Semester-long Project

Math242: Applied Statistics — Fall 2016
Instructor: David White — Proposal Due: October 17
Presentation: Last week of classes — Paper Due: Monday, December 19

You will work on a semester-long project to apply the modeling techniques we learn to a dataset of your choosing. It will be worth 10% of your grade in the class. You should apply the concepts as we learn them, and should seek to learn now modeling techniques as necessary to conduct your analysis. You will present your findings in the last week of classes, and will turn in a written paper (via email!) on the day of the final exam. The first step is to find a data set of sufficient complexity to cover topics from chapters 1-15. The data set itself is up to you; it should have a large number of entries (more than 200) and a large number of columns (more than 10). You should think of the final project as being about as difficult as three of the weekly projects. You may team up with one other person for this project, but then your joint project should be approximately four regular labs.

Your dataset should have some categorical explanatory variables, so that you can demonstrate modeling techniques using them. Remember that you can always *make* a quantitative variable categorical, e.g. turning number of hours of sleep per night into the binary variable that tells whether or not an individual got more than 7 hours of sleep per night. You will be allowed to use this method if you find a dataset you want to analyze that lacks categorical variables.

Please begin looking for your data set on the internet as soon as possible. Since this is a course on the analysis of data, not the gathering of data, I do not want you to try to gather the data from scratch. Instead, use data that has already been gathered by someone else. The most difficult part will be getting data, reading it into R, and cleaning the data. Depending on how difficult this is, it can be worth up to 20% of the grade for the project. I also have data sets (already clean) based on the work of Denison professors and I am happy to share these with you. Please see the last page for a list of possible applications. Two different groups should not analyze the same data set. Above all, choose something that is interesting to you.

Your project will be graded on the following criteria:

- Proposal & data set - did you successfully get data and explain clear, statistical questions you wish to answer.
- Models - did you analyze your data set using tools from each chapter.
- Conditions - did you clearly state all your assumptions, both regarding the data set (e.g. how it was gathered) and each test you conducted.
- Assumptions - did you clearly state all assumptions required for your analysis to work, e.g. regarding the data set (e.g. how it was gathered), outliers, distributions, etc.

- Language - did you articulate your goals and results both mathematically and in English, so that a non statistician could read your paper.
- Visualization - did you include all graphs, analyze them correctly, and make effective use in your write-up of data visualization.
- Presentation - how well you present your project.
- Writing quality for the paper.

You will present your project in the last week of class. Make it something interesting!

Project Proposal

A typed-up two-page proposal of what you would like to do your project on is due in class on **Monday, October 17th**. Give as much detail as you can about exactly what you would like to accomplish. You should start by telling me if you're working alone or with a partner (and who your partner is). You should have a data set by this date and should have already looked at the data. Your proposal should discuss

- How to get your data set into R.
- What statistical questions you hope to answer using this data set.
- What models you hope to build, and for which response variable. Also, which variables are quantitative and which categorical.
- Possible sources of bias in the data and how you plan to deal with bias if it occurs.
- How the data was gathered originally and how you would change this methodology if you were gathering the data, to correct for selection bias and measurement bias. If you do not know how the data was gathered, then discuss how you would gather it yourself if you had to, and how you will be certain the data set does not contain bias.

This last item will serve as a rough draft for the methods section of your paper, and will give you a chance to build in feedback from me in the form of rewriting. As we move into the unit on statistical tests you will conduct each test on your data set and collect the results for use when you write the final project.

Paper

The deadline to submit your paper is 9am on December 19 for both the R code, the cleaned data set, and the written part. Please make sure that it is copied over to your shared drive space and that the code can be run from there (e.g. if it's reading in data then the data should be there too). Please additionally email me with the final paper, which should be possible for me to read without looking at the R code or data (i.e. the paper should include any graphics you use). You should write your report in R Markdown and should include the Knitted HTML or PDF file so that I can read it.

Your paper should include an **Introduction** discussing the field of study the data set came from, the questions you answered using the dataset, why these questions are important, and what the answers were. Your introduction should be written to an audience of interested professionals in the field the data came from, but not necessarily people with a background in statistics. Your introduction should make it easy for such individuals to use your report. Do not write with me as the intended audience.

Your paper should include a **Methods Section** with a discussion of how you got the data (e.g. what website, how you cleaned it), how it was originally gathered (i.e. the original study that produced the data), any bias in the data (whether it's there, how to deal with it, and how to reduce it in future studies), and any assumptions you had to make in order to conduct your analysis. You should include citations to your references section, showing me where you learned what you learned about the data.

Your paper should include a **Results and Conclusions Section** with the models you built, model utility tests demonstrating their use, results of all the statistical tests you conducted, how to interpret these models and results in English, demonstrations of why the conditions for each model and test are valid, and assessments of the models. Please include graphs and plots to visualize these models, but put them in an **Appendix** and reference them from the Results and Conclusions section as Figure 1, Figure 2, etc.

Your paper should include a **Conclusion** written for my eyes. This is where you should tell me about difficulties you faced in the project, information about how well you worked with your partner, and what you learned from the project.

Your paper should include a **References Section** and in-line citations for information you learned from external sources. At the minimum, this references section should include the URL where you got the data, but I expect it will also include citations to the sources you used to gather information about the data itself. Using external sources without citation is always a violation of the honor code, so please don't forget to cite your sources, even if you don't quote directly from them.

I would expect the written portion of the report to take 8-10 pages and the graphs to take several extra pages. It would be best to intersperse the graphs with the exposition to help you make your points.

Your final paper **must** analyze your data using tools from each of the units of our course. In particular, you must:

- Use graphs to present your data visually
- Present numerical summaries of your data (mean, median, IQR, etc)
- Attempt to fit a probabilistic model somewhere (e.g. a normal model for residuals)
- Use Q-Q plots to analyze whether or not the probabilistic model fits
- Build a multivariate linear regression model including some categorical explanatory variable(s). Use one of the model building techniques to find the best set of explanatory variables (and explain your process in the Methods Section).
- Use confidence intervals to give a range of estimates for some unknown parameter(s)
- Make and use prediction intervals to make predictions.
- Use hypothesis testing (e.g. to test if an estimate from a previous study still holds, to test if a regression model is useful)
- Record and justify any assumptions used
- Interpret your final model in context, i.e. writing sentences about the meaning of the coefficients in real-world terms.
- Use VIFs to check for multi-collinearity.
- Use the test for randomness.
- Use hypothesis testing and model utility tests (with the appropriate degrees of freedom). For multiple linear regression you MUST use the ANOVA F-test rather than testing each slope individually, to avoid drastically increasing your probability of a Type I error.
- Discuss R^2 and adjusted R^2 and include in your paper the meaning of these numbers in context.
- Test for outliers, influential points, and points with large leverage. Decide what to do about such points and justify the assumption.
- Use transformations if necessary.
- If necessary, use polynomial regression or regression with interaction terms.

Each method should appear somewhere in your report. You may also (but do not have to) use more advanced topics such as weighted linear regression, robust standard errors, robust regression, principal component analysis, etc. that we didn't cover in depth in class. You'll know when you need an advanced topic at the moment when you face a problem you don't know how to handle with what we covered in our course. At that point, please come talk to me so I can suggest an appropriate way to handle the situation and give you a reading.

If done correctly, your paper will help you when you apply for jobs, because you can share your abilities in statistics and with statistical software with potential employers. Write the introduction with that audience partially in mind.

Final Project Presentation

Individuals will have 10 minutes to present their projects while groups of 2 will have 15 minutes. The presentation is part of the grade for each individual, so you should try to find a way to tag-team the presentation so that both get to present. Part of your grade is to stay within the time limits, and this will be the hardest part for many of you. I'll give you a 1 minute warning, but it would be good to practice your talk beforehand to work on timing. Focus on:

- Presenting the data set and where it came from.
- Presenting the questions you hoped to answer
- Presenting the statistical tests and models you used, their results, and why the conditions were satisfied.

You do not need to run R code as part of the presentation and you should avoid showing code. Instead focus on what you are excited/passionate about with your project. The talk should be aimed at your peers, not me; you will get your chance to communicate with me in the written portion. You should think of your presentation as a report on what you found to a board of directors. If you use slides please submit them to the shared drive after your presentation. If done correctly, your final paper and presentation will help you when you apply for jobs, because you can share your abilities in statistics and with statistical software with potential employers.

Sample Interdisciplinary Data for Projects

- Social Justice - dataset from Franklin County (Columbus) featuring people on welfare and jobs willing to hire them. What is the best matching?

- Data from around the world - Chicago on crime, locations, salaries, food safety, hospitals, etc. Chinese Census Data, CIA Factbook, Global Terrorism Database, Religion data, UK government data.
- Bureau of Labor Statistics. Data on traffic stops (including race).
- Human interactions and preferences (e.g. from Google, Facebook, Twitter, Uber, Netflix, Wikipedia, Pew Research Center, Salary Data, American Time Use Survey, teaching evaluations, academic publications), entrepreneurship (data on MBA programs, world bank data, GEM data), activity generated data (e.g. web tracking, crowd sourcing, flu trends, apps, Pandora, driving data, MathOverflow).
- Biology - data on the numbers and types of fauna in a particular park over many years, human movement data, genomics, ecological forecasting, epidemiology, healthcare data (NESARC, adolescent data), NIST data, Biotechnology data, ecology data.
- Physics - protein folding, human movement data, statistical mechanics applies statistics to small particle theory. Astronomy - Radio jet data, FITS image data
- Geoscience - ice core data for global temperatures, Paleobiology Database.
- Neuroscience - neuron firing rates as objects move in field of vision, connectomics, determining what traits are predicted by IQ, classifying personality types
- Chemistry - understanding reactions and energy via simulation, analyzing concentrations of a chemical in a sample
- Psychology - PsycINFO repository, data on brain disorders and genetic correlations
- Political science - data on polarization in Congress; socially supplied data, data from religious leaders and congregations, bureau of labor statistics, voting patterns
- Economics - Conference Board, World Bank, financial data, linear regression to relate economic quantities, confirmation of supply and demand in various sectors of the market, gender vs. pay, race vs. pay, education vs. crime, etc.
- Computer science - analysis of social networks and how information spreads, dataset of 10 million leaked username/password pairs.
- Education and Sociology - general social survey, NAEP database, CUPP report, CCD (common core) data, Ohio Dept of Education.
- History - trans-Atlantic slave trade database.
- Sports (sabremetrics in baseball, the role of data in football, soccer, and basketball)
- Computational linguistics - using data to translate dead languages, natural language processing, Google Ngram data.

Data Repositories Online:
https://archive.ics.uci.edu/ml/index.html
http://libguides.denison.edu/ECON407F2015
http://www.kdnuggets.com/datasets/index.html
https://www.kaggle.com/
http://www.bradthiessen.com/html5/data/
http://rdatasciencecases.org/Data.html
http://www.icpsr.umich.edu/icpsrweb/landing.jsp
http://deeplearning.net/datasets/
http://www.icpsr.umich.edu/index.html
http://lemire.me/blog/archives/2012/03/27/publicly-available-large-data-sets-for-database-research/

`http://www.datasciencecentral.com/m/blogpost?id=6448529%3ABlogPost%3A307383`

Check-Ins

Every Monday in November and December, you should upload a check-in to the Semester-long Project folder in your Assignment Inbox on the shared drive. Different check-ins should have different dates, and should be saved as RMarkdown files. Each check-in should include a partial analysis, as you apply what you are learning each week to your pet dataset. This means you should plan to have your dataset cleanly read into R by November 1. If you need help with this, please reach out to me or to Gege Tian (tian_g2@denison.edu).

Grade

Here is an approximate rubric I will use to assign your grade.

Proposal	10%
Check-ins	15%
Presentation	15%
Data cleaning & Analysis	10%
Paper	50%

If your data situation requires more cleaning than usual (e.g. extracting data from a PDF) then it can be worth more, and the paper slightly less.