STAT 601_002: Statistical Methods I, Fall 2019
24hour Take-Home Exam due on 5:00 pm, Nov. 10

Problem:

Air Pollution and Mortality. Does pollution kill people? Data in one early study designed to explore this issue came from five Standard Metropolitan Statistical Areas (SMSA) in the United States, obtained for the years 1959–1961. (Data from G. C. McDonald and J. A. Ayers, "Some Applications of the 'Chertoff Faces': A Technique for Graphically Representing Multivariate Data," in Graphical Representation of Multivariate Data, New York: Academic Press, 1978.) Total age adjusted mortality from all causes, in deaths per 100,000 population, is the response variable. The explanatory variables are the mean annual precipitation (in inches); median number of school years completed, for persons of age 25 years or older; percentage of 1960 population that is nonwhite; relative pollution potential of oxides of nitrogen, NOX; and relative pollution potential of sulfur dioxide, SO2.

"Relative pollution potential" is the product of the tons emitted per day per square kilometer and a factor correcting for SMSA dimension and exposure. The first three explanatory variables are a subset of climate and socioeconomic variables in the original data set. (Note: Two cities—Lancaster and York—are heavily populated by members of the Amish religion, who prefer to teach their children at home. The lower years of education for these two cities do not indicate a social climate similar to other cities with similar years of education.) Is there evidence that mortality is associated with either of the pollution variables, after the effects of the climate and socioeconomic variables are accounted for? Analyze the data and write a report of the findings, including any important limitations of this study.

As a statistician, you propose two best models and describe why these two models are the best to explain the relationship between two the mortality and the pollution variables. You also make an inference in your proposed models and interpret them.

Instruction:

- You can communicate about this problem with your instructor only. Don't discuss about this problem with other students except your partner.
- Both you and your partner will submit ONE report in addition to the R-code for your analysis.
- You should make written report including introduction, model/methods, result, conclusion/discussion, and appendix.
  - ✓ Introduction: you explain about the summary of data and give the goal of data analysis
  - ✓ Model/Methods: you explain how you find your best two models
  - ✓ Result: you summarize your results. If you need, you can use table or figure
  - ✓ Conclusion/discussion: you summarize and discuss about your finding.
  - ✓ Appendix: you attach your code in appendix.

Your written report should be submitted on the deadline

This is an example of a table of results for the different models

The model	Response	Predictor-(s)	R^2 %	$adj\ R^2$ %	Constant variance	Normality	Residuals independence	Influential points
1	Original	original	33.4	32.1	Holds	Not valid	Holds	Exist
2	B-C tran	original	46.1	44.95	Holds	Not valid	Holds	Exist(7,9,32)
3	B-C tran	original	50.89	49.82	Holds	Not valid	Holds	NO
4	B-C tran	Log tran	51	49.94	Holds	Not valid	Holds	NO
5	B-C tran	X, X^2, X^3	51.26	47.9	Holds	Not valid	Holds	NO
6	B-C tran	$X, \sqrt{X}, X^2, logX$						
7	B-C tran	$\sqrt{X}$	37	35.6	Holds	Not valid	Holds	NO
8	Original	$X, \sqrt{X}, X^2, logX$						
9	Original	$\sqrt{X}$	37	36	Holds	Not valid	Holds	NO
10	B-C tran	$X, \sqrt{X}, X^2, logX$	37	36				