HW 3 Winter 2020 Due Friday Feb. 14, 2020 @ 5 pm

The North Carolina Data:

Background:

In 2009, the state of North Carolina released to the public a large data set containing information on births recorded in this state. This data set has been of interest to medical researchers who are interested in studying the relation between habits and practices of expectant mothers and the birth of their children.

The unit of observation is each birth. There are 10,000 observations and 132 variables recorded in the data set. Some of the predictors are categorical and others are numerical.

Some of the Numerical Variables:

Birth Weight (g), Age of Father, Age of Mother, Visits, Wt Gain Education of father (years), Education of mother (years), Gest Age and AveCigs.

Some of the Categorical Variables:

Gender, Premie, LowBirthweight, Marital, Racemom, Racedad, Hispmom, Hispdad, Habit, MomPriorCond, BirthDef, DelivComp, BirthComp

Question 1: Using North Carolina data set posted on CCLE week six.

Create the table of correlation (4 decimal places) among the following quantitative variables. (Remember you cannot compute Pearson's coefficient of correlation among qualitative variables).
Birth Weight (g), AveCigs, Age of Father, Age of Mother, Visits, Wt Gain and Gest Age.

1) You may name your correlation matrix as corrmat and then use the library corrplot If you don't have the corrplot package please install it
2) You may use the scatterplot.matrix function to study the densities of the numerical predictors
3) You may use the corrplot function to create a visual representation of the correlation matrix. What do you notice?

Question 2: Use Birthweight to represent your response variable.

1) As the first block of predictors, pick three numerical predictors based on the table of correlation (correlation to the response variable). " Don't worry about the multicollinearity issue at this point". Keep in mind (Those predictors might be highly correlated with each other).
2) Create a multiple linear model call it Model1 using the selected three numerical predictors. Test the relevant assumptions, and check its diagnostic.

3) Create another MLR model based on standardized partial coefficients call it model1.1 (see week six lecture notes for information on partial standardized residuals for the relevant library that you need to install and the relevant commands). Library: beta.lm is a good candidate. What are the most important predictors based on model1.1?

Question 3: As the second block for your MLR model1, pick three other predictors:

1) Add three numerical predictors to the MLR model1 in question 2. Call it model 2. So, model 2 should now have six numerical predictors.
2) Test the relevant assumptions, and check its diagnostic.
3) Create the linear model for model2 based on standardized coefficients. Call it model2.1. What are the most important predictors based on model2.1?

Question 4: Present the summary of results for model1 and model2:

1) Compare the models that you reported in question 2 and question 3 above.
2) Is it worth having six predictors instead of three?
3) How much does your R^2 increase?
4) Is the change significant? Interpret the results within context.

Question 5: Pick a categorical predictor with two categories: (smart selection of the categorical predictor)

1. Use the selected categorical predictor (alone) to predict your response variable (as a SLR) call it model3, the check assumptions. Does that categorical predictor have a decent R^2? Explain.
2. Add the picked categorical predictor to model2 and call it model4 (total of seven predictors; model4 has six numerical predictors and one categorical predictor). Does that categorical predictor enhance model2 and its R^2? Explain.