**Assignment 4**

**Author:** Vidhya Venugopal, 8902970

**Course:** PROG8430

**Background**

Canadian Government wanted to understand the determination factors which are associated with the Canadian people about their political engagement and their political knowledge.

**Data Source**

Data was obtained from 2033 residents of Canada to determine the key factors associated with political engagement. A variety of variables were measured and recorded including some tests they were asked to complete. One group of respondents (“Treat”) were given additional education on political matters while the other (“Control”) were not.

1. Reduce Dimensionality
2. Apply the Missing Value Filter to remove appropriate columns of data:

Considered only the numeric values from the provided datasets. Based on the observation on NA/missing values then time2 column seems to have missing value which is been filtered to remove appropriate columns of data.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A white background with black and white clouds

Description automatically generated

1. Apply the Low Variance Filter to remove appropriate columns of data:

Based on the observation age column seems to have low co-efficient variance value. Therefore, age column has been filtered to remove appropriate columns of data.

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A number of numbers and a few digits

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

1. Apply the High Correlation Filter to remove appropriate columns of data:

Based on the observation time1 and time3 column seems to have strong positive correlation values. Therefore, time1 and time3 columns are highly correlated but only one time1 column has been filtered to remove appropriate columns of data.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

1. Data Transformation

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

The data identifying the group, graduated high school, nation, gender, marital status and political affiliation responsible for the political engagement was transformed to fourteen dummy variables.

A screenshot of a computer screen

Description automatically generated

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

From the summary statistics we conclude that the transformation of group, graduated high school, nation, gender, marital status, and political affiliation worked properly. Also, all the data looks reasonable (that is, there are no values that seem like they are necessarily wrong). No of Child seems very tightly clustered as does income but both seem to have some extreme values.

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

A screenshot of a computer code

Description automatically generated



A group of graph of income

Description automatically generated

To generate the Histogram only the numeric and non-binary variables. The histograms show reasonable distributions. But No of Child is having extreme values that will influence the outcomes.

1. Comment on any outliers you see and deal with them appropriately:

A computer code with blue text

Description automatically generated



A group of diagrams with lines and text

Description automatically generated

There seem to be outliers in No of Child. As of now, leave this column in until for now carefully to determine the effect on the next outcome model.

1. 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:

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A computer code with blue text

Description automatically generated

A math equation with blue text

Description automatically generated

A group of graphs showing the value of income

Description automatically generated

A screenshot of a computer

Description automatically generated

Income, housing & other all appear to be approximately normally distributed.

Also, No of Child really does seem to have a significant outlier.

A computer code with text

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer code

Description automatically generated

A graph of a political engagement

Description automatically generated

Treat Group seem to be more strongly positively correlated with Measure of Political Involvement and Food is negatively correlated with marital status option ‘Never’.

Other correlations to notice are:

1. Standardised Score Test and Income
2. Measure of Political Involvement and gender option ‘Male’
3. Food and Other income
4. Food and Housing
5. Control Group and Measure of Political Involvement
6. Control Group and Treat Group

Yes, it is surprising as they are multiple correlations and these correlations like Measure of Political Involvement, Score on Political Awareness test and Standardised Score Test may have confounding effects in the model.

1. Simple Linear Regression
2. 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:

A screenshot of a computer code

Description automatically generated

A screen shot of a computer screen

Description automatically generated

1. 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:

A screenshot of a computer program

Description automatically generated

A graph of blue dots

Description automatically generated

1. Compare the models. Which model is superior? Why?

A screenshot of a computer error

Description automatically generated

A screenshot of a computer error

Description automatically generated A close-up of a computer screen

Description automatically generated

A group of graphs showing different types of data

Description automatically generated

Model 1 – p-value < 0.05, therefore reject null hypothesis and accept alternative hypothesis and the variable score on political awareness is significant and F-statistics is 36.3 and R-squared – 0.0176 and adjusted R-squared – 0.0171.

Model 2 – p-value > 0.05, therefore accept null hypothesis and reject alternative hypothesis and the variable standardised score test is not significant, and F-statistics is 1.26 and R-squared – 0.000618 and adjusted R-squared – 0.000126.

After comparing the F-statistics, R-squared & adjusted R-squared is bigger in Model 1. Therefore, Model 1 is better than Model 2.

1. 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. (Your commentary should be yours, not simply copied from my example):

1. F-Stat

2. R-Squared value

3. Residuals

4. Significant variables

5. Variable Co-Efficients

**Model 1: Baseline Model - All Variables included:**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 96% of variation is explained by the model.
3. The residuals look symmetrical.
4. Seven variables (and the intercept) look significant (p-values of t-test <0.10)
5. Variable housing, control-group, higher grades, nation Europe, nation Southern, marital status Marrier, marital status Never, political affiliation Liberal, political affiliation – political-other is negatively correlated with measure of political involvement instead of positively.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

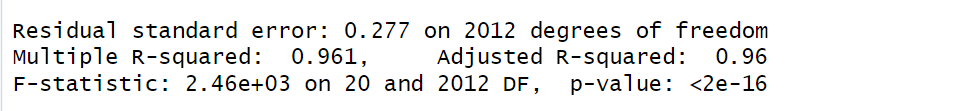
Description automatically generated

A close-up of numbers

Description automatically generated

A screenshot of a computer

Description automatically generated



**Model 2: Backward Selection**

1. Overall, the model is significant (p-value of F-Stat < 0.05)
2. 96% of variation is explained by the model.
3. The residuals look symmetrical.
4. Seven variables (and the intercept) look significant (p-values of t-test <0.10). Income, score, Control-Group, HS.Gradyes, Male, Married, Liberal & Political-Other.
5. Variable Control-Group, HS.Gradyes, Married, Liberal & Political-Other is still negatively correlated with measure of political involvement instead of positively.

A screenshot of a computer code

Description automatically generated

A screenshot of a computer program

Description automatically generated

A screenshot of a computer

Description automatically generated

A close-up of a number

Description automatically generated

1. Model Development - Multivariate
2. 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.
3. **Independence of Predictors**

The Spearman rho value for score & food (0.06), New\_Democrat & Time1 (0.07),

Married & Food (0.08) are all very low suggesting that the predictors are independent.

1. **Distribution of Error Terms**

The error terms seem to be not normally distributed as p-value < 0.05.

1. **Homoscedasticity**

Based on Residuals vs. Fitted and Scale-Location, there appears to be constant over the predicted values and not correlated with predictors. Residuals are normally distributed and Based on Residuals vs. Leverage and Cook’s Distance, there is one data point-67 which is an outlier which has high influence or leverage on the observation model.

A close-up of a computer screen

Description automatically generated A screenshot of a computer

Description automatically generated Blue text on a white background

Description automatically generated

A group of graphs showing different types of data

Description automatically generated

**Final Model, Recommendation, and Interpretation**

Based on the above, I recommend the following model developed with backward selection which is Model 2.

Political Engagement=

(0.000000728) \* Income +

(0.123215723) \* score +

(-1.964756435) \* Control-Group +

(-1.046055455) \* Graduated High School +

(1.117747350) \* Gender of Male +

(-0.029390855) \* Married status who is Married +

(-1.200991747) \* Political Affiliation with Liberal +

(-0.679221638) \* Political Affiliation with Other