Statistics 108, Project 2, *Due: December 4th by 5pm*

Turn in the report in electronic form (word or pdf) through Canvas

Instruction: This project is to analyze a dataset, from start to finish, based on the multiple linear regression model. It is an *individual* project. Students could discuss with each other to get better understanding of the project. Copying solutions or computing codes from other students or other sources is plagiarism. At a minimum, all students involved will receive a 0 on this project for any type of academic dishonesty.
R codes: Attach the entire R codes you used to analyze the data at the end of the report.

Data description: The data "diabetes.txt" contains 16 variables on 366 subjects who were interviewed in a study to understand the prevalence of obesity, diabetes, and other cardiovascular risk factors in central Virginia for African Americans. We will consider building regression models with `glyhb` as the response variable as Glycosolated Hemoglobin > 70 is often taken as a positive diagnostics of diabetes. The goal is to find the "best" model for later use.

Data exploration and split data for validation later on.

1. Among all the variable, which of the variables are quantitative variables? Which are qualitative variables? Draw histogram for each quantitative variable and comment on its distribution. Draw pie chart for each qualitative variable and comment on how its classes are distributed. Draw scatterplot matrix and obtain the pairwise correlation matrix for all quantitative variables in the data. Comment on their relationships.

2. Regress `glybh` on all predictor variables (Model 1). Draw the diagnostic plots of the model and comment.

3. You want to check whether any transformation on the response variable is needed. You use the function '`boxcox`' to help you make the decision. State the transformation you decide to use. In the following, we denote the transformed response variable to be `glyhb`*. Regress `glyhb`* on all predictor variables (Model 2). Draw the diagnostic plots of this model and comment. Apply `boxcox` again on Model 2; what do you find?

4. Randomly split data into two equal halves: a training data set and a validation data set.

Selection of first-order effects. We now consider subsets selection from the pool of all first-order effects of the 15 predictors. `glyhb*` is used as the response variable for the following problems.

5. Fit a model with all first-order effects (Model 3). How many regression coefficients are there in this model? What is the MSE from this model?

6. Consider best subsets selection using the R function `regsubsets()` from the `leaps` library with Model 3 as the full model. Return the top 1 best subset of all subset sizes (i.e., number of X variables) up to 16 (because `frame` has 3 levels). Get SSE_p,R^2_p, $R^2_{a,p}$, C_p, AIC_p,BIC_p for each of these models, as well as the none-model (the model with only an intercept). Identify the best model according to each criterion. For the best model according to C_p criterion, what do you observe about its C_p value? Do you have a possible explanation for it?

   Denote the best models according to AIC, BIC, and adjusted R^2 be Model 3.1, Model 3.2, Model 3.3, respectively. (It is possible that some of the three models are the same.)

Selection of first- and second- order effects. We now consider subsets selection from the pool of first-order effects as well as 2-way interaction effects of the 15 predictors.

7. Fit a model with all first-order and 2-way interaction effects (Model 4). How many regression coefficients are there in this model? What is the MSE from this model? Do you have any concern about the fitting of this model and why?

8. Apply the `forward stepwise procedure` using R function `step()` (or `stepAIC()`), starting from the none-model and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs1. Compare its AIC value with that of Model3.1. What do you find?

9. Apply the `forward stepwise procedure` using R function `step()` (or `stepAIC()`), starting from the full model (Model 3) and using the AIC_p criterion. What is the model being selected? Denote this model by Model fs2. Compare its AIC value with that of Model fs1. What do you find?

10. Compare the BIC values of Model fs1 and Model fs2. What do you find? Do AIC and BIC choose the same model among these two models or not? Denote the model selected by AIC among the two models by Model 4.1 and that selected by BIC be Model 4.2. (It is possible that Model 4.1 and Model 4.2 are the same model.)

Model validation. We now consider validation of the models (Model 3.1, Model 3.2, Model 3.3, Model 4.1, Model4.2) you selected in the previous studies.

11. Internal validation. We use $PRESS$ for this purpose. Calculate $PRESS$ for each of these models. Comment.

12. External validation using the validation set. For each of these models (Model 3.1, Model 3.2, Model 3.3, Model 4.1, Model4.2), calculate the mean squared prediction error (MSPR), i.e., you use the model to predict the 183 observations in the validation set and calculate the averaged squared prediction error. How do these MSPRs compare with the respective $PRSSE/n$ (here n is the sample size of the training data, i.e., 183). Which model has the smallest MSPR?

13. Based on both internal and external validation, which model you would choose as the final model? Fit the final model using the entire data set (training and validation combined) (Model 5). Write down the fitted regression function and report the R summary() and anova() output.