**Executive Summary**

**Findings**

**Statistical methodology**

**Fields of study:**

**Response Variable:**

**Happiness score:**

Every country is placed on a ‘Happiness ladder’, the higher up the ladder that the country is, the more happy the population. The score is calculated as an average rating between 1 and 10 taken from 1000 people.

**Predictor Variables:**

**LGDP:**

The natural logarithmic value of Gross Domestic Product.

**Support**

The proportion of ‘Yes’ responses to the question: “If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?”

**HLE**

The Healthy Life Expectancy, life expectancy discounting freak accidents or chronic illness.

**Freedom**

The proportion of ‘Yes’ responses to the question: “Are you satisfied with your freedom to choose what you do with your life?”

**Corruption**

The proportion of ‘Yes’ responses to the questions: “Is corruption widespread throughout the government in this country?” and “Is corruption widespread within businesses in this country?”

**Simple Linear Regression**

(X is one of LGDP, HLE, Support, Freedom and Corruption)

Initial investigations into the relationship between the happiness score and these 5 other measurements were carried out through simple linear regression. This allowed us to examine which of these 5 was most crucial to a happiness score globally. In order to carry linear regression out, however, the validity of the 4 basic assumptions below must be ensured through residual diagnostics.

1. The mean of all residuals is 0.
2. Homoskedasticity, or equal variance, is true across the dataset.
3. All points are independent.
4. The data is normally distributed.

Through residuals plots on R, it became clear that these assumptions would broadly hold true across the datasets. For example, figure 1 shows the 4 residual plots for the Support dataset.

A group of graphs and diagrams

Description automatically generated

*Figure 1. The residual plots for the Support dataset.*

A graph with numbers and letters

Description automatically generatedThe variety of graphs here show that all assumptions will hold. Assumption 1 can be examined through the Residuals vs Fitted graph and is largely aligned. The slight bend to it suggests we will underestimate our small and large values and overestimate the middle values. The largely straight line in the scale location graph, proves Assumption 2, of homoskedasticity, an equal variance across residuals. Finally, the Q-Q plot clearly shows a normal distribution proving number 4. In order to verify Assumption 3, of each residual being independent to the next, an autocorrelation plot can be used as in figure 2.

The lack of any noticeable pattern evident within the autocorrelation immediately suggest that our data can be treated as independent, and this is proven by the low correlation score calculated, of just 0.259. Therefore, we can conclude that for the Support dataset at least, the assumptions hold. The full range of these graphs for each dataset is shown in the appendix however assumptions 1, 2 and 4 broadly hold for all of them. The autocorrelation plots for each predictor variable, however, begins to suggest assumption 3 may not be valid for all datasets and this is further proven by their residual correlation.

|  |  |
| --- | --- |
| Predictor Variable | Residual Correlation |
| LGDP | 0.338 |
| Support | 0.259 |
| HLE | 0.491 |
| Freedom | 0.543 |
| Corruption | 0.818 |

In order to amend the high correlation values seen, especially in Freedom and Corruption, we can randomize the variable selection used in the simple linear regression, which although random and so with different results each time, consistently produces correlations below 0.1, which therefore aligns the data to assumption 3 and allows us to perform simple linear regression. Through simple linear regression, the R2 of each predictor variable can be compared, to reveal the most influential predictor variable. This was carried out twice, once with the full dataset, and once with any outliers, calculated from boxplots, removed.

|  |  |  |
| --- | --- | --- |
| Predictor Variable | R2 values (full dataset) | R2 values (outliers removed) |
| LGDP | 0.6152 | 0.6843 |
| Support | 0.6964 | 0.6445 |
| HLE | 0.5575 | 0.5917 |
| Freedom | 0.4395 | 0.2756 |
| Corruption | 0.2227 | 0.0795 |

This table shows that with the full dataset, support can be considered the most influential variable, however with all outliers removed, this changes to LGDP. Overall, however the values for LGDP, Support and HLE do not undergo significant levels of change with the outliers being removed. The influence of Freedom and Corruption has changed significantly however. The Freedom dataset has now changed from having similar values to other predictor variables, and now shows little clear relationship with the Happiness Index. Corruption has undergone the largest change. The R2 value of 0.2227 was already the lowest R2 calculated and demonstrated little relationship between Corruption and the Happiness Index of a country, however, this has dramatically lowered to just 0.0795 with the reduction of the outliers. This suggests a few countries have significant influence over the Corruption dataset. Therefore, with these results we can conclude that Support is the best single predictor of any given country’s Happiness index, however LGDP and to a lesser extent HLE can give a reasonable indication. Freedom does demonstrate some correlation with the Happiness Index however this is clearly predicated upon some points with high influence and therefore this relationship is far from dependable. The Corruption database has already revealed little relation to the Happiness Index and the removal of outliers has only further reinforced this, suggesting this measure cannot be used as an estimate for Happiness Index individually.

**Multiple Linear Regression**

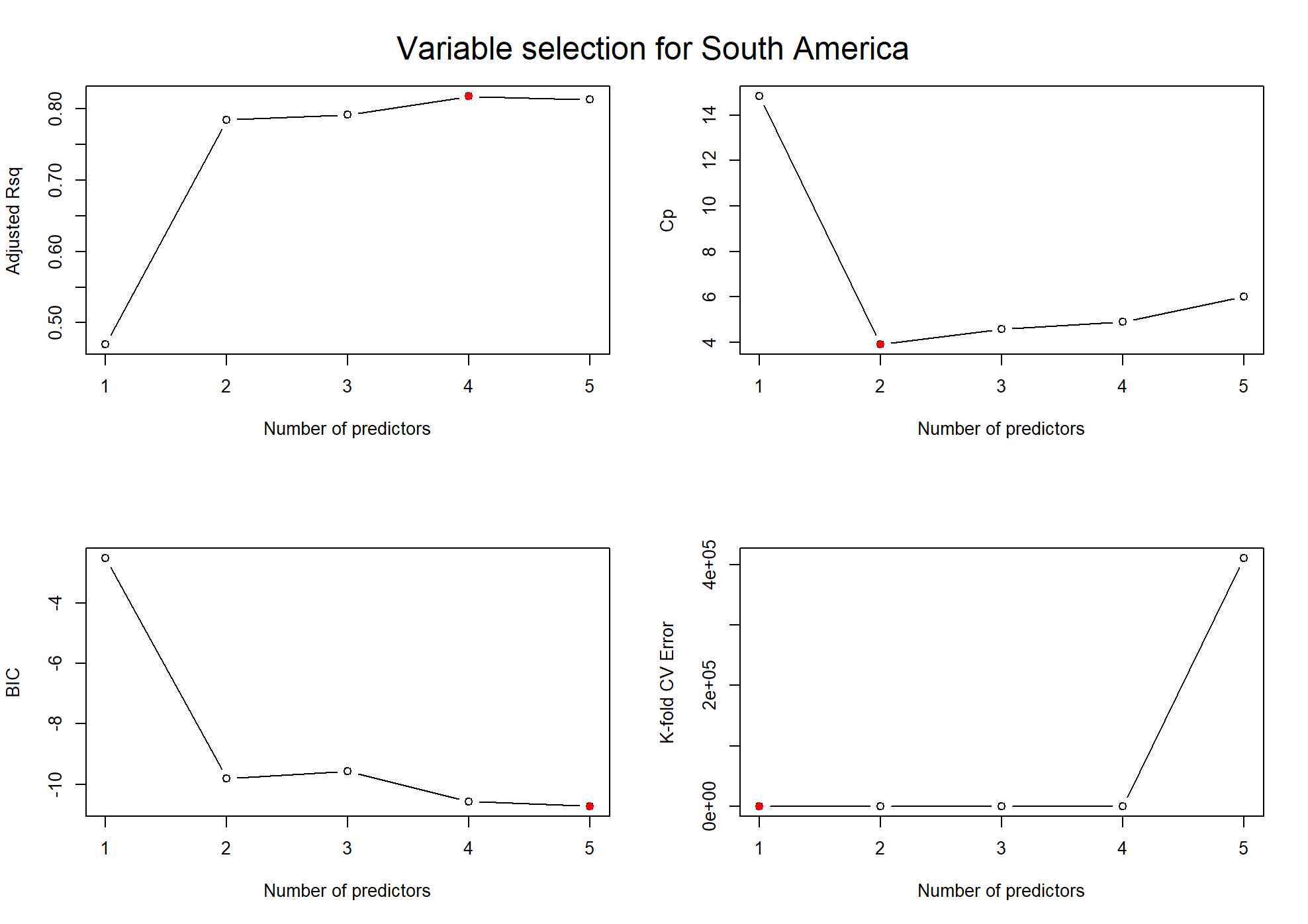
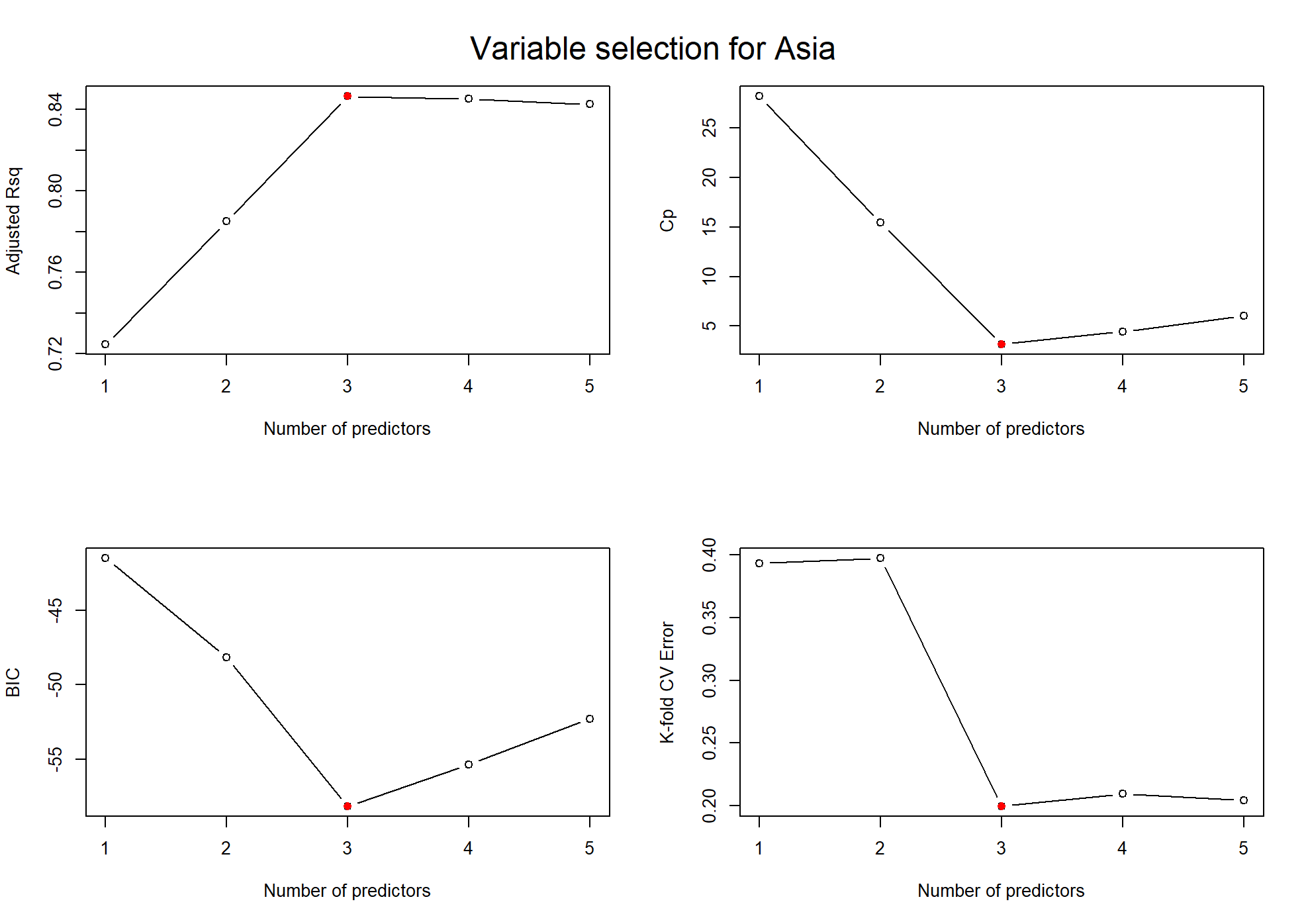
Check the Multi-collinearity of variables with VIF:

A common method to analyse multiple linear regression is through the Variance Inflation Factor (VIF) to test for multicollinearity among the variables LGDP, Support, HLE, Freedom and Corruption. VIF measures the extent to which an independent variable can be linearly predicted by other independent variables.

We set the threshold of , for variables with larger values methods such as deleting highly correlated variables, merging variables, or using regularization can be applied.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | LGDP | Support | HLE | Freedom | Corruption |
| Happy\_general | 3.95 | 2.84 | 3.66 | 1.55 | 1.37 |
| Africa | 1.90 | 1.51 | 1.58 | 1.12 | 1.16 |
| Asia | 2.69 | 2.09 | 2.79 | 1.44 | 1.48 |
| Europe | 4.78 | 1.61 | 3.03 | 1.82 | 2.87 |
| North America | 2.55 | 2.84 | 1.69 | 1.58 | 1.60 |
| South America | 7.37 | 1.54 | 3.05 | 5.28 | 1.43 |

As these vif factors are all below 10, there is no need for any larger value methods to be applied and therefore, all these datasets can be applied to multiple linear regression models.

With the simple linear regression having produced a broad relationship between the Happiness index and each predictor variable established for any country worldwide, multiple linear regression can now be carried out to assess the predictor variables as a whole, as opposed to simply individually. Analysis was also carried out by continent as we hypothesised that the relative importance of predictor variables would likely vary geographically, and this would allow governments to have a far more accurate way to estimate and develop their Happiness index. In order to the carry out the multiple linear regression, a ‘best subset’ analysis was carried out, to identify the combination of predictor variables which had the most influence. The outcome of this analysis varied by continent. Some, such as Asia, showed a clear pattern across our four best subset measures of Adjusted R2, Mallow’s Cp, BIC and k-fold CV error, whilst other continents, such as South America, showed widespread dissonance (figure 4).

Through adjusted R2 values, these subsets could be quantified in terms of how accurate they were for predicting the Happiness Index worldwide and for specific continents. For continents such as South America, with such a wide range of possible ‘best subsets’, individual multiple linear regression was carried out for each possible combination with the best R2 value taken. This was therefore found to be made up of 4 variables.

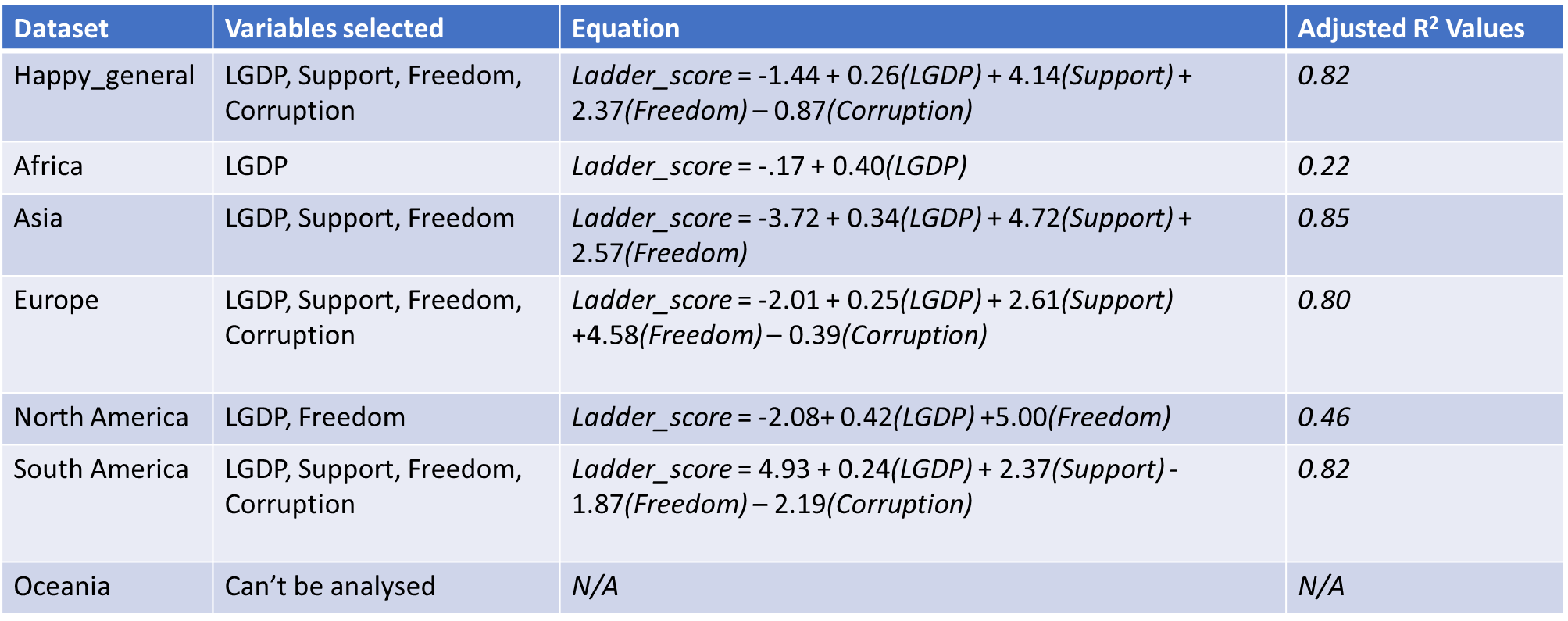
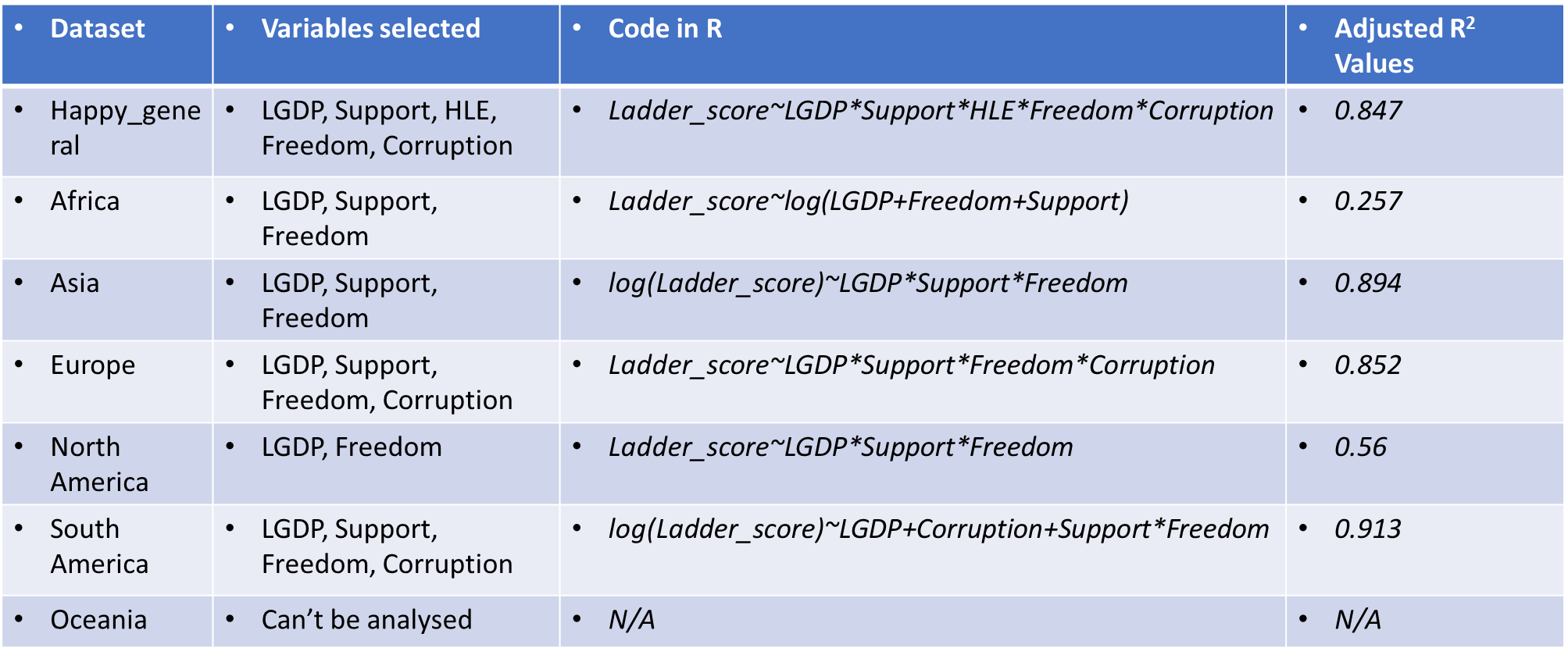


Figure 5 here shows the equations for each of the multiple linear regressions performed so far. Whilst these broadly give high R2 values, some such as North America and Africa continue to yield low results. In order to evaluate the general dataset and each continent further, interaction terms and transformations of the variables can be used to further correlate our two datasets. For these, the ‘best subset’ was once again returned to, as for some continents, where the measures showed little difference between multiple subsets, the interaction terms revealed that an alternative subset would be a better choice.



These interactions terms show that the adjusted R2 improve, however without any dramatic changes. Africa stills shows a low R2 whilst Europe, Asia and the general dataset show little change. Some significant change was shown to both North and South America, which will help with Happiness Index understanding and prediction within these areas.

Oceania has so far not been investigated as a continent and this is as there are only two countries within it’s dataset, making it difficult to quantify. However, it can be compared to other continents in order to understand which continent, its data is most similar to.

Correlations between indicators from Oceania and other regions were calculated using Pearson's correlation coefficient or Spearman's rank correlation coefficient.

For each country, the relative difference between its performance on these indicators and other regional averages can be calculated. This can be achieved by calculating the standardized difference for each indicator and then summarizing the differences.

The closest continent with Oceania is Europe, we combine them into one dataset.

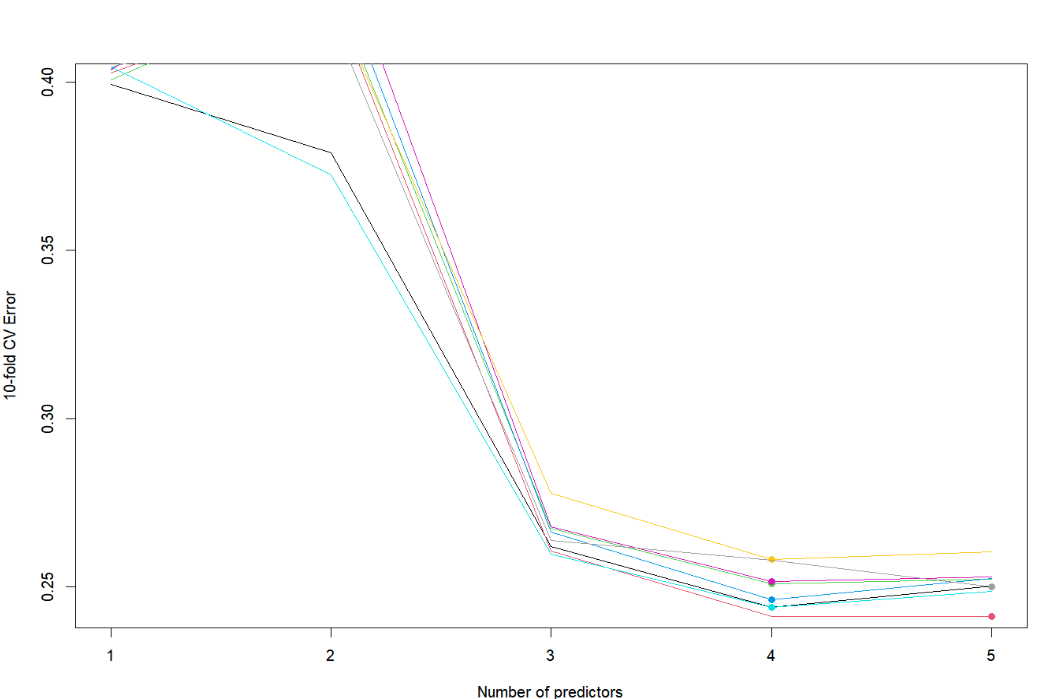
Then we bring the Europe\_Oceania dataset back to the multiple linear regression model as Oceania’s Analysis, the result is below:

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Variable selected | Equation | Adjusted R2 values |
| Europe\_Oceania (Oceania) | LGDP, Freedom, Support, Corruption | -2.5232+0.5183\*LGDP+1.4282\*Freedom+3.2440\*Support-0.8032\*Corruption | 0.81 |

**K-fold cross validation:**

Randomly assign the data set by 10 fold 8 times, thus creating 8 different training and validation set combinations, and calculate their test errors. Find the model with the smallest test error in each fold and store the results.

After multiple rounds of experiments, it was found that in the data set Happy\_general, when the predictor variables are 4 (LGDP, Support, Freedom, Corruption), the model error is the smallest.



**Figure 10-fold cross validation**