# Quiz #4. AMS 580

# Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_SBU ID:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Dear all, take home Quiz #4 is due on Tuesday, April 6, 2021, at 8:00am. Quiz #4 solutions will be provided right afterwards. Please submit your quiz to the Blackboard in a word or PDF document generated by the RMD file.

# Please include (1) R code; (2) Output from R; (3) Answers to all the questions asked.

#### Logistic Regression (\* A type of Generalized Linear Model when the response is binary) with the bwt Data

The bwt data is a built in R dataset containing 9 variables and 189 cases – please see the following link for detailed description:

<http://garthtarr.github.io/mplot/articles/birthweight.html>

Your first task is to use all the other 8 variables to predict the binary dependent variable ‘low’ that indicates whether a given infant was born with low birth weight or not. You will split the data into 80% training and 20% testing, using seed = 123, and follow the procedures from the following website:

<http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/>

You second task is to perform variable selection using (a) Stepwise variable selection, and (b) Best-subset variable selection, based on detailed descriptions of the related methods in the following paper:

<http://atm.amegroups.com/article/view/9706/pdf>

Please note that for the logistic regression model, one does not use the coefficient of determination for model selection, but rather, one should use the information criterion, typically either the AIC, or the BIC.

They are both functions of the likelihood albeit with the opposite signs from the likelihood:

<https://en.wikipedia.org/wiki/Bayesian_information_criterion>

<https://en.wikipedia.org/wiki/Akaike_information_criterion>

Since a larger likelihood corresponds to a better model, a smaller AIC or BIC will correspond to a better model. The BIC will usually results in a smaller model, hence more robust and preferred than the AIC.

1. Task 1.
2. Please split the data into 80% training and 20% testing using seed =123.
3. Then you shall fit a logistic regression model with all the other 8 predictors using the training data.
4. Please use this fitted model based on the training data to predict the response variable ‘low’ (baby would be born with low birth weight or not) for the testing data. Please generate the confusion matrix, and report (i) the overall accuracy, (ii) the sensitivity (that is, the probability a baby is predicted to be born with low birth weight given that he/she was in fact born with low birth weight), and (iii) the specificity (that is, the probability a baby is predicted to be born with normal birth weight given that he/she was in fact born with normal birth weight).
5. Task 2 – Part I: Stepwise Variable Selection.
6. Please split the data into 80% training and 20% testing using seed =123.
7. Then you shall fit a logistic regression model that best predicts whether the infant would be born with low birth weight or not using the stepwise variable selection method (both directions) and the BIC based on the training data. (\*Please do not consider interactions.)
8. Please use this fitted model based on the training data to predict the response variable ‘low’ (baby would be born with low birth weight or not) for the testing data. Please generate the confusion matrix, and report (i) the overall accuracy, (ii) the sensitivity (that is, the probability a baby is predicted to be born with low birth weight given that he/she was in fact born with low birth weight), and (iii) the specificity (that is, the probability a baby is predicted to be born with normal birth weight given that he/she was in fact born with normal birth weight).
9. Task 2 – Part 2: Best-Subset Variable Selection.
10. Please split the data into 80% training and 20% testing using seed =123.
11. Then you shall fit a logistic regression model that best predicts whether the infant would be born with low birth weight or not using the best subset variable selection method and the BIC based on the training data. (\*Please do not consider interactions.)
12. Please use this fitted model based on the training data to predict the response variable ‘low’ (baby would be born with low birth weight or not) for the testing data. Please generate the confusion matrix, and report (i) the overall accuracy, (ii) the sensitivity (that is, the probability a baby is predicted to be born with low birth weight given that he/she was in fact born with low birth weight), and (iii) the specificity (that is, the probability a baby is predicted to be born with normal birth weight given that he/she was in fact born with normal birth weight).
13. Among all three models obtained in questions 1, 2 and 3 above, which one is the best and why?

