STAT 6337
Advanced Statistical Methods I (Fall 2024)
Project 3

This project is individual work. So do not consult with anybody in or out of class. You can ask me or TA questions if something is not clear.

Complete this page below and attach with your project. Your project will not be graded without it.

This project is entirely my work. I have not discussed about this project with anybody in or out of class. I understand and have complied with the academic integrity policies written in the *Handbook of Operating Procedures* of UT Dallas https://policy.utdallas.edu/utdsp5003**.**

YOUR NAME _____

DATE _____

YOUR SIGNATURE (NOT just typed name) _____

Stat 6337 (Advanced Statistical Methods) Fall 2024

Project #3

Notes:

- You are supposed to work on this project entirely on your own. So, do not consult with anyone within or outside the class.
- You are welcome to ask me or TA questions. However, first try to find the answer on your own. Don't be afraid to google! It is a necessary skill for becoming a successful programmer. If you ask us for help with SAS commands, we will first ask you what you found by googling (and what code you have written so far).
- The dataset is comma separated (.csv extension). To read it in SAS, use the options DSD and FIRSTOBS = 1 in the INFILE statement (e.g., INFILE auto DSD FIRSTOBS = 1;)
- For any test of significance, include the hypotheses, test statistic values, p-values, and conclusion.

Consider the health dataset, which consists of data from 54 cities. We want to build a multiple linear regression for predicting death rate per 1000 residents.

normality assumptions?
const var?

1. Fit a multiple linear regression model using all available predictors. Carry out regression diagnostics (including plot of absolute residuals). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. Clearly state each assumption, the diagnostic tools used to check it, and the conclusion. Use all tests and plots that were discussed in class.

2. If an assumption is not met, attempt to remedy the situation. Explain the steps used to obtain the transformed model. Comment on the fit the transformed model using appropriate tests and statistics.

   Note: For the remaining parts, continue in transformed scale, if a transformation was applied earlier.

3. Use the principle of extra sum of squares (type I and III SS) to determine which variables can be removed from the model (try removing one variable at a time — the least significant one). Once a tentative model is obtained, compare it with the initial model (with all variables but in transformed scale obtained earlier) using appropriate extra sum of squares. Clearly state at each step the hypotheses being tested, the appropriate extra sum of squares, test statistic, p-value, and conclusion.

4. Starting with the full model, find the best model(s) using adjusted R^2, C_p, and BIC criterion. Also, find models using stepwise, forward, and backward selection methods. Compare all these models.

5. For one of the "final" models obtained from the previous part, report coefficients of multiple determination, multiple correlation, partial correlation, and partial determination.

6. Consider the largest coefficient of partial determination. Show that its alternative interpretation in terms of coefficient of simple determination holds by fitting appropriate models and calculating R^2. Similarly, show that the alternative interpretation of the coefficient of multiple determination holds.

7. Use the final model to obtain 95% interval estimates for the mean response for the entire range of predictor and plot them against each predictor in the final model separately (by fixing other predictors' value to their averages). In the same plot, add two corresponding sets of intervals — (i) 95% interval estimates for death rate of a new city (not in the dataset) (ii) 95% simultaneous confidence bands for the entire regression line. Compare the three sets of intervals.

(i) PI
(ii) Bonferroni?

Submit your report with the following components (in this order):

- at most 5 pages of answers;
- relevant parts of SAS output with relevant numbers highlighted (label each part of the question);
- your SAS code including brief typed comments of main steps (with label for each part).