Project 1: Regression Analysis

This project is about doing a "complete" regression analysis, and presenting your results in an IDMRAD paper. Your IDMRAD paper should have all the elements that we have discussed in class and in homework:

- Title
- Author & Contact Info
- Abstract
- Introduction
- Data
- Methods
- Results
- Discussion
- References
- Technical Appendix

Further details on how each part of the paper will be graded appear at the end of this assignment document. ***(Please review the guidelines and reference materials from week02 for writing an IDMRAD paper.)***

As preparation for completing the IDMRAD paper I will ask you to develop the data analyses for the technical appendix as part of hw06.

Your final technical appendix should include any code you want me to see, and any results that support your reasoning in the IMRAD paper itself. Each time you assert a fact in the main part of the paper based on your work in the technical appendix, **please** indicate right there in the paper what page number(s) in the appendix I should look at, for further details, evidence, etc., justifying your assertion. 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, and you will lose points.

The technical appendix will be graded as part of hw06. You can (and should) continue to work on your technical appendix after you turn in hw06. (Your final technical appendix should not look like a homework assignment, but rather like well-organized and informative lab/data analysis notes.)

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] 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)

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? Does it matter if you use number of crimes, or (number of crimes)/(population), in your analysis?
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.
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? (You will not be able to answer this question by simply fitting a linear regression model to the data in `cdi.dat`!)

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 mentioned in the question (and/or any transformations or interactions of them that you may find useful), but not other variables in the data set. At some point you may wish to consider partial F tests for groups of variables or 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:
    - It is a good idea to use automatic/algorithmic methods for variable selection and transformations *combined with* your personal observations and your understanding of what can be communicated successsfully to the client, in arriving at a final model.
    - Think about the problem and the meanings (both verbal/scientific and mathematical) of the variables, as you work to choose a good set of variables & transformations, and ***look*** at the raw data, transformed data, diagnostic plots, etc., frequently throughout the process. You will find it helpful to use some of the variable selection and transformation tools we've talked about in class, but don't rely on them exclusively—make sure your final model is interpretable and explainable as well.
    - 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.
  - Although you may consider several models in the technical appendix, please don't present 10 different models in the main paper. Ideally, present your one best model. If necessary, discuss one or two close competitors.

- You should turn in a single pdf containing
  - A complete IDMRAD report ***(Please review the guidelines and reference materials from wek 02 for writing an IDMRAD 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 will want to 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—If I can't find it, I will assume it's not there, and you will lose points.*

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 (any static written resources are fine), 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.*** *Any posts on Piazza should be private to the instructors. Please remember to cite all the sources that you used, including webpages, in the reference list at the end of your report.*

Due dates

Mon Oct 11 First draft of technical appendix work will be due with hw06.

Mon Oct 18 Rough draft pdf of entire paper due on Canvas [not gradescope].

Mon Oct 25 Peer reviews of rough drafts due on Canvas.

Fri Oct 29 Final draft of paper due on Canvas by 11:59pm.

Grading

Below is a summary of what I will be looking for.

The percentages in the table on the next page assume that all parts of the paper are there. If one or more parts is missing, it may result in a much lower grade than the percentages suggest.

Part	Looking For...	Percent
Title	*Clear, interesting, focused.*	5%
Author/Contact Info	*Your name & email addr!*	∞!!
Abstract	*Summarize I, D, M R and D sections of the paper* (typically one sentence each).	5%
Introduction	Brief, clear, to the point; context for the problem; What is the problem/aim of the study? *Why would anyone want to read this paper? What questions will be addressed?*	10%
Data	*What data set was used in this study?* Typically, include variable definitions, sample size, quick numerical summaries of the variables and initial EDA, but *no model fitting, transformations, or analysis*.	5%
Methods	*What did you do, to address these questions?* List the methods and/or analyses that will be used to answer eacb question stated in the **Introduction**. *No data analysis, graphing, model fitting, etc. appears here*; you just say what methods and analyses you will use with which variables, to answer each question.	5%
Results	*Statistical analysis & results* in order parallel to **Introduction** and **Methods** sections. Here you *finally* get to show the data analyses (model fitting, graphics, etc.) that you did, and what the results were. Don't overload the reader: put the highlights here so the reader understands what you did and why, and refer the reader to specific pages or sections of the **Technical Appendix** for more details. It should be clear which data analyses and results go with which question from the **Introduction**. *Every analysis that is presented here should have been mentioned in the* **Methods** *section.*	10%
Discussion	*What does it all mean? Recap findings; address main problem/question; strengths & weaknesses; implications, unanswered questions, future research.* Typically you will say, for each question from the **Introduction**, how the analyses that you did the **Results** section answers that question. You might also mention EDA and so forth from the **Data** section if that makes clearer to the reader what answers you found for one (or more) of the questions. Then you will talk about the big picture, what future work or generalizations of your work might look like, and any limitations of your study. But *there should be no additional analyses or results in this section; just use the analyses you did for the* **Results** *section (and possibly the* **Data** *section).*	10%
Mechanics	Follows C-C-C[2] as much as possible (sentences, paragraphs & sections); *Grammatical; Complete sentences and paragraphs; Easy to follow.*	5%
Statistical Content	*Correctly and appropriately uses technical and non-technical material* we have learned in class. Easy to follow; Analyses makes sense/not crazy (roughly 10% per research question)	40%
References & Citations	*Follow ASA guide*[3]*, "The Reference List".* *Follow ASA guide, "Reference Citations".* Be sure to cite **all** sources!	5%
Technical Appendix	*Contains complete versions of the analyses listed in the* **Methods** *section and presented in the* **Results** *section*: R code, output, graphs, tables, and comments explaining what you did and why. There may be additional analyses here (e.g. to support the **Data** section of the paper, or to show why the methods and analyses that you chose for the paper were the right ones). Make it easy for me to follow.	0%[4]

[4]See Rule 3 in `10 rules for better organized papers.pdf`.

[4]See `ASA Style Guide.pdf`.

[4]You will get credit for this as part of hw06 instead.