SAS Programming 3

Submission Instructions: Welcome to SAS Programming 3. Upon completion, you *should upload* your **Word Document** and **SAS codes** to Blackboard by 11:59 PM of the due date.

Multiple Regression and more!

This problem is a long data analysis problem for you to do on your own. The assignment covers all activities from the first module to the current module. Have fun!

Problem: 2016 Election

The data set ’county_level_election.csv’ includes many possible predictor variables to attempt to predict ’votergap’ (trump-clinton) from the 2016 election across $n = 3,141$ counties in the US. The variables in the data set are:

state:	the state in which the county lies
fipscode:	an ID to identify each county
county:	the name of each county
population:	total population
hispanic:	percent of adults that are hispanic
minority:	percent of adults that are nonwhite
female:	percent of adults that are female
unemployed:	unemployment rate, as a percent
income:	median income
nodegree:	percent of adults who have not completed high school
bachelor:	percent of adults with a bachelor’s degree
inactive:	percent of adults who do not exercise in their leisure time
obesity:	percent of individuals with BMI > 30
density:	population density, persons per square mile of land
cancer:	prevalence of cancer per 100,000 individuals
votergap:	percentage point gap in 2016 presidential voting: trump-clinton

*Note: there are some missing values throughout this data set. You can just ignore this issue as you fit your models since the rate of missingness is pretty low, $\approx 2\%$...sorry Alaska. To simplify your life, you may want to remove the observations with missingness from the get go.

1. Explore the data set. What factors appear to be associated the most strongly with the response? Do any associations appear non-linear?

2. Fit a regression model to predict *votergap* from all the main effects of the quantitative predictors (*population* through *cancer*, 12 of them). Call this **model1**. Include the summary output, from SAS for this model. Which variables are not significant at the 0.05 level?

3. Check and comment on the assumptions for model1. Provide at least 2 plots for this question and use them in your comments.

4. Fit a quadratic regression model to predict *votergap* from *minority* and this variable squared. Call this **model2**. Interpret the results of this model. Provide a visual of this quadratic relationship with the scatterplot of the data. What is the approximate minimum/maximum votergap based on minority alone (and at what value of minority)?

5. Fit a regression model to predict *votergap* from *minority*, *obesity*, *female*, and the interactions of *female* with the other two predictors. Call this **model3**. Briefly interpret the results (you can interpret it in terms of transformed variables).

6. Fit a regression model to predict *votergap* from 7 predictors: *obesity*, *minority*, $minority^2$, *female*, and the interactions between *female* with the other three predictors, and call this **model4**. Include the summary SAS output for this model. Perform a formal hypothesis test to determine if model4 is doing a significantly better job at predicting *votergap* than model3.

7. Use a sequential model selection technique to select a best model to predict *votergap* from all 12 quantitative predictor main effects. Call this **model5**. Include the summary SAS output for this model.

8. Fit a regression model to predict *votergap* from the variable *state* alone. Call this **model6**. Include the summary SAS output for this model. Interpret the coefficient associated with counties in Massachusetts.

9. Fit a regression model to predict *votergap* from the resulting variables in **model5** along with *state*. Call this **model7**. Interpret the coefficient associated with counties in Massachusetts.

10. Dozens of counties were left out of this dataset. Determine which of all the models 1-7 should be best for predicting these left out counties

Your Word document should include the results of your analysis and answers to the above questions. Images, tables, and references should be formatted using APA. Use the Journal Style when outputting your results in SAS to Word.