Project 1: US 2016 Presidential Primary Election

Step 1: Simple Regression Model

In this step, I have to develop a regression model that can be used to predict the result in percentage of Hillary Clinton against Bernie Sanders in the US 2016 presidential primary elections. I can only use 5 predictors to explain the data, and as such, will use a multiple linear regression to predict the results in the column HilaryPercent. To decide which model is best, I set the program to calculate all the possible models for 1 to 5 predictors. The considered predictors are those from column 2 to 51, as they are numerical values. The total number of models computed per step are as such: $\binom{51}{nbrofpredictors}$. As more predictors imply more models to compute, it could be more efficient to limit oneself to a calculation time of maximum 2-3 hours. This would have led me to compute models for 1 to 4 predictors. However, as I had time to run the model, I decided to let it run for 5 predictors as well.
I then want to minimize the AIC coefficient to select the best model among those computed.
The best model is as follows:

Predictor	EDU635213	HSG495213	INC910213	RHI225214	RHI625214
Meaning	High school graduate or higher, percent of persons age 25+, 2009-2013	Median value of owner-occupied housing units, 2009-2013	Per capita money income in past 12 months (2013 dollars), 2009-2013	Black or African American alone, percent, 2014	Two or More Races, percent, 2014
Coefficient	-7.898796e-3	-3.982381e-7	9.710374e-6	6.855099e-3	-1.301030e-2

Which is represented by the equation:

$$HilaryPercent = -7.898796e-3 * EDU635213 - 3.982381e-7 * HSG495213 + 9.710374e-6 * INC910213 + 6.855099e-3 * RHI225214 - 1.301030e-2 * RHI625214$$

This model gives a warning that there might be a strong multicollinearity in between some predictors. I notice that HSG495213 and INC910213 have insignificant coefficients compared to the other three predictors. According to the theory, if the simple linear regression of these predictors with HilaryPercent is significantly different from zero, then there is multicollinearity in the multiple regression. As such, I compute simple linear regression for HSG495213 and INC910213. As the coefficient for these regression are respectively -3.494870e-07 and -9.710374e-6, I conclude that they do not indicate a multicollinearity. I moreover also tested the other variables, as their coefficients are not relatively large, and their coefficients for simple linear regressions are not significantly different from zero either. I also computed the variance inflation factors for all the predictors. As they all are all in between 1 and 3, we can consider they are small and the multicollinearity is quite low. The reason why the condition number is large then must be coming from another factor.

I find that the per capita money income in the past 12 months from 2009 to 2013 and the Black or African American alone population percentage in 2014 have significant effects on the percentage of votes for Hillary Clinton in the primary elections as they have a t statistic of respectively 13.944 and 47.620. I moreover can note that the Black or African American alone population percentage in 2014 has more than 3 times the impact of the per capita money income in the past 12 months from 2009 to 2013 on the percentage of votes for Hillary Clinton.
The percentage of people age 25+ that are high school graduate or higher from 2009 to 2013, the median value of owner-occupied housing units from 2009 to 2013 and the percentage of people of two or more races in 2014 seem to contribute negatively to the percentage of votes for Hillary Clinton in the primary elections with t statistics of respectively -19.729, -9.088 and -8.439. I can note that the percentage of people age 25+ that are high school graduate or higher from 2009 to 2013 has twice the impact of the two other predictors. However, these negative impacts have overall less influence that the positives impacts induced by the per capita money income in the past 12 months from 2009 to 2013 and the Black or African American alone population percentage in 2014.

The assumptions made for this model are:

1. The relationship between the outcomes and the predictors is (approximately) linear.
2. The error term ε has zero mean.
3. The error term ε has constant variance.
4. The errors are uncorrelated.
5. The errors are normally distributed or we have an adequate sample size to rely on large sample theory.

The model assumptions can be checked against the property of the residuals.

- Residual against the fitted values:

I plotted the residuals against the fitted values.

The residuals are quite symmetrically distributed across the Y-axis and a bit less symmetrically across the X-axis. Nevertheless, I can consider that the assumption that the relationship between the outcomes and the predictors is (approximately) linear is verified. If I can see some outliers in the data, there are not a large number of them, and as such, I don't consider them to be a big problem in the model.

I can also note that the residuals are all grouped in the center of the plot, which can indicate that the variance of the error terms are all the same.

- Autocorrelation:

The Durbin-Watson test is at 1.938 so I can conclude that there is little autocorrelation of the residuals.

- Normality check:

I plotted the quantiles of the residuals against the theoretical quantiles of a normal distribution.

I see that the residuals stay on a straight line on the normal quantile quantile plot as such, the residuals follow a normal distribution.

- Scatter plot:

I computed several useful plots for each predictor (see the results in the Jupiter notebook).

These plots can help identify non-constant variance, violation of the assumption of linearity and potential outliers.

I found that there is a relatively strong correlation in between the model's predictions and its actual results.
For the predicted versus residual plot : we can see that the values are symmetrically distributed across the Y-axis. However they are all less or more unbalanced across the X-axis, which shows that the model is not quite accurate. However, as I desire a simple model in this step, I will not proceed to any transformation.

- **Influential points and outliers:**

I plot the influence plots which show the studentized residuals versus the leverage of each observation as measured by the hat matrix.

As can be seen from the graph, the leverage of all points vary from a scale of 10^1 to 10^3, as such some points have higher influence that some other, which could be from the influence that some county have over some other. In fact, some county might have more delegates than some other compared to their population, which mean they would have more influence on the final results. According to the wikipedia article, some county also have more influence over the final result as their primary are earlier than those of other county. As such, it does not seem particularly abnormal to have points more influential than others.

I also remark that the values of the Student's t distribution are all null up to 3 decimals. This indicates strong evidence against the null hypothesis, so I reject the null hypothesis.

Step 2: Full Regression Model

For this second step, I decided to use forward selection, as I predicted that I would add enough variables so that the numbers of predictors becomes large. I first started with a forward selection on the original predictors. Then I added interactions to the model and recomputed a forward selection. Then, I checked the model results and tried to find ways to improve.
I started by checking the normality of the data, then the residuals against fitted values, the leverage and influential points and the variance inflation factor. To stabilize the variance I could have transformed some of the variables, however, from the plot of the residuals against the fitted values, I found that the variance of the error terms seems to be equal.
I then used cross validation to evaluate the model performance. As P1Test has 300 rows, I decided to use 6 folds so that each sample was of 60 values, which seemed like a slightly low number of values per fold, but I still wanted to have multiple folds.
I then computed the fitted model for the result of the forward selection above to have access to all the statistics about the model.
Finally I predicted the results for HilaryPercent for the test data set, and exported it to csv.
I finally calculated the model's weighted mean squared errors :

Final note: I did not have time to fully run the last forward selection, so I used the formula which I got after 2 hours of running the program : 'RHI125214:RHI225214 + RHI625214:RHI825214 + HSG096213:RHI825214 + AGE775214:HSG096213 + EDU635213 + INC910213:SEX255214 + HSG495213:POP715213 + HSD410213:RHI225214 + HSG495213:INC110213 + POP645213:RHI825214 + LFE305213:RHI725214 + AGE295214:AGE775214 + HSD310213:RHI225214'