Linear Relationship Between Happiness and Region, Economic Production, Social Support, Life Expectancy, Freedom, Corruption, and Generosity

Sifei Li
Department of statistic, University of Toronto
STA302: Methods of Data Analysis 1
Professor Katherine Daignault
December 17, 2021

I. Introduction

There are many external factors such as the living environment could affect happiness beside internal emotions. As people pay more attention on pursuing happiness, what external factors impact their happiness and what they can do to improve happiness become important. Since 2012, the World Sustainable Network began to analyze the relationship between happiness and the "6 keys" including Economic Production (evaluated by logged GDP per capita), Social Support, Life Expectancy, Freedom to make life choices, Perception of Corruption, and Generosity using data on Gallup world survey in social science and environmental science point of view (Helliwell et al., 2020). The goal of this study is to statistically show a simple reliable formula that explains how the happiness score is related to some or all the "6 keys" plus regions.

II. Method

1 response and 7 predictors indicated in previous section in the row dataset were selected. 60% observations were randomly selected to form a training dataset, and the rests formed a test dataset without replacement. The reason for this 60:40 ratio of dividing data was to provide enough data for building the model.

All possible subset method was used to build 6 models with 6 different number of predictors using numerical variables only, because including categorical variables would increase the complexity of model and would not meet the goal of study. Comparing to other selection methods, this method was chosen because the number of predictors were small enough to fit and compare all possible models. The selected models were compared to each other by adjusted R^2, multicollinearity and number of significant predictors. Multicollinearity was evaluated by variance inflation factor (VIF) of each predictor, with a cut-off 5 for severe issue. Categorical variables were ignored during calculating VIF. Significant predictors were indicated by p-value of t test smaller than 0.05. For selected models, adding back categorical variable was considered in each model and partial F tests were conducted between all models and the full model with cut-off p-value 0.05. Some reasonable interacting terms were also added based on content of variables.

Model diagnostic was conducted on all selected models. Condition 1 and 2 were checked using response vs. fitted value plot and scatterplots between each pair of predictors respectively. Condition 1 held if the dots were randomly distributed around function y=x. Condition 2 held if there were not any strange non-linear pattern between any pair of predictors. The assumption of linearity, uncorrelated errors, constant variance, and normality were assessed by residual vs. all predictors, residual vs. fitted value plots and normal QQ plots. Standard residuals were used due to potential existence of problematic points. If both conditions held, systematic patterns in plots were identified for violation of linearity assumption, fanning patterns were identified for violation of constant variance assumption and obvious separated groups in plots were identified for violation of uncorrelated errors assumption. If any of the conditions did not hold, non-uniform patterns in residual plots only indicated the existence of a problem. Normality assumption held if dots on QQ plots form a straight line with formula y=x. BoxCox transformation was used for violation of normality or linearity. Reasonable transformations were needed if any other violations found. Categorical variables were ignored during drawing residual plots and checking assumptions. Problematic points including leverage points with cut-off $2\frac{p+1}{n}$, outliers with cut-off $[-2,2]$ (as the dataset was small), and influential points to regression surface with cut-off $F(p+1, n-p-1)$, to fitted value with cut-off $2\sqrt{\left(\frac{p+1}{n}\right)}$ and to slopes with cut-off $[-\frac{2}{\sqrt{n}}, \frac{2}{\sqrt{n}}]$ were identified, where n was the number of observations and p was the number of predictors in the model. Points with value exceeded the cut-off or out of the range would be problematic. Removing these points were considered if appropriated.

In model validation, models with the same formulas were fitted using test dataset and were compared to the ones using training data by adjusted R^2, signs and values of estimated regression coefficients, significance of variables and violation of assumptions. The models were validated if similar properties resulted using the 2 different datasets and no worse assumption violations or multicollinearity appeared. 2 standard error was used in evaluating the similarity of coefficients as a cut-off. A final model was selected comprehensively plus consideration of Akaike's Information Criterion (AIC).

III. Results

1. Exploratory Data Analysis (EDA)

The analysis began on splitting dataset to train and testing dataset randomly, numerical summaries (refer Appendix table 4) did not show a big difference between 2 datasets. In the EDA of training dataset, the ladder score is a subjective score up to 10. Except logged GDP per capita (unit: dollars), life expectancy (unit: age) and generosity (residuals), other variables are measured in percentage therefore 1 is upper limit. The distribution of ladder score was slightly left skewed and the distributions of other 6 numerical variables were also slightly skewed, which indicated a potential violation of linearity and normality assumption. Figure 1.9 - 1.14 illustrated the relationship between ladder score and each numerical predictor was linear.

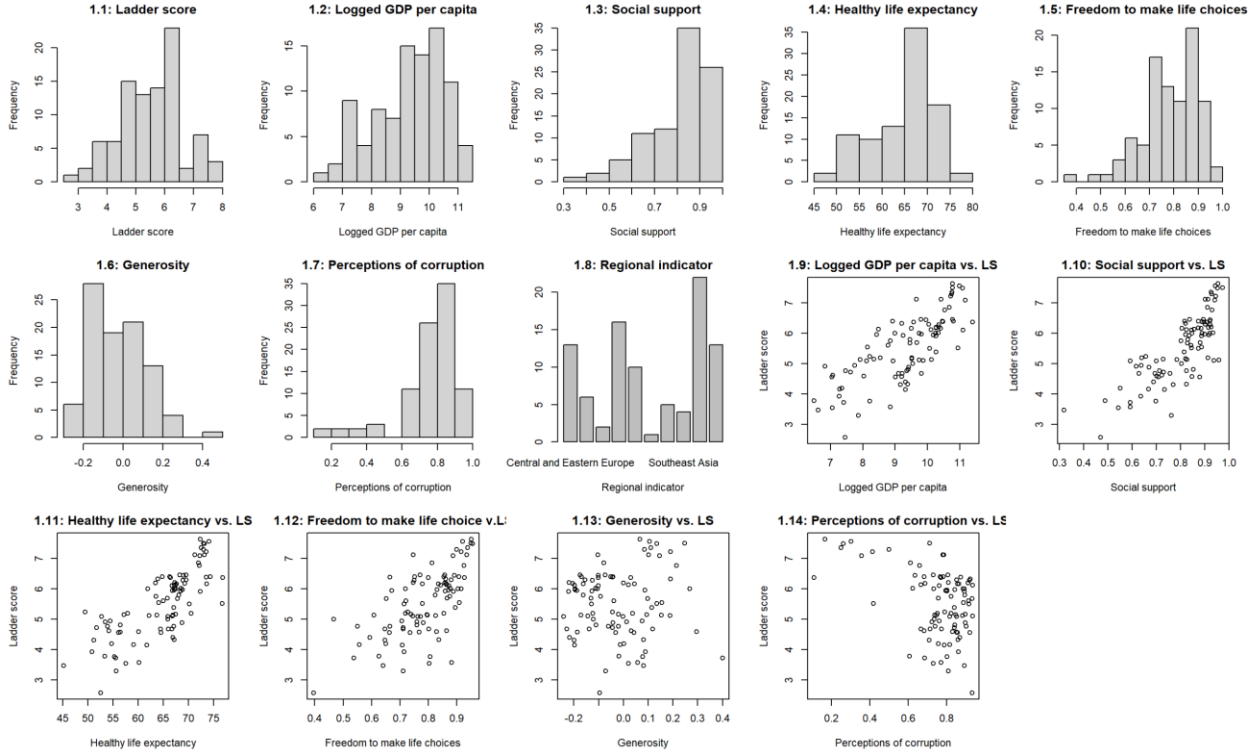

Figure 1

Distributions of ladder score and 7 predictors, and scatterplots between ladder score and each of 6 numerical factors.

To avoid overfitting, I merged "regional indicator" by states. The number of categories reduced from 10 to 5, named "regional2".

2. Model selection

Using all subset selection, I got 6 models with 6 different number of numerical predictors. The summary of the models was indicated in table 1. By comparing, mod1, mod2 and mod3 were

eliminated by low adjusted R^2. Mod4 was selected by high adjusted R^2, no severe multicollinearity, all significant predictors and passed partial F test (p-value 0.0003). mod5 and mod6 were eliminated by severe multicollinearity and non-significant variables. I also tried adding regional2 to mod4, but it did not pass partial F test. Further, I interacted regional2 with the slope of "freedom to make life choices" in mod4, because changing both intercept and slope would lead an over fitted model. The model with interactive term increased adjusted R^2 to about 81% were selected and named mod4_inter.

Table 1:

Summary statistics of models from all possible subsets model selection

model name	number of predictors	adjusted R square	Number of severe multicollinearities	number of significant predictors
mod1	1	62.96%	-	1
mod2	2	69.91%	0	2
mod3	3	74.36%	0	3
mod4	4	76.00%	0	4
mod5	5	76.03%	1	4
mod6	6	75.75%	1	4

3. Model diagnostic

In model diagnostic, I firstly checked condition 1 and condition 2 (refer Appendix figure 3). The plots showed both 2 models satisfied 2 conditions well. From the residual plots (refer figure 2), there was not any obvious systematic pattern, fanning pattern or separated groups of dots, which showed the 3 assumptions held. Furthermore, normality assumption reasonably satisfied in mod4 by normal QQ plot because the 2 endings for mod4 were slightly tilted, consistent with the observation in EDA. A transformation that squared "ladder score" was done to normalize the distribution. 2 new models were fitted by the same predictors and squared ladder score. There were not any new issues appeared by looking at their statistics and conditions (plots refer Appendix figure 3). In residual plots shown in figure 2, the normal QQ plots became straighter, and no new violations appeared.

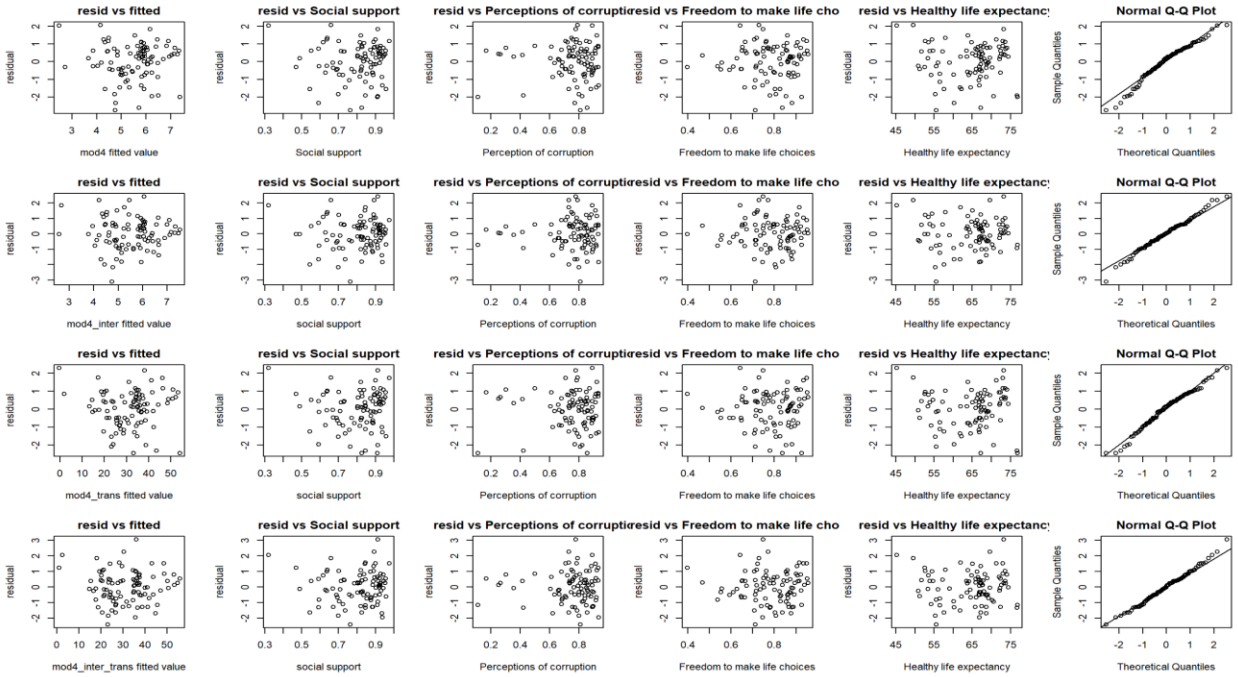

Figure 2:
Residual plots and normal QQ plots of mod4, mod4_inter, mod4_trans and mod4_inter_trans respectively.

Problematic points were identified by different cut-offs. From table 2, all models have similar number of leverage points, outliers, and influential points. Mod4_inter had more influential points, but within a reasonable range. Overall, each model had about 20% problematic observations, therefore removing was not considered.

Table 2:
Summary of problematic points for 5 models.

Model name	# Leverage points	# Outliers	# Influential points on regression surface	# Influential points on fitted value	# Influential points on coefficients
mod4	6	6	0	7	6
mod4_inter	6	5	0	10	4
mod4_trans	2	6	0	5	8
mod4_inter_trans	0	4	0	7	6

4. Model validation

4 new models were fitted using same formulas and test dataset. Overall, the comparable estimated coefficients did not change too much except 1 slope in mod4_trans. The significance of predictors in all models dropped and some appeared as non-significant in test model. In terms of adjusted R^2, mod4_trans reduced 5% only, mod4_inter_trans and mod4 reduced 8%, mod4_inter reduced 11%. Model diagnostic was also conducted on all test models. No new severe assumption violation or multicollinearity appeared (refer Appendix figure 4). Number of problematic observations slightly decreased for all models. Overall, none of the models were validated due to the drop in predictor significance.

5. Final model

The summary of 4 models was shown in table 3, mod4_inter was selected to be the final model because it held assumptions, had higher adjusted R^2 and smallest AIC, even though it did not validate model perfectly.

Table 3:

Summary of statistics of 4 models

Model name	Assumptions and conditions	# Problematic point	Adjusted R^2	AIC
mod4	A little violation in normality	25	75.99%	154
mod4_inter	Hold	25	81.05%	136
Mod4_trans	Hold	25	74.84%	598
Mod4_inter_Trans	Hold	23	80.98%	576

IV. Discussion

The final model had the formula

$$\text{Laddar score} = \beta_0 + \beta_1 \cdot \text{Social support} + \beta_2 \cdot \text{Freedom to make life choices} \cdot \text{regional2} + \beta_3 \cdot \text{Perceptions of corruption} + \beta_4 \cdot \text{health life expectancy} + \text{error},$$

where $\beta_0 = -0.90, \beta 1 = 3.12, \beta_2 = \begin{pmatrix} 2.42\ \mathbb{1}_{Aferica} \\ 1.78\ \mathbb{1}_{Asia} \\ 1.91\ \mathbb{1}_{Commonwealth} \\ 2.73\ \mathbb{1}_{Europe} \\ 2.69\ \mathbb{1}_{North\ and\ South\ America} \end{pmatrix}, \beta 3 = -0.94, \beta 4 = 0.04$ are regression coefficients, they are all significant. β_2 is a vector because each number represents the slope of "freedom to make life choices" on the specific region only. For example, with all other variables remaining unchanged, if a country in Asia increases 1% in freedom to make life choices, its happiness score is expected to increase 0.0178. Only perception of corruption is disproportional to happiness, others' associations are all positive.

There are 3 limitations in this model. Firstly, the model was not validated. The adjusted R^2 dropped in test model and many predictors appeared less significant. The reason may be the existence of problematic observations in both datasets. Secondly, the model had many influential points, which potentially stretches the regression surface. Both limitations are hard to fix because removing these points would cause a big change in data statistics. Both limitations lead a potential problem that the model may not be generalizable on other data. Finally, I ignored the existence of categorical variable during assessing VIF. This may omit existence of multicollinearity between regions and other factors and results in an unstable regression surface. More actions on multicollinearity such as assessing GVIF could be done.

Even though it has limitations, this formula is simple enough to show the linear relationship between happiness ladder score and other significant factors because it does not include any transformations, it is reliable because it satisfies all assumptions therefore the estimates are likely to be unbiased. It shows that only these 4 factors in the "6 keys" significantly contribute to happiness. People who live in the countries included in the training dataset would increase their expectation on happiness score by decreasing perception of corruption or increasing the other 3 factors.

Appendix

Table 4:

Summary statistics in training and test dataset, train size 92, test size 61.

Variable	mean (s.d.) in training	mean (s.d.) in test	categories in categorical variable	number in training	number in test
Ladder score	5.507 (1.1)	5.422 (1.138)	Central and Eastern Europe	13	4
Logged GDP per capita	9.318 (1.219)	9.262 (1.184)	Commonwealth of Independent States	6	6
Social support	0.805 (0.129)	0.815 (0.109)	East Asia	2	4
Healthy life expectancy	64.719 (7.121)	64.034 (7)	Latin America and Caribbean	16	5
Freedom to make life choices	0.784 (0.114)	0.783 (0.124)	Middle East and North Africa	10	7
Generosity	-0.021 (0.138)	-0.005 (0.171)	North America and ANZ	1	3
Perceptions of corruption	0.756 (0.172)	0.699 (0.175)	South Asia	5	2
			Southeast Asia	4	5
			Sub-Saharan Africa	22	17
			Western Europe	13	8

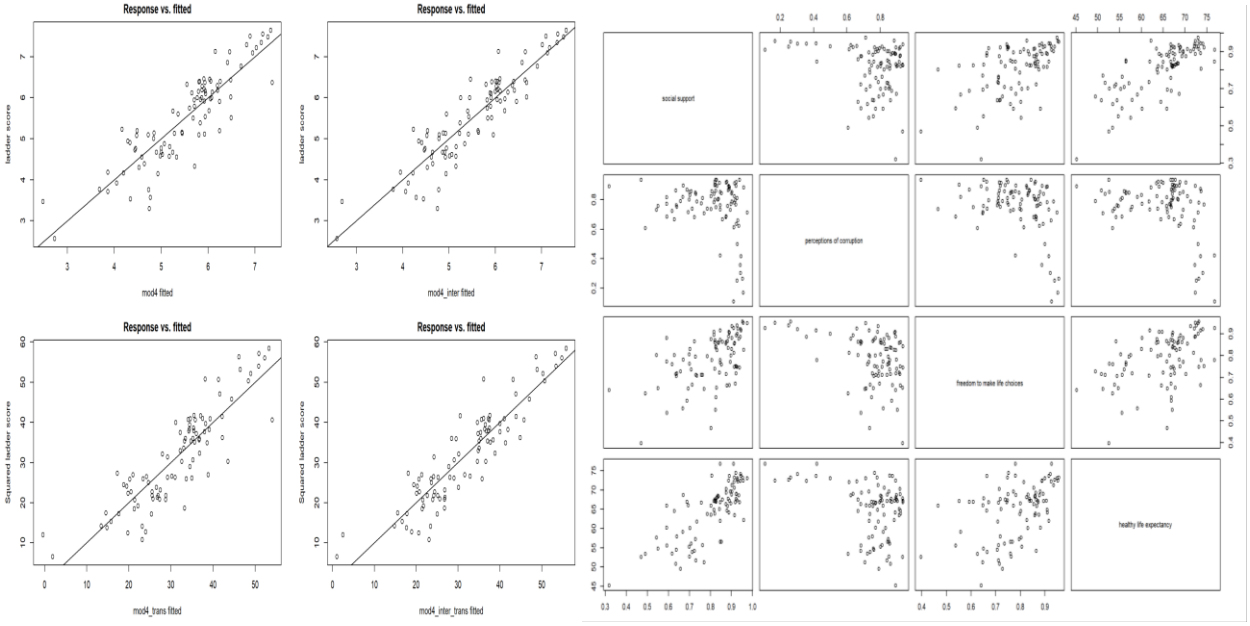

Figure 3:
Condition 1 (left) for mod4, mod4_inter, mod4_trans and mod4_inter_trans respectively, and condition 2 (right) for both models.

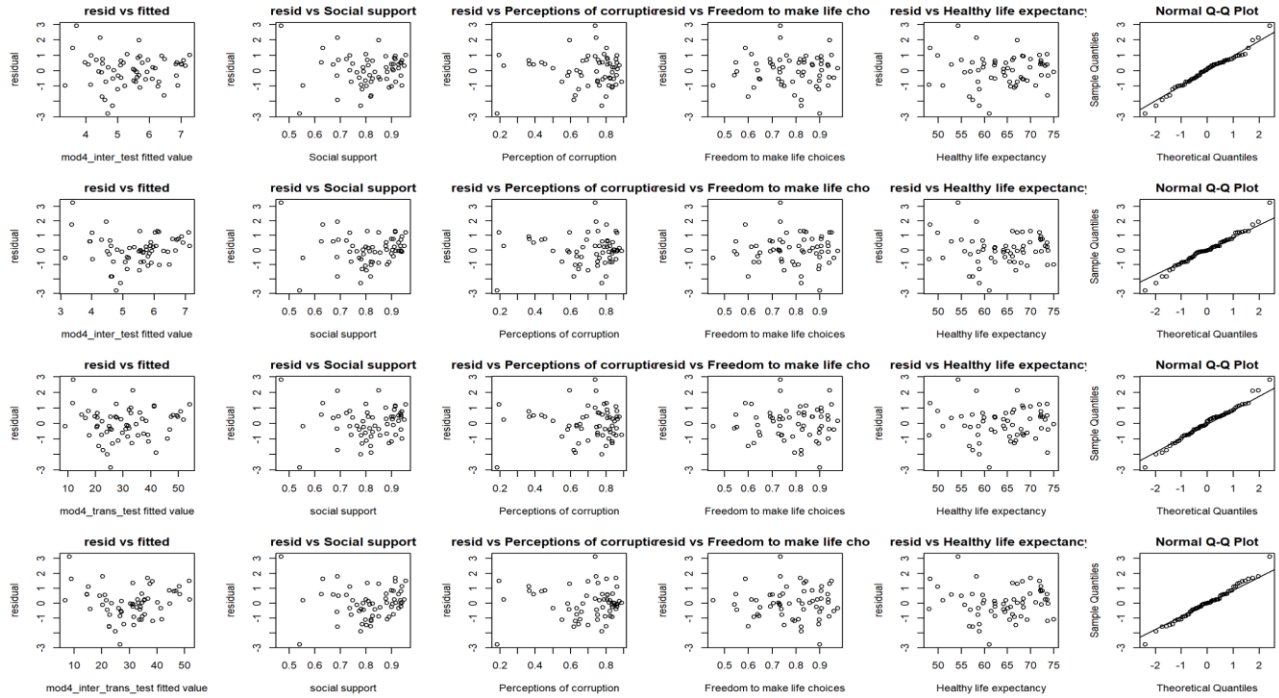

Figure 4:
Residual plots and normal QQ plots of mod4_inter_test, mod4_inter_test, mod4_trans_test and mod4_inter_trans_test respectively

Reference

Helliwell, John F., Richard Layard, Jeffrey Sachs, and Jan-Emmanuel De Neve, eds. (2020). *World Happiness Report 2020*. New York: Sustainable Development Solutions Network. https://worldhappiness.report/ed/2020/