Data Analysis Exam 1

36-401, Modern Regression, Fall 2015

Due at 3:00 pm on Thursday, 15 October 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.

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.

The exam project will look at economic mobility across generations in the contemporary USA. The data come from a large study[1], based on tax records, which allowed researchers to link the income of adults to the income of their parents several decades previously. For privacy reasons, we don't have that individual-level data, but we do have aggregate statistics about economic mobility for several hundred communities, containing most of the American population, and covariate information about those communities. We are interested in predicting economic mobility from the characteristics of communities.

1 The Data

The data file `mobility.csv` has information on 729 communities[2]. The variable we want to predict is economic mobility; the rest are predictor variables or covariates.

1. 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.

[1] The solutions will say which. In the meanwhile, tracking it down would not actually help you very much.

[2] Technically, "commuting zones". These include cities and their suburbs and exurbs, but also many rural areas with integrated economies.

Individuals are assigned to the community they grew up in, not the one they were in as adults.

2. Commute: Fraction of workers with a commute of less than 15 minutes.
3. Longitude: Geographic coordinate for the center of the community
4. Latitude: Ditto
5. Name: the name of principal city or town.
6. State: the state of the principal city or town of the community.

An important hypothesis for the researchers who gathered this data is that short commuting times lead to higher rates of social mobility. In this assignment, we will not be concerned with their explanation of how this might work, but just with whether there really is such a connection.

2 Formatting Instructions

Your answer should be written in a report-style format. **You have a four page length limit.** Nothing over four pages will be read. Do not try to game this: fonts should be no smaller than 10 points, margins should be reasonable, graphs should be embedded in the report and count against the length.

This first DAP gives specific prompts for each section of your report. However, as the semester progresses, there will be less structure. 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 c orrectly ananlyze the data, not force the analysis to match a preconceived idea . Discuss whether or not your results match your hypothesis. The following are som e prompts for each section

Introduction Write a short introduction describing the research problem. Clearly state the research hypothesis at the end. Cite any sources you use for backgroud information

Exploratory Data Analysis/Inital Modeling Examin the two variables individually. Report summary measures and describe any interesting featurs these measures indicate. Graphically display the data (think about what types of graphs would be the most useful first). Describe the graph. Examine the two variables together. Graph the data together. Describe any trends or interesting features that you see. Fit a simple linear regression model to the data. Find the estimated regression function and display your regression line appropriately (can combine with EDA graph if necessary)

Modeling and Diagnostics Creat diagnostic plots to determine the appropriateness of your model. Discuss whether the assumptions are met. If not, what steps do you take to transform the variable(s)? If you decide transformations are necessary, do them; recheck your diagnostics.

Inference and Results Report your final estimated regression function and interpret the parameters in context. Are your parameters significantly different

from zero? At what level? Report $(1-\alpha)\%$ confidence intervals (choose an appropriate α). Is there statistical significant linear relationship bewteen the two variable of interest? Explain in context of problem. Create plots contrasting the model's predictions with the actual data.
Conclusion and Discussion Summarize your main findings in the analysis. What is the final conclusion with regards to the original research hypothesis? Make some recommendations for future work or studies. What can be done to improve the research study?

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.

3 Rubric

As usual, this describes the ideal.

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 main text of the report is free of intrusive blocks of code.

Explotory Data Ananlysis (15) Variables are examined individually and bivariately. Features/observations are discussed with Figure/Tables.

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 question about social mobility is 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.

4 Two Hints

Maps

Because the data here are spatial, it may be helpful to make exploratory plots (or plots of model predictions, model residuals, etc.) in the form of maps, with a dot at the latitude and longitude of each community, and the value of the variable being plotted represented by the color or size of the dot. For instance, this code will make a map where the color of each city reflects its population (presuming the data is loaded into `mobility`).

```
# Compute the quintiles of community population
population.quintiles <- quantile(mobility$Population, c(0,0.2,0.8,0.4,1.0))
# Assign each community to its quintile
population.categories <- cut(mobility$Population, population.quintiles)
# Make up a color scale
five.shades.of.grey <- grey((5:1)/5)  # So darker == larger
# Plot each town with corresponding color
plot(x=mobility$Longitude, y=mobility$Latitude,
     col=five.shades.of.grey[population.categories], pch=19, cex=0.5)
```

Improving the labels, adding a legend, etc., is left as an exercise, and no promises are made that this will be a good thing to do.

Hiding Code in R Markdown

The report is supposed to be a humanly-readable document, and big (or even small) chunks of computer code do not improve readability. Fortunately, it is easy to tell R Markdown to run a piece of code, and include its output, but *not* include the code in the PDF or HTML document:

````
```{r, echo=FALSE}
summary(mobility[,c("Mobility", "Population")])
plot(Population ~ Mobility, data=mobility)
```
````

will run the code, produce a table of summary statistics for two of the variables (which?) and a scatter-plot of two variables, but will not "echo" the code into the report. (Again, no promises that you will want to do exactly this thing.)

`echo=FALSE` is actually one of many options which control how R Markdown processes your code; see `http://yihui.name/knitr/options/` for a complete list, and the examples at `http://yihui.name/knitr/demos/`.