Project 2: Regression Analysis

This project is about doing a "complete" regression analysis, and presenting your results in an IMRAD paper. Your IMRAD paper should have all the elements of the paper you wrote for Project 1. ***(Please review the guidelines and reference materials from Project 1 for writing an IMRAD paper.)***

In addition, please include a "Technical Appendix" after the Reference section of your IMRAD paper. The technical appendix should include any code you want me to see, and any results that support your reasoning in the IMRAD paper itself. *I do not promise to look at the appendix, but if I have questions about your work, I may look there to see details of what you did, so* **please** *make it easy for me to find the part(s) of the appendix that relate to each part of the IMRAD paper.* If you don't make it super-easy for me to find supporting work for each part of your analysis in your appendix, I will simply assume you didn't do it.

There are no "extra points" for the appendix. It is just there to make it easier for me to see what you did, if I need to.

The Data

The file `cdi.dat` in the `project02` folder in the files area for our course on Canvas is taken from Kutner et al. (2005)[1]: It provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992. The definitions of the variables are given in Table 1 on p. 2 of this document.

The Research Questions

Some (fictional!) social scientists are interested in looking at this historical data, to learn how average income per person was related to other variables associated with the county's economic, health and social well-being. Here are **four** questions that the researchers are interested in (continues on p. 3):

1. Looking at the data one pair of variables at a time, which variables seem to be related to which other variables in the data? Which are not? Are all of the relationships what a reasonable person would expect, or are there some surprises? Can you explain these findings in terms of the meanings of the variables?
2. There is a theory that, if we ignore all other variables, per-capita income should be related to crime rate, and that this relationship may be different in different regions of the country (Northeast, North-central, South, and West). What do the data say?
3. Find the best model predicting per-capita income from the other variables (including possible transformations, interactions, etc.). Here "best" means a good compromise between
    - Best reflects the social science and the meaning of the variables
    - Best satisfies modeling assumptions
    - Is most clearly indicated by the data
    - Can be explained to someone who is more interested in social, economic and health factors than in mathematics and statistics.

[1] Kutner, M.H., Nachsheim, C.J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models, Fifth Edition*. NY: McGraw-Hill/Irwin.

Table 1: Variable definitions for CDI data from Kutner et al. (2005). *Original source:* Geospatial and Statistical Data Center, University of Virginia.

Variable Number	Variable Name	Description
1	Identification number	1–440
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged 18–34	Percent of 1990 CDI population aged 18–34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 or old
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990, including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990 CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI population that is unemployed
15	Per capita income	Per-capita income (i.e. average income per person) of 1990 CDI population (in dollars)
16	Total personal income	Total personal income of 1990 CDI population (in millions of dollars)
17	Geographic region	Geographic region classification used by the US Bureau of the Census, NE (northeast region of the US), NC (north-central region of the US), S (southern region of the US), and W (Western region of the US)

4. A county is a governmental unit in the United States that is bigger than a city but smaller than a state. There are 50 states in the US, plus the District of Columbia, which is usually coded as a 51^{st} state in data like this. There are 48 states represented in the data. There are approximately 3000 counties in the US, and 373 represented in the data set. Should we be worried about either the missing states or the missing counties? Why or why not?

Further Directions And Hints

- Feel free to use transformations, interactions, etc. as needed, for each of the research questions.
- For question #2, please only use the variables per-capita income, crime rate, and region (and/or any transformations or interactions of them that you may find useful). At some point you may wish to consider partial F tests for groups of variables, and vif's to help explain your findings; and/or you may find other tools to be useful.
- For question #3, you cannot possibly search the whole model space: including main effects and all possible interactions, there's something like $2^{2^{16}}$ possible models, even if you don't consider transformations. So:
  - Think about the problem and the meanings (both verbal/scientific and mathematical) of the variables, to choose a good set of variables and a good path through the model space, to work with, and ***look*** at the raw data, transformed data, diagnostic plots, etc., early and often. There's nothing wrong with using some of the variable selection tools we've talked about in class, but they are no substitute for also looking at and thinking about the data.
  - Don't transform for the sake of showing that you know how to make transformations. Choose transformations that (a) help with modeling in some way; and (b) are still explainable to the social scientist. If there are no transformations that satisfy these criteria, don't transform.
  - Generally, main effects (the original input variable) are easier to explain than interactions (products of input variables). Two way interactions (product of two variables) are easier to explain than three way interactions (product of three variables), which are easier to explain than 4-way interactions, etc. So don't go too wild with interactions unless you are really getting somewhere that you can explain to the social scientists.
  - Review the advice for model building at the end of lecture 11, from Gelman & Hill.
  - Please do not present 10 different models. Ideally, present your one best model. If necessary, discuss one or two close competitors.
- You should turn in a single pdf containing
  - A complete IMRAD report ***(Please review the guidelines and reference materials from Project 1 for writing an IMRAD paper)***.
  - A technical appendix after the Reference section of your IMRAD report. The technical appendix should include any code you want me to see, and any results that support your reasoning in the IMRAD paper itself. *If I have questions about what you did in the main IMRAD paper, I may look at the appendix to see details of what you did, so please make it easy for me to find the part(s) of the appendix that relate to each part of the main paper.*

You are to do this project on your own, without collaborators. If you are unsure of what something means, feel free to look it up on the web or elsewhere, but you may not post questions on discussion websites or

blogs like stackexchange, etc. ***You are welcome to discuss this project with me or the TA (office hours are also fine), but no one else.*** *Please remember to cite all the sources that you used, including webpages, in the reference list at the end of your report.*

Due date

Optional: I will hold extra office hours on Friday October 11, 9:30-10:30 and 12:00-1:30. You may use these office hours to discuss your initial work on the project with me.

Required: You must submit a final pdf on Canvas by midnight (11:59pm) Friday October 18. "Grace" for late submissions until Saturday at 11:59pm.

Grading

Below is a summary of what I will be looking for. See the two pdf's listed on p. 1 of this project assignment for more detail.

Part	Looking For...	Percent
Title	Clear, interesting, focused	5%
Author/Contact Info	Your name & email addr!	∞!!
Abstract	Summarizes I,M,R,D in 3–5 brief, clear sentences.	5%
Introduction	Brief, clear, to the point; context for the problem; What is the problem/aim of the study? What questions will be answered?	10%
Methods	Study design; how was the data collected? Definition of variables & outcome measure(s); statistical methods; ethical considerations	10%
Results	Statistical analysis & results in order parallel to Intro & Methods; no new methods or data; no big picture discussion	10%
Discussion	Recap findings; address main problem/question; strengths & weaknesses; implications, unanswered questions, future research	10%
Mechanics	Follows C-C-C as much as possible (sentences, paragraphs & sections); Grammatical; Easy to follow	5%
Statistical Content	Correctly and appropriately uses technical and non-technical material we have learned in class. Easy to follow; Analysis makes sense/not crazy (roughly 10% per research question)	40%
References & Citations	Follow ASA guide, "The Reference List" & "Reference Citations" (be sure to cite all sources!)	5%
Technical Appendix	Helps me to understand your paper and give you max points above; Easy to follow	0%