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 in this assignment, 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.

Describe in your own words what a multiple linear regression is and why one would be used.

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 in your R environment and build a full correlation matrix ,i.e. a matrix that shows the correlations between all variables. Do you detect any multicollinearity that would affect the construction of a multiple regression model? Comment on the distribution of each field. Do you anticipate that there are fields that may not be useful for the model? If yes, provide an example.

Question 3 — (5 points)

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

Question 4 — (25 points)

Build a multiple regression model in which all coefficients are significant — use stepwise elimination based on coefficients with the p-value > 0.05. Show each step as you eliminate the coefficients and clearly state the reason for their elimination. At each step, determine 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.

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.