STAT306 Project
Jiaqi Zhao 75619916

STAT306 Project Final Report

1. Introduction

1.1 Motivation

It is generally accepted that commerce is based on sound industry and agriculture, and that education is an important factor in economic development. In this project, we wanted to find useful insight into business development. We also wanted to see if the economic distribution of commerce, industry, and agriculture shares differ between high and low-income countries. Since, the economic positions of the two kinds of countries are different, we can also add an interaction term.

In the following research experiment we will be answering the research question: How does the market value of domestic companies in a country relate to the development of industry, agriculture, expenditure on education and income level of the country. We aim to find a linear model between firm market capitalization and the four explanatory variables. We assume it is positively correlated with industry, negatively correlated with agriculture, and positively correlated with education.

1.2 Data collecting

The World Bank Group works in every major area of development, and it provides free and open access to global development data (https://data.worldbank.org/). DataBank is an analysis and visualisation tool that contains collections of time series data on a variety of topics. We can create our own queries by this tool (https://databank.worldbank.org/home and https://databank.worldbank.org/source/world-development-indicators). For this research, we chose the database *World Development Indicators* as the resource and our five variables are retrieved from DataBank tool by selecting the corresponding indicators. Specifically the original source of our education data is UNESCO Institute for Statistics (http://uis.unesco.org/) which is also found in the database *World Development Indicators*. Data are end of year values randomly chosen from the last 5 to 10 years and 233 sets of data are chosen in this research. Specific descriptions of the variables are as follows:

1) Response variable: Market capitalization of listed domestic companies (% of GDP) is measured by the weighted average of share price times the number of shares outstanding (including their several classes) for listed domestic companies.

2) Explanatory continuous variable: Industry, value added (% of GDP) is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It comprises value added in mining, manufacturing (also reported as a separate subgroup), construction, electricity, water, and gas. Total GDP is measured at purchaser prices. Value added by industry is normally measured at basic prices.

3) Explanatory continuous variable: Agriculture, value added (% of GDP) is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It includes forestry, hunting, fishing, and cultivation of crops and livestock production. Total GDP is measured at purchaser prices. Value added by industry is normally measured at basic prices.

4) Explanatory continuous variable: Government expenditure on education, total (% of GDP) is calculated by dividing total government expenditure for all levels of education by the GDP, and multiplying by 100. All the data are mapped to the International Standard Classification of Education (ISCED) to ensure the comparability of education programs at the international level.

5) Explanatory dummy variable: Income, indicating whether the data source country is a high-income country. The high-income country defined by the World Bank (https://data.worldbank.org/country) is taken as 1, the low-income country is 0 (baseline).

2. Analysis

2.1 Pre-Processing

In *figure 0*, as there is a big difference between the firm market capitalization of high-income and low-income countries, we decided to analyze high income and low income as a categorical variable to ensure the accuracy of the model.

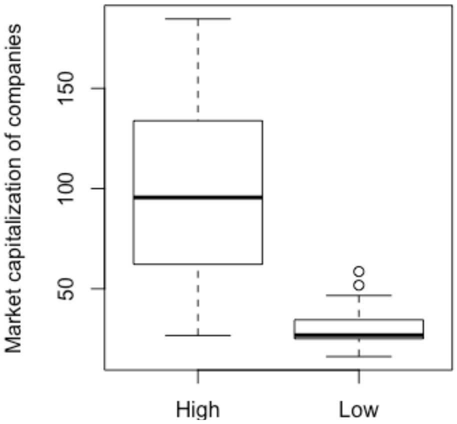


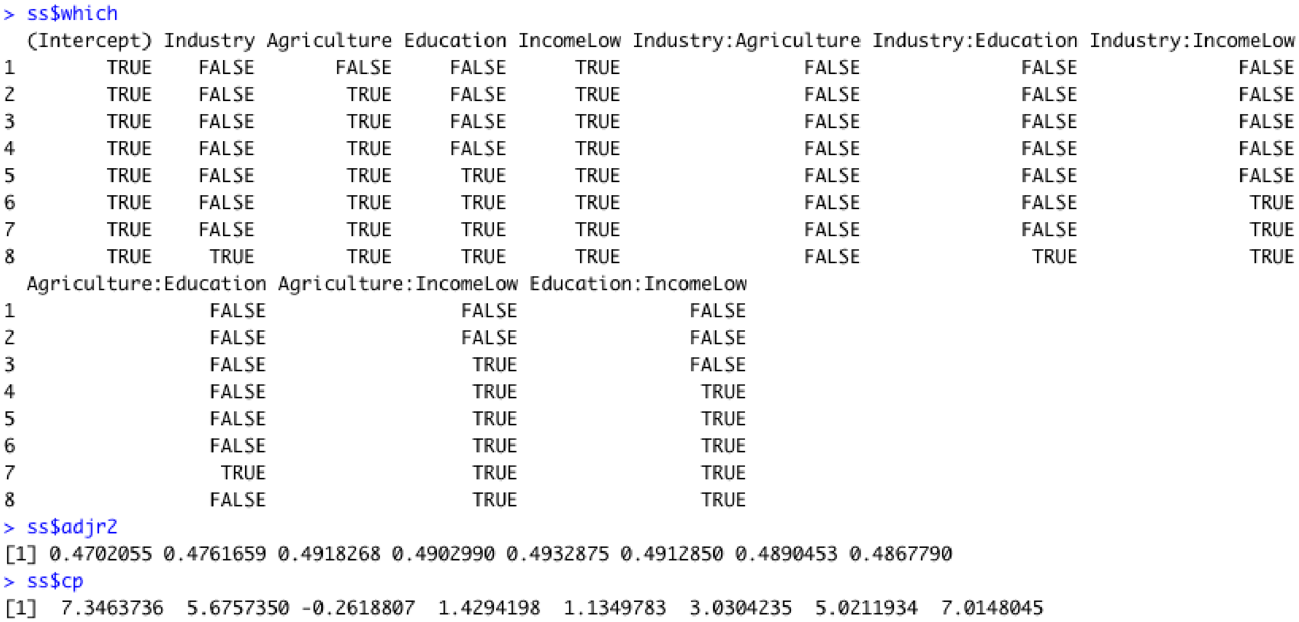
Figure 0: box plot of market capitalization for low-income and high-income country

Before studying which model can best represent the relationship between market value of domestic companies and development of industry, agriculture, expenditure on education and income level of the country, we need to define a full model first. We have four explanatory variables, which means we actually have 11 interaction terms to consider($\binom{4}{2} + \binom{4}{3} + \binom{4}{4} = 11$). Considering that the parameter quantity of the model is too large after adding the 11 terms, we first focus only on the interaction between two variables like we always encountered in class, which are 6 extra possible terms.

2.2 Model Selection

We use *regsubsets* command to do the exhaustive search of all models (4 explanatory variables with 6 interaction terms). From the R output in *Figure 1*,

1. The largest adjR2 gives size 5, with variables Agriculture + Education + Income + Education*Income + Agriculture*Income
2. The smallest Cp gives size 3, with variables Agriculture + Income + Agriculture*Income
3. The Cp which is closest to p(number of parameters) gives size 8, with Industry*Agriculture+Industry*Education+Industry*Income+Agriculture*Education+Agriculture*Income+Education*Income



```
> ss$which
  (Intercept) Industry Agriculture Education IncomeLow Industry:Agriculture Industry:Education Industry:IncomeLow
1        TRUE    FALSE       FALSE     FALSE      TRUE                FALSE              FALSE              FALSE
2        TRUE    FALSE        TRUE     FALSE      TRUE                FALSE              FALSE              FALSE
3        TRUE    FALSE        TRUE     FALSE      TRUE                FALSE              FALSE              FALSE
4        TRUE    FALSE        TRUE     FALSE      TRUE                FALSE              FALSE              FALSE
5        TRUE    FALSE        TRUE      TRUE      TRUE                FALSE              FALSE              FALSE
6        TRUE    FALSE        TRUE      TRUE      TRUE                FALSE              FALSE               TRUE
7        TRUE    FALSE        TRUE      TRUE      TRUE                FALSE              FALSE               TRUE
8        TRUE     TRUE        TRUE      TRUE      TRUE                FALSE               TRUE               TRUE
  Agriculture:Education Agriculture:IncomeLow Education:IncomeLow
1                 FALSE                 FALSE               FALSE
2                 FALSE                 FALSE               FALSE
3                 FALSE                  TRUE               FALSE
4                 FALSE                  TRUE                TRUE
5                 FALSE                  TRUE                TRUE
6                 FALSE                  TRUE                TRUE
7                  TRUE                  TRUE                TRUE
8                 FALSE                  TRUE                TRUE
> ss$adjr2
[1] 0.4702055 0.4761659 0.4918268 0.4902990 0.4932875 0.4912850 0.4890453 0.4867790
> ss$cp
[1]  7.3463736  5.6757350 -0.2618807  1.4294198  1.1349783  3.0304235  5.0211934  7.0148045
```

Figure 1: R output for exhaustive method for model with 11 parameters

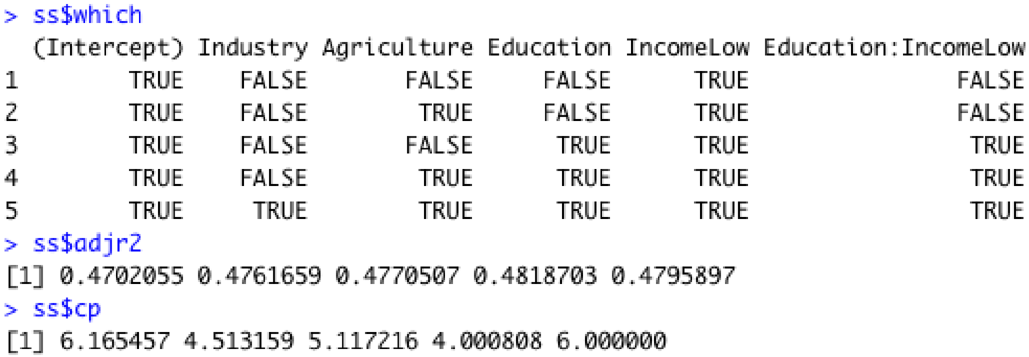
But we also have further insight. According to some of our literature sources it implies that Industrial and Agriculture do not directly affect education as much as the income level of a country's population. We also chose a model with all 4 explanatory variables and an interaction between Income and Education, which are 6 parameters in total. Then we also did the model fitting starting from this model.

We found that there are many studies on the relationship between government expenditures and education levels and most scholars agree that there is a positive correlation between the two (Blankenau, Simpson & Tomljanovich, 2007). And Awaworyi Churchill & Yew (2017) concludes that education spending is an important investment for the government. Government spending has a direct impact on education development (Gutiérrez-Garrido & Acuña-Duarte 2020). Therefore, it is reasonable to assume that the development of education will be different in high-income and low-income countries, thus affecting the development of commerce. Therefore, we add an interaction term between education and income level.

Further, we don't think industry and agriculture have a direct effect on education as much as the income level does. Our reasoning behind this is that nearly half of the funding for public educational institutions comes from local government taxes, followed by tuition and public donations (Biddle & Berliner 2002). In contrast, the development of industry and agriculture is shown more in the difference between high and low income. We can say if industry and agriculture do have an impact on education, it can also be reflected in the indicator of income, which means we don't need interaction terms between industry, agriculture, and education. This avoids making the model more complex and is also consistent with the reality that the government regulates the development of education through taxation.

Then we use *regsubsets* command to do the exhaustive and forward search respectively of those models (4 explanatory variables with 1 interaction term between Income and Education). From R output with exhaustive method in *Figure 2*,

(1) The largest adjR2 gives size 4 , with variables Agriculture + Education + Income + Education*Income
(2) The smallest Cp gives size 4 as well and it is also close to p=5
(3) The Cp which is closest to p(number of parameters) gives size 5, with variables Industry + Agriculture + Education + Income + Education*Income

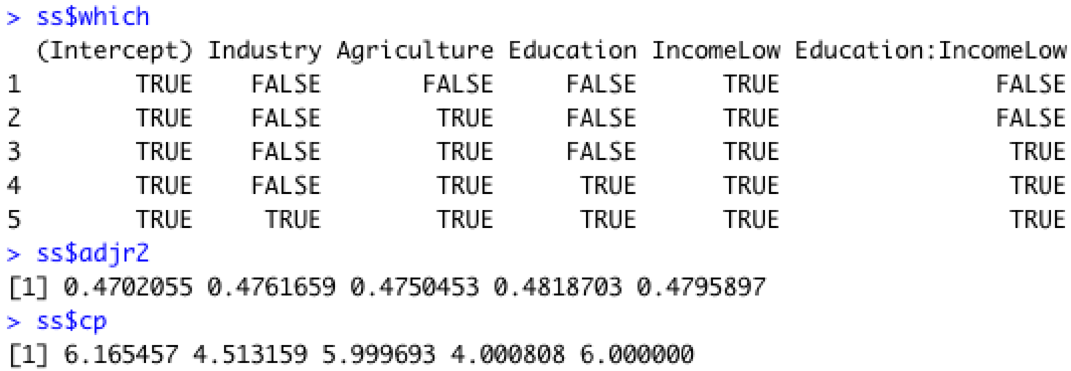


```
> ss$which
  (Intercept) Industry Agriculture Education IncomeLow Education:IncomeLow
1        TRUE    FALSE       FALSE     FALSE      TRUE               FALSE
2        TRUE    FALSE        TRUE     FALSE      TRUE               FALSE
3        TRUE    FALSE       FALSE      TRUE      TRUE                TRUE
4        TRUE    FALSE        TRUE      TRUE      TRUE                TRUE
5        TRUE     TRUE        TRUE      TRUE      TRUE                TRUE
> ss$adjr2
[1] 0.4702055 0.4761659 0.4770507 0.4818703 0.4795897
> ss$cp
[1] 6.165457 4.513159 5.117216 4.000808 6.000000
```

Figure 2: R output for exhaustive method for model with 6 parameters

From R output with forward method in *Figure 3*,

(1) The largest adj_R2 gives size 4, which is the same model as exhaustive method
(2) The smallest Cp gives size 4, which is the same model as exhaustive method
(3) But the Cp which is closest to p (number of parameters) gives size 5, with variables Industry + Agriculture + Education + Income + Education*Income



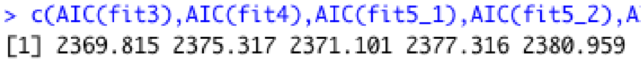
```
> ss$which
  (Intercept) Industry Agriculture Education IncomeLow Education:IncomeLow
1        TRUE    FALSE       FALSE     FALSE      TRUE               FALSE
2        TRUE    FALSE        TRUE     FALSE      TRUE               FALSE
3        TRUE    FALSE        TRUE     FALSE      TRUE                TRUE
4        TRUE    FALSE        TRUE      TRUE      TRUE                TRUE
5        TRUE     TRUE        TRUE      TRUE      TRUE                TRUE
> ss$adjr2
[1] 0.4702055 0.4761659 0.4750453 0.4818703 0.4795897
> ss$cp
[1] 6.165457 4.513159 5.999693 4.000808 6.000000
```

Figure 3: R output for forward method for model with 6 parameters

So we compare 5 model's performance finally, and the 5 models are respectively(name is based on size):

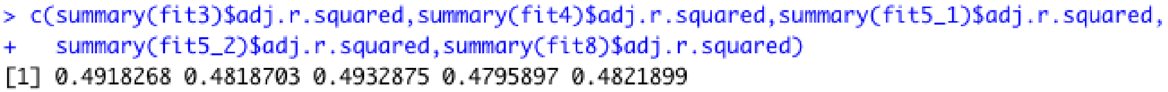
- Model3: Agriculture + Income + Agriculture*Income
- Model4: Agriculture + Education + Income + Education*Income
- Model5_1: Agriculture + Education + Income + Education*Income + Agriculture*Income
- Model5_2: Industry + Agriculture + Education + Income + Education*Income
- Model8: Industry*Agriculture+Industry*Education+Industry*Income+Agriculture*Education+Agriculture*Income+Education*Income

First, we compare their value of AIC, R output(*Figure 4* & *Figure 5)* tell us that the model 3 has the smallest AIC and the largest adjR2, so model3 has the best performance.



```
> c(AIC(fit3),AIC(fit4),AIC(fit5_1),AIC(fit5_2),AIC(fit8))
[1] 2369.815 2375.317 2371.101 2377.316 2380.959
```

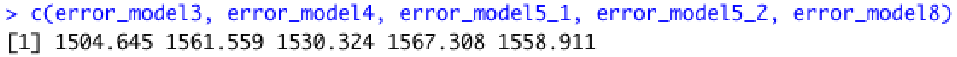
Figure 4: AIC comparisons



```
> c(summary(fit3)$adj.r.squared,summary(fit4)$adj.r.squared,summary(fit5_1)$adj.r.squared,
+   summary(fit5_2)$adj.r.squared,summary(fit8)$adj.r.squared)
[1] 0.4918268 0.4818703 0.4932875 0.4795897 0.4821899
```

Figure 5: adjR^2 comparisons

Since we can't get the best model from adjR2 and AIC, we do the Leave-One-Out-Cross-Validation (LOOCV) and get the error of the 5 models containing different explanatory variables. Given the R output(*Figure 6)*, model3 has the best performance with least error.



```
> c(error_model3, error_model4, error_model5_1, error_model5_2, error_model8)
[1] 1504.645 1561.559 1530.324 1567.308 1558.911
```

Figure 6: LOOCV error comparisons

2.3 Log Model Selection

So far, the Model3 performed best among these models we compared. Then we check the Normal QQ plot (*Figure 7)*, and we realize that taking the log of response variable may seem better.

After taking the log of the response variable, the adjR2 of model3 increased from 0.4932875 to 0.6046465 and AIC dropped from 2360.676 to 380.1911, so new log_model3 is indeed performing much better.
To make this fitted model more convincing, we rebuilt the model with all four explanatory variables and an interaction between Income and Education, while taking the log of the response variable and utilizing the *regsubsets* command to do an exhaustive and forward search of all models. Based on the adjR2 and Cp from the R output(*Figure 8* & *Figure 9)*, log_Model3 and log_Model4 both seem to be a good fit.

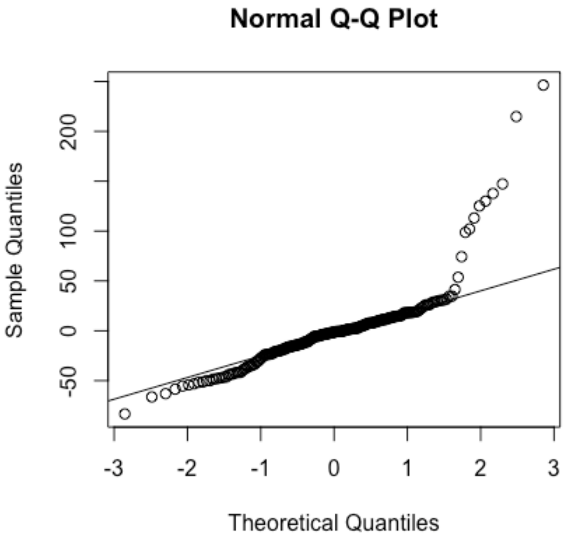
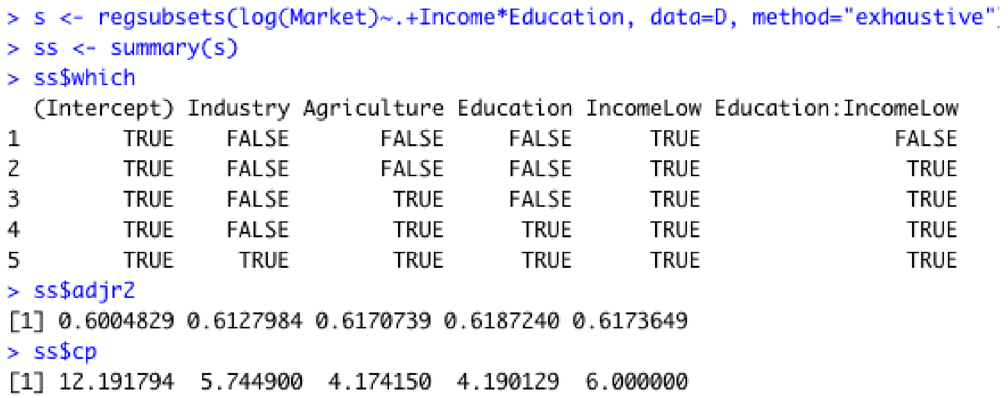
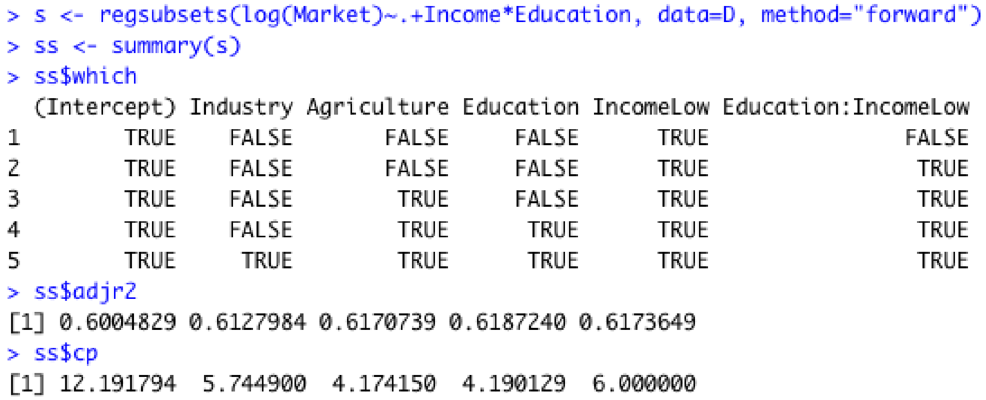


Figure 7: Normal Q-Q plot for the best model we selected(Model3)



```
> s <- regsubsets(log(Market)~.+Income*Education, data=D, method="exhaustive"
> ss <- summary(s)
> ss$which
  (Intercept) Industry Agriculture Education IncomeLow Education:IncomeLow
1        TRUE    FALSE       FALSE     FALSE      TRUE               FALSE
2        TRUE    FALSE       FALSE     FALSE      TRUE                TRUE
3        TRUE    FALSE        TRUE     FALSE      TRUE                TRUE
4        TRUE    FALSE        TRUE      TRUE      TRUE                TRUE
5        TRUE     TRUE        TRUE      TRUE      TRUE                TRUE
> ss$adjr2
[1] 0.6004829 0.6127984 0.6170739 0.6187240 0.6173649
> ss$cp
[1] 12.191794  5.744900  4.174150  4.190129  6.000000
```

Figure 8: R output for exhaustive method for new log model



```
> s <- regsubsets(log(Market)~.+Income*Education, data=D, method="forward")
> ss <- summary(s)
> ss$which
  (Intercept) Industry Agriculture Education IncomeLow Education:IncomeLow
1        TRUE    FALSE       FALSE     FALSE      TRUE               FALSE
2        TRUE    FALSE       FALSE     FALSE      TRUE                TRUE
3        TRUE    FALSE        TRUE     FALSE      TRUE                TRUE
4        TRUE    FALSE        TRUE      TRUE      TRUE                TRUE
5        TRUE     TRUE        TRUE      TRUE      TRUE                TRUE
> ss$adjr2
[1] 0.6004829 0.6127984 0.6170739 0.6187240 0.6173649
> ss$cp
[1] 12.191794  5.744900  4.174150  4.190129  6.000000
```

Figure 9: R output for forward method for new log model

However, we noticed that there is no Education term in log_Model3 yet it has an interaction term of Education and Income, which makes no sense of the interaction. And we also notice that the Cp of log_Model6 is close to p, so we use LOOCV to compare log_Model4 and log_Model6. Based on the result, log_Model4 "log(Market) ~ Agriculture + Income + Education Education*Income" is our final model.

Then we check the residual plot and QQ plot of log_Model4. There are two clusters in the fitted value vs residual plot in *Figure 10*, which might be an indication of some underlying and unaccounted for variables. We believe this might be due to the categorical variable 'income', which has two levels, as shown in *figure 11*. We think this is a reasonable explanation, because business development in high and low-income countries is different indeed, both in reality and in the results of the research data. Further, the QQ plot of log_Model4 seems to be good.

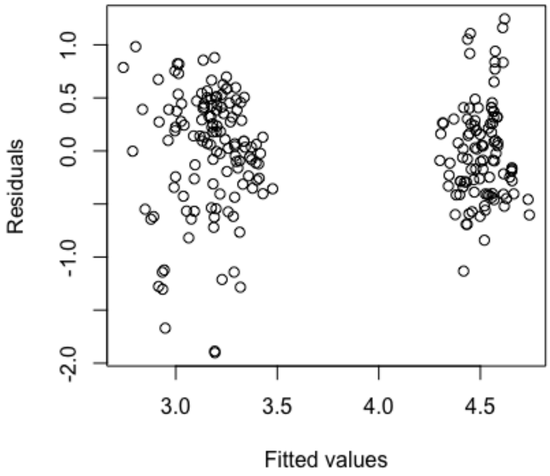


Figure 10: Residual plot vs fitted value

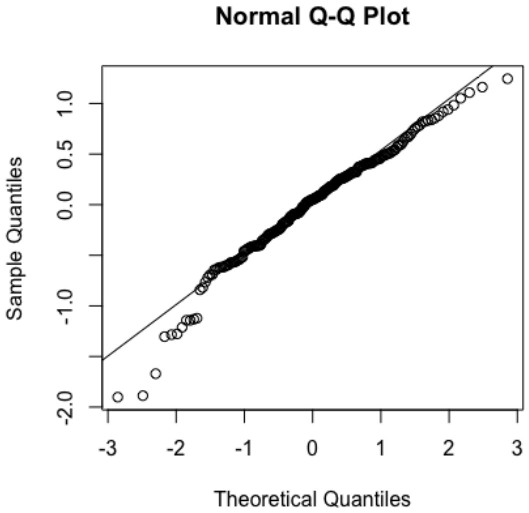


Figure 12: Normal QQ plot for log_Model4

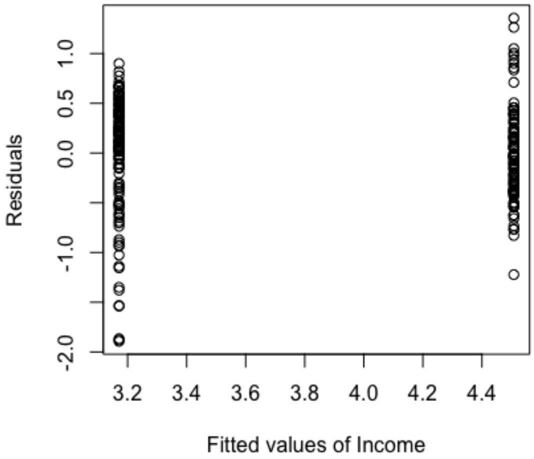
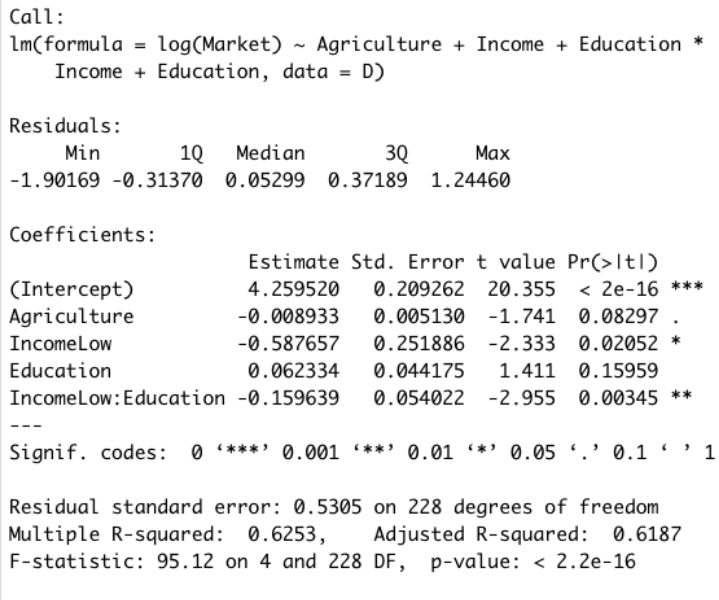


Figure 11: Residual plot vs fitted value of income

3. Conclusion

We have concluded that the relation between market capitalization of select companies and our explanatory variables is logarithmic. Specifically, the relation is described by the model in *Figure 13*. As the plots in section 2 show, a logarithmic model, and in particular, log_model4 is the best choice for our data. This model also suggests that in predicting market capitalization, industry is not as relevant as the other explanatory variables.

Our fitted model of market capitalization of companies is



```
Call:
lm(formula = log(Market) ~ Agriculture + Income + Education *
    Income + Education, data = D)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.90169 -0.31370  0.05299  0.37189  1.24460 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)           4.259520   0.209262  20.355  < 2e-16 ***
Agriculture          -0.008933   0.005130  -1.741  0.08297 .  
IncomeLow            -0.587657   0.251886  -2.333  0.02052 *  
Education             0.062334   0.044175   1.411  0.15959    
IncomeLow:Education -0.159639   0.054022  -2.955  0.00345 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5305 on 228 degrees of freedom
Multiple R-squared:  0.6253,	Adjusted R-squared:  0.6187 
F-statistic: 95.12 on 4 and 228 DF,  p-value: < 2.2e-16
```

Figure 13: Summary of log_Model4

$Log(Market) = 4.25952\ -\ 0.008933 * Agriculture\ +\ 0.062334 * Education\ - 0.587657 * IncomeLow - 0.159639 * Education * IncomeLow + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2\ )$ and IncomeLow =1 for low-income countries and 0 for else

$$\widehat{Market} = e^{4.25952\ -\ 0.008933 * Agriculture\ +\ 0.062334 * Education\ -\ 0.587657 * IncomeLow\ -\ 0.159639 * Education * IncomeLow}$$

It shows the biggest influence on the market capitalization is Income-level, the high-income countries have on average 31.5 (percentage of GDP) more on market capitalization than low-income countries, while keeping
other variables constant. Further, it shows that for high-income countries, market capitalization will increase as the expenditure on education increases. But, for low-income countries, market capitalization will decrease as the expenditure on education increases. Last, agricultural growth has a negative effect on market capitalization, but the power of influence is not as much as other variables.

4. Discussion

4.1 Limitation

Though it seems that we have found a suitable model to explain the relationship between the market value of domestic companies in a country and the development of industry, agriculture, expenditure on education and income level of the country. But the model we selected still has some limitations in the process.
Firstly, in order to make the model relatively simple, we excluded the possibility of interaction with more than two variables, but in real life this cannot be excluded. It is possible that the interaction between industry, agriculture, expenditure on education and income level of the country together may affect the fitting of the model, which might undermine a little bit the accuracy of the models that we've chosen now.
Besides, this relationship may not be fixed, but may change with time or some major events. However, since the data we have collected now is limited, we cannot completely guarantee that this is an accurate model, but at least it will provide some guidance to our research question.

4.2 Future Work

For further exploration, we can gather more data and take a more holistic view starting from the real full model which contains all the explanatory variables and all possible interactions. Further, as mentioned above, there are two clusters in Figure 9, which might be an indication of some underlying and unaccounted for variables. In this research we assume this is because the two levels of Income, but in future studies, if we want a more detailed analysis, we can increase the level of Income, say low income, low and middle income, middle and high income, and high income countries. Besides, we can analyze the underlying relationship between the explanatory variables. For example, we can do the logistic regression and take Income as a response variable. Or we could break down Education into several indicators, like the share of higher education, or the population of secondary education, etc, and that would give us a better understanding of what it entails to promote economic growth.

5. Reference

Awaworyi Churchill, S., Ugur, M., & Yew, S. L. (2017). Government education expenditures and economic growth: A meta-analysis. The B.E. Journal of Macroeconomics, 17(2) doi:10.1515/bejm-2016-0109

Blankenau, W. F., Simpson, N. B., & Tomljanovich, M. (2007). Public education expenditures, taxation, and growth: Linking data to theory. The American Economic Review, 97(2), 393-397. doi:10.1257/aer.97.2.393

Biddle, B. J., & Berliner, D. C. (2002). A Research Synthesis / Unequal School Funding in the United States. *Beyond Instructional Leadership, 59*(8).

Gutiérrez-Garrido, F. M., & Acuña-Duarte, A. A. (2020). Local government expenditure on education and its effect on income inequality at the county level in chile. Ecos De Economía, 23(49), 4-28. doi:10.17230/ecos.2019.49.1