very high power (especially interesting if you fail to reject null hypothesis... this means it has very low probability of a type II error... and even though we can’t officially *accept Ho,* this is stronger evidence that Ho may really be true.)

Example weaknesses:

* Had to combine two or more categories of responses into one in order to use your chosen statistical method (e.g. had to make the explanatory variable binary for logistic regression), which means you now can’t make inferences about those groups separately.
* We treated a variable as continuous even when it was recorded as {1, 2, 3, 4, 5, 6 or more}. The numeric value of the last group is questionable (would analysis change if we replaced it with the numeric value of 7 instead of 6? 6.5? 8? which of these is most representative of the answers “6 or more”?... we don’t and can’t know).

These limitations and weaknesses are natural suggestions for further research. When suggesting further research, feel free to make suggestions of things you can’t or don’t know how to do (yet!), e.g. include another variable (not in your data set), or examine a non-linear relationship.