DA5020 – Assignment 10

In the last assignment, you performed simple linear regression. This assignment provides you with an opportunity to build a causal forecasting model using multiple regression. Even though the dataset has changed, the concept is the same. However, we are now integrating more variables in the model and we will evaluate their significance when making predictions.

Question 1 — (5 points)

In your own words, provide a clear definition of the confidence interval and the prediction interval, and state their respective significance.

Install the **openintro** R package and load the library in your R environment. Use the **ncbirths** dataset to answer the following questions

Question 2 — (5 points)

Load the data, build a full correlation matrix and explain the chart. Do you detect any multicollinearity that would affect the construction of a multiple regression model? Are there any fields that may not be useful (or redundant) when building the model? If yes, provide an example.

Question 3 — (10 points)

Build a full multiple regression model that predicts the birth weight i.e **weight**. Display the model's summary and comment on the: R-squared, Standard Error, F-Statistic, p-values of coefficients.

Question 4 — (20 points)

Using the model from the previous question, perform stepwise elimination based on coefficients with the p-value > 0.05 to ensure that <u>all coefficients are significant</u>. Demonstrate each step as you eliminate the coefficients and indicate if the model is improving.

Question 5 — (10 points)

Use the following data to predict the birth weight using the final model from question 4 above: fage = 40, mage = 32, mature = 'mature mom', weeks = 42, premie = 'full term', visits = 12, marital = 'married', gained=22, lowbirthweight = 'not low', gender = 'female', habit = 'nonsmoker', whitemom = 'white'. After which, derive the 95% confidence and prediction intervals for the forecasted birth weight. Comment on the results. <u>Note: you can use a library to make your predictions.</u>

Submission Details

- Your submission must contain two files: the .Rmd file and a knitted PDF or HTML (from the .rmd). Name your .Rmd file, DA5020.A10.FirstName.LastName.Rmd and your PDF/HTML DA5020.A10.FirstName.LastName.{pdf,html}, where *FirstName.LastName* is your first and last name.

- The .Rmd file must be fully commented and properly "chunked" R code and detailed explanations. Make sure that it is easy to recognize which question you answer and that your code runs from beginning to end (because that is how we will test it). Code that doesn't execute, stops, throws errors will receive no points. If the TAs have to "debug" your code or spend any effort getting it to run, substantial points will be deducted.

- Not submitting a knitted PDF or HTML will result in reduction of 30 points.

- Not submitting the .Rmd file (or both) will result in a score of 0.

** Note: Do not merely print the output for each question but always explain the results from your code. Points will be deducted for any missing explanation and any visualizations that are not clearly labeled.