Building Regression Models for Happiness Rate Predictions

048DSARM3 (Applied Regression and Time Series Analysis)

Saint Joseph University

24-10-2022

Doctor Abbas Mourad

Mohammad Falha

Contents:

Abstract …… 2

Literature Review …… 2

Methodology …… 3

Data preparation and Descriptive Statistics …… 3

1- Data source and Variables Definition …… 3

2- Cleaning and Preparation Steps …… 3

3- Summary of the Data …… 4

4- Normality of the Dependent Variable(Y)…… 4

Correlation Analysis …… 4

Building Models and Testing Linear Regression Assumptions…… 6

Comparing Between Models …… 10

Prediction and Validation …… 11

Result …… 12

Discussion …… 12

Conclusion and Recommendations …… 12

References …… 13

Appendices …… 14

Abstract:

This project aims to conduct statistical analysis on survey data that contains happiness scores in different countries and other factors, to study the relationship between those variables and how they may affect the happiness score. In the end, a model will be established using multiple regression tools to predict happiness scores based on features that influence this score, moreover, the recommendations proposed by this regression model will help in focusing and working on the real predictors controlling the happiness score.

Keywords: Happiness index, life ladder, best-fitted model, countries, predictors

Literature review:

The “Happiness index/score” term was used popularly decades ago, many surveys and studies focused on topics relevant to this title, interpreted factors affecting it, and proposed solutions to enhance the feeling of happiness among people.

Shedding light on this topic returns to its importance in affecting people’s mental and emotional health, which is reflected in crucial phenomena in societies such as suicide, crime, and family problems (divorce, domestic violence…) rates.

Ahtesham (2020) [1] studied in its paper the position of India in the “world happiness report”, and what are the factors that are causing the decline in its happiness index (thirty ranks down in less than a decade), it presented the reasons behind this decaying such as economic inequality (India’s richest 1% has almost 73% of the total wealth created in the economy), the public health spending in insufficient in a way that millions of Indians are deprived of health care services, absence of trustworthy social support system and others… consequently, Ahtesham suggested a set of instructions that can help in raising the happiness score in India such as spreading the spirit of volunteering social service, improving public health, and working towards a more equitable society.

HU's paper [2] analyzed the factors influencing happiness, he concluded that gender (females are happier than males, which is maybe due to the roles associated with both genders in Chinese society), employment situation (employed, retired, learning people have better feelings than unemployed ones), education, marital status (single, widowed and divorced individual’s happiness level is lower than married individuals), health and social interaction have positive significant correlation with the Chinese happiness rate, while the research revealed that the location of residence (urban or suburban) had little bearing on people's happiness. After his research, he examined the impact of China's GDP on the happiness index and concluded that there is no meaningful correlation, and that the happiness paradox also exists in China.

Alba (2019) [3] conducted a statistical study of the "World Happiness Report 2017" and concluded that the socioeconomic North-South split has a strong impact on the happiness rate.

Additionally, he demonstrated that the ten happiest nations are in the global north, whereas the ten unhappiest nations are located in the global south.

Methodology:

In the beginning, data will be cleaned in excel, out-of-interest data will be dropped, the missing value will be filled, and then data will be imported to R studio to conduct statistical analysis on it, to choose the best-fitted model to predict the happiness index, so finally the model can be used to estimate this index in different countries.

Data Preparation and Descriptive Statistics:

1- **Data Source and Variables Definition:**

The statistical analysis was done on World Happiness Report 2022 data [4], the variables definition is (See Appendix B for detailed description):

- Y: Happiness score (or Life Ladder)
- X1: Log GDP per capita
- X2: Social support
- X3: Healthy life expectancy at birth
- X4: Freedom to make life choices
- X5: Generosity
- X6: Perceptions of corruption
- X7: Positive affect
- X8: Negative affect
- X9: Confidence in national government

2- Cleaning and Preparation Steps:

The following steps were performed into the train data in excel (117 records) to become ready for the analysis:

- The table in the report contains records for countries over many years, for that the data was filtered to extract the 2021 records only so each country represents a record.
- Country and Year columns were dropped; now the table contains only the dependent and independent variables.
- Some values were missing from the data, for that, we filled the values either by calculating the average of the column or by the nearest record. (See Appendix B for details).

Note that we did not split the data into train and test data since the test data is a separate file, but the same process was applied on the test data (31 record which forms 21% of the overall data), but the difference that it was chosen from the year 2018.

3- Summary of the data:

- After introducing data into R studio and taking a brief look at the variables we summarized it to obtain the minimum value (Min), the value of the 1st quartile (1st Qu.), median value (Median), the value of the 3rd quartile (3rd Qu.), maximum value (Max) for each variable, the results are shown in Figure1 below.

```
> summary(my_data)
       Y                X1              X2               X3              X4               X5                X6              X7               X8              X9
 Min.   :2.179   Min.   : 5.527   Min.   :0.4355   Min.   :51.30   Min.   :0.3943   Min.   :-0.28553   Min.   :0.1449   Min.   :0.1789   Min.   :0.1161   Min.   :0.1067
 1st Qu.:4.921   1st Qu.: 8.808   1st Qu.:0.7170   1st Qu.:61.90   1st Qu.:0.7171   1st Qu.:-0.07591   1st Qu.:0.6695   1st Qu.:0.5968   1st Qu.:0.2303   1st Qu.:0.3269
 Median :5.721   Median : 9.562   Median :0.8483   Median :66.30   Median :0.8075   Median : 0.01918   Median :0.7699   Median :0.6663   Median :0.2747   Median :0.4939
 Mean   :5.608   Mean   : 9.534   Mean   :0.8044   Mean   :65.30   Mean   :0.7860   Mean   : 0.02999   Mean   :0.7228   Mean   :0.6540   Mean   :0.2909   Mean   :0.4952
 3rd Qu.:6.436   3rd Qu.:10.536   3rd Qu.:0.8971   3rd Qu.:69.35   3rd Qu.:0.8710   3rd Qu.: 0.13060   3rd Qu.:0.8515   3rd Qu.:0.7337   3rd Qu.:0.3451   3rd Qu.:0.6176
 Max.   :7.794   Max.   :11.545   Max.   :0.9799   Max.   :74.35   Max.   :0.9651   Max.   : 0.54583   Max.   :0.9463   Max.   :0.8344   Max.   :0.6067   Max.   :0.9298
```

Figure 1: the summary of each variable

4- Normality of the dependent variable(Y):

Although the normality of the dependent variable is not a necessary assumption in linear regression models, checking the distribution of Y is necessary because if it is not normal it may affect the distribution of prediction error, which is assumed normal. Therefore, to test if Y is normally distributed, we plotted the boxplot for Y (Figure 2) which showed that Y seems to be normally distributed. Moreover, to ensure this observation, we plotted the distribution of Y as a histogram; we can notice that it is slightly skewed to the left (negatively skewed), so we applied transformations to Y (Figure 3) to see if we get a better distribution but the non-transformed Y is the best normally distributed among them.
In addition to the graphical representation, we performed Shapiro–Wilk test on Y, and the p-value was significant (0.0556 > 0.05) which means that the null hypothesis (normality) is not rejected.

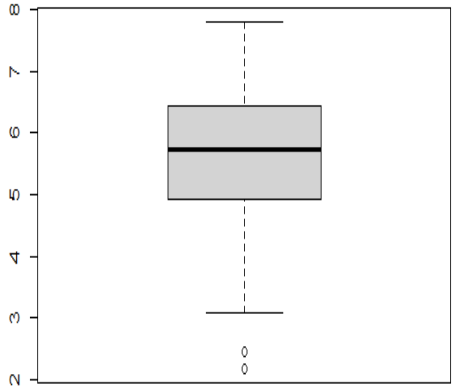

Figure 2: Boxplot of Y distribution

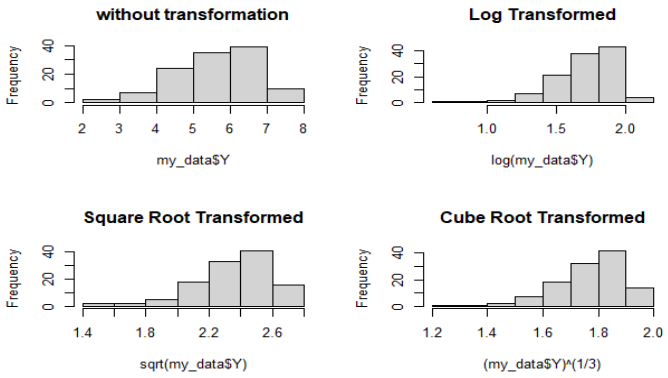

Figure 3: Histograms of Y distributions

Correlation Analysis:
To study if there are two or more correlated independent variables, and to ensure the absence of multicollinearity in our model, we plotted the matrix of scatterplots between each pair of variables (Figure 4) and a correlation heat map (Figure 5).

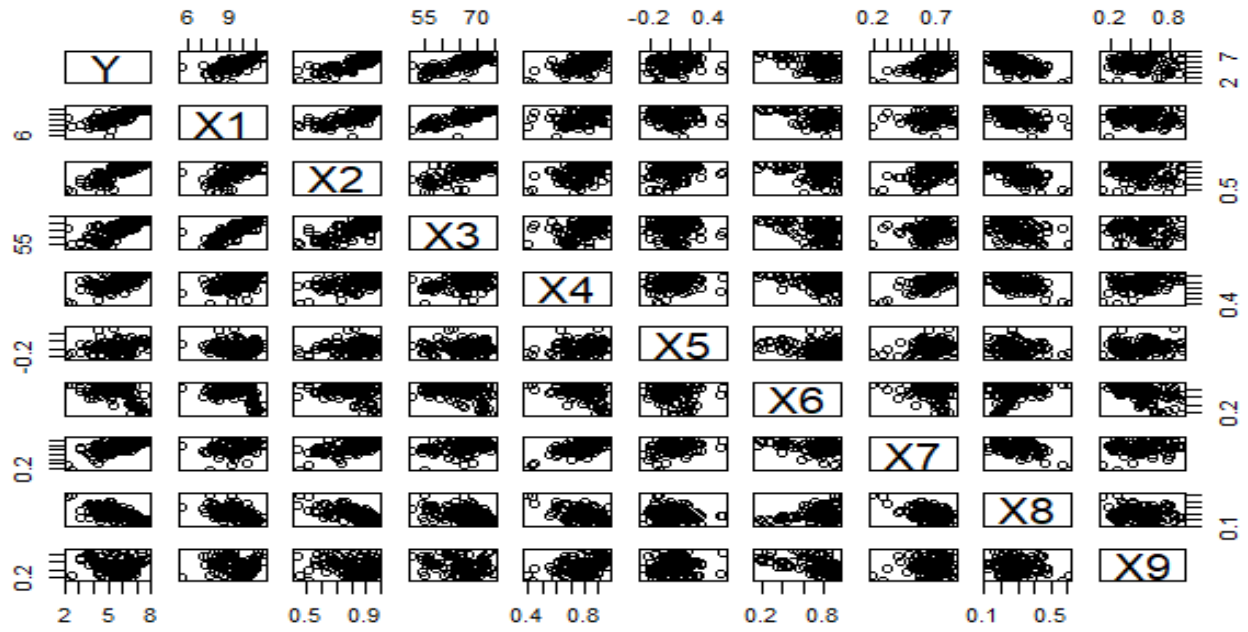

Figure 4 matrix of scatter plots between each pair of variables

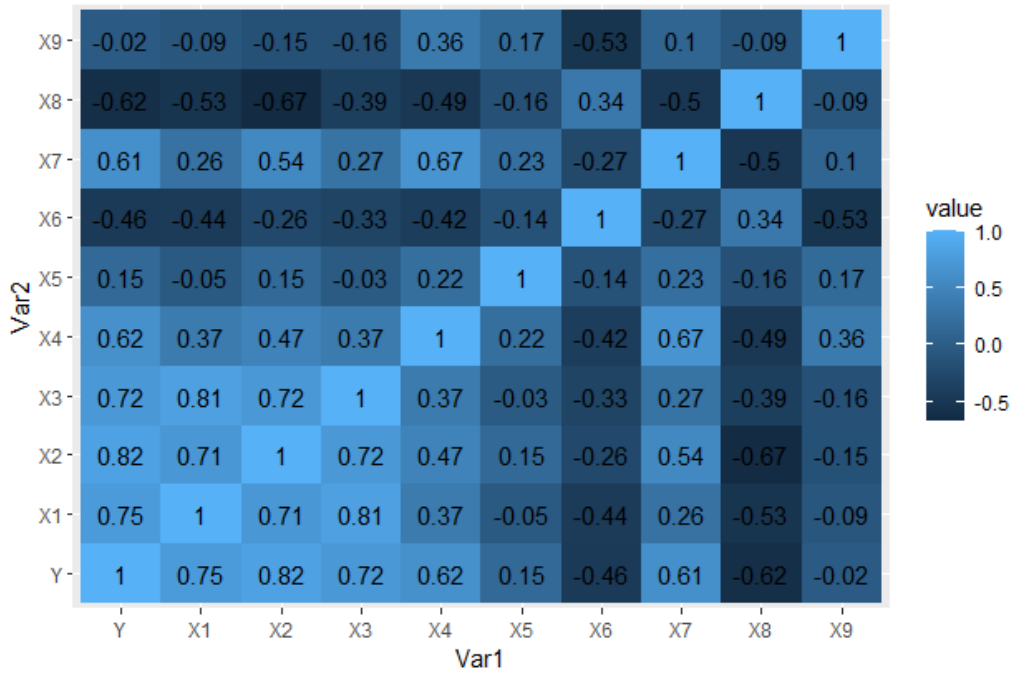

Figure 5 Correlation Heatmap with coefficients

Plots Interpretation: The matrix of scatter plots shows that there is a possible correlation between (X1 and X3), (X2 and X1), and (X2 and X3) variables, but since it is somehow vague it is not sufficient to determine multicollinearity. However, the heatmap (with coefficients) gives more accurate assumptions where we can notice that the only high correlation is obtained between X1 and X3 variables (r=0.81), so we should consider this while building our model by dropping one of these two variables.

Building Models and Testing Linear Regression Assumptions:

To determine the best-fitted model, we will start building a model containing all the variables, after that, we will eliminate insignificant variables one by one, finally, we will test the linear regression assumptions in models that contain only significant variables to choose the best one among them.

For each model with only significant variables, the below measures will be calculated to help in comparing between models:

- AIC: Akaike information criterion, it is a score (estimator of prediction error) that is used to compare models, the lowest the AIC the best the model is.
- R^2: indicates how well the model can explain the observed data.
- RMSE: Root Mean Square Error, estimates the difference between actual y-values and the regression line, so models with higher RMSE are inaccurate more than others.

As well, as to study the listed linear regression assumptions:

- Assumption of linearity: states that the relationship between Y and the predictors is linear.
- Assumption of normality: states that the error is normally distributed, and its mean is zero.
- Assumption of homoscedasticity of variance: states that the residuals have constant variance

the following plots will be plotted:

- Residuals vs fitted: the plot tests the linearity of the model and non-constant variance.
- Q-Q plot: assess if the residuals are normally distributed.
- Scale-location plot: checks the assumption of homoscedasticity.
- Residuals vs leverage plot: tests if there is an influential observation in the data.

The models studied are as below (Appendix C contains all the models summary in R studio):

Model 1: containing all independent variables
Analyzing Results: p-value for the F-test (overall significance of the regression model) is significant (<2.2e-16), which means that there is at least one significant variable in the model, but the p-value for each of X3, X5, and X8 variables are insignificant, so it is not the best-fitted model.

Model 2: containing all independent variables except X3
Analyzing Results: X5 and X8 variables are insignificant, so it is not the best-fitted model.

Model 3: containing all independent variables except X3 and X5
Analyzing Results: X8 variable is insignificant, so it is not the best-fitted model.

Model 4: containing all independent variables except X3 and X5 and X8

Analyzing Results: all variables are significant, so assumptions will be checked by appropriate plots (Figure 6) and tests.

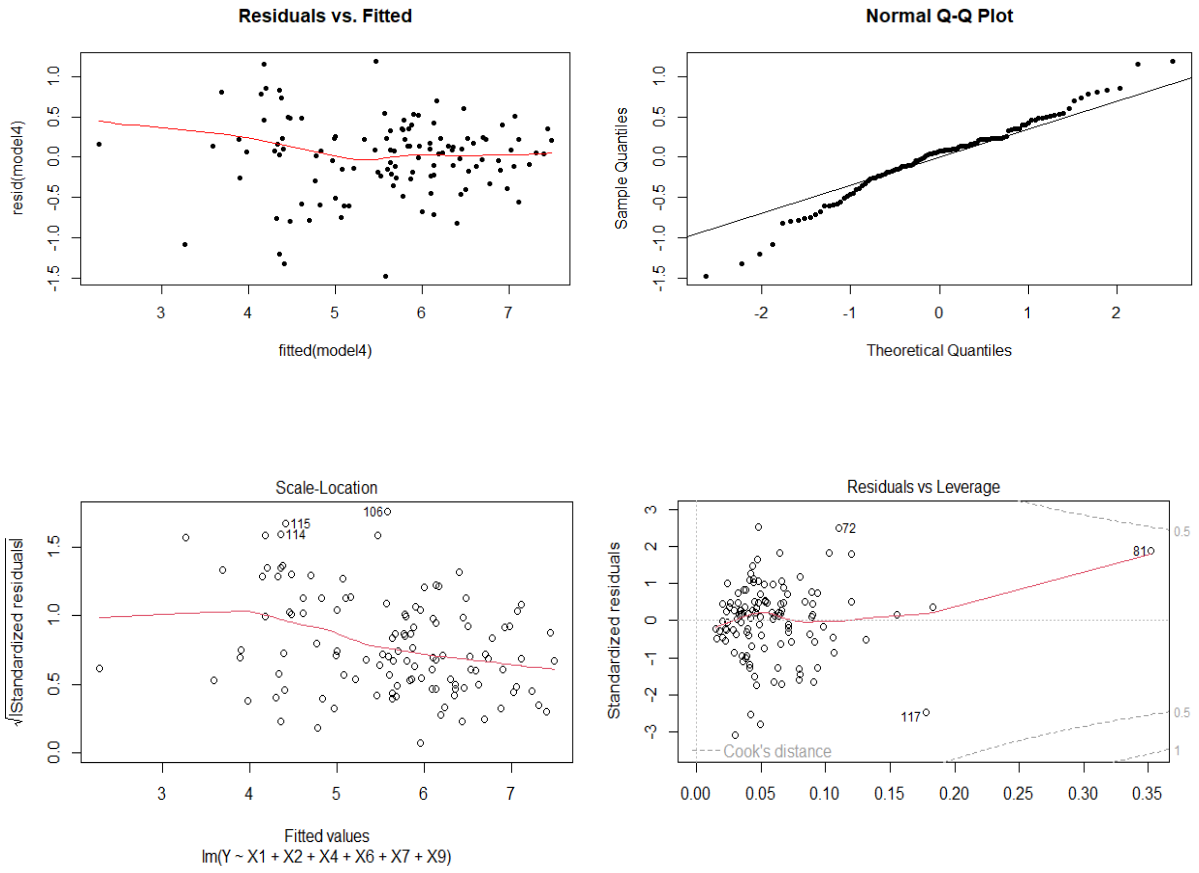

Figure 6

Some measures of the model:

AIC	p-value of Shapiro test	R-squared	RMSE
170.5528	0.03729	0.8167	0.6772187

Results interpretation:

1- Assumption of linearity: the relation is linear since the red line in the 'residuals vs fitted' plot is roughly straight at zero and the points are randomly spread around the line which indicates that the variance does not vary as a function of the fitted values.
2- Assumption of normality: The normality of errors seems to hold since data points in the Q-Q plot are forming approximately a straight line.
3- Assumption of homoscedasticity of variance: since the line in the scale location does not show many fluctuations, and the spread of standardized residuals around the red line doesn't vary so the assumption is true.

4- No influential points since there are points outside of Cook's distance in the Residuals vs leverage plot (no need to eliminate the outliers).
5- No multicollinearity: X1 is highly correlated with X3 as mentioned above so by eliminating X3 from this model, we will solve the problem of multicollinearity.
6- The independent variables are 81.67% able to explain the variation of the Y variable.

Note: although the p-value in the Shapiro test is insignificant (0.037<0.05) which means that the null hypothesis (normality) is rejected but since it is not much less than 0.05 and the Q-Q plot is approximately straight we can assume the errors are distributed normally.

Model 5: includes X2, X4, X6, and X7 as independent variables
Analyzing Results: X7 is insignificant, so it is not the best-fitted model.

Model 6: includes X2, X4, X6, and X7 as independent variables
Analyzing Results: all variables are significant, so assumptions will be checked by appropriate plots (Figure 7) and tests.

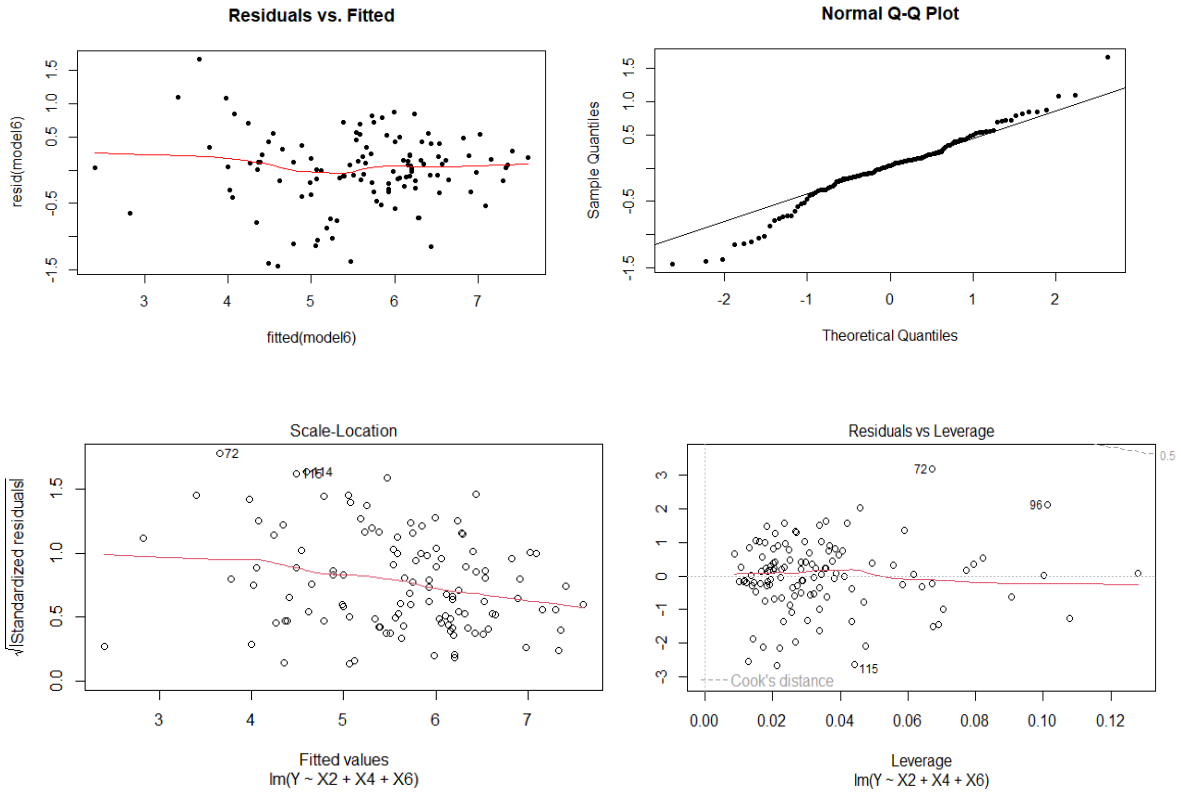

Figure 7

Some measures of the model:

AIC	p-value of Shapiro test	R-squared	RMSE
196.5889	0.02474	0.7653	0.7489101

Results interpretation:

1- Assumption of linearity: The red line in the 'residuals vs fitted' plot is straight at zero and so the assumption is valid.
2- Assumption of normality: The data points in the 'Q-Q plot' seems to form a straight line, so the assumption is true.
3- Assumption of homoscedasticity of variance: Since the line in the 'scale location plot' does not show many fluctuations, and the spread of standardized residuals around the red line does not vary so the assumption is true.
4- No influential points since there are no points outside of Cook's distance in the 'Residuals vs leverage plot' (no need to eliminate the outliers).
5- No multicollinearity
6- The independent variables are 76.53% able to explain the variation of the Y variable.

Model 7: includes X2, X3, X4, and X6 as independent variables
Analyzing Results: all variables are significant, so assumptions will be checked by appropriate plots (Figure 8) and tests.

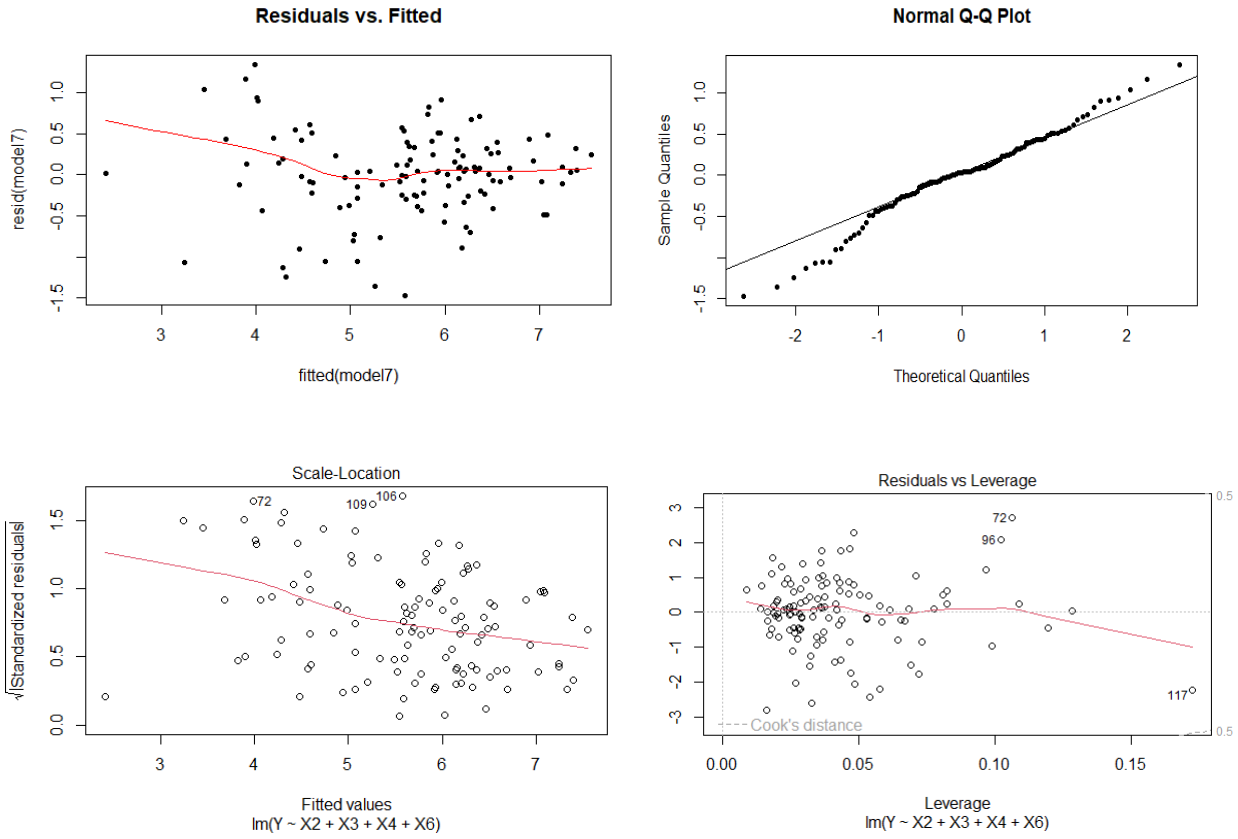

Figure 8

Some measures of the model:

AIC	p-value of Shapiro test	R-squared	RMSE
188.5828	0.04175	0.7826	0.7275301

Results interpretation:

1- Assumption of linearity: the linearity of the relation is somehow weak since the red line in the 'residuals vs fitted' plot is not straight at zero.
2- Assumption of normality: The normality of errors seems to hold since data points in the 'Q-Q plot' are forming approximately a straight line.
3- Assumption of homoscedasticity of variance: The line in the scale location is not horizontal and shows a steep angle so the assumption is not valid since variance is not constant.
4- No influential points since there are no points outside of Cook's distance in the 'Residuals vs leverage' plot (no need to eliminate the outliers).
5- No multicollinearity.
6- The independent variables are 78.26% able to explain the variation of the Y variable.

Comparing between models:

As it is demonstrated above, all the variables in Models 4,6, and 7 are significant so we must choose the best-fitted model among them, and since the assumptions of linearity and homoscedasticity are not valid in model 7, they will be omitted. And between models 4 and 6, model 4 will be chosen for the following reasons:

- AIC of model 4 (= 170.5528) < AIC of model 6 (= 196.5889)
- RMSE of model 4 (= 0.6772187) < RMSE of model 6 (= 0.7489101)
- Model 4 is more capable of explaining the variation of the Y variable since R^2 of model 4 (81.67%) > R^2 of model 6 (76.53%)

Moreover, in affirmation of the obtained results, model 4 (Figure 9 shows its summary) was chosen among all models by the "StepAIC" function in all directions (Backward, Forward, Both).

```
> summary(model4)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X6 + X7 + X9, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47597 -0.23196  0.07083  0.23669  1.18796

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.16936    0.77378  -1.511  0.13360
X1           0.27733    0.06638   4.178 5.92e-05 ***
X2           3.36612    0.60567   5.558 1.93e-07 ***
X4           1.86423    0.62675   2.974  0.00361 **
X6          -1.11551    0.34235  -3.258  0.00149 **
X7           1.75515    0.60282   2.912  0.00436 **
X9          -0.76988    0.33543  -2.295  0.02362 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4831 on 110 degrees of freedom
Multiple R-squared:  0.8261,    Adjusted R-squared:  0.8167
F-statistic: 87.11 on 6 and 110 DF,  p-value: < 2.2e-16
```

Figure 9 summary of model 4 in R studio

Finally, the outliers of model 4 will be tested, where the results are shown in figure 10, it shows that there are 6 outliers in model 4, but none of them is an influential point, so their elimination is not necessary, especially since they are near the interval [2, -2].

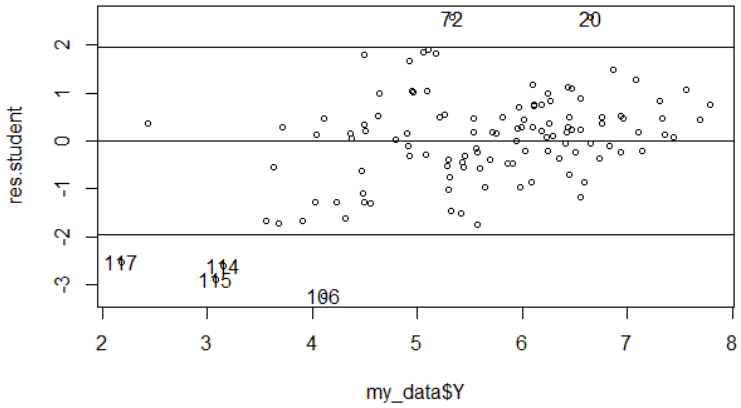

Figure 10 Outliers of model 4

Predictions and validation:

After the diagnosis of model 4, predictions on the test data must take place, the predicted value for each of the 31 records and their lower and upper bounds respectively are obtained in figure 11. While figure 12 shows whether each prediction is true or false.

```
> pred <- predict(model4,newdata=test_data,interval="prediction",level=0.9)
> print(pred)
       fit      lwr      upr
1  7.534793 6.693579 8.376007
2  6.872738 6.058611 7.686865
3  7.124910 6.301290 7.948530
4  7.164951 6.345769 7.984133
5  6.456364 5.640292 7.272436
6  6.279611 5.464532 7.094690
7  6.627378 5.817292 7.437465
8  6.263161 5.450451 7.075871
9  5.971877 5.156089 6.787665
10 6.064524 5.244369 6.884679
11 6.110035 5.293905 6.926166
12 6.196481 5.380466 7.012495
13 5.792242 4.980078 6.604407
14 5.310211 4.496105 6.124318
15 3.862839 3.031056 4.694622
16 6.348648 5.534353 7.162943
17 5.698757 4.873987 6.523528
18 4.332404 3.517963 5.146846
19 6.148200 5.339766 6.956635
20 4.531943 3.715235 5.348651
21 5.399154 4.554809 6.243498
22 5.208796 4.383809 6.033783
23 5.807924 4.984337 6.631512
24 5.366963 4.548364 6.185563
25 5.002931 4.181285 5.824576
26 4.363734 3.546342 5.181127
27 4.552520 3.690053 5.414986
28 3.667128 2.845406 4.488849
29 4.271986 3.427071 5.116902
30 5.473184 4.652020 6.294347
31 2.707412 1.854862 3.559963
```

Figure 11 Predictions of the test data

```
> predicted_or_not <- (pred[,'fit'] >= pred[,'lwr']) & (pred[,'fit'] < pred[,'upr'])
> print(predicted_or_not)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  21   22   23   24   25   26   27   28   29   30   31
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
>
```

Figure 11 checking if the predictions are true

Analyzing the figures:

- Predicted Upper Bound Average – predicted Lower Bound Average = 1.65, which is a rational and accepted range since:
  (Y median – Y 1st quartile) + (Y 3rd quartile – Y median) = 0.8 +0.715 = 1.515.
- Figure 12 shows that all the predicted values are within the 90% confidence interval.

Result:

We can conclude that model 4 is valid, consequently, Log GDP per capita, social support, Freedom to make life choices, Perceptions of corruption, Positive affect, and Confidence in national government variables have a significant impact on the happiness rate according to the following equation:

$$Y = 0.28 * X1 + 3.36 * X2 + 1.87 * X4 – 1.12 * X6 + 1.76 * X7 – 0.77 * X9 – 1.17$$

OR

Happiness rate = 0.28 * Log GDP per capita + 3.36 * social support + 1.87 * Freedom to make life choices – 1.12 * Perceptions of corruption + 1.76 * Positive affect – 0.77 * Confidence in national government – 1.17

Discussion:

After determining the optimal model, it is important to know that excluding variables from the model does not imply that they are insignificant or without effect on the dependent variable. For example, life expectancy (X3) is very influential on the happiness score(Y0, but since it is highly correlated with another independent variable (X1: Log GDP per capita), this variable has more impact on the dependent variable Y, life expectancy (X3) was dropped.

On another hand, the model showed a similarity with Ahtesham's [1] results regarding the importance of social support and life expectancy in determining the happiness score. However, the model conclusion disagrees with HU's [2] conclusion, which underestimated the influence of the GDP on the happiness rate, although they share some results concerning the significant effect of health and social interaction.

Conclusion and Recommendations:

Based on the importance of happiness and what consequences it may have among individuals and societies, it is necessary to take great care in enhancing the predictor controlling it. For example, launching awareness campaigns to promote community solidarity, reshaping the economy in a way that rises productivity –while preserving the mental health of workers -, and raising the budget of health care services are practical procedures to boost the happiness rate.

Lastly, upcoming research efforts must study other factors that may affect happiness, such as family status, social media consumption, and faith in God/atheism to develop a better vision of happiness and the predictors controlling it.

References:

[1] Ahtesham, Sarah (2020). ANALYSING HAPPINESS INDEX AS A MEASURE ALONG WITH ITS PARAMETERS AND STRATEGIES FOR IMPROVING INDIA'S RANK IN WORLD HAPPINESS REPORT. https://ictactjournals.in/paper/IJMS_Vol_6_Iss_1_Paper_5_1170_1173.pdf

[2] HU, ZIMU (2012). Chinese Happiness Index and Its Influencing Factors Analysis. https://www.diva-portal.org/smash/get/diva2:517428/FULLTEXT01.pdf

[3] Alba, Charles (2019). A Data Analysis of the World Happiness Index and its Relation to the North-South Divide. https://digitalcommons.iwu.edu/uer/vol16/iss1/6/

[4] Helliwell, J. F., Layard, R., Sachs, J. D., De Neve, J.-E., Aknin, L. B., & Wang, S. (Eds.). (2022). World Happiness Report 2022. New York: Sustainable Development Solutions Network. https://worldhappiness.report/#:~:text=World%20Happiness%20Report%202022,bright%20light%20in%20dark%20times.

Appendices:

Appendix A:

The variables and how they are calculated by the World Happiness Report 2022 are explained in detail below:

- Happiness score or subjective well-being (variable name ladder): The survey measure of SWB is from the Feb 18, 2022 release of the Gallup World Poll (GWP) covering years from 2005 to 2021. Unless stated otherwise, it is the national average response to the question of life evaluations. The English wording of the question is "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you feel you stand at this time?" This measure is also referred to as the Cantril life ladder or just the life ladder in our analysis.
- The statistics of GDP per capita (variable name GDP) in purchasing power parity (PPP) at constant 2017 international dollar prices are from the December 16, 2021 update of the World Development Indicators (WDI). The GDP figures for Taiwan, Syria, Palestine, Venezuela, Djibouti, and Yemen are from Penn World Table 10.0. – GDP per capita in 2021 are not yet available as of January 2022. We extend the GDP-per-capita time series from 2020 to 2021 using country-specific forecasts of real GDP growth in 2021 first from the OECD Economic Outlook No 110 (December 2021) and then, if missing, forecasts from World Bank's Global Economic Prospects (Last Updated: 01/11/2022). The GDP growth forecasts are adjusted for population growth with the subtraction of 2019-20 population growth as the projected 2020-21 growth.
- Healthy Life Expectancy (HLE). Healthy life expectancies at birth are based on the data extracted from the World Health Organization's (WHO) Global Health Observatory data repository (Last updated: 2020-12-04). The data at the source are available for the years 2000, 2010, 2015, and 2019. To match this report's sample period (2005-2021), interpolation and extrapolation are used.
- Social support (or having someone to count on in times of trouble) is the national average of the binary responses (either 0 or 1) to the GWP 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?"
- Freedom to make life choices is the national average of responses to the GWP question "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
- Generosity is the residual of regressing the national average of response to the GWP question "Have you donated money to a charity in the past month?" on GDP per capita.
- Corruption Perception: The measure is the national average of the survey responses to two questions in the GWP: "Is corruption widespread throughout the government or

not" and "Is corruption widespread within businesses or not?" The overall perception is just the average of the two 0-or-1 responses. In case the perception of government corruption is missing, we use the perception of business corruption as the overall perception. The corruption perception at the national level is just the average response to the overall perception at the individual level.

- Positive affect is defined as the average of three positive affect measures in GWP: laughing, enjoyment, and doing interesting things in the Gallup World Poll waves 3-7. These measures are the responses to the following three questions, respectively: "Did you smile or laugh a lot yesterday?", and "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Enjoyment?", "Did you learn or do something interesting yesterday?"
- Negative affect is defined as the average of three negative affect measures in GWP. They are worry, sadness, and anger, respectively the responses to "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Worry?", "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Sadness?", and "Did you experience the following feelings during A LOT OF THE DAY yesterday? How about Anger?".

Appendix B:

The criteria for dealing with the missing values are: If records from previous years are available, they will be used to replace missing values; otherwise, the average of the column's seven values will be used to fill them in (cells of yellow background in the dataset).

In our data there are 28 missing values with previous records, so they are replaced as follows:

- 10 values were replaced by 2020 records (values in red in the attached dataset)
- 9 values were replaced by 2019 records (values in blue in the attached dataset)
- 9 values were replaced by 2011 or older records (values in green in the attached dataset)

Appendix C:

Summary of all models in R:

```
> summary(model1)

Call:
lm(formula = Y ~ ., data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.51749 -0.20088  0.04926  0.23081  1.11060

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.65066    1.02518  -1.610  0.11032
X1           0.23549    0.08228   2.862  0.00506 **
X2           2.98579    0.73405   4.068 9.11e-05 ***
X3           0.01880    0.01633   1.151  0.25229
X4           1.71125    0.64364   2.659  0.00905 **
X5           0.19801    0.32800   0.604  0.54733
X6          -1.05489    0.34825  -3.029  0.00307 **
X7           1.83525    0.61779   2.971  0.00367 **
X8          -0.17175    0.74206  -0.231  0.81741
X9          -0.71587    0.34505  -2.075  0.04042 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4861 on 107 degrees of freedom
Multiple R-squared:  0.8287,    Adjusted R-squared:  0.8143
F-statistic: 57.53 on 9 and 107 DF,  p-value: < 2.2e-16
```

```
> summary(model2)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X5 + X6 + X7 + X8 + X9, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47224 -0.22538  0.05021  0.22989  1.17437

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.25594    0.96757  -1.298  0.19704
X1           0.28686    0.06923   4.144 6.81e-05 ***
X2           3.32879    0.67184   4.955 2.70e-06 ***
X4           1.84989    0.63322   2.921  0.00424 **
X5           0.17519    0.32789   0.534  0.59424
X6          -1.10024    0.34653  -3.175  0.00195 **
X7           1.75281    0.61455   2.852  0.00520 **
X8           0.08781    0.70802   0.124  0.90152
X9          -0.77664    0.34150  -2.274  0.02493 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4868 on 108 degrees of freedom
Multiple R-squared:  0.8266,    Adjusted R-squared:  0.8138
F-statistic: 64.36 on 8 and 108 DF,  p-value: < 2.2e-16
```

```
> summary(model3)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X6 + X7 + X8 + X9, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47845 -0.23026  0.06347  0.23779  1.20115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.23372    0.96351  -1.280  0.20311
X1           0.27811    0.06704   4.149 6.65e-05 ***
X2           3.39454    0.65831   5.156 1.13e-06 ***
X4           1.86757    0.63028   2.963  0.00374 **
X6          -1.11662    0.34404  -3.246  0.00156 **
X7           1.76524    0.61209   2.884  0.00473 **
X8           0.07975    0.70554   0.113  0.91021
X9          -0.76502    0.33969  -2.252  0.02632 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4852 on 109 degrees of freedom
Multiple R-squared:  0.8262,    Adjusted R-squared:  0.815
F-statistic:    74 on 7 and 109 DF,  p-value: < 2.2e-16
```

```
> summary(model4)

Call:
lm(formula = Y ~ X1 + X2 + X4 + X6 + X7 + X9, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47597 -0.23196  0.07083  0.23669  1.18796

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.16936    0.77378  -1.511  0.13360
X1           0.27733    0.06638   4.178 5.92e-05 ***
X2           3.36612    0.60567   5.558 1.93e-07 ***
X4           1.86423    0.62675   2.974  0.00361 **
X6          -1.11551    0.34235  -3.258  0.00149 **
X7           1.75515    0.60282   2.912  0.00436 **
X9          -0.76988    0.33543  -2.295  0.02362 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4831 on 110 degrees of freedom
Multiple R-squared:  0.8261,    Adjusted R-squared:  0.8167
F-statistic: 87.11 on 6 and 110 DF,  p-value: < 2.2e-16
```

```
> summary(model5)

Call:
lm(formula = Y ~ X2 + X4 + X6 + X7, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50076 -0.22929  0.08559  0.25429  1.68412

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06101    0.55412  -0.110   0.9125
X2           5.67563    0.48252  11.763  < 2e-16 ***
X4           1.56237    0.64396   2.426   0.0169 *
X6          -1.19941    0.29325  -4.090 8.15e-05 ***
X7           1.13536    0.64232   1.768   0.0799 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5415 on 112 degrees of freedom
Multiple R-squared:  0.7776,    Adjusted R-squared:  0.7696
F-statistic: 97.89 on 4 and 112 DF,  p-value: < 2.2e-16
```

```
> summary(model6)

Call:
lm(formula = Y ~ X2 + X4 + X6, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.44314 -0.24996  0.03605  0.30984  1.67025

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.06326    0.55931  -0.113  0.91015
X2           5.96424    0.45830  13.014  < 2e-16 ***
X4           2.18891    0.54265   4.034  0.00010 ***
X6          -1.17158    0.29556  -3.964  0.00013 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5465 on 113 degrees of freedom
Multiple R-squared:  0.7714,    Adjusted R-squared:  0.7653
F-statistic: 127.1 on 3 and 113 DF,  p-value: < 2.2e-16
```

```
> summary(model7)

Call:
lm(formula = Y ~ X2 + X3 + X4 + X6, data = my_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.46626 -0.24219  0.03594  0.31500  1.34115

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.95280    0.80423  -2.428  0.01677 *
X2           4.73262    0.58842   8.043 1.00e-12 ***
X3           0.04144    0.01311   3.162  0.00201 **
X4           2.22821    0.52240   4.265 4.19e-05 ***
X6          -0.97342    0.29127  -3.342  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.526 on 112 degrees of freedom
Multiple R-squared:  0.7901,    Adjusted R-squared:  0.7826
F-statistic: 105.4 on 4 and 112 DF,  p-value: < 2.2e-16
```