36-617: Applied Linear Models
Fall 2022
Take-home midterm. Due Weds Oct 5, 11:59pm

- Please turn this midterm in to Gradescope via the Canvas "take-home midterm" assignment link as usual.
- Rules and guidelines for the take-home midterm:
  - Covers material through Ch 7 (variable selection)
  - Open-book, open-notes, etc., but not open-people
    - Do the work on your own, no collaborators
      - In particular, no current or former MSP students or solutions
    - Feel free to use web resources (incl stackexchange etc) but you ARE NOT ALLOWED to post questions or interact on the web
  - Office hours (me or TA) or private Piazza questions are fine (ok to learn from any answers you hear me or Lorenzo give to other students)
  - You are free to talk with each other about general questions in the class, etc. But you are not allowed to ask questions like "what kind of graph did you make for question 3(b)?", "What do you do for question 1(a)?", etc.
  - If you are unsure whether something is allowed, or in the spirit of these guidelines, ASK US.
  - *You are on your honor to follow these rules & guidelines.*
- There are two problems (with parts of course) below.

Exercises

1. Return again to the `beauty` data. We will use the variables in the reduced data set `beauty.red`, with the `profevaluaion` variable also removed, and whatever transformations you decided to use in HW04.

   (a) Print a summary of the fitted model you obtained after completing problems 3(c) and (d) on HW04, and write a short paragraph interpreting the fitted model for a college dean who is trying to understand what factors, other than teaching quality, might affect course evaluations.

   (b) Using the version of `beauty.red` that you used to obtain the model in part (a) [including any transformations that you applied for HW04], apply the lasso to select variables for predicting `courseevaluation` (or a transformation of `courseevaluation` if that's what you used in HW04) for this data set.

   - Is it feasible to use shrinkage plots for the lasso, as in lectures 08 and 09? If so, try it. If not, explain why not.
   - The function `cv.glmnet` in `library(glmnet)` tries to find an optimal λ by cross-validation using mean-squared prediction error. Read the documentation and try variable selection using `cv.glmnet`. Note that `cv.glmnet` produces both `lambda.min` (the best value found by cross-validation) and `lambda.1se` (the value of λ that is one SE larger than `lambda.min`, which many people use to protect against capitalization on chance).
   - You can compare the results of the two values of λ with code like this:

     ```
     result <- cv.glmnet(x,y)
     plot(result)
     c(lambda.1se=result$lambda.1se,lambda.min=result$lambda.min)
     cbind(coef(result),coef(result,s=result$lambda.1se),coef(result,s=result$lambda.min))
     ```

(c) Compare the model in part (a) with the model in part (b): Make a table showing which variables remain in the final model for each of two models, and then write a brief paragraph saying which model you would use to help the college dean understand factors other than teaching quality that affect course evaluations, based on this table and any other evidence that seems relevant, and explain your reasoning.

2. The file `cdi.dat`, in the same Canvas folder as this midterm assignment sheet, 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.

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

(a) Data description.

- Make a table or tables showing appropriate summary statistics for each variable in the data set. Note that summary statistics for continuous variables should be different from the summary statistics for categorical variables.
- Indicate where (in which variables) there is missing data (NA's), if any, how much there is (in each variable) and why it might be there.
- Make some appropriate descriptive EDA plots to illustrate any important features of the variables or possible important relationships among them.

(b) Build a regression model that predicts per-capita income from crimes and region of the country (using ony these three variables, not the full set of variables in the data set). Should there be any interactions in the model? What does your model say about the relationship between per-capita income and crimes? Do your answers change, depending on whether you use number of crimes, or "per-capita crime" = (number of crimes)/(population) as a measure of crimes? If so, which one best describes the relationship between per-capita income and crimes? Why? Show the fitted model results and explain your answer to these questions in terms of those results (as well as any economics knowledge you may have).

(c) Use methods we have discussed in class and/or methods from Sheather Chapters 5, 6 & 7 (including, as needed: transformations, interactions, variable selection, residual diagnostic plots, fit indices, added-variable or marginal model plots, etc.) to find the multiple regression model predicting per-capita income from the other variables, that makes the "best" tradeoff between the following criteria:

- Reflects the social science and the meaning of the variables
- Satisfies modeling assumptions
- 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.

Organize your work so that it tells an interesting data analysis story that a statistican (like me!) might like, and be sure to explain why and how you arrived at a final model.

(**Note:** No matter what you do, you are likely to be unhappy with some or all of these criteria; the better you make one criterion, the worse another is likely to get. So you will have to find a compromise or tradeoff between these criteria. Explain how you decided to make the tradeoff(s) you made.)

(d) Provide a careful and easy-to-follow interpretation of your final model for a client or collaborator who is more interested in social, economic and health factors than in mathematics and statistics. This should be long enough that you hit all the important points, but not so long that your collaborator gets bored and stops reading.