1. **Final R (or Rmd) file** and **complete report (word of pdf) file** should be submitted by **Saturday, Oct 9, 11:59 pm** on Blackboard. Please NO zip files.
2. It should be the **individual work, NO team work**
3. Please write your **name** at the top of the report file

Use significance levels of 0.05 unless the instructions state otherwise.

Data Sets:

You need to download dataset **birthweight.csv** for Exercise 1-4. The **birthweight** 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)
- **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

Exercise 1

(a) Generate Boxplot for infant birth weight (**Weight**) and comment on the general features of the distribution. Generate a normal QQ-plot and perform Shapiro-wilk test to check whether normality is a reasonable assumption for **Weight**. Make a conclusion.

(b) Generate a boxplot of **Weight** by **MomSmoke** and compare infant birth weights between smoking levels.

(c) For each level in **MomSmoke**, perform Shapiro-wilk test for checking the Normality of **Weight**. Make a conclusion.

Exercise 2

We want to test if there is a significant difference in birth weights between infants from smoking mom and non-smoking mom.

Perform a hypothesis test of whether infants from smoking moms have different weights than infants from non-smoking moms. Which test do you choose? Use the answer in Exercise 1 for choosing the proper test. Specify null and alternative hypotheses and state your conclusion

NOTE: If you decide to use the parametric test, perform **two-sample t-test** rather than **ANOVA.**

Exercise 3

Now perform one-way ANOVA on **Weight** with **MomSmoke**.

(a) Check homogeneity of variance assumption. Does it hold and okay to perform ANOVA?

(b) Make a conclusion on the effect of **MomSmoke**. Compare your result with the conclusion of Exercise 2.

Exercise 4

Using **Black, Married, Boy,** and **MomSmoke,** and **Ed** variables as possible effects, find the best ANOVA model for **Weight**. Manually perform backward selection based on type3 SS result with **0.05** criteria on p-value. Perform **backward selection only with main effects** and then check the **interaction** effects only based on significant main effect terms.

NOTE: For backward selection, you remove a variable from the least significant one, ONE BY ONE, until there is no more variable with a p-value larger than the criteria.

(a) Write down step by step how you perform backward selection and how you find the final model. Please do NOT include all intermediate tables and graphs in the report. Just describe each step which variable you delete and why.

(b) Specify the final model and report the amount of variation explained by the model. Also, check the Normality assumption through diagnostics plots.

(c) State conclusions about significant differences in **Weight** across groups. For each significant variable, state specifically which level has a larger or smaller mean value of **Weight**.