Data Analysis Exam 1
36-401, Section B
Due at 3:00 pm on Friday October 13

1 General Instructions

This exam is a 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.

Submit two files to Canvas. The first is the PDF of your report. The other is your R code. This can be a text file or an R markdown file. Your code should be 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.

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.

Your report should have the following sections:

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. Examine the two variables individually. Report summary measures and describe any interesting features 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. Create 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 a 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 Writing a Good Report

The text should be laid out cleanly, with clear divisions and transitions between sections and sub-sections. The writing itself should be well-organized, free of grammatical and other mechanical errors, divided into complete sentences logically grouped into paragraphs and sections, and easy to follow.

All numerical results or summaries should be reported with appropriate measures of uncertainty attached when applicable.

Figures and tables should be easy to read, with informative captions, axis labels and legends.

Your code should be organized so that it is easy for others to read and understand. Use comments and meaningful names. Only include computations which are actually needed to answer the analytical questions, and avoid redundancy. Code borrowed from the notes, from books, or from resources found online should be explicitly acknowledged and sourced in the comments.

Variables should be examined individually and bivariately.

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

All estimates or derived quantities should be accompanied with appropriate measures of uncertainty.

The substantive question about social mobility should be 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 must be clear and convincing. If uncertainties in the data and model mean the answers to some questions must be imprecise, this too should be reflected in the conclusions.

For those who use R Markdown: The report is supposed to be a humanly-readable document, and big chunks of computer code do not improve readability. 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. For example,

````
```{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 and a scatter-plot of two variables, but will not "echo" the code into the report.

4 The Data

The data is about economic mobility. The data come from a large study 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 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.

The data file

`www.stat.cmu.edu/~larry/=stat401/mobility.csv`

has information on 729 communities. The variable we want to predict is economic mobility; the rest are potential predictor variables (covariates). The variables are:

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. 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 longitude for the center of the community.
4. Latitude. Geographic latitude for the center of the community.
5. Name: the name of principal city or town.
6. State: the state of the principal city or town of the community.

The Research Question. The researchers who gathered this data want to know if short commuting times lead to higher rates of social mobility. Your report should be devoted to addressing this question.