Model the Gini Index across countries by using multiple regression

-homework 3 for explanatory model

Mengchuan Fu (Mike)
A14008047
Session: 3:40

Executive Summary

An increasing tendency of income inequality among the world has caused a big discussion about the factors affecting inequality. The present study investigates a number of factors that influence income inequality in countries among the world. This study uses multiple regression model to analyze many different indicators. As a result, three components are formed from the initial indicators. The impact of these components on income inequality are then analyzed and the regression model are formed.

Introduction

Gini index measures the extent to which the distribution of income among individuals or households within an economy deviates from a perfectly equal distribution. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual or household. The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. Thus a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality.

The purpose of the study is to find out the factors (other measurement index) that related to the income inequality (Gini Index) for countries among the world.

Data

According to the article – "Economic inequality" in Wikipedia. The main causes of the income inequality are: Labor market, Taxes, Education, Globalization, Economic development, Individual preferences, which correspond to the indicators on the world bank of "Wage and salaried workers, total", "Total tax rate", "Adjusted savings: education expenditure (% of GNI) ", "Trade", "Adjusted net national income", "Labor force, total".

As gini index is a kind of economic index, the countries that have great economic impact to the world should be included to the analysis. The top ten countries in the 2015 GDP rankings are: United States, China, Japan, Germany, United Kingdom, France, India, Italy, Brazil and Canada, so at least these ten countries should be included in the analysis.

As the year 2010 is the latest year that has relatively sufficient Gini data for "large countries" and other countries, we pick the Gini data in 2010, and for Japan, India and Brazil which missed the data in 2010, we pick the data in close year. From the metadata, the Gini data for one specific country did not changed a lot among years, so this method should not affect the result a lot.

Download the indicator data for countries in 2010 and merge the read the two table and again make up the indicator data for "large countries".

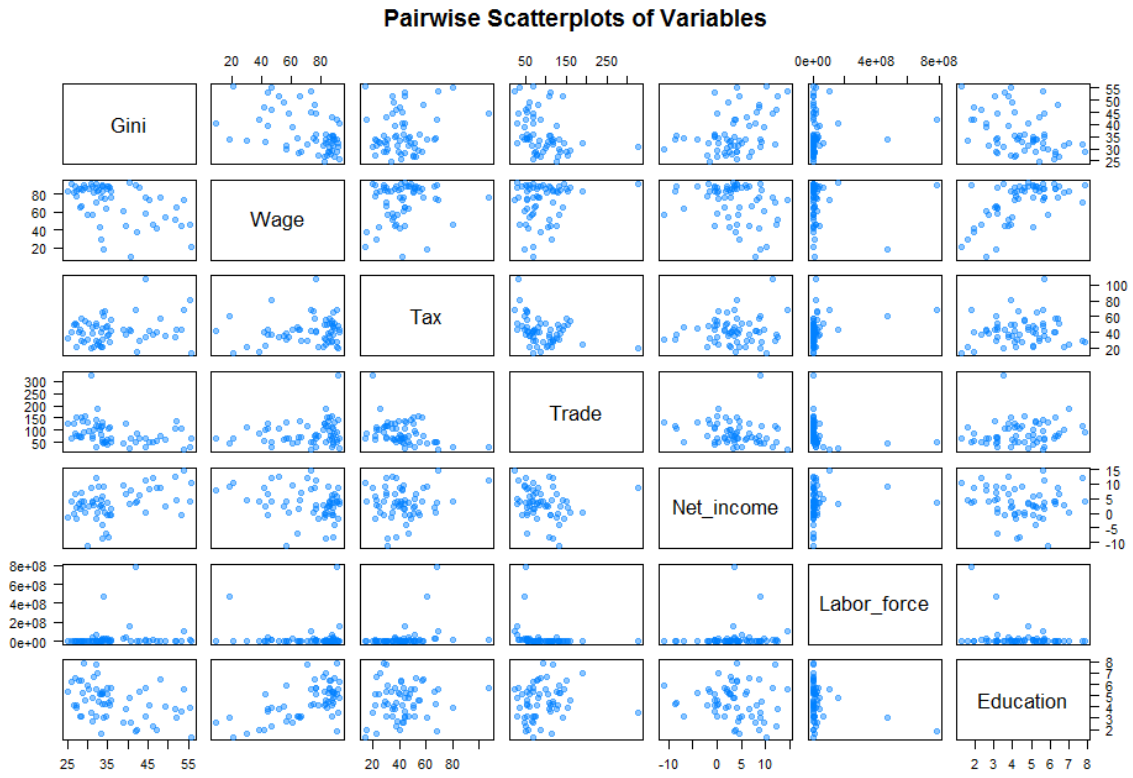
There are 60 countries are included in the data and the summery statistics are as follow:

Statistic	Gini	Wage	Tax	Trade	Net_income	Labor_force	Education
Minimum	24.94	9.30	14.30	22.51	-10.99	237978	1.30
Median	33.88	79.80	40.75	77.59	3.57	4705756	4.42
Maximum	55.62	93	107.40	326.14	12.48	781054640	7.88
Mean	36.47	71.25	42.17	90.22	3.46	34266458	4.46
Standard deviation	8.30	20.99	16.38	49.08	5.24	117280900	1.52

Methods

We first use scatterplot and compute the correlations to get an intuitive sense. The overview of the variables are as follows:

Scatterplot:



Correlation:

	Gini	Wage	Tax	Trade	Net_income	Labor_force	Education
Gini	1.000	-0.483	0.180	-0.336	0.393	0.096	-0.386
Wage	-0.483	1.000	0.142	0.192	-0.287	-0.038	0.599
Tax	0.180	0.142	1.000	-0.370	0.116	0.308	0.099
Trade	-0.336	0.192	-0.370	1.000	-0.186	-0.236	0.150
Net_income	0.393	-0.287	0.116	-0.186	1.000	0.121	-0.128
Labor_force	0.096	-0.038	0.308	-0.236	0.121	1.000	-0.254
Education	-0.386	0.599	0.099	0.150	-0.128	-0.254	1.000

From the scatterplot and the correlation table, we see that the variable “Labor_force” has little to do with Gini index. To double check, we calculate the correlation between Gini and the log form of “Labor_force”, which is still very low (0.194), so we get rid of the unrelated variable “Labor_force” and make the following regression with the rest five variables.

Fit the linear model for the five variables and the result are as follow:

```
Call:
lm(formula = Gini ~ Wage + Tax + Trade + Net_income + Educatior
    data = new_gini)

Residuals:
     Min       1Q   Median       3Q      Max
-14.9487  -4.2991   0.1217   4.1359  14.6922

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46.81940    4.46013  10.497 1.19e-14 ***
Wage        -0.12320    0.05593  -2.203   0.0319 *
Tax          0.07990    0.06037   1.324   0.1912
Trade       -0.02538    0.02026  -1.253   0.2158
Net_income   0.37580    0.17995   2.088   0.0415 *
Education   -0.88475    0.73147  -1.210   0.2317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.805 on 54 degrees of freedom
Multiple R-squared:  0.385,     Adjusted R-squared:  0.3281
F-statistic: 6.762 on 5 and 54 DF,  p-value: 5.844e-05
```

From the R-squared and adjusted r-squared, the fitted model has a relatively weak correlation with Gini. From the overall F-test, the model is significant. From the T-test of each variable, only “wage” and “net_income” is significant to the Gini. As the t-test for variables “Tax”, “Trade”, “Education” are not significant, we use VIF function to compute variance inflation factor.

```
   Variables      VIF
1       Gini 1.626096
2       Wage 1.913990
3        Tax 1.285395
4      Trade 1.296197
5 Net_income 1.223203
6  Education 1.613171
```

From the result of VIF function, we don’t see distinctive multicollinearity among variables.

As variable “Tax”, “Trade”, “Education” didn’t show linear relationship to Gini, we try some transformations to study. We use the log x and log y models to the three variables respectively. The p-value of T-test for each variable are as follows:

p-value of T-test	Tax	Trade	Education
Log x	0.43522	0.00171	0.00161
Log y	0.1786	0.00624	0.00185

For Tax, both log x and log y models are not significant. For Trade, both models are significant and the log x model has lower p-value. For education, both models are significant and the log x model has lower p-value. As a result, we get rid of the variable of “Tax” and transform the variables “Trade” and “Education” into log form in the linear model. After transforming, we fit the linear regression model again.

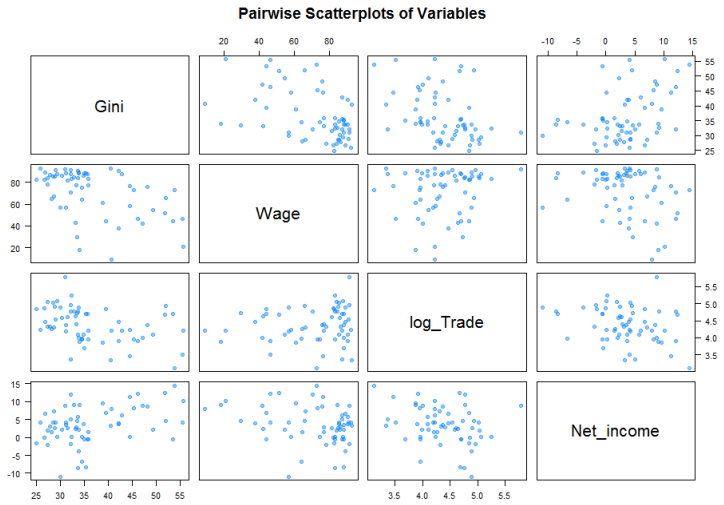
```
Call:
lm(formula = Gini ~ Wage + log_Trade + Net_income + log_Education,
    data = gini_2)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9471  -4.2368  -0.7228   3.8603  16.2257

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.39672    8.58927   7.847 1.55e-10 ***
Wage         -0.12560    0.05565  -2.257   0.0280 *
log_Trade    -4.50716    1.79754  -2.507   0.0151 *
Net_income    0.31732    0.18118   1.751   0.0855 .
log_Education -2.34946    2.91916  -0.805   0.4244
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.733 on 55 degrees of freedom
Multiple R-squared:  0.3869,    Adjusted R-squared:  0.3423
F-statistic: 8.677 on 4 and 55 DF,  p-value: 1.672e-05
```

```
          Variables      VIF
1              Gini 1.631020
2              Wage 1.940691
3         log_Trade 1.232816
4        Net_income 1.237580
5 log_Education 1.730858
```

As the log form of the variable “Education” is still not significant and there is still no multicollinearity among variables.

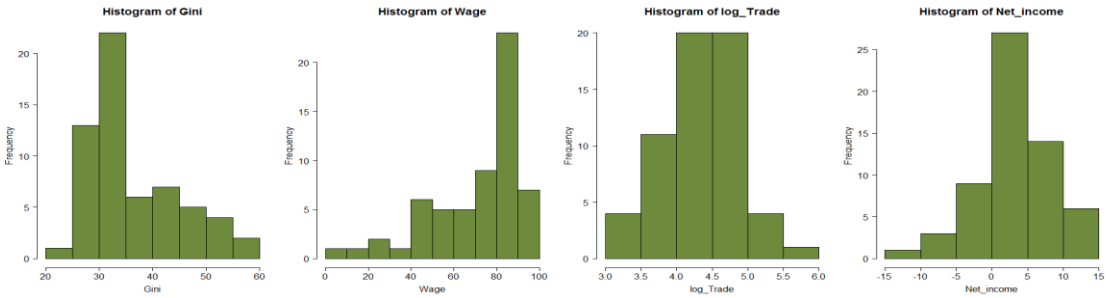
```
Analysis of Variance Table

Model 1: Gini ~ Wage + log_Trade + Net_income
Model 2: Gini ~ Wage + log_Trade + Net_income + log_Education
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     56 2522.7
2     55 2493.3  1    29.366 0.6478 0.4244
```

From the partial F-test, there is no significant sign that the beta of log_Education not equal to one, so we get rid of the variable “log_Education”, and fit the model by the rest variables.

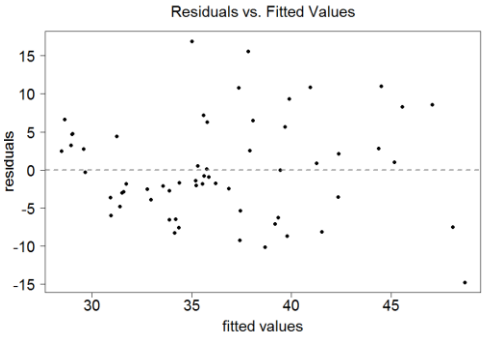



```
Call:
lm(formula = Gini ~ Wage + log_Trade + Net_income, data = gini_2)

Residuals:
    Min      1Q  Median      3Q     Max
-14.789  -4.128  -1.153   4.490  16.913

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 66.68679    8.51694   7.830 1.47e-10 ***
Wage        -0.15338    0.04352  -3.525 0.000853 ***
log_Trade   -4.65791    1.78213  -2.614 0.011483 *
Net_income   0.31517    0.18059   1.745 0.086435 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.712 on 56 degrees of freedom
Multiple R-squared:  0.3797,    Adjusted R-squared:  0.3464
F-statistic: 11.42 on 3 and 56 DF,  p-value: 5.981e-06
```

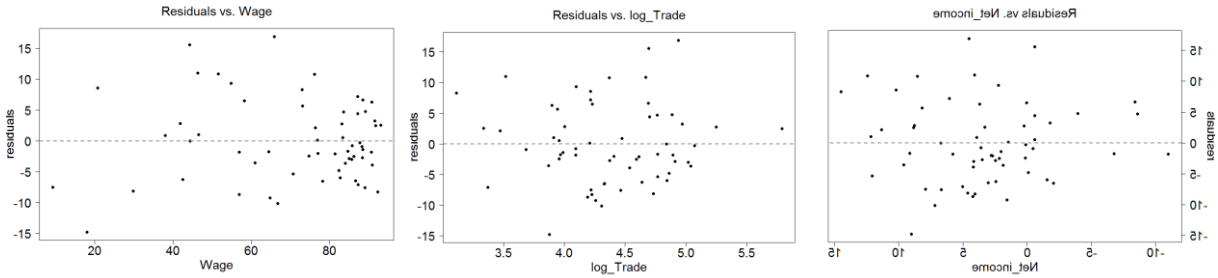
Finally, we derive a model that each variable is significant.

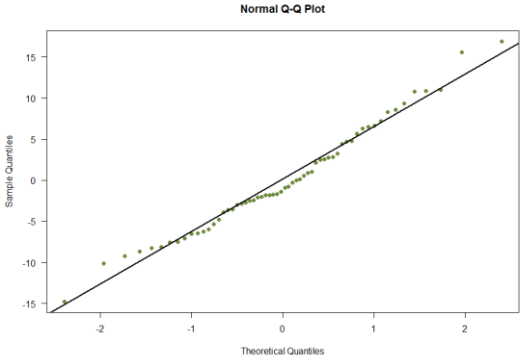


	Variables	VIF
1	Gini	1.612034
2	Wage	1.335570
3	log_Trade	1.227833
4	Net_income	1.235696

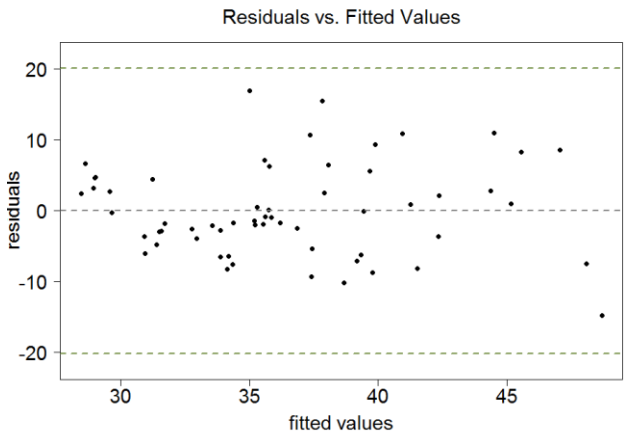


From the residual plot, there is no distinct linear trend, but variable of residuals increase as Gini increase, the hypothesis of constant variance is violated.





From the normal quantile plot, residuals are right-skewed, not normally distributed.



From the summery of the fitted linear model, the RMSE=s=6.712. From the residuals vs. fitted values plot, there is no point go beyond positive or negative 3s. As a result, there is no outlier.

As all the data is collected from year 2010, there is no possible that the data is a time series.

Results

The conclusions drawn by previous research give us some different factors that influence income inequality. They can be concluded into three main groups: the labor market, globalization and economic development factors.

After the presented analysis, a multiple regression analysis was made to study how the estimated components influence income inequality. It is also possible that some components covering factors presumed to influence inequality in actual fact do not influence income inequality at all. For a regression analysis, the Gini coefficient is selected as a dependent variable. The three index are included into the regression model as independent variables.

The estimates of the regression model are as follows:

Gini index = -0.15338 * Wage and salaried workers, total index -4.65791 * ln (Trade index) + 0.31517 * Adjusted net national income index + 66.68679

The model is significant with 100% confidence. The coefficient of the third component is significant with 95% confidence, the coefficient of the second component with 99% confidence and the coefficient of the first component with 100% confidence.

However, the fraction of variability in response variable which can be explained by regression model is somewhat low (which can be reflect by that R^2=0.3797 and adjusted R^2=0.3464), it illustrated that the regression model just has a weak correlation with Gini index.

In the future study, the model may be improved by adding more relative index into the regression, like the “GDP index”, “GDP growth index” and so on and the other form of transformation should be tried in the variable filtering process.

References

1. http://data.worldbank.org/indicator
2. https://en.wikipedia.org/wiki/Economic_inequality
3. https://en.wikipedia.org/wiki/Gini