R: Income Inequality Analysis with Multi-linear Regression

Goal

In order to explore what factors can have a strong influence on the unequal distribution of wealth of a country, I build a multi-linear regression in R using several selected country-level features as independent variable and GINI index as dependent variable. Here, the GINI index represents the income or wealth distributed to the nation's residents, and is usually used to measure the inequality of income. The country-level features I selected are Co2 emission per capital, electric power consumption, annual GDP growth rate, expense (%) of GDP, government expenditure on education, agricultural land percentage, high-tech exports, population of young people, total population and unemployment rate.

Data

Data Source: World Bank data: http://data.worldbank.org2. explanation for each variable
Variable Description and Reason for Selection:

- **Co2 emission per capital:** this variable is in the environment section. A high co2 emission amount can represent that a county is industrialized but not developed very well. Therefore, the income gap between workers and business owners may be huge which leads to the inequality.
- **Electric power consumption:** this variable represents the life quality of residents in a country. high electric power consumption can show that most people in the country are able to use electric devices.
- **Annual GDP growth rate:** this variable is an important economic indicator that measure the development of a country. Increased GDP growth rate may contribute to lower GINI index.
- **Expense (%) of GDP:** Expense represent whether the economy is booming or not. High expense represents that people have more money than they need for basic living, therefore it may indicate a low GINI index.
- **Government expenditure on education(% of government expenditure):** People with higher education level tends to earn higher wages. Therefore, high expenditure on education may indicate that more people can receive basic education and result in a lower GINI index.
- **Agricultural land percentage:** Since many poor country's main labor force are farmers; they have a high agricultural land percentage. Therefore, I guess there is a correlation between agriculture and GINI index.
- **High-tech export:** High high-tech export represents that a country is has advanced technology. Since nowadays the economy is technology motivated, it is a good measure for a country's achievement and development.

- **Population ages 15-64, total:** More young people in a country represents more labors which means most of the population can work and earn a living. Therefore, the number of people in poverty can be decreased.
- **Total population:** Large population may indicate a low GINI index because it is difficult to solve poverty issue in that case.
- **Unemployment rate:** This variable tells how many people are under the unemployed state. High unemployment rate means the economy is in recession which leads to a high GINI index.

Model Building

First, I use backward elimination on the dataset to get a raw model with 9 variables using R’2 adjusted, cp and BIC as criteria. Then, I performed individual t test on the model to see the significance of each variable. The result shows that several variables are not significant with p-value smaller than 5%. Therefore, I removed those variables and conducted a partial F-test to compare the second model with the first model and null hypothesis is not rejected.

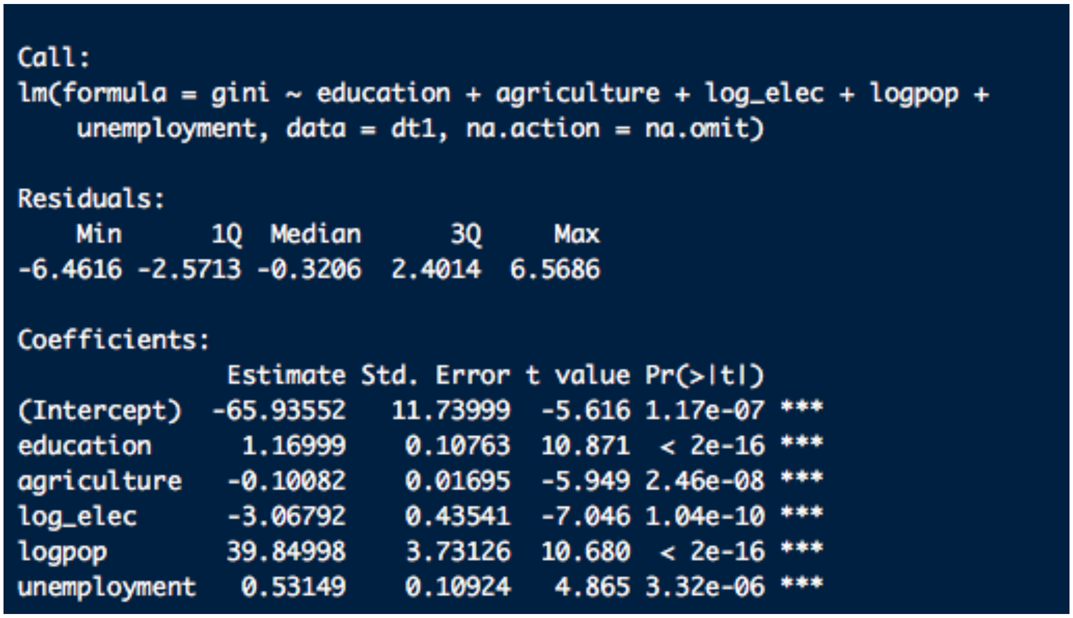
Then I checked the regression assumption of the second model. It has collinearity and constant variance problem. To address the collinearity problem, I removed the variable **co2_emission** which have high correlation with **electric power consumption**. To solve the constant variance problem, I transformed variables and removed outliers. I use log transformation on **total population** and **electric power consumption**, because these two are highly skewed variables. Also, I removed outliers from dataset using cook distance because the box plot and scatter plot both show that many variables have outliers.

Finally, I build the final model with five selected and transformed variables. The assumptions for linear regression are all checked with plots. The only violation is the normality violation but it’s not severe.

Final Model

Gini = 1.16999***Government expenditure on education(%)** - 0.10082***Agricultural land percentage** - 3.06792*log(**Electric power consumption**) + 39.84998log(**Total population**) + 0.53149**Unemployment rate**

The final model shows that the there is a positive correlation between Gini and **Government expenditure on education, Total population and Unemployment rate.** There is a negative correlation between Gini and **Agriculture land percentage**, **Electric power consumption**.



```
Call:
lm(formula = gini ~ education + agriculture + log_elec + logpop +
    unemployment, data = dt1, na.action = na.omit)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4616 -2.5713 -0.3206  2.4014  6.5686

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -65.93552   11.73999  -5.616 1.17e-07 ***
education      1.16999    0.10763  10.871  < 2e-16 ***
agriculture   -0.10082    0.01695  -5.949 2.46e-08 ***
log_elec      -3.06792    0.43541  -7.046 1.04e-10 ***
logpop        39.84998    3.73126  10.680  < 2e-16 ***
unemployment   0.53149    0.10924   4.865 3.32e-06 ***
```

What Next

The higher the Gini Coefficient, the greater the degree of inequality. The model indicates that a country with high government expenditure on education(% of total), large population and high unemployment rate tends to have inequality in the wealth distribution, while country with high agriculture land percentage and high electric power consumption tend to have equality income distribution. The findings seems to be interesting to me. When I selected those variables, I assumed that country with government expenditure on education(% of total) would have more equal incomes. The next step is to collect data related to education expenditure to evaluate why and how does this correlation happen.