**Multivariate Linear Regression (MLR) Using R Studio**

**Author Name: Mohammed Adeen Shaik**

**Background:**

A survey of 2033 residents of Canada was conducted 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”) was given additional education on political matters while the other (“Control”) was not

The task is to use multiple linear regression to determine the factors that contribute to Political Awareness (variable: Pol).

**Task 1: Dimensionality Reduction**

Removing the ‘ID’ columns since they hold no value in our analysis

1. **Missing value Filtering:**

A screenshot of a computer screen

Description automatically generated

Fig. 1

The column ‘time2’ has the highest null values of 1973 out of 2033, that’s 97.05% of the column filled with null values meaning they don’t contribute much to the data frame.

1. **Low Variance Filtering:**

A screenshot of a computer screen

Description automatically generated

Fig. 2

The column ‘housing’ has the lowest variance of 4.959966e – 02, This can also be

further confirmed by calculating IQR in Fig. 3:

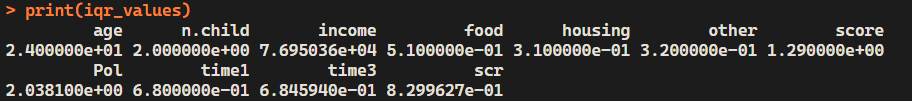


Fig. 3

1. **High-correlation filter:**

A screenshot of a computer program

Description automatically generated

Fig. 4

The columns time1 and time3 are colinear having a correlation coefficient of 0.9986

from the correlation matrix in the Fig. 4. This is because **time1 is the percent of time taken [ time2 (which we removed because of null values) + time3 ].**

The best solution here is that we take the overall time taken variable ( time1) and leave the less significant variable (time3). Hence removed ‘time3’ from the data frame.

**Task 2: Data Transformation and Cleaning**

**A graph of a graph

Description automatically generated with medium confidence**

Fig 5.1

**A graph of a graph

Description automatically generated with medium confidence**

Fig 5.2

These histograms show

* Normal distribution in Age, Income Score Pol and scr
* A right-skewed distribution for n.child, Others, and time1
* A uniform distribution of food (percentage of food expenses )

**Task 3: Outliers**

**1. Outlier Detection using Boxplots:**

**A group of diagrams with different sizes and shapes

Description automatically generated with medium confidence**

Fig. 6

1. **Age, Income, Food, Pol, and Time1 Columns:**

The boxplots for the "age," "income," "food," "Pol," and "time1" columns show no apparent outliers. The data for these variables appears to be within a reasonable range, with no values deviating significantly from the central tendency.

1. **Number of Children (n.child) Column:**

There is a significant outlier in the "n.child" column, indicating a value of 200 children. This value is exceptionally high and unrealistic, likely due to data entry error. Considering it is an extreme outlier among 2033 observations, it would be reasonable to consider removing this observation to prevent it from skewing the analysis.

1. **Other, Score, and Standardized Score (scr) Columns:**

The boxplots for the "other," "score," and "scr" columns show multiple outliers. These outliers indicate values that are either significantly higher or lower than the rest of the data distribution.

The "other" column suggests a right-skewed distribution, with several data points extending beyond the upper quartile.

**2. Dealing with Outliers:**

1. **n.child:** Remove the observation of n.child data point with value = 200. Or Correct them to value n.child = ‘2’. Either can be done.
2. **Other:** Cap the outliers to the 95th percentile value to reduce the impact of the outliers. ( replacing the abnormally high values to the value at 95th percentile)
3. **Score and scr:** Capping the extreme values to upper cap = 0.95 (95th percentile) and lower cap =0.05 (5th percentile) to reduce the impact of the outliers.

**3. Boxplot after Dealing with the Outliers:**

**A screenshot of a computer

Description automatically generated**

Fig. 7

It can be noted that there is still an abnormal value (outlier) in n.child. That is a having 6 children at home, but that is acceptable and does not impact the analysis when compared to the 200 children. So we keep the value.

**Task 4: Exploratory Analysis (Correlations)**

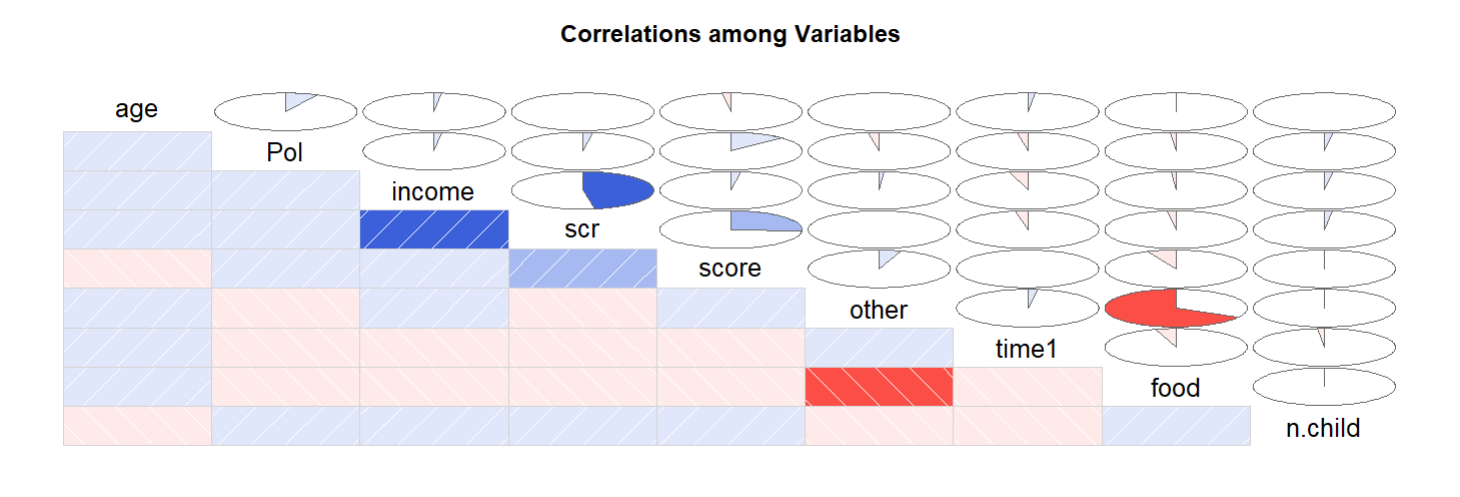


Fig. 8

Based on the correlation matrix visualization in the above figure there are some noteworthy correlations: (Orange-Red means negative and deep blue means positive)

1. **Income and SCR**: There is a strong positive correlation between income and scr. This means that as income increases, the SCR (possibly a score or rating related to income) also tends to increase. This makes sense because higher income can often lead to better scores in various financial or socioeconomic metrics.
2. **Income and Score**: There is also a noticeable positive correlation between income and score. Similar to the previous point, higher income can lead to better overall scores in many contexts (e.g., credit scores, quality of life scores, etc.).
3. **Food and Other:** The strong negative correlation between food and other suggests that individuals who spend a higher percentage of their income on food expenses tend to spend a lower percentage on other expenses, and vice versa. This makes sense because household budgets are often constrained, and spending more in one category typically means less available for others. They have a correlation coefficient of **-0.671**

These correlations are not particularly surprising when considering typical relationships between income and socioeconomic scores or ratings.

**Task 5: Simple Linear Regression (SLR)**

1. **Pol** – dependent variable, **score** - predictor / independent variable.

**SLR Model: Model1**

A computer screen shot of a number

Description automatically generated

Fig 9.1

A screen shot of a graph

Description automatically generated

Fig. 9.2

1. **Pol** – dependent variable, **scr** - predictor / independent variable.

**SLR Model: Model2**

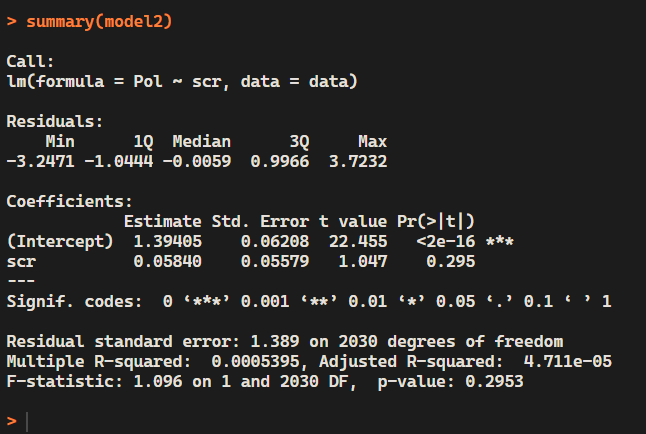
****

Fig 10.1

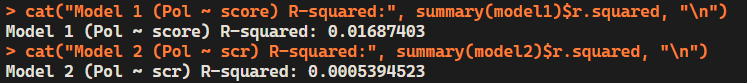
A screen shot of a graph

Description automatically generated

Fig. 10.2

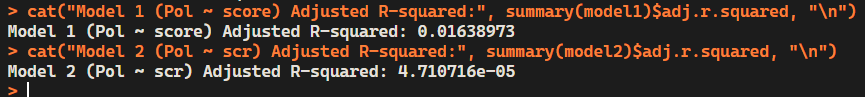
**Comparing Models:**

1. R squared comparison:



This suggests that Model 1 has a marginally better fit to the data compared to Model 2, though both models explain very little of the variance.

1. Adjusted R squared comparison:



The Adjusted R-squared for Model 1 is notably higher than that for Model 2, suggesting that even after accounting for the number of predictors, Model 1 still offers a marginally better fit.

**Conclusion:**

1. With the summary of *SLR models* in Fig 9.1 and Fig 10.1, It is evident that both models have failed the p test and cannot predict the dependent variable ‘Pol’.
2. With the *scatterplots* shown in Fig 9.2 and 10.2, It is evident that it is not possible to predict the outcome (Pol) using either of the predictor variables. But the question is not whether the model is accurate enough it is about which is better ( Score or scr )

By comparison: Model 1 outperforms Model 2, as it shows slightly better predictive accuracy with higher R-squared and Adjusted R-squared values. Despite the overall low explanatory power of both models, Model 1 / Score is the better choice for predicting the outcome of Pol.

**Task 6: Model Development – Multivariate**

**Creating a linear model using all predictors for the dependent variable ‘Pol’ named ‘full.model’**

A screenshot of a computer

Description automatically generated

**Calculate and display percentage error for a better context of the RMSE value for full model**

A black screen with orange text

Description automatically generated

**Full model’s Summary of Measures:**

**A screenshot of a computer

Description automatically generated**

**Creating a linear model using backward elimination to remove predictors that do not significantly contribute to the model named ‘ back.model’**

**A screenshot of a computer

Description automatically generated**

**Calculate and display percentage error for a better context of the RMSE value for the back model**

**A computer screen with orange text

Description automatically generated**

**Back Model’s Summary of Measures:**

**A computer screen shot of a black screen

Description automatically generated**

**Comments, Comparison and Overview on the measures of both the models:**

|  |  |  |  |
| --- | --- | --- | --- |
| Measure | Full Model | Back Model | Better  Model |
| F-stat | The F-statistic of 2520 with 21 predictors on 2010 degrees of freedom shows the model is highly significant, indicating effective variance explanation by the predictors. | With an F-statistic of 4818 for 11 predictors on 2020 degrees of freedom, the model is even more significant, demonstrating a stronger impact from the remaining predictors after backward elimination. | **Full Model** |
| R-squared value | The model explains 96.34% of the variance in the dependent variable, which indicates a strong fit. | Nearly identical to the full model at 96.33%, maintaining strong explanatory power with fewer predictors. | **Full Model** |
| Residuals | Range from -0.899 to 1.112, with a median close to zero, suggesting that the model predictions are generally accurate. | Similar range as the full model, from -0.910 to 1.117, with median very close to zero, confirming the accuracy of model predictions. | **Back Model** |
| Significant variables | Multiple variables significantly impact the dependent variable, including **grouptreat**, **hs.gradyes**, **gendermale**, and **PoliticalLiberal** | Easily It is the winner when compared to full model as this one retains only the significant predictors and drops the nonsignificant ones**.** | **Back Model** |
| Variable Co-Efficients | Coefficients for significant variables are robust, indicating strong effects | Similar coefficient values to the full model, ensuring that the strength of relationships is maintained even with fewer variables. | **Tie** |

**Task 7: Model Evaluation – Verifying Assumptions- Multivariate**

1. **Error terms mean of zero:**

Finding the mean of residuals

full.res <- residuals(full.model)

back.res <- residuals(back.model)

**Full Model:**

** = −5.87×10^(-18)**

**Back Model:**

** = 6.33×10^(-18)**

The mean of the residuals for both models is very close to zero. This is consistent with the assumption that error terms have a mean of zero. Henceforth, this indicates that the models are a good fit.

1. **Constant Variance (Residuals vs Fitted Plot)**

**Full model:**

A screen shot of a computer generated image

Description automatically generated

Fig 11.1

**Back Model:**

**A graph of dots and lines

Description automatically generated with medium confidence**

Fig 11.2

**Conclusion:**  The assumption for equal variance is met here. As both the Figures (Models) have similar variations throughout the axis.

1. **Normal Distribution**

**Q-Q Plot:** We will evaluate the residuals of the two models based on the Q-Q plot.

**Full model:**

A graph of a normal q-q plot

Description automatically generated

Fig 12.1

**Back Model:**

A graph showing a line

Description automatically generated

Fig 12.2

With both plots being equally similar to each other it is difficult to to conclude which model is better by visualization alone so we do one more normality test: **The Shapiro test**

**A screenshot of a computer program

Description automatically generated**

**Conclusion:**

In both cases, the p-values are extremely small, leading us to reject the null hypothesis. Therefore, both "full.res" and "back.res" deviate significantly from a normal distribution

**Overall Evaluation:**

1. **Error Terms Mean of Zero:**

The difference from 0 for the Full Model is lesser than back-model. It is nearer to a mean of 0 than the model.

**Conclusion: Full Model wins**

1. **Constant Variance (Homoscedasticity):**

Overall, the plot suggests that the model assumptions of linearity and constant variance are reasonably satisfied, but no clear winner can be decided.

**Conclusion:** Tie

1. **Normally Distributed Residuals:**

Both Residuals of the model do not follow a normal distribution. Hence cannot say which is the better model.

**Conclusion:** Tie

**Final Recommendation:**

Based on the evaluation and the test scores of the models, I am picking the **Full Model** as the win. Though Both have performed equally well. If I could pick both, I would have picked both because they gave also similar results and numbers with even their RMSE values where the same. Lastly, the only reason I picked this model is based on the Error Terms Mean of Zero’s result.

------------------------------------------THANK YOU---------------------------------------------------------