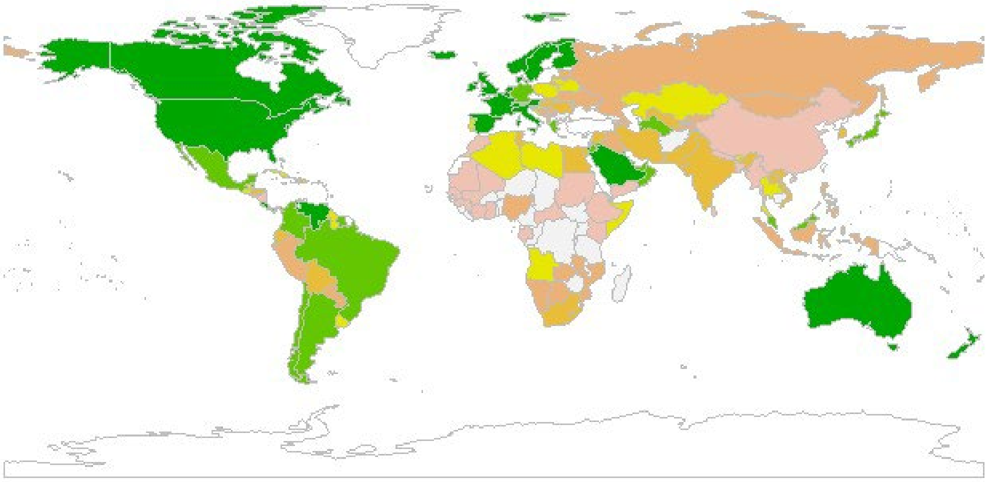
World Happiness Report Analysis Using Multiple Linear Regression



Zhiqi PANG
Xintong YU

Data 603: Statistical Modeling with Data
University of Calgary
Winter 2023

CONTENTS

1 INTRODUCTION

With the latest World Happiness Report, it is evident that being happy has become a worldwide priority. Happiness differs systematically across societies and over time, for reasons that are identifiable, and even alterable through the ways in which public policies are designed and delivered. Therefore, happiness scores are very important for the formulation of government policies and the development and progress of society (Yamini, 2022).

World Happiness Report reflects a new worldwide demand for more attention to happiness and absence of misery as criteria for government policy. It reviews the state of happiness in the world today and shows how the new science of happiness explains personal and national variations in happiness. Our objectives are to analysis of the World Happiness Report from 2008 to 2022 using R, description and understanding of the six parameters affecting a country's people's happiness, and finally establish a model to explore how to calculate world happiness scores.

2 STATISTICAL METHODOLOGIES

2.1 DATA SOURCE

The dataset is from the World Happiness Report (2023), an annual publication of the UN Sustainable Development Solutions Network. It is open on the website of World Happiness Report.

2.2 VARIABLE EXPLANATIONS AND DATA ASSUMPTIONS

The following is a complete list of variables used in our modelling process.

Dependent variable

"ladder", numeric, qualitative data

Independent Variable

1) "country": character, qualitative data

2) “regional”: character, qualitative data

3) “year”: numeric, quantitative

4) “Log GDP per capita”: numeric, quantitative data

5) “Social support”: numeric, quantitative data

6) “Healthy life expectancy at birth”: numeric, quantitative data

7) “Freedom to make life choices”: numeric, quantitative data

8) “Generosity”: numeric, quantitative data

9) “Perceptions of corruption”: numeric quantitative data

10) “Confidence in national government (Positive/Negative affect)”: numeric, quantitative data

2.3 MODELLING PLAN

We plan to approach this project using the methods we have learned in Data 603. We will first run a linear regression model using all predictors and test the variables for multicollinearity. Once we have removed the variables with high multicollinearity, we will use stepwise regression to recommend a model of main effects. We will then perform a partial F-test to compare our full model and reduced model. Once we are satisfied with our main effects, we will use the individual t-test to check for significant higher-order terms and interactions. We intend to test this model with another F-test to evaluate if the higher order terms and interactions are significant. Any significant higher-order or interaction terms will be added to our main effects to produce our final model. Our model will then test for the following 5 assumptions as shown below:

1) Linearity Assumption: Review residual plots

2) Normality Assumption: Using Shapiro-Wilk normality test

3) Equal Variance Assumption (heteroscedasticity): Using Breusch-Pagan test

4) Multicollinearity: Using variance inflation factors (VIF)

5) Outliers: check Cook’s distance and leverage

If our model does not satisfy any of these assumptions, we will review our workflow above to see if any improvements can be made. Once we confirm our model to satisfy all of the assumptions, we will use the model to predict future happiness ladder.

3 RESULTS

3.1 EVALUATING OVERALL MODEL UTILITY

Full model:

$$\widehat{ladder} = \beta_0 + \beta_1 factor(country) + \beta_2 year + \beta_3 GDP + \beta_4 social + \beta_5 healthy + \beta_6 freedom + \ \beta_7 generosity + \beta_8 corruption + \beta_9 positive + \beta_{10} negative$$

To testing a relationship between the response and predictors, an overall F-test is firstly performed to check if the multiple regression model is useful. To address the overall question, the hypothesis is:

$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_p = 0$

$H_a\colon\ at\ least\ one\ \beta_i\ is\ not\ zero\ (i = 1,2, \dots, p)$

Table 1 The ANOVA Table for Happiness Data

Source of Variation	df	Sum of Squares	Mean Square	F-Statistic
Regression	164	2313.7	14.1079	108
Residual	1793	234.20	0.1306	
Total	1957	2547.9		

Compare the NULL model (Model with only intercept) with the full model, the output shows that *Fcal*=108 with *df*= 164 (*p*-value < 2.2e-16 < α = 0.05), indicating that we should clearly reject the null hypothesis. Therefore, the large F-test and *p*-value suggests that at least one of the independent variables must be related to the dependent variable ladder.

Once we check the overall F-test and reject the null hypothesis, we can check the test statistics for the individual coefficients and particular subsets of the full model test.

3.2 MODEL SELECTION

Step-wise regression and other regressions procedure like backward elimination procedure and forward selection procedure are used to screen for significance variables. Then we compare the models by using adjusted R^2, RMSE, AIC (Akaike's information criterion) and Mallows's C_p Criterion. Normally, higher adjusted R^2 and lower RMSE, AIC, *Cp* values indicate better fit models.

Individual Coefficients Test (t-test), let:

H_0: $\beta_i = 0$

H_a: $\beta_i \neq 0$ $(i = 1,2,\dots,p)$

From the summary of the full model, the output shows that the healthy has t$_{cal}$ = -1.468 with the *p*-value = 0.142216 > 0.05, indicating that we should clearly not reject the null hypothesis that the negative has a non-significant influence on the ladder. The other independent variables including year, GDP, social, healthy, freedom, generosity, corruption, and positive can remain in the regression model because of their significant p-values negative at *α* = 0.05.

3.2.1 Stepwise Regression Procedure

penter = 0.1, prem = 0.3 is specified to follow the same procedure of stepwise regression. The first order model is:

$$\widehat{ladder} = \beta_0 + \beta_1 factor(country) + \beta_2 year + \beta_3 GDP + \beta_4 social + \beta_5 freedom + \beta_6 generosity + \beta_7 corruption + \beta_8 positive + \beta_9 negative$$

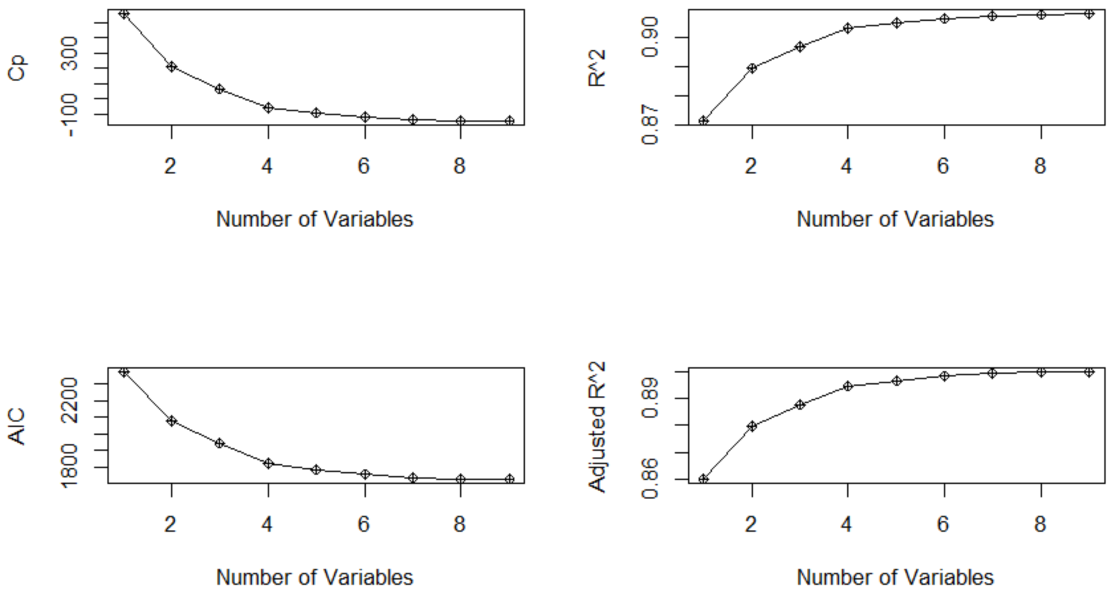


Figure 1 *Cp*, *AIC*, R^2 and adjusted R^2

A small value of *Cp* and AIC means that the model is relatively precise. Adjusted R^2 increase only if RMSE decreases. From the output, we can find when the number of independent variables improves to 6, the model is the best.

3.2.2 Testing for Interaction in Multiple Regression

After using model selection by automatic methods or all possible regression methods, we might not have the best fit model yet, as we consider only main effects on independent variables. After eliminating some variables that are not important out of the model, we consider interaction terms and/or high order multiple regression model to improve the model.

$$\widehat{ladder} = \beta_0 + \beta_1 factor(country) + \beta_2 year + \beta_3 GDP + (\beta_4 social + \beta_5 freedom + \beta_6 generosity + \beta_7 corruption + \beta_8 positive + \beta_9 negative)^2$$

To test for multicollinearity in our models, we check the variance inflation factors (VIF) to determine which variables should remain in our best fitted model.

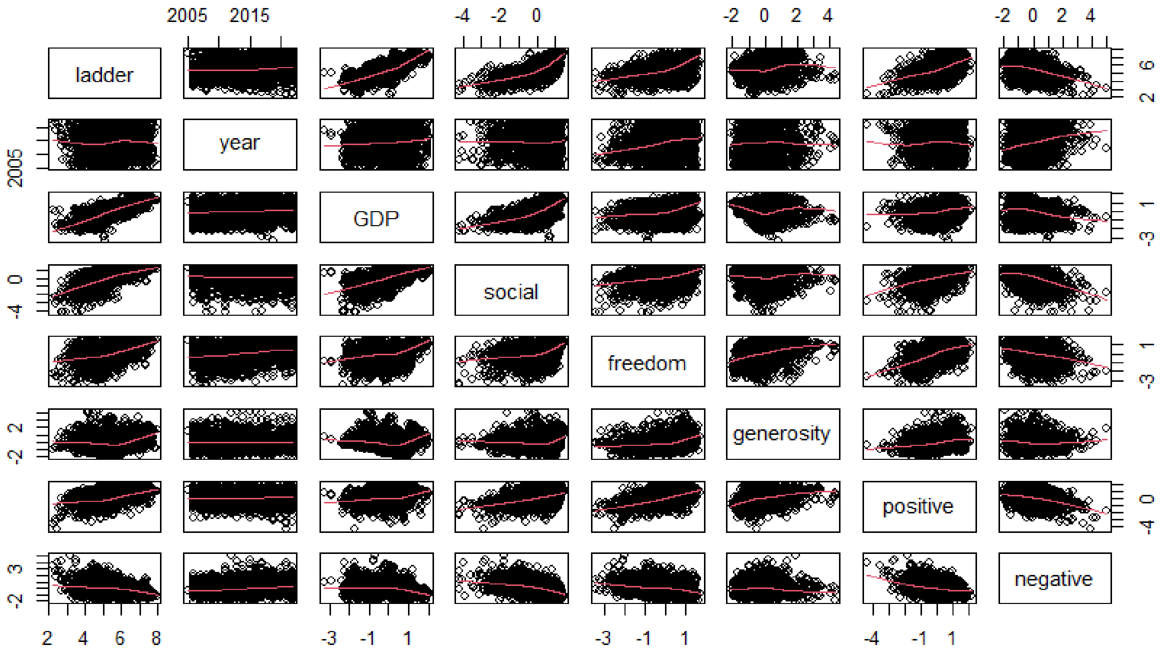


Figure 2 Pairs plot of variables in happy dataset

The output of VIF shows that multicollinearity is tested due to GDP. There are no interaction variables with GDP, so we think multicollinearity is due to countries with GDP. But we cannot remove or combine countries, because we need the country factor for prediction. GDP is also very important, so we keep it in the model.

3.2.3 The best fit model

Trying more regression selection methods, the interaction terms of social*freedom and positive*negative are determined to be consistently significant. Using t test, the nonsignificant terms of year*social, year*negative, negative, social, corruption are dropped off from interaction models. We also test for high order model between predictors and Y to improve the model. The final best fit model is as below.

$$\widehat{ladder} = \beta_0 + \beta_1 factor(country) + \beta_2 year + \beta_3 GDP + \beta_4 social + \beta_5 freedom + \beta_6 generosity + \beta_7 positive + \beta_8 negative + \beta_9 freedom^2 + \beta_{10} negative^2 + \beta_{11} social * freedom + \beta_{12} positivef * negative$$

The final best fit includes country, year, GDP, social, freedom, generosity, positive and negative, quadratic terms of freedom and negative, the interaction term of social * freedom and positive*freedom. Except for the year, all variables in the best fit model are significant. Compared with the first order model, after including the interaction terms and quadratic terms, they led to such a big improvement in the model. adjusted R^2 increases from 0.899671 to 0.9034, and the standard error of residuals (RMSE) decreases from 0.3614163 to 0.3547. Therefore, it is clear that adding the additional terms really has led to a better fit to the data.

3. 3 REGRESSION MODEL DIAGNOSTICS

Once we have fitted the multiple regression model, model Diagnostics are performed, including linearity Assumption, normality assumption, equal Variance Assumption (heteroscedasticity) using Breusch-Pagan test, multicollinearity using variance inflation factors (VIF), and outliers by checking Cook's distance and leverage.

3.3.1 Linearity Assumption

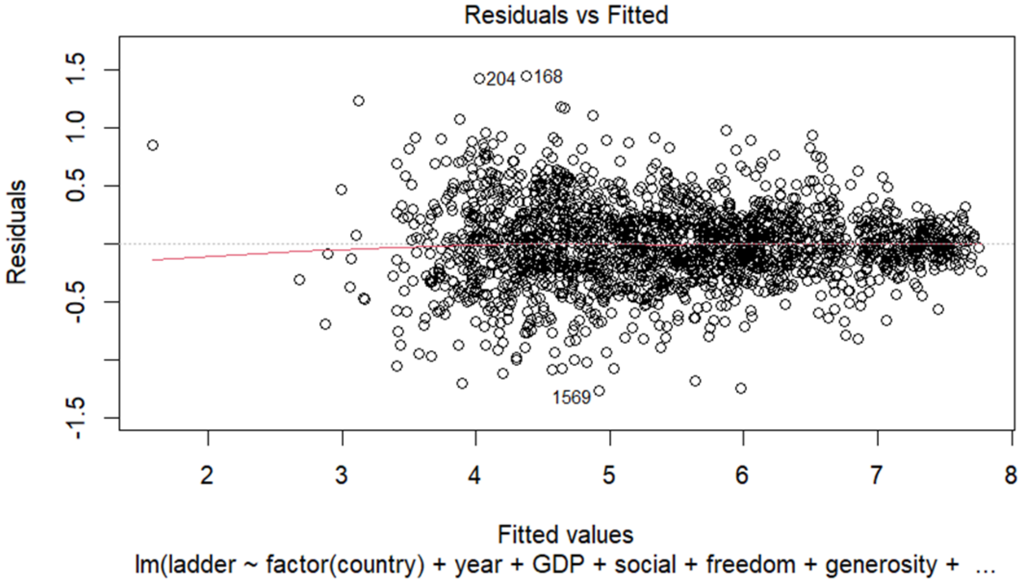
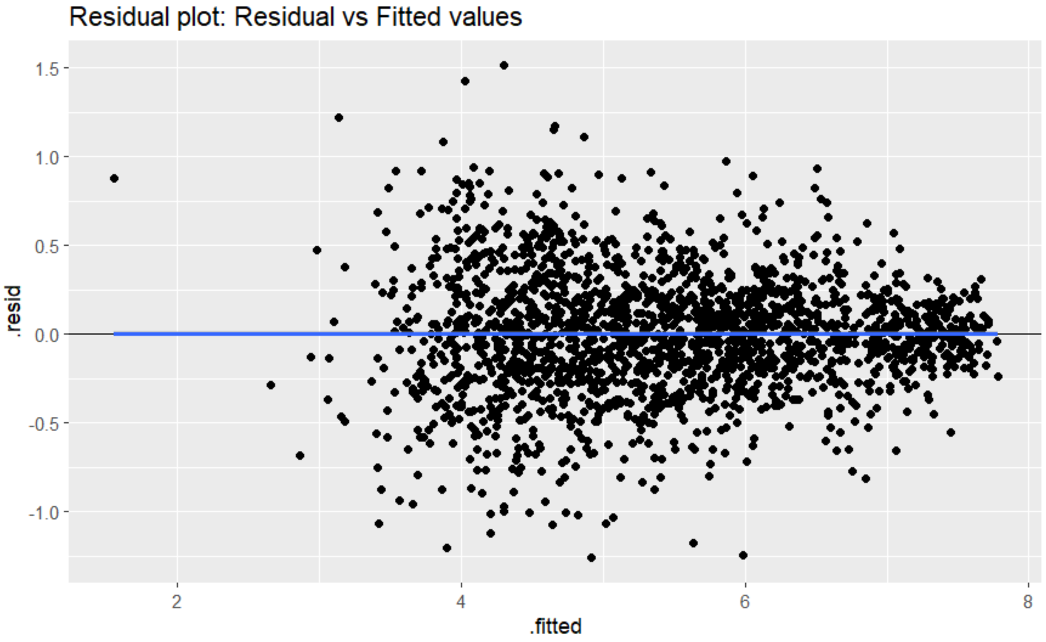


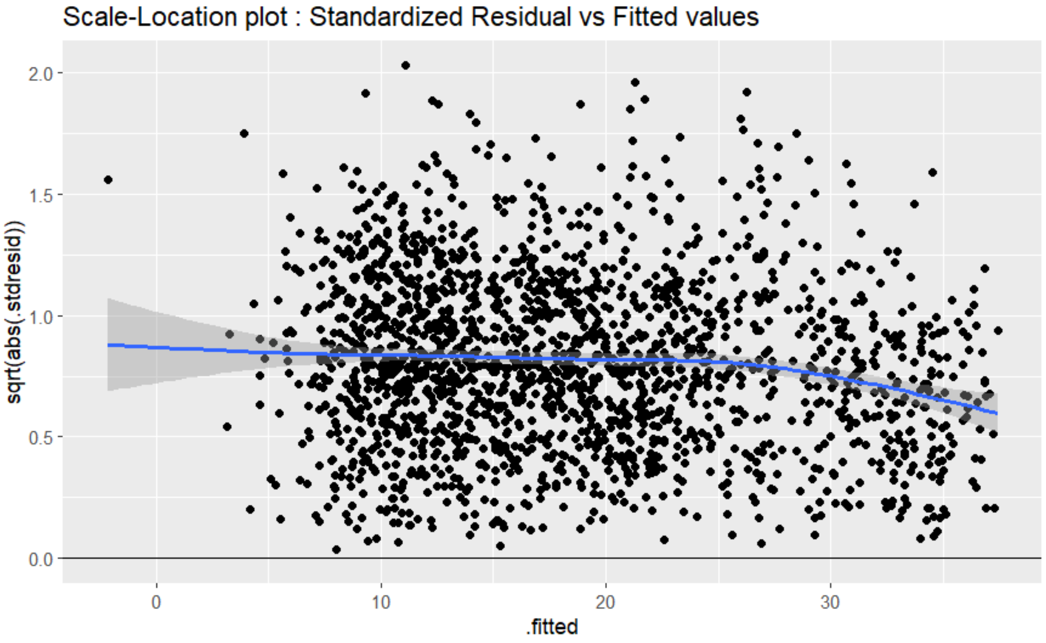
Figure 3 Residuals vs fitted values

For checking the linearity Assumption, the residual plots is reviewed. The plot of residuals vs. values shows that there is no pattern of the residuals in the best fit model.

3.3.2 Equal Variance Assumption



(a)



(b)

Figure 4 Scale-location plot

The output displays the residual plot and Scale-Location plot that result from the best fit model. In our case, the residuals tend to form a horizontal band-indicates that the plot does not provide evidence to suggest that heteroscedasticity exists.

Breusch-Pagan test, let:

H_0: heteroscedasticity is not present (homoscedasticity)

H_a: heteroscedasticity is present

The output displays the Breusch-Pagan test the p-value = 2.2e-16 < α = 0.05, indicates that null hypothesis is rejected. Therefore, the test provides evidence to suggest that Heteroscedasticity does exist. We've been trying models with more power on (e.g., power of 11), it still shows heteroscedasticity.

3.3.3 Normality Assumption

Shapiro-Wilk normality test, let:

H_0: the sample data are significantly normally distributed

H_a: the sample data are not significantly normally distributed

The Shapiro-Wilk normality test shows that the residuals are not normally distributed as the *p*-value = 2.86e-9 < 0.05.

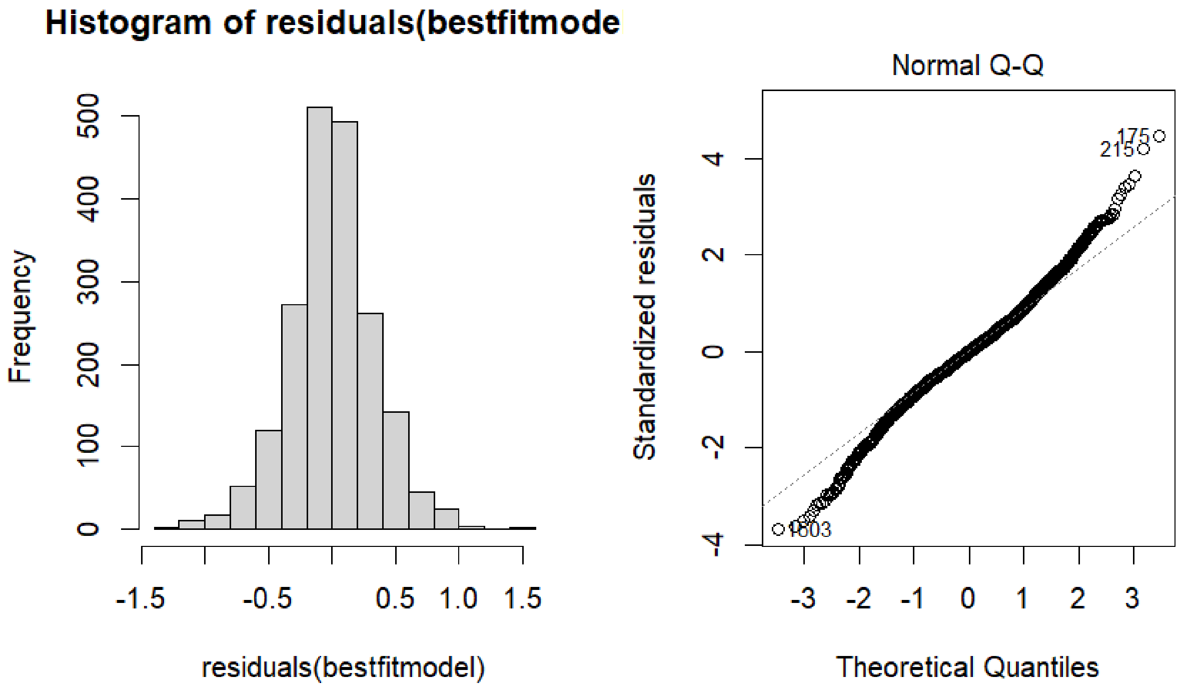


Figure 5 Histogram and Q-Q plot

The geom_qq_line and stat_qq_line compute the slope and intercept of the line connecting the points at specified quartiles of the theoretical and sample distributions. The outputs show the residual data of the best fit model. have a normal distribution (from the histogram and Q-Q plot).

3.3.4 Outlier

In the cook's distance and leverage plot the above outliers are detected. A new dataset was created by removing these points. After dropping the outliers, adjusted R square increases from 0.9034 to 0.9054, and the standard error of residuals (RMSE) decreases from 0.3547 to 0.3521, compared with using the old dataset.

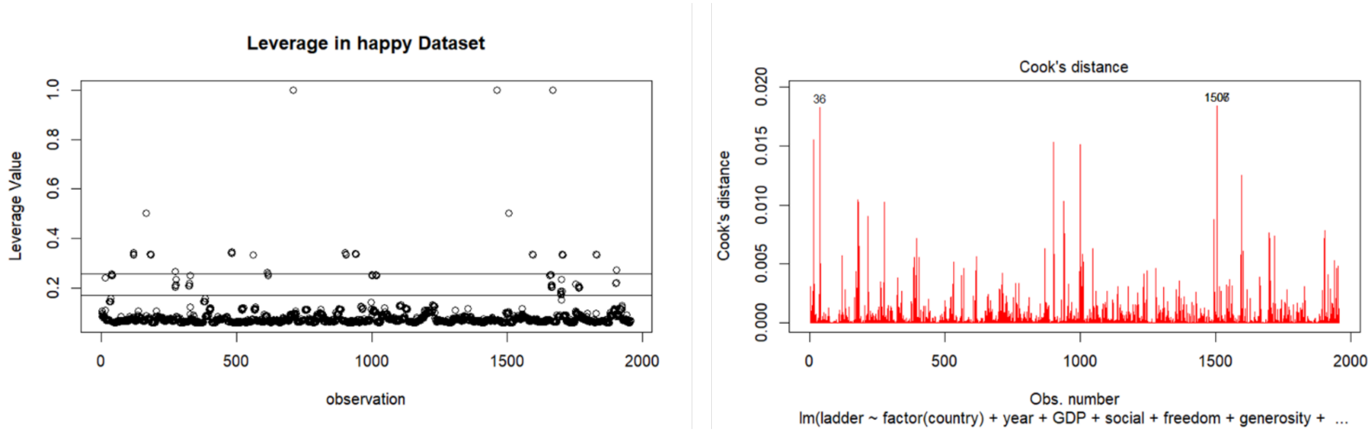


Figure 6 Plots to check for outliers

From the output we can find there are some points of high leverage and far from the line. These points can strongly influence the slope of the least squares line.

We drop the outliers that $h > \frac{3p}{n}$. For our case, p = 167 and n = 1958.

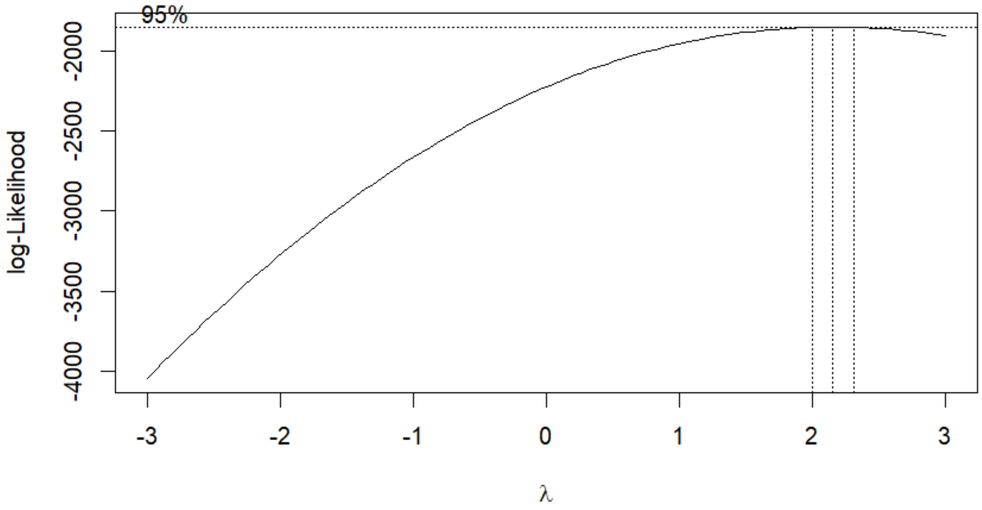


Figure 7 Plot of best lambda

Using the the Box-Cox method we calculated the best lambda is 2.151515. After Box-Cox Transformations, the scale-location plot shows a better pattern than the previous one. The histogram and Q-Q plot reflect that the residual data have a normal distribution. However, the Shapiro-Wilk normality test still not pass the normality assumption as the 1.053e-07 < 0.05.

3.4 INTERPRETING COEFFICIENTS

The best fit model:

$$\begin{aligned}\widehat{ladder} &= 0.588069 + 0.002488 year + 0.620407 GDP + 0.230608 social \\ &+ 0.063943 freedom + \ 0.051502 generosity + 0.218000 positive \\ &- 0.080402 negative - 0.035960 freedom^2 + 0.033643 negative^2 \\ &+ 0.066725 social * freedom + 0.092379 positive * negative \\ &+ \begin{cases} -0.794434 & if\ the\ country\ is\ Angola \\ -0.0409952 & if\ the\ country\ is\ Azerbaijan \\ -1.793368 & if\ the\ country\ is\ Botswana \\ & \dots \\ & \dots \\ -0.0726104 & if\ the\ country\ is\ Zimbabwe \end{cases}\end{aligned}$$

The final best fit model includes the predictors of year, GDP, social, freedom and generosity, positive, negative, quadratic terms of freedom and negative, interaction terms of social* freedom, positive*negative.

The interception, β_0 =-22.1403. In this model, variables of year, GDP, social, freedom, generosity and positive are positively related to the dependent variable Y (the ladder). The coefficient of negative = -0.1890 means 1-unit negative attitude towards the government, the ladder decreases 0.1890. Among the coefficients, GDP has the biggest influence towards the *Y*, β_{GDP} = 4.4252, which indicates that 1-unit GDP per capita higher leads to 4.4252 ladder increase.

Adjusted R^2 and RMSE of Best Fitted Model

- Adjusted R^2 = 0.9222, this value indicates that 99.22% of the variation of the response variable “ladder” can be explained by the final model containing the predictors year, GDP, social, freedom, generosity, positive, negative, the quadratic terms of freedom and negative, as well as the interactions between social* freedom, positive*negative.

- RMSE = 2.292, this value indicates that the standard deviation of the unexplained variation in estimation of response variable “ladder” is 2.292.

4 PREDICTION

Predicted happiness score in South Africa in 2030

To predict future atmospheric happiness score in South Africa, we used the projected data from South Africa from 2006 to 2021. In addition to the year and country, other independent variables can be seen as time-series data. In 2030, the expected GDP is 0.1562965, expected social is 0.9538768, expected freedom is 0.2512453, expected generosity is -0.2624, expected positive is 1.516889 and expected negative is 0.3723583. The fit prediction value of happiness score in South Africa in 2030 is 5.299911 and 95% prediction interval is between 4.581515 and 6.018308.

$$\begin{aligned}
&\widehat{ladder}_{South\ Africa} \\
&\qquad = 0.588069 + 0.002488 year + 0.620407 GDP + 0.230608 social \\
&\qquad + 0.063943 freedom + 0.051502 generosity + 0.218000 positive \\
&\qquad - 0.080402 negative - 0.035960 freedom^2 + 0.033643 negative^2 \\
&\qquad + 0.066725 social \ast freedom + 0.092379 positive \ast negative - 1.030479
\end{aligned}$$

$$\begin{aligned}
&\widehat{ladder}_{South\ Africa} | year = 2030 \\
&\qquad = 0.588069 + 0.00248 \ast 2030 + 0.620407 \ast 0.1562965 + 0.230608 \\
&\qquad \ast 0.9538768 + 0.063943 \ast 0.2512453 + 0.051502 \ast (-0.2624189) \\
&\qquad + 0.218000 \ast 1.516889 - 0.080402 \ast 0.3723583 - 0.035960(0.2512453)^2 \\
&\qquad + 0.033643(0.3723583)^2 + 0.066725 \ast 0.9538768 \ast 0.2512453 + 0.092379 \\
&\qquad \ast 1.516889 \ast 0.3723583 - 1.030479 \\
&\quad = 5.299911
\end{aligned}$$

5 CONCLUSIONS

To summarize our findings from the analysis, the main effects of country, year, GDP, social, freedom, generosity, positive and negative are shown be significant for happiness score. And higher order terms of freedom and negative are significant ($p - values < \alpha = 0.05$). Interactions between social and freedom and positive and negative are significant (based on individual coefficient test: p-values).

To examine the practical implications of our model, we first must identify how each term affects happiness score. The GDP per person as a predictive variable confirms that higher GDP in a country lined to higher happiness score. Similarly, when a country has better social support, and residents have more freedom to make life choices, their happiness score will be higher. People in the country are more likely to donate to charities and they feel more enjoyable, the country' ladder can be high. By contrast, if people always feel worried, sadness or anger, the country is more likely to get low scores. The coefficient about the country can reflect the average score of the country compared with the world average score. If it is higher than the world average score, the coefficient is positive, otherwise it is negative. From the distribution of the map (Figure 8), we found that Northern Europe, North America, and Australia got high score, so countries in these regionals, the coefficients are greater than 0.

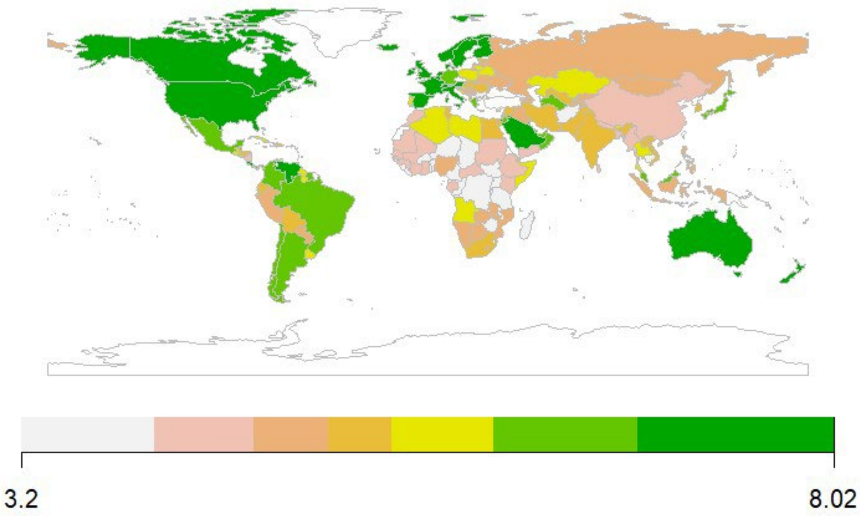


Figure 8 World happiness map of 2023

6 DISCUSSIONS

In a first-order model, the VIFs of all independent variables are 0. But as long as the interaction term and higher power term are added, the multilinearity of most of the terms are detected. After the data standardization, only items related to GDP have collinearity problems. We deleted the interaction items related to GDP and optimized the model. Because in the end the happiness score of a specific country should be predicted, so we keep the factor of country in the model although it is related to GDP.

After dozens of attempts with different models, the unequal variance and non-normality of the error term always appeared together in our model. In the final best fit model, after dropping the outliers and box-cox transformation, the residuals of the model still do not fit normal distribution. *p*-value=1.053e-07 of Shapiro-Wilk normality test is lower than 0.05. However, the histogram and Q-Q plot show that the data are normally distributed. Maybe in the future we can use more advanced methods to solve the nonnormality and heteroscedasticity.

For prediction of the happiness score, all independent variables are predicted by year. If we can find other factors of independent variables, we can remove the time-series effects and have much lower multicollinearity between variables. This would improve prediction accuracy.

REFERENCES

- Kaggle. (2022) ‘World Happiness Report up to 2022’ [Online]. Available at: https://www.kaggle.com/datasets/mathurinache/world-happiness-report?select=2021.csv/ (Accessed 11 March 2023).
- World Happiness Report. (2023) [Online]. Available at: https://worldhappiness.report/archive/ (Accessed 21 March 2023).
- Yamini, A. (2022) ‘Analysing World Happiness Report (2020-2022)’. Available at: https://www.analyticsvidhya.com/blog/2023/01/analysing-world-happiness-report-2020-2022/ (Accessed 11 March 2023).