HW2\_Template

Lina Cao , Thursdays 3:30PM

SDGB 7844; Prof. Nagaraja; Fall 2017

# Executive Summary

This report discusses facters that can influence Gini Index and how. Ten variables are picked at first, in order to build a linear regression model that can explain Gini Index. I collect 264 observed countries’ data, and use 70 of them(with complete data set) to fit a muiti-linear regression model. I also try to improve the model and make appropriate choice by checking Cook’s distance and doing partial F-test. After deciding variables in final model, I update the data set to countries with complete data in all four variables, as well as the final regression. The final regression result shows that Gini Index can be influenced by GDP(dollars), growth rate of GDP(%), inflation rate(%) and health expenditures per capita(dollars).

# Introduction

Gini Index is a measure of statistical dispersion that intended to represent the income or wealth distribution of a nation’s residents, and is the most commonly used measure of inequality.(source: wikipedia) It is widely used for a nation’s politicians or economists for making economic policies and in turn narrow the gap between the rich and poor. Therefore, understanding what factors may influence Gini coefficient and build an efficient model to monitor and predict Gini Index become crucial topics.

In a mathematical way, Gini coefficient is usually defined based on the Lorenz curve, which plots the proportion of the total income of the population (y axis) that is cumulatively earned by the bottom x% of the population.

Like all the other indices, Gini Index also has defects. For example, because the Gini coefficient measures relative, not absolute, wealth, same Gini coefficient can have different income distributions in empirical world. This means the number of Gini index may sometimes treat politicians about one nation’s income inequality status. Therefore, in this report, multi-variable linear regression is introduced to model Gini Index in another aspect.

# Data

## Variables

All data sets in this paper are downloaded from the website of world bank (<http://data.worldbank.org>). In addition to Gini Index, 10 more variables are introduced to build a decent regression model. As is known to all, Gini index is an integrated economic parameter that can be used in one nation’s macroeconomic regulation and control process. I consider the variable choosing categories from three steps for a nation’s income distribution process: total wealth of a country, the distribution of the wealth and the redistribution process. Based on the three aspects above, many variables are considered at the very begining, such as “govenment expenditure”, “net official development assistance”, “infrastruction spending”, etc. But taking the amount of total effective data into consideration, i.e, to make the data set combines more data in the chosen year, I chose the following variables.(The specific reason for each data is attached in appendix.)

1. Gini Index (response variable)
2. GDP($)
3. GDP annual growth rate(%)

(3)Inflation rate (consumer prices, annual %)

1. Population
2. Unemployment
3. Tax revenue (% of GDP)
4. Refugee population by country or territory of origin (% of urban population)
5. Urban population (% of total)
6. Labor force(people)
7. Mobile cellular subscriptions (per 100 people)
8. Health expenditure per capita

## Data Cleaning

The basic rule to clean and find proper year for further study is quite simple. I first wrote a function to calculate, for each file (variable), which column contains most data and then chose the column number that appear the most times. (see reference part)

The result is a bit discussion-needed though. There are two columns, 35th and 55th,

which indicates for year 1990 and 2010 respectively, have largest data amount in some files. Since our goal is to find regression model fitted with Gini Index, I focused on which column in the data set of Gini index can give most data(the 55th column).

Therefore, I finally used 55th column, i.e year 2010, in each variables’ data set to keep year consistency, and make a reasonable regression for interpretation.

## [1] 70

After merging data into one data set and compute data set without missing data. There are 70 countries, including Australia, Canada, Switzerland and other countries that are included in the final data set.

## Statistical Description

## gdp r.gdp inf pop uem tax  
## [1,] 0.0248872 0.4193799 0.08034412 0.06435897 0.1396177 -0.1577088  
## refuge urban lbf mobile hlt  
## [1,] 0.2110205 -0.1744891 0.08342701 -0.2728585 -0.3604632

First check the statistical parameters(mean, median, max, min, sd) of variables and correlation between Gini Index and each variable. Among 11 variables, “tax”,“urban”,“mobile”,“hlt” have negative correlation with Gini Index. When considering the absolute value for how much the linear correlation each variable has, “r.gdp”(0.4193) has the largest linear correlation while “gdp”(0.0249) has least linear correlation with Gini Index.

For each variable, histograms are introduced to see the distribution of the data. Most variables are right skewed, while “urban” and “mobile” are both left skewed; and “r.gdp” and “tax” are almost symmetric. Variables “gdp”, “pop”, “refuge” and “lbf” have outliers that are extremely large. I suppose these demographic-related data may influenced by total population of a country, therefore once population has extreme data, the other will correspondingly have problem.

# Methods

## Model Building Summary

I used backward elimination approach to build my regression model. After the first round of variables selection, four variables, namely “gdp”, “r.gdp”, “inf” and “hlt” are left in my regression model. Then the regression assumptions are checked with four-variable model. From Cook’s distance chart, the distance of several countries sticks out, however the values are all less than 1. After trying some attempts for improving model, the first-fit model is proved to be a better one. At last, new data set, with only three variables’ complete data, is introduced to refit the model and derive the final estimates for the linear regression model.

## Variables Determination —- first-fit model

I firstly included all candidate variables and get the first regression result. The Cook’s distances are all smaller than one, not influential point has to be removed.

## [1] FALSE

The core of backward regression is to delete one variable at a time, the criterion of which is choosing the one with least deterioration when the variable is not included in the model. Take the result in the first fit as an example, “tax” has the largest p-value, implying a least significance of its estimate. Therefore I attempted to drop “tax” in my model.

Before that, I checked the situations in which I dropped other insignificant variables in the first model, in order to confirm the deletion of “tax” did cause least loss of r-squared.

## [1] 0.3173495 0.3058077

## [1] 0.3171199 0.3159198 0.3121234 0.3020413 0.2977341 0.2869300

From the result above, the R-squared increases after dropping “tax”, which is a good sign for deletion, since R-squared tend to decrease with fewer variables. Additionally, the same parameter in trial models range from (0.2869, 0.3171), none of which are as high as 0.3173 in model 2. Hence, it is a safe choice to drop “tax” in the second round.

With the same principle, the final variables are chosen after the work of round three to eight. The model can be temporarily written as :

## [1] "gini(%) = 38.88(%)+1.006e-12(%/$)\*gdp($)+1(%/%)\*r.gdp(%)-0.8145(%/%)\*inf(%)-0.001909(%/$)\*hlt($)"

The first fit model- to decide which variables are included in my model- is as shown above. All variables have a significant estimate of beta, the R-squared of the model is 0.2901. Though R-squared is not large enough, all the estimates are significant. Further improvements are tried to increase R-squared.

## Regression Assumption Check

Assumptions need to be checked to make a valid linear regression. The first two, “fixed X” and “linearity” are acknowledged and checked in appendix.

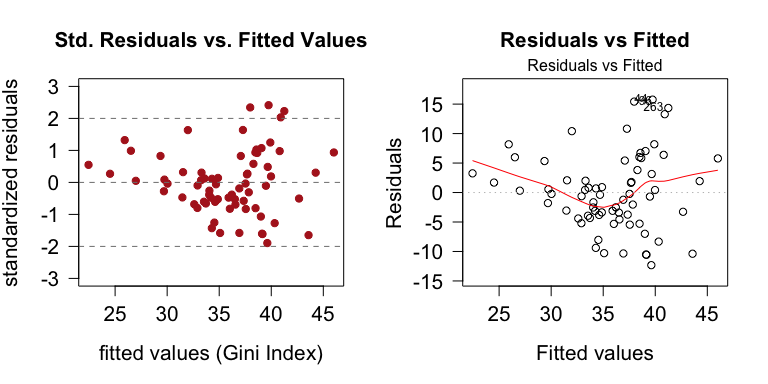
The following assumptions need further computation to determine whether it is true in this regression:

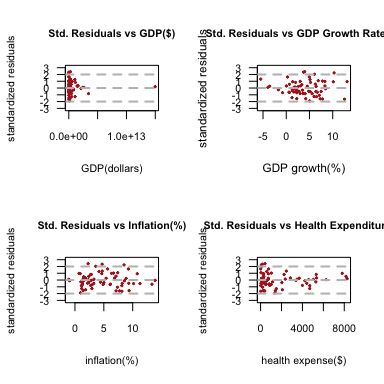
1. The conditional mean of error should be zero;

## [1] 1.618745e-16

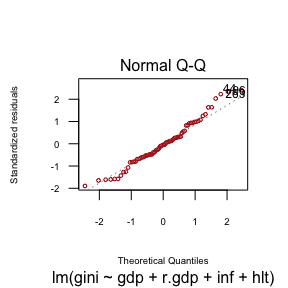
From the calculated mean of regression residuals, the answer is 1.619e-16, which is close enough to zero to regard the value as zero. Therefore this assumption meets.

1. The variance of the error term should be constant regardless of x values;

The red line in the right side plot remains around +/-5 and becomes flat when fitted value varies. This means the volatility fluctuates but tends to remain at a certain level. 



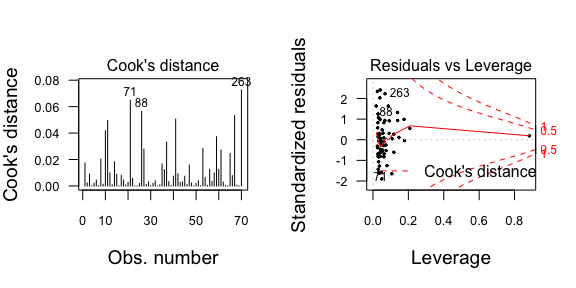
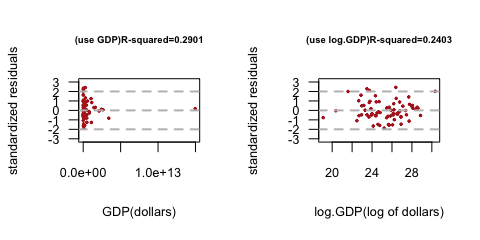
From the residual plot, “r.gdp” and “inf” has no pattern of their residual plots, with their residuals ranging from indicating a relatively good fit in this model. While the residuals of “hlt” has larger deviation when the x-value becomes smaller. As for “gdp”, there is an obvious outlier. Cook’s distance is used later to see the result after omiting the outlier.

1. error(residuals) should be normally distributed;  
   

This assumptions can be checked by normal Q-Q plot of residuals as above, the residuals are close to the normal quantile line, especially at the middle part. However there do have some outliers on both sides. A good sign for the sided data is that they have tendency to return to the normal Q-Q line.

1. additionally, there is no autocorrelation. However, this part together with whether there is multi-collineary problem need to be discussed later.

## Model Improvement

Inspiring from the result above, use Cook’s distance to see influential data. The first plot shows all data’s Cook’s distance are below 1, in addition to the flat solid line in second talbe, there is no need to remove influential data and refit.  According to a narrower standardized deviation when GDP increases, as well as outliers, I tried to use log of GDP to derive a better fit model. 

After transforming the data, it is controversial about whether remain “lg.gdp” in the model or not. The good sign is the x-values are closer, with a smaller range, also the dots in standardized residuals plot converts to randomly distributed below and above zero. However, the R-squared of new model drops a lot to 0.2403. Therefore, I used partial F-test to see if log(gdp) should be removed if “gdp” is in log format.

## Analysis of Variance Table  
##   
## Model 1: gini ~ r.gdp + inf + hlt  
## Model 2: gini ~ lg.gdp + r.gdp + inf + hlt  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 66 3093.2   
## 2 65 3091.7 1 1.4515 0.0305 0.8619

The F-value is 0.8619, therefore, we should not reject the null hypothees– the coefficient of “lg.gdp” is zero–which means “lg.gdp” should be deleted in the final model if we take the logarithm for “gdp”. Considering a higher R-squared, I finally use the first-fit model as the final model.

## Model Evaluation

Since the variables have been cut into four, I reorganized the data just in case of misdeleting country that is completed in the last four variables yet has missing data in the dropped ones.

##   
## Call:  
## lm(formula = gini ~ gdp + r.gdp + inf + hlt, data = ccc)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.3054 -4.1667 0.0557 4.6417 15.0187   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.888e+01 1.990e+00 19.543 < 2e-16 \*\*\*  
## gdp 1.006e-12 5.023e-13 2.003 0.048954 \*   
## r.gdp 1.000e+00 2.531e-01 3.952 0.000179 \*\*\*  
## inf -8.145e-01 2.816e-01 -2.893 0.005048 \*\*   
## hlt -1.909e-03 4.572e-04 -4.174 8.26e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.899 on 72 degrees of freedom  
## Multiple R-squared: 0.3545, Adjusted R-squared: 0.3187   
## F-statistic: 9.887 on 4 and 72 DF, p-value: 1.967e-06

With the final regression result, the following evaluation steps are considered.

### (1) adjusted R-squared

Adjusted R-squared is 0.3187, which means 31.87% of variability can be explained by the model.

### (2) Overall F-test

The F-statistic is 9.887, with its p-value really closes to zero. It is a relatively small number to reject the null hypothesis– all variables’ beta is equal to zero.

### (3) t-test

Focusing on t-statistic for each variable, and their corresponding p-value, I found every beta hat is significant. To be more specific, “r.gdp” and “hlt” are significant at 0.001 level, “inf” and “gdp” are at 0.01 and 0.05 level respectively. Therefore I should reject the null hypothesis of the slope is equal to zero.

### (4) RMSE

## [1] 6.899001 8.358205

RMSE is about 6.899(%), while the standard deviation of Gini index(using the updated data set) is 8.358(%). They are close, though more steps can be done to make a better fit, but we have reached a nice result for the regression model.

### (5) Cook’s distance

As the first plot shown above in “Model Improvement”, all the Cook’s distance are less than 0.3, no influential data point need to be moved.

# Discussion

## Results

The final result of the estimated model is:

## [1] "gini(%) = 38.88(%)+1.006e-12(%/$)\*gdp($)+1(%/%)\*r.gdp(%)-0.8145(%/%)\*inf(%)-0.001909(%/$)\*hlt($)"

The intercept means if all the other variables are zero, Gini index will be 38.88%, but this point will never be met. It is unreasonalbe that a country has zero GDP, with zero growth rate of GDP, no inflation and no health expenditure per capita.

After converting GDP to trillion dollars, the slope of it means, keep other variables stable, when GDP increase 1 trillion dollars, Gini index will increase 1.006%.

In the same way, remain other variables the same, 1% increase in GDP growth rate will increase 1% of Gini index.

Remain other variables the same, if inflation rate increase 1%, Gini index will deteriorate 0.8145%.

Remain other variables the same, health expenditure per capita increase 1 dollar, Gini index will decrease 0.0019%.

## Usefullness of model

The adjusted R-square is only 31.87%, so the model can only explain a certain cases. It just give us a simple concept of how these selected variables influence Gini index, using only 77 countries data in 2010. These all show the limitation of this model. Surely, we cannot use this model to predict the future Gini index, but it can somewhat show how Gini index will move when variables fluctuate.

## Further improvement

I noticed that the linear relationship among respond variables and explanatory variables, as well as different explanatory variables are not obvious. Therefore, next step, other variables, which have stronger linear relationship with Gini (but not used to derive Gini) can be introduced. Also, collinearity and autocorrelation should be checked in order to reduce their influence to the model. Further discussion is needed for a better model.

# Appendix

## Reference

1. World Bank: http://data.worldbank.org
2. Naguib, Costanza. The relationship between inequality and GDP growth: An empirical approach. No. 631. LIS Working Paper Series, 2015.
3. Haas, Christina. “Income Inequality and Support for Development Aid.” (2013).
4. Coburn, David. “Beyond the income inequality hypothesis: class, neo-liberalism, and health inequalities.” Social science & medicine 58.1 (2004): 41-56.
5. Coburn, David. “Income inequality, social cohesion and the health status of populations: the role of neo-liberalism.” Social science & medicine 51.1 (2000): 135-146.

### (0) Gini Index (response variable)

### (1) GDP($)

As an important economic indicator, GDP is the variable that can be considered as how much a nation can produce in one year and to be an indicator of total wealth. Costanza Naguib(2015) concluded in his paper—The relationship between inequality and GDP growth—that higher inequality levels are associated to higher levels of per capita GDP and per capita GDP growth.

### (2) GDP annual growth rate(%)

With the inspiration of Costanza Naguib’s finding(2015), GDP’s growth rate is a possible variable to explain Gini Index. Also, dividing data by former year’s data can help build a more stationary data in time series cases.

### (3)Inflation rate (consumer prices, annual %)

Inflation is a problem occuring also under the aspect of total wealth. It reflects a general increase in prices and fall in the purchasing value of money. Since this is a model for inequality, I chose CPI related inflation computing data in my model.

### (4) Population

Population is also an important variable in macroeconomic problems.A larger population produces larger amount of total wealth, while it also gives politician a higher level difficulty to give proper resource allocation.

### (5) Unemployment

Income inequality may also as a result of people not have own work to earn enough money. Unemployment rate, to some extend give information about it.

### (6) Tax revenue (% of GDP)

For a government, income inequality is another name of social benefit allocation and redistribution. For a government, their “revenue” comes from tax, which is a preparation of income reallocation. Therefore I include this variable in the model. By dividing by corresponding GDP of each country, this data series gets rid of the scale influence to some extend.

### (7) Refugee population by country or territory of origin (% of urban population)

Refugees are “product” of the failure of equal income distribution. A larger size of refugee in one country always indicates a heavier burden for a government to allocate more alms for refugees, which will lead to a higher Gini Index.

### (8) Urban population (% of total)

Urban area always tends to feed wealthy people, also government may implement better allocation policies with the help of a more advanced ideology, in which case the Gini index will be smaller than that in suburban area.

### (9) Labor force(people)

Just an inverse version of unemployment, labor force is a sign for people that can work, which may have opposite direction of the way Gini index goes.

### (10) Mobile cellular subscriptions (per 100 people)

This is a sign for the development of mobile devices. In this high-tech era, cellular subscription is a sign for advanced society. And an advanced society tends to have smaller gap between the wealthy and the poor.

### (11) Health expenditure per capita

Health care is a symbol of a higher level of consumption. I regard it as also as a way to separate people who have high and low amount of disposable income, which implies the status of people’s income level in one certain area.

## Data Cleaning

The basic rule to clean and find proper year for further study is quite simple. I first wrote a function to calculate, for each file (variable), which column contains most data and then chose the column number that appear the most times. (see reference part)

## temp  
## 35 36 43 48 51 55 57 58   
## 3 1 1 1 1 3 1 1

The result is a bit discussion-needed though. There are two columns, 35 and 55,

## [1] "X1990" "X2010"

which indicates for year 1990 and 2010 respectively, have largest data amount in some files. Since our goal is to find regression model fitted with Gini Index, I focused on which column in the data set of Gini index can give most data.

## [1] "55" "X2010"

Therefore, I finally used 55th column, i.e year 2010, in each variables’ data set to keep year consistency, and make a reasonable regression for interpretation.

## [1] 70

## [1] "Argentina" "Armenia"   
## [3] "Australia" "Austria"   
## [5] "Belgium" "Bangladesh"   
## [7] "Bulgaria" "Belarus"   
## [9] "Canada" "Switzerland"   
## [11] "Colombia" "Costa Rica"   
## [13] "Cyprus" "Czech Republic"   
## [15] "Germany" "Denmark"   
## [17] "Dominican Republic" "Egypt, Arab Rep."   
## [19] "Spain" "Estonia"   
## [21] "Ethiopia" "Finland"   
## [23] "France" "United Kingdom"   
## [25] "Georgia" "Greece"   
## [27] "Honduras" "Croatia"   
## [29] "Hungary" "Ireland"   
## [31] "Iceland" "Israel"   
## [33] "Italy" "Jordan"   
## [35] "Kazakhstan" "Korea, Rep."   
## [37] "Lesotho" "Lithuania"   
## [39] "Latvia" "Moldova"   
## [41] "Madagascar" "Mexico"   
## [43] "Macedonia, FYR" "Mongolia"   
## [45] "Malawi" "Netherlands"   
## [47] "Norway" "Nepal"   
## [49] "Pakistan" "Peru"   
## [51] "Poland" "Portugal"   
## [53] "Paraguay" "Romania"   
## [55] "Russian Federation" "El Salvador"   
## [57] "Serbia" "Sao Tome and Principe"  
## [59] "Slovak Republic" "Slovenia"   
## [61] "Sweden" "Thailand"   
## [63] "Tunisia" "Turkey"   
## [65] "Ukraine" "Uruguay"   
## [67] "United States" "Vietnam"   
## [69] "Vanuatu" "Zambia"

There are 70 countries, including Australia, Canada, Switzerland and other countries that are included in the final data set.

## Regression Assumption Check

The first two assumptions

1. x values are fixed and are measured without error;

* Some of the variables, for example, population(“pop”), urban population(“urban”) and labor force percentage(“lbf”) are not fixed since the usage of census method or estimation of the sample. Therefore we cannot say they are measured without any error;

1. variables are linearly related;

* Check the scatter plot between every two variables as below. The result turns really same as that from residuals plots. It can be intuitively seen that variables hardly show linear relationship with each other 