EAS 508 (Section 002): Homework 1

Due October 1, 2021 by 11:59 PM

Notes: The homework should be submitted via UBLearns. 25% of the grade will be removed for each day late (including 12:00 AM October 2, so make sure that you give sufficient time to submit). The answers to these problems should be succinct answers, accompanied by figures as needed. We do not want to see code as the coding is not the end result but is a means to answer the questions. We are interested in the answers and how you understand the techniques. The overarching motivation of this homework is to practice thinking like a data scientist and negotiating with messy real-world data.

Question:

In class, we have discussed multiple linear regression techniques. On the UBLearns page, a dataset named "CANCER.csv" has been uploaded. As a part of our practice to get introduced to messy real-world data, I uploaded the data dictionary for this data set which will describe the purpose of each independent and dependent variable. In class, we discussed three approaches, Plug All Independent Variable In, Parsimonious, and Comprehensive, based on the situation we may face. For HW1, we intend to develop a multiple linear regression model using the Comprehensive approach. Using this data and based on our discussions, answer the following questions:

1. What is your suggested model for predicting TARGET_deathRate? Give equation and method with final parameters. **(5 points)**
2. What descriptors did you select and why? What impact should that have on the model? **(5 points)**
3. Why did you select the model as your final answer? **(5 points)**
4. Discuss your result in a maximum of 500 words and 3 supporting plots. The discussion needs to be coherent and understandable to a non-expert (Assuming that we are communicating our analysis to a higher official who is not a data scientist). **(10 points)**

Note that there is not one definitive answer, so there are many acceptable answers as long as you justify your model on the above points.