STAT 6012: Linear Models for Data Science
Class Activity 7

Due date: Wednesday, July 24 by 10:50 am Via Canvas.

Complete the following questions an R Markdown file and submit your compiled HTML file. If you are working in a group, list the names (last, first) of the group members in alphabetical order of last names.

The attached dataset contains information compiled by the World Health Organization and the United Nations to track factors that affect life expectancy in 2015.

Our goal is to build a multiple linear regression model to predict `life.Expectancy` using:

`Adult.Mortality,infant.deaths,HIV.AIDS,BMI, GDP,Schooling`.

To make calculations easier, the following sample R code subsets the data on the two variables, as well as removes all missing values.

```
life_data3<-select(life_data,Life.expectancy, Adult.Mortality,
                infant.deaths,HIV.AIDS,BMI, GDP,Schooling)%>%
            na.omit()
```

1. [3] Data exploration: make a facet wrapped scatterplot to display the relationship between all the predictors and the response variable. Does a linear model seem appropriate for predicting `life.Expectancy` using all these predictors?

2. [5] Build a multiple linear regression using the `lm()` function and use it to investigate whether the multiple regression model meets all the assumptions for linear regression. Be sure the state the reasons why the model meets or does not meet the assumptions for a linear model.

3. [2] The code `summary(life_data3)` will summarize all the variables in `life_data3`. From this summary, which variable might benefit from a transformation before modeling? Justify your reasoning.

4. Homework/Practice: proceed to do a log transformation of the variables identified in Question 3, and rebuild the model. Is this model better, in terms of meeting the assumptions for linear regression?