Stat 21 Final Project - Part 2

Swarthmore College

Due: May 6, 2021

Contents

This assignment is due on to be submitted on Moodle by Monday **April 26**. This is the second of four parts that will comprise your final project for this class. This part of your final project is worth 50 points. The purpose of this exercise is to decide upon a final project topic and to spend some time developing a strategy for completing your project step-by-step.

There are three broad options for you to choose for your final project. Each of these three options has their own statistical merit *and* each option is customizable so that you can spend time researching something that is important to you. My hope is that this project is more enjoyable than laborious. For each of the final project options I've provided a guide on what to include in your report.

To receive credit for this part of your final project, you must submit a final project proposal to Moodle by April 26th. In this proposal you must specify

- Which option are you choosing for your project
- At least two peer-reviewed resources you think you may want to reference in your bibliography.

This information will be submitted in a text box of no more than 300 words.

Option 1

Present a critique of a published paper that presents either an ANOVA model or a MLR model. The objectives for this option are for you to gain experience critically reading and analyzing scientific articles, apply what you have learned in class to understand a published statistical analysis, and to practice communicating statistical information. Note that just because a paper is peer-reviewed, does not mean that it is without error or even that it is a quality application of a statistical method. You are meant to evaluate the work as much as learn from it, you may find an application that you believe is an inappropriate application of a statistical model we covered in class. If this is the case, justify your critique!

One way you can find articles that are of interest to you, is by going to https://www.jstor.org/ and in the search bar enter "linear regression" (for example) in quotation marks. Then refine your search on the left hand side to filter only by "Journals" and then filter again by the subject area of your choice. This is an opportunity to further your knowledge in any area while simultaneously practicing statistical reasoning.

What should the project include?

1. A explanation of the problem or purpose presented in the paper. What problems or questions did the researcher set out to investigate?
2. Background or literature review. How did the researcher build their statistical model? How was the data collected and why were these variables chosen?

3. Methodology. What model did the researcher use and why? What software did they use? What kind of statistical summaries and graphical techniques did they use?
4. Results and conclusions. What conclusions did the researchers draw from their analysis? How do you interpret the results of their model?
5. Discussion and critique. What did you learn from this analysis? Were there any weaknesses to the methods used? Do all the necessary assumptions seem reasonable? Do you agree with the conclusion or do you think there is a better way to address the research question?
6. References. You must cite your sources throughout your report and list them all at the end of your document. (APA or MLA format is fine.)

Option 2

Find a data set and build your own statistical model: ANOVA, SLR, or MLR. The objectives for this option are for you to develop an appreciation for the difficulties of sharing and interpreting raw data, to practice building a statistical model from scratch, and to practice communicating statistical information.
If you do this option you may want limit your analysis to estimation and data exploration (rather than inference). It is incredibly difficult, time consuming, and costly to obtain a random sample of data to use for inferential conclusions. This is really only verifiable with experiments but experimental data is often kept private. It can be really fun however to find a data set on a topic that interests you so I want to keep this option open to you. My advice to you is this: keep the model simple (you don't need to have a bunch of predictors for example) and supplement your estimated model with many plots! If you do preform any inferential procedures, then make sure you clearly articulate under which circumstances are the results generalizable and to which population.

There are many places online to find data on a topic that may be of interest to you. Feel free to use your own data set but I recommend not spending more than a couple of days looking for a data set to use. (It's far too easy to get distracted by all the cool data you could analyze and then become overwhelmed by too much data with too many issues.) To help you find some interesting and relatively clean data sets, here are some resources you may wish to explore:

- Data from the city of Philadelphia
- Inter-university Consortium for Political and Social Research
- Criminal justice data in the US
- Data on fatal police shootings in the US.

What should the project include?

1. A explanation of the problem or purpose. In what context is the model useful and applicable?

What problems or questions did you set out to investigate? What are the key issues raised? How were the data collected?

2. Background or literature review. What sources or background readings did you consult? What literature exists on your topic?
3. Methodology. What did you do, and how did you do it? What statistical and graphical techniques did you use?
4. Results and conclusions, the summary and presentation of your data analyses. What did you find out? This might include tables, graphs, or verbal summaries.
5. Discussion and critique. What did you learn about the problem or question you set out to investigate? What were weaknesses and strengths of your analysis? If you had more time or resources, how could your project be improved?

6. References. You must cite your sources throughout your report and list them all at the end of your document. (APA or MLA format is fine.)

Option 3

Use the sills you have learned this semester to explore a statistical model that we did not cover in class. Some topics you may wish to cover could include

- Probit regression;
- Logistic regression;
- Principle component analysis;
- Random effects models;
- ANCOVA models;
- Two-way ANOVA models;
- Time series models;
- and more!

The objectives for this option are for you to understand how the fundamental modeling techniques we learned in class can be generalized to more complicated settings and to communicate the big-picture of a new statistical model to a general audience.

What should the project include?

1. An introduction to the model. When is this model useful? What questions does this method allow us to answer? What kind of relationships does this method allow us to investigate.
2. Background and example. Which textbook did you consult to understand this method? What is a recently published example application of this model and how was it applied?
3. Methodology. How does this method compare to those that we covered in class? Describe any similarities and/or differences. What assumptions are necessary?
4. Application in R. What functions or packages exist that allow users to build this type of model? How do you extract the relevant pieces of information about this model (like p-values or estimated model parameters)? Can you provide a toy example and interpret the output? (You may use examples from the R documentation if you wish.)
5. Results and conclusions. What are some useful estimates that can be derived from this model? What type of visual summaries are useful in applications of this model? What are some limitations of the method?
6. References. You must cite your sources throughout your report and list them all at the end of your document. At least one of your references must be a textbook and another must be a peer-reviewed paper that applies the method. (APA or MLA format is fine.)