What are the contributing factors for Log GDP?

Amber Shao, Haotian Li, Jiaxuan Li, Roxie Wei, Yike Chen

12/9/2020

Introduction

According to the definition published by International Monetary Fund, Gross Domestic Product (abbreviated as GDP) is a measurement of a country's monetary value of final goods and services produced within its for a given time. Compared to its numerical value, the growth rate of GDP seems more important because it could provide not only a horizontal comparison with other country's economic status but also a vertical comparison for its own economy. Thus, this commonly used index, Log(GDP), has been considered as "one of the most important indicators of the general health of the economy". Owing to its significance, our group decided to analyze the Log GDP value of 120 countries in 2018 provided by the World Bank and all the possible factors that affect these values. Among all the aspects we explored, we find several promising factors that may impact the Log GDP: the country's population, geographic location, labor, and people's life expectancy of the country.

Exploratory Data Analysis:

During the data exploration stage, we are finding our general exploratory directions by trying to answer the following questions.

1. What topic in general are we looking at? GDP? Environmental protection? Living Standard?
2. What may be a contributing factor of the topic we want to discuss? what problem may we meet? How do we solve it?
3. To what extent do demographic impact a country's GDP?
4. Can we measure one country's economy by any indicator? If so, do we have it in the graph or we need multiple steps of calculation to obtain it using the data we have?
5. Are there any bias/influential aspects in the responses that could skew data? If so, among them, which has the comparatively the most impact on the data?
6. What is the relationship between certain variables?

After carefully explore the data, we finally set our topic to be GDP and here are the variabled chosen for our specific and detailed exploration graphics.

- GDP (Quantitative): As mentioned in introduction, GDP measures a country's monetary value of final goods and services produced for year 2018. This data is given by the world bank data in the column called GDP. However, since we interest in the average GDP for each continent and in the growth rate, we have two other aggregated data from the GDP column.

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's 
## 1.770e+09 2.663e+10 1.003e+11 6.987e+11 4.220e+11 1.790e+13         6 
```

- Continent (Qualitative): This variable contains the geographic information for each country. For example, United States locates in North America Continent. Therefore, it is labeled North America in the Continent column. Since this piece of information is not provided by the World bank, we collected information from

United Nation and use it to conduct further explorations. There are six continents in total, which is shown in the chart below

- Average GDP (Quantitative): This variable measures the average GDP for each continent. We obtain this data by first filter out the GDP for each continent. In this process, there are few countries that do not have GDP so we use na.rm to remove them. Then, we find the mean of GDPs for each continent. The data is presented below. Among all six continents, North America has the highest average GDP while Africa has the least.

```
##       Continent  GDP_Average
## 1        Africa 8.774767e+10
## 2          Asia 8.864308e+11
## 3        Europe 7.163176e+11
## 4 North_America 1.946778e+12
## 5       Oceania 2.818424e+11
## 6 South_America 3.864478e+11
```

- Log GDP (Quantitative): This variable measures the growth rate for each GDP for each country. We obtain this data by first collect the GDP for each country, then apply the algorithm Log() to all the GDPs. In this process, there are few countries resulted in NA so we use !is.na to remove them.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.29   24.01   25.32   25.51   26.77   30.52
```

- Telephone Usage on Average (Quantitative): This variable measures the telephone usage for each continent. We obtain this data by first filter out the telephone usage for each continent. Then, we find the mean of telephone usage for each continent. The data is presented below. Among all six continents, North America has the highest average GDP while Africa has the least.

```
##       Continent Telep_ave
## 1        Africa   1026693
## 2          Asia  12330496
## 3        Europe   7869987
## 4 North_America  13845113
## 5       Oceania   2905069
## 6 South_America   6470076
```

- Female Life Expectancy (Quantitative)： This variable is used to measure the expected life span of female in each country. That is how long on average would we expect a female live in this country. We filtered the life expectancy of female in 120 countries through the World Bank data table.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   54.99   72.01   78.23   76.51   82.75   87.70
```

- Male Life Expectancy (Quantitative): This variable is used to measure the expected life span of male in each country. That is how long on average would we expect a male live in this country. We obtain this data by filtering the life expectancy of male in 120 countries through the World Bank data table.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   50.65   66.66   72.93   71.63   77.31   82.30
```

- Total Life Expectancy (Quantitative): This variable is used to measure the expected life span of a human being in each country. That is how long on average would we expect a human being live in this country. We obtain this data by filtering the total life expectancy in 120 countries through the World Bank data table.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   52.80   69.66   75.53   74.03   79.80   84.93
```

- Labor Per Capita (Quantitative): This cariable is used to measure the ratio of labor force older than 15 in a country's population. This information is obtained by dividing the labor column by the pop_total column

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.2297  0.4111  0.4760  0.4640  0.5256  0.7012
```

Exploratory Graphics

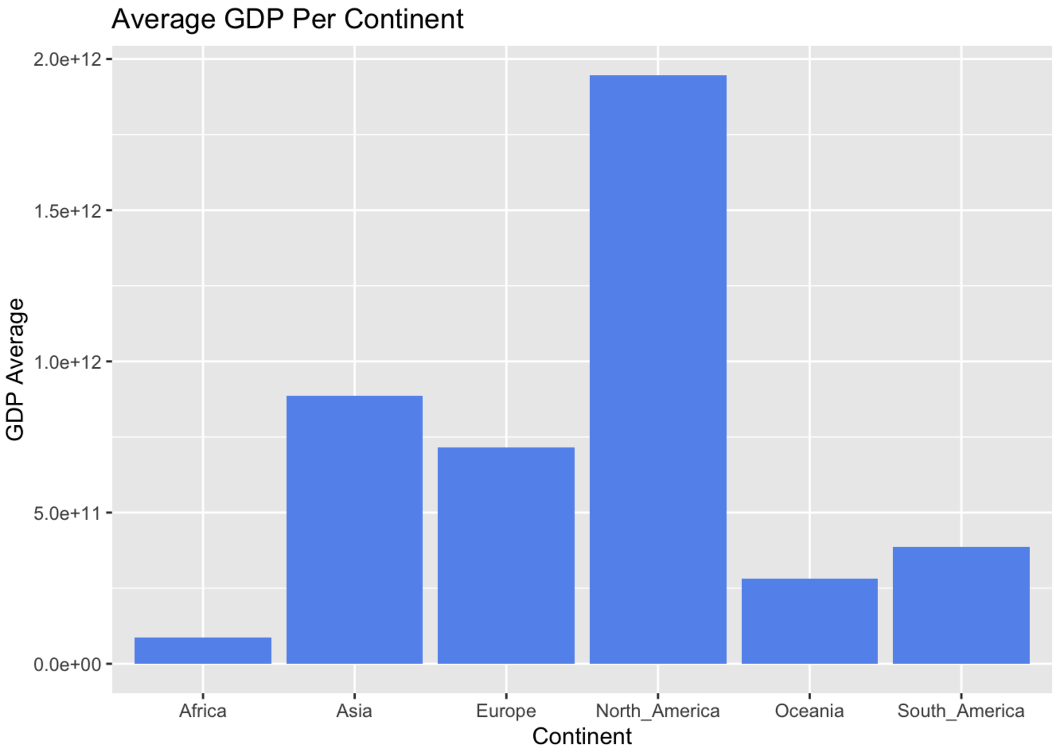


Figure 1: By Yike Chen

The data shows that there are a total of 120 countries. To facilitate the compilation of the data, we attribute the countries into six categories based on their geographic locations: Africa, Asia, Europe, North America, Oceania, and South America. Figure 1 shows the average GDP values of these six continents. As can be seen from the figure, North America, Asia, Europe are among the top three, among which the average GDP of South America far exceeds the data of other continents. On the contrary, Africa's average GDP is at the bottom of the ranking. Curious about the reason behind this considerable GDP gap, we will continue to study what are the contributing factors of GPD.

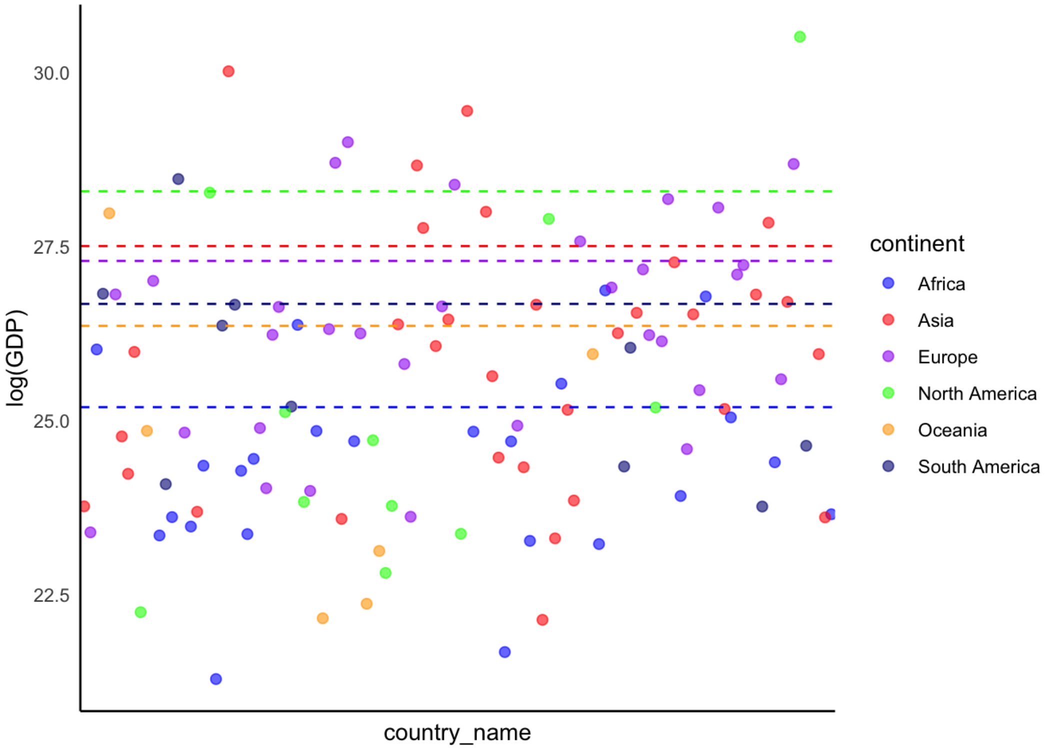


Figure 2: By Roxie Wei

GDP of all countries in the world is plotted in the Figure 2 above and colored by continent. To manifest the effect of continents on GDP, we draw reference lines of average GDP of all continents in the same plot. It is seen that North America ranks the first in terms of GDP while Africa ranks bottom, which is consistent with our expectation and Figure 1. At the same time, the average GDP of other continents, including Asia, Europe, Oceania, and South America, are close to each other. Moreover, the distribution of different countries' GDP from the same continents is evenly distributed along the average line. Therefore, we suspect that the location of a country would affect GDP. To further investigate this question, we conduct a hypothesis testing in part 3. It is noted that for simplicity, we select two samples deviates most from the average GDP of the world, including Africa and North America, to do the test.

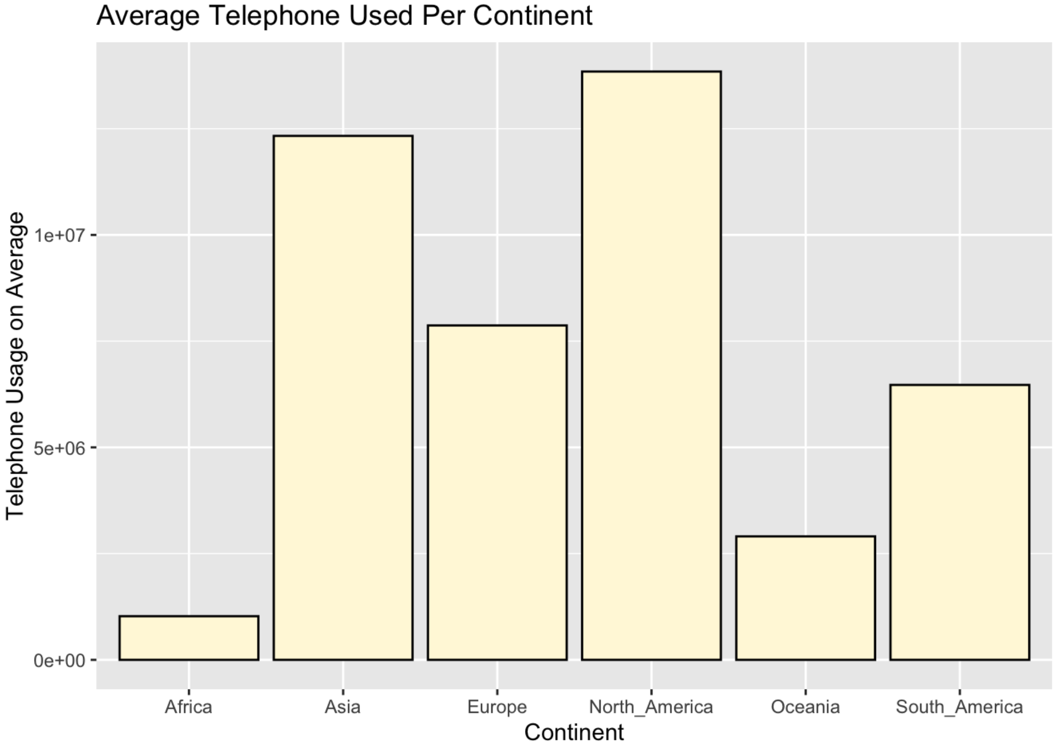


Figure 3 By Jiaxuan Li

Exploring whether GDP and living conditions are correlated, we draw the bar graph for average telephone used per continent. From this chart, we can see that the mobile phone usage rate in North America, Asia and Europe far exceeds that of other continents. Comparing to the Figure 1, we observe that these three continent also has the top three GDP average. Thus, we conjecture that this increment in the average telephones usage could be correlated with GDP. Being more specific, it is possible that high economic status, which is indicated by high average GDP, lead to this high telephone usage. However, it is also possible that the high usage of mobile phone lead to other factors, like easier communication when doing business or higher level of manufacture, has driven the overall economic and technological development of this continent, thereby bringing GDP to a new height. According to the 2018 global smart phone shipment statistics released by market statistics company Counterpoint, we learn the fact that the top three global mobile phone sales are Samsung, iPhone, and Huawei. All three brand originate from Asia and North America Continent, which correspond to the trend observed in the figure above.

```
##      male_r  female_r
## 1 0.4868874 0.4943384
```

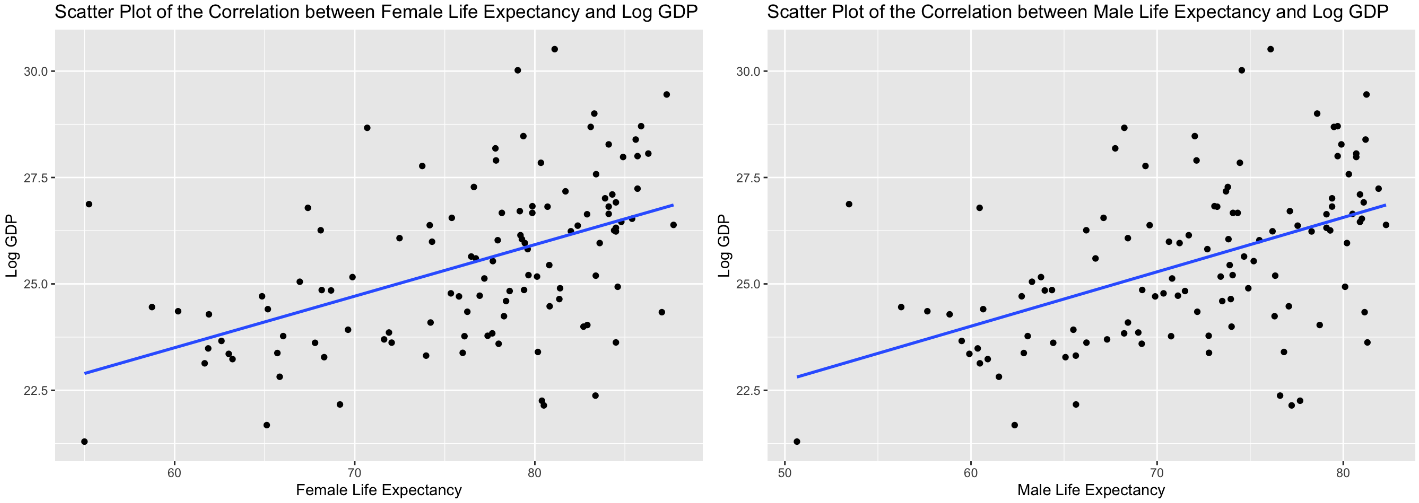
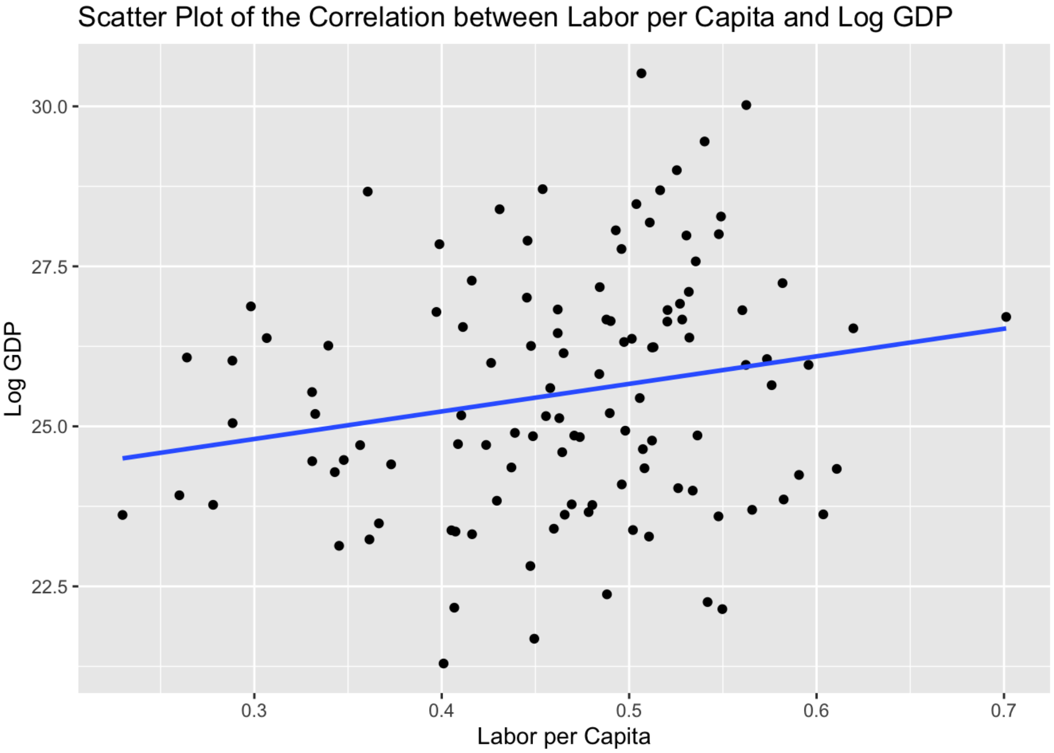


Figure 4: By Amber Shao

Figure 4 composes two linear regression models: one is for analyzing the correlation between the female life expectancy and the Log GDP in 2018; the other one is for analyzing the correlation between the male life expectancy and the Log GDP in 2018. In this analysis, all variables used are quantitative. According to the calculation presented in the Rmd file, both correlation coefficients female_r = 0.4943384 and male_r = 0.4868874 indicate pisitive correlation between the explanatory and response variables: As the life expectancy for male/female increases, on average, the log GDP increases as well; Vice versa. However, this positive correlation does not indicate any causation relationship between the two variables. Furthermore, comparing the regression analysis and the r, we observed a small difference between female and male life expectancy. Thus, we will continue explore and investigate on whether this difference is statistically significant on part 3.



```
##
## Call:
## lm(formula = loggdp ~ labor, data = laborvsgdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.943 -1.546 -0.111  1.274  4.824
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  23.5102     0.9481  24.796   <2e-16 ***
## labor         4.3068     2.0025   2.151   0.0337 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.861 on 112 degrees of freedom
## Multiple R-squared:  0.03966,    Adjusted R-squared:  0.03109
## F-statistic: 4.625 on 1 and 112 DF,  p-value: 0.03365
```

```
## [1] 0.1991476
```

Figure 5: By Haotian Li

Since we want to see whether there exists a correlation between labor per Capita and Log GDP, we decide to draw a scatter plot with a regression line to see the overall relationship. In Figure 5, we use labor per capita as the explanatory variable and the Log GDP as the response variable. As we calculated, r = 0.1991476, which is considered as a weak positive correlation between these two variables. However, this does not imply any causation between these two variables. Because r = 0.1991476 means a fairly weak correlation, we would like to conduct a hypothesis test in part 3 to discuss whether there exist such a correlation between Labor per Capita and Log GDP.

Formal Data Analysis - Contributing Factors to Log GDP

Geography

Inspired by Figure 1, 2, and 3, we decide to conduct a hypothesis test on whether geography affects GDP. Here, since we focus more on the GDP averages rather than GDP growth, we would use the GDP rather than Log GDP.

Hypothesis Test on Whether Continent Affects GDP

```
##
##  One Sample t-test
##
## data:  log_africa_gdp_no_na
## t = -4.1275, df = 23, p-value = 0.0004097
## alternative hypothesis: true mean is not equal to 25.51462
## 95 percent confidence interval:
##  23.75260 24.92915
## sample estimates:
## mean of x
##  24.34087
```

```
##
##  One Sample t-test
##
## data:  log_europe_gdp_no_na
## t = 2.9739, df = 30, p-value = 0.005754
## alternative hypothesis: true mean is not equal to 25.51462
## 95 percent confidence interval:
##  25.77293 26.90538
## sample estimates:
## mean of x
##  26.33915
```

```
##   continents se_continent z_continent  p_continent less_than_cl
## 1     Africa   0.05682562   -20.65526 8.754716e-95         TRUE
## 2     Europe   0.04898629    16.83192 0.000000e+00         TRUE
```

Null hypothesis: Which continent a country is does not affect its GDP.$\bar{x} = \mu$

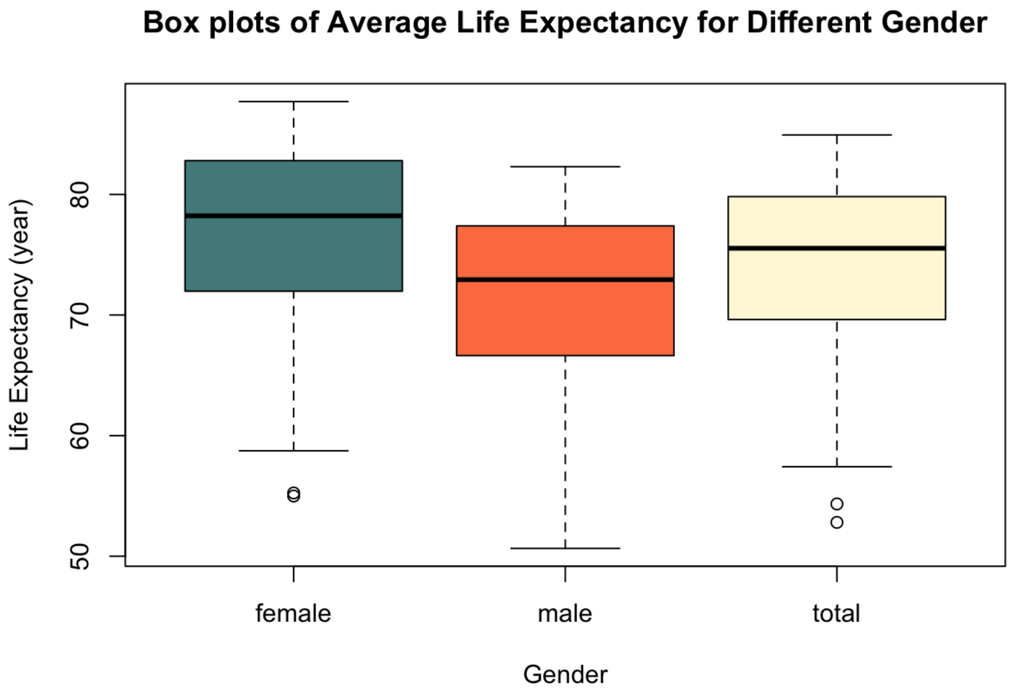
Alternative hypothesis: Which continent a country is does not affect its GDP$\bar{x} \neq \mu$

predetermined confidence level: 0.05, which is 5%

Then the z-score for African countries' GDP is -20.65526, which gives a p-value close to 0 using pnorm function $pnorm(-20.65526) * 2$, while the z-score for European countries' GDP is 16.83192 yielding a p-value close to 0 using pnorm function $(1 - pnorm(16.83192)) * 2$. Noting that both p values is far less than the predetermined confidence level 0.05, we conclude that there is enough evidence for us to reject the null hypothesis in favor of the alternatives. Therefore, we think that geographic factors, like which continent a country is located, has a impact on countries' GDP growth.

Life Expectancy:

Curious about the impact of life expectancy on Log GDP, we took a closer look at the world bank data and found an interesting fact: the data regarding life expectancy is presented in three categories, which are female, male, and total. Plotting the distributions of these three categories of data and calculating their means, we observed a noticeable difference between average males' and females' life expectancy. It is possible that this difference in average life expectancy between males and females is due to chance. Thus, to investigate whether this statement is true, we decided to perform a hypothesis test first and then conduct our regression analysis.



Hypothesis Test on the difference between Male and Female Life Expectancy

```
##   Gender     Mean       SD        SE
## 1 female 76.51231 7.663472 0.6995761
## 2   male 71.63431 7.178113 0.6552691
```

This is an two-sided, two sample, z-test for difference between means

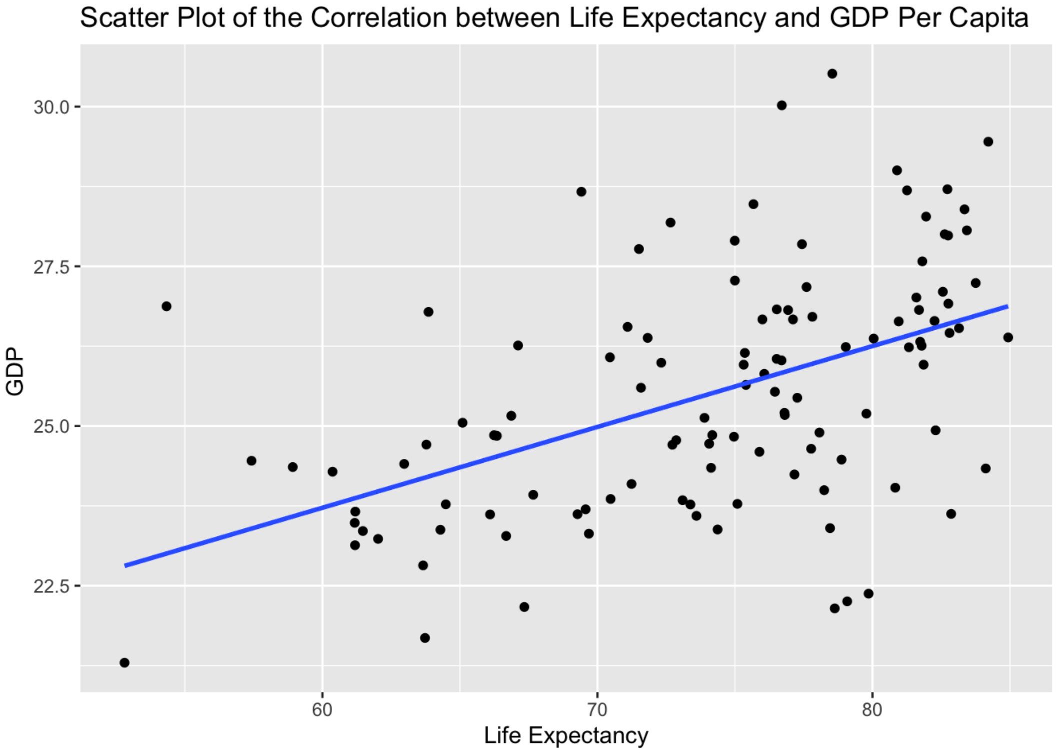
- H0: There is no statistical significant difference between average female and males’ life expectancy. That means, the difference in average life expectancy observed is due to chance. That is, X_bar_female = X_bar_male.
- H1: There is a statistical significant difference between average females’ and males’ life expectancy. That means, the difference in average life expectancy observed is not due to chance. That is, X_bar_female ≠ X_bar_male.
- predetermined confidence level:5%, which is 0.05

```
##          se       z      p_value
## 1 0.9585324 5.08903 3.598995e-07
```

From the computation above, we obtain this extremely low p-value 3.598995e-07, which is approximately 0. Since it is lower than the predetermined confidence level 0.05, there is enough evidence to reject the null hypothesis in favor of the alternatives. This rejection of the null hypothesis means that it is highly likely that the difference between average life expectancy in male and female cannot be explained by chance variation. And from the box plot drawn above, it is reasonable to infer that female has longer life expectancy than male. Therefore, using either variable to conduct the regression analysis is unfair. Thus, in order to better analyze the correlation between life expectancy and Log GDP, we decided to use the life_expectancy_total column as our index for life expectancy.

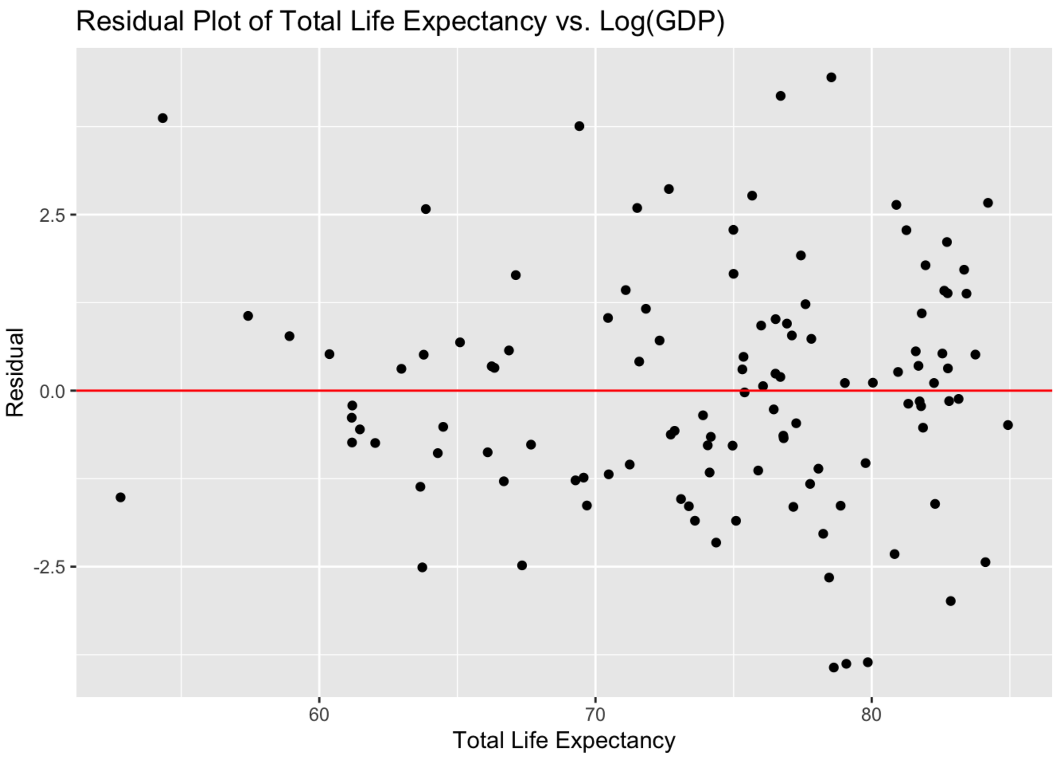
Regression Analysis:

drawing out the data and conducting the regression analysis, we found a positive correlation between the total life expectancy and Log GDP for these countries. r = 0.493734 suggests a comparatively strong association between the total life expectancy and the Log GDP. Being more specific, as the total life expectancy increases, the log GDP on average increases as well. However, since r will not change if we switch the explanatory and response variable, another possible interpretation of this positive correlation is as the Log GDP increases, on average, the total life expectancy increase as well. Similar to the analysis done to figure 2, this positive correlation does not imply any causal relationship between these two variable.



```
##
## Call:
## lm(formula = y ~ x, data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9324 -1.0940 -0.0714  0.9452  4.4499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.12887    1.56968  10.275  < 2e-16 ***
## x            0.12652    0.02106   6.009 2.38e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.651 on 112 degrees of freedom
## Multiple R-squared:  0.2438, Adjusted R-squared:  0.237
## F-statistic:  36.1 on 1 and 112 DF,  p-value: 2.378e-08
```

```
##
##  Pearson's product-moment correlation
##
## data:  x and y
## t = 6.0086, df = 112, p-value = 2.378e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3407621 0.6212374
## sample estimates:
##      cor
## 0.493734
```



Drawing out the residual plot, we observed that the residual plots distribute randomly and do not display any obvious pattern. Therefore, we conclude that a linear regression model is appropriate. Thus, we could use the formula y_hat = m_hat * xi + b_hat learned in class to estimate the parameter y_hat, the expected Log GDP for certain total life expectancy. Being more specific, that is log GDP = 16.128 + (0.1265 * life expectancy total)

Parameter Prediction:

The following chart is the predicted y_hat using the linear regression model.

```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##     16.1289       0.1265
##   life_expectancy_total predicted_log_GDP
## 1                    60          23.72030
## 2                    65          24.35292
## 3                    70          24.98554
## 4                    75          25.61816
## 5                    80          26.25078
```

Labor Per Capita:

Hypothesis Test: Since the correlation coefficient appears to show a weak correlation, we wonder whether this correlation really exist. Therefore, we would like to conduct a hypothesis test on the coefficient.

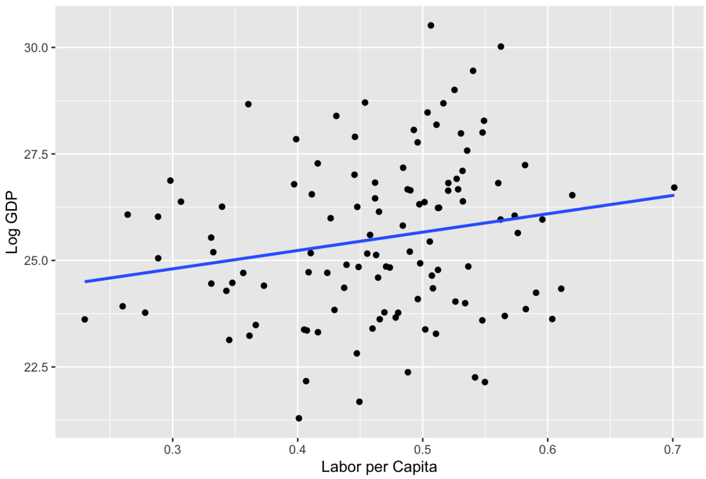
- Null Hypothesis(H0): There is no correlation between the labor per Capita and the Log GDP. The correlation observed is purely due to chance.
- Alternative Hypothesis(H1): There is a correlation between the labor per Capita and the Log GDP.
- Predetermined confidence level = 5%, which is 0.05.

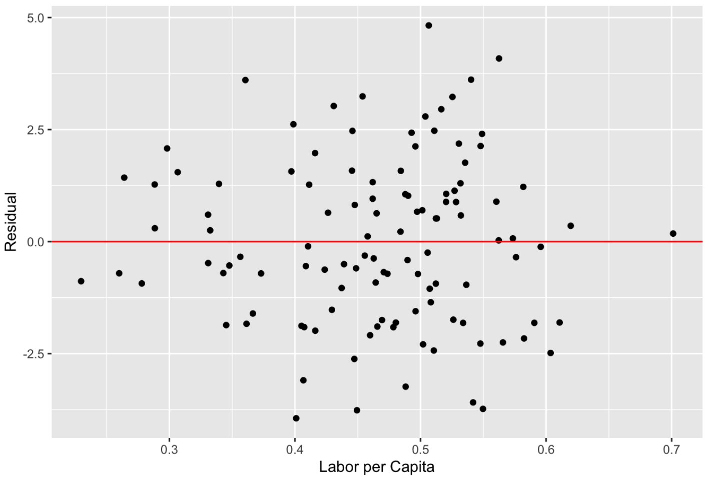
```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 23.510231  0.9481439 24.796057 2.676051e-47
## labor        4.306772  2.0025361  2.150659 3.365283e-02
```

According to the computation above, we can see that the t = 2.150659 and p_value = 3.365283e-02. Since the p_value is less than the predetermined confidence level 0.05, there is enough evidence to reject the null hypothesis in favor of the alternatives. This rejection of the null hypothesis means that even though r = 0.1991476 represents a fairly weak correlation, this positive correlation between the labor per capita and the log GDP is not due to chance. Therefore, we can further conduct the regression analysis and parameter prediction.

Regression Analysis:

```
## 
##  Pearson's product-moment correlation
## 
## data:  laborvsgdp$labor and laborvsgdp$loggdp
## t = 2.1507, df = 112, p-value = 0.03365
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01581192 0.36952799
## sample estimates:
##       cor 
## 0.1991476
```


Scatter Plot of the Correlation between Labor per Capita and Log GDP


Residual Plot of Labor per Capita vs. Log(GDP)

Looking at the residual plot, we find that the the residual plots distribute randomly in a oval shape. This indicates that the regression analysis could be a good model. Thus, we would like to construct the parameter predictor. That is Log(GDP) = 23.510 + (4.307 * Labor per Capita)

Parameter Prediction:

The following chart is the predicted log GDP using the linear regression model.

```
## 
## Call:
## lm(formula = laborvsgdp$loggdp ~ laborvsgdp$labor)
## 
## Coefficients:
##      (Intercept)  laborvsgdp$labor  
##           23.510             4.307
```

```
##   Labor_per_Capita predicted_log_GDP
## 1             0.30          24.80226
## 2             0.35          25.01760
## 3             0.40          25.23294
## 4             0.45          25.44828
## 5             0.50          25.66362
## 6             0.55          25.87896
## 7             0.60          26.09429
```

Conclusion

Starting from exploring the provided dataset, we eventually narrow down our target contributing factors to GDP growth to three categories: location, the number of laborers, and life expectancy of the population. We think that where a country is located may impact its GDP growth rate from our initial data exploration. To testify our hypothesis, we conduct a hypothesis test that results in rejecting the null hypothesis in favor of the alternative that the location of a country has an impact on its GDP. Built upon this conclusion, we guess that the difference in natural resources may explain this difference. For example, According to Diane Boudreau’s article on introducing the natural resources of North America, North America benefits greatly from its fertile soils, plentiful fresh water, oil and mineral deposits, and forests." All these factors mentioned in the article may explain why North America has the highest GDP in 2018, as illustrated in figure 1.

Verified the first factors, we move on to other quantitative variables: the labor per capita and total life expectancy. The hypothesis test we conducted verifies that the correlation coefficient r = 0.1991476 is statistically significant for labor per capita. Then, we explore the regression analysis and construct a parameter prediction mechanism to predict the GDP growth given labor per capita. Addressing the positiveness of the coefficient, we hypothesize that promoting more people in the population to work (increase the labor per capita) may positively impact a country's GDP growth rate. Of course, the recommendation's uncertainty comes from the fact that there should not be any causal relationship drawn from a correlation coefficient.

For life expectancy, we first discover through hypothesis test that there is a difference between females' and males' life expectancy. Therefore, we decide to use total life expectancy as the explanatory variable to conduct the regression analysis. Finding out r = 0.493734, we also explore the regression analysis by drawing out the residual plot and construct a parameter predictor for finding the GDP growth given total life expectancy. Based on this reasonably strong coefficient, we also hypothesize that the high GDP growth may allow countries to invest more and improve their living condition such as health care system, public infrastructure, and environment, which all contribute to a high life expectancy.

Even though we obtained several conclusions from our data analysis and exploration, some drawbacks may impact the research's accuracy. One limit would be the absence of data. During the research, we need to remove some N/A data resulted from computation in almost every step. This may result in some extent of data loss, especially those influential points. However, since regression analysis is particularly susceptible to those influential points, the absence of such points may drastically change the results. Another constraint we met is the lack of analysis models. For example, when exploring data, we found an exponential relationship between certain two variables. However, unknowing the specific analyzing method for an exponential relationship, we have to focus on other aspects. All these limitations, along with those interesting hypotheses resulting from our data analysis, motivate us to explore data and statistics.

Reference:

Boudreau, Diane Boudreau, et al. "North America: Resources." National Geographic Society, 9 Oct. 2012, www.nationalgeographic.org/encyclopedia/north-america-resources/.

Callen, Tim. "Finance & Development." Finance & Development | F&D, 24 Feb. 2020, www.imf.org/external/pubs/ft/fandd/basics/gdp.htm.

Team Counterpoint. "Global Smartphone Market Share: By Quarter." Counterpoint Research, 24 Nov. 2020, www.counterpointresearch.com/global-smartphone-share/.

"UNSD - Methodology." United Nations, United Nations, unstats.un.org/unsd/methodology/m49/.