MA678 Midterm Report

Qiannan(Nancy) Shen

12/8/2021

Abstract

This report is based on the World Happiness Report which is a landmark survey collecting the happiness score by Gallup World Poll from 2015 to 2019. I observed that the happiness score varies in different countries or regions and related variables would have various effects on the happiness score. Therefore, I build a multilevel model based on the ten geographical regions as categories to detect how the predictors affect the scores. The result shows that the effect of the absence of government corruption on happiness score has a distinguishable distinction. This report illustrates the main parts containing the data analysis, model fitting, and discussion of the result.

Introduction

The happiness score is a meaningful indicator to explore the satisfaction and euphoria of the society you live in when we are pursuing to be in an imaginary and desirable community the utopia. The happiness score varies in different geographical regions, and the related features have a complicated effect to determine the score. For example, we could easily obverse that the overall happiness and satisfaction level in those developed countries would be higher than those developing countries with poor fundamental facilities and suffering problems. Additionally, the features in different countries like the economic production GDP, social support and family contributions, the life expectancy and health, freedom and restrictions vary and represent multiple effects in different regions. For some features, it would reasonable to think that the higher economic production, social support, and freedom may lead to a higher happiness score, and citizens who lived in these countries would be able to easier feel satisfied. But for some variables, it may make the problem complicated to interpret. So I am interested in how these factors have a different impact on happiness and whether there are some factors that stand out which show an obvious and distinguishable effect.

Hence, the report decides to use the geographical regions as the categorical level to conduct an in-depth analysis and investigate which regions would have a higher happiness score and how the predictors affect happiness in different geographical regions.

Method

Data Cleaning and Processing

The data set is from Kaggle: World Happiness Report (https://www.kaggle.com/unsdsn/world-happiness). Firstly, I just downloaded the five CSV documents collected the data from 2015 to 2019. Then I found the countries in each document are quite different and the region information is missing in some data frames. So, I use inner_join() to add the region column for exiting countries and combine them into one data frame. I clean the data frame to make it appropriate to process follow-up analysis. Finally, I got a data frame with 762 observations and 10 columns. Here are the explanations of columns:

Column names	Explanation
year	The year of the happiness survey
Country	The name of the country
Region	The region that the country belongs to
score	The happiness score
GDP	The score of economic production GDP
social	The score of social and family support
life_ex	The score of life expectancy
freedom	The score of freedom
trust	The score of absence of government corruption
generosity	The score of generosity

Exploratory Data Analysis

As categorizing the data into ten region levels, the report in the first place focuses on the distributions of happiness scores in these ten regions illustrated in Figure 1 and observes the difference. It indicates that the mean and distribution vary in different regions. Countries in Australia and New Zealand have the highest happiness score since we know that they are the most livable countries in the world. And Western Europe and North America also rank high in the charts and it is reasonable that most of the countries in these regions are developed with high living standard and their life are peaceful. While, by contrast, Sub-Saharan Africa and Southern Asia have the lowest happiness score because of the low living standard and unstable and turbulent society.

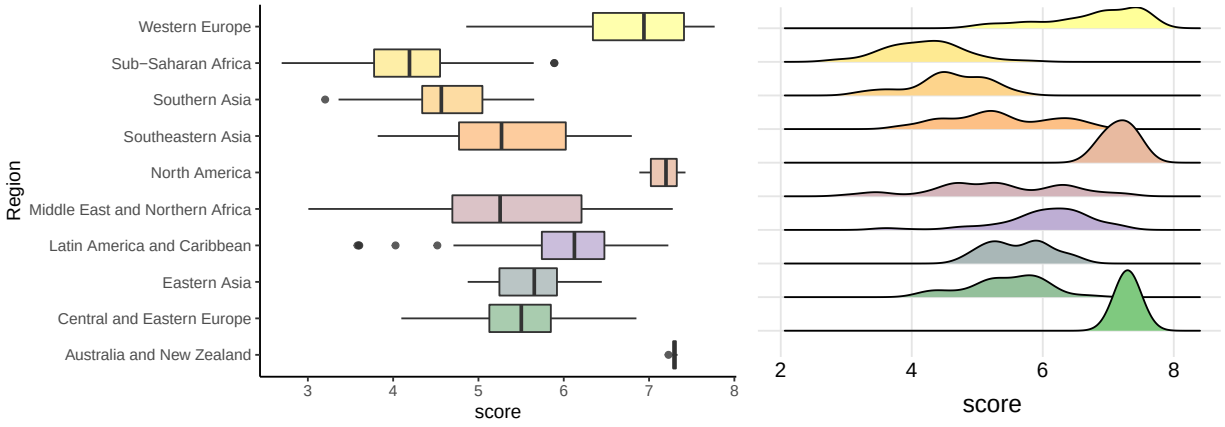


Figure 1: Distribution of Happiness Scores in Ten Regions

In the following part, I draw a series of graphs to display the effect of various factors on happiness scores. Figure 2 (a) shows the relationship between economic production GDP and happiness score. It indicates that except Australia and New Zealand and North America, most of the regions have an increasing trend as GDP grows but with different slopes and intercepts. Similarly, as Figure 2 (a), Figure 2 (c) is the relationship between life expectancy and happiness score and Figure 2 (d) is the relationship between freedom and happiness score. most of the regions have positive correlations but with different slopes and intercepts. Figure 2 (b) illustrates the relationship between social support and happiness score. Apparently, the overall slopes and intercepts don't have much difference and have positive correlations.

However, in Figure 2 (e) and Figure 2 (f), which is the relationship between trust, generosity, and happiness score. There is no obvious increasing trend as above graphs. The change in happiness score varies as each unit change on the absence of corruption (trust) or generosity and the change is slight, while three regions still have an apparent increasing trend.

As expected, the features have an overall positive effect on happiness but are vary in different regions. It interests me that factors may decrease the happiness score in some regions. So I next fit the model and check the details of the effect on each variable in ten regions.

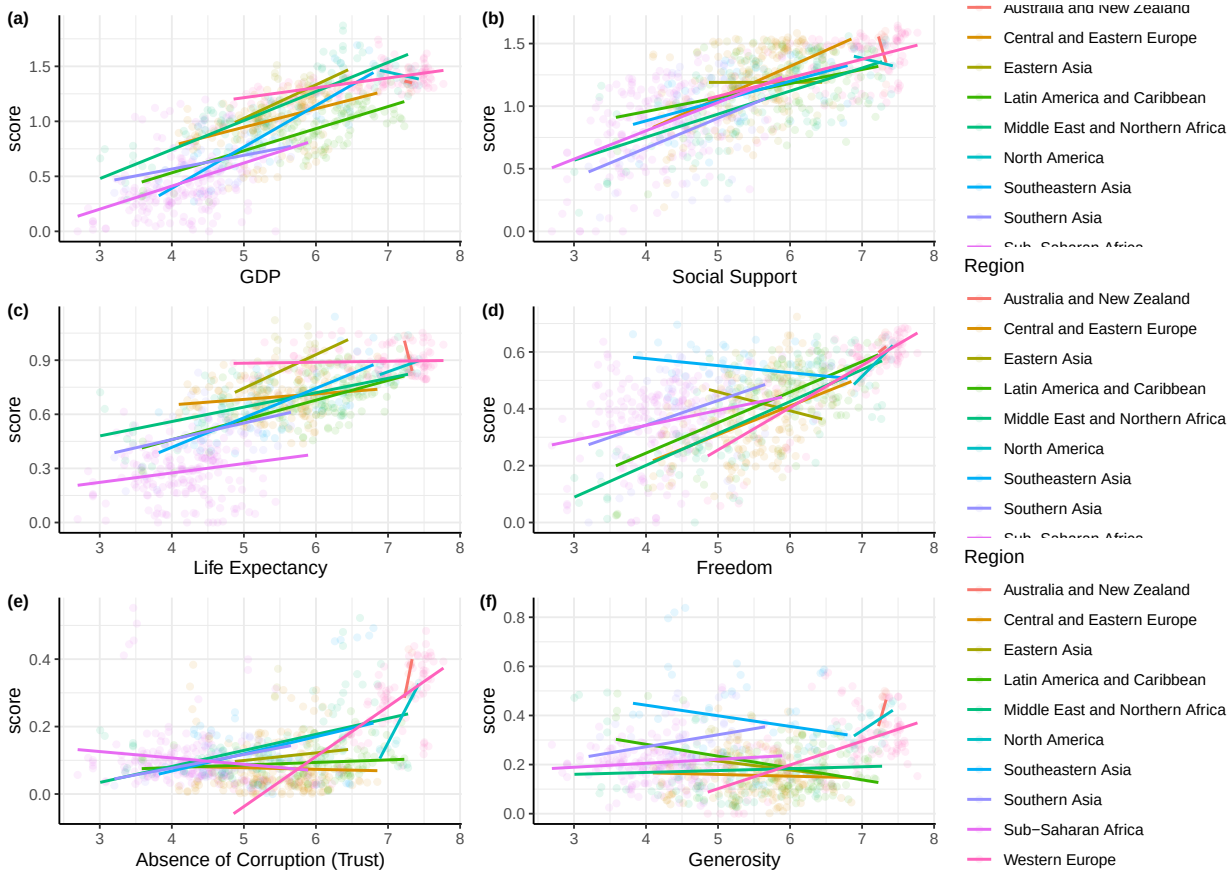


Figure 2: Happiness Scores v.s. predictors in Ten Regions

Model Fitting

Considering different categories, I will use the multilevel model to fit the data. Since the number of variables collected by the yearly happiness survey is reasonable and looks highly relative, I choose to use all six variables in the dataset. And to see the fixed effects below, all variables are significant at alpha = 0.05 level.

```
Model <- lmer(score ~ (GDP + social + life_ex + freedom + trust + generosity | Region) +
                GDP + social + life_ex + freedom + trust + generosity, data = happiness)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.43	0.20	12.34	6.48e-32
GDP	1.26	0.23	5.56	3.74e-08
social	0.28	0.14	2.03	4.30e-02
life_ex	1.08	0.24	4.49	8.30e-06
freedom	1.15	0.33	3.49	5.11e-04
trust	0.85	0.94	0.90	3.68e-01

	Estimate	Std. Error	t value	Pr(>|t|)
generosity	0.55	0.37	1.50	1.33e-01

Result

Model Coefficients

As the result above, I am able to get the following formula of fixed effect:

$$score = 2.43 + 1.26 * GDP + 0.28 * social + 1.08 * life\_ex$$
$$+1.15 * freedom + 0.85 * trust + 0.55 * generosity$$

Then let's take *Southern Asia* as a example. Adding the random effect to slopes and intercept, the estimated formula is:

$$score = 2.90 + 0.1 * GDP + 0.38 * social + 1.56 * life\_ex$$
$$-0.85 * freedom + 6.24 * trust + 0.5 * generosity$$

In the fixed affect formula, combining the relationships shown above part, we could conclude that these factors have a positive correlation with happiness score in general. And it is reasonable that with higher economic production, social support, freedom, generosity, and longer life expectancy, citizens would be more likely to be happy and reach sanctification, and the happiness score increase.

However, when focusing on the random effect, it is interesting to find that the fluctuation of the random effect on the predictor **trust** is bigger than expected. While the other five predictors on different regions vary in a small range. The result indicates that the effect of the absence of government corruption on happiness score has a distinguishable distinction. In detail, especially shown in the *southern Asia* formula, for each 1% increase of the absence of government corruption, the happiness score would have an expected 6.4% increase. In contrast, the absence of corruption would have a negative effect on the happiness score. For instance, the happiness score in Eastern Asia would have an expected 3.67% decrease for each 1% increase of the absence of government corruption. It makes sense that countries in different regions have their own national condition in the political system, government administrative policy, style, and social system, and so forth. So the effect of the absence of corruption varies so much in different regions.

Remarkably, freedom has a negative correlation with the happiness score in *southern Asia* differentiated from the positive coefficient in the other nine regions. These results also appear in the predictor **social** and **generosity** in some countries.

	(Intercept)	GDP	social	life_ex	freedom	trust	generosity
Australia and New Zealand	2.59	1.21	0.15	1.23	1.42	1.88	0.42
Central and Eastern Europe	2.56	1.17	0.51	0.48	1.61	-0.24	1.22
Eastern Asia	1.75	2.10	0.28	0.94	1.28	-3.67	0.29
Latin America and Caribbean	2.67	1.60	-0.31	1.57	2.13	1.84	0.00
Middle East and Northern Africa	1.61	1.69	0.47	1.78	1.11	0.15	-1.07
North America	2.56	1.45	0.03	1.22	1.87	1.14	0.34
Southeastern Asia	2.22	1.66	0.18	1.05	1.28	-0.91	0.50
Southern Asia	2.90	0.10	0.38	1.56	-0.85	6.24	0.77
Sub-Saharan Africa	2.75	0.79	0.61	0.06	0.44	-0.70	2.25
Western Europe	2.73	0.78	0.48	0.94	1.23	2.75	0.80

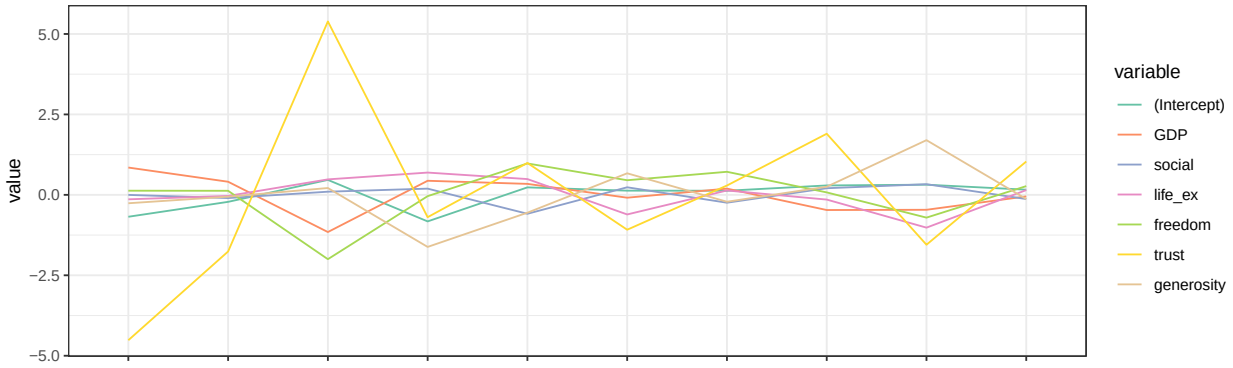


Figure 3: Random Effects Plot

Model Validation

For the model checking part, I mainly draw a residual plot, Q-Q plot, and residual leverage plot which are shown in the Appendix. In the residual plot, all the points are randomly dispersed around the horizontal 0 line which indicated the model is appropriate. In the Q-Q plot, data appears as roughly a straight line and the normality is approved. The residual leverage plot indicates that there is no abnormal point which is good.

Discussion

In this report, the result of the model is reasonable to some extent. For most features of different regions, it represents a positive correlation with the happiness score. It makes sense that, with higher economic production, the region would have a high living standard and adequate fundamental facilities which improve the quality of their life and lead to euphoria. And the higher degree of freedom, social support and generosity, and longer life expectancy would bring more physical and mental pleasure. However, the result indicates that the effect of the absence of government corruption on happiness score has a distinguishable distinction. Based on my acknowledgment, I could say the different result is caused by the different national conditions of countries. But investigating the specific political system, government administrative policy, and social conditions is another long and complicated story, and is hard to give a clear explanation based on my current knowledge.

This report also has some limitations. First, the data of the happiness score is from 2015 to 2019 and the happiness score prior to these years is not included. So the model can only apply to the current year happiness study and the effect of time would not show in the model which I think is another important factor to affect happiness. In a further study, we could find an appropriate way to get the data in the previous year and consider the time factor into the model. Second, the trend of variables shown on the EDA part is quite different from the model-fitting result I got and it is worthy to figure out the reason. Third, as we could see that the variables we use are different from the normal measurable data, all the happiness score, degree of freedom, social support are invisible and measurement is ambiguous. This is probably to cause some bias and lead to inaccuracy. So it is important to check whether the way of collecting data and the computing method are appropriate.

Citation

[1] *World Happiness Report*. https://www.kaggle.com/unsdsn/world-happiness

[2] VAMSI KRISHNA. *Happiness Index and Terrorism!*. https://www.kaggle.com/vamsikrishna/happiness-index-and-terrorism

Appendix

More EDA

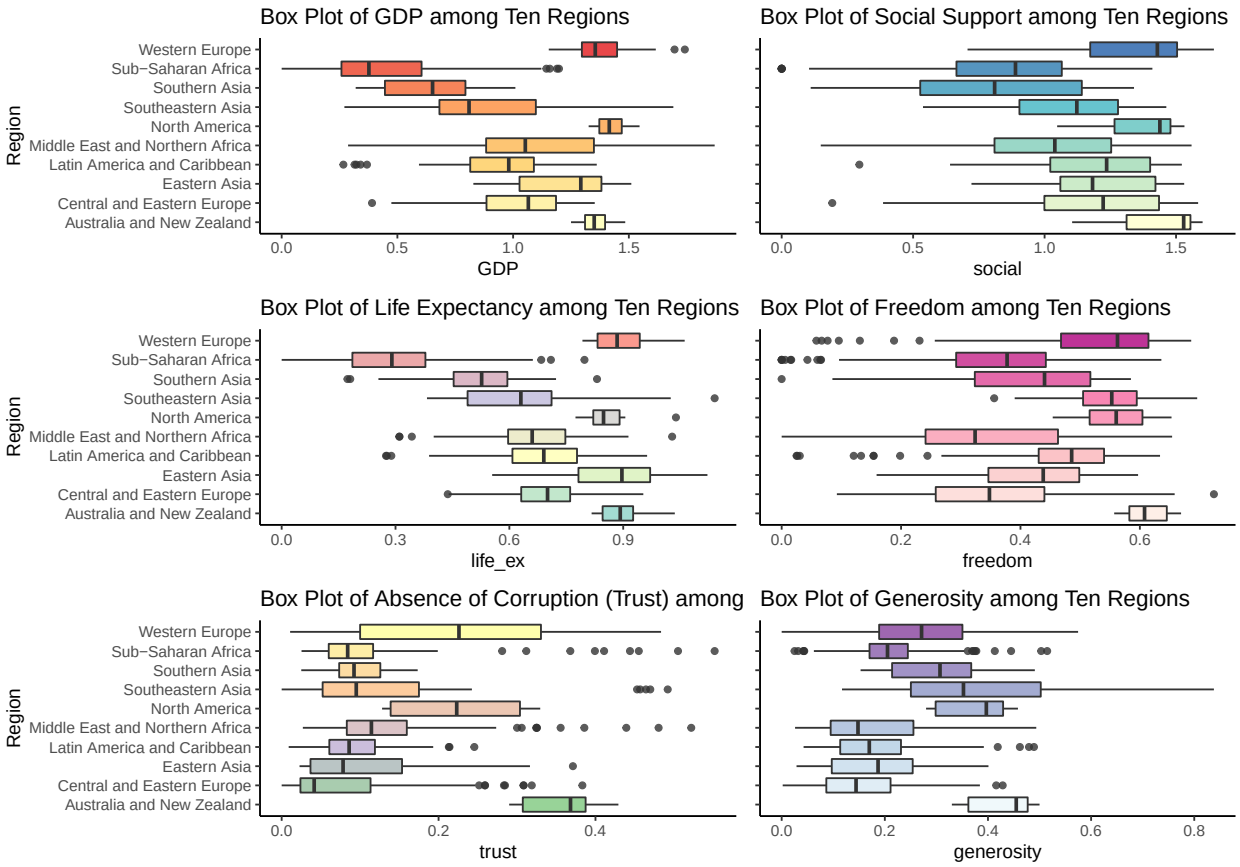
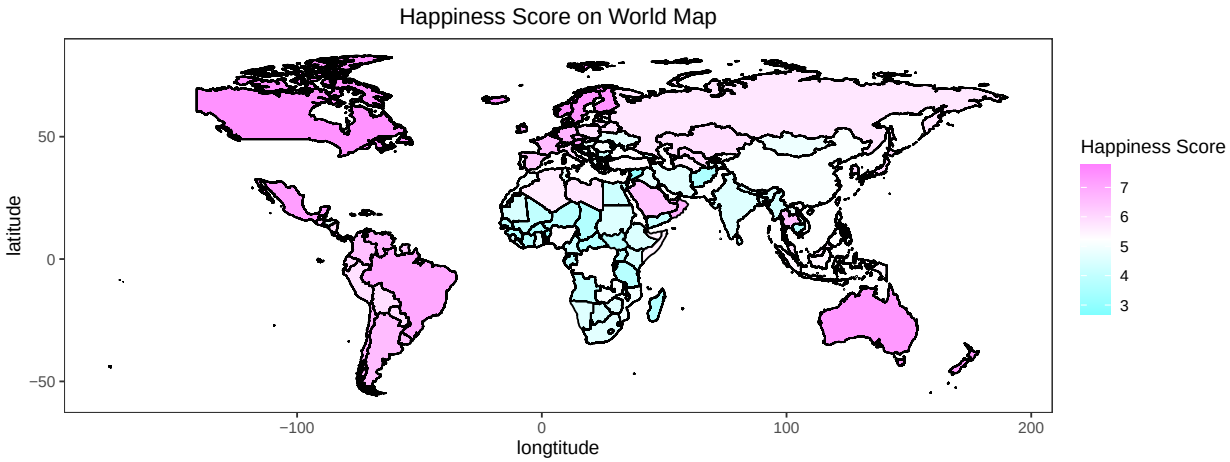


Figure 4: Variables Box Plot



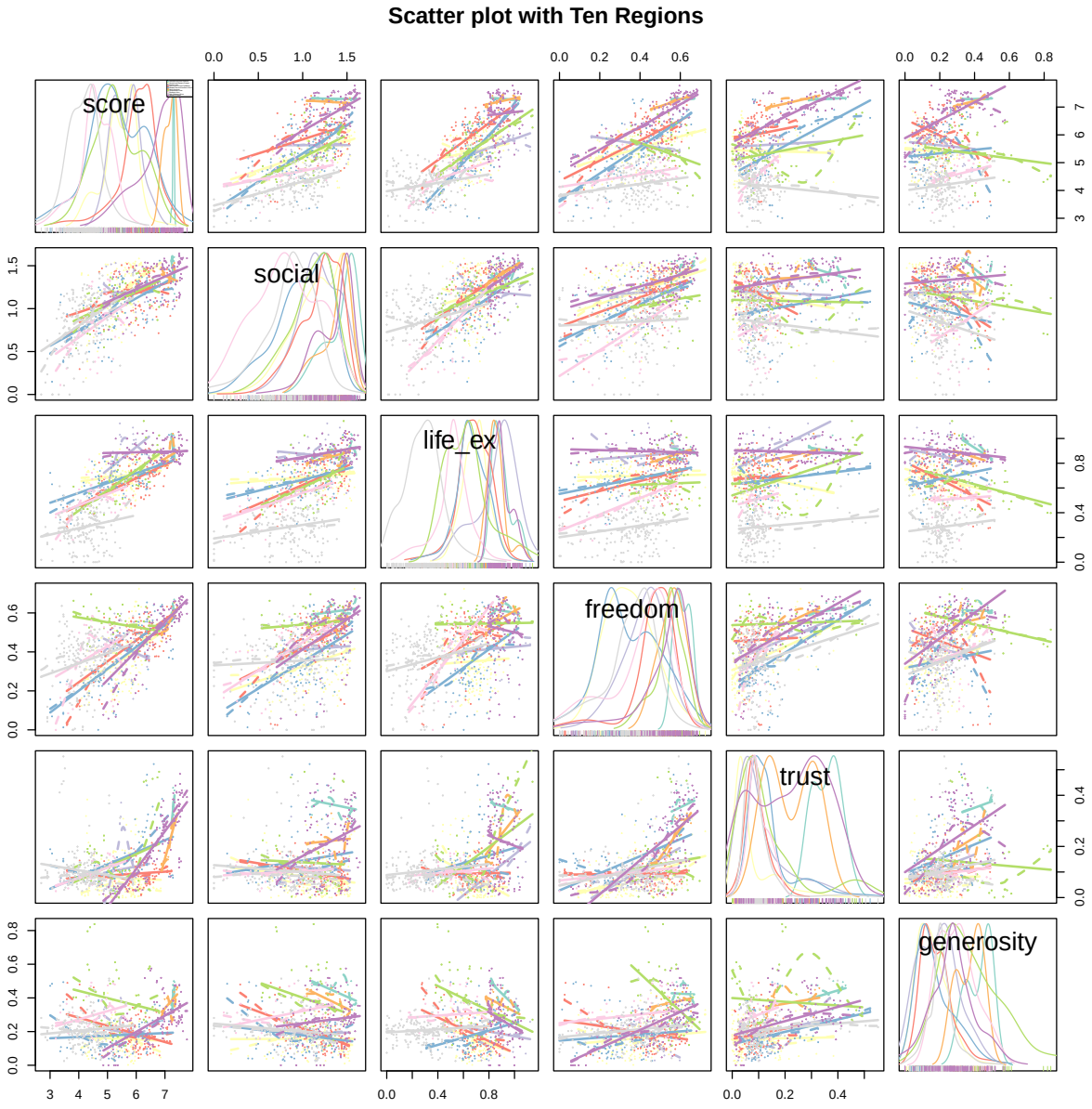


Figure 5: Distribution and Scatter Plot

Model Checking

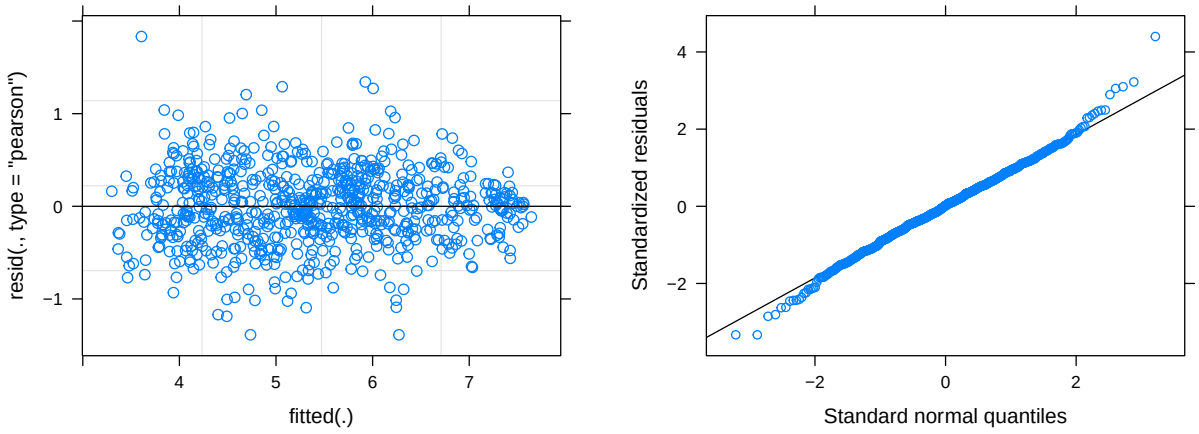


Figure 6: Residual Plot and Q-Q Plot

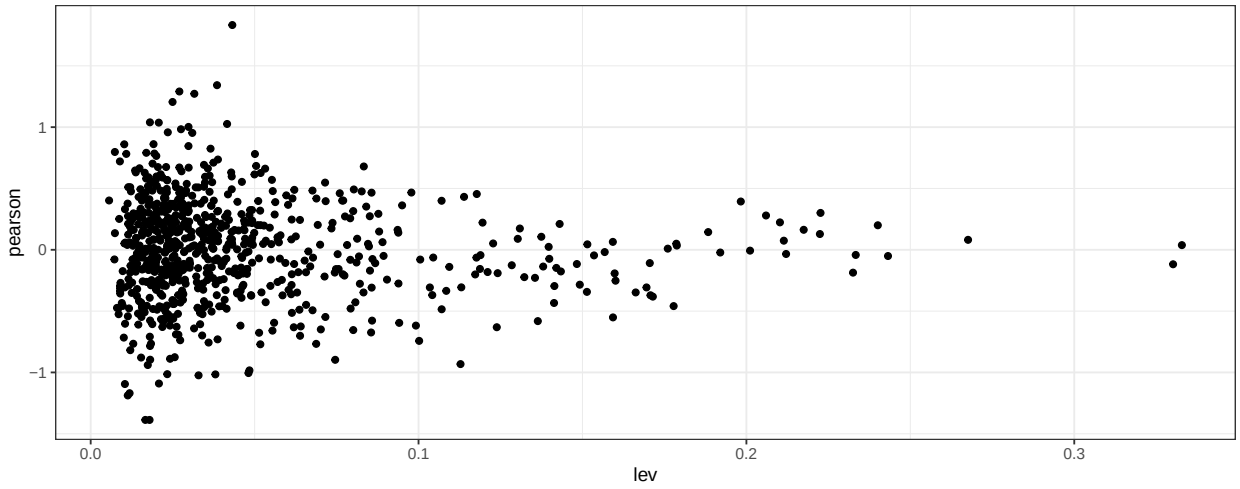


Figure 7: Residuals v.s. Leverage Plot

Full Results

Random effects of model

```
## $Region
##                                 (Intercept)         GDP       social
## Australia and New Zealand         0.1600457 -0.04827252 -0.130240289
## Central and Eastern Europe        0.1305754 -0.08800131  0.235016113
## Eastern Asia                     -0.6808074  0.84863339 -0.001312132
## Latin America and Caribbean       0.2323603  0.34095118 -0.588109376
## Middle East and Northern Africa  -0.8252230  0.43759779  0.192877944
## North America                     0.1275420  0.19336906 -0.243831269
## Southeastern Asia                -0.2172693  0.40764217 -0.099434784
## Southern Asia                     0.4652214 -1.15591543  0.098348467
## Sub-Saharan Africa                0.3155923 -0.46408908  0.330736949
## Western Europe                    0.2919626 -0.47191524  0.205948376
##                                    life_ex     freedom      trust  generosity
## Australia and New Zealand        0.14847192  0.27083204  1.0359188 -0.12892182
## Central and Eastern Europe      -0.60892568  0.45422767 -1.0839489  0.66983407
## Eastern Asia                    -0.13899542  0.12907813 -4.5178616 -0.26132578
## Latin America and Caribbean      0.49048931  0.97565623  0.9914944 -0.55481323
## Middle East and Northern Africa  0.69232806 -0.04090809 -0.6940679 -1.61827246
## North America                    0.13600553  0.71675657  0.2901518 -0.21414233
## Southeastern Asia               -0.03441828  0.12594563 -1.7604553 -0.05662017
## Southern Asia                    0.48077309 -1.99983653  5.3897129  0.21241031
## Sub-Saharan Africa              -1.01898968 -0.70795319 -1.5502544  1.69969220
## Western Europe                  -0.14673886  0.07620155  1.8993101  0.25215920
##
## with conditional variances for "Region"
```

Fixed effects of model

```
## (Intercept)         GDP      social     life_ex     freedom       trust
##   2.4337725   1.2563042   0.2772453   1.0839509   1.1520112   0.8471683
##  generosity
##   0.5525946
```

Coefficients of model

```
## $Region
##                                 (Intercept)       GDP      social    life_ex
## Australia and New Zealand          2.593818 1.2080317  0.14700503 1.23242278
## Central and Eastern Europe         2.564348 1.1683029  0.51226143 0.47502519
## Eastern Asia                       1.752965 2.1049376  0.27593319 0.94495545
## Latin America and Caribbean        2.666133 1.5972554 -0.31086406 1.57444017
## Middle East and Northern Africa    1.608550 1.6939020  0.47012326 1.77627892
## North America                      2.561314 1.4496732  0.03341405 1.21995639
## Southeastern Asia                  2.216503 1.6639463  0.17781054 1.04953258
## Southern Asia                      2.898994 0.1003887  0.37559379 1.56472396
## Sub-Saharan Africa                 2.749365 0.7922151  0.60798227 0.06496118
## Western Europe                     2.725735 0.7843889  0.48319370 0.93721200
##                                    freedom      trust   generosity
## Australia and New Zealand        1.4228432  1.8830871  0.423672753
```

```
## Central and Eastern Europe        1.6062388 -0.2367805  1.222428635
## Eastern Asia                      1.2810893 -3.6706932  0.291268785
## Latin America and Caribbean       2.1276674  1.8386628 -0.002218657
## Middle East and Northern Africa   1.1111031  0.1531004 -1.065677889
## North America                     1.8687677  1.1373201  0.338452243
## Southeastern Asia                 1.2779568 -0.9132869  0.495974397
## Southern Asia                    -0.8478254  6.2368812  0.765004884
## Sub-Saharan Africa                0.4440580 -0.7030861  2.252286771
## Western Europe                    1.2282127  2.7464785  0.804753773
##
## attr(,"class")
## [1] "coef.mer"
```

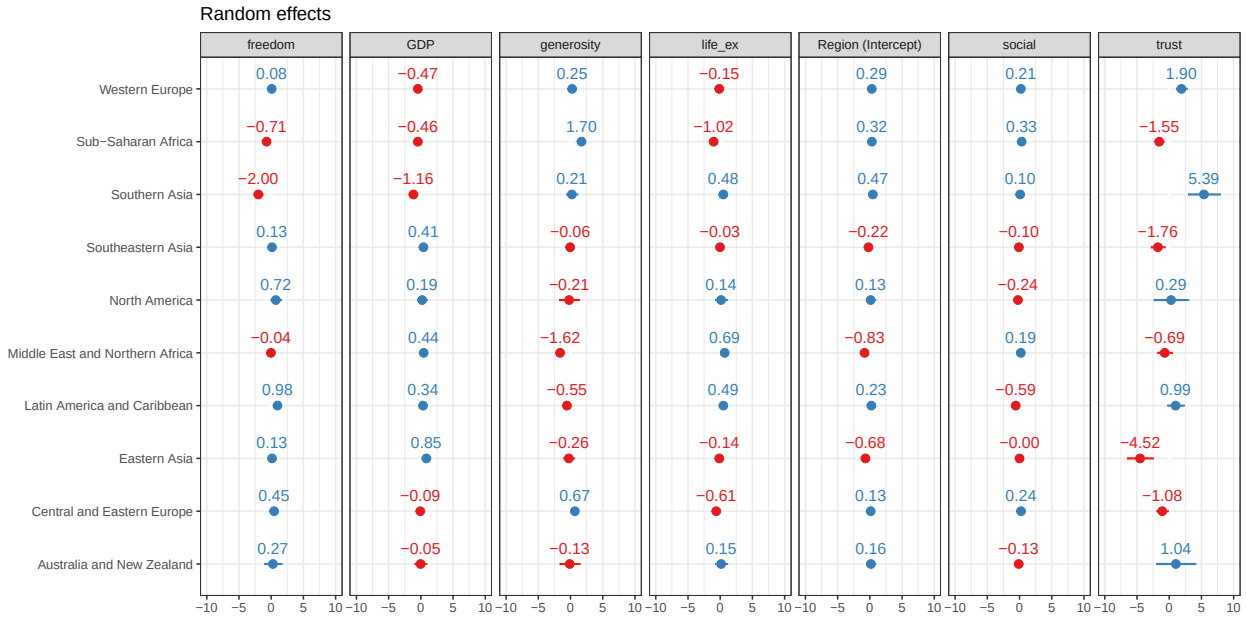


Figure 8: Random Effect Plot