STA 6443 Final Exam

1. **Final code (.R or .Rmd)** and **complete report (.pdf or .docx)** should be submitted by Thursday, Dec 7, 11:59 pm on Canvas (No .zip files). Prepare your report file containing relevant tables or plots.
2. **Use significance levels of .05** unless the instructions state otherwise.

Data Sets:

You need to download the 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 a total of 400 observations in **birthweight,** and the 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 fitting 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 a stay
- Adjusted R-squared criteria

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

Answer the 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 the diagnostics plot. Do not leave an observation that has Cook's distance larger than **0.115**. Re-fit the model if necessary. Finally, how many observations did 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 the following model selection methods and compare their best models. Comment how they differ or are similar in terms of selected variables

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

Answer the 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 an observation that has Cook's d larger than **0.1**. Re-fit the model if necessary. Finally, how many observations did 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 of delivering a low birthweight infant? For example, the answer will be like "a married, high-educated, older woman has a high chance of delivering a low birth weight infant."

(5) What is the sample proportion of low birth weight infants in the dataset?

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

(7) Comment on the 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 birth weight is a risk factor that can lead to infant mortality. If you want to implement a low-birthweight prevention program, what would you suggest to pregnant women?