Data Analysis Project 3

36-401, Modern Regression, Fall 2015, Section B

Due at 5:00 pm on Tuesday, 15 December 2015

This exam is week-long take-home data analysis exam. You are allowed to use your textbook as well as other reference books you feel you might need. You should use the statistical software R to perform your analysis. **You are under no circumstances allowed to consult with any person other than your professor and your teaching assistants.** You are expected to comply with the CMU policy on academic integrity. Unauthorized help will result in failing the exam, and possibly more severe disciplinary action.

Please submit two files to Blackboard: one is the PDF of your report. If you use R Markdown (or knitr) to prepare your report, also submit your .Rmd (or .Rnw) file. If you did not use R Markdown, submit a separate plain-text file with all of your R code, clearly commented so that it is clear which parts of your code go with which parts of your report. DO NOT SUBMIT WORD FILES; THEY WILL NOT BE GRADED.

While there are wrong answers, there are many possible right answers. Any data analysis decisions or conclusions that you make should be justified and explained. Your job is to correctly ananlyze the data, not force the analysis to match a pre-conceived idea.

1 Formatting Instructions

Your answer should be written in the format of a scientific report. **You have an eight page length limit.** Nothing over the limit will be read. Do not try to game this: fonts should be no smaller than 9 points, margins should be reasonable, graphs and tables should be embedded in the report and count against the length.

Data and Research Problem

We return to the data on economic mobility used in DAP 1. In that project, the only real covariate as the fraction of workers with a commute of less than 15 minutes. The researchers who collected the data, however, also included a large number of other covariates, and in this project, we provide an expanded set of

variables. Your personalized data file will be e-mailed to you, at your Andrew e-mail address[1]. The variables are as follows:

1. ID: A numerical code, identifying the community.
2. Name: the name of principal city or town.
3. State: the state of the principal city or town of the community.
4. Longitude: Geographic coordinate for the center of the community
5. Latitude: Ditto
6. Mobility: The probability that a child born in 1980–1982 into the lowest quintile (20%) of household income will be in the top quintile at age 30. Individuals are assigned to the community they grew up in, not the one they were in as adults.
7. Population in 2000.
8. Is the community primarily urban or rural?
9. Black: percentage of individuals who marked black (and nothing else) on census forms.
10. Racial segregation: a measure of residential segregation by race.
11. Income segregation: Similarly but for income.
12. Commute: Fraction of workers with a commute of less than 15 minutes.
13. Mean income: Average income per capita in 2000.
14. Gini: A measure of income inequality, which would be 0 if all incomes were perfectly equal, and tends towards 100 as all the income is concentrated among the richest individuals (see Wikipedia, s.v. "Gini coefficient").
15. Gini bottom 99%: Gini coefficient among the lower 99% of that community.
16. Fraction middle class: Fraction of parents whose income is between the *national* 25th and 75th percentiles.
17. Local tax rate: Fraction of all income going to local taxes.
18. School expenditures: Average spending per pupil in public schools.
19. Test scores: *Residuals* from a linear regression of mean math and English test scores on household income per capita.

[1]Each student will receive a slightly different subset of the data, so that your results will not quite match that of any other student. But since you will not collaborate with another student on this exam, that will not matter to you.

20. High school dropout rate: Also, *residuals* from a linear regression of the dropout rate on per-capita income.
21. Manufacturing: Fraction of workers in manufacturing.
22. Chinese imports: Growth rate in imports from China per worker between 1990 and 2000.
23. Foreign: fraction of residents born outside the US.
24. Religious: Share of the population claiming to belong to an organized religious body.
25. Violent crime: Arrests per person per year for violent crimes.
26. Divorced: Fraction of adults who are divorced.
27. Married: Ditto.

Specific Analytical Questions

1. Is commuting time still an important predictor of economic mobility, even when other variables are also considered?
2. Which variables are, in fact, the most important variables for predicting economic mobility?
3. To what extent do measures of better education predict higher levels of economic mobility?
4. To what extent do measures of integration across social groups predict economic mobility?
5. To what extent do variables which can be *directly* affected by government policy predict economic mobility?

Suggested Outline

1. *Introduction* Write four to five sentences introducing the research problem and describing specific research hypotheses. Cite any information sources in parentheses or foot- or end- notes.
2. *EDA* How many observations do you have?
    - Examine the (predictor and response) variables univariately and multivariately.
    - Provide graphical displays or numerical summaries for all variables.
    - You need EDA for all pairs of numerical variables and at least for the categorical variables and the response variable. Describe your results.

- Which variables seem associated with economic mobility?

3. *Initial modeling* Start by building a multivariate linear regression to the data predicting usage from the predictor variables. Address the specific questions of the client when building the model. Be sure to justify the choices you made in building this initial model.

4. *Diagnostics/model selection*

   - Are the basic assumptions met for your multivariate linear regression model? Why or why not?
   - What transformations do you choose (if any)? Why?
   - Are there any outliers in your sample overly influencing your model? Identify any outlier candidates and decide whether or not to remove them. Give details.
   - Do you exclude any variables? Why? All exclusions/inclusions must be justified.

5. *Final model inference/results* Create a table that summarizes your final model (coefficients, standard errors, confidence intervals, p-values). Provide interpretations of all your coefficients in the context of the problem. Be sure to address the specific questions of the client (above).

6. *Discussions/results*: What are your conclusions? Identify a few key findings, and discuss, with reference to the supporting evidence. Can you come up with explanations for the patterns you have found? Suggestions or recommendations for the client? How could your analysis be improved? (6–8 sentences)

Rubric

Words (10) The text is laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself is well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow from the presumed level of knowledge.

Numbers (5) All numerical results or summaries are reported to suitable precision, and with appropriate measures of uncertainty attached when applicable.

Pictures (5) Figures and tables are easy to read, with informative captions, axis labels and legends, and are placed near the relevant pieces of text.

Code (10) The code is formatted and organized so that it is easy for others to read and understand. It is indented, commented, and uses meaningful names. It only includes computations which are actually needed to answer the analytical questions, and avoids redundancy. Code borrowed from the notes, from books, or from resources found online is explicitly acknowledged and sourced in the comments. Functions or procedures not directly taken from the notes have accompanying tests which check whether the code does what it is supposed to. The text of the report is free of intrusive blocks of code. If you use R Markdown, all calculations are actually done in the file as it knits, and only relevant results are shown[2]. If you do not use R Markdown, the code in your appendix must generate exactly the results you show in your report, and must have comments making it clear which parts of your code go with which results.

Explotory Data Ananlysis (15) Variables are examined individually and bivariately. Features/observations are discussed with appropriate figure or tables. The relevance of the EDA to the modeling is clearly explained.

Model formulation and checking (30) The initial model's formulation is clearly related to the substantive questions of interest. The model's assumptions are checked by means of appropriate diagnostic plots or formal tests; if the model is re-formulated, the changes are both well-motivated by the diagnostics, and still allow the model to answer the original substantive question. Limitations from un-fixable problems are clearly noted.

Estimation, Inference and Uncertainty (15) The actual estimation of model parameters or predictions is technically correct. All calculations based on estimates are clearly explained, and also technically correct. All estimates or derived quantities are accompanied with appropriate measures of uncertainty.

Conclusions (10) The substantive questions are answered as precisely as the data and the model allow. The chain of reasoning from estimation results about models, or derived quantities, to substantive conclusions is both clear and convincing. Contingent answers ("if X, then Y, but if Z, then W") are likewise described as warranted by the model and data. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too is reflected in the conclusions.

Extra credit (10) Up to ten points may be awarded for reports which are unusually well-written, where the code is unusually elegant, where the analytical methods are unusually insightful, or where the analysis goes beyond the required set of analytical questions.

[2]See the model report for DAP 1 for examples of how to do this.

Writing Advice

Your language should be very clear and precise. Do not make claims for which you have no evidence. Do not say "will" or "would" when you really mean "may" or "might". Do not use language that implies causation; you are studying associations between variables only. When revising, look for "throat-clearing" phrases ("This is because", "this is due to", "the reason for this is", "this means that", "I believe that", "I think the reason is that") and either cut them or replace them with more concrete and informative phrases. Make sure pronouns have clear referents ("these confidence intervals show" vs. "this shows").

You may assume that the reader has a general familiarity with the contents of 36-225 and 36-226, and with the models and methods we have covered so far in the course, but will need to be reminded of any details. The reader should not be assumed to have any prior familiarity with the data set.