Midterm - Spring 2017

Modern Data Mining

20 March 2017

Instructions

This exam is open-book, open-notes, open-internet, but is strictly no collaboration. TA's are in the room to answer questions, and information that needs to be shared with the class will be shared via the whiteboard in front or via Piazza (for code or similar).

Data needed

Available on Canvas: `/canvas/Data/exam_data_final.csv`

We have also added an optional template Rmd file as well, available on the exam assignment page. That file has a web url to the exam data, which you can use to load the data as well.

Electronic Submission

In the 'Assignments' section of Canvas, go to the 'Midterm' assignment and you will be able to upload your completed files. If you use knitr, upload your `.rmd` and a compiled file (e.g. `.pdf`, `.html`) for knitr users. If you don't, attach your writeup with R code attached as an appendix for non-knitr users. Make it clear which part of your code corresponds with which part of your analysis. **A write up inside R with comments is not acceptable!**

The submisison assignment will close at 8:00 pm.

FAQ

- If you have trouble uploading your files, email them to `lzhao@wharton.upenn.edu` before 8:00 pm
- As always skip any part you have trouble with and you may come back to finish it if you have time. Ask one of us for help if you are stuck somewhere.
- None of the questions are related, so skip any part you have trouble with and finish it later.

You can read in the data by specifying the whole path of the data, or by setting the working directory.

```
# setwd('Dropbox/STAT471')
election <- read.csv("exam_data_final.csv")
```

The data for this exam is aggregated from various sources, but all describes demographic, social, and political makeups of various US counties, which are roughly understood as regional administrative areas of US states. A few variables are included to describe state-level behaviors, and these are clearly labelled. Unless specified, assume that the data was collected in 2012 and reflects the state of the counties at that time.

This data can be used for various purposes. For this exam we will mainly focus on understanding and predicting the results of the 2012 presidential election.

See the full description of the dataset in the appendix first so as to get familiar with the variables.

Part 1: Data Exploration

1) How many states are included in the data? Which states?

2) Despite us focusing on election results, let us first take a quick look at a few variables that seem to provide disturbing evidence that the country is not in a great shape. Show histograms of `uninsured` and `poverty_frac`. Write a brief summary of the two variables.

3) Make side by side boxplots of `poverty_frac` by `state`. Do you have graphical evidence to show that wealth distribution varies among states? No formal tests are needed. You only need to describe what you see from the graph. (Hint: using parameter `las=2` in the `plot` function will print the x-axis label vertically to show all the state names)

Part 2: Election Prediction

We will build a model step by step to predict the result of the 2012 presidential election `dem12_frac` will be used as the response.

1) Fit a simple linear model predicting `dem12_frac`, using only `median_income`. Note that `median_income` is a categorical variable.

1a) According to this model, does median income have a significant effect (at the 0.05 level) on `dem12_frac`?

1b) Describe the effect of `median_income` on `dem12_frac` - which counties are most likely to vote for the Democratic candidate?

2) Fit a new model to predict `dem12_frac` using `median_income` and `state`.

2a) Is the `median_income` significant at the 0.05 level in this new model? Provide a proper test.

2b) Describe the effect of `median_income` on `dem12_fracm` if we control `state`. How is the effect of `median_income` in this new model different from the one in the model of **1)**?

3) Parsimonious model to predict `dem12_frac`.

3a) The following variables should not be included in the model: `county`, `rep12_frac`. Explain why.

3b) Build a model using LASSO excluding `county` and `rep12_frac`. For simplicity, exclude `state` as well. To ensure reproducibility, you *must* use `set.seed(34)` at the top of your block and use 10-fold cross-validation. Under this model, what are the non-zero coefficient variables under `lambda.1se`?

3c) Fit a ordinary least squares (OLS) model using the non-zero coefficient variables from **3b)**. Are all the selected variables significant at the 0.05 level? (Hint: each categorical variable should be treated as one variable, for significance testing)

3d) Starting from the model we got from **3c)**, use backward elimination (manually) to create a smaller model, where all variables are significant at the 0.05 level. Use this as the final model. Report the summary.

3e) Do a model diagnosis for the model of **3d)** by providing the residual plot and the qqnorm plot. Reason whether the assumptions of the linear model are met.

3f) Write a brief summary of your findings based on the final model. Describe how each variable affects the prediction. Address your concerns, if any.

Part 3: Turnout

Voter turnout is an important element of election analysis, specifically, what percentage of registered voters will actually vote. Political campaigners pay close attention to it and try to raise turnout in areas where they think their candidate will be favored.

1) Some political theories suggest that a higher margin in 2008 will cause a lower turnout in 2012. Fit a simple linear model predicting 2012 turnout, using only `margin_state_08`.

1a) According to this model, does `margin_state_08` have a significant effect (at the 0.05 level) on voter turnout in 2012?

1b) In the 2008 election, Washington state was carried by Barack Obama with 57.65% of the vote, to John McCain's 40.48%. Can we provide the 95% confidence interval for `turnout` in 2012 of King County, located in Washington? If yes, provide the 95% confidence interval; if no, explain why.

Appendix

Here is a variable description table. Unless otherwise specified the data is for the year 2012.

% latex table generated in R 3.3.1 by xtable 1.8-2 package % Mon Mar 20 12:18:35 2017

	variable_names	definition
1	county	Name of the county (note: this is not necessarily unique, e.g. can have multiple counties with same name)
2	state	Which state the county is in
3	total_votes_12	# of votes cast in the 2012 presidential election, for a county
4	dem12_frac	% of votes cast in 2012 presidential election for the Democrat (Barack Obama), for a county. Note this and the Republican % are not guaranteed to add to 1.
5	rep12_frac	% of votes cast in 2012 presidential election for the Republican (Mitt Romney), for a county
6	total_population	# of residents in the county, in 2012
7	registered_voters	# of registered voters in the county, in 2012
8	reg_vote_frac	% of total population registered to vote (calculated as fraction of #7/#6)
9	turnout	% of registered voters that voted in 2012 election (calculated as fraction of #3/#8).
10	hs_diploma_degree	% of county residents which have at least a high school diploma
11	median_income	Measure of median income, "Low", "Medium", "High" represent the 0-33rd percentile, 34-66th percentile, and up.
12	poverty_frac	% of residents living under the federal poverty level
13	caucasian_frac	% of residents who are non-Latino White
14	adult_smoking	% of adult residents who identify as cigarette smokers
15	gini_coefficient	Measure of income inequality; higher = more inequal
16	median_age	Median age of county residents
17	uninsured	% of residents who are do not have health insurance
18	unemployment	% of residents who are unemployed
19	violent_crime	Violent crime incidence rate; measured as violent crimes per 1000 residents
20	homicide_rate	Homicide incidence rate; measured as homicides per 1000 residents
21	infant_mortality	Infant death incidence rate; measured as infant mortalities per 1000 residents
22	african_american_frac	% of residents who are African-American
23	latino_frac	% of residents who are Latino
24	bachelors_degree	% of residents who have at least a bachelors degree
25	graduate_degree	% of residents who have at least a graduate degree
26	competitive_state_08	TRUE if the margin of victory in the 2008 presidential election was less than 10%
27	margin_state_08	Margin of victory in the 2008 presidential election for the county's state (absolute value of Democrat - Republican vote)