MSDS 6372 Project Guidelines

Important Dates

Project 1: Analysis of a data set using regression techniques
Due date: End of unit 5

Project 2: Analysis of a data set using dimension reduction techniques
Due date: End of unit 10

Project 3: Group project applying various techniques with data sets given by the instructor
Due date: End of unit 15

Purpose/Objectives

With these projects, you will demonstrate that (a) you can execute a data analysis using appropriate techniques, and (b) interpret the results in a practical or realistic context. Stated another way, you will *apply* a variety of processes that you have learned to a single set of data in a clear and well-organized way. Please recognize that this is not a term paper. It is an *exercise in data analysis*, and so if you have a reasonable understanding of how to do the analysis, *the data will tell you what to write*.

Do not lose sight of this crucial reality: ***I do not intend for you to spend every waking hour on each project. You do not want to get bogged down in any one section of it, and you do not want to spend 8 hours a day for 4 weeks on it. There are limitations to what I can expect and to what you can reasonably accomplish. So, stay focused on moving forward as you work on it.***

Basic Requirements

Each project should be **no more than seven pages** of single-spaced or double-spaced text (or anywhere in between). Please use 11-point font and 1-inch margins. The page limitation **includes graphics and tables**. Graphics are both necessary and inevitable. Use only the ones that are helpful. A handful of useful and clear visuals are vastly superior to a lot of confusing or unnecessary ones. Graphics should be interspersed throughout the text. Each graphic or table should be clearly labeled and discussed in the text. They should also appear as close as possible to the text where they are discussed

If it is necessary, you may include more tables and figures in an appendix of no more than 10 pages. Your SAS code should also be included in the appendix. (The code is included in the 10 pages.) Any table or figure that is included in the appendix must be referenced in the text. **I do not want 60 pages of SAS output attached to the project, and anyone doing so will be asked to rewrite his or her**

project and be deducted half a letter grade for every day the rewriting takes. Learning what to include and what not to include is part of becoming a good writer.

Obtaining Data

For projects 1 and 2, you are to analyze a preexisting data set that you find on the Internet, in a textbook, or in some other source. You can even collect one yourself if you want. The data set should be large enough that software is necessary for the analysis. It is OK with me if you use data from work or another research project, but please clear the use of the data with the owner of the data.

For the third project, the instructor will supply several data sets from which you may choose. You may work on the third project with a group or individually. It is OK if two individuals or groups choose the same data set. The data sets will be rich enough that several types of analyses are reasonable. The first two projects are individual projects.

Evaluation

In general, expect the essay to be weighted about 80% on content and 20% on organization and presentation. Keep in mind that projects are generally much more sensitive to sincere and obvious effort over and above actual results.

Written Paper Structure

Please note that this outline gives you an idea of what the minimum content ought to be for your essay. You are free (and encouraged) to supplement this as you see fit. Please recognize that there is a high degree of flexibility here as long as the basic objective (described above) is being accomplished.

Introduction

Briefly introduce the data and give some background on the experiment/study. Again, this should not be a comprehensive literature review. It should be just enough to motivate the problem you are addressing with these data.

Section I. Descriptive Statistics

Begin by describing your data set (but don't repeat anything from the introduction). *Among other possible things to address are the following:* Where did the data come from? How was the data generated (i.e., what kind of collection method is being used)? Identify the specific variables and how they are measured (if you know this). Do you have continuous numerical data or proportions? Briefly explain what the expected relations will be between them (i.e., which one(s) are dependent and which are independent—this will be based on whatever semilogical relation you happen to see there), etc. If you've done any manipulations to the data—like sampling it to reduce the size of the data you're working with—you should identify that here.

Write a descriptive summary for each of the variables you are working with. It is up to you to decide exactly which strategies you want to use and how you want to present them. At a minimum, you want to do some form of numerical summary and some form of graphical summary for each one. Depending on the nature of your data, these might be a five-number summary, calculation of the mean and standard deviation, a time series chart, a frequency histogram, and/or any of the various other data summary strategies that we have discussed. ***It is important to recognize that the objective here is quality, not quantity***. I don't want you to do one of everything you learned; choose a plan that briefly but usefully summarizes your different variables.

Based on the introduction of variables and the descriptive summaries, write a brief statement about what kinds of inferences or hypotheses you think the data are suggesting. These might relate to predictive powers and probabilities for single variables, to the relationship between dependent and independent variables, or other inferences you might find. The point is that you can (and want to) use this to lead into sections II and III.

Section II. Analysis

In this section, you will evaluate the relationship between the variable(s) you presume to be dependent and independent.

Do an appropriate analysis for the associations that you anticipate in your data. How you organize this is up to you, but you want to utilize an organizational logic that makes sense.

The analytical model you use should be determined by whichever makes the most sense given the form your data takes (for project 1: regression, for project 2: dimension reduction, for project 3: anything reasonable goes). Please be sure that you include all of the appropriate analytical components in your analysis (e.g., checking assumptions, residual analysis). The presentation of your analysis should include both a tabular summary of the key coefficients and calculations (please don't put the entire output into your paper; just the stuff you are actually going to use) and a discussion of what they are and what they mean. In other words, show the key calculations and explain what they tell us.

Section III. Interpretation & Conclusion

Address whether the outcomes you have found with this analysis are consistent with the general descriptive interpretations you identified in section I. Also briefly discuss the applicability of the results to a general population. (If the subjects were all college students, do the results apply to general population of adults?) Mention any other strengths and weaknesses of the data and your analysis.

Note: References are not required, but if you use material from a book, an article, or a website, you must cite the source and use quotation marks for any paragraph (or part thereof) that is quoted word-for-word from the source(s). I expect you to do this properly. Plagiarism is intolerable at the upper-class/graduate student level. ASK if you are unsure how to quote something from an outside source. Please look up how to cite references if you have to. Any style is acceptable, as long as you are consistent.

General Writing Skills

This is not a writing class, but you are definitely expected to write with clear prose and good organization. Written communication is important in statistics and in the workplace. A couple of things to keep in mind:

1. **Failing to proofread is intolerable at the graduate level**. Even the best writers must write, check, and rewrite until they have edited themselves into a well-written essay. Three revisions of a completed draft seem to be a **minimum** to produce good work, but you are certainly encouraged to do more! (Aside: I had a mentor at my first job that asked to see EIGHT to TEN drafts of a grant before he would even think of submission!) Please ask someone else to read your work once you have a working draft. Everyone has a tendency to become so familiar with his/her work that the mistakes do not present themselves, even on multiple rereadings.
2. **Typographical errors are intolerable, especially with modern spell- and grammar checkers**. Expect to be penalized if you leave more than three of these in your essay. Be careful about the proper use of words like “sever” and “severe,” and “to,” “too,” and “two,” which would not be caught by a spell-checker.
3. Good organization means that (a) each paragraph is ordered logically with respect to both the one preceding it and the one following, and that (b) paragraphs should always begin with some form of a logical transition from the idea that precedes it.
4. Watch out for redundancy in both word usage (using the same word multiple times in the same or consecutive sentences) and in ideas (stating the same thing more than once or twice).
5. Avoid just doing a “list essay,” or one where every paragraph and/or section is organized with “First…,” “Second…,” “Third…,” etc., types of transitions. You can and probably will use some of this, but you want your prose to flow smoothly and to vary its approach enough that reading it doesn’t feel terribly rigid.

*** In following through on point 1 above, you should be focusing on points 2 through 5.

In addition to the three broad sections of the paper, “good writing” necessitates that you also have an introduction and a conclusion. Given the nature of the project, these can both be relatively brief, but they should serve an appropriate purpose (i.e., to introduce and to conclude).