Statistical Analysis Report

BIOST 579

Ivy Zhang

June 8th, 2021

***Source of Data***

The dataset I plan to analyze is from the World Happiness Report project. Every year, the project has collected national average happiness scores and various factors (such as GDP, healthy life expectancy, social support score) of countries in that year from 2005 to 2020, and the generate a world happiness ranking of countries all over the world. We start to wonder what factors of a country are associate with its national average happiness score and what improvement a country can make to improve the happiness level of its citizens. Therefore, this analysis report focus on discovering the relationship between different factors and the national average happiness report. This dataset only includes countries with valid happiness scores and the national average happiness scores will be the main interest of this analysis. A small number of countries or regions may have missing value in one or multiple factors due to geographical or political reasons. The quantitative data of subjective rating factors mostly come from the Gallup World Poll(GWP) by conducting interviews and surveys to worldwide participants in that year and calculate the average scores in each country. This study is a fully observational study.

***Participants***

The study population is countries with valid national average happiness scores at least for one year, which means there are citizens responding to the survey or the interview of Gallup World Poll in the country.

***Outcome***

The national average happiness scores(variable name: ‘Life Ladder’) in the dataset are collected from the Gallup World Poll. Happiness score is a range from 1 to 10, with 10 at the top. At the top of happiness score means the participants reporting themselves as currently living the best possible life, and 1 means the worst. Happiness scores are recorded at the national average level in our study.

***Predictors***

* **GDP**: GDP per capita
* **Year:** Year of the information is taken.
* **Healthy Life Expectancy**: Data extracted from WHO, interpolation and extrapolation are applied since only WHO only offers data in 2000, 2005, 2010, 2015, and 2016.
* **Social Support:** National average of a binary response to the 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**: National average of a binary response to the question:” Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"
* **Generosity:** residual of regressing national average of response to the GWP question “Have you donated money to a charity in the past month?" on GDP per capita.
* **Positive Effect:** average of three positive effect measures in GWP: happiness, laugh, and enjoyment. The effect will be evaluated by the question:” Did you experience (the positive effects) A LOT OF THE DAY yesterday?”
* **Negative Effect:** average of three negative effect measures in GWP: worry, sadness and anger. The effect will be evaluated by the question:” Did you experience (the negative effects) A LOT OF THE DAY yesterday?
* **Corruption:** National average of a binary response to the question either is “Is corruption widespread throughout the government or not" and “Is corruption widespread within businesses or not?" Which question is asking is based on the government status in that country.

***Interested Scientific Questions***

1. What factors of a country are associated with the national average happiness score and how important are the factors?
2. Does the country tend to have a lower national average happiness score in COVID time compared to before (2019 v.s. 2020)?
3. Does COVID-19 change people’s mindset on the importance of several factors of the country that are associated with the national average happiness score?

***Sample Size***

There are a total of 1708 entries in the dataset after removing the entries with missing value, with 126 entries in 2019, 81 entries in 2020, and 1501 entries from 2005 to 2018. 93 countries have valid national happiness scores in 2019 and 2020, but 15 have one or multiple missing values in other factors.

***Missing Data***

Entries with missing values in one or multiple factors will be removed from the dataset when doing linear regression models. No entries will be considered as missing values in doing the paired T-test.

***Statistical Analysis and Methods***

One Generalized Estimating Equations(GEE) model, three linear regression models and one paired t-test are considered to be applied in this data analysis. For each predictor, a 95% Wald confidence interval will be used to determine whether the predictor is statistically significantly associated with the national average happiness score. Attribution of each factor will be helped in analyzing the factor’s effect on happiness score.

The GEE model is for answering the first scientific question. National average happiness scores will be the response variable. All other factors will be the predictors, and the country will be adjusted in the model. The model will be used to determine which factors are associated with the national average happiness score and how important they are.

The paired-T test will be applied to the 93 countries which have valid national average happiness scores in 2019 and 2020 to compare whether the difference in national average happiness exists between these two years. The test will be performed with a significance level of 0.05.

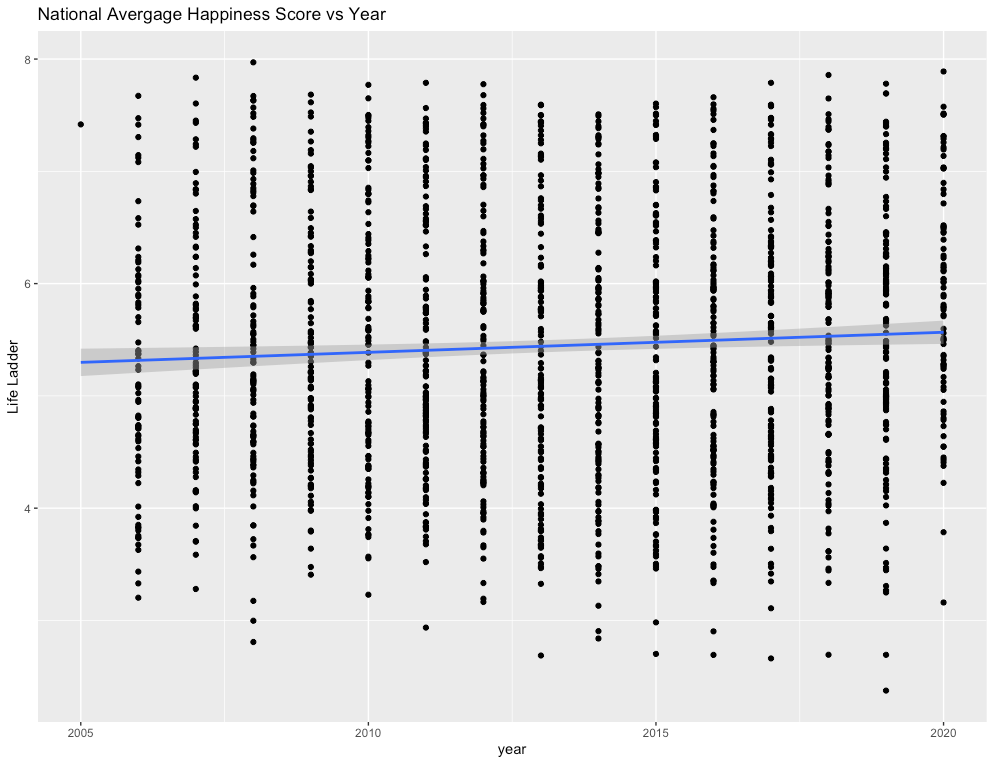
The second, third, and fourth linear regression model will be applied to the 126 entries in 2019, 81 entries in 2020, and a total of 207 entries. All entries in these three models have no missing value in any of the factors. AIC with k equals 2 will be applied to help determine the predictors for the 2019 model and 2020 model. AIC here is defined by:

where L means the likelihood. The smaller AIC is, the better fit the model is.

After determining the significant predictors associated with the national average happiness score in the two models, the significant predictors will be included in the two-year model. The interaction term of the year 2020 and the independent variables will also be included in the two-year model to see whether there is a significant difference between 2019 and 2020.

***Results***

Figure 1： Linear Trend of National Average Happiness Scores on Year

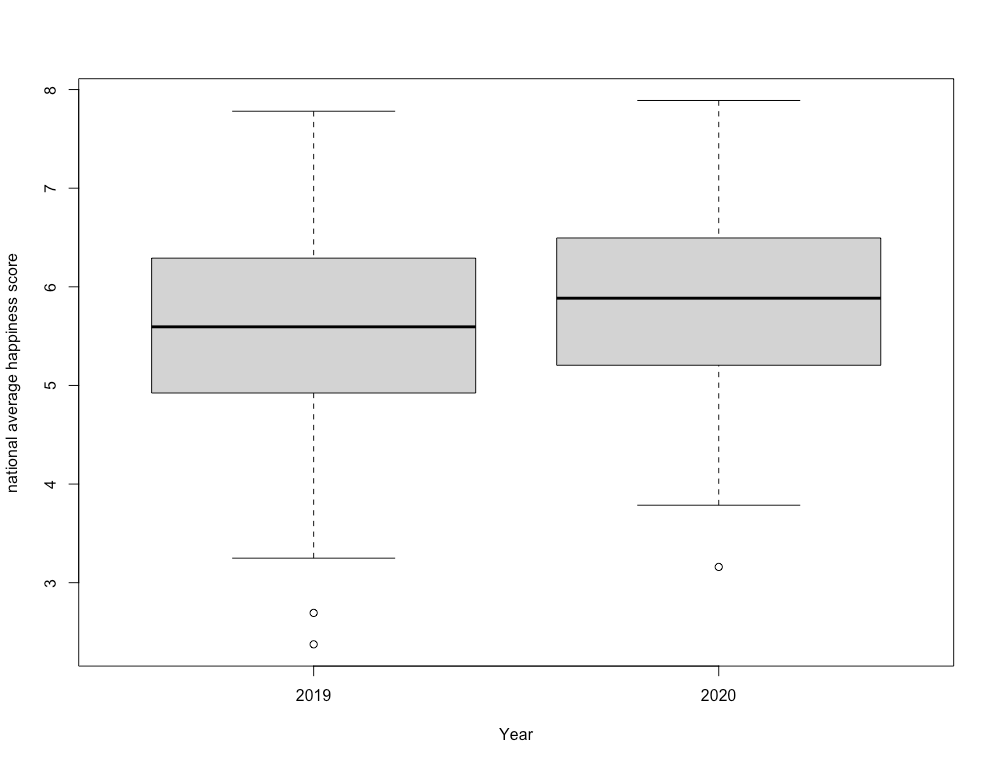


To research what factors of a country are associated with the national average happiness score, we fit a GEE model using robust SE with national average happiness score as the response, and all other factors in the dataset as the predictor variables, and we adjust for the countries. We also assume the national average happiness scores of a country in different years are independent. At the significance level of 0.05, we can conclude that there is statistically significant evidence of the association between the national average happiness score and all predictor variables except for the negative effect. Social support, log GDP per capita, healthy life expectancy at birth, freedom to make life choices, positive affect, and negative affect are positively associated with the national average happiness score, with negative perceptions of corruption. Table 1 will show the estimated coefficients of all predictors in the model. All predictors have a p-value that is much smaller than 0.05. From table 1, we can see people weigh social support and positive affect most when evaluating their happiness level. One of the interpretations of our result is between countries who are differed by 1 score in social support but have the same value at all other factors, the countries which have higher social support score will have 1.8445 higher in average national average happiness score. for It seems to be surprising that negative affect is also positively associated with the national average happiness score. It may due to citizens who are more emotional, they tend to feel more negative effects but also are easier to feel happy compared to citizens who are less emotional.

Table 1: Estimated Model for Question 1

| Variables | Estimated Coefficients |
| --- | --- |
| Intercept | -2.53 |
| Log GDP per capita | 0.38 |
| Social Support | 1.84 |
| Healthy life expectancy at birth | 0.03 |
| Freedom to make life choices | 0.39 |
| Generosity | 0.42 |
| Perceptions of corruption | -0.70 |
| Positive affect | 1.99 |
| Negative affect | 0.19 |

Figure 2: Boxplot of National Average Happiness Scores between 2019 and 2020



Paired T-test has been applied to 93 countries with valid national average happiness scores in 2019 and 2020. Although it seems national average happiness scores in 2020 are higher than 2019(estimated: 0.08), we do not have statistically significant evidence proving this difference (p = 0.9). In addition, from the boxplot, we also can see the average and median of national average happiness scores in 2020 seems to be higher than in 2019. The national average happiness scores in 2019 seem to vary in a larger range compared to 2020. It may be can explain by in 2020, people have more time spending with their families, or in their personal lives instead of on traffic or work due to COVID-19.

To evaluate the difference between the year 2019 and the year 2020 in how people evaluating their happiness level, I applied three different linear regression models. The model selection is made by AIC. Our 2019 model has AIC value of -138.14:

Life Ladder =

Our 2020 model has AIC value of -135.46:

Life Ladder =

All estimated coefficients of these two models will be shown in table 2.

Then our third model is using the national average happiness scores as the response variable, and all other variables appeared in the previous two models as the predictors. The third model also includes the interactive term between these variables and a dummy variable for the year 2020 to analyze the difference between these two years.

The result is shown in table 2. From the table, we can clearly see there are three significant interactive terms at the significant level of 0.05. It seems that people evaluate healthy expectancy at birth, log GDP per capita, and negative affect differently between these two years. Based on table 2, in 2019, negative affect is positively associated with the national average happiness score, but it became negatively associated with the national average happiness score in 2020. In 2019, citizens feel more negative effects will tend to make the country have a higher national average happiness score. However, in 2020, this trend becomes inverse. In addition, people attributed health life expectancy at birth less in 2020 but weigh log GDP per capita more compared to 2019 when they are evaluating their happiness level. All other factors seem not to have a significant change in their attributions between these two years.

Table 2: Estimated 2019 Model, 2020 Model, Two-Year Model for Question 3

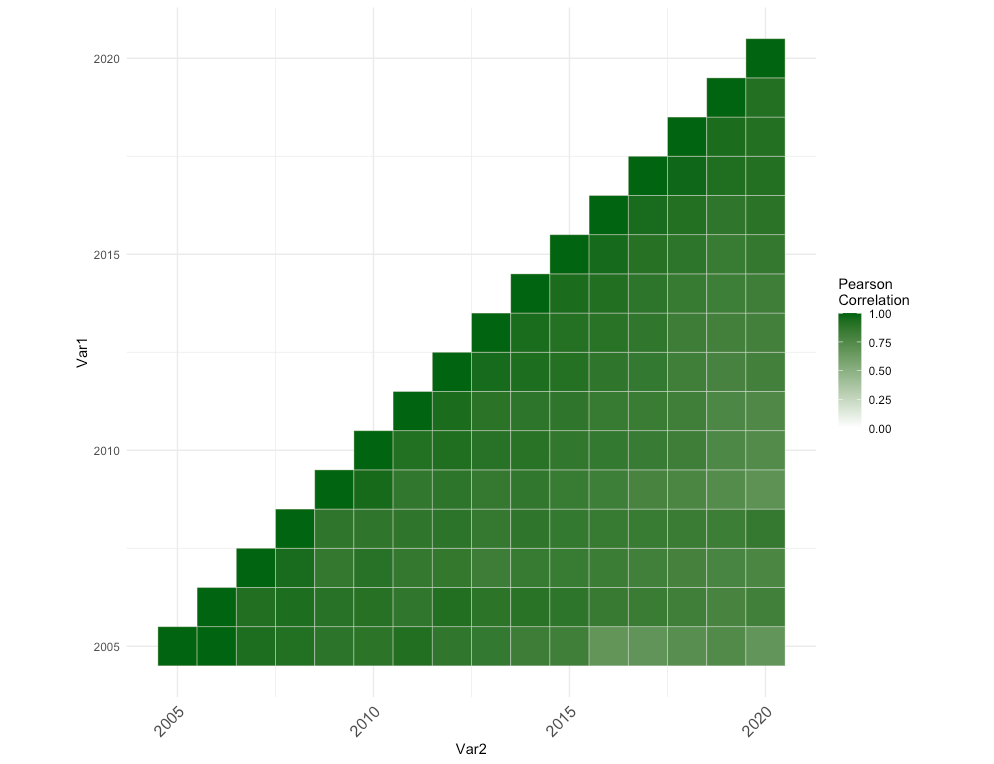
(\*means significant at the 0.05 level)

(Blank means the model in the column does not contain the covariate of that row)

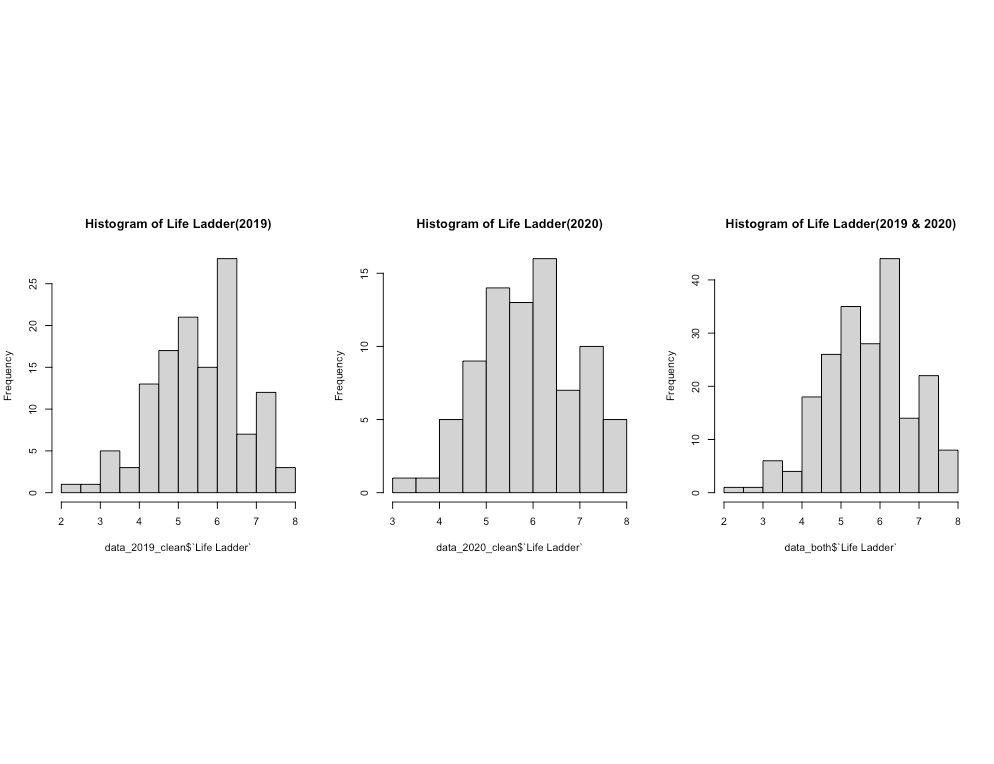
| Variable | 2019 | 2020 | Both |
| --- | --- | --- | --- |
| Intercept | -3.78\* | -2.86\* | -4.12\* |
| Social support | 4.18\* | 2.35\* | 3.79\* |
| Healthy life expectancy at birth | 0.07\* |  | 0.05\* |
| Perceptions of corruption | -1.08\* | -0.60\* | -0.93\* |
| Positive affect | 2.27\* |  | 1.88\* |
| Negative affect | 2.14\* |  | 2.25\* |
| Log GDP per capita |  | 0.56\* | 0.13 |
| Freedom to make life choices |  | 2.07\* | 0.83 |
| Year 2020 |  |  | 1.35 |
| Log GDP per capita x 2020 |  |  | 0.43\* |
| Social support x 2020 |  |  | -1.64 |
| Healthy life expectancy at birth x2020 |  |  | -0.05\* |
| Perceptions of corruption x 2020 |  |  | 0.40 |
| Positive affect x 2020 |  |  | -1.38 |
| Negative affect x 2020 |  |  | -2.78\* |

***Development and Limitations***

In answering the first question, we make strong assumptions that each country’s happiness level is independent, and the factors associated with national average happiness scores do not change over time which may not be valid in this case. The following graph shows the correlation of the national average happiness scores of a country between different years. The correlation is calculated using the pairwise method. Although based on the following graph, the correlation structure of first-order autoregressive may fit better since the observations of different years are positively correlated, the dataset contains too many missing values making it difficult to fit the first-order autoregressive correlation structure in certain variables. Future researches and studies can investigate the correlation between years and may give a better correlation structure to fit the GEE model or applying specialized methods that can accommodate spatial correlation between countries. The following graph also shows that the national average happiness score of the same country in 2019 is also possible to be positively correlated to the score in 2020. Therefore, the required independent assumption for our fitted linear models to answer question three may also be violated in this case.



To do linear regression and paired t-test, we also require the assumption that the national average happiness scores as the response variable to be normally distributed. Based on the following graphs, we cannot really conclude the national average happiness scores in the three models are not normally distributed. Therefore, I do not conclude these assumptions in the three models and paired t-test are not valid.



The paired t-test requires the observations between these two years are sampled independently. Based on the description from the GWP, they did interviews and survey independently in two years with different randomly selected citizens from the participant country. Also, the national average happiness score of the same country in two years should be counted as paired since they are evaluating at the same country. Thus, I think paired t-test is a reasonable approach to answer the second question.

Also, removing the missing data will lead to selection bias. Since the more heavily affected countries by the COVID-19 are more likely to not respond to the GWP survey. In other words, the countries that have a lower national average happiness score due to COVID-19 tend not to respond to the survey. We may have a higher mean national average happiness score in 2020 than the truth is. It may be one of the explanations why we have a higher mean national average happiness score in 2020 compared to 2019.

The change in happiness score may also be explained by factors outside this research focus and dataset. Most of the factors in our dataset are the average of binary questions. Future research can be designed to have multi-level responses to furtherly research the association between different variables and the national average happiness score.

At last, we assume each country’s national average happiness score is independent of other countries, which may not be true. The countries geographically close to each other may somehow correlate with each other in national average happiness scores. We also assume that the national average happiness score of a country but in the different years are independent when we are doing the first question.