HW3-Yidi Kang-0330\_0530Thu

Yidi Kang-03:30-05:30

March 28, 2019

# Abstract

This study mainly explore possible explanatory factors that are related with literacy rate across countries. We will use multi-regression models to find the hidden relationships. According to missing data and model hypothesis, we finally include 49 countries for our study. Our result shows that 3 factors are strongly related with literacy rate for a country, they are “School complete rate”, “GDP per capita” and " Fertility rate“.

# Introduction

Nowadays, education has become a popularly debated topic for most countries, and literacy rate is one of the basic part of education achievement. A country/ area’s literacy rate is calculated by total number of literate persons divided by total number of population for a given age range. In our study, the literacy rate is for many countries across the world and the age range is above 15. We are doing this study about literacy rate is because literacy rate is a very important sign for a country’s primary education level, people’s quality level and even people’s life-quality level. In our study, we are trying to use data from World Bank to explore important variables that can have impact on literacy rate. By doing so, we hope we can give some insights to government about which part we should work on in current condition, and find the best and most practical solution for a country.

# Data

## Data Source:

All data in our research are from World Bank :<https://data.worldbank.org/indicator>

## Variables:

To determine which variables we want to test in our study, I decided to analyzing low literacy rate from 3 aspects.

First is the literacy cost and current education situation for people to become literacy. This part I include 3 variables.

1.Government expenditure per student, primary (% of GDP per capita) From government side, one of the non-literate reason for a people is her/his family cannot afford the education cost (tuition fee or time spent), so the government expenditure per student may give us some insights on whether the education-spent is sufficient for reducing literacy rate.

2.Primary completion rate, total (% of relevant age group) The literacy rate is a sign for primary education of a country/ area, so direct variable such as primary completion rate may be the most significant reason for literacy rate.

3.School enrollment, primary (% gross) The literacy rate is a sign for primary education of a country/ area, but I also wanna compare the difference between primary completion and primary enrollment, to see if there are obvious difference.

Second is whether the living environment (like economic or cultural condition) disable or discourage people in a country to literate. There are 5 variables in this part.

4.Rural population (% of total population) Comparing to Rural area with Urban, one of the most significant difference is education for students, the rural area may lack of teachers or other resources to support all children’s education. So more percent of rural population may lead to lower literacy rate.

5.Population, total If a country’s population is way beyond it’s capacity, then how to survive will become the primary concern for both people and government, so the education part may be overlooked.

6.GDP per capita (current US$) GDP is an important sign for a country’s economic development, and we should more focusing on GDP per capita, because what we really care about is how is everyone’s economic situation in a country, like whether the GDP per capita is sufficient for one to finish primary education.

7.Fertility rate, total (births per woman) In a way, the fertility rate is similar to population of a country, but what is different is that it focusing more on whether the population growth is beyond control or capacity, which will affect a country’s education ability a lot.

8.Life expectancy at birth, total (years) Comparing 50 years’ life expectancy to 80 years expectancy, there is more chance that it stands for the whole society’s progress, people will more likely not spend their whole life struggling with hunger and disease, but have more time to receive education.

Third part includes self-willing and self-oriented, which includes 2 variables

9.Labor force, total The labor force rate (or we can use unemployment rate) is an indicator whether of education level and labor demand. For example, if a country’s labor rate is really low, surviving will become the first concern and people will lost motivation to learn or let children learn.

10.Income share held by lowest 20% Low income for a family usually leads to not enough money for children’s education or need children to work early. Focusing on the lowest 20% part may give us better insights of whether the lowest income contribute to non-literacy most.

## Summary of our explainatory variables

First, we check NA values of our data.

Second, we check summary information for all variables.

## Rural p School r Fertility r Pop t GDP p expend p  
## Min. 0.0 37.77333 1.205000 10908 273.5383 5.32241  
## 1st Qu. 194524.2 87.63051 1.733250 762391 2043.6952 12.23067  
## Median 1929259.0 93.16648 2.249500 6109053 6492.0532 15.60742  
## Mean 15768733.9 89.67680 2.783082 33533263 17418.1192 16.49177  
## 3rd Qu. 9725792.2 96.31179 3.643250 22312766 20015.5516 19.83039  
## Max. 874855900.0 99.90800 7.338000 1364270000 185152.5272 43.49877  
## NA's 3.0 85.00000 17.000000 1 14.0000 127.00000  
## labor t complete r eLife 20%low income Literacy rate  
## Min. 38638 48.10355 50.62100 2.400000 32.00386  
## 1st Qu. 1147617 84.66885 66.21500 5.500000 78.88474  
## Median 3967675 96.64122 73.48300 7.100000 92.47476  
## Mean 17840524 91.42138 71.75957 6.882667 85.06006  
## 3rd Qu. 11733520 100.58729 77.36546 8.400000 97.94686  
## Max. 786573327 122.56149 83.98049 10.500000 99.98055  
## NA's 31 88.00000 18.00000 142.000000 166.00000

## Data Cleaning

For literacy rate, since this is our must-have response variable, we are going to drop NA in this variable.

For GDP per cap, population total, labor force, we are going to remove countries which has NA in these variables. Because I think these are very important basic that we can look at a specific country’s literacy rate.

For school enrollment rate, primary completion rate, gorvernment expend, lowest 20% income we are going to use mean to make up all the NA values

Also, since Rural rate is a better index than Rural total, we also need to change this variable.

## Study year I choose

I choose 2014’s data as my resource, first, I wanna focus more on current literacy rate, second, in this year, we can have more non-NA literacy rate examples in study.

## Study countries I choose

“Indonesia”“Brazil”“Pakistan”“Bangladesh”“Mexico”“Iran,IslamicRep.”“Turkey”“SouthAfrica”“Colombia”“Spain”“Kenya”“Argentina”“Peru”“Uzbekistan”“Angola”“Coted’Ivoire”“BurkinaFaso”“Malawi”“Guatemala”“Ecuador”“Zimbabwe”“Cambodia”“Guinea”“Rwanda”“Tunisia”“DominicanRepublic”“Burundi”“Hungary”“Azerbaijan”“Honduras”“Tajikistan”“Paraguay”“Singapore”“Turkmenistan”“WestBankandGaza”“Oman”“Georgia”“Uruguay”“Jamaica”“Botswana”“Lesotho”“Slovenia”“GuineaBissau”“Mauritius”“EquatorialGuinea”“Guyana”“Maldives”“Barbados”“Vanuatu”

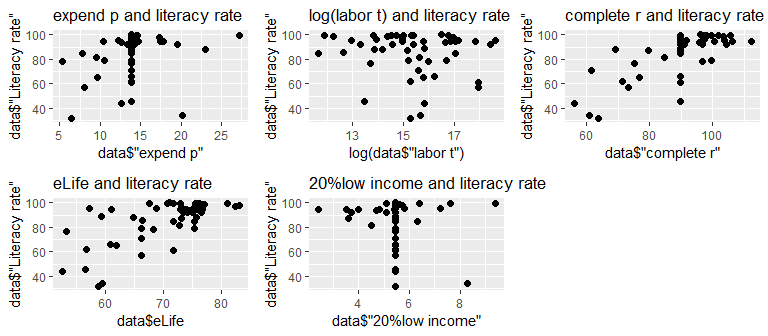
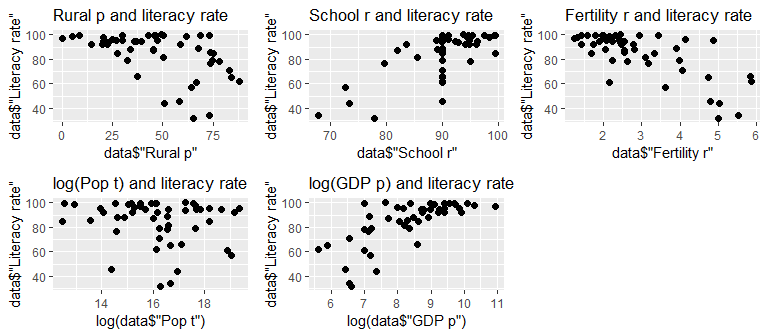
# Methods

## Model building summary

transformation:

According to the scatter plot of all explainatory variables and response variable, the population, GDP and labor are not linearly related with literacy rate.So I decide to use log(Pop t), log(GDP p), and log(labor t) to transform data. We can see after transformation, the relationship between these 3 variables and y are more linearly related.

Relevent Plots:



Collinearity check: correlation matrix

## Rural p School r Fertility r Pop t GDP p expend p labor t  
## Rural p 1.00 -0.34 0.49 -0.13 -0.60 -0.23 -0.16  
## School r -0.34 1.00 -0.54 -0.03 0.31 0.13 0.02  
## Fertility r 0.49 -0.54 1.00 -0.16 -0.47 -0.23 -0.18  
## Pop t -0.13 -0.03 -0.16 1.00 -0.09 -0.06 0.99  
## GDP p -0.60 0.31 -0.47 -0.09 1.00 0.30 -0.06  
## expend p -0.23 0.13 -0.23 -0.06 0.30 1.00 -0.01  
## labor t -0.16 0.02 -0.18 0.99 -0.06 -0.01 1.00  
## complete r -0.47 0.73 -0.59 0.10 0.29 0.11 0.14  
## eLife -0.55 0.57 -0.82 0.08 0.55 0.24 0.10  
## 20%low income 0.16 -0.02 0.04 -0.03 0.13 0.23 -0.04  
## complete r eLife 20%low income  
## Rural p -0.47 -0.55 0.16  
## School r 0.73 0.57 -0.02  
## Fertility r -0.59 -0.82 0.04  
## Pop t 0.10 0.08 -0.03  
## GDP p 0.29 0.55 0.13  
## expend p 0.11 0.24 0.23  
## labor t 0.14 0.10 -0.04  
## complete r 1.00 0.64 -0.01  
## eLife 0.64 1.00 0.06  
## 20%low income -0.01 0.06 1.00

From the correlation matrix, we can see there are 3 pairs of variables have high correlation

school r with complete r of 0.734; eLife and Fertility r of -0.819; Pop t and labor r of 0.986

by comparing 3 pairs RMSE, I decide to remove 3 variables (school r, eLife, Pop t) with slightly higher RMSE.

Variable selection:

So I started with a 5 variavble model, check the significance for each variable, and reduce not significant variables.

Finally we got a model of 3 variables with all significance smaller than 0.05

## Assumption check

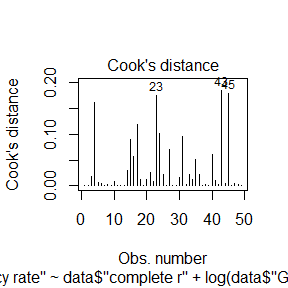
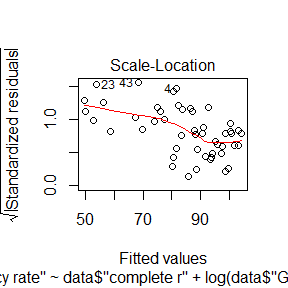
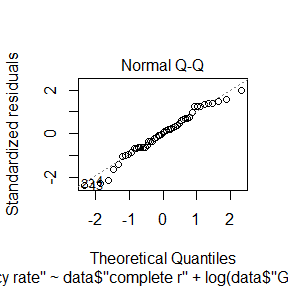
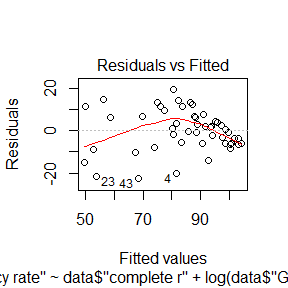
1. x variables are fixed and no error

All the x variables are from gorvernment investigation and with the same scale, which ensured it’s accuracy

1. E(e) = 0

Our OLS ensured this

Below is a plot to check constant variance and Normality Assumption.



3.constant Variance

From 1st and 3rd of p.a, we can see that after standerlization, the reasidual still follows a tend of decreasing

so we kind of voilated the constant variance assumption

1. Normality of standarlized residuals

According to 2nd normal Q-Q plot, we can see that the standardized residuals are almost normally distributed.

1. Independence

Our data are collected from different countries, so we can say that they are independent to each other

1. correlation of x variables

We already moved variables that have high correlation, so we meet this assumption.

## Model Evaluation

1.Estimate sigma:

sigma = RMSE = 10.04; However, since we didn’t meet constant variance assumption, so we cannot conclude that 95% y values are in Y +/- 2 sigma

2.R^2 and Ra^2

The R^2 of our model is 70.58%, means 70.58% of the variance can be explained by our model; Ra^2 is 68.61%.

## Hypothese test (alpha=0.05)

1. overall F-test

H0: all beta equals to 0; H1: at least 1 beta is not 0. According to summary of our model, we can see that the Fc is 35.98 with 45 degree of freedom, and the p-value is 5.126\*(10^(-12)), which is much smaller than alpha=0.05

1. t-test

H0: beta i equals to 0; H1: beta i is not 0.

for complete r: test statistic is 3.817 with 48 degree of freedom, and p-value is 0.00041, for log(GDP): test statistic is 2.7 with 48 degree of freedom, and p-value is 0.00964, for fertility r: test statistic is -2.35 with 48 degree of freedom, and p-value is 0.02, all 3 p-values are smaller than 0.05(alpha)

so in 3 hypothesis test, we all reject H0, every slope is statistically significant basing on other variables are already in model.

# Discussion Final

## Final Model:

## [1] "Final model is y = 0.56 \*complete r+ 4.53 \*log(GDP p) -4.03 \* fertility r+ 8.43"

## Interpretation:

Hold other 2 variables stay, 1 unit(percent) increase in complete rate will increase literacy rate 4.53 units(percent)

Hold other 2 variables stay, if GDP per capita increase by 1% ,literacy rate will increase 0.0453 units(percent)

Hold other 2 variables stay, 1 unit(percent) increase in fertility rate will decrease literacy rate 4.03 percent.

When complete rate, fertilyty rate equal to 0, and GDP per = 1, literacy rate is 8.43%, while this situation is not practical.

## Usefulness

By building this model, we can get some insights on what is important for a countries literacy rate, and we may use this insights to actually publish relevant rules or find practical solution to improve a countries literacy rate.

## Improvement

Since our variables did not fully meet the assumptions (violated the constant variance assumption), next time, we can find more ways to transform our data so that we fit all assumptions.

Also, our model’s R^2 now is 71%, we may try more variables to improve the part that can be explained by our model. (Also we need to pay attention to not overfitting)

# Reference

World Bank : <http://data.worldbank.org/indicator>

Literacy Rate : <https://ourworldindata.org/literacy>