Individual Assignment 1

Air Pollution and Mortality

Researchers at General Motors collected data on 60 U.S. Standard Metropolitan Statistical Areas (SMSA's) in a study of whether air pollution contributes to mortality. The dependent variable for analysis is age adjusted mortality (called "Mortality"). The data include variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. The task is to determine whether air pollution is significantly related to mortality.

Goal: Carefully analyze this set of data using the following steps.

1. **Data Exploration**:
    a. First, consider each of the variables individually. Compute **summary measures** (statistics like mean, median, variance, etc) and **graphical displays** (histograms, boxplots, etc) and **conclude** about the distribution of each of the variables. If a variable is highly skewed, suggest (and investigate) a suitable transformation that eases the skew.
    b. Next, perform ***pairwise investigations*** using **correlation analysis** and **scatterplots** (including trellis graphs, matrix plots, etc). **Conclude** about the *form* and *strength* of the relationship between individual predictors and the response. Determine which variables appear to have only a weak relationship and may not be very useful in explaining the response. Also investigate (and conclude) whether ***transformations*** between some variables may strengthen the relationships.
2. **Data Modeling**
    a. Consider ***different*** regression models for mortality and recommend ***three*** models that fit the data best. These three "best" models may be a consequence of measures of model fit, common sense, intuition and/or other considerations (or a combination of all). Please explain your rationale.
    b. Then, investigate the ***model assumptions*** of those three models. Are the errors normally distributed? Are the errors independent or is there evidence of serial correlation? Do the errors exhibit constant variance? Is the linearity assumption of the model satisfied? Attach exhibits as necessary to support your arguments. Use your findings to revise your models (e.g. use transformations to strengthen the assumptions). This may require going back to part a.
    c. Check each model for ***unusual*** **and** ***influential*** observations. Be careful in deleting influential observations. Investigate which of the coefficients are mostly influenced by different influential data points.
    d. Based on steps a-c (and possibly iterating among some of the steps) recommend your ***final model***. The final model should represent the data in a best possible way and it should also adhere to all model assumptions.

Also, based on this final model, ***conclude*** whether pollution affects mortality and which other factors also affect mortality.

Your report should consist of 3-4 pages (plus attachments, added to the end of the report). The report should clearly show your train of thought, it should bring across the different steps that you have taken, and **why** you have taken them. Do not just write a set of stand-alone, unconnected paragraphs or key words. Consider the report as a "story" in which you want to bring your argument across as to why your final model is the best one and why all the intermediate steps (transformation, data elimination, etc) were necessary. Attach all relevant exhibits (Figures, Tables) that support your findings. Label, format and refer to all exhibits appropriately throughout your report. Only attach exhibits if you use them to bring across a particular point. Then, do not forget to describe what you have done to create that exhibit and what can be learned from it. As always, use proper English language, be specific but concise. Writing style and quality of presentation matters!

Grading will take into account completeness of your analysis, correctness, and style.