

**ST4248 Statistical Learning II Report**

**The Degrees of Freedom**

*AY18/19, Semester 2*

**Group C1**

Carel Ong Shi Ting (A0157667M)

Swan Htet Lin (A0149926N)

**Abstract**

Having freedom – personal and economic, allows individuals to make choices plainly and have a sense of ownership in life. This is central to the role in human progress and ultimately, a nation’s growth. However, the intrinsic value of freedom is often overlooked by policymakers.

This report attempts to get a broader picture of “freedom” in the world by mainly looking at these two components: Personal Freedom Index (PFI) & Economic Freedom Index (EFI). Using statistical tools, we hope to uncover their relationship to any social and economic phenomena, providing ways to optimize countries’ PFI and EFI in the process.

**Objectives**

**1** – There are many factors that determines a country’s PFI & EFI and it is not feasible for a country to focus in every area pertaining to freedom. We aim to narrow down the factors to include only those that countries should concentrate on to maximise their PFI & EFI.

**2** – Does having Economic Freedom ensures Personal Freedom, vice versa? We will be analysing on the relationship between them.

**3** – How free are we actually? We will be investigating the changes in PFI & EFI in Singapore over the years.

**Description of data**

Our dataset[[1]](#footnote-1) consists of 1458 observations, 78 predictors and 2 response variables, of various countries from the year 2008 to 2016. The predictors & response variables are quantitative values ranging from 0 to 10, with 10 representing the most positive outlook regarding the context.

Here are brief descriptions on some of the variables:

**pf\_score** (Response) – Personal Freedom Index (PFI). It represents the ability of an individual to take any plausible actions free from external restraints or entities.

**ef\_score** (Response) – Economic Freedom Index (EFI). It represents the ability of an individual to control his or her own finance and property.

**pf\_rol** (Predictor) – Measure of Rule of Law. Rule of Law encompasses mainly government integrity and judicial effectiveness and fairness. Higher value of the measure means that everyone is subjected to publicly disclosed legal codes and processes with greater equality.

**ef\_money\_currency** (Predictor) – Measure of Sound Money. It represents the freedom in having money that has a purchasing power determined by the markets and not by governments and political parties.

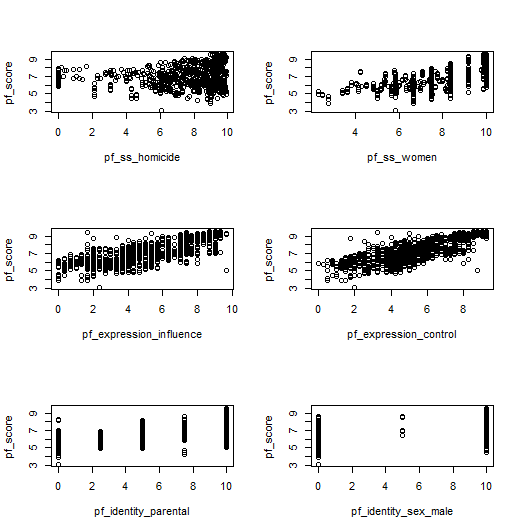
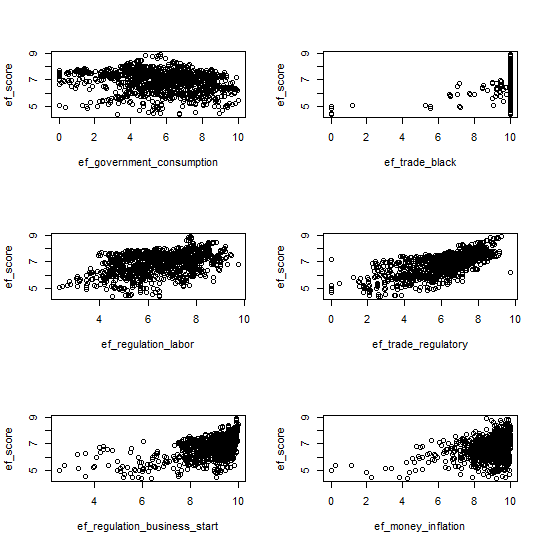
**Data Exploration**

To start with the 1st objective, we will have to build an ideal model that can sufficiently explain the response variable from its predictors. Since there are 2 response variables, we will have to build 2 models with each variable as the response. There are various approaches to fit the data and their effectiveness depend on characteristics of the data. We will have to learn some knowledge on the data before building the appropriate models.

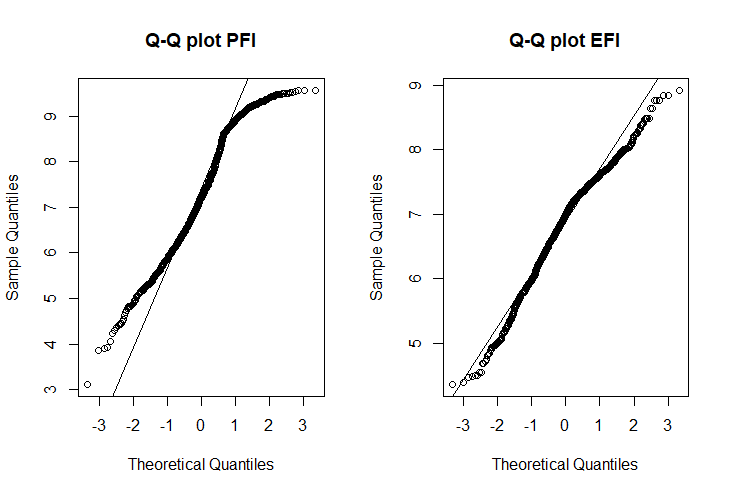
Data Pre-processing

First, we will be cleaning our data. We start by removing predictors with significant amount of missing values. This is because a predictor with many missing values might not be reliable in explaining the variation of the response variable. Next, we removed observations with missing value in any of the predictors. This results in our data to be reduced to 1184 observations and 47 predictors.

Preliminary Plots

 **Figure 1: Scatter plots of PFI/EFI against predictor**

We plotted PFI/EFI against some of the predictors to get some idea on their underlying relationship. From Figure 1, most of the scatter plots appear to be linear to a certain extent and there is no obvious non-linearity in the plots. This may tempt us to use linear regression to fit the data. However, individual plots do not entirely tell us whether a linear model with multiple predictor is suitable if the assumption of additivity of predictors do not hold. This issue can be avoided with non-regression methods.

 **Figure 2: Q-Q plots**

We performed Q-Q plot on the response variables to check on their normality. From Figure 2, it is observed that PFI deviate largely from normality while EFI conforms closely to normality. This hints the need for non-parametric approach to predict PFI.

The plots of normality and scatterplot alone did not provide substantial evidence for us to use a specific method over any other. To optimally select the best method, we will be building models of suitable approaches, afterwards which we will evaluate their performances and rank them.

Since our data contains significant number of features (47), it will not be parsimonious to build a model with all of them. Large number of predictors can result in multicollinearity which can affect the interpretability of the predictor and the accuracy of model. We will specifically focus on methods embedded with feature selection techniques.

**Models Construction**

After careful consideration, we will be constructing models with the following approaches: (1) Multiple Linear Regression, (2) Smoothing GAM, (3) Bagging, (4) Random Forest, (5) Boosting, (6) Neural Network.

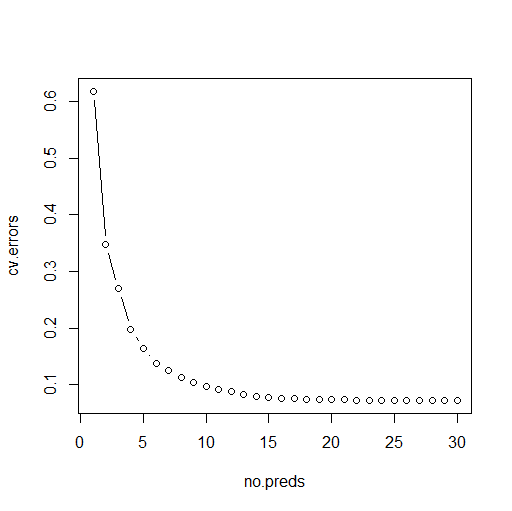
Linear Regression is the most basic method involving linearity assumption. Smoothing GAM introduces non-linearity to regression problems with back-fitting approach. (3), (4) & (5) are non-parametric ensemble methods involving trees. Lastly, neural network is a non-parametric method that is very versatile and powerful in its prediction, with the cost of interpretability.

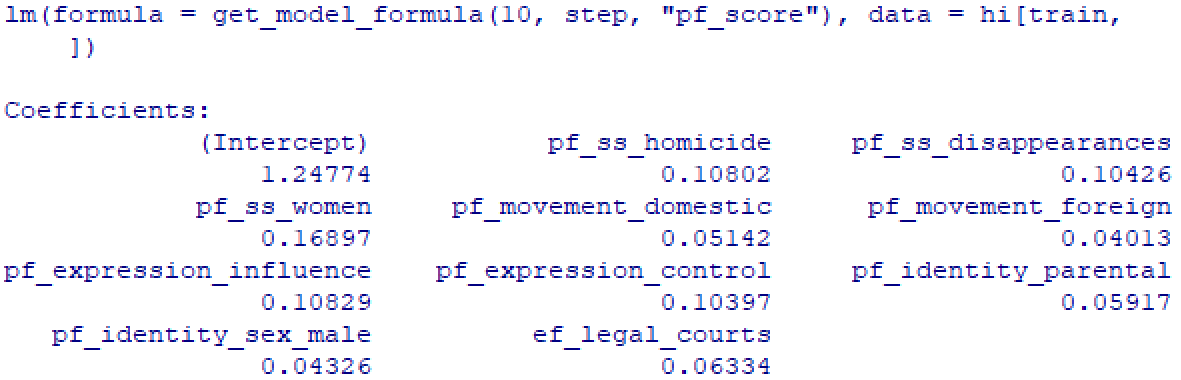
We will perform 80/20 training-test split and evaluate the models based on their mean squared errors (MSE) on the test set. To obtain their best representation, we will tune specific parameters of the models by 10-fold Cross Validation (CV).

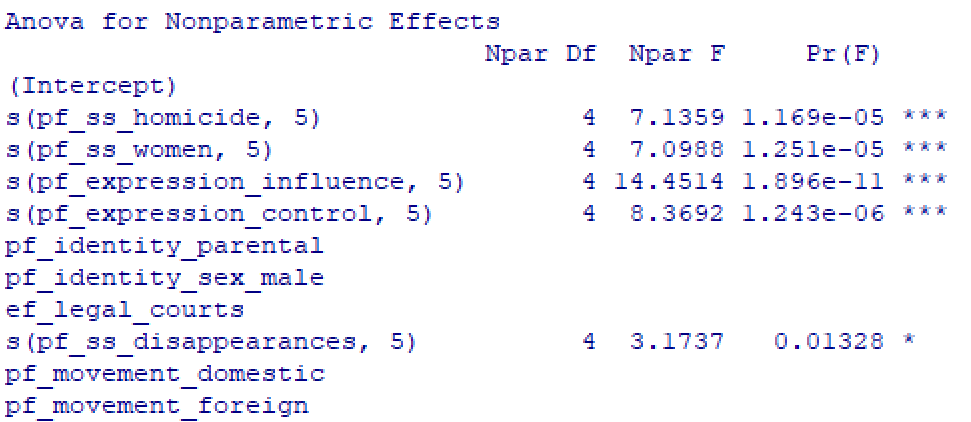
**PFI** as response variable

For multiple linear regression approach, we first performed stepwise forward selection on the predictors. Then, we tuned the number of predictors to be selected in the regression model by their cross-validation error.

To keep error low while maintaining the interpretability of the model, we will select the model with 10 chosen predictors from the forward selection procedure, shown below in Model (1). From Figure 3, increasing the number of predictors from 10 do not reduce the CV errors significantly.

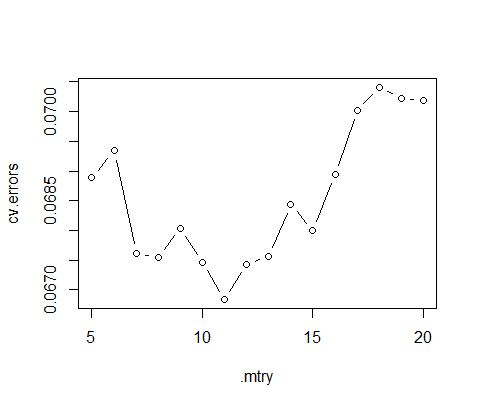
 **Model (1) Figure 3**



 **Model (2)**

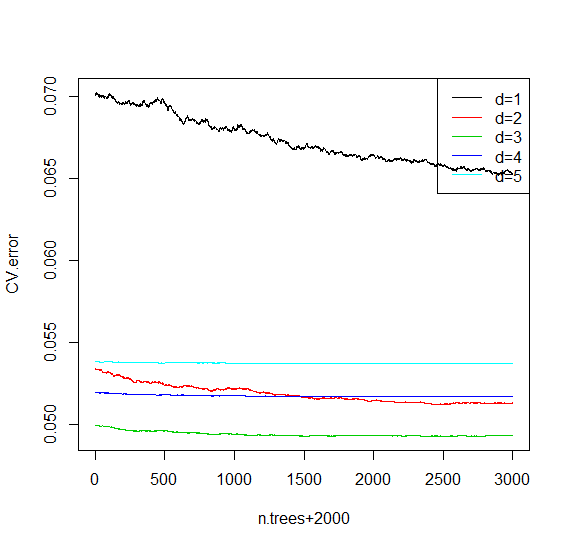
For smoothing GAM, we will extend Model (1) by introducing smoothing spline function to 5 of the predictors with degree of freedom 5 each. From Model (2), the p-values of the smoothed variables are reasonably low (<0.05), implying some non-linear relationship with the response variable.

For tree methods, we will not be performing any feature selection procedures prior to the model construction. They have their own built-in feature selection process, which deal with the problem of multicollinearity better than other methods. For bagging, we will construct with 1000 trees, which is a reasonable amount for our dataset.

 **Figure 4**

For random forest, we will construct with 1000 trees and by considering only 11 features at each split.

From Figure 4, the optimal mtry is 11 with the lowest CV errors. As mtry increases from 11, CV errors start to have an increasing trend, most likely due to increasing variance of the model. Therefore, we can expect random forest to outperform bagging, which considers all the features available at each split.

 **Figure 5**

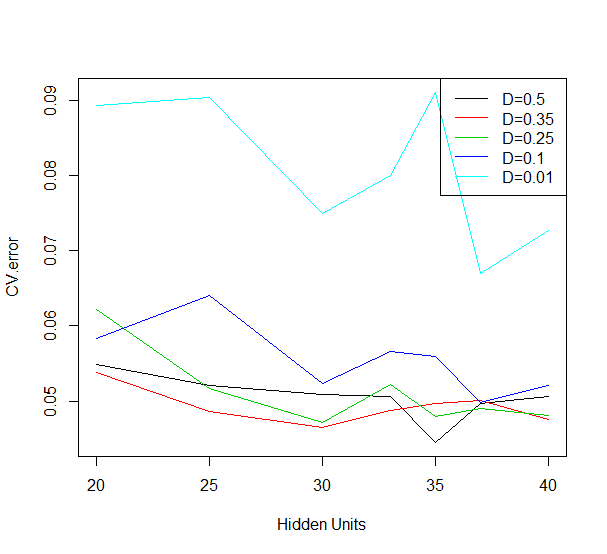
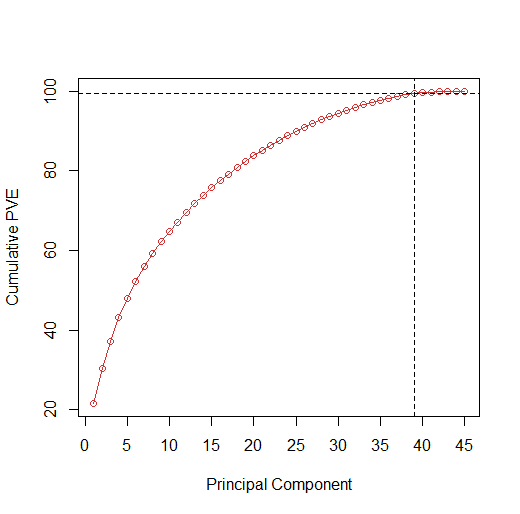
For boosting approach, we will construct with 3000 trees, learning rate (lambda) = 0.1 and interaction depth (d) =3.

Generally, a larger lambda will require lesser trees while a smaller lambda will require more trees to produce an optimal model. Hence, we tuned only number of trees and interaction depth, while fixing lambda=0.1.

From Figure 5, model of d=3 clearly outperforms other models given any tree size, as seen from its lower CV errors.

Note: Figure 5 is zoomed into 2000 – 5000 trees range.

For neural networks approach, we first performed Principal Component Analysis (PCA) to reduce the dimensions of the data. We selected the set of principal components that sum up to at least 99.5% of variance explained, which is a reasonable variation. This amounts to 39 principal components as shown by the cumulative variance plot in Figure 6. We will be using these 39 principal components as input to the neural network model with 1 hidden layer.

 **Figure 6** **Figure 7**

We tuned the number of hidden units (M) and weight decay (D) for a range of values. From Figure 7, parameters are optimally tuned at M=35 & D=0.5, which incurs the lowest CV error.

|  |  |
| --- | --- |
|  | Test MSE |
| Linear Model (p=10) | 0.0984 |
| Smoothing GAM (p=10) | 0.0942 |
| Bagging (trees=1000) | 0.0784 |
| Random Forest (mtry=11) | 0.0642 |
| Boosting (trees=3000, lambda=0.1, d=3) | 0.0450 |
| Neural Nets (p=39, M=35, D=0.5) | **0.0411** |

**Evaluation of Models**: **Table 1**

From Table 1, Neural Networks incurs the lowest test MSE, followed closely by the Boosted model. These 2 models outperform the rest by a respectable margin, while linear model performs the worst. Surprisingly, smoothing GAM did only slightly better than the linear model, implying smoothing was not truly necessary. Also, ensemble tree methods did significantly better than the regression methods.

|  |  |
| --- | --- |
|  | Test MSE |
| Linear Model (p=10) | 0.0760 |
| Smoothing GAM (p=10) | 0.0682 |
| Bagging (trees=1000) | 0.0347 |
| Random Forest (mtry=10) | 0.0278 |
| Boosting (trees=3000, lambda=0.1, d=3) | **0.0181** |
| Neural Nets (p=39, M=35, D=0.5) | 0.0223 |

**EFI** as response variable **Table 2**

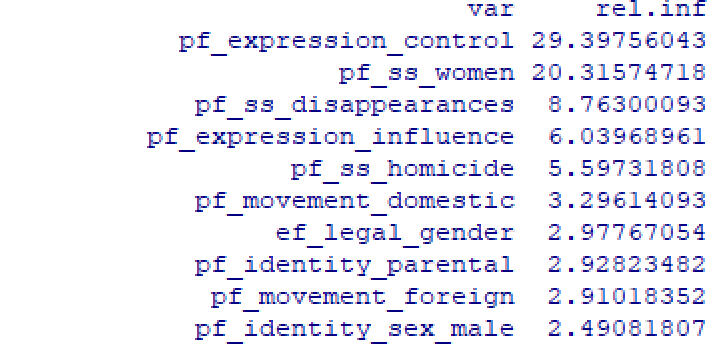
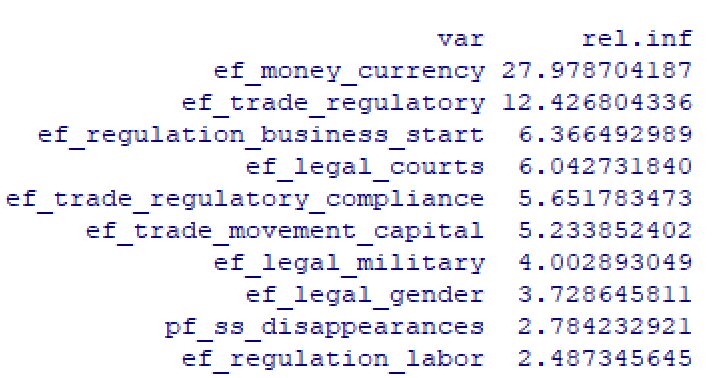
The same procedures above are performed for EFI as the responseand similar results are produced. In this case, the Boosted model comes out top, and is now followed closely by neural networks. This clearly underlines the No Free Lunch Theorem where there is no one best method in every problem. Also, the models did a better job in predicting EFI than PFI, as seen by their lower test MSE in Table 2 that that of in Table 1.

In conclusion, considering the high dimensionality of our data, this is where more complex and robust methods really shine. In contrast, a basic method such as linear model simply does not have the flexibility to learn the data as well. Performance-wise from the results, we can use neural network and boosting approach interchangeably for prediction purposes.

Beside the performance, it is also important to be able to interpret the results of the model well. As powerful and relatable neural networks is, it suffers greatly in explaining why and how it works. Our report requires identifying of important predictors which neural network cannot provide conveniently. As a result, we will be using the results of the Boosted models for further analysis.

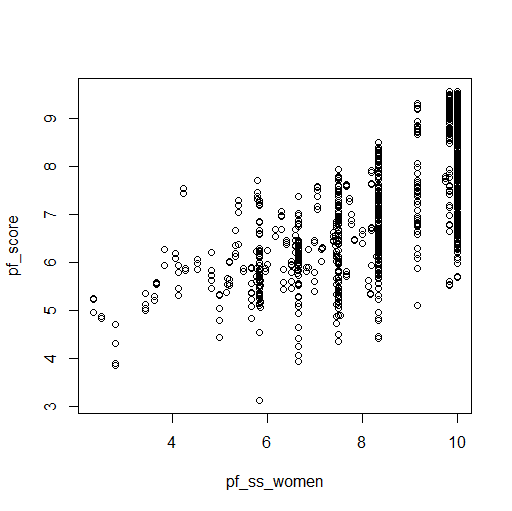
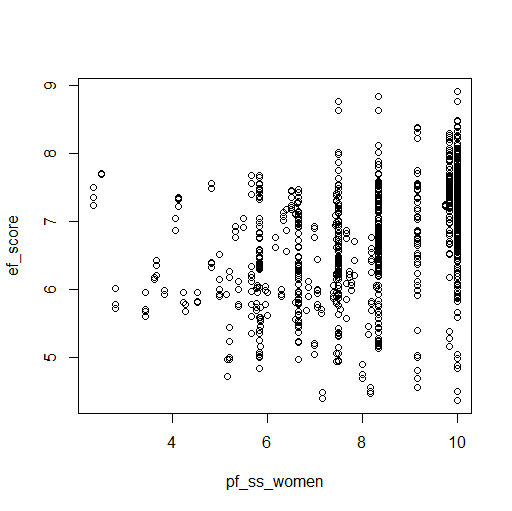
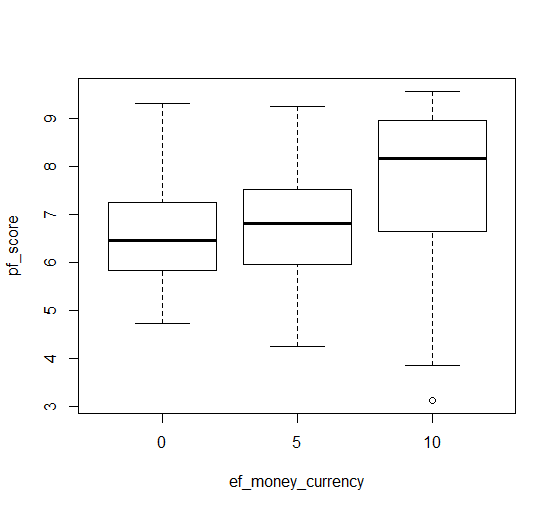
**Variable Importance**

We will be analysing on the behaviour of certain influential predictors from the boosted models.

**Table 3: Variable Importance (EFI, Top 10) Table 4: Variable Importance (PFI, Top 10)**

From Table 3, ef\_money\_currency is the most important variable in predicting EFI, followed by ef\_trade\_regulatory. They have relative influence value that are significantly greater than the rest of the predictors. Likewise, from Table 4, pf\_expression\_control is the most important variable in predicting PFI, followed by pf\_ss\_women.

Interestingly, only 2 out of 47 predictors, appeared in both top 10 variable importance list in predicting PFI/ EFI. They are pf\_ss\_disappearances & ef\_legal\_gender. This hints that most variables that are crucial to development of either personal freedom or economic freedom, might not be crucial to that of the other. We proceed to investigate this notion.

 **Figure 8** **Figure 9 Figure 10**

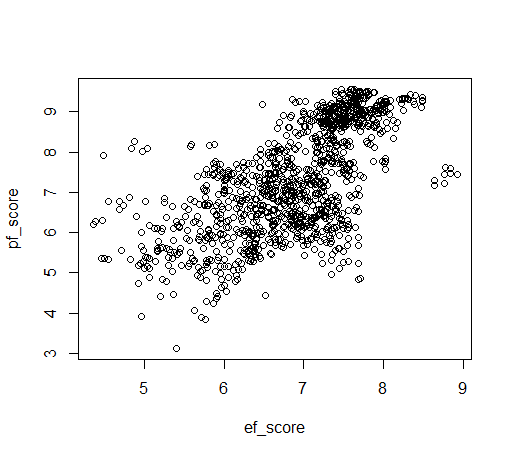
From Figure 8, pf\_ss\_women is shown to have a clear positive relationship with PFI. This is consistent with the claim that pf\_ss\_women is the 2nd most important variable to PFI.

In contrast to Figure 9, there is a weaker positive relationship between pf\_ss\_women and EFI. There are points where achieving a significantly high score in pf\_ss\_women result in significantly low score in EFI. Furthermore, as pf\_ss\_women increases, the spread in EFI also increases, implying some unreliability of pf\_ss\_women as a predictor to EFI. This is also the case for unreliability of ef\_money\_crrency, the most important variable to EFI, as a predictor to PFI shown in Figure 10.

These analyses showed that an important variable alone might not be enough to solve problems pertaining to different aspects of freedom: Personal, Economic, etc. Perhaps, a viable option for a country to increase ‘freedom’ for the residents is to shift their focus on improving combination of scores of top few predictors presented in both Table 3 and Table 4.

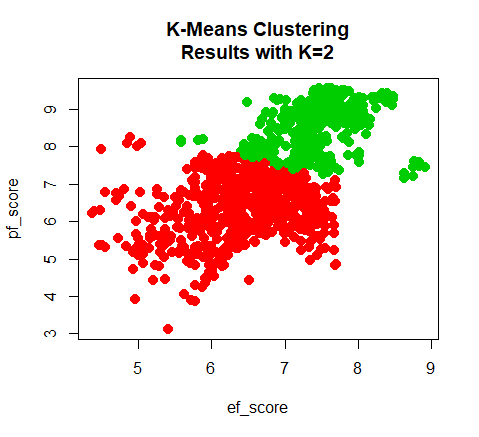
Referring to No Free Lunch theorem, just like there is no single best model that excels in everything, there is also no single best variable that excels in solving any freedom issues. This brings up the concern on the balance between Personal Freedom and Economic Freedom in countries. Does ensuring Economic Freedom ensure Personal Freedom, vice versa? We proceed to find out.

**Relationship between EFI & PFI**

 **Figure 11**

On first sight, Figure 11 tells us that there is a strong positive relationship between PFI & EFI. With the exceptions of few extreme points, it can be even observed that with every unit increase in EFI, there is greater increase in PFI. This is promising to be a positive news and might be consistent with the claim that Economic Freedom ensures Personal Freedom.

However, this result is not entirely convincing given just a scatter plot. We further performed cluster analysis to look into grouping behaviour.

 **Figure 12**

|  |  |  |  |
| --- | --- | --- | --- |
|  | # of Observations | Mean (PFI) | Mean (EFI) |
| Red Cluster | 689 | 6.345292 | 6.426226 |
| Green Cluster | 495 | 8.611209 | 7.465899 |
| Overall | 1184 | 7.292614 | 6.860887 |

**Table 5**

We performed K-means clustering (K=2) and Figure 12 shows the grouping with the most optimal local minimum out of 50 initial cluster assignments. From the green cluster, there is a moderately positive relationship between EFI and PFI. The cluster is mainly occupied by points with high EFI & PFI, where the points are closely packed around the center.

In contrast, red cluster shows that there is no obvious relationship between EFI and PFI. The cluster is mainly occupied by points with greater range of EFI and PFI from low to moderately high score. The points are rather loosely packed around the center. From Table 5, red cluster contain majority of the points (~60%) but has a significantly lower mean of PFI and EFI than that of green cluster respectively.

Based on the above information, we can deduce that points in red cluster represents majority of under-developed and developing countries while points in green cluster represents majority of developed nations.

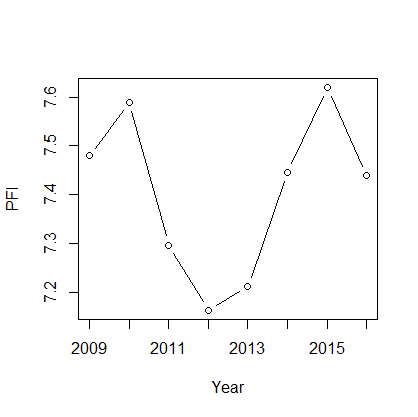
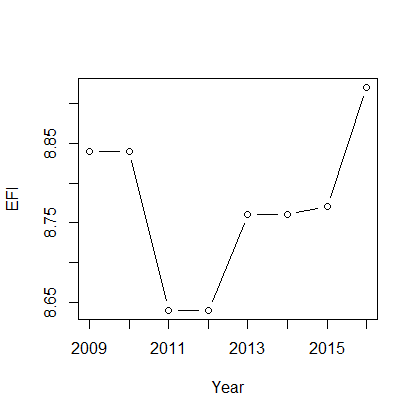
Most countries in green cluster tend to enjoy exceptional amount (>7.5) of Economic and Personal Freedom simultaneously. A possible reason is that individuals, who experience higher Economic/Personal Freedom, tend to make more rational and calculated decisions. These eventually lead to better economy and governing of a country, which in turns generate more freedom in every aspect. Afterall, freedom begets freedom.

On the other hand, it is worth noting that some points in red cluster with moderately high EFI or PFI suffer greatly in the other. Perhaps, there are trade-offs between improving Economic Freedom and Personal Freedom in some developing countries, hindering them to maximise both at once. Perhaps, countries just do not have the ability to improve PFI/EFI possibly due to their geographical location (Lack natural resources, Nearby Warzone), history and culture, governing regime, etc.

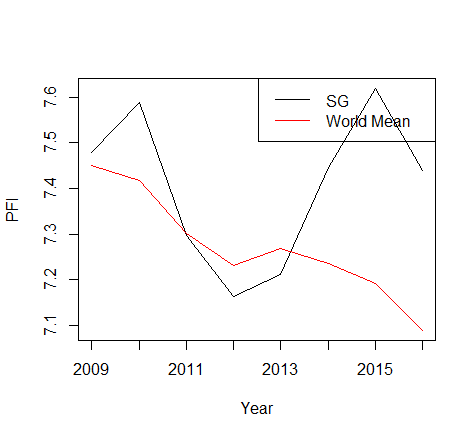
All in all, the cluster analysis shows that we need to be careful in making generalisations as correlation presented in Figure 11 does not imply causation. Achieving high Economic Freedom does not necessarily lead to high Personal Freedom, vice versa. This depends on the various circumstances of a country.

**Freedom at Home**

For our last objective, we will briefly discuss the trend of PFI & EFI specifically in Singapore and the implications that can be made.

  
 **Figure 13 Figure 14**

From Figure 13 and Figure 14, Singapore’s Freedom indexes have been on a general increasing trend over the years, particularly EFI. Economic outlook in Singapore tend to be strong and this is not surprising considering its world class reputation as commodity trading and business hub. While consistently ranking at 2nd place in EFI, only behind Hong Kong, it comes in at a staggering 60th place on average in PFI, out of 150 participating countries. ­­

 **Figure 15**

From Figure 15, it can be observed that Singapore’s PFI is only slightly above the global mean in the recent years. There was a period (2011-2013) where the PFI even fall below the mean.

From these findings, Personal Freedom in Singapore seems to be rather scarce given that Singapore is a developed nation. While this may not be a cause for concern yet, there should be some room for improvements. Living here, we may not notice these subtle changes or may be just satisfied with current state, but we should not underestimate the intrinsic value of freedom.

According to the Cato Institute, higher freedom promotes participation and collaboration and that freedom is around 54 times more effective than democracy!

**Limitations**

Main difficulty we faced was on data cleaning where we had to decide between having less observations or less predictors. Initially, there were significant number of predictors (78) so we opted to remove more of predictors with large incomplete data, preserving more on observations. Furthermore, our dataset only comprises of data from year 2008 to 2018. For future studies, we can try to obtain more data for years before 2008 in order to get a better understanding of the overall trend of countries’ PFI, and hence be able build a better model.

Besides Economic Freedom and Personal Freedom, Social Freedom is also another important component of Freedom, which we left out in our report. This means that even with high Economic & Personal Freedom, it may still not be well enough. Given more time, we would also like to explore more on the relationship between developed countries and freedom. Particularly, estimating the fine line between being a developed country and a developing country and the level of freedom expected to be developed.

In conclusion, we hope that this report will give a constructive and transformed view of Freedom, incorporated with the methodologies learnt in ISLR.

1. https://www.kaggle.com/gsutters/the-human-freedom-index/version/2#hfi\_cc\_2018.csv [↑](#footnote-ref-1)