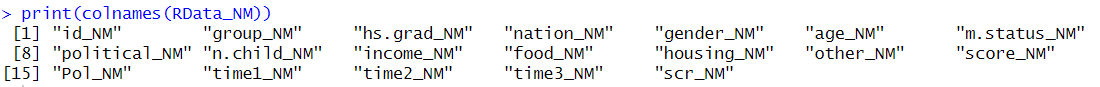
**Data Loading and Descriptive Analysis**

**Rename all variables with your initials appended (just as was done in assignment 1)**

I have renamed all the variables by appending them with my initials NM.



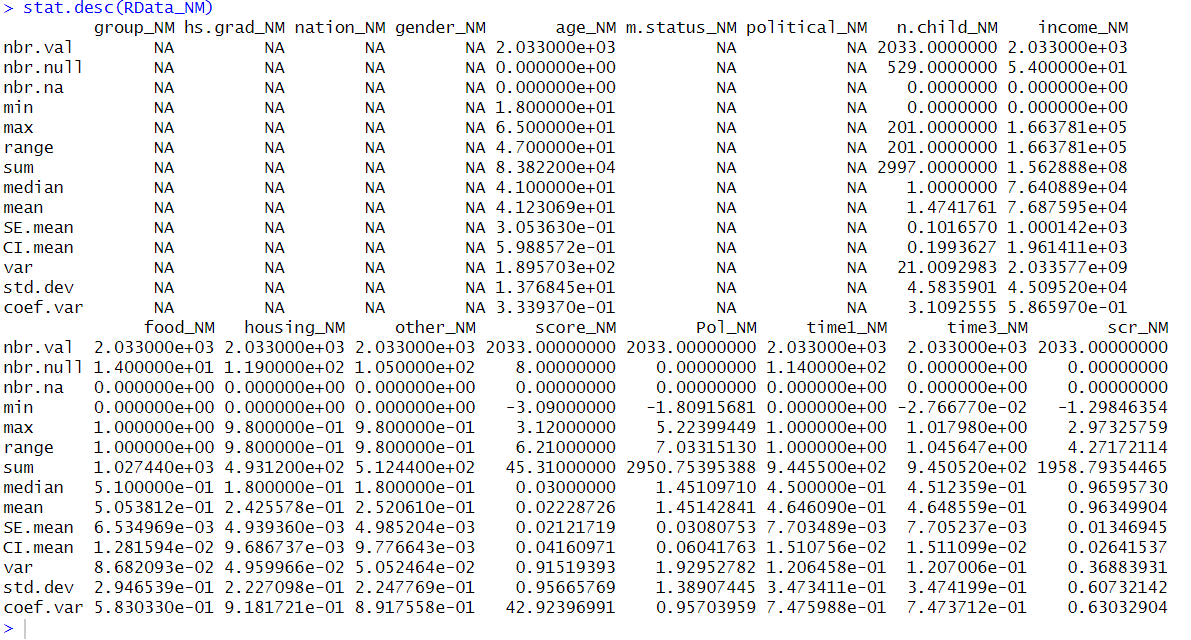
**1. Reduce Dimensionality**

**1) Apply the Missing Value Filter to remove appropriate columns of data.**

****

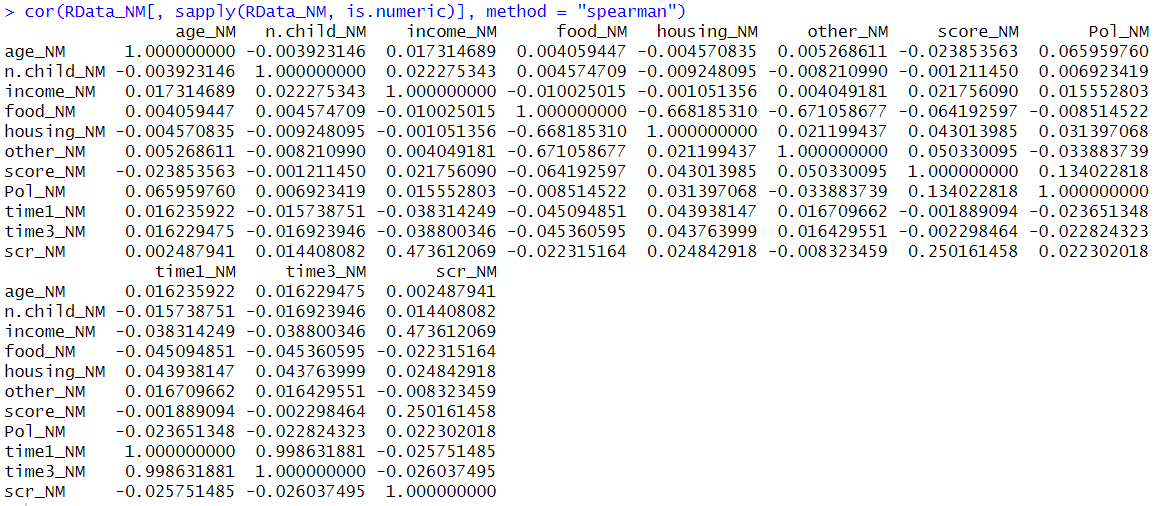
After applying the missing values filter, we can see that time2\_NM the column for the amount of Time Taken on Section-1 for the Standardized test has a lot of missing values. Since the amount of time taken on the Standardized test does not affect the Political Awareness score, we will be removing this column.

**2) Apply the Low Variance Filter to remove the appropriate columns of data.**

****

Based on this we can conclude that housing\_NM (Percentage of Income to Housing), other\_NM (Percentage of Income to Other Expenses) and food\_NM (Percentage of Income to Food) all have low variance. We will be keeping these for now until further revelations.

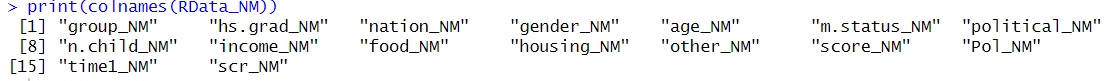
**3) Apply the High Correlation Filter to remove the appropriate columns of data.**

****

Based on the High Correlation Filter we can clearly see that time1\_NM and time3\_NM are very highly correlated, which makes sense since time1\_NM and time3\_NM are variables for time taken on the Standardized test. time1\_NM being the percentage of time taken and time3\_NM being the time taken on section 2. It makes sense that these are correlated since time2\_NM was a variable with many missing values. Hence, time1\_NM percentage is completely calculated out of time\_3. Therefore, we will be removing time3\_NM as time1\_NM and time3\_NM are very highly correlated.

Furthermore, we can see that housing\_NM, and other\_NM are also all highly negatively correlated with food\_NM. We can remove them but we are not entirely sure yet how they might affect the analysis later on. For now, we will keep them and see the significance they hold later on. Based on that we can remove them or keep them.

Therefore, we have removed the id column as that doesn’t contribute much to the analysis. Moreover, based on the missing value filter and the low variance filter we have removed the time3\_NM and the time2\_NM columns. We will be keeping the housing\_NM, other\_NM and the food\_NM variable for now since we are not sure if they can affect the analysis or not.



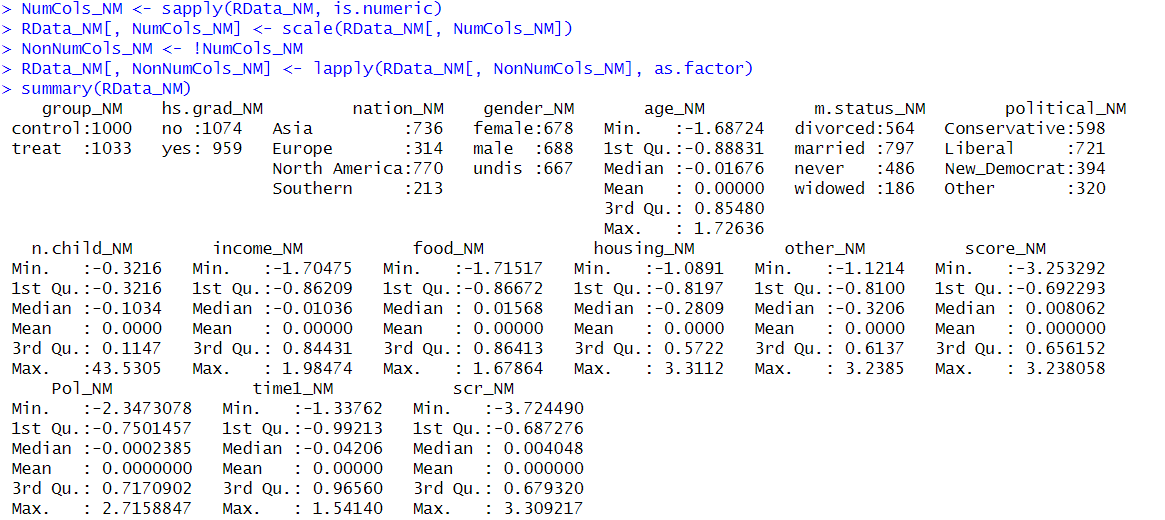
Furthermore, according to my understanding and a bit of analysis variables such as food\_NM, nation\_NM (Nationality (Region)), m.status\_NM (Marital Status), n.child\_NM (Number of Child), housing\_NM, other\_NM, time1\_NM and scr\_NM (Standardized Score Test) won’t affect the Political Awareness score by a lot but we will be keeping them for now and check their significance level later on.

**2. Data Transformation**

Here there are three possible ways which come across for me. Either transform the categorical variables to dummies (One Hot Encoding) or convert the categorical variables into factors. Another option is to do half and half based on the total number of categories in a particular variable. Since the number of “levels” affects the categories and that might subsequentially affect the models and the analysis. I will be converting all the categorical variables into factors so it’s easier to conduct.

**1) As demonstrated in class, transform any variables that are required to conduct the regression analysis (e.g. categorical variables to dummies).**

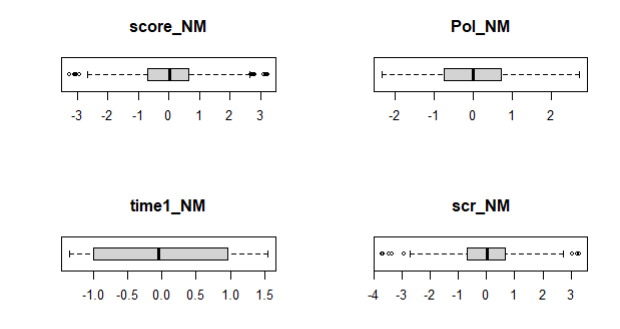
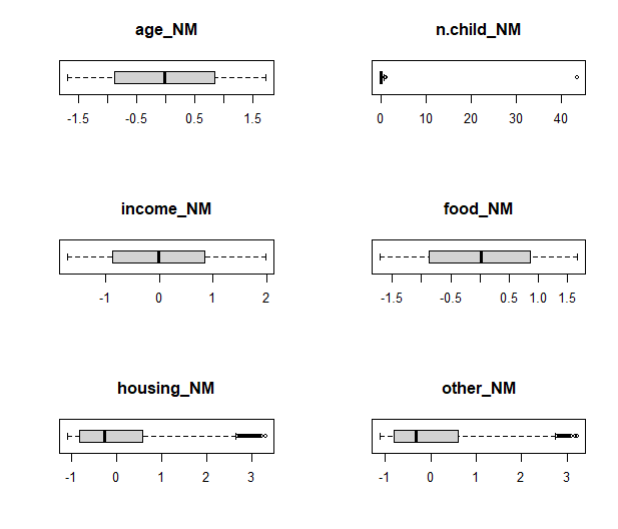
Converted all the categorical variables to factors. Standardized the data along with it as well. Standardizing the data means to format the numerical data in such a way that it is easier for us to analyse the data.



Here we can see that the mean for the numerical variables is 0 that confirms that the data has been standardized. Moreover, we have also converted the categorical variables to factors so that our job is easier and that it doesn’t interfere with the regression algorithm.

**3. Outliers.**

**1) Create boxplots of all relevant variables (i.e. numeric, non-binary) to determine outliers**

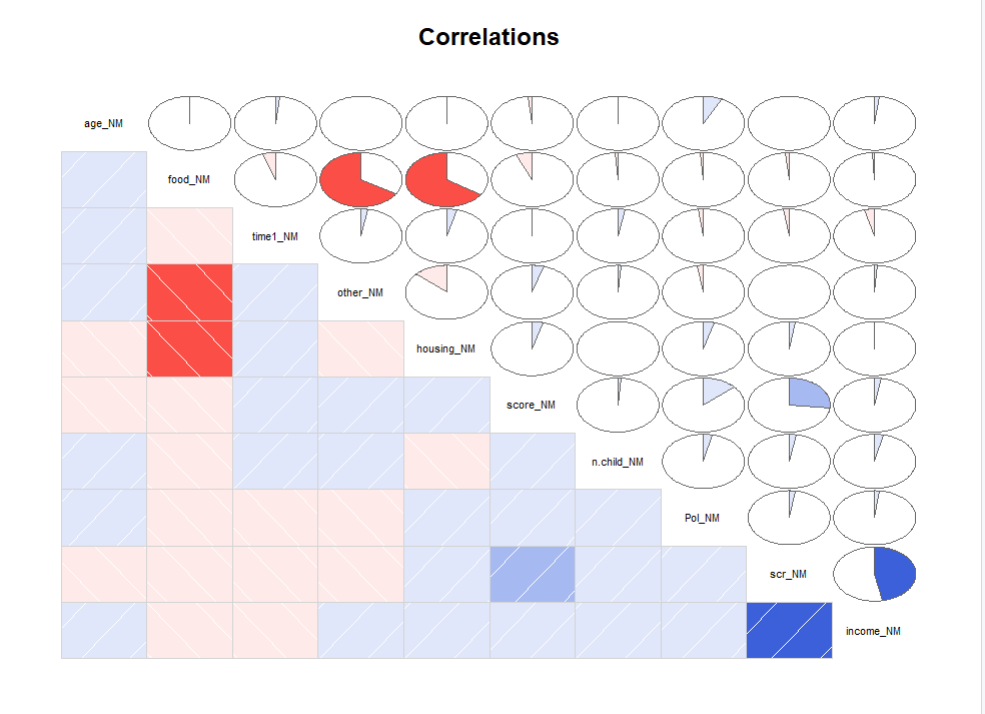
****

Here we can see that the Box Plots for all the variables are very inline. There are a few outliers here and there but they all seem reasonable and can be explained well. For example, the outliers in the housing\_NM variable are all above the upper bound and this is expected as some people might leave in more expensive houses than others.

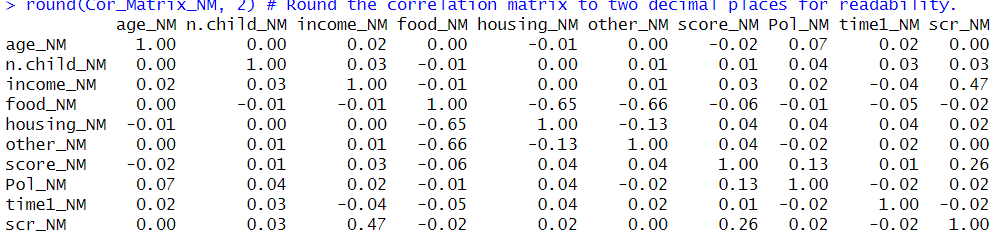
On the other hand, there are quite a few outliers in the score\_NM variable and they might affect the overall analysis. But since these outliers are very few in number, we are keeping them in as of now to see if they affect the analysis later on and remove them later if need be.

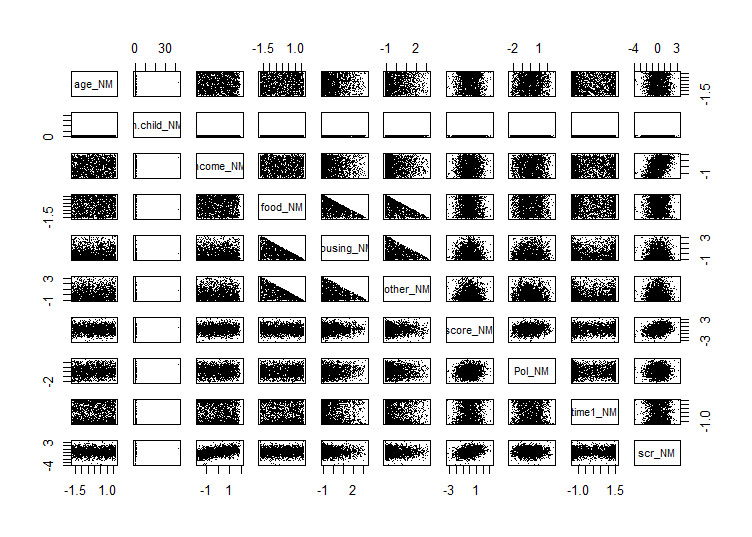
**4. Exploratory Analysis.**

**1) Correlations: Create both numeric and graphical correlations (as demonstrated in class) and comment on noteworthy correlations you observe. Are these surprising? Do they make sense?**

****

From this Pictorial representation we can clearly see which variables are correlated to what. Blue means positively correlated and red means negatively correlated. Here we can see that Housing and Food is negatively correlated. On first look here people who spend more on food tend to decrease their spending on housing. This can hint at budgeting and financial planning. This can also indicate that there are strict budge allocating and that some households might prioritize something else above the other. Moreover, since this survey is being done in Canada, seasonal effects such as winter and summer can also be an effective factor on this correlation. Similar can be said for other spending and food as well as they show a similar trend for both the variables.

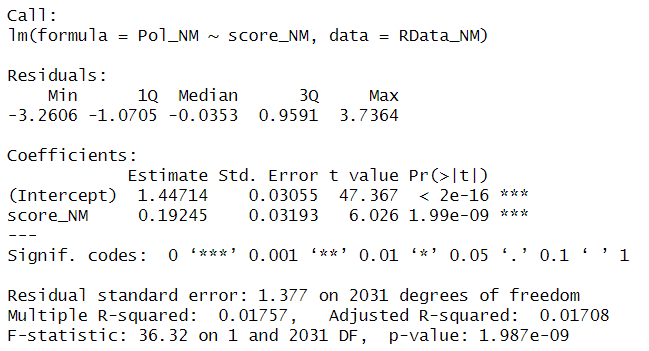


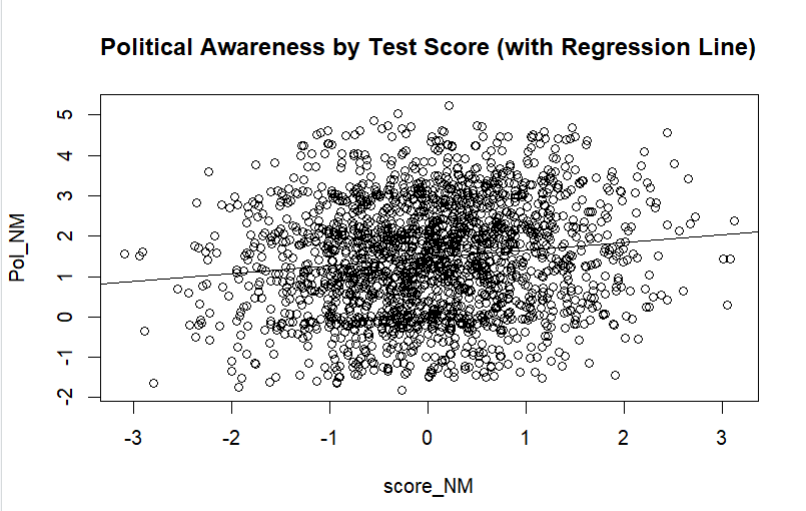
****

The findings in the correlation matrix above are supported by this correlation matrix as well where we make use of the pairs methods. We can see that income and standardized score test have a positive correlation which is to be expected as people with higher income might have more access to resources and better access to education overall. Moreover, Higher-Income families also have a social stigma of putting their children into high-class schools where education is a strong focus.

**5. Simple Linear Regression**

**1) Create a simple linear regression model using Pol as the dependent variable and score as the independent. Create a scatter plot of the two variables and overlay the regression line**

****

****

This is the Simple Linear Regression for Political Awareness across the Political Awareness Test Score. Let’s explain the important elements step by step.

Residuals are the difference between actual values and the predicted values. With the help of the residuals, we can see how good the model actually is and how well does it fit the data. Furthermore, they also indicate potential skewness or errors in the prediction. Here values indicate the range of the data such as minimum value, maximum value, the median and the 1st and 3rd quadrant. Our Median is very close to zero which is the expected value. Moreover, the first and third quartiles are fairly balanced which is a good thing.

The intercept is the value of the dependant variable when the independent variable value is zero. Here the p-value should be less than 0.05 for us which is the case which means that score is a significant predictor for Political Awareness. Coefficients indicate the direction that the observed value fall on the regression line. Here as our p-value is less than 0.05 it is significant. This indicates that there is a statistically significant impact. T-value measures the strength of the relationship between the dependant and independent variable. Our t-value is high so that means the relationship between political awareness and the political awareness score variable is statistically very significant. T-value is the test statistic for a hypothesis test that the coefficient is different from 0.

The Significant Codes indicate the level of significance. The more stars there are the better. Here the p-value of 1.99e-09 is very small smaller than 0.05. Hence, there are three stars besides it. This indicates that the relationship between these two variables is very significant at the level of 0.001. They indicate the confidence level.

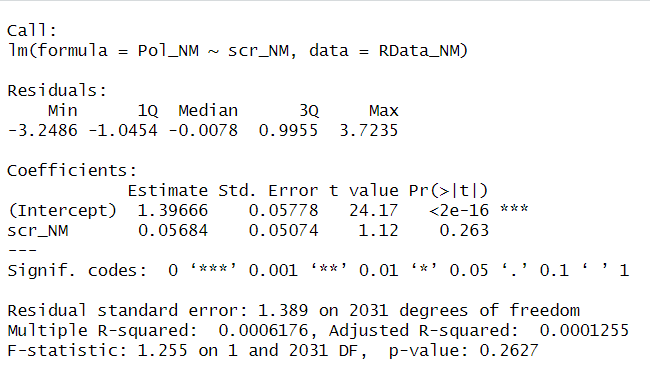
Residual Standard error is the average distance between the regression line and the observed values. The average amount the response will deviate from the line. When the value is lower this means that the model is a good fit to the data.

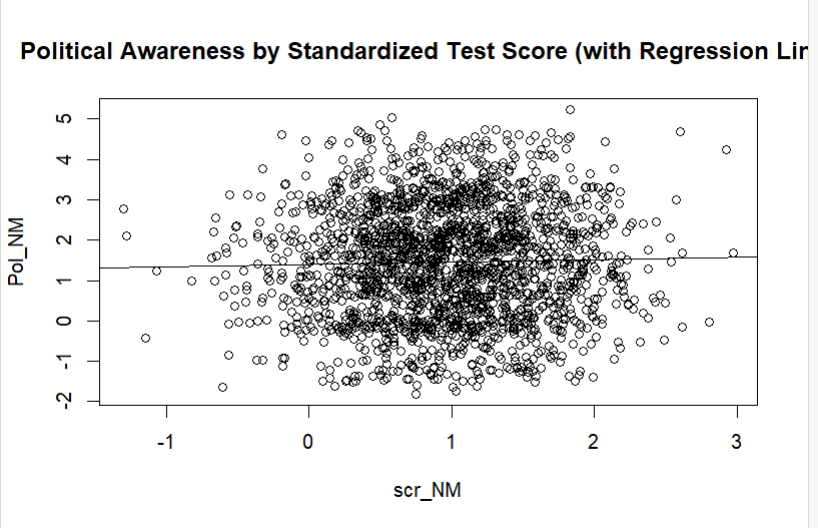
R-Squared and Adjusted R-squared are the variance in the predicted value derived from the independent variables. Adjusted R-squared adjusts the R-squared value based on the number of predictors and the sample size of the data. This is useful when comparing against various models. The value closer to 1 indicates that the model has good fit and that it is strong. Here our R-squared values are closer to 1 indicating a string model.

F-statistic test the null hypothesis that all the regression coefficients are equal to zero. It measures the overall significance of the model. The p-value assess the evidence against the model containing no predictors. A smaller p-value (< 0.05) indicates strong evidence against the null hypothesis. Suggesting that the model is statistically significant. A large F-value that the model is better than the model with 0 predictors. Which is true here for both of our values.

All of these values indicate that our model is statistically significant and that the relationship between these two variables is strong. Overall meaning that the model is strong and is a good fit for the data.

**2) Create a simple linear regression model using Pol as the dependent variable and scr as the independent. Create a scatter plot of the two variables and overlay the regression line.**

****

****

This is the Simple Regression model for Standardized test score and Political Awareness.

As we discussed and learnt earlier about the values and predictors that indicate if the model is good or not. Let’s take a look at the values provided by this model and check the fit of the model.

First of all, p-value for the scr\_NM variable is 0.263 which is greater than 0.05. This suggests that the predictor might not be significant. Moreover, the t-value here can be considered high but it is not as high as the previous model. In addition, the confidence level against the null hypothesis here is 0 in case of scr\_NM as there are no stars.

In case of the R-value, it is very close to zero indicating that the model is not a good fit and is not strong according to the adjust r-squared value. The F-statistic value is also not as high as expected and p-value is greater than 0.05.

All of these indicators show us that the model is not a good fit for that data and that there is not a strong statical significance and that the relationship between scr\_NM and Pol\_NM is rather weak significantly.

**3) Compare the models. Which model is superior? Why?**

Comparing the models it is clear the model between Political Awareness and the Political Awareness score variable is superior than the model between Political Awareness and Standardized Test Score.

To further explain the reasoning for this, the t-value for model-1 is 6.026 suggesting that it is significant and when comparing it to model-2 where the t-value is 1.12 it shows that model-2 is weaker when comparing the t-values.

Furthermore, the p-values for both the models are wildly different. For model-1 the p-value is very low way lower than 0.05 indicating that the variable has a statistically significant impact. Whereas, when we compare it to the p-value of the 2nd model which is 0.2627 which is greater than 0.05 it shows that the coefficient might not be statistically significant.

Moreover, looking at Residual standard error for 1st model is 0.01757 and for model 2 its 1.389. As we looked earlier, the model is a good fit if the residual standard error is closer to 0. Here that is apparent for model 1 and is not the case for model 2.

R-Squared value and the Adjusted R-Squared value should be closer to 1 so that we can infer the model has a good fit and that it is strong respectively. When looking at model 1 and model 2 the R-values for model 1 are all very close to 1 when compared to model 2. This supports our analogy that the 1st model is superior than model 2.

F-statistic for model 1 suggests that the model is statistically significant as the f-value is very higher than 1. When comparing it to model 2 where the f-value is 1.255 it suggests that the 1st model is much more statistically significant than model 2.

Due to all these values and many more considerations, we can tell that model 1 is superior and statistically more significant than model 2 and that we should go with the 1st model when we are predicting Political Awareness.

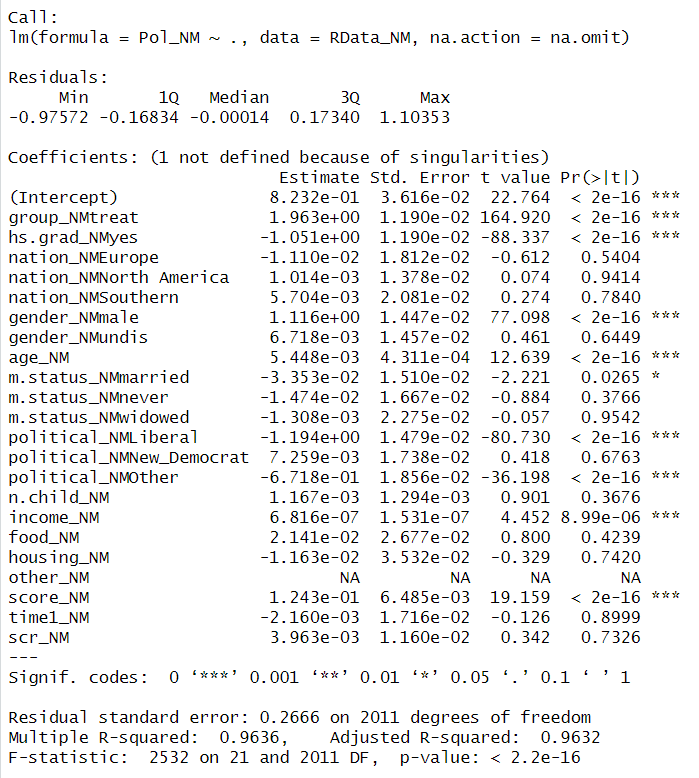
**6) Model Development – Multivariate**

**As demonstrated in class, create two models using two automatic variable selection techniques discussed in class (Full(baseline), Backward). For each model interpret and comment on the five main measures we discussed in class.**

**1. F-Stat, 2. R-Squared value, 3. Residuals, 4. Significant variables, 5. Variable Co-Efficient**

**1) Full Model**

Multi-Linear Regression is about predicting one variable with the help of more than one independent variable. The full model, as the name suggests, makes use of all the variables regardless of their significance level.



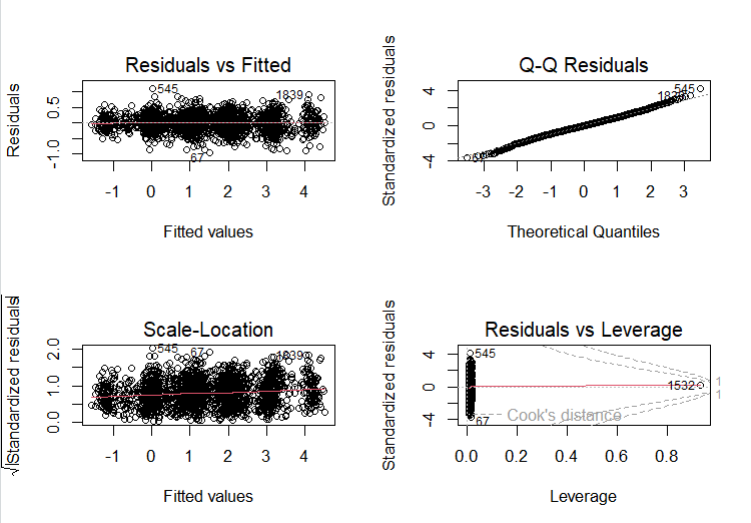
The understanding for all the components in this Multi-Linear Regression is pretty much the same as the one I explained above in Simple Linear Regression. This is the Full Model where we use all the variables to make our regression model. Here we can see what variable impacts the Political Awareness variable in what significance. From a brief glance we can see that score\_NM, group\_NMtreat, hs.grad\_NMyes, gender\_NMmale etc are some of the variables that are significantly effecting the regression analysis.

This also shows that variables like food\_NM, other\_NM, housing\_NM, n.child\_NM, scr\_NM are variables which are not of significance can be removed. But since we will be doing backwards selection, we can keep those for now.

F-statistic is a ratio which compares the model’s explained variance to the unexplained variance. If tests the overall significance of the model. Here the f-stat value is 2523 with a very low p-value of 2.2\*10-16 which indicates that the overall model is very good and has good significance.

The R-Squared value 0.9636 are variance in the predicted value which we get from the observed values. Our R-Squared value is closer to 1 which means that our model is a good fit.

Residuals are the values which showcase the difference between the predicted values and the actual values. The Residuals summary showcase that they are normally distributed. This can be seen by these plots as well. When the residuals are normally distributed it means that the model is a good fit. Residuals look symmetrical as well.



Residuals vs Fitted Plot check the homoscedasticity of residuals. The plot is ideal as the points are all scattered in a randomly around the horizontal line dictating that the variance is constant.

Q-Q plot shows us that the residuals are normally distributed.

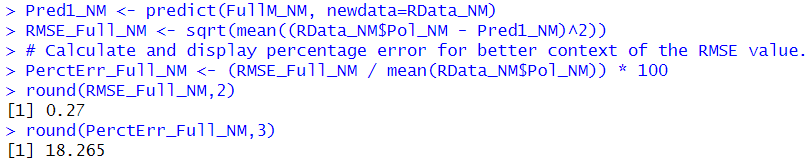
Scale Location Plot is used to check the homoscedasticity of residuals. Ideally there should equally spread points across the horizontal line.

Residuals vs Leverage Plot is used to identify special cases that might affect the regression analyses. Ideally most data points should have low average and should be within the Cook contour lines.

Significant variables here are variables that have a significant impact on our dependant variable Pol\_NM. These variables have a good relationship with Pol\_NM and are statistically significant to it. The variables with 3 stars next to them are all significant.

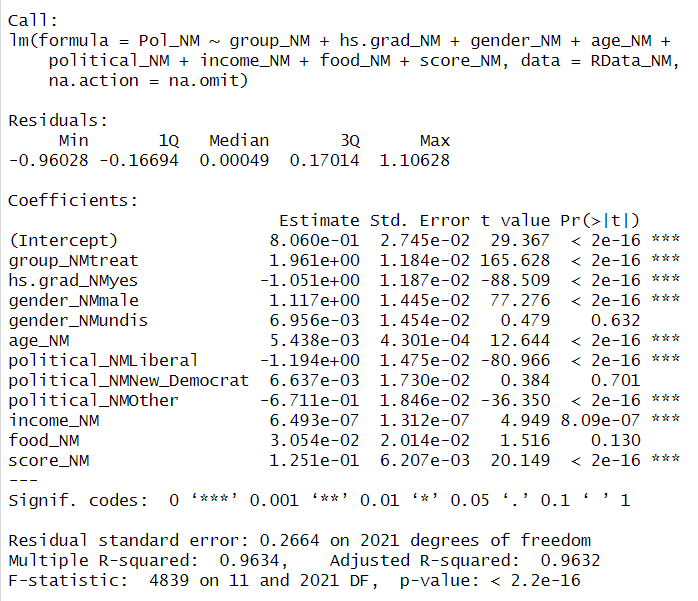
Variable Coefficient represent the changes in the response variable based on the predictor variables. We assume here that all the other variables except one predictor variable are constant. With the help of variable coefficients, we can figure out the actual value or magnitude of a relationship between each independent variable and the dependant variable.

The RMSE value or the Root Mean Squared Error value for Full Model comes out to be 0.27 and the percentage of error is 18% for the model. The Root Mean Squared Error measures the distance between the predicted value predicted by the model and the actual value that is present within the dataset.



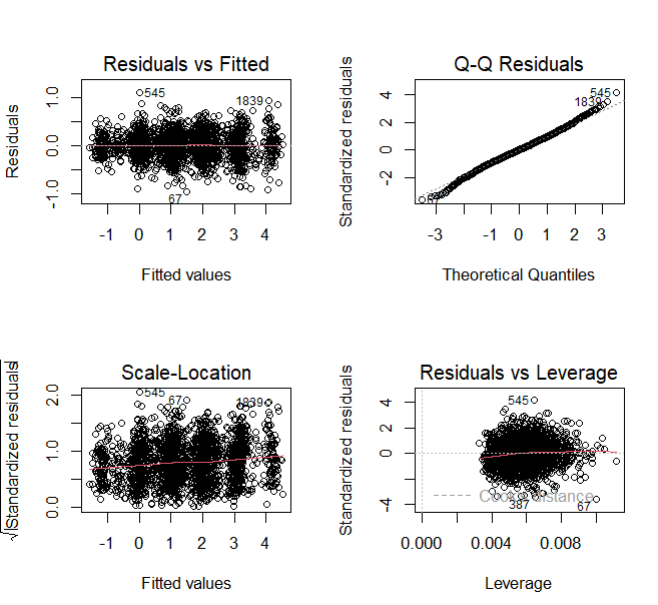
**2) Backwards Selection Model:**

The Backward Elimination/Selection model selects variables based on their significance to our dependant variable and then keeps them if they are statistically highly significant. If they are not significant than the model gets rid of the variable and moves on. This process is conducted from the back hence the name. This process is also conducted for every variable and based on their significance rating the variables are either removed or kept in the regression model and its analysis.



Once again, the components are the same as explained earlier in the SLR section. But here the model has selected fewer variables based on the backwards elimination method. The variables which are selected here are all based on the significance level and how it might affect the model and the relationship between the dependant variable and the independent variables. Based on the statistical significance the model selects the variables.

Now taking a look at the plots



We can see that the residuals are normally distributed by the q-q plot and we can also infer that the residuals vs fitted plot is ideal as the variables are having a constant variance and are all present in a dotted manner along the line. Furthermore, all the dots are within the cook’s distance which means that the abnormal points aren’t affecting the regression model and that the homoscedasticity is met by the scale location and that the variables are equally spread.

Taking a look at the F-statistic the value is 4839 with a very low p value of 2.2 \* 10-16, indicating that the overall model is significant. As the f-value is greater than one.

Now looking at the R-Squared which is 0.9632. Since it is very close to 1, we can infer that the model is strong and that it has a good fit.

We have explained the residuals with the help of the plots above. The summary of residuals indicate that the residuals are normally distributed, suggesting a good fit of the model with the data. The residual standard error is also at 0.2664 which shows that the deviation with observed and predicted variables is low.

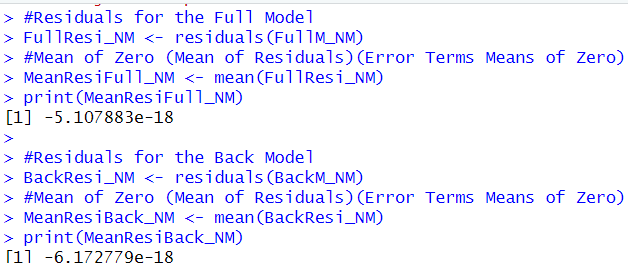
Significant variables are the variables that are impacting our dependant variable. Variables selected here are based on the backwards elimination model and are all statistically significant on the model. The variables with 3 stars are all highly significant and impact our dependant variable in a statistically significant manner.

Variable coefficient are the changes that occur within the response variable based on the independent variables as explained earlier. They are associated with t-value and the p-value. In this model all the variables bar 2 are statistically significant.

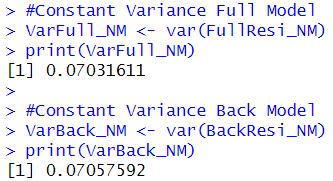
The RMSE value for the backwards selection is 0.27 which is the same as the full model. Furthermore, the percent error is also the same at 18.3% which is also very similar to the Full model.

**7. Model Evaluation – Verifying Assumptions- Multivariate**

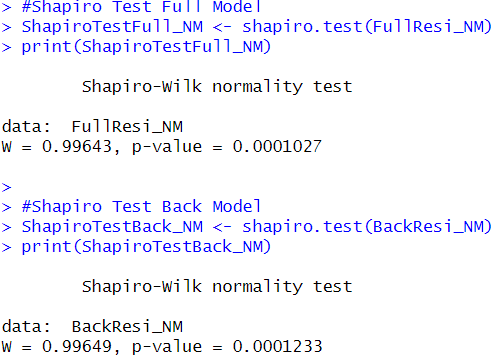
**1) For both models (as discussed and demonstrated in class) evaluate the main assumptions of regression: Error terms mean of zero, constant variance and normally distributed.**

****

Error terms means of zero or the mean of the residuals should ideally be zero in a well-fitted regression model. Here for both the models the values are very close to zero meaning that the models are unbiased and are making good predictions. That means that the models are a good fit for the data.



Constant variance of the residuals is known as homoscedasticity. Here the variance for full model and back model are relatively close. This suggests that the assumptions of constant variance are met.



Normality in Residuals is checked so that we can see if the errors are within our expectations and are normally distributed. We do this to figure out if the errors are within the bell curve and can be dealt with easily. A low p-value indicates the evidence against the null hypotheses. Here our p-vales are small but we should consider that the sample size is very large and small changes can lead to differences from normality. Therefore, its good to note that the residuals, may not follow the normal distribution based on the Shapiro Wilks test but the departures do not note that the normal distribution is not present. Moreover, based on the plots as well we can see that the residuals are normally distributed.

**8. Final Prediction – Multivariate**

**1) Based on your preceding analysis, recommend which of the models should be used.**

Based on our previous analysis we can see that both the models are pretty much similar in all the departments. When it comes to p-value or the t-testboth state that the models are both very well fitted and are strong. Furthermore, even after removing the un-significant variables with the help of backwards selection model we can see that the difference in the accuracy and the percent of error is not that different than the full model.

Moreover, R-Square value and many other factors all indicate towards our model being string and being a good fit for our data. In addition, the error percent for both our models is about 18%.

Based on all these considerations recommendations for either the full model or the back model can make sense and can be explained against.

Nevertheless, I would like to make the recommendation of backwards selection model. As both the models and their parameters are similar. Its pretty meaningless to keep the un-significant variables and increase the overall computational weight of the model which in hindsight can impact future computations.   
  
Hence, I would like to recommend the second model which is the Backwards Selection model.

References

1. Khalil .S.(04-04-2024) [PROG8435-SLR-Demo](https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/22990283/View) R File V1 [R]. <https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/22990283/View>
2. K.Samer (04-04-2024) [PROG8435-24W-MLR-Demo](https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/22990283/View) R File V1 [R]. <https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/23010921/View>
3. Khalil. S. (2024, April-06). Regression Analysis Summary, Elaborated Statistics Explanation, [PROG8435-L09-24W](https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/22988959/View), [Multiple Linear Regression Analysis Guide](https://conestoga.desire2learn.com/d2l/le/content/1001783/viewContent/22988953/View)

https://conestoga.desire2learn.com/d2l/home/1001783