PROJECT

Statistics 108
Summer Session 1, 2020

[This project is due by 8.00pm on Friday, July 31. You may form a group of three students (including yourself) registered in the course. Please turn in only one report for your group and write the names of the group members on it.]

Does pollution have effect on mortality? Data in one early study designed to explore this issue came from 60 Standard Metropolitan Statistical Area (SMSA) in the United States, obtained for the years 1959-1961. [Source: GC McDonald and JS Ayers, "Some applications of the 'Chernoff Faces': a technique for graphically representing multivariate data", in Graphical Representation of Multivariate Data, Academic Press, 1978.

Total age-adjusted mortality [MORT] from all causes, in deaths per 100,000 population, is the response variable. The predictor variables are listed below:

Mean annual precipitation (in inches) [PRECIP],
Median number of school years completed by persons of age 25 or over [EDUC],
Percentage of population in 1960 that is nonwhite [NONWHITE],
Percentage of households with annual income under $3000 in 1960 [POOR],
Relative pollution potential of oxides of nitrogen (NO_X) [NOX],
Relative pollution potential of sulphur dioxide (SO_2) [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 goal of the analysis would be to relate mortality to all the variables. Thus mortality is the response variable. It would be also important to see if pollution is related to mortality. Since the variables NO_X and SO_2 are skewed, it will be a good idea to transform them by using the natural logarithm. Also the variables NONWHITE and POOR are skewed, and it may be reasonable to transform them using a cube root.

Analyze the data set given below using the regression method after transforming the variables. Perform a complete analysis including residual analysis, variable selection etc. Prepare a thorough report as you would do for a client. This report should include all the steps used in the analysis, their justifications (relevant plots, analysis of residuals, diagnostics etc.), and your conclusions. Also comment on the possible improvements that can be made on your analysis if you detect nonlinearities, unequal variance, outliers etc. Please cut and paste the relevant portions from your computer printouts. Please attach your R codes in an appendix of your report.

You may want to follow the steps given below with a summary of your findings for each step.
1. There should be an introduction with a brief description of the data and goal of the analysis.

2. Obtain a matrix plot of the data, calculate the correlation matrix, fit the regression, obtain the ANOVA table, estimates of the parameters and their standard errors etc.
3. Do the diagnostics: plot the observed Y values against the fitted Y-values, plot the residuals against the independent variables, histogram of residuals, normal probability plot of residuals etc.
4. If you believe there is some nonlinearity in the data from your analysis in steps 2 and 3, then include the nonlinear terms (such as squares), fit the regression, obtain the ANOVA table, estimates of the parameters along with their standard errors, plot of observed against fitted Y values, plot of residuals against the fitted values, histogram of residuals, normal probability plot etc. [There may be no need to do Box-Cox transformations if you begin the data analysis after transforming the variables as suggested.]
5. If you believe some variables can be deleted from your model (either from step (2) if you do not suspect nonlinearity or the model from step (4) if you suspect nonlinearity), then use all subsets regression (if possible using the computer) and stepwise procedures for model selection. If you have used both procedures (all subsets and stepwise), then comment on the differences between the results, if any.
6. Summarize your findings. Briefly discuss if further analysis is needed for this data.
7. Attach all the R codes in an Appendix of the report.

Format
- Reports should be typed.
- The report should include a title page.
- The main body of the report should contain no R code.
- All R code should be included in an appendix.