Economic Freedom, how do We Achieve It?

Sally Shue Yan Pang (14174354), Richard Reynard (14170539)

Table of Contents

ABSTRACT 3
INTRODUCTION 3
DATA 3
METHODS 4
EXPLORATORY ANALYSIS 5
RESULTS AND DISCUSSION 6
PRINCIPAL COMPONENT ANALYSIS (PCA) 6
LOGISTIC REGRESSION (LR) 7
MULTIPLE LINEAR REGRESSION (MLR) 10
CONCLUSION 12
REFERENCE 13
APPENDIX 14

Figure 1. Heatmap of Selected Variables 5
Figure 2. PCA Correlation Matrix 6
Figure 3. PCA Biplot 7
Figure 4. Logistic Regression Result 8
Figure 5. ROC Plot 9
Figure 6. MLR Matrix Scatterplot 10
Figure 7. MLR Linear Model Summary 11
Figure 8. Residual Histogram 11

Abstract

It is a consensus that economic freedom is desirable due to the increased productivity and prosperity induced. However, due to the differences in resource ownership by countries around the world, it is often achieved in a non-inclusive manner. In this study, we adopt statistical methods such as Principal Component Analysis, Multiple Linear Regression and Logarithmic Regression to investigate several factors that contribute to the economic freedom. We conclude that property rights and government integrity are the two most significant variables that contribute to economic freedom and therefore government may influence their national and macroeconomic policies to improve on these two areas efficiently. However, it should be noted that the dataset used is last updated in 2019, in which several macroeconomic events have happened since then, and therefore may alter the impact of each factor on economic freedom.

Introduction

Economic freedom is closely defined as the ability to possess "personal choice, voluntary exchange, freedom to compete in markets, and protection of person and property" (Lawson, 2018). As a result, economic freedom is linked to the subject of living standard and prosperity. Theoretically, this means that the existence of economic freedom will further enhance the productivity of the resources and factor input such as human capital. This is because the labours are more motivated to work, given the sense of security induced by economic freedom.

Therefore, it is evident that economic freedom is desirable. However, in the real world, there has been a discrepancy in the economic freedom enjoyed by people from different countries. Though it is dynamic in nature, the inequality between different countries is staggering. For example, Singapore is ranked first while neighbouring countries such as Malaysia and Indonesia are ranked 42 and 60 respectively in the index compiled by The Heritage Foundation (The Heritage Foundation, 2023). Therefore, this study aims to investigate which of the components contribute most significantly to the world rank. With this, it may be possible to focus on some areas to improve the economic freedom to achieve inclusive growth.

Data

The study is conducted by assigning a score between 0 to 100 on each of the variables, which are categorised under 4 umbrellas - **Rule of Law** (Property Rights, Judicial Effectiveness, Government Integrity), **Government Size** (Tax Burden, Government Spending, Fiscal Health),

Regulatory Efficiency (Business Freedom, Labor Freedom, Monetary Freedom) and **Market Openness** (Trade Freedom, Investment Freedom, Financial Freedom). The score 0 means no economic freedom while 100 means total economic freedom. Software such as R and Excel are used to compile, construct, filter and manage the massive datasets to ensure that the output produced will be able to bring value-add to the population worldwide.

The dataset is obtained from Kaggle (https://www.kaggle.com/datasets/lewisduncan93/the-economic-freedom-index) and was last updated in 2019. Further inspection of the dataset shows that there are some missing values on several variables, such as inflation rate and the GDP growth rate for the past 5 years. As a result, data cleaning and pre-processing would need to be done. The dataset consists of 6,324 sample sizes, 186 rows and 34 columns. The definitions of all the attributes will be illustrated in appendices for detailed knowledge.

Data cleaning is the process of identifying incomplete, inaccurate, and inappropriate data. It is a phase for fixing any errors and inconsistencies that are found to improve the quality (Ridzuan & Zainon, 2019). Hence, the na.omit method was used to remove from all rows with NA data.

Methods

In this case study, RStudio is used to aid with the data processing (RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.) RStudio is being used in this study to process the framework, including processing data, analysing statistics, and visualising data. It provides a collection of tools for both simple and complex statistical procedures, enabling the development of conditional expressions, iterative operations, insight directions, and built-in technicians for estimations utilising arrays and matrices. The research uses several distinct approaches to establish its goal.

There is an enormous number of machine learning methods available that assist in summarising the data included in massive data tables. But to accomplish our objective, we have chosen three alternative approaches: Principal Component Analysis (PCA), Logistic Regression (LR), and Multiple Linear Regression (MLR). Our first option is PCA, which reduces the number of dimensions in massive data sets by condensing a large number of variables into a smaller set while maintaining the majority of the large set's data. Additionally, odds ratios with several explanatory variables were obtained using LR. With the exception of the response variable being a binomial, the process is relatively similar to multiple linear regression. The final type of technique, MLR, is frequently used to clarify the relationship between a continuous dependent variable and one independent variable. All methods and codes are detailed in the appendices.

Exploratory Analysis

Exploratory analysis is a kind of data analysis that identifies broad trends in the data. In our analysis, we utilized a heatmap along with correlation and dendrogram grouping of the allocation on several important metrics that an analyst employs to spot potential relationships between variables and understand the strength of these linkages.

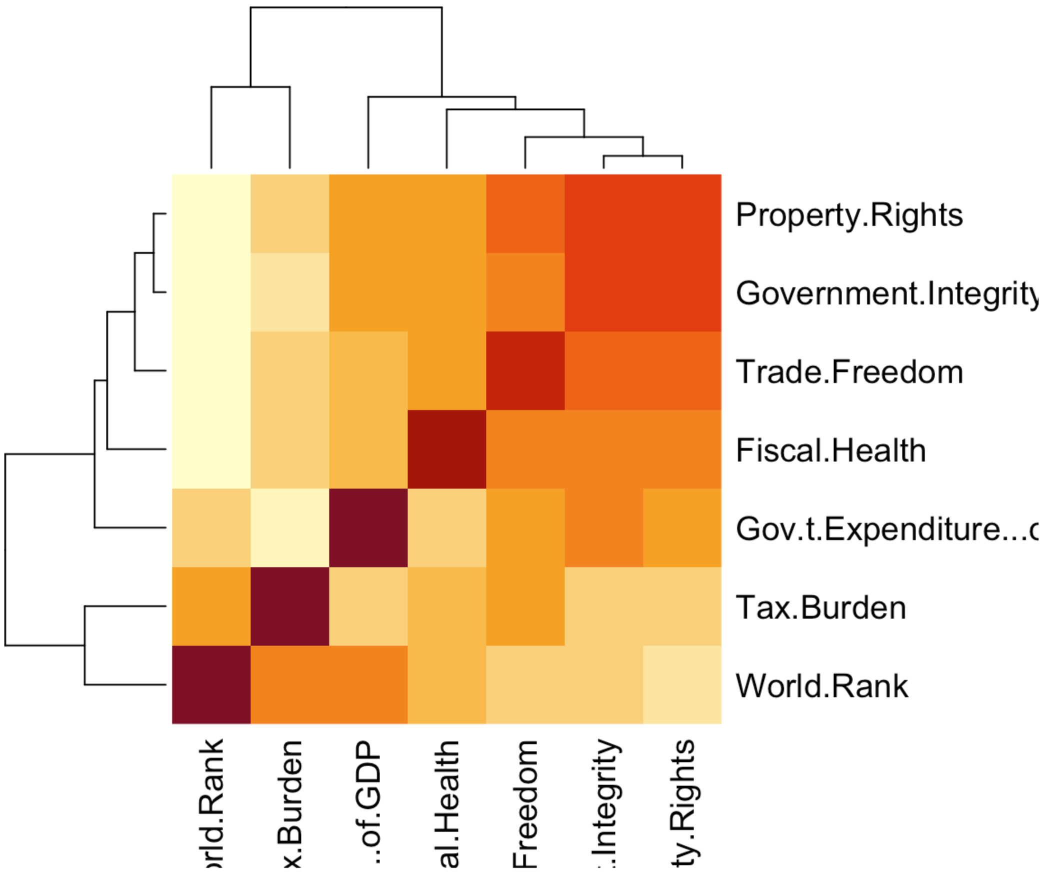

Figure 1. *Heatmap of Selected Variables*

The heatmap displays the correlation for the selected variable in this dataset. The correlation for the seven factors used in this study is shown in the second heatmap. According to the intensity of the colours, the graphs clearly reveal that World.Rank has the highest correlation with *Tax.Burden* and *Govt.Expenditure* and lower connection with *Fiscal.Health*, *Trade.Freedom*, *Government.Integrity*, and *Property.Rights*.

The dendrogram indicates that the variables *Tax.Burden* and *World.Rank* is the most comparable, but it fused later with the rest of the variables, it becomes clear that this variable is substantially distinct from the rest of the variables in terms of their relationships to and affectivity on world ranking.

Results and Discussion

A subcategory of predictive analytics called innovative insights employs historical data, analytical modelling, data mining, and machine learning to forecast future outcomes (Tucci, 2021). The conclusions of this study will be demonstrated in this section using a variety of machine learning techniques to spot trends and foresee likely outcomes.

Principal Component Analysis (PCA)

```
##                       World.Rank Property.Rights Government.Integrity
## World.Rank            1.00000000      -0.8616857           -0.7765615
## Property.Rights      -0.86168565       1.0000000            0.8529087
## Government.Integrity -0.77656150       0.8529087            1.0000000
## Trade.Freedom        -0.72590049       0.6937529            0.5740697
## Tax.Burden            0.06513135      -0.2117329           -0.2769190
## Fiscal.Health        -0.54305149       0.3125907            0.2864529
##                      Trade.Freedom  Tax.Burden Fiscal.Health
## World.Rank             -0.72590049  0.06513135   -0.54305149
## Property.Rights         0.69375291 -0.21173292    0.31259066
## Government.Integrity    0.57406969 -0.27691897    0.28645290
## Trade.Freedom           1.00000000 -0.02995695    0.27396795
## Tax.Burden             -0.02995695  1.00000000   -0.06103768
## Fiscal.Health           0.27396795 -0.06103768    1.00000000
##                      Gov.t.Expenditure...of.GDP
## World.Rank                          -0.05796776
## Property.Rights                      0.27376276
## Government.Integrity                 0.30584127
## Trade.Freedom                        0.19427719
## Tax.Burden                          -0.28721121
## Fiscal.Health                       -0.03984859
```

Figure 2. *PCA Correlation Matrix*

Correlation matrix is produced to show the correlation coefficients between the different variables. It should be noted, however, that the larger the numerical value of the World Rank, means that the country is ranked further down the list. By creating this correlation matrix, it can be observed that *Property.Rights* and the other variables have high correlation with *World.Rank* (>0.60). As *Tax.Burden* and *Fiscal.Health* have correlations of 0.0651 and 0.543, respectively, suggesting that both variables have a weaker link with *World.Rank* in comparison to the rest.

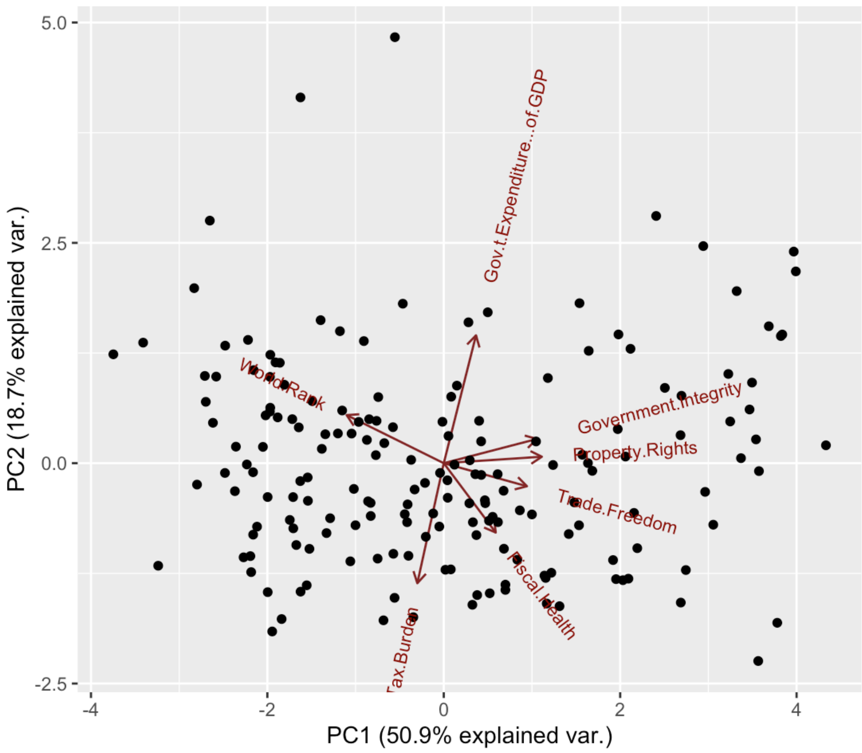

Figure 3. *PCA Biplot*

Comparing with the biplot output above, the initial loading vector shows almost identical weights for *Government.Integrity*, *Property.Rights, Trade.Freedom* and *Fiscal.Health* with much less weight for *Tax.Burden* and *Gov.t.Expenditure.* PC1 corresponds for 50.9% of total variance, whereas PC2 accounts for 18.7%, with a combined proportion of 69.6% within both PCA. This demonstrates that this PCA is only displaying 69.6% of the information being displayed. It is an approximation, but if the sum of ratios is less than 85%, it is said that the representation is not very accurate as much information is missing (Roy, 2020)

Logistic Regression (LR)

The values of one dependent variable are predicted using logistic regression using one or more independent variables. Using this relationship, the value of one of those parameters is then predicted based on the other.

```r
pred <- predict(model, newdata = test_set_scaled, type = "response")

pred_class <- ifelse(pred > 0.5, 1, 0) # Evaluate the model

# Confusion Matrix
confusion_matrix <- table(pred_class, test_set_selected$World.Rank_binary)
print(confusion_matrix)
```

```
##
## pred_class  0  1
##          0  9  0
##          1  4 24
```

```r
# Accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: 0.8918919
```

```r
# Precision, Recall, F1 Score
precision <- confusion_matrix[2, 2] / sum(confusion_matrix[, 2])
recall <- confusion_matrix[2, 2] / sum(confusion_matrix[2, ])
f1_score <- 2 * (precision * recall) / (precision + recall)

cat("Precision:", precision, "\n")
```

```
## Precision: 1
```

```r
cat("Recall:", recall, "\n")
```

```
## Recall: 0.8571429
```

```r
cat("F1 Score:", f1_score, "\n")
```

```
## F1 Score: 0.9230769
```

Figure 4. *Logistic Regression Result*

Precision is a measure of how well the model performs when the forecast is accurate. Recall gauges how well a model predicts positive classifications. The recall in this scenario is roughly 0.9259, suggesting that the model is able to properly identify 92.59% of the actual positive instances. The precision in this situation is 1, showing that all positive predictions are true positives. The F1 score is a composite metric that takes into account both recall and precision. It is the precision and recall harmonic mean. The F1 score in this instance is roughly 0.9615, demonstrating an excellent balance between recall and precision.

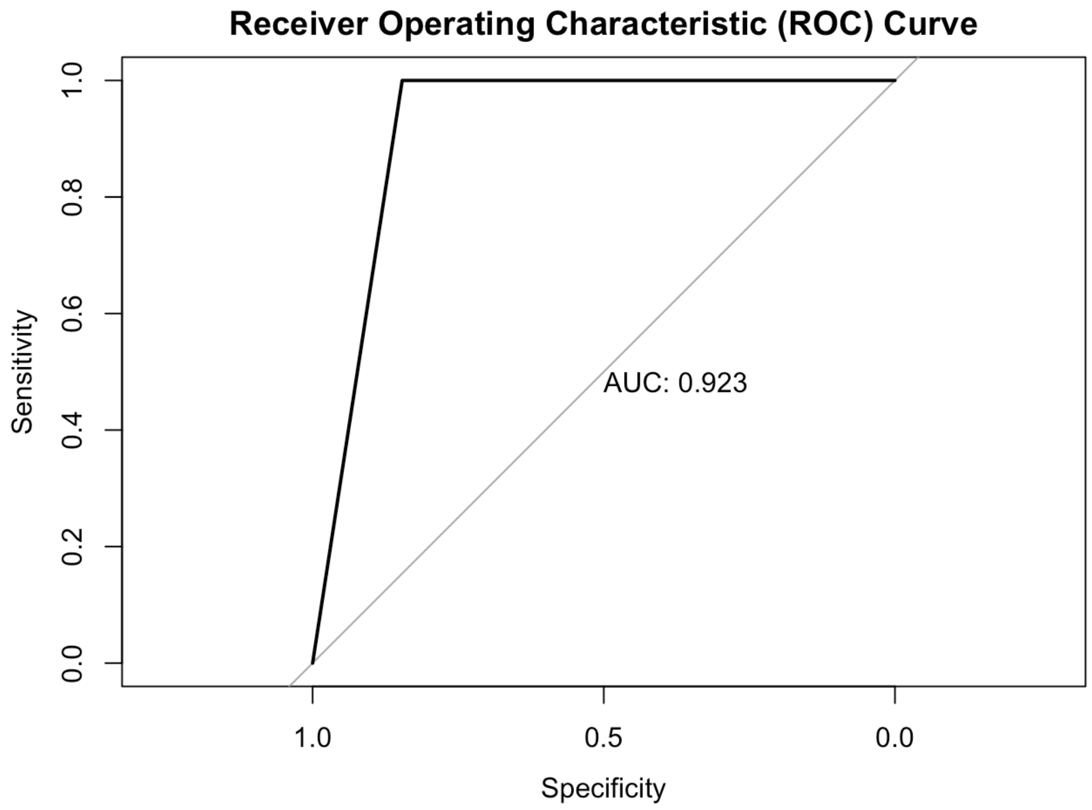

Figure 5. *ROC Plot*

A performance indicator for classification problems at various threshold levels is the Receiver Operating Curve (ROC) curve. ROC is a probability curve, and area under the curve (AUC) represents the degree or measurement of separability. It demonstrates how well the model can distinguish between classes (Chan, 2023). Based on our findings, the area under the curve (AUC) of 0.923 and a curve that is more closely positioned in the top-left corner suggest that the performance was effective.

Overall, the test set accuracy for the logistic regression model was quite high at about 89.19%. It does well in classifying the positive class, judging by its great recall, outstanding precision, and high F1 score.

Multiple Linear Regression (MLR)

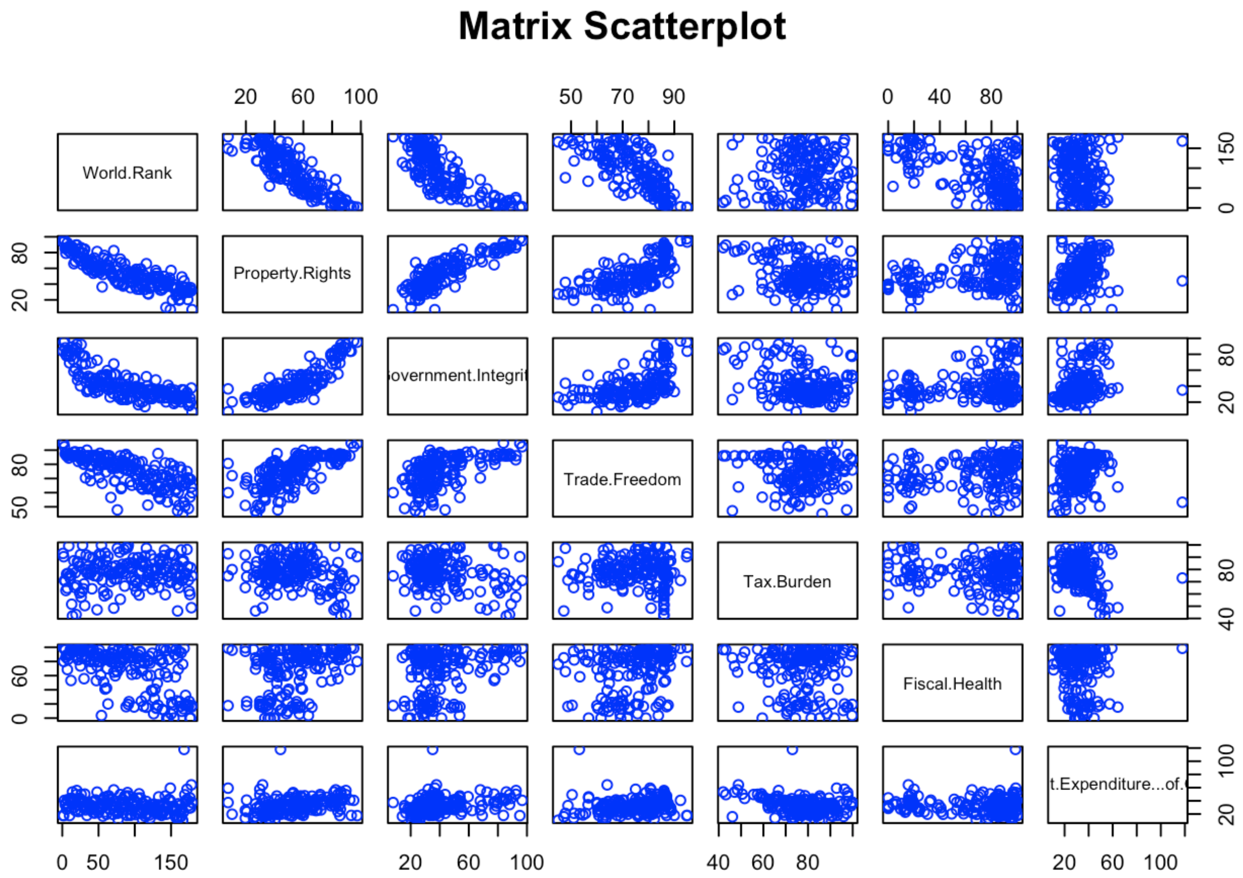

Figure 6. *MLR Matrix Scatterplot*

As mentioned before, *World.Rank* is utilised as the dependent variable to assess the importance of the independent variables. Therefore, *World.Rank* is used as the y-axis in the first row of Figure 6, with the other variables on the x-axis. It can be seen that the variables of *Property.Rights*, *Government.Integrity*, and *Trade.Freedom* indicate a perfect negative correlation, which means that these 3 factors will decrease the numerical value of *World.Rank* and therefore increasing the economic freedom; but, for *Gov.t.Expenditure...of.GDP*, the majority of the points are grouped at the bottom and widely scattered, indicating outliers.

```
##
## Call:
## lm(formula = World.Rank ~ ., data = pcaDf[3:9])
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.412 -11.949  -1.393  11.181  61.172
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               297.35258   13.93274  21.342  < 2e-16 ***
## Property.Rights            -1.36763    0.15044  -9.091 2.33e-16 ***
## Government.Integrity       -0.57178    0.13281  -4.305 2.80e-05 ***
## Trade.Freedom              -0.99712    0.17342  -5.750 4.00e-08 ***
## Tax.Burden                 -0.35090    0.12086  -2.903  0.00418 **
## Fiscal.Health              -0.44206    0.04595  -9.621  < 2e-16 ***
## Gov.t.Expenditure...of.GDP  0.64232    0.11613   5.531 1.17e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.7 on 172 degrees of freedom
## Multiple R-squared:  0.8873, Adjusted R-squared:  0.8833
## F-statistic: 225.7 on 6 and 172 DF,  p-value: < 2.2e-16
```

Figure 7. *MLR Linear Model Summary*

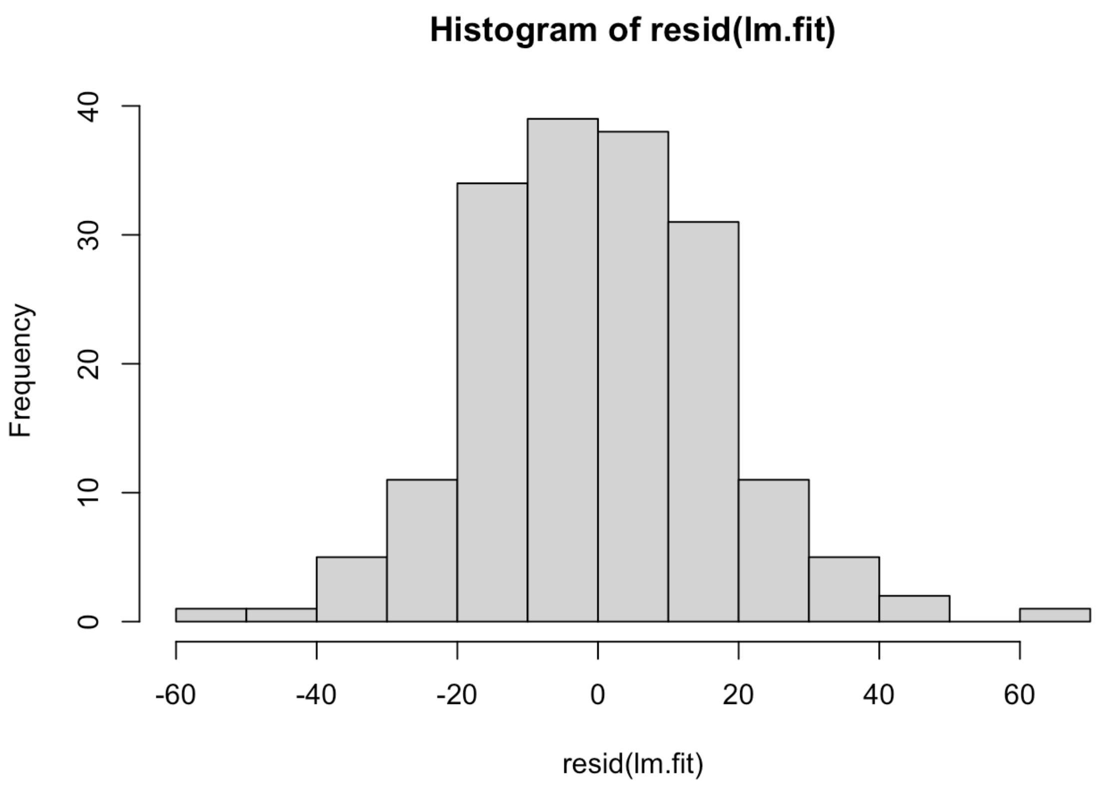

Figure 8. *Residual Histogram*

The equation of the linear model is:

$$World.Rank = -1.36763(Property.Rights) - 0.57178(Government.Integrity) - 0.99712(Trade.Freedom) - 0.35090(Tax.Burden) - 0.44206(Fiscal.Health) + 0.64232(Gov.t.Expenditure...of.GDP) + 297.35258$$

Looking at the output above, a regression model should be evaluated for a number of key factors, including R2, significance F, variable coefficients, and p-value. As shown in Figure 7, a regression model is used to determine the most frequent factor influencing the dependent variables, using the hypothesis of economic freedom world ranking as the dependent variables. Based on these values, it appears that the residuals are distributed symmetrically around zero as the min of -53.412 and 1Q of -11.949, the result shows that the residuals produced a symmetrical output. By looking at the estimated coefficients column and the Signif.codes (***), Tax.Burden does not have a significant influence on the World Rank as compared to other 4 independent variables. The codes are shown as asterisks (), with a greater number of asterisks denoting greater levels of significance.

The model's residual standard error is 17.7, which is far from zero and indicates that the model is not particularly excellent at forecasting the predicted output because the data points are more loosely dispersed around the fitted regression line (Zach, 2021).

The coefficient of determination, represented by this Multiple R-squared value (0.8873), shows the percentage of the variance in the dependent variable (World.Rank) that can be explained by the independent variables in the model. However, the adjusted R-squared (0.8833), offers a more precise indication of the model's goodness of fit. In our case, we had an accuracy rate of 88.3%. Lastly, based on the F-statistic (225.7), which is used to test the model's overall significance. It determines whether the regression model adequately accounts for the variation in the dependent variable. Strong evidence is presented against the null hypothesis that all coefficients are zero by the related p-value (2.2e-16). Based on this finding, it is able to reject the null hypothesis while accepting the alternative hypothesis.

Conclusion

In conclusion, deriving the insights from PCA, LR and MLR, we can conclude that each of the methods implemented show some similarities and differences. The difference lies in the accuracy of each method, in which PCA ranks the lowest (69.6%), followed by MLR (88.3%) and LR (89.2%). Despite the differences in the accuracy level, the methods produce the same result on the factors that contribute the most to the economic freedom, including Property Rights and Government Integrity. As a result, it may be advised that the government were to put more emphasis on these two areas to improve their economic freedom. This is not to say that they should solely allocate their resources on these areas, but it. Only serves as a tool for the government to address this issue more efficiently.

However, it should be noted that the dataset used in this report is obtained from 2019, in which some macroeconomic events have happened, including the outbreak of SARS-Cov-2 and the Russia-Ukraine War, which change the world's dynamic and policy. Therefore, there may be difference in the result obtained and the actual outcome in the real world. Consequently, further research is required to study the changes in government's behaviour in response to the events above.

Reference

Chan, C. (2023) What is a ROC Curve and How to Interpret It. *Displayr*. https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/. Accessed on 22 May 2023.

Lawson, R. A. (2008). Economic freedom. *The Concise Encyclopedia of Economics*.

Ridzuan, F., & Zainon, W. M. N. W. (2019). A review on data cleansing methods for big data. *Procedia Computer Science*, *161*, 731-738.

Roy, A. (2020). Data Preprocessing And Interpreting Results: The Heart Of Machine Learning: Part 2- PCA, Feature Selection And Result Analysis. *Towards Data Science.* https://towardsdatascience.com/data-preprocessing-and-interpreting-results-the-heart-of-machine-learning-part-2-pca-feature-92f8f6ec8c8. Accessed on 22 May 2023.

The Heritage Foundation (2023). https://www.heritage.org/index/. Accessed on 16 May 2023.

Tucci, L. (2021). What is predictive analytics? An enterprise guide. *TechTarget Business Analytics*. https://searchbusinessanalytics.techtarget.com/definition/predictive-analytics. Accessed on 22 May 2023.

Zach (2021). How to Interpret Residual Standard Error. *Statology*. https://www.statology.org/how-to-interpret-residual-standard-error/. Accessed on 22 May 2023.

Appendix

Field	VariableType	VariableMeasure	Description
CountryID	Qualitative	Explanatory	Country number
CountryName	Qualitative	Explanatory	Country name
Webname	Qualitative	Explanatory	Web name
Region	Qualitative	Explanatory	Country region
WorldRank	Quantitative	Response	Country's economic freedom ranking in the world
RegionRank	Quantitative	Explanatory	Country's economic freedom ranking in the region
2019Score	Quantitative	Explanatory	Country's economic freedom ranking in the world in 2019
PropertyRights	Quantitative	Independent	Legal ownership of resources
JudicalEffectiveness	Quantitative	Explanatory	effectiveness of fair justice system
GovernmentIntegrity	Quantitative	Independent	Government's act of serving the public
TaxBurden	Quantitative	Explanatory	Total tax revenue received as a percentage of GDP
GovtSpending	Quantitative	Explanatory	Total government spending in a fiscal year
FiscalHealth	Quantitative	Independent	The ability to continue current service levels for the forseeable future without jeopardising the financial stability of the organization or suddenly increasing the price of government
BusinessFreedom	Quantitative	Explanatory	Overall indicator of the efficiency of government regulation of business
LaborFreedom	Quantitative	Explanatory	Legal and regulatory framework of a country's labor market
MonetaryFreedom	Quantitative	Explanatory	Price stability with an assessment of price controls
TradeFreedom	Quantitative	Independent	The absence of tariff and non-tariff barriers that affect imports and exports of goods and services
InvestmentFreedom	Quantitative	Explanatory	Regulatory restrictions that are imposed on investment
FinancialFreedom	Quantitative	Explanatory	The ability to have financial cushion to afford a certain lifestyle
TariffRate	Quantitative	Explanatory	Tax imposed on imported goods and services
IncomeTaxRate	Quantitative	Explanatory	Proportion of income paid as tax to the government
CorporateTaxRate	Quantitative	Explanatory	Proportion of profit made by companies paid as tax to the government
TaxBurdenOfGDP	Quantitative	Independent	Proportion of tax to GDP
GovtExpenditureOfGDP	Quantitative	Independent	Proportion of government spending to total GDP
Country	Qualitative	Explanatory	Country
Population(mil)	Quantitative	Explanatory	Country population in millions
GDPPPP(bil)	Quantitative	Explanatory	Country GDP adjusted for PPP in billions
GDPGrowthRate	Quantitative	Explanatory	Year-on-year GDP growth rate
5YearGDPGrowthRate	Quantitative	Explanatory	GDP growth rate for the past 5 years
GDPPerCapita	Quantitative	Explanatory	GDP per person adjusted for PPP
Unemployment	Quantitative	Explanatory	Unemployment rate
Inflation	Quantitative	Explanatory	Inflation rate
FDI Inflow	Quantitative	Explanatory	Inflow of investment from abroad in millions
PublicDebtOfGDP	Quantitative	Explanatory	Amount of debt owed as a proportion of GDP