PS1

In this problem set, you will be working with a sample from the National Center for Health Statistics (NCHS) data. The objectives are: 1) know what a dataframe is and how to open a csv file, ii) plot a histrogram that shows the distribution of a continuous variable, iii) run a simple regression, and iv) perform some minor data cleaning.

Please work with your group members and discuss thoroughly. If someone in your group is struggling, be sure to offer help. Note that you need to turn in both R code and written document describing what you did.

1. Open the data set "NCHS_birthweight2000_sample.csv" located in the **gitpage**. Note that this is a csv file. You will need to save the file in a particular location in your computer and state the path before you open the file. For example, if you store your file in this path: /home/desktop, then use the command:

   data <- read.csv("/home/desktop/NCHS_birthweight2000_sample.csv") to read the file. Comment on what does read.csv do.

2. Next, plot a histogram using hist(data$dbirwt). Note the difference between hist(dbirwt) and hist(data$dbirwt). Try it and comment why one command [hist(data$dbirwt)] works and the other [hist(dbirwt)] does not. What can you comment regarding the distribution of the birthweight?
3. The sample data has 5 variables: 1) dbirwt (birthweight of an infant in grams), 2) dmeduc (mother's education), 3) race_white (an indicator for if mother's race is white), 4) cigar6 (if smoked during the time of pregnancy), and 5) dmage (mother's age). Using the **lm** function in R, regress infant birthweight on mother's education.
4. Use the command table(data$dmeduc) to tabulate education values in the sample. Note that the education values goes from 0 to 17. However, 88 observations have education value of 99, which is wrongly coded. Write a command to delete observations with education value of 99. Use: data <- subset(data, educ != 99). Comment on what this command is doing and specifically what **!=** represents.
5. Using the corrected data from 4, re-run the regression asked in point 3.
6. Let's run a multivariate regression. Add race (race_white) and mother's age (dmage) in the specification. Describe what the coefficients on: i) mother's education, ii) mother's race, and iii) mother's age are telling you.