STAT 215: Group Final Project

Colin Reimer Dawson

Presentations and Writeups Due 12/17, 7pm

This project is very much in line with MP1 and MP2, except that

(a) You will work in groups of 2 or 3

(b) I'm not providing example topics

(c) You will ideally use more than one model type. This could be linear and logistic regression, if you have a binary and a quantitative response that are of interest; it could be ANOVA and regression, to examine a quantitative response in two different ways. Alternatively you could stick with one model type, but explore some technique that goes just a bit beyond what we did in class.

(d) You should provide some more background and discussion on the topic and how your findings connect to other questions of interest than you likely did for previous projects.

Your goal is, quite simply, to build models using one or more datasets to help you say something about a topic of interest to you.

You should use any appropriate tools and methods drawn from those you have learned about in class. If you want to do a bit of exploration into a modeling topic we haven't talked about, even better! (But this is not required)

As always, document the process and write about the results in Markdown.

As with MP2, strive for a writeup that looks like a research paper, not a lab notebook. The following snippet added in your first code chunk will suppress code, unwanted messages, and text output from the final Knitted document. Note that this means that you cannot rely on raw R output to convey your results: you will need to refer to key results in the text. (If there is particular output you want to include, you can put it in an Appendix section, and set `results = 'markdown'` for those chunks.)

```
opts_chunk$set(
    echo = FALSE,
    message = FALSE,
    results = 'hide'
)
```

The Write-Up

You should turn in a paper that describes your questions, why they are of interest, and (very briefly) what others have said about it before. You should document your data exploration, and the CHOOSE, FIT, ASSESS cycle, making use of any metrics, tools, code, etc. that we have used in class (or perhaps dipping a bit into other methods that are “adjacent” to what we’ve done).

There is not a set path of model-fitting and assessing that you must follow, but here are some things to be sure to include:

1. An introduction section, setting up your question, your data, and how you will approach that question using statistical modeling
2. Before fitting any models, set aside a random subset of your dataset (perhaps 20% of cases) that you will use only to evaluate prediction error for your handful of “finalist” models (that is, don’t estimate coefficients on the test set, just use your fitted model to predict the test set and compare reality to predictions).
3. Do some exploration and visualization of your training set (the 80% of cases used for fitting), possibly fitting some single predictor models, to determine whether any transformations are needed, and perhaps to narrow down the space of possible models
4. Check model conditions using appropriate tools
5. Report any relevant fit measures that you used to decide among models.
6. Some assessment of multicollinearity in your (nearly?) final model, and if appropriate, doing something about it
7. Evaluation of your models on the 20% of held out cases
8. Report and interpret confidence intervals at a couple of representative values of the predictors.

Throughout, discuss what you are doing in text! Where possible, interpret the components of the models you are considering (particularly the "finalists"). It may not always be possible to give an intuitive interpretation of every coefficient, but try to do this when you can.

There is no formal length requirement; just try to balance thoroughness with conciseness. My guess is that once all the code and raw output is suppressed, depending on how big your graphs are, you will probably wind up with around 15-ish pages, give or take a few pages.

The Presentation

Each group will present for about 10 minutes on 12/17, with an additional 3-4 minutes for questions and group discussion. We will have a maximum of 8 groups (if everyone works in pairs), so this would bring us to just under the 2 hours alotted. In your presentation you will not be able to go through every detail of your modeling process, but show us the highlights, and make some slides with key graphs, summary statistics, diagnostics, etc. If you are so inclined, you could explore using RMarkdown to make your slides (this option is available in RStudio when you create a new document). A good guideline is to aim for no more than one slide per minute of talking time, ideally less. So we're talking no more than 7-10 slides.

Some Places to Look for Data

- The federal government has as lot of data on different topics here: `http://data.gov`
- The American Psychological Association maintains a repository of datasets related to (surprise) psychology, here: `http://www.apa.org/research/responsible/data-links.aspx`
- The data science competition Kaggle makes datasets available here: `https://www.kaggle.com/datasets`
- UC Irvine has a repository of datasets for machine learning models here: `http://archive.ics.uci.edu/ml/index.php`
- Links to economic dataset repostories collected here: `http://www.economicsnetwork.ac.uk/data_sets`