Mengjie Shen

Jeffrey Simonoff

Regression and Multivariate Data Analysis

1st March 1, 2021

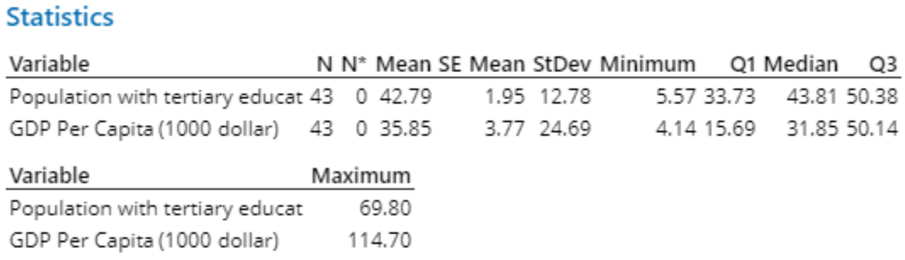
Linear Regression Analysis: the Education Level and the Economic Status

With the development of technology, there is an increasing demand for higher education and broader knowledge. Usually, the popularization of education can lead to the further development of the country. Therefore, many countries work hard on providing more opportunities for their citizens to go to college. It will be very interesting for us to know which factor is associated with citizens' education level. In my assumption, if countries' wealth level is higher, more citizens could afford and complete higher education, so there might be a linear relationship between them. I use the ratio of citizens who completed tertiary education to represent citizens' education level of a country and GDP per capita to represent the wealth level. Therefore, the regression model should be:

Population ratio with tertiary education = $\beta_0 + \beta_1 \times$ GDP per capita + random error

I gathered the sample of data of 43 countries about the population who completed tertiary education in the age group of 25-34 years old in 2019 from the International OECD Statistics(https://data.oecd.org/eduatt/population-with-tertiary-education.htm). The corresponding data about GDP per capita in 2019 is obtained from the website of The World Bank(https://data.worldbank.org/indicator/NY.GDP.PCAP.CD?end=2019&most_recent_value_desc=false&start=2019). In order to make the slope more meaningful, we divide the GDP per capita data by a thousand.

First, let's take a look at the data. Here are the summary statistics.



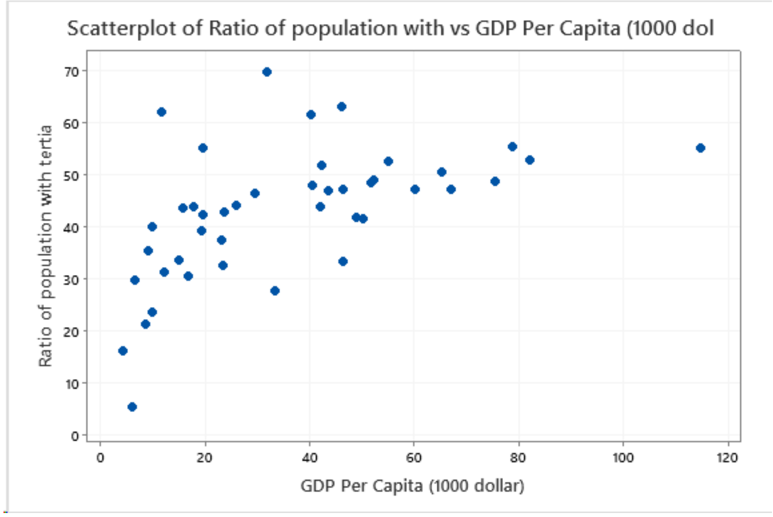
Statistics

Variable	N	N*	Mean	SE Mean	StDev	Minimum	Q1	Median	Q3
Population with tertiary educat	43	0	42.79	1.95	12.78	5.57	33.73	43.81	50.38
GDP Per Capita (1000 dollar)	43	0	35.85	3.77	24.69	4.14	15.69	31.85	50.14

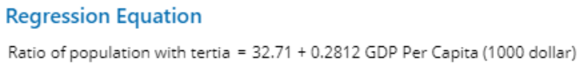
Variable	Maximum
Population with tertiary educat	69.80
GDP Per Capita (1000 dollar)	114.70

The mean ratio of the population with tertiary education is around 42.79%, with the highest ratio of 69.80% in the sample achieved by Korea. The mean GDP per capita is 35.85 thousand dollars, with the highest GDP per capita of 114.70 thousand dollars by Luxemburg.

Here is a scatter plot of the two variables:

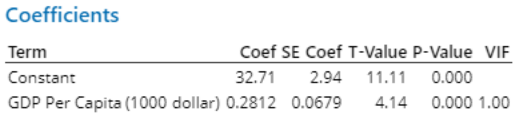


Some points follow the pattern that a higher GDP per capita leads to a higher ratio of the population with tertiary education. However, it doesn't seem to have a strong linear relationship. There is a very obvious point far away from the others in terms of its value of GDP per capita. This is Luxembourg. We use all the data point for this moment and conduct a least-square linear regression with all the data points from the sample with GPD per capita as the dependent (target) variable, and population ratio with tertiary education among 25-34 years old age group as the independent (predicting) variable:



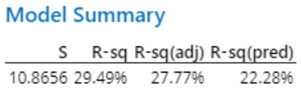
Regression Equation

Ratio of population with tertia = 32.71 + 0.2812 GDP Per Capita (1000 dollar)



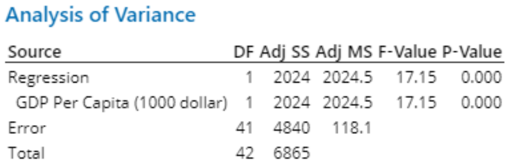
Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	32.71	2.94	11.11	0.000	
GDP Per Capita (1000 dollar)	0.2812	0.0679	4.14	0.000	1.00



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
10.8656	29.49%	27.77%	22.28%



Analysis of Variance

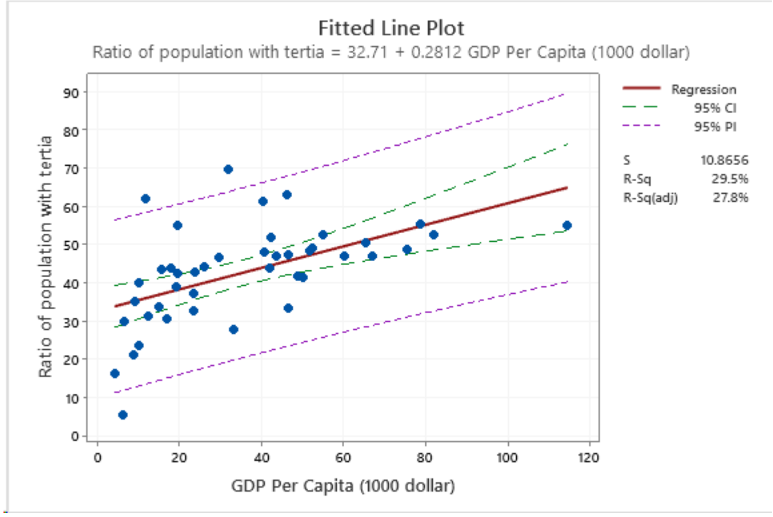
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2024	2024.5	17.15	0.000
GDP Per Capita (1000 dollar)	1	2024	2024.5	17.15	0.000
Error	41	4840	118.1		
Total	42	6865			

The R^2 is 29.49%, which means that this model accounts for 29.49% of the variability in the target variable. The F-statistic is very significant, and the p-value is very low, which means that there is extremely strong evidence that the relationship between GDP per capita and the ratio of population with tertiary education is strong. The estimated slope coefficient means that an estimated one unit increase in GDP per capita is associated with a 0.2812 percentage point increase of the ratio of the population with tertiary education in 25-34 years old group. The intercept means that with zero GDP per capita, the country will have 32.71% of the population completing tertiary education in the 25-34 years old. However, this is not meaningful, since there will be a very rare case that a country has 0 GDP per capita. The standard error of the estimate of 10.86 tells that this model could be used to predict the ratio of the population with tertiary education within ±21.72 percentage points, roughly 95% of the time.

The t-statistics for GDP per-capita tests the same thing as the F-statistics in simple linear regression, which is the null hypothesis of $\beta_1 = 0$. Since the p-value is less than 0.001, it is statistically highly significant to reject the null hypothesis, meaning that there is a strong relationship between the GDPs per capita and the ratio of the population with tertiary education. The t-statistics for the constant is also statistically highly significant, therefore, we reject the null hypothesis of $\beta_0 = 0$.

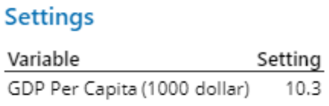
Next, we analyze the confidence intervals and prediction intervals. The confidence interval of this case represents an interval estimate for the true average ratio of the population with tertiary education for all countries that have the GDP per capita equal to a certain value. The prediction interval is used to provide an interval estimate for the population ratio with tertiary education of one particular country given GDP per capita equal to a certain value.

The following graph illustrates the confidence interval and prediction interval of our case:



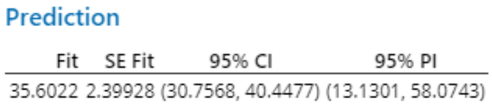
We can get several conclusions from this graph. First, we can see that the pointwise confidence interval (the inner pair of lines) is narrower than the pointwise prediction interval (the outer pair of lines) because the prediction interval accounts for both the variability in estimating the population means as well as the random variation of the individual values. Second, both intervals are narrowest near the mean value of GDP per capita equals 35.85 thousand US dollars and the population ratio with tertiary education equals 42.79 thousand US dollars. It gets wider on both sides. This indicates that we can make a rather precise prediction if the x value is near the mean, and less accurate if the x value gets extreme. Last, we can see most of the points are within the prediction interval. There are three points outside the prediction interval. From left to right, the two above the prediction interval are Russia and Korea. The one below the prediction interval is South Africa.

We can also get the prediction interval and confidence interval for a specific value of GDP per capita. We use GDP per capita equals 10.3 as an example, this is roughly the value for China in 2019.



Settings

Variable	Setting
GDP Per Capita (1000 dollar)	10.3

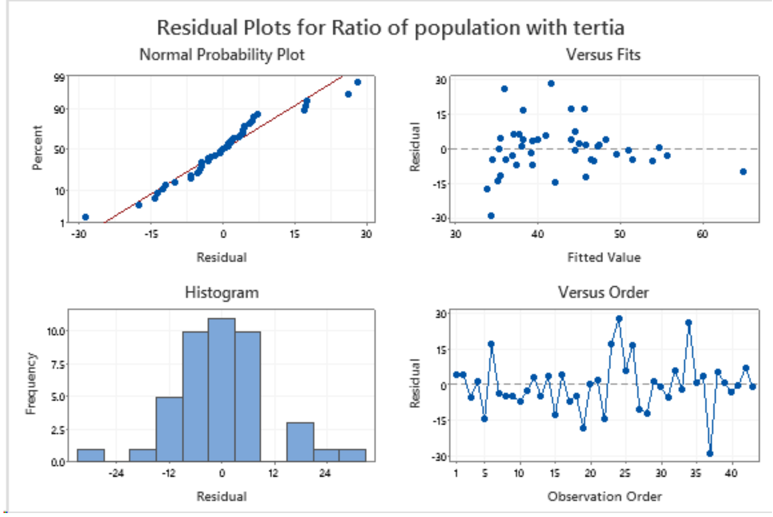


Prediction

Fit	SE Fit	95% CI	95% PI
35.6022	2.39928	(30.7568, 40.4477)	(13.1301, 58.0743)

We can see that for all countries with GDP per capita equals 10.3 thousand dollars, the average population ratio with tertiary education is (30.76, 40.44). For one particular country with GDP per capita equals 10.3 thousand dollars, the population ratio with tertiary education is (13.13, 58.07).

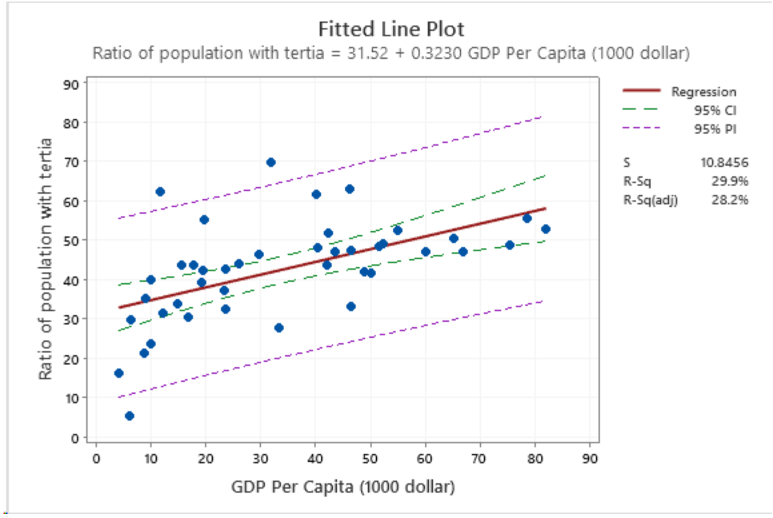
Next, we use the residual plots to check the assumption of least square regressions hold as well as see whether the points we noticed earlier are really a special case. The graph down below shows a normal probability plot of residuals, a plot of residuals versus fitted values, a histogram of the residuals and a plot of residuals versus orders.



From the graph, we can see that it violates the assumption that there is no pattern on the plot of residuals versus fitted values. There are four unusual observations in these plots. Luxemburg is a leverage point. South Africa, Russia and Korea are the outliers.

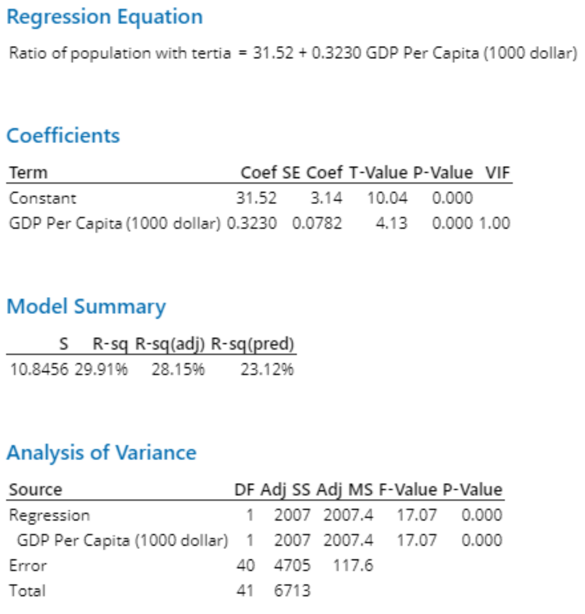
In the plot of residuals versus fitted values, one point is far right from the other point. This is Luxemburg. It has the GDP per capita as high as 114 thousand US dollars. It can have a strong effect on the slope of the fitted regression. Let’s dive down to see the situation of this special country. The reason for such a small country to have this unusually high GDP per capita is that it achieved a huge success in its economy and is a highly developed country. Its economy largely developed in banking and steel. It is the largest steelmaker in the world. There is a company called ArcelorMittal, which produces 8 percent of the world’s entire steel output. Moreover, even though its country area is smaller than Delaware and population less than Wyoming, it owns 155 bank companies. With the highly developed industry plus its smaller population, it generates the highest GDP per capita all around the world. With a highly developed social system, the government encourages the education of its citizens and therefore the country consistently spends the third largest portion of its budget on education.

Let's run the regression again with the model extracting the data point of Luxemburg. Here is the fitted line plot now:



We can see there are still three points outside the prediction interval. The two above are Korea and Russia and the one below is South Africa.

Here is the regression output:



Regression Equation

Ratio of population with tertia = 31.52 + 0.3230 GDP Per Capita (1000 dollar)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	31.52	3.14	10.04	0.000	
GDP Per Capita (1000 dollar)	0.3230	0.0782	4.13	0.000	1.00

Model Summary

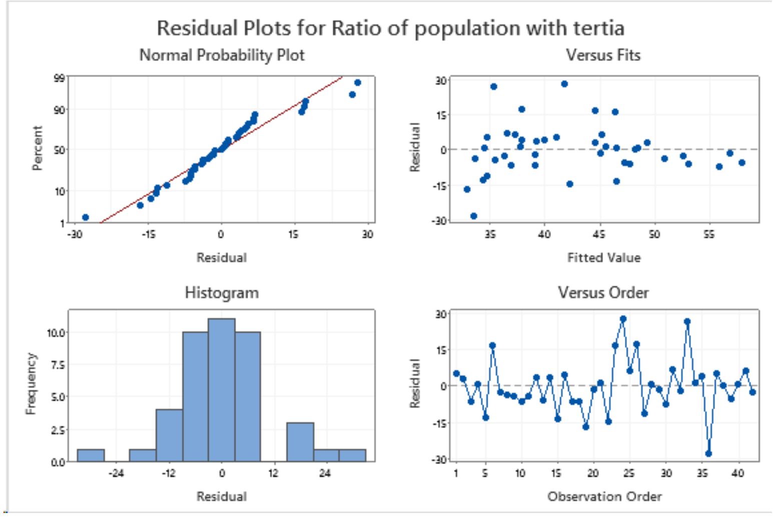
S	R-sq	R-sq(adj)	R-sq(pred)
10.8456	29.91%	28.15%	23.12%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	2007	2007.4	17.07	0.000
GDP Per Capita (1000 dollar)	1	2007	2007.4	17.07	0.000
Error	40	4705	117.6		
Total	41	6713			

The regression relationship doesn't change a lot. The R^2 increases a little bit (from 29.49% to 29.91%). The slope and the intercept don't have a statistically significant difference from the original model.

Here is the residual plot and we can see clearly that Korea, Russia and South Africa, being the outliers, are far away from other points in the plot of residuals versus fitted value:



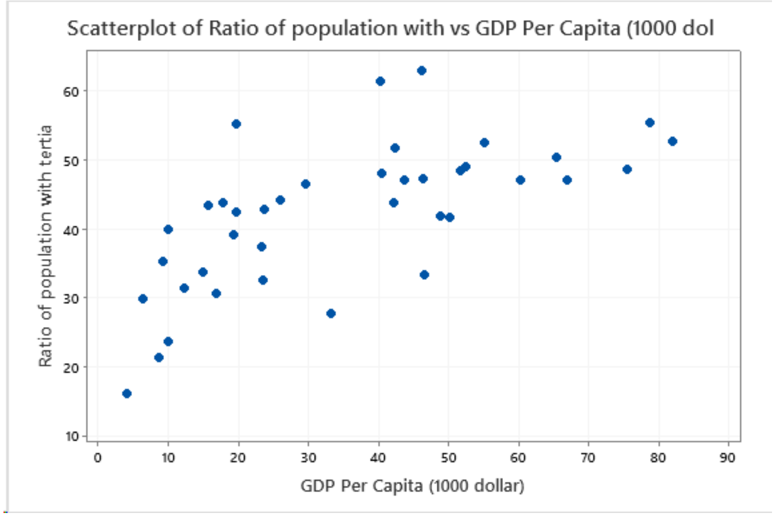
Both the original model and the model without the point Luxemburg have three outliers Korea, Russia and South Africa. Let's try to dive down to see the conditions in those countries.

Korea is a very special case. Although it doesn't achieve an extremely high GDP per capita, it is ranked the highest ratio of the population with tertiary education by the OECD in 2019. Part of the reason is that the country spent a large portion of the budget on education. In 2016, the country spent 5.4% of its GDP on all levels of education – roughly 0.4 percentage points above the OECD average. Culture also plays an important factor. Society has the an obsession with education because many families think that graduating from a top-ranked university leads to a higher socioeconomic position, promising marriage prospects, and a better career path. Russia's high ratio of the population with tertiary education is probably due to its strong investment in education in history.

For South Africa, it has a significantly low ratio given its GDP per capita. It is partly because tertiary education is not common. According to the statistics given by the OECD, in 2018, over half of 25-64 year-olds in South Africa had attained an upper secondary education as the highest level achieved. It is significantly above the G20 average of 32% and the OECD average of 38%. However, only 7% of adults have a tertiary education, the lowest among all OECD countries.

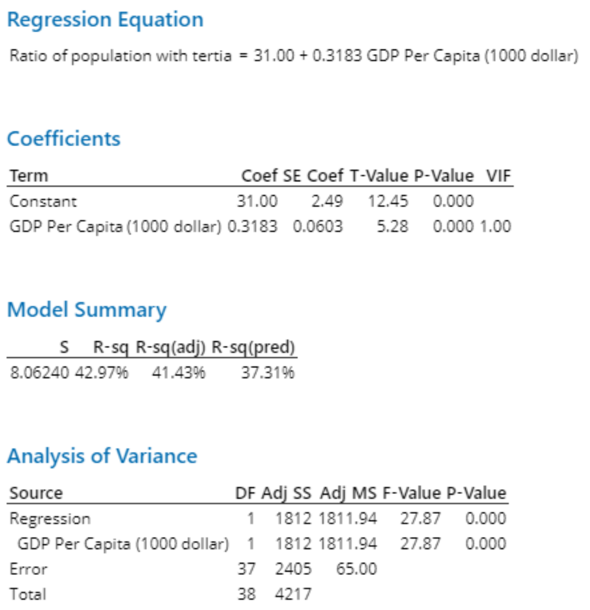
Since outliers can have a strong effect on the regression result, so we omit those outliers and run the regression again.

The scatter plot below is the model that we omit both the leverage point(Luxemburg) and the outliers(Korea, Russia and South Africa):



We can see that the pattern that with a higher GDP per capita, there is a higher ratio of the population with tertiary education, although the pattern is not strong. There are three points above that seem unusual. From left to right, it is Lithuania, Japan and Canada.

Here is the regression output:



Regression Equation

Ratio of population with tertia = 31.00 + 0.3183 GDP Per Capita (1000 dollar)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	31.00	2.49	12.45	0.000	
GDP Per Capita (1000 dollar)	0.3183	0.0603	5.28	0.000	1.00

Model Summary

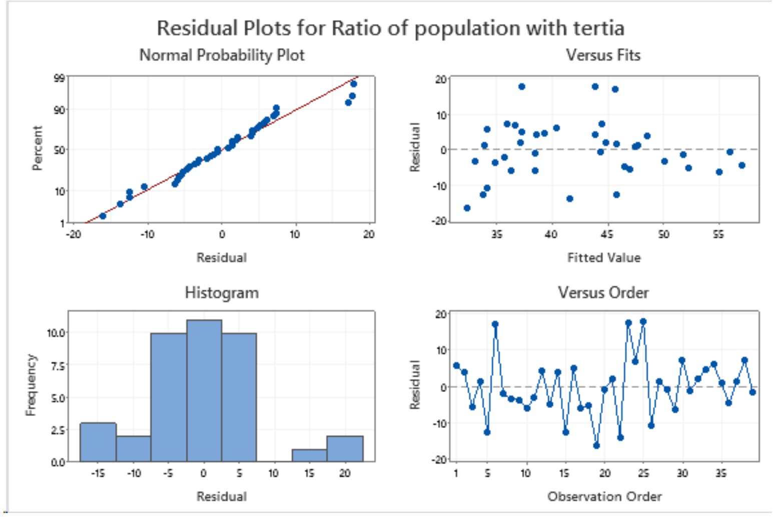
S	R-sq	R-sq(adj)	R-sq(pred)
8.06240	42.97%	41.43%	37.31%

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1812	1811.94	27.87	0.000
GDP Per Capita (1000 dollar)	1	1812	1811.94	27.87	0.000
Error	37	2405	65.00		
Total	38	4217			

The R^2 increases by over 10%. Given P-Value equals zero, we see that the regression relationship is still very strong. Both the intercept and the slope coefficient don't change significantly from the original model. The standard error of the estimate decrease by roughly two percentage points.

Here is the residual plot:



Since the linear pattern between GDP per capita is not significant. New outliers will keep emerging if we keep extracting points from the model. So we will stop here for the final conclusion.

When I made the assumption, I was hoping for a stronger relationship. Since logically speaking, the stronger the country's economy is, the more likely it that people can afford higher education. However, from the statistics shown in the report, R^2 only accounts for around 30% of the variability in the target variable in the original model. Moreover, another limitation of this model is that its prediction interval is nearly 40%. This is nearly half of the range of the entire data. Therefore, although the model is statistically significant, it is not practically important. Perhaps, one single variable is not enough to predict this target variable. Indeed, from the case of Korea and Luxemburg we've talked about before, the influence factor of the population ratio with tertiary education may depend on the importance that the society places on education and government policy of education subsidy. In this case, multiple linear regression might construct a better model.