STAT 6337
Advanced Statistical Methods I (Fall 2024)
Project 4

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)

Project# 4

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.

Questions 1 to 5 are based on the cardio dataset. It contains data on 998 individuals on the following variables.

#	Variable	Description
1	uric	Uric acid level
2	dia	Diastolic blood pressure
3	hdl	High-density lipoprotein cholesterol
4	choles	Total cholesterol
5	trig	Triglycerides level in body fat
6	alco	Alcohol intake (ml per day)

1. Fit a full model for predicting uric acid levels using all other explanatory variables. Find the best model(s) using adjusted R^2 criterion.

2. For the best models chosen above, check all assumptions (using all plots and tests discussed in class), detect any outliers and influential points, and check for collinearity using appropriate diagnostic tools.

3. Fit a weighted least squares regression model using the same variables as the best model above. Check if iterating the process of estimating weights improves the estimates. Comment on the final results.

4. For the above selected model, use iteratively reweighted least squares approach to robust regression for dampening the influence of outlying cases. Use Bisquare weight function and MAD (for calculation of scaled residuals). The initial residuals may be obtained from OLS model. Carry out at least three iterations (excluding iteration 0) and compare the residuals, weights, and regression coefficients across iterations including the ones from OLS (Note: Don't use any in-built SAS procedures for robust regression).

5. Fit a simple linear regression model of uric acid on triglycerides. We would like to perform inference on the regression coefficient β_1. Use nonparametric bootstrap with at least 1,000 resamples and two methods of resampling in regression set-up: (1) Resampling of residuals (fixed X sampling) and (2) Resampling of (X, Y). For each method, report the following:

   - histogram and Q-Q plot of the bootstrap distribution of $\hat{\beta}_1$ with comments about the shape of the distribution
   - bias and standard error (SE) of $\hat{\beta}_1$

- 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\beta}_1$
- 2.5th and 97.5th percentiles of the sampling distribution of $\hat{\beta}_1 - \beta_1$
- 95% confidence interval (CI) for β_1 using three bootstrap methods — normal approximation, basic bootstrap, and percentile bootstrap

Compare the results from the two methods of resampling and the results (SE and CI) from the simple linear regression output. (Note: Don't use any in-built SAS procedures for bootstrap).

6. Consider the breast tumor dataset. It has the following variables (and their possible values) on breast tumors:

1 Clump Thickness: 1 - 10

2 Uniformity of Cell Size: 1 - 10

3 Uniformity of Cell Shape: 1 - 10

4 Marginal Adhesion: 1 - 10

5 Single Epithelial Cell Size: 1 - 10

6 Bare Nuclei: 1 - 10

7 Bland Chromatin: 1 - 10

8 Normal Nucleoli: 1 - 10

9 Mitoses: 1 - 10

10 Class: 0 for malignant, 1 for benign

The goal is to develop a logistic regression model for predicting whether a tumor is malignant or not. For each predictor, fit a separate simple (univariate) logistic regression model. Select the predictors that are significant at univariate p-value of 0.1. Fit a multiple logistic regression model with these selected predictors and use a likelihood ratio test (LRT) to test if these predictors are significant jointly. If they are not found to be signficant jointly, remove the least significant variable and test the significance of the reduced model using LRT. Repeat the steps until you have a model with all significant predictors. Compare this model with the full model using LRT and if necessary, add any of the predictors that were removed. Interpret the coefficients of the final model. Compare this model with the one obtained using stepwise procedure.

Useful links:

- `http://blogs.sas.com/content/iml/2016/08/10/bootstrap-confidence-interval-sas.html`
- `http://blogs.sas.com/content/iml/2014/01/29/sample-with-replacement-in-sas.html`

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

- at most 10 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).