STA 6443 Final Exam

1. **Final code (.R or .Rmd)** and **complete report (.pdf or .docx)** should be submitted by Wednesday, Dec 9, 11:59 pm on Blackboard. Please NO zip files.
2. Submit the complete report file **with only relevant graphs and tables**.

Use significance levels of .05 unless the instructions state otherwise.

Data Sets:

You need to download dataset **birthweight_final.csv**. The data record live, singleton births to mothers between the ages of 18 and 45 in the United States who were classified as black or white. There are total of 400 observations in **birthweight,** and variables are:

- **Weight:** Infant birth weight (gram)
- **Weight_Gr;** Categorical variable for indication of low birthweight; 0 is normal, **1** is **low birthweight**
- **Black:** Categorical variable; 0 is white, 1 is black
- **Married:** Categorical variable; 0 is not married, 1 is married
- **Boy:** Categorical variable; 0 is girl, 1 is boy
- **MomSmoke:** Categorical variable; 0 is non-smoking mom, 1 is smoking mom
- **Ed:** Categorical variable for Mother's education Level; 0 is high-school grad or less; 1 is college grad or above
- **MomAge:** Mother's age (centered to zero)
- **MomWtGain:** Mother's weight gain during pregnancy (centered to zero)
- **Visit:** number of prenatal visits

Exercise 1 (25 points)

Consider to fit a multiple linear regression to model **Weight** using possible explanatory variables; **Black**, **Married, Boy, MomSmoke, Ed, MomAge, MomWtGain**, and **Visit** (all predictors excluding **Weight_Gr**).

(1) Perform the following four model selection methods and compare their best models. Comment on how they differ or similar in terms of selected variables in the final model. No need to interpret outputs.
   - Stepwise selection with **0.01 p-value criteria** for both entry and stay
   - Forward selection with **0.01 p-value criteria** for entry
   - Backward selection with **0.01 p-value criteria** for stay
   - Adjusted R-squared criteria

NOTE: R output from Backward selection displays variables "removed" from each step.

Answer following questions from the best model determined by Stepwise selection with **0.01 p-value criteria**

(2) Fit the linear regression with the best model determined by stepwise selection and comment on diagnostics plot. Do not leave observation which has Cook's distance larger than **0.115**. Re-fit the model if necessary. Finally how many observations you use in the final model?

(3) How much of the variation in **Weight** is explained by the final model?

(4) Interpret the relationship between predictor variables (in the final model) and Weight value specifically.

Exercise 2 (30 points)

Now we consider fitting a logistic regression for low birthweight (**Weight_Gr**=1). Again consider **Black**, **Married, Boy, MomSmoke, Ed, MomAge, MomWtGain**, and **Visit** as possible explanatory variables.

(1) Perform following model selection methods and compare their best models. Comment how they differ or similar in terms of selected variables

- Stepwise selection with AIC criteria
- Stepwise selection with BIC criteria

NOTE: stepwise selection with BIC criteria can be performed by step() function by adding an option **k=log(n),** where n is a sample size. Check Week 15 respiratory data example - how to use this option.

Answer following questions from the best model determined by stepwise selection with BIC criteria

(2) Fit the logistic regression with the best model determined by stepwise selection with BIC criteria. Do not leave observation which has cook's d larger than **0.1**. Re-fit the model if necessary. Finally how many observations you use in the final model?

(3) Based on your final model, interpret the explicit relationship between response and predictors using Odds Ratio.

(4) Which woman has the high chance to deliver a low birthweight infant? For example, answer will be like "a married, high-educated, and older woman has the high chance to deliver a low birthweight infant."

(5) What is the sample proportion of low birthweight infant in dataset?

(6) Perform classification with probability cut-off set as sample proportion you answer in (5). What is misclassification rate?

(7) Comment on Goodness of fit test and make a conclusion

Exercise 3 (15 points)

Compare results from Exercise 1-2 and comment on different or similar conclusions from each analysis.

Low birthweight is a risk factor that can lead infant mortality. If you want to implement a low-birthweight prevention program, what would you suggest to pregnant women?