King Abdullah University of Science and Technology, CEMSE, STAT

Project Report:

World Happiness

By:

Mohammed Al Saleem 188301

Abstract

Even when the happiness is mainly under the individual control, but the outer world can affect it. Like what happening to the country and the social live and culture. World Happiness Report team did a great job collecting data and making surveys to give every country a happiness score, and they providing the dataset that used. Multiple linear regression used to fit a model that can predict the happiness score of countries. The model depends on the social support, life expectancy, freedom, and corruption perception. Also, to classify the countries to happy and not happy, logistic model used where the dependent variable became binary. The logistic model depends on the social support, life expectancy, and freedom.

Table of Contents

Table of Figures

INTRODUCTION

Happiness is not just a good feeling and a smiley face. It is the feeling of enjoying life and satisfaction, and the desire to make the best of your life. Happiness is what make us be and do our best.

Our happiness is under our control. Researches and psychologists found that happiness mainly depends on our mindset, our habits, and our way of live. So, everyone can be happy if he/she decide!

Data Description

The dataset is about the world happiness of 146 countries to 158 countries, from 2015 to 2022. The dataset is taken from Kaggle (kaggle.com/datasets/mathurinache/world-happiness-report).

The world happiness report calculated the happiness score from variance sources, like statistical data of logged GDP (Gross domestic product) per capita for economy, and a national survey of the main life evaluation questions using the Cantril Ladder in the Gallup World Poll (10 for the best possible life, and 0 for the worst). The formula of GDP per capita = Real GDP of the country / Population of the country.

From these sources we get the following factors:

- Economy: Logged GDP per capita
- Health: Life Expectancy (based on study that use historical data along with the previous survey)
- Family: Social support (from 0 to 1)
- Freedom: to make life choices (from 0 to 1)
- Trust: Corruption Perception (from 0 to 1)
- Generosity (from -1 to 1)

The data of most years are explained by something called pooled OLS regression, which make it unsuitable to model. But fortunately, the original data of year 2021 are available and suitable to create a model.

The scientific goals

- Explore the dataset across the years and regions, and if COVID-19 left an effect on world's happiness.
- investigate the collinearity and the correlation between factors.
- Create a model from the original data of year 2021.
- Simple prediction for happy and non-happy countries (logistic regression).

STATISTICAL METHODS

First investigate data the correlation between the variables. Collinearity and Multicollinearity happens when there are regressors are linearly dependent, like if there are exist constants c1, c2 and c3, not all equal to zero, such that $c_1X_1 + c_2X_2 = c_3$.

Multicollinearity is a problem where having highly correlated independent variables ($\rho(X,Y) = cov(X,Y) / (\sigma X \sigma Y)$) cause significant fluctuate in the model results if one of the correlated variables changed, which make the model unstable and vary due to any small change.

To check multicollinearity, the first method is to plot the correlation of all the variables. Good function for this is "ggpairs" from "GGally" library. The second method is to use VIF (Variance Inflation Factor). It measures the multicollinearity of multiple regression variables in a model. The higher VIF value for variable means higher correlation between it and the rest variables.

The methodology to fix the Multicollinearity issue is to remove one or some variables that have high correlation and VIF.

Where the dataset has multiple independent variables, Multiple Linear Regression (MLR) is used to fit a model. The basic MLR model is:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon_i$$

where y_i the dependent variables, $x_1, \dots, x_n$ are the independent variables, ε_i random error, $\beta_0, \dots, \beta_n$ are the coefficients that will be estimated.

As arrays: $Y = \beta X + \varepsilon$

After fitting the model with all variables (**lm(Happiness.Score~. , data = Od2021)**), backward stepwise selection will be used to select the final model that have all variables significant. Deleting biggest p-values that not significant (a=0.05) one by one.

For comparing the models adjusted R^2 is used instead of R^2, since R^2 always increase with adding variables to the model.

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

For residual analysis, the used methods are diagnostic plots, residual plots, Shapiro test, and non-constant variance test. To test normality, linearity, homogeneity of the variance, and cook's distance.

To classify countries to happy and not-happy (above or below the mean of hippieness score) we need to use logistic regression, where the dependent variable became binary variable.

Logistic regression model defined as:

$$\pi(\boldsymbol{x}) = \frac{e^{\beta_0 + \boldsymbol{\beta_1 x}}}{1 + e^{\beta_0 + \boldsymbol{\beta_1 x}}}.$$

Also, it can be written as:

$$g\big(\pi(\boldsymbol{x})\big) = logit\big(\pi(\boldsymbol{x})\big) = log(odds) = \log\left(\frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})}\right) = \beta_0 + \boldsymbol{\beta_1 x}.$$

This transformation is called the logit transformation or log-odds. Also, in logistic regression the expected value of the dependent variable given $\boldsymbol{x}$ is between 0 and 1 ($0 \leq E(y|\boldsymbol{x}) \leq 1$). Note, in the simple linear regression (i.e., $y = \beta_0 + \boldsymbol{\beta_1 x}$) one unit change in $\boldsymbol{x}$ gives a change in y by $\boldsymbol{\beta_1}$ units. On the other hand, in the logistic regression changing one unit in $\boldsymbol{x}$ does not change the value of y directly, but gives a change in the log-odds by $\boldsymbol{\beta_1}$.

Same as MLR, after fitting the model with all variables (**glm(happy~. , data = Ld2021, family = binomial)**), backward stepwise selection used to select the final.

To measure the goodness of fit, AIC, ROC, and Confusion Matrix are used. AIC (Akaike Information Criteria), it is an analogous to adjusted R^2, where the model with a minimum AIC value is preferred.

$$AIC = -2l + 2(p + 1),$$

where l is the maximized log likelihood of the model.

ROC (Receiver Operating Characteristic) curve is a plot with y-axis of the sensitivity values and against x-axis with 1-specificity values, and evaluated at different threshold values from 0 to 1.

AUC is a performance measure of a model across all possible classification thresholds and equal to the area under the ROC curve. 1 represent the perfect classifier, and 0.5 represent the random classifier.

Confusion matrix is a tabular representation of observed vs predicted data, and used to quantify the accuracy of the model. Specificity is the proportion of correct identified of negatives. Sensitivity is the proportion of correct identified of positives.

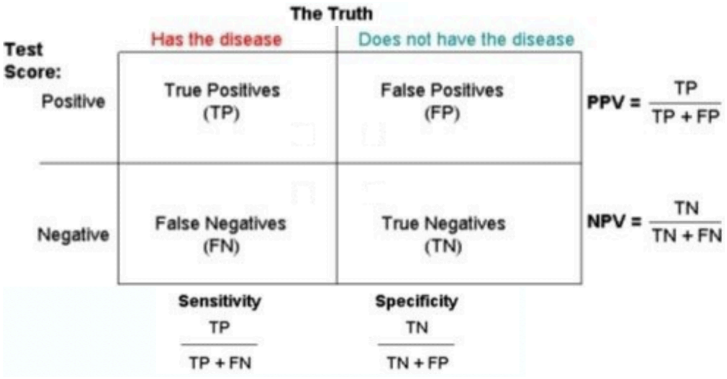

Figure 1. The Confusion matrix

A probability threshold (c) used to classify data as $\hat{Y}=1$ if $\hat{\pi} > c$ and $\hat{Y} = 0$ if $\hat{\pi} \leq c$.

APER = (FN + FP)/(TN+TP+FN+FP)

Accuracy = 1-APER = (TN + TP)/(TN+TP+FN+FP)

APER (Apparent Error Rate) is the fraction of misclassified observations by the classification function.

RESULTS

Exploring data, we found that Finland have the highest hippieness score from 2018 until now which make it the happiest country. On the other hand, Burundi was one of the least happy two countries from 2015 to 2018, and from 2020 to now Afghanistan is the lowest happy country.

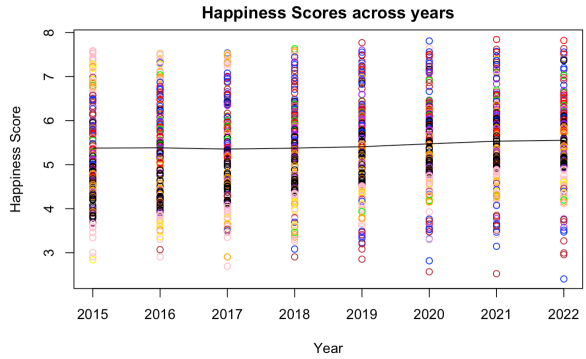

Figure 2. Happiness Scores across years, colored based on regions

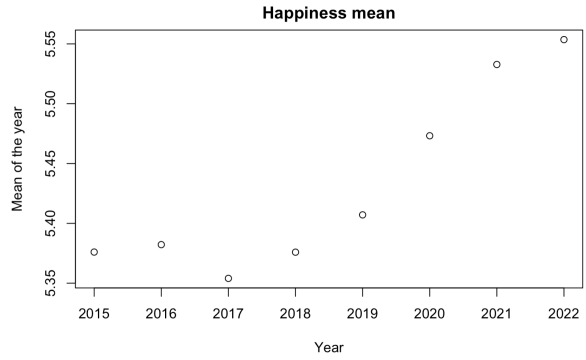

Figure 3. Happiness means across years

From Figure 2 and Figure 3 we see there are no significant difference between the years except small non-linear increasing.

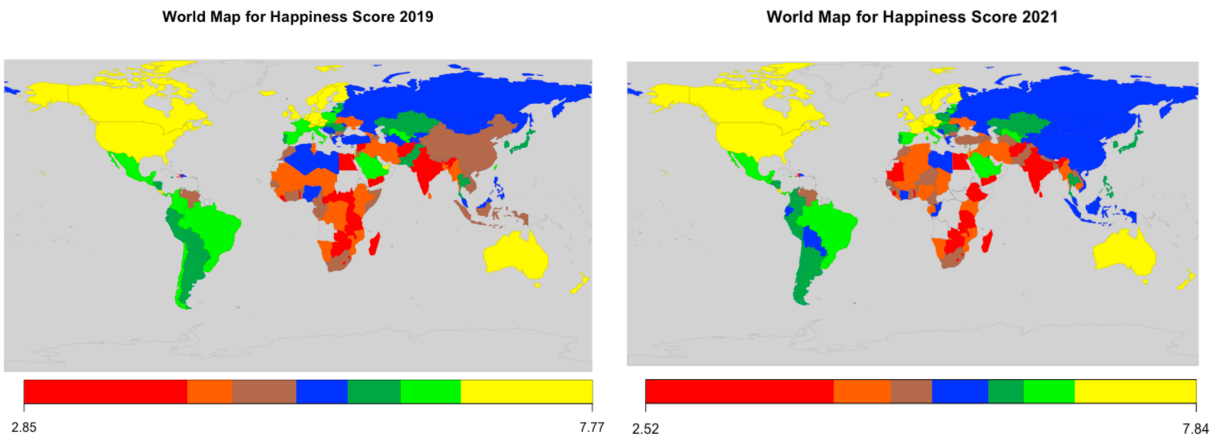

Figure 4. World map for Happiness score in 2019 and 2021 (Before and after COVID-19)

There is no sign that COVID-19 left an effect on world's happiness.

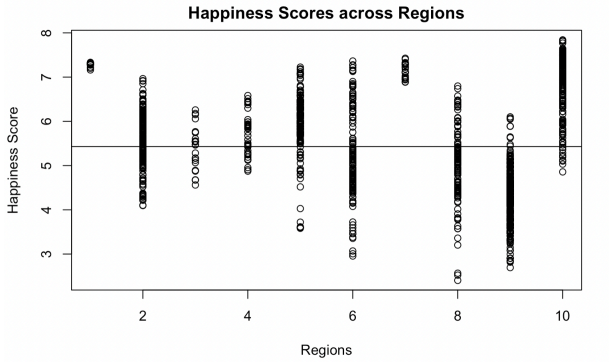

Figure 5. Happiness Scores across Regions

From Figure 5, Australia and New Zealand (1), North America and ANZ (7) always had a high happiness score. Sub-Saharan Africa (9) vary from average countries to sad countries. Opposite to Sub-Saharan Africa, Western Europe (10) vary from average countries to happy countries.

Regions will be added to the model to test if their effect is significant.

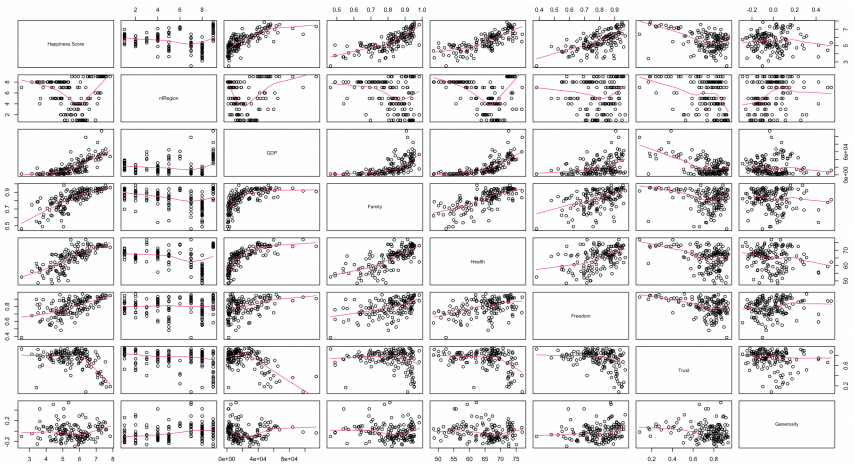

Figure 6. Dataset Scatter Matrix

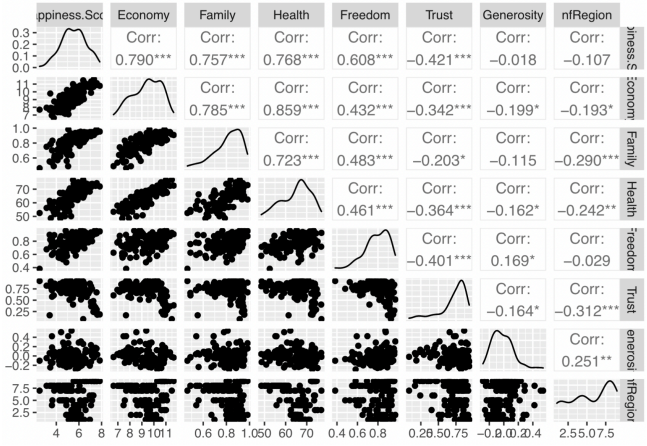

Figure 7. Correlation Scatter Matrix

By illustrating the happiness scores against the variables (Figure 6), we can see GDP need to be transformed by logged (to Economy variable). Also, from Figure 6 there is collinearity. The high correlations are Economy:Health (0.859), Economy:Family (0.785), and Family:Health (0.723).

VIF values of model0: Economy (5.1656), Family (3.0635), Health (4.2707), Freedom (1.5864), Trust (1.5961), Generosity (1.2097), nfRegion (1.3929). Economy have the highest VIF, so Economy will be deleted from the model. After deleting it, VIF of Family and Health decreased to 2.3476 and 2.5395 respectively, which are reasonable.

Furthermore, Figure 6 shows that Family have curved fit against the happiness, and transform it by square it gave a little higher adj. R2.

Continuing backward selection, we end with:

Happiness.Score ~ Family^2 *+ Health + Freedom + Trust – 1*, with adj. $R^2 = 0.9904$. While model0's adj. $R^2 = 0.7303$

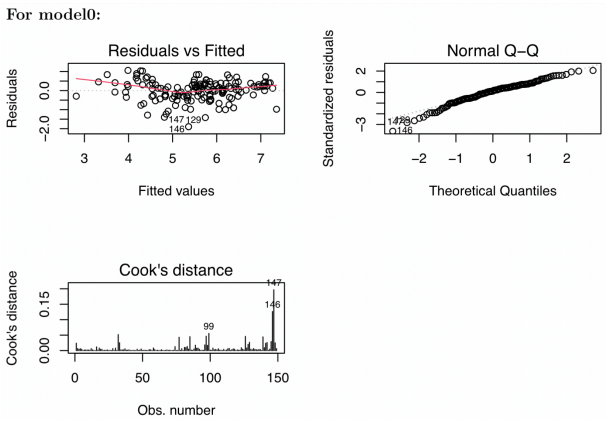

Figure 8. Diagnostic plots for model0

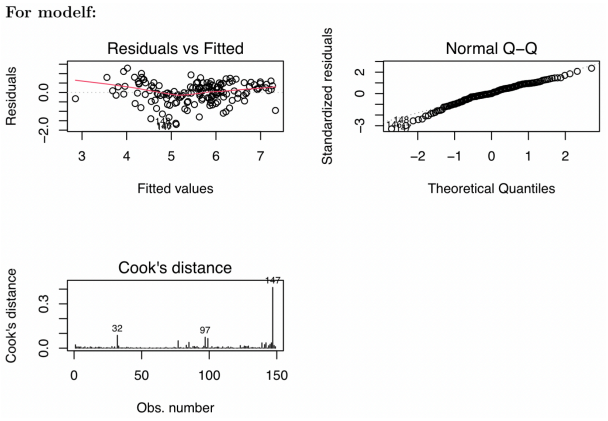

Figure 9. Diagnostic plots for modelf

Shapiro-Wilk normality test model0: W = 0.968, p-value = 0.0015
Shapiro-Wilk normality test modelf: W = 0.9746, p-value = 0.0073
Non-constant Variance Score Test model0: Chi-square = 5.1996, p = 0.0226
Non-constant Variance Score Test modelf: Chi-square = 5.5263, p = 0.0187
And for anova(model0, modelf), p = 0.05657

Anova fail to reject modelf different than model0, but modelf have fewer variables and all of them are significant, and higher adj.R^2. Moreover, Shapiro test rejected the null of normality for the model, and ncvTest also rejected the null of constant variance, but it still better than other models.

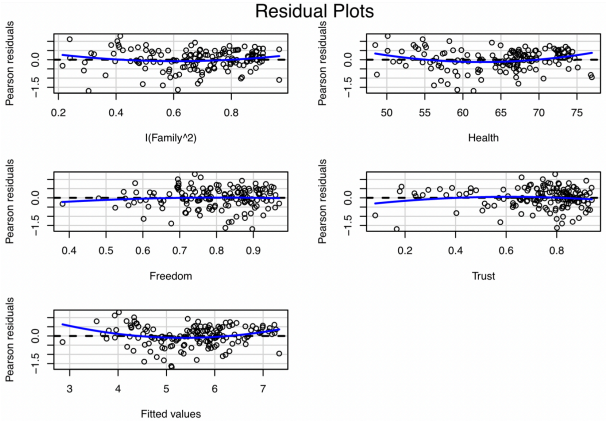

Figure 10. Residual Plots

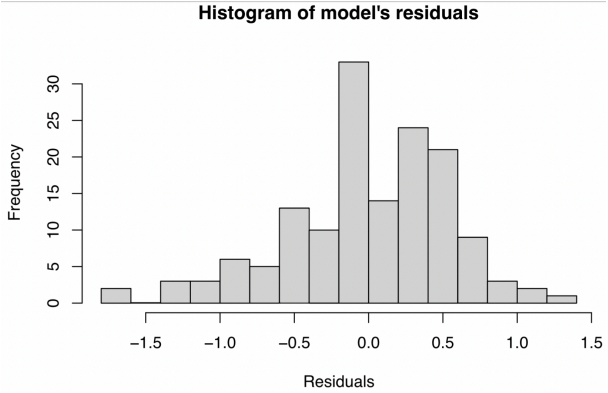

Figure 11. Residual Histogram of model's residuals

Residual plots, and the histogram of residuals give acceptable graphs.

The Final model:

$$Happiness.Score = 2.5624 \times Family^2 + 0.0467 \times Health + 1.7974 \times Freedom - 0.9048 \times Trust$$

Logistic Regression: To classify countries to happy and not-happy:

Fitting Lmodel0 with all variable (**glm(formula = happy~. , data = Ld2021, family = binomial**) returns multiple variables not significant. AIC = 89.4376

Again, the backward stepwise selection applied, $AIC_{(Lmodel1)} = 87.5586$, $AIC_{(Lmodel2)} = 85.7576$, AIC (Lmodel3) = 84.9234, until we end with $AIC_{(Lmodel4)} = 84.2212$

Lmodel4 variables are all significant, and it have better results of AIC compared with Lmodel0.

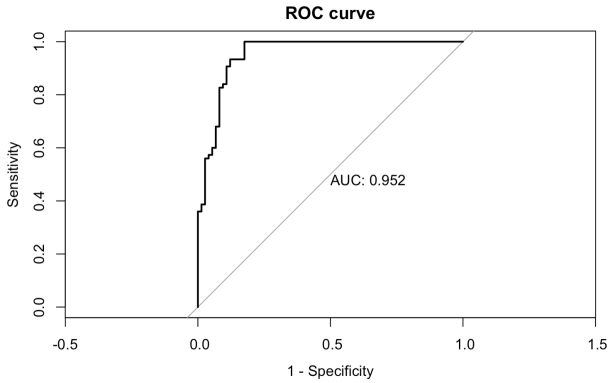

Figure 12. The ROC plot and AUC

From the ROC plot curve and the high value of AUC (0.952), we conclude that the Lmodel4 have a good fit.

The final model (Lmodel4):

$$log(odds) = -44.3549 + 18.3826 \times Family + 0.3361 \times Health + 8.4726 \times Freedom$$

Confusion matrix with probability threshold c = 0.5:

Lmodel4.pred	0	1
No	64	5
Yes	10	70

Accuracy = 1-APER = 0.8993
Sensitivity = 0.8649
Specificity = 0.9333

Figure 13. Lmodel4 confusion matrix

All Accuracy, Sensitivity, and Specificity have big value which tell the predict is good.

DISCUSSION AND CONCLUSIONS

This report went through analyzing the World Happiness Report dataset to understand it and to come out with a model that can predict the happiness score with high level of accuracy.

The data consisted of multiple independent variables (Family, Health, …etc.) and one dependent variable (Happiness.Score). For further analysis it is important to choose the most appropriate model. Since there are multiple independent variables, with one dependent variable with continues range, it is

obvious that MLR is the most appropriate model in our case. After backward stepwise selection, the final model with significant variables is depending on four independent variables. That are Family, Health, Freedom, and Trust.

(Happiness.Score = 2.5624×Family^2 + 0.0467×Health + 1.7974×Freedom - 0.9048×Trust)

This model has some problems like that the normality hypotheses is rejected, and the graph of residuals is a little curved. Moreover, happiness of a country could be affected by much more variables. If World Happiness Report team provided more detailed data, like the whole survey and the methodology of calculating the life expectancy, it will be possible to study the dataset deeper to get better model.

Classifying countries to happy and not-happy by Logistic Regression went great. Where it minimized AIC, all variables are significant, and have high scores for AUC (0.952), Accuracy (0.8993), Sensitivity (0.8649), and Specificity (0.9333)

$$\boldsymbol{log(odds) = -44.3549 + 18.3826 \times Family + 0.3361 \times Health + 8.4726 \times Freedom}$$

Finally, collecting the data is just started recently (2015), in the future there will be larger dataset that will be enough to fit a model that can predict the countries' happiness for next years.