36-617: Applied Linear Models
Fall 2019
HW06 – Due Mon Oct 11, 11:59pm

- Please turn the homework in to Gradescope via the Canvas hw06 assignment link as usual.
- Please finish reading Chapter 7 of Scheather this week. For next week, read Chapter 8, on logistic regression.
- There are only 2 problems (with parts of course) below. They can be done in either order, and it might be worth considering doing #2 before #1.

Exercises

1. For this exercise we will consider some of the data analyses needed for Project 01. That way, you can use your work here as part of the technical appendix for the paper for Project 01. (There are 3 parts to this problem.)

   Consider the data file `cdi.dat` in the `Project 01` folder in our files area on Canvas. The variables are defined in Table 1; a more complete description of the data can be found in the `project-01.pdf` assignment sheet.

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)

(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 will 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 crime rate and region of the country. Should there be any interactions in the model? What does your model say about the relationship between per-capitqa income and crime rate? Do your answers change, depending on whether you use number of crimes, or "per-capita crime" = (number of crimes)/(population) as a crime rate measure? If so, which one best answers the question? Why? Show the fitted model results and explain your answer to these questions in terms of those results.

(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 analysis, fit indices, 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.

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.

2. Return again to the `beauty` data that you have worked on for several assignments. For this exercise, use the transformed variables that you found for HW05. (There are 4 parts to this problem.)

(a) Use the "all subsets" method to choose the best model for `coursevaluation` (or a transformation of it, if you found a good transformation for HW05), considering all the other variables (with whatever transformations you found for HW05) except for `profevaluation`, `profnumber`, `multipleclass` and the 30 class variables (`class1` through `class30`.

(b) Repeat part (2a) using Stepwise Regression instead.

[continued on next page]

(c) Repeat part (2a) using the lasso instead.

- Is it feasible to use shrinkage plots for the lasso, as in lecture 11? 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))
```

(d) *Briefly* compare the models in parts (2a), (2b) and (2c), with the model you obtained for HW05, part (2c): Make a table showing which variables remain in the final model for each of the four methods, and then write a sentence or two saying which model or models makes the most sense to you, based on this table.