Mingwei Huang
Eco 203 Project
12/8/2019

The effect of the economic statistics on inequality

Introduction

The wealth inequality has been seriously concerned since substantial polarity will trigger many issues and impede general welfare. But it is simply occurring with the determined property rights. There are almost infinite factors contributing to the problem, and they are interrelated in fact. It makes the phenomenon too complex to find its internal causes. Anyway, intuitively, there may be some correlations between wealth inequality in a country and the general economic statistics. And such information may shed light on how inequality is affected by macro economies.

In this study, I measured the inequality within countries by the Gini index which is used to measure the economic inequality of wealth distribution with range 0, perfect equality, to 100%, perfect inequality, and performed data analysis of the relationship between inequality and other economic indicators. To be specific, I wanted to see which of them are actually correlated to inequality.

Utilizing the data from the World Bank, I chose ten criteria of 267 countries based on my economic knowledge and some common sense and conducted a regression model of those data against the Gini coefficient in the last forty-one years (1979-2019). I collected the annual data of ten independent variables in total as my initial regression model. Since these data were not intact, I gathered, through programming, 367 sets of samples that had all the ten independent variables and the Gini index presented. By running a multiple regression on them, I could be informed which and how these independent variables are related to the Gini coefficient.

As a result, the final model satisfies all assumptions of linear regression. The error distribution is generally normal by the Jarque-Bera test, Omnibus test, and observing the shape of the histogram of standardized residuals. The model is Homoscedastic by Goldfeld-Quandt test and observing the scatter plot of residual over fitted values, which means I have no need to transform my dependent variable. The model has no multicollinearity by the VIF test and reading the correlation matrix. And it either has no serious outliers. Since the sample are collected through some algorithms from the original incomplete data set, it is not ordered in time, thus it does not have an autocorrelation.

The result implies that real interest rate, tertiary education enrollment rate, taxation, unemployment, and trade deficit do significantly relate to inequality. Most results confirm my expectation except the coefficient of interest rate which is positive. My hypothesis is a higher interest rate decreases the investment and therefore, the market becomes less active, solidifying the economic stratification. The rest shows quite reasonable patterns. Unemployment and trade deficit are positively correlated to inequality while tertiary education enrollment rate, taxation is negatively correlated. Last but worth mentioning, there tends to be little correlation between secondary education enrollment, population, and inequality.

Data Description

The data of this study is imported from the World Bank through the World Development Indicator. The dependent variable is the Gini index (*gini*), measuring wealth inequality. Even though there has been some criticism about the Gini coefficient, it is still one of the best and most famous measurements nowadays. And the independent variables include GDP per capita (constant LCU in million, denote as *gdp*), gross savings (% of GDP, denote as *gsav*), population (in million, denote as *pop*), real interest rates (*interest*), secondary education enrollment rate (*secedu*), tertiary education enrollment rate (*teredu*), tax revenue (% of GDP, denote as *tax*), unemployment rate(*unempl*), import of goods and services (% of GDP, denote as *imp*), and export of goods and services (% of GDP, denote as *exp*).

Table 1 shows the descriptive statistics of my sample which contains 367 observations of countries in years. The average Gini index of sampled countries in years is 37.507 with a relatively large variation. The mean unemployment rate is 8.09 percent is quite out of my expectation. There seem to be some outliers with have extreme unemployment but I have to see if the high unemployment has a significant influence on the regression model in order to drop them. Figure 1 presents the scatter plot between all individual independent variables against the dependent variable. Everything looks normal, and further analysis will be done looking at the partial regression plot after establishing a regression model.

Regression Analysis

Based on the data collected, I prompt the following regression model as the initial linear model:

$$\mathbf{gini} = \beta_0 + \beta_1 \boldsymbol{gdp} + \beta_2 \boldsymbol{gsav} + \beta_3 \boldsymbol{pop} + \beta_4 \boldsymbol{interest} + \beta_5 \boldsymbol{secedu} + \beta_6 \boldsymbol{teredu} + \beta_7 \boldsymbol{tax} + \beta_8 \boldsymbol{unempl} + \beta_9 \boldsymbol{imp} + \beta_{10} \boldsymbol{exp} + \varepsilon$$

Those independent variables are all possibly contributing to determining the dependent variable. For instance, the higher the unemployment rate there is, the more people suffer and thus the more relative poverty and higher the Gini index there may be.

The result of the initial regression model is shown in Table 2. The model is valid since the p-value for the overall F test is very insignificant. The R-square indicates that there is about 45.7 percent of the variation of the Gini coefficient is explained by the independent variables. However, there are some independent variables with a not very statistically important coefficient: *gsav*, *pop*, and *secedu*, which implies I should consider dropping these variables since they are too insignificant.

Besides, through the correlation matrix which is presented in Table 3, the correlation between import and export is over 0.8 which makes sense intuitively. Therefore, I substitute the two variables with the total trade (total), the sum of import and export, and the trade deficit (defic), the difference between the export and import.

Based on the previous analysis, I further develop this model into:

$$\mathbf{gini} = \beta_0 + \beta_1 \mathbf{\textit{gdp}} + \beta_2 \mathbf{\textit{interest}} + \beta_3 \mathbf{\textit{teredu}} + \beta_4 \mathbf{\textit{tax}} + \beta_5 \mathbf{\textit{unempl}} + \beta_6 \mathbf{\textit{total}} + \beta_7 \mathbf{\textit{defic}} + \varepsilon$$

And the statistics of the new regression model are displayed in Table 4. The increase of adjusted R-square and F statistic for the overall validity test is presented. Figure 2 demonstrates the standardized error distribution which is almost normal with some acceptable distortion. In addition, both the Jarqua-Bera test and the Omnibus test show very large possibilities of the distribution being normal: 95.4% and 93.5%, with a skew close to zero. Figure 3 and Figure 4 show the scatter plot between the residuals and predicted values, and the partial regression plot. All of them look fine as the residuals do not depend on the independent variables except that it looks like there is some tendency for the variance of the residuals increases as the predicted value increase though it is not obvious. By running the Goldfeld-Quandt test, the p-value with the null hypothesis of homoscedasticity is large, implying the small default is bearable and thus omitted. Hence, I can say the assumption of homoscedasticity holds statistically at the 10% level.

Referring to Table 3, there is no high correlation between studied independent variables now, and all the coefficients are quite statistically significant. To further test the multicollinearity, I choose not to use the conditional number since it does not perform well in dealing with high dimensional matrices. Instead, I implement the VIF test with a tolerance of 1.835 (1/(1-R^2)). The result is shown in Table 5, and all coefficients of independent variables are less than 1.835, meaning there is no multicollinearity in this model. At last, since the data is not ordered in time, the Durbin-Waston test is disabled. However, because the original data is collected annually, there is a possibility of autocorrelation. I add the time variable into the dataset and re-run a regression, generating the result shown in Table 6. As the time variable is not statistically important at alpha = 10%, I perform a partial F test dropping the variable time, resulting in an F statistic 1.03 which has a very high p-value. Thus, the autocorrelation does not exist in the model either.

The most serious outlier, detected by the outlier test from the python package, Statsmodels, is Honduras in 2008. It makes sense because one-fourth of the GDP of the country had relied upon foreign countries, and there started to be some political turbulence before the 2009 Honduran coup d'état. However, there is not any need to drop the outlier since does not have much influence on the whole model. I cannot visually display the leverage and the influential points graphs since the sample size is too large, but the outlier test shows that there is no serious outlier in my sample.

Therefore, all assumptions for linear regression models are satisfied and the model is valid while all independent variables are statistically important. My final model is:

$$\textbf{gini} = \beta_0 + \beta_1 \textbf{\textit{gdp}} + \beta_2 \textbf{\textit{interest}} + \beta_3 \textbf{\textit{teredu}} + \beta_4 \textbf{\textit{tax}} + \beta_5 \textbf{\textit{unempl}} + \beta_6 \textbf{\textit{total}} + \beta_7 \textbf{\textit{defic}} + \varepsilon$$

Empirical Results

The final regression model, depicted in Table 4, shows that GDP per capita, tertiary education enrollment, taxation, and total trade are negatively related to wealth inequality, but real interest rate, unemployment, and trade deficit are positively related. However, since the unit of GDP per capita is in million, and the mean is merely close to 3, while the magnitude of its coefficient is small, about one-tenth, GDP per capita does not play a very determinate role of inequality, though it is related somehow.

It is difficult to interpret the change in a way that can be directly perceived. And almost half of the countries have the Gini index between 30 and 40. Thus, it is better to keep in mind that even a change of 5 can be very magnificent.

It is interesting to see the real interest rates contribute, or at least related to the increase of the Gini index. The real interest rates actually have a sizable effect on the increase of the dependent variable since it has a large standard deviation of 9.7, with a coefficient of approximately 0.19, meaning it strongly influences the Gini index which has a mean of just about 37.5. I suspect this is because a less real interest rate causes more investment and flourishes economic activities which bring more opportunities and interactions between classes.

In addition, tertiary education does play a role in mitigating inequality. Even though the education system has been criticized for creating gaps among classes by high thresholds, it actually promotes the average level of capabilities, creating more opportunities and spaces for more human potentials. Of course, taxation does play a similar role since it redistributes wealth in some way that is not very polarized, theoretically.

Summary and Discussion

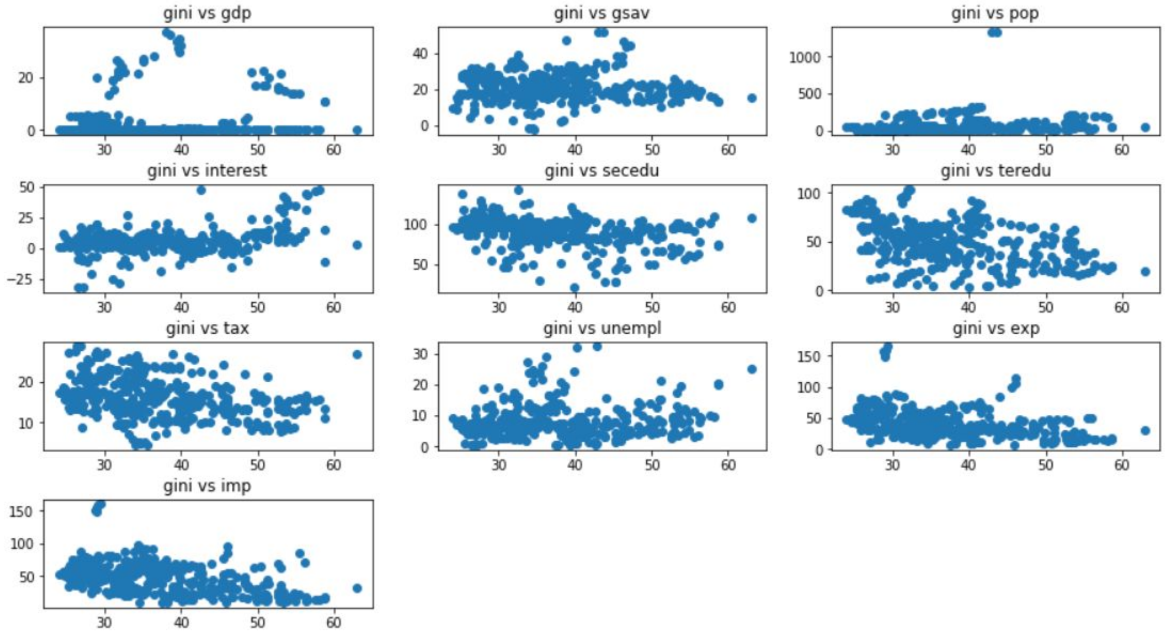
This study has explored some determinants of inequality within countries, measured by the Gini coefficients, focusing on the national economic statistics. As a result, I have found GDP per capita, real interest rate, tertiary education enrollment rate, tax revenue, the unemployment rate, total trade among countries, and trade deficit are statistically related to the Gini coefficient.

There are some shortfalls in the study. Firstly, the dependent variable, the Gini index, does not have a clear meaning in the linear regression model since it is difficult to figure out what really means by “the Gini index increases 2”, for example. This makes it hard to interpret the result of the final model, though there are some indications on the sign of the coefficients which are useful. I should have chosen the dependent variable more considerately.

Secondly, a simple statistical correlation does not reveal much information. And it is risking to make some hypothesis with the reference of the model for there is no clear cause-effect relation or even a direct relationship.

Lastly, the sample I am using only represents those having very complete recorded economic statistics. That excludes countries that are not as developed as to gather sufficient statistics. This can be modified with some techniques to analyze the incomplete data sets in the future.

Figure 1: Scatter plots of the Gini index against independent variables



Generated by python package, Matplotlib

Figure 2: Residuals Histogram

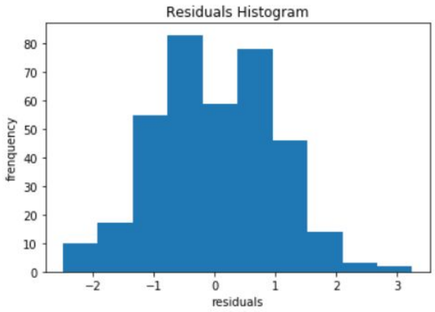
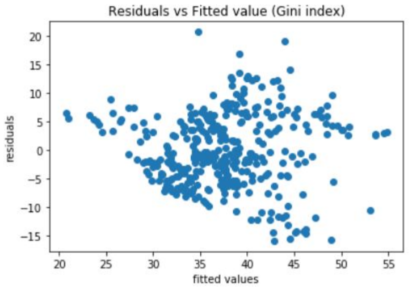


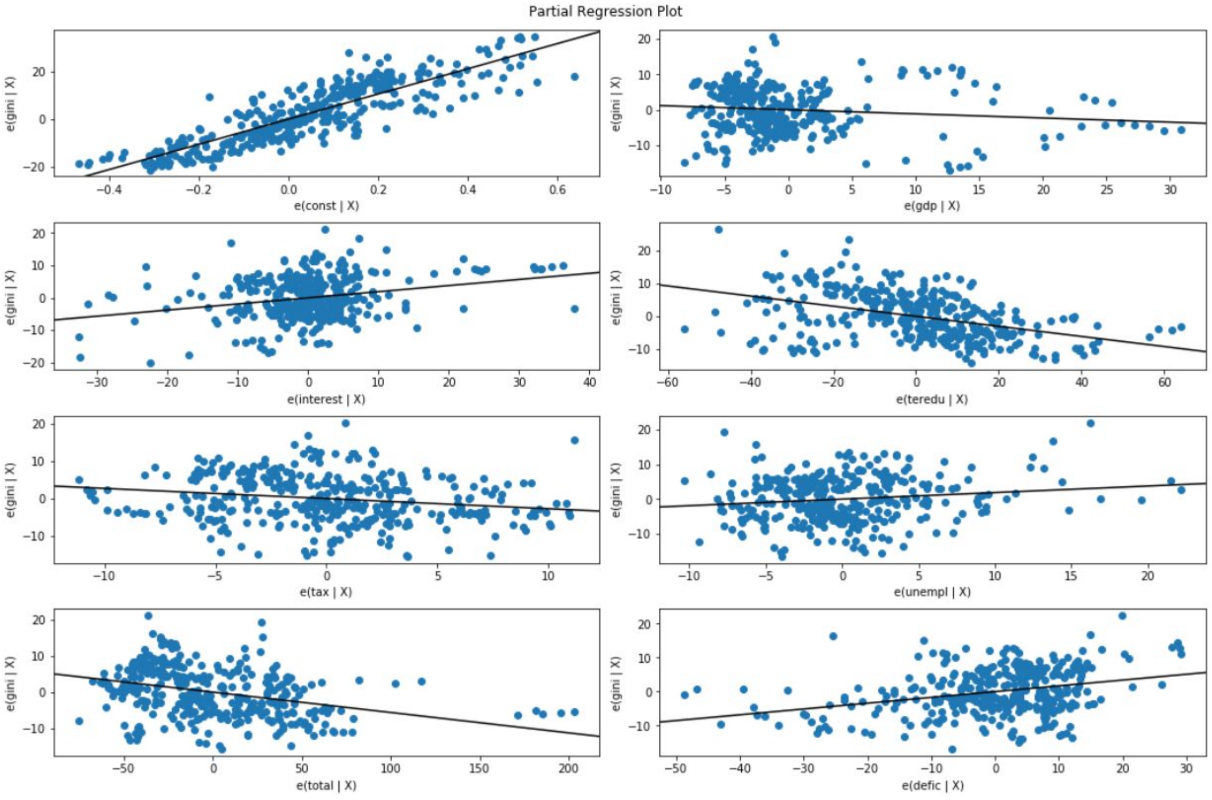
Figure 3: Residuals vs Fitted values

```
Goldfeld-Quandt test
[('F statstic', 0.4667039509589737), ('p-value', 0.9999996815426171)]
```



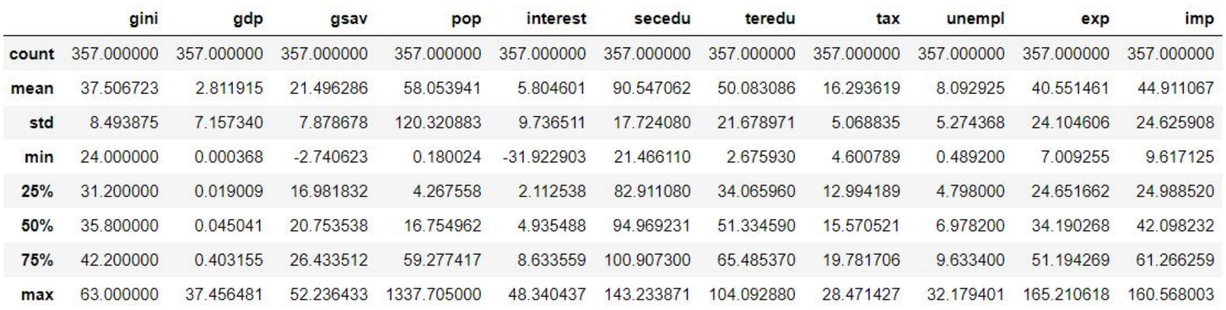
Generated by python package, Matplotlib

Figure 4: Partial regression plot



Generated by python package, Matplotlib

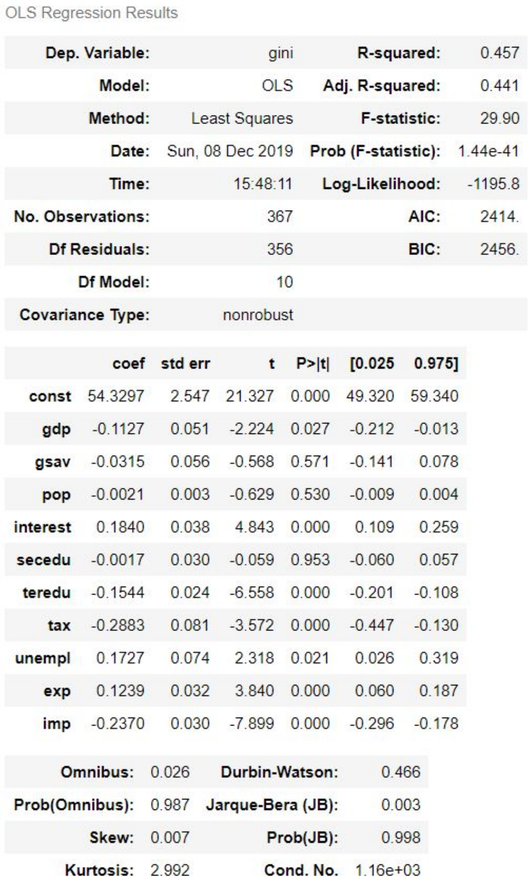
Table 1: Descriptive statistics



	gini	gdp	gsav	pop	interest	secedu	teredu	tax	unempl	exp	imp
count	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000	357.000000
mean	37.506723	2.811915	21.496286	58.053941	5.804601	90.547062	50.083086	16.293619	8.092925	40.551461	44.911067
std	8.493875	7.157340	7.878678	120.320883	9.736511	17.724080	21.678971	5.068835	5.274368	24.104606	24.625908
min	24.000000	0.000368	-2.740623	0.180024	-31.922903	21.466110	2.675930	4.600789	0.489200	7.009255	9.617125
25%	31.200000	0.019009	16.981832	4.267558	2.112538	82.911080	34.065960	12.994189	4.798000	24.651662	24.988520
50%	35.800000	0.045041	20.753538	16.754962	4.935488	94.969231	51.334590	15.570521	6.978200	34.190268	42.098232
75%	42.200000	0.403155	26.433512	59.277417	8.633559	100.907300	65.485370	19.781706	9.633400	51.194269	61.266259
max	63.000000	37.456481	52.236433	1337.705000	48.340437	143.233871	104.092880	28.471427	32.179401	165.210618	160.568003

Generated by python package, Pandas

Table 2: Initial model



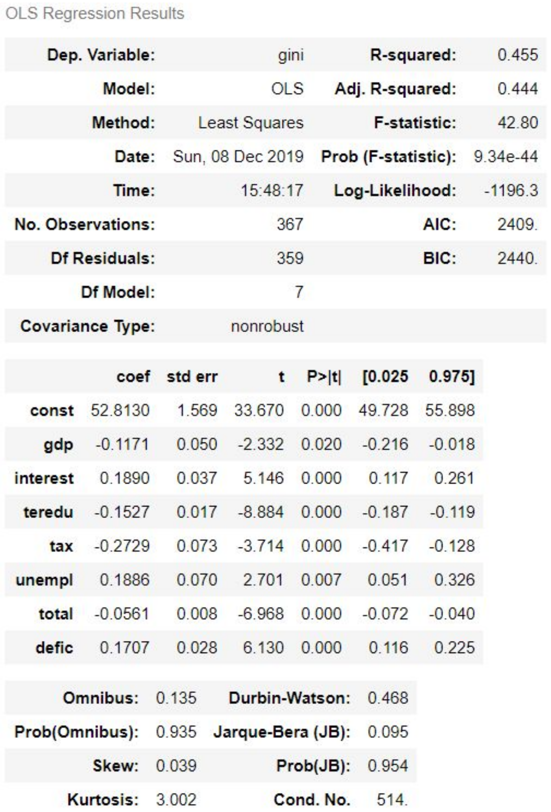
OLS Regression Results

Dep. Variable:	gini	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.441
Method:	Least Squares	**F-statistic:**	29.90
Date:	Sun, 08 Dec 2019	**Prob (F-statistic):**	1.44e-41
Time:	15:48:11	**Log-Likelihood:**	-1195.8
No. Observations:	367	**AIC:**	2414.
Df Residuals:	356	**BIC:**	2456.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	54.3297	2.547	21.327	0.000	49.320	59.340
gdp	-0.1127	0.051	-2.224	0.027	-0.212	-0.013
gsav	-0.0315	0.056	-0.568	0.571	-0.141	0.078
pop	-0.0021	0.003	-0.629	0.530	-0.009	0.004
interest	0.1840	0.038	4.843	0.000	0.109	0.259
secedu	-0.0017	0.030	-0.059	0.953	-0.060	0.057
teredu	-0.1544	0.024	-6.558	0.000	-0.201	-0.108
tax	-0.2883	0.081	-3.572	0.000	-0.447	-0.130
unempl	0.1727	0.074	2.318	0.021	0.026	0.319
exp	0.1239	0.032	3.840	0.000	0.060	0.187
imp	-0.2370	0.030	-7.899	0.000	-0.296	-0.178

Omnibus:	0.026	**Durbin-Watson:**	0.466
Prob(Omnibus):	0.987	**Jarque-Bera (JB):**	0.003
Skew:	0.007	**Prob(JB):**	0.998
Kurtosis:	2.992	**Cond. No.**	1.16e+03

Table 4: Final model



OLS Regression Results

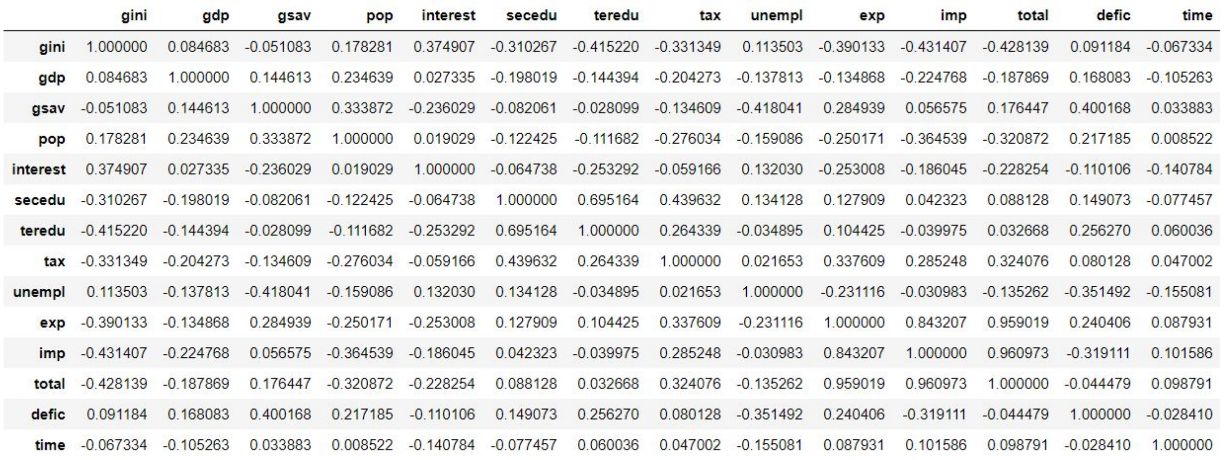
Dep. Variable:	gini	**R-squared:**	0.455
Model:	OLS	**Adj. R-squared:**	0.444
Method:	Least Squares	**F-statistic:**	42.80
Date:	Sun, 08 Dec 2019	**Prob (F-statistic):**	9.34e-44
Time:	15:48:17	**Log-Likelihood:**	-1196.3
No. Observations:	367	**AIC:**	2409.
Df Residuals:	359	**BIC:**	2440.
Df Model:	7		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.8130	1.569	33.670	0.000	49.728	55.898
gdp	-0.1171	0.050	-2.332	0.020	-0.216	-0.018
interest	0.1890	0.037	5.146	0.000	0.117	0.261
teredu	-0.1527	0.017	-8.884	0.000	-0.187	-0.119
tax	-0.2729	0.073	-3.714	0.000	-0.417	-0.128
unempl	0.1886	0.070	2.701	0.007	0.051	0.326
total	-0.0561	0.008	-6.968	0.000	-0.072	-0.040
defic	0.1707	0.028	6.130	0.000	0.116	0.225

Omnibus:	0.135	**Durbin-Watson:**	0.468
Prob(Omnibus):	0.935	**Jarque-Bera (JB):**	0.095
Skew:	0.039	**Prob(JB):**	0.954
Kurtosis:	3.002	**Cond. No.**	514.

Generated by python package, Matplotlib

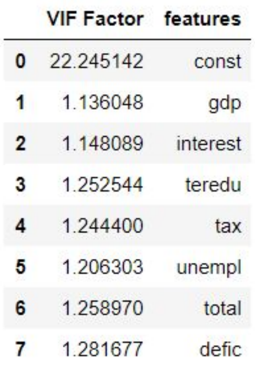
Table 3: Correlation matrix



	gini	gdp	gsav	pop	interest	secedu	teredu	tax	unempl	exp	imp	total	defic	time
gini	1.000000	0.084683	-0.051083	0.178281	0.374907	-0.310267	-0.415220	-0.331349	0.113503	-0.390133	-0.431407	-0.428139	0.091184	-0.067334
gdp	0.084683	1.000000	0.144613	0.234639	0.027335	-0.198019	-0.144394	-0.204273	-0.137813	-0.134868	-0.224768	-0.187869	0.168083	-0.105263
gsav	-0.051083	0.144613	1.000000	0.333872	-0.236029	-0.082061	-0.028099	-0.134609	-0.418041	0.284939	0.056575	0.176447	0.400168	0.033883
pop	0.178281	0.234639	0.333872	1.000000	0.019029	-0.122425	-0.111682	-0.276034	-0.159086	-0.250171	-0.364539	-0.320872	0.217185	0.008522
interest	0.374907	0.027335	-0.236029	0.019029	1.000000	-0.064738	-0.253292	-0.059166	0.132030	-0.253008	-0.186045	-0.228254	-0.110106	-0.140784
secedu	-0.310267	-0.198019	-0.082061	-0.122425	-0.064738	1.000000	0.695164	0.439632	0.134128	0.127909	0.042323	0.088128	0.149073	-0.077457
teredu	-0.415220	-0.144394	-0.028099	-0.111682	-0.253292	0.695164	1.000000	0.264339	-0.034895	0.104425	-0.039975	0.032668	0.256270	0.060036
tax	-0.331349	-0.204273	-0.134609	-0.276034	-0.059166	0.439632	0.264339	1.000000	0.021653	0.337609	0.285248	0.324076	0.080128	0.047002
unempl	0.113503	-0.137813	-0.418041	-0.159086	0.132030	0.134128	-0.034895	0.021653	1.000000	-0.231116	-0.030983	-0.135262	-0.351492	-0.155081
exp	-0.390133	-0.134868	0.284939	-0.250171	-0.253008	0.127909	0.104425	0.337609	-0.231116	1.000000	0.843207	0.959019	0.240406	0.087931
imp	-0.431407	-0.224768	0.056575	-0.364539	-0.186045	0.042323	-0.039975	0.285248	-0.030983	0.843207	1.000000	0.960973	-0.319111	0.101586
total	-0.428139	-0.187869	0.176447	-0.320872	-0.228254	0.088128	0.032668	0.324076	-0.135262	0.959019	0.960973	1.000000	-0.044479	0.098791
defic	0.091184	0.168083	0.400168	0.217185	-0.110106	0.149073	0.256270	0.080128	-0.351492	0.240406	-0.319111	-0.044479	1.000000	-0.028410
time	-0.067334	-0.105263	0.033883	0.008522	-0.140784	-0.077457	0.060036	0.047002	-0.155081	0.087931	0.101586	0.098791	-0.028410	1.000000

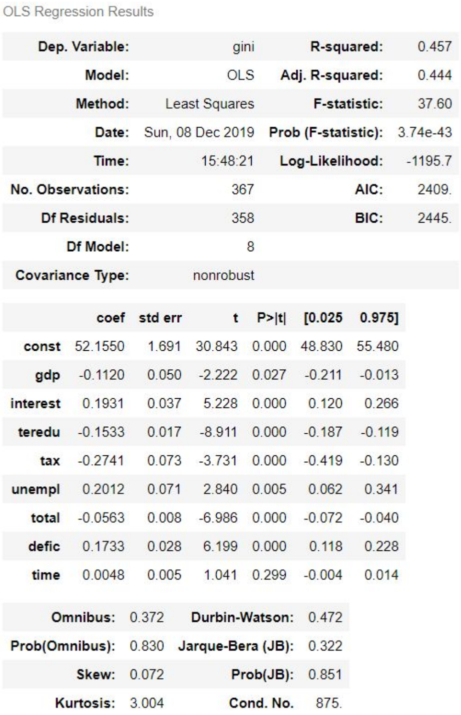
Generated by python package, Pandas

Table 5: VIF test



	VIF Factor	features
0	22.245142	const
1	1.136048	gdp
2	1.148089	interest
3	1.252544	teredu
4	1.244400	tax
5	1.206303	unempl
6	1.258970	total
7	1.281677	defic

Table 6: Model with time variable



OLS Regression Results

Dep. Variable:	gini	**R-squared:**	0.457
Model:	OLS	**Adj. R-squared:**	0.444
Method:	Least Squares	**F-statistic:**	37.60
Date:	Sun, 08 Dec 2019	**Prob (F-statistic):**	3.74e-43
Time:	15:48:21	**Log-Likelihood:**	-1195.7
No. Observations:	367	**AIC:**	2409.
Df Residuals:	358	**BIC:**	2445.
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	52.1550	1.691	30.843	0.000	48.830	55.480
gdp	-0.1120	0.050	-2.222	0.027	-0.211	-0.013
interest	0.1931	0.037	5.228	0.000	0.120	0.266
teredu	-0.1533	0.017	-8.911	0.000	-0.187	-0.119
tax	-0.2741	0.073	-3.731	0.000	-0.419	-0.130
unempl	0.2012	0.071	2.840	0.005	0.062	0.341
total	-0.0563	0.008	-6.986	0.000	-0.072	-0.040
defic	0.1733	0.028	6.199	0.000	0.118	0.228
time	0.0048	0.005	1.041	0.299	-0.004	0.014

Omnibus:	0.372	**Durbin-Watson:**	0.472
Prob(Omnibus):	0.830	**Jarque-Bera (JB):**	0.322
Skew:	0.072	**Prob(JB):**	0.851
Kurtosis:	3.004	**Cond. No.**	875.

Generated by python package, Matplotlib

The github source of the entire project: https://github.com/Scott-Huang/Eco203-Project