World Happiness Analysis

STAT GR 5291 Advanced Data Analysis

Final Project

Name
Yiqi Lei (yl4353)

Submitted to
Professor Ronald Neath

Department of Statistics, Columbia University
May, 2020

I. Introduction

1. Topic

The topic of this project is world happiness. Although happiness is a relatively subject matter, we can still use statistical methods to investigate the factors that may affect happiness at a broad view, estimate the magnitude of those effects, and understand the quality of lives in the world.

2. Data

The data was released by the Gallup Company from their World Poll study. The data were collected through surveys from participants all around the world. The data used for this project were the national average responses to life evaluation questions for the years 2017 to 2019.

Here are our data variables:

- Ladder score: the step of the ladder the participant personally feels standing at this time for a ladder with steps numbered from 0 at the bottom representing the worst possible life, to 10 at the top representing the best possible life. We can simply understand this as the happiness score.
- Name: the country/territory participated in the World Poll. There are a total of 153.
- Region: the region the country/territory belongs to. There are 9 of them.
- GDP per capita: the GDP per capita of the country.
- Healthy Life Expectancy (HLE) at birth: the equivalent number of years of good health that a newborn can expect, computed by applying disability weights to health states, according to the World Health Organization.
- Social support: the national average of the binary responses (either 0 or 1) to the question if a person has someone to count on in times of trouble.
- Freedom: the national average of the binary responses (either 0 or 1) to the question if one is satisfied with the freedom to make choices in life.
- Generosity: the residual of regressing national average of response to the question if one has donated money to a charity in the past month, on GDP per capita.
- Corruption Perception: the national average of the survey responses (either 0 or 1) to two questions, if there is corruption widespread throughout the government or not, and if there is corruption widespread within businesses or not.

3. Interest

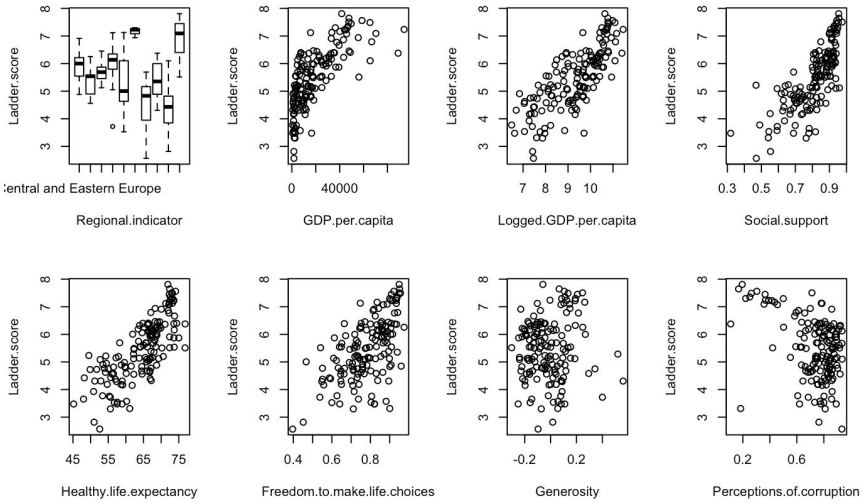
There are lots of interesting questions we can ask with these variables. The response variable is the ladder score, the others are the potential predictors for the ladder score. According to Gallup, they define the present life evaluation scored less than or equal to 4 as suffering, scored 5 or 6 as struggling and scored greater than or equal to 7 as thriving. With this standard, is the majority of

countries in the world struggling or thriving? Is region a significant factor for happiness score? If yes, how do the scores differ in different regions in the world? Intuitively, one would think that GDP per capita, healthy life expectancy, social support, freedom and generosity have positive effects on the happiness score, while corruption perception acts negatively. Is this really the case statistically? What are the remaining factors that are statistically significant after accounting for other variables? What are the estimates of those effects?

II. Analysis

1. Preprocessing

Data collected in our real life is not always perfect. From plots of ladder score corresponding to variables, at a closer look, it is obvious that For GDP per capita, the scores are highly skewed. It is proper to do a log transformation on GDP per capita and we can see that now there might be a linear relationship between the logged GDP per capita and ladder score.



2. Ladder Score Worldwide

Ranking by ladder score, there are 15 countries/territories categorized as suffering, 17 as thriving. The majority, which is 121 countries/territories are struggling. The result could be surprising to some idealists.

3. Ladder Score and Region

Besides country, region is the only categorical variable in the dataset. From the boxplot shown above, ladder scores differ region by region, and there are also gaps within each region. Testing the null hypothesis that the mean scores in regions are all equal, by fitting the one-way analysis of the variance model, I get a p-value of <2e-16 which is way smaller than 0.01. This indicates

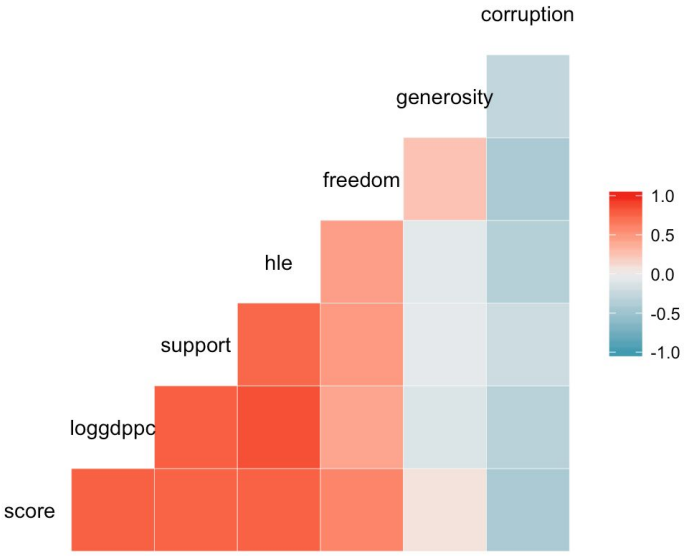
that at least some regions have different ladder scores so the score does differ by region. Tukey method can help make it easier to see the differences more clearly.

	diff	lwr	upr	p.adj
Sub-Saharan Africa-North America and ANZ	-2.7900301	-3.975690	-1.60436985	1.955176e-10
South Asia-North America and ANZ	-2.6980822	-4.113568	-1.28259592	3.663079e-07
Southeast Asia-North America and ANZ	-1.7901583	-3.147249	-0.43306794	1.596055e-03
Sub-Saharan Africa-Latin America and Caribbean	-1.5982908	-2.209545	-0.98703642	1.845302e-12
South Asia-Latin America and Caribbean	-1.5063429	-2.491961	-0.52072479	1.036361e-04
Sub-Saharan Africa-Central and Eastern Europe	-1.5003228	-2.156658	-0.84398778	6.358334e-10
South Asia-Central and Eastern Europe	-1.4083749	-2.422568	-0.39418220	6.668679e-04
Sub-Saharan Africa-East Asia	-1.3313551	-2.321700	-0.34100974	1.168290e-03
South Asia-East Asia	-1.2394072	-2.495829	0.01701436	5.652171e-02
Sub-Saharan Africa-Southeast Asia	-0.9998718	-1.835004	-0.16473930	6.678123e-03

Here, I selected the top 10 differences between two regions. The table shows the estimated mean value of score differences between the two regions each row (diff), with the 95% confidence interval shown. For instance, for the first row, we are 95% confident that the mean score of happiness in Sub-Saharan Africa is 2.7900301 lower than the score in North America and ANZ (Australia and New Zealand) with a 95% confidence interval from 1.60436985 to 3.975690. The adjusted p-value for multiple comparison is very small, meaning this difference is statistically significant.

4. Model Building

Now, try to build a model to predict the ladder score using independent variables introduced before.



From the correlation plot, we can see that each independent variable has some correlation with the dependent variable, and the independent variables themselves are somewhat correlated. There might be some predictors that we do not need.

First, try a full model. I compare the full model without region, and the full model with region. By a chi-squared test on these two models, I get a p-value of 1.995e-05 which is quite smaller than 0.01. Hence, region is actually important for our prediction model.

Our m.full = lm(score~region+loggdppc+support+hle+freedom+generosity+corruption,df). Because score is quantitative, multiple regression is used here.

Then, I compared m.full.1 with m0=lm(score~1,df) to test the null hypothesis that all coefficients of the predictors are 0, ie. NH: $\beta_1 = \beta_2 ... = \beta_7 = 0$. With a p-value<2.2e-16<<0.01, there is strong evidence that at least some of the predictors affect the ladder score.

Next, since the region factor has 9 categories, it is reasonable that not all of them are important. To treat them separately so that we can know the importance of each category, I transfer them into 8 dummy variables. Do the full model again, but this time, replace the region by the 8 dummy variables.

Now, we need to eliminate some unnecessary predictors. I choose to use backward stepwise elimination based on AIC.

```
Step:  AIC=-187.85
score ~ loggdppc + support + freedom + generosity + corruption +
    region.Commonwealth.of.Independent.States + region.East.Asia +
    region.Middle.East.and.North.Africa + region.South.Asia +
    region.Southeast.Asia + region.Sub.Saharan.Africa
```

Hence, we get our best model with an AIC of -187.85. To have a closer look at the model, I get the following results.

```
Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)            2.5 %      97.5 %
(Intercept)                             0.35565    0.73244   0.486 0.628030     -1.092339812  1.80363521
loggdppc                                0.28462    0.07097   4.010 9.79e-05 ***  0.144318272  0.42491908
support                                 2.19763    0.63530   3.459 0.000717 ***  0.941674010  3.45357946
freedom                                 2.01681    0.48511   4.157 5.56e-05 ***  1.057782792  2.97584620
generosity                              0.63518    0.31992   1.985 0.049033 *    0.002728268  1.26763975
corruption                             -0.67290    0.29204  -2.304 0.022678 *   -1.250246002 -0.09554843
region.Commonwealth.of.Independent.States -0.55335    0.17116  -3.233 0.001526 **  -0.891725569 -0.21497790
region.East.Asia                       -0.49609    0.22668  -2.189 0.030279 *   -0.944229724 -0.04795919
region.Middle.East.and.North.Africa    -0.51031    0.15409  -3.312 0.001178 **  -0.814939221 -0.20567159
region.South.Asia                      -0.92136    0.22857  -4.031 9.05e-05 *** -1.373233724 -0.46948835
region.Southeast.Asia                  -0.91847    0.20288  -4.527 1.26e-05 *** -1.319542243 -0.51739242
region.Sub.Saharan.Africa              -0.65513    0.16151  -4.056 8.22e-05 *** -0.974422392 -0.33583609
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5213 on 141 degrees of freedom
Multiple R-squared:  0.7963,    Adjusted R-squared:  0.7804
F-statistic:  50.1 on 11 and 141 DF,  p-value: < 2.2e-16
```

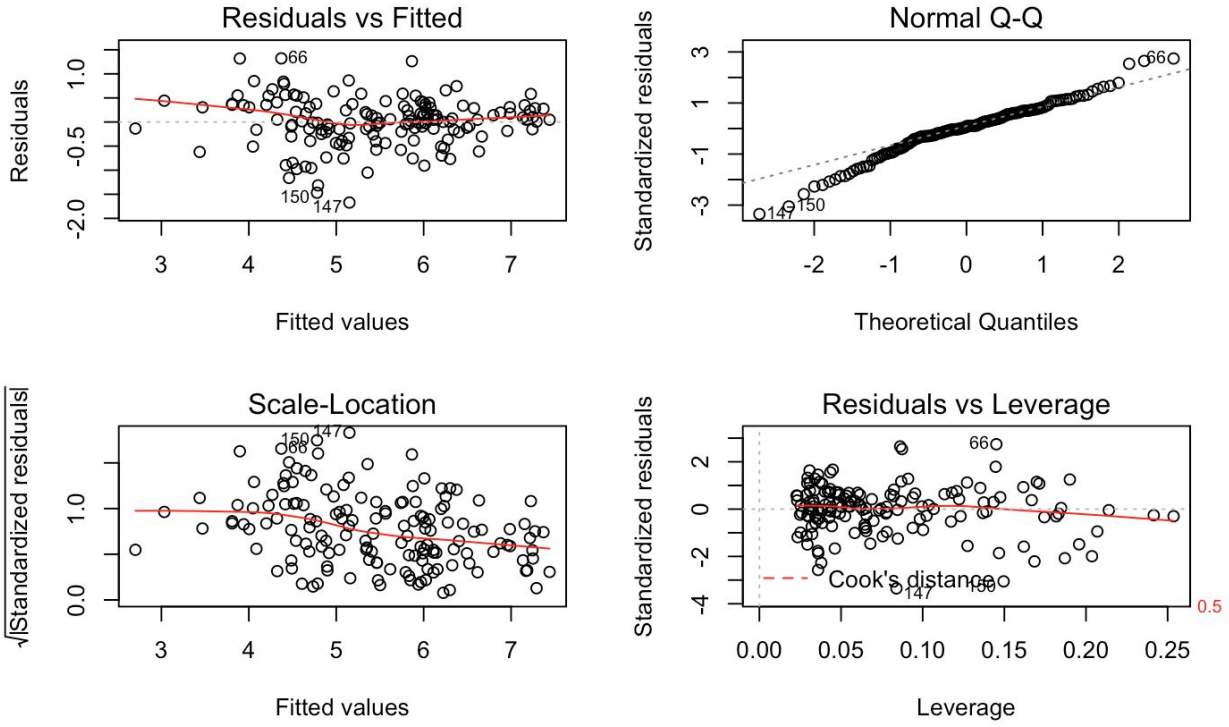
From the summary, we can see that each predictor is statistically significant in the mode. We can also know from the estimates and confidence intervals about the specific performance of each predictor.

Loggdppc is estimated to be 0.28462×log(x) higher if one gdp is x times of another. As support increases by 0.1, the ladder score will increase by 0.219763 with a 95% confidence interval of 0.091674 to 0.345358. For freedom, the one who voted yes to the freedom question would have a 2.01681 higher ladder score with a 95% confidence interval of 1.05778 to 2.97585. Similar for other predictors.

Note that the predictors loggdppc, support, freedom and generosity have positive effects on the score and predictors including and after corruption all act negatively, which matches common senses. Moreover, the adjusted r squared is 0.7804 which means 78.04% of the response variables can be explained by the model which is good.

5. Assumption Check

The multiple regression model assumes linear relationships, normality for residuals, no multicollinearity among independent variables and similar variance for error terms.



From Residuals vs. Fitted plot, the red line is not quite horizontal but there is no obvious pattern so there might be slight heteroscedasticity problems. From Q-Q plot, residuals are not perfectly normally distributed due to some skewness, but at least the majority of them satisfy normality. The line in the Scale-Location plot is not completely horizontal but close, indicating that the points may not spread very randomly. In the Residuals vs Leverage plot, we can barely see the Cook's distance line which means all of our data are well inside the lines so there are no outliers.

III. Conclusion

In this project, we look into the factors that might be used to predict the ladder scores, ie. happiness score. The majority of the countries and territories in the world are struggling towards happiness. All the variables have some correlations between each other. Region, as a categorical variable, has strong effects on the happiness score. Other variables, when acting together, might be eliminated from the full model. With backward stepwise elimination, we can get a model that explains 78% of the data. We can also know specifically how each of the predictors influence the ladder score and the corresponding 95% confidence intervals. Although there are heteroscedasticity, normality and homoscedasticity problems, they are minor and our model is still acceptable.

IV. Limitation

The data are from surveys, so that most of the predictors are subjective responses but not objective measures. Also, the model I build in this project is good for predicting group happiness, like country, regionwide or worldwide happiness, but not individual happiness. For further study, if there are complete individual survey responses available, the model could be more accurate and can be used to predict individual happiness scores.