MTH-245 Project Proposal

Topic & Task: The MTH-245 project centers on multiple linear regression analysis. You will be tasked with, working in a team of two-three students, finding an interesting and robust dataset, graphically and numerically summarizing your data, choosing the "best" multiple linear regression model, and making appropriate interpretations and conclusions. The "best" model could be based on fit, significance, predictive ability, etc. or a combination of the above. To this point, having more than one best model is permitted.

Proposal: More details about the project will soon be provided. For now, gather with your team and have one member fill in this form and submit the PDF to Box by **Monday, November 7 by 11:59pm.** *This is your project proposal and must be approved before moving on to the next stage.*

1. ***Team.***

   - Name 1: Marissa Patel

   - Name 2: Michael Peeler

   - Name 3:

2. ***Dataset.*** Find an interesting and robust dataset. I've provided a few resources at the end of this document, but you are encouraged to utilize other resources as well. Using datasets from other regression-based projects is not allowed.

   (a) In one or two paragraphs, provide a description/background about your data.

   The data were obtained from the Stat2 textbook's datasets. The data contains information about a number of babies born in North Carolina in 2001, as well as information about their mothers. The variables are either continuous quantitative, discrete quantitative, or categorical, and are related either to the mother of the baby or the baby itself. An example would be our selected continuous quantitative response variable, the birth weight of the child.

   (b) Choose **one** continuous quantitative response variable.

   - **Response Variable: Birth Weight**

   (c) Choose **at least 5** predictor variables; they can be a mixture of quantitative and categorical, but don't have to be. **For each variable that you pick, determine if it is categorical, continuous quantitative, or discrete quantitative.**

   Remark: Be careful in choosing between discrete quantitative and categorical. If your variable is discrete quantitative, but has "few" categories, then it should be treated as categorical. If your variable is discrete quantitative, but has "many" categories, then it should be treated as continuous quantitative.

- **Predictor Variable 1: Race. Catagorical with 4 levels (White, Black, Hispanic, Other).**
- **Predictor Variable 2: Gestation Period. Discrete Quantitative.**
- **Predictor Variable 3: Sex. Catagorical with 2 levels (Male, Female).**
- **Predictor Variable 4: Mother is a Smoker. Categorical with 2 levels (Yes, No).**
- **Predictor Variable 5: Mother's age. Discrete quantitative.**
- More predictor variables?

3. ***Observations.*** How many observations (rows) are in your dataset? (**50 or more**)

   There are 1,450 observations in the dataset.

4. ***Questions.*** What research questions are of interest to you? Come up with **at least two**.

   Remark: These research questions are subject to change as you progress through the project, which is okay. Consider these preliminary questions of interest.

   - **Question 1: Which of the predictor variables, if any, strongly predict the response variable?**
   - **Question 2: How influential are outlier instances, if any, on the model?**
   - More questions?

Some Resources:

- http://users.stat.ufl.edu/ winner/datasets.html
- https://guides.emich.edu/data/free-data
- http://www.statsci.org/datasets.html
- http://jse.amstat.org/jse_data_archive.htm
- http://www.hawkeslearning.com/Statistics/dis/datasets.html
- https://www.kaggle.com/