International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:07, Issue:06 "June 2022"

A REGRESSION ANALYSIS ON THE GLOBAL ECONOMY

Michael Cho

Hong Qiao International School - Rainbow Bridge International School

DOI: 10.46609/IJSSER.2022.v07i06.018 URL: https://doi.org/10.46609/IJSSER.2022.v07i06.018

Received: 28 June 2022 / Accepted: 5 July 2022 / Published: 8 July 2022

ABSTRACT

The economics of a nation has usually been regarded as one of the key indications of a nation's compressive national power and as a factor to gauge a country's quality of life. In this study, I use a linear regression model to investigate the statistical connection between many national characteristics and national GDP levels. According to the study's findings, the technology factor and the health factor have the biggest beneficial effects on economic development among all the other potential national characteristics. 2. Significantly hindering economic growth are the birth rate and agricultural-related factors. A comprehensive statistical analysis is conducted in the result section to consolidate the linear regression conclusion.

Keywords*:* Linear regression, economic development

1. Motivation

As a factor to measure the quality of life, the economy of a country has been considered one of the most crucial indicators of a country's compressive national power. Economics studies over different countries thus help us to better understand not only our country's position in the world but also the relative welfare level from a bigger picture. Since graduating from middle school, I have become increasingly interested in understanding how the economy operates. After reading an academic poster qualitatively analyzing the driving forces behind economic development, I developed an idea to quantify the impact of each factor using the statistical knowledge I learned from the AP courses in statistics and microeconomics. Specifically, by using a linear regression model, this study analyzes the statistical correlation between national characteristics and national GDP. The conclusion of this study hopes to provide quantitative evidence to a different model of economic development, which will help us to understand how the economy operates and further provide useful suggestions for policymakers to make feasible development plans that consider the characteristics of the nation.

2. Relevant Economy Data

Gross domestic product (GDP), measuring a nation's total output over the span of a season or year, is one of the most widely used national economy representations. The calculation of GDP considers every aspect of the national output. No matter whether the output is tangible goods (such as food or vehicles) or intangible goods (such as service or music), it is converted into monetary values and added to the GDP calculation. By investigating more carefully each of the aspects of GDP calculation, I found the following potential contributor to national GDP:

- Education level: Education raises people's capability in GDP contribution. Residents with higher education backgrounds will promote the cost of human capital, which in turn stimulates economic development.

- Crop production: For certain counties, crop production is one of the major contributors to their national GDP. However, crop production is playing a less important role in modern society, especially in developed countries.

- Population: people of different ages play very different roles in economic development. Due to the different high consumption power, it is often agreed that a high percentage of the labor population within the society is important for economic development.

- Coastline ratio: Due to the business convenience brought by the geographical location, the cities near the sea are often considered more economically important. Consequently, if a country has a high coastline ratio, it might imply convenience in international transportation, which facilitates economic development.

- Technology: Technology has been considered the primary productive force in the modern economy. Historically, the technology revolution tends to generate new business industries, which further brings new jobs for economic development.

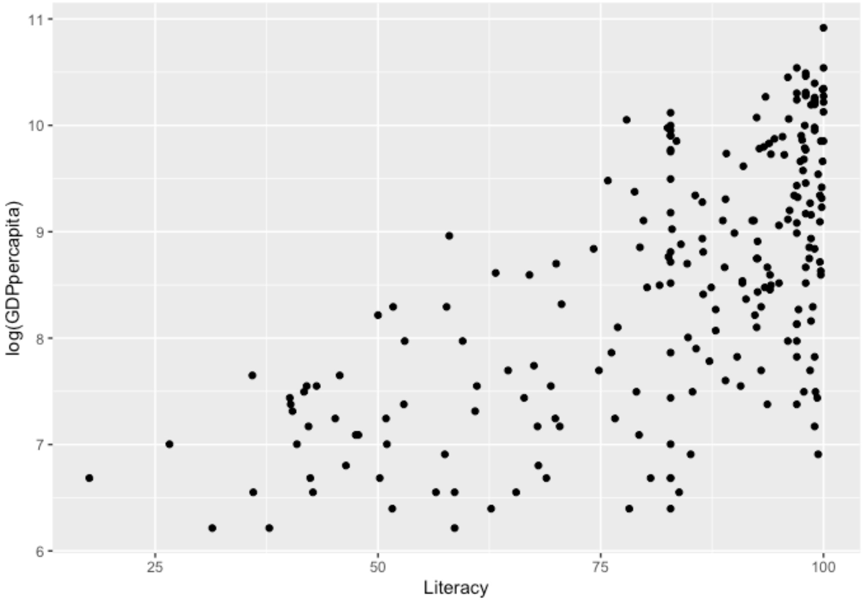
Comprehensive data containing all the above factors are difficult to obtain. The US Census Bureau once published the relevant statistics of all the major counties in 2006. The statistics cover19 different aspects of nearly 200 countries. Although the data is from 2006, it contains many important economic dependence factors that can be conceptually borrowed in modern economic development. Taking the cell phone access rate as an example, in 2006, the cell phone was not widely accessible in many counties. A country with a high percentage of the population with access to cell phones can be considered as the 2006 technology entity. In this study, I expect the technology entity might have changed in the year 2022 but the importance of the technology factor has not.

3. Data Visualization

In this section, I provide some visualization of the country dataset published by US Census. Specifically, to visually inspect the economic contribution of each of the factors, I plotted the log(GDP) concerning different factors.

- Education factor

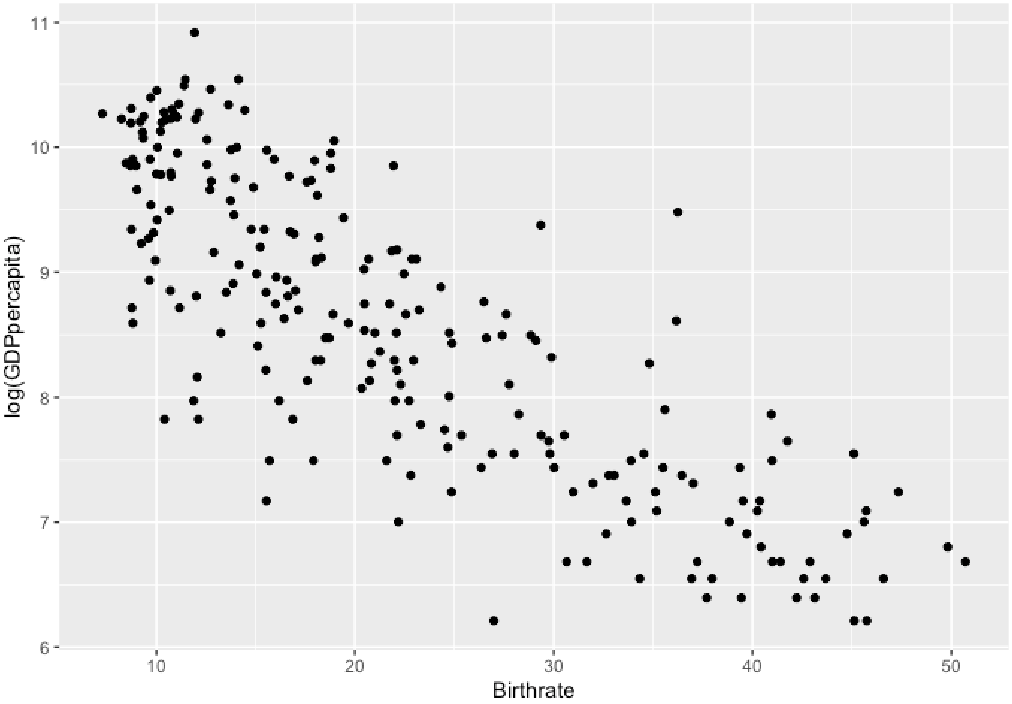
To investigate the hypothesis that "Residents with higher education background will promote the cost of human capital, which in turn stimulates economy development", I first plotted the GDP against the literacy ratio of the 200 counties:



As we can see from the above plot, the literacy ratio tends to have a positive effect on GDP development in the sense that the higher the literacy ratio within the country, the higher the GDP is of the country. The result is consistent with my background research that education frequently increases productivity and creativity, which diversifies the GDP contribution.

- Fertility factor

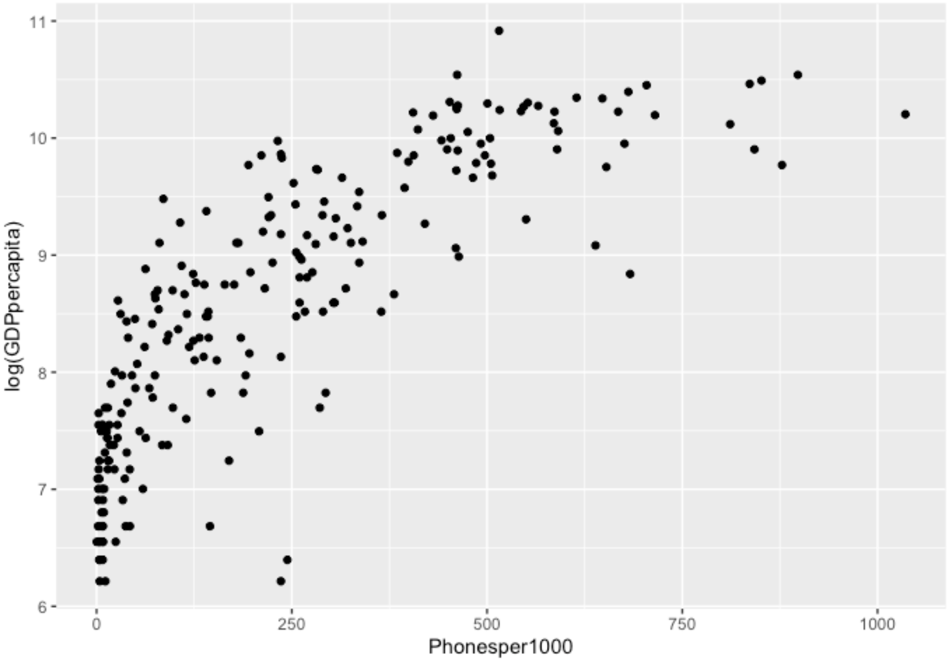
Lower birth rates are linked to slower population growth, faster population aging, and slower economic growth. To investigate the birth rate effect on economic development, I plotted the log of GDP against the birth rate, which is defined as the number of births per thousand of the population per year.



However, As the above plot shows, the birthrate has a clear negative effect on GDP development. After some background research, I found that a moderate birth rate that is sufficient to maintain population growth is best for economic growth, however, an excessive birth rate usually indicates 1.high dependence on the labor-intensive industry, 2. Low society welfare and 3 moderate education resources.

- Technology factor

Economic development also depends on the more efficient production of more and better goods and services, which is made possible by technological advancement. Back in 2006, access to mobile phones can be considered the advancement of technology. Below, I plot the association between phone access rate and the log of national GDP:



Here, GDP development is positively correlated with phones per 1000 people, exhibiting that the more people with phones within the country, the higher the GDP is of the country. However, the technology factor tends to have a smaller positive effect after a certain threshold. In this data example, after the threshold of 500 phone access rate, the contribution to log(GDP) becomes much smaller as it is compared at the level of 50 phone access rate. To quantitatively evaluate the contribution of each of the factors, I studied the linear regression model, which has been widely used in economic analysis.

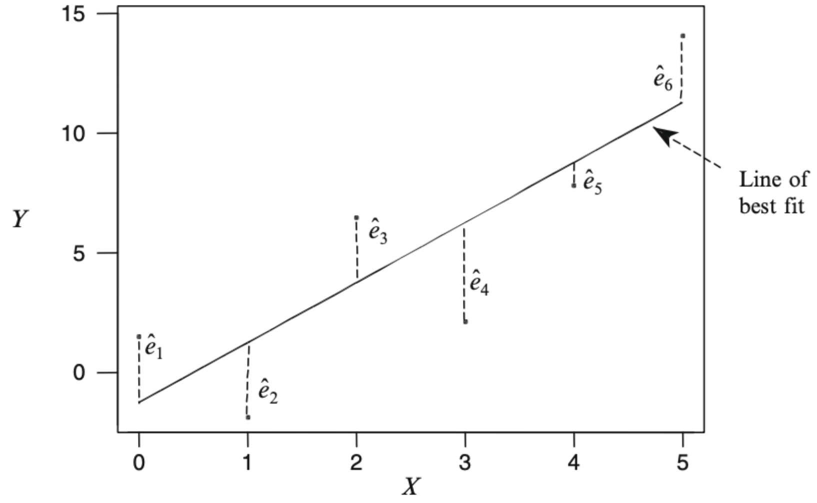
4. Model

4.1 Model optimization setup

In this section, I summarize the basic model setup from an optimization perspective. Taking a bivariate regression as an example, the model aims at finding an interceptβ_0 and a slopeβ_1 parameter that minimizes the distance between observations and the model.

$$Y = \beta_0 + \beta_1 X_1$$

where Y is the GDP level and X_1 is the variable that is linearly correlated to the GDP levels.



Mathematically, the optimal intercept β_0 and slope β_1 can be found via the following:

$$\widehat{\beta_0},\widehat{\beta_1} = argmin_{\beta_0,\beta_1}\sum_{i=1}^{n}(Y_i-\beta_0-\beta_1X_{1i})^2 = argmin_{\beta_0,\beta_1} f(\beta_0,\beta_1)$$

Using calculus to take the derivatives of $f(\beta_0,\beta_1)$ w.r.t β_0,β_1, we can solve two linear equations$\frac{dL(\dots)}{d\beta_0}=0$, $\frac{dL(\dots)}{d\beta_1}=0$ with two unknowns (β_0,β_1):

$$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}X_iY_i - n\,\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\,\bar{X}^2}$$

$$\widehat{\beta_0} = \bar{Y} \text{ - } \widehat{\beta_1}\bar{X}$$

Following the same logic, one can thus generalize the model to multiple regression, which has p different variables $(X_1, X_2, \ \dots X_p\ )$ that are linearly correlated to GDP level.

$$\vec{\beta} = argmin_{\vec{\beta}}\sum_{i=1}^{n}\left(Y_i-\beta_0-\beta_1X_{1i}-\beta_2X_{i2}-\cdots\beta_pX_{ip}\right)^2 = argmin_{\vec{\beta}} f\left(\beta_0,\beta_1,\dots\beta_p\right)$$

The model is thus assuming one unit of change of X_1 will change β_1 unit of Y.

4.2 Model statistics setup

However, the optimization setup in the previous section has limited practical implication. For example, one can gather some randomly sampled numbers as the covariate $X_1, \dots X_p$ and solve the optimization to obtain $\beta_0, \beta_1 \dots \beta_p$. To take the randomness of the data into consideration, a solution from a statistics perspective has been proposed, which assumes

$$\log(Y_i) = \beta_0 - \beta_1 X_{1i} - \beta_2 X_{i2} - \cdots \beta_p X_{ip} + \epsilon_i$$

Where:

- Y is the GDP level
- $X_1, X_2, \ \dots X_p$ are p variables relevant to GDP level
- $\epsilon \sim N(0, \sigma^2)$ is the residual, which is assumed to follow normal distribution

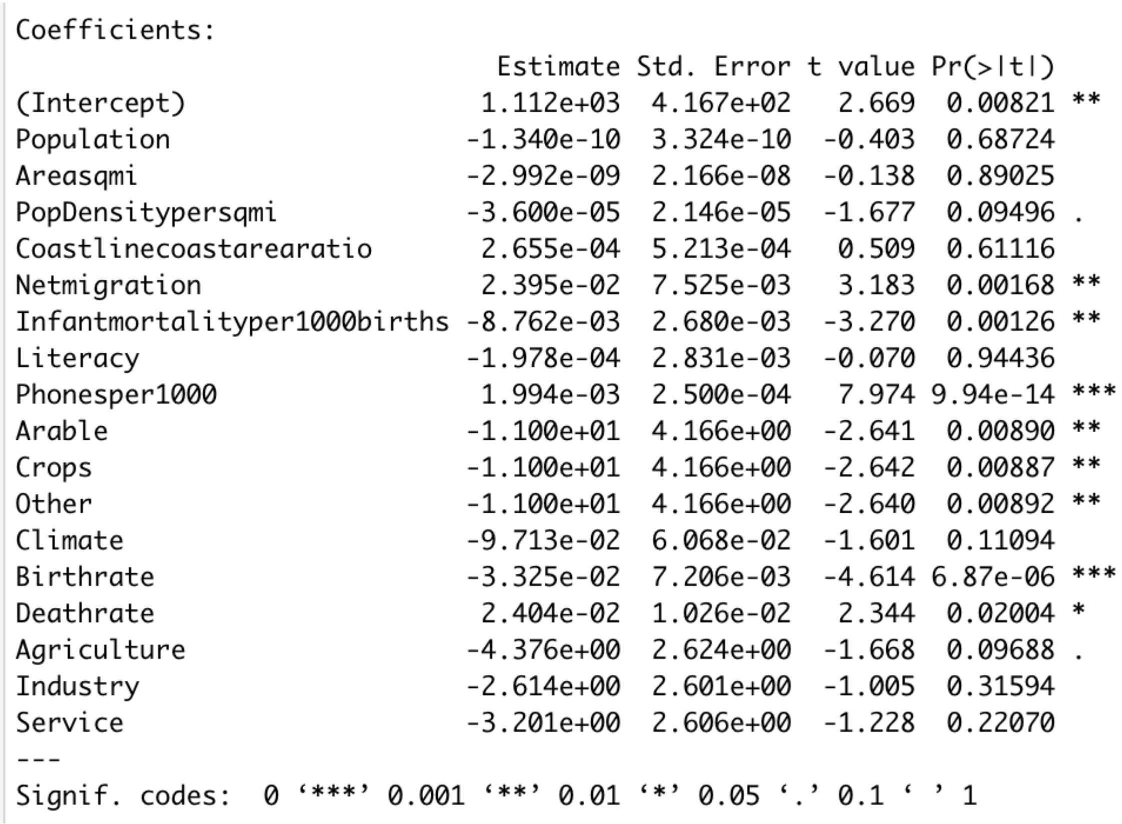
Instead of finding the $\vec{\beta}$ that minimizes the distance, the model solves the problem by maximizing the probability of data occurrence:

$$argmax_{\vec{\beta}} \prod_{i=1}^{n} P(Y_i|X_i) = argmax_{\vec{\beta}} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} ex\,p\left(\frac{\log(Y_i) - \beta_0 - \beta_1 X_{i1} \dots - \beta_p X_{ip}}{-2\sigma^2}\right)$$

It turns out that the $\vec{\beta}$ that maximize the likelihood is equivalent to the $\hat{\beta}$ that we obtained in section 4.1. However, this probability interpretation provides us with an extra parameter estimate σ^2, which can be quite useful to construct the student t statistics to test if the model is effective in explaining log(GDP). The calculation of the t statistics is implemented in R summary functionwith the interpretation that a large derivation of 0 represents the effectiveness of the factor.

4.3 Model solution and interpretation

To implement such a model, I used R to solve the optimization and obtain the solution. A summary of the coefficient estimates, and their corresponding t statistics areobtained in the following plot:



```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  1.112e+03  4.167e+02   2.669  0.00821 ** 
Population                  -1.340e-10  3.324e-10  -0.403  0.68724    
Areasqmi                    -2.992e-09  2.166e-08  -0.138  0.89025    
PopDensitypersqmi           -3.600e-05  2.146e-05  -1.677  0.09496 .  
Coastlinecoastarearatio      2.655e-04  5.213e-04   0.509  0.61116    
Netmigration                 2.395e-02  7.525e-03   3.183  0.00168 ** 
Infantmortalityper1000births -8.762e-03  2.680e-03  -3.270  0.00126 ** 
Literacy                    -1.978e-04  2.831e-03  -0.070  0.94436    
Phonesper1000                1.994e-03  2.500e-04   7.974 9.94e-14 ***
Arable                      -1.100e+01  4.166e+00  -2.641  0.00890 ** 
Crops                       -1.100e+01  4.166e+00  -2.642  0.00887 ** 
Other                       -1.100e+01  4.166e+00  -2.640  0.00892 ** 
Climate                     -9.713e-02  6.068e-02  -1.601  0.11094    
Birthrate                   -3.325e-02  7.206e-03  -4.614 6.87e-06 ***
Deathrate                    2.404e-02  1.026e-02   2.344  0.02004 *  
Agriculture                 -4.376e+00  2.624e+00  -1.668  0.09688 .  
Industry                    -2.614e+00  2.601e+00  -1.005  0.31594    
Service                     -3.201e+00  2.606e+00  -1.228  0.22070    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the fitted model, we could conclude that

- Among all the factors, the phone accessible percentage is of critical importance by showing significance with a positive impact. Specifically, one unit increase of technology factor corresponds to $e^{1.993599x\ 10^{-3}} - 1$ =0.001996 unit of GDP increment.

- The birth rate has a significant negative impact on GDP development while the death rate has a positive impact on GDP development. Specifically, with a similar calculation, one unit increase of the birth rate will decrease the GDP by -0.032 units.

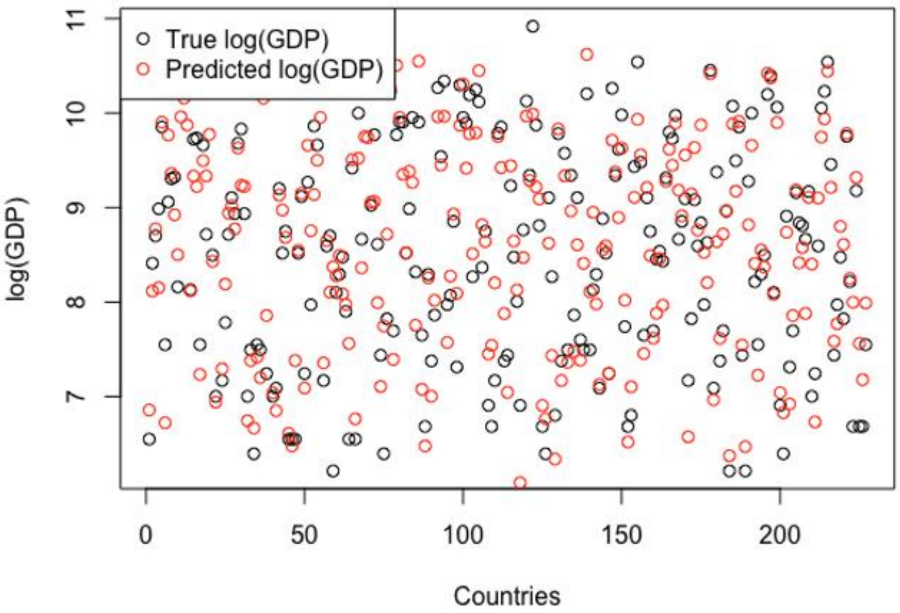
- Factors related to agriculture such as crops production; percentage of arable land and agriculture industry percentage have a significant negative impact on GDP. The results are not surprising because agriculture production represents alabor-intensive industry. A high percentage of the service usually corresponds to the developed economy.

- Other factors such as infant mortality rate and net immigration tend to have significant effects. This is consistent with my background research that the health condition within the country is positively correlated with economic development and a positive net immigration number typically corresponds to better social welfare.

5. Result Analysis

5.1 Predicted Versus True

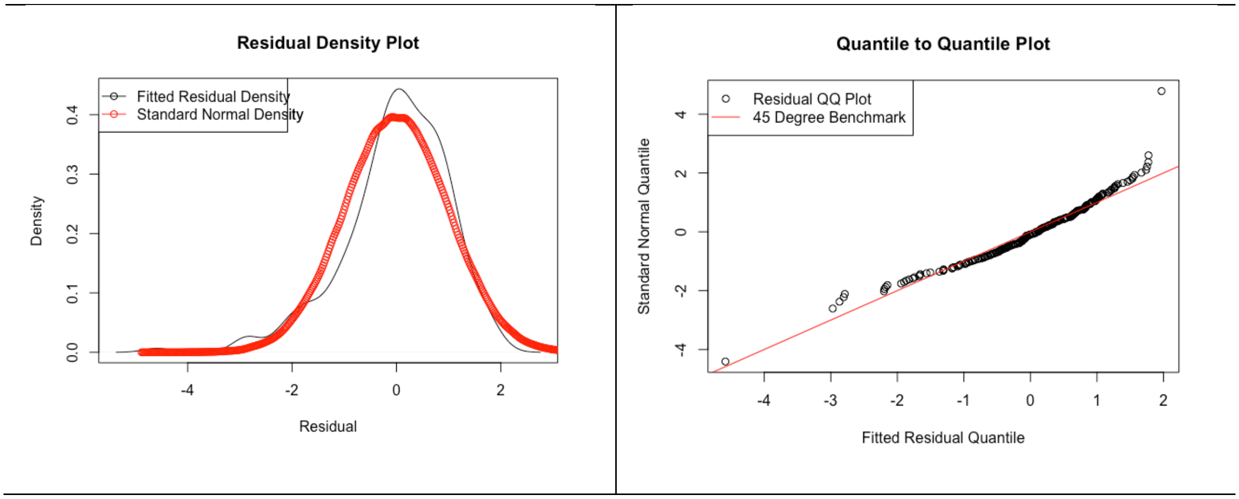
To check the fitness of the model, I plotted the real log(GDP) across 227 counties with the predicted log(GDP) across the 227 counties:



As it is indicated by the above plot, the predicted value is always nearby the true value. The predicted residual seems to be small but to robustly testing the model assumption. I supplemented the plot with two commonly used residual analyses.

5.2 Density plot and QQ plot

One important assumption in linear regression mode is that $\epsilon \sim N(0, \sigma^2)$ .I here first estimate the σ^2 with the residual $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-p}$ where p is the number of parameters included in the linear regression model. Then with the estimated $\hat{\sigma}^2$, I simulate the normal distribution with $N(0, \hat{\sigma}^2)$ and overlay it with the density plot of the residual $Y_i - \hat{Y}_i$ to compare how similar they are:



As it is indicated by the density overlapping plot, the model residual closely mimics the standard normal density, which further confirms the validity of the model assumption. However, to further examine the normality assumption, the quantile-to-quantile plot is also provided on the right-hand side. With the quantile line being in alignment with the 45-degree benchmark, the normality assumption is clearly satisfied.

5. Conclusion

In this project, I examine the statistical relationship between national attributes and national GDP using a linear regression model. The study concludes that Among all the factors influencing economic development,1. The technology factor and the health factor have the largest positive impact on economic development. 2. The birth rate and agriculture factors have a significant negative impact on economic development. The studies aid in our understanding of how the economy functions and further offer helpful recommendations for policymakers to create workable development plans that take the nation's characteristics into account.

Reference

- Linear Regression Using R: An Introduction to Data Modeling, David J. Lilja, 2016
- Country of the world. https://gsociology.icaap.org/dataupload.html

- All of Statistics: A Concise Course in Statistical Inference, Larry Wasserman, Springer, 2010
- R for Data Science, Hadley Wickham, Garret Grolemund, 2017https://r4ds.had.co.nz/