**Regression Project**

**QMB-6304 Analytical Methods for Business**

Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

1. Load the file “6304 Regression Project Data.xlsx” into R. This file contains information on the population and other factors of 437 counties in the American Midwest. This is your full data set. Variable names are self-explanatory with those beginning with a “pop” prefix being numbers of population and those with a “per” prefix being percentages of the total population.
2. Using the “poptotal” variable in combination with “percollege” and “perprof” calculate new variables “popcollege” and “popprof”. These of course are the population in each county with a college degree, and the population with a professional job. Add these variables to the data frame.
3. Using the “popchild” and “popadult” variables calculate a new variable which will be the ratio of children to adults in each county’s population. Add this variable to the data frame.
4. Using the “popchild” and “perchildpoverty” variables calculate a new variable which will be the number of children living in poverty in each county. Add this variable to the data frame.
5. Subdivide the full data set to create two smaller data frames which include only rural and metropolitan counties, respectively. Use the “inmetro” variable for this.
6. Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 60 counties from the rural poverty data set.
7. Using the numerical portion of your U number as a random number seed and the random selection method presented in class, take a random sample of 30 counties from the metro poverty data set.

**Analysis**

1. Using the “perelderlypoverty” as the dependent variable apply any or all of the remaining numerical variables (except “id”) to parameterize the best possible fit multiple regression model. Use the “some.rural.poverty” data frame for this and apply only main-effects variables. Where needed feel free to apply any data transforms to improve this fit. Show the results of this best fit model using the summary(*df.out*) command. Describe the methodology you used to arrive at the selection of independent variables you used in your model.
2. Assess your best fit model’s conformity to the LINE assumptions of regression. State your conclusions and show appropriate graphs and/or analytical output to support those conclusions.
3. Determine whether you believe multicollinearity exists in your best fit model. State your conclusions and show appropriate graphs and/or analytical output to support those conclusions.
4. Determine if any of the counties in your “some.rural.poverty” data set have an outsized leverage in influencing your best fit model. If so, state which counties (county name and state) have this outsized influence.
5. Assess how well your best fit model predicts “perelderlypoverty” when applied to the “some.metro.poverty” data frame. Tell whether you believe the fit is better or worse than when the model is used with the “some.rural.poverty” data. Show appropriate graphs and/or analytical output to support your conclusions.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is the “Regression Project” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.