Draghicescu Alexandra-Elena
No etud. 22218574

Academic year 2022/2023
L3, IEM
Statistics for Economics and Business
Student project

Human capital and economic growth

- **Background**

King Arthur and the knights of the round table searched for the holy grail. Long after the mystic island Avalon immersed in the floods, Solow and the neo-classical economists entered the quest for long-run economic growth. If we only understood the determinants of long-run economic growth — the story goes — we could implement policies to foster wealth across all nations. Your task it to contribute to that noble endeavor.

Assume wealth is a social product, i.e. the outcome of the joint effort of all individuals in a society. How could we increase wealth? Well, probably the simplest way would be to increase the society — double the number of people and everything else, and you double the output. But that doesn't make anybody richer because the doubled output is going to serve twice as many people.

Thus, in order to grow the wealth of each individual in the society, something needs to grow that helps the members of that society to achieve their goals. The general term for such athing in economics is capital. At the beginning the focus was very much on physical capital,i.e. installed machinery, and how these improve over time due to technical progress. Only later,human capital as an important productive factor has been emphasized, with schooling being one important input to human capital.

- **Task**

If we would increase years of schooling by one year, how would that affect economic growth?

- The data set provided for the analysis is panel data compiled from several sources: the schooling attainment data is from (Barro and Lee, 2013), henceforth BL, and economic data from the Penn World Tables (Feenstra et al., 2013), henceforth PWT. The data set covers 146 countries over 14 five-year periods and provides the following variables:

Variable name	Description	Source
countrycode	iso country code	PWT
year	year of measurement	PWT/BL
country	name of country	PWT/BL
region_code	some 'world-region'	BL
rgdpo	Output-side real GDP at chained PPPs (in mil. 2017US$)	PWT
emp	Number of persons engaged (in millions)	PWT
pop	Population (in millions)	PWT
cn	Capital stock at current PPPs (in mil. 2017US$) at current PPPs (in mil. 2017US$)	PWT
yr_sch	average years of schooling for 15 to 99 years	BL

- **Results**

To answer this question I was thinking of two approaches of the issue.

1) In a general way,there might be a DIRECT relation between education and Economic Growth of a country -more precisely number of years of schooling and GDP.

2) On the other hand,we can consider an INTERMEDIATE between education and economic growth and that is the *Employment*-we can discuss about the influence of employment on the GDP growth as well as having a *multiple regression*.

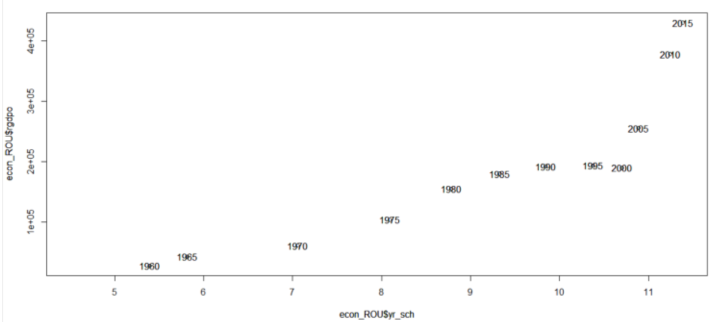
I. Let's now take the 1st approach and consider a DIRECT relation between Education and Economic Growth.

1st thing that comes in mind is to plot the data and view the relationship. Now,without having any idea of how the image is going to be,we might think that there might be a *direct positive* relationship between years of schooling and GDP growth.

For this step of our analysis,we use the variables ***yr_sch*** and ***rgdpo*** from the dataset.

As there are so many values(observations),let's focus on one specific country to have a better view on the plot.I will take for example Romania,my nattive country.

- *1st step is to get data specifically only from that country,by creating a subset of the dataset.*



$$\boldsymbol{rgdpo = \beta_0 + \beta_1 * yr_{sch} + \varepsilon}$$

$\beta_0 = intercept\ for\ base\ year(average\ GDP\ of\ Romania\ in\ 1950)$
$\beta_1 = coefficient\ of\ independent\ variable(effect\ of\ education\ on\ Romania's\ GDP)$
$\varepsilon = random\ error\ term(captures\ the\ effect\ of\ external\ factors)$

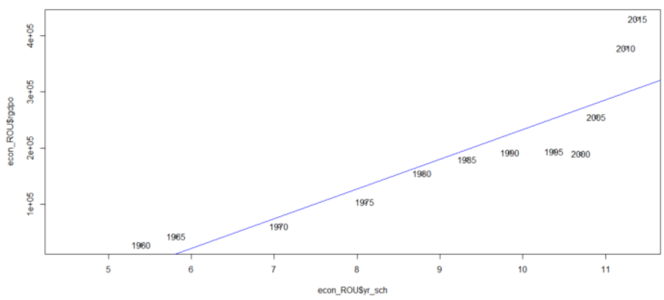
```
Call:
lm(formula = rgdpo ~ yr_sch, data = econ_ROU)

Residuals:
   Min     1Q Median     3Q    Max
-80262 -29145 -16737  32584 123427

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -294295      81900  -3.593 0.004902 **
yr_sch         52751       8815   5.984 0.000135 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60850 on 10 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.7817,    Adjusted R-squared:  0.7599
F-statistic: 35.81 on 1 and 10 DF,  p-value: 0.0001349
```

We can interpret the regression results as being *statistically significant*,as the *p-value* approaches to 0. (see the signif.codes section)

On average,the ***GDP of Romania*** has a negative value arround -294295 mil USD(=β_0).

Further ***education*** has a positive effect on ***GDP***,so with every ***year of schooling*** added,the ***GDP*** level is growing with 52751 mil USD(=β_1).

- *Let's observe the effect of β_1(slope of the regression).*



The observations are close to the line-the tendancy is to follow it-but not exactly pointing on it. We might interpret the results as follows:
- *Adjusted R-squared*: 0.7599 means that almost 76% of the variation of the output is explained by the input.(In other words,76% of the Real ***GDP of Romania*** can vary on the years of schooling, for the period between 1960-2015).
-The value of 0.7599 has another meaning as well-it is greater than 0 and close to 1,so we can interpret the result as an existing *strong positive relation* between years of schooling and GDP.

- *We also may think about the effect of education over years.*

In this case,we introduce *year dummies*(variables that can take 0 or 1 values) to capture the effect of ***education*** over time and also be capable to establish time trends,based on the results.Also,creating a dummy for the country ROU as we are going to use the ECON dataset.

- *Creating a new regression model:*

$$\boldsymbol{rgdpo = \beta_0 + \beta_1 * yr_{sch} + \beta_2 * ROU + \delta_j * years\_i + \varepsilon}$$

$\beta_0 = intercept\ for\ base\ year(average\ GDP\ in\ 1950)$
$\beta_1 = coefficient\ of\ independent\ variable(effect\ of\ education\ on\ GDP)$
$\beta_2 = coefficient\ of\ the\ country\ dummy(being\ Romania\ or\ not)$
$\beta_0 + \beta_2 = intercept\ for\ base\ year\ in\ Romania(average\ GDP\ in\ Romania\ in\ 1950)$
$\delta_j = coefficient\ of\ the\ time\ dummy(GDP's\ difference\ between\ base\ year\ and\ year\ i)$
$\beta_0 + \delta_j = intercept\ for\ the\ year\ i(average\ GDP\ of\ a\ random\ country\ in\ year\ i\ )$
$\beta_0 + \beta_2 + \delta_j = intercept\ of\ the\ year\ i\ for\ Romania(average\ GDP\ of\ Romania\ in\ year\ i)$
$\varepsilon = random\ error\ term(captures\ the\ effect\ of\ external\ factors)$

$$years\_i = \begin{cases} 0\,; year! = i \\ 1\,; year == i \end{cases} \quad ;\ \mathrm{i} = \overline{1955,2015} \quad ; j = \overline{1,13}\ *$$

$$ROU = \begin{cases} 0\,;\ country\_code! = "ROU" \\ 1\,;\ country\_code == "ROU" \end{cases}$$

*(each j corresponding to one year i we have pairs as follows {(i,j)}={(1955,1),(1960,2)…(2015,13)}

```
lm(formula = rgdpo ~ yr_sch + ROU + years_, data = econ)

Residuals:
     Min       1Q   Median       3Q      Max 
 -809334  -286491  -118464    28267 17309656 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -114567     126171  -0.908  0.36401    
yr_sch         46134       8031   5.744 1.11e-08 ***
ROU          -237847     248188  -0.958  0.33805    
years_1955     18418     161850   0.114  0.90941    
years_1960     38029     150904   0.252  0.80107    
years_1965     45664     150013   0.304  0.76087    
years_1970     39266     145862   0.269  0.78781    
years_1975     47091     145963   0.323  0.74703    
years_1980     58439     146214   0.400  0.68945    
years_1985     47020     146591   0.321  0.74844    
years_1990     57840     144755   0.400  0.68953    
years_1995     78443     145465   0.539  0.58979    
years_2000    114310     146181   0.782  0.43435    
years_2005    188551     147032   1.282  0.19991    
years_2010    314445     147929   2.126  0.03369 *  
years_2015    388838     148816   2.613  0.00907 ** 
---
Signif. codes:  
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 852600 on 1510 degrees of freedom
  (518 observations deleted due to missingness)
Multiple R-squared:  0.06962,	Adjusted R-squared:  0.06037 
F-statistic: 7.532 on 15 and 1510 DF,  p-value: < 2.2e-16
```

Years of schooling still have a *high statistical significance* on the regression model.(***)

According to the result of the regression,the average ***GDP*** value for Romania in the base year is equal to -352414 mil USD $(=\beta_0 + \beta_2)$.

We can observe a sharp increse in the country's ***GDP*** in the 2010's years period.On average, ***GDP of Romania*** was higher with 388838 mil USD(=δ_{13}) in 2015 ,compared to the base year 1950 when it was equal to -352414 mil USD .

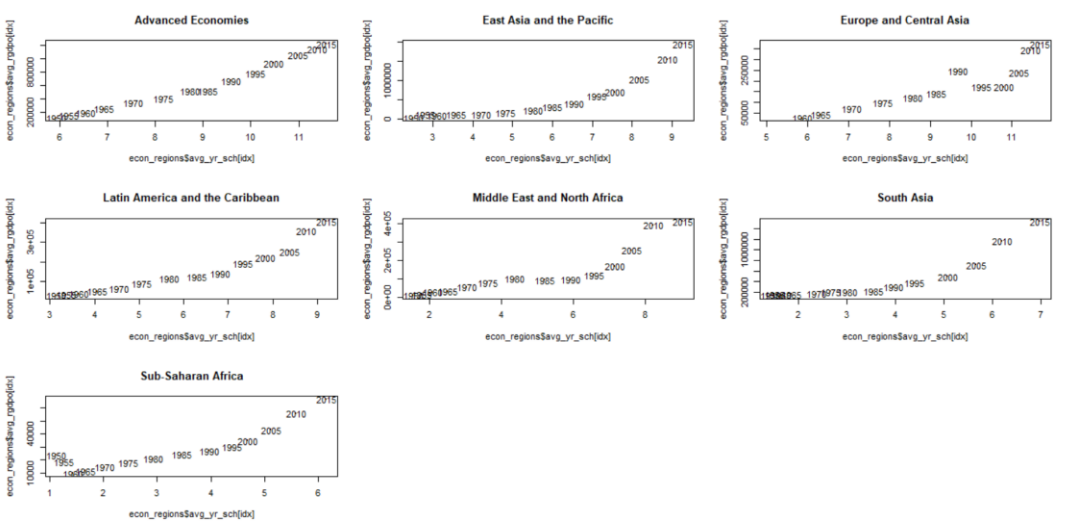
For example, with 2 ***years of schooling*** added,the average values of ***GDP in Romania*** in latest period is arround the value of 128692 mil USD(=$\beta_0 + 2 * \beta_1 + \beta_2 + \delta_{13}$) compared to the value of 82558 mil USD(=$\beta_0 + \beta_1 + \beta_2 + \delta_{13}$) which was obtained by adding only one year of further schooling.The difference between these 2 values is exactly the coefficient of yr_sch β_1 . As written in the summary,it is *statistically significant* and has a *positive effect* on the ***GDP*** of the country over years.

By doing the *Linear Hypothesis Test* in R Studio,as *F statistic* value is arround 8.05 and *p-value* corresponding is very close to 0,we can conclude that even thougth the year dummies are not statistically significant,they are *jointly significant*.In this case,we reject the null hypothesis with the coefficients being 0,as their values have a practical meaning.

- *2nd step that comes in mind is to have a broaden view over the world.*

Now,we can consider to plot these two variables for all the 7 regions from the dataset In this case,we might consider to plot the average values of ***yr_sch*** and ***rgdpo*** .

```
dev.new()
par(mfrow=c(3,3))
regions <- sort(unique(econ_regions$region_code))
for(region in regions){
  idx <- econ_regions$region_code==region
  plot(econ_regions$avg_yr_sch[idx], econ_regions$avg_rgdpo[idx],
       cex=0.3,main=region)
  text(econ_regions$avg_yr_sch[idx], econ_regions$avg_rgdpo[idx], econ_regions$year[idx])
}
```



As we can observe the figures,there is a tendancy of *positive linear relationship* ***between years of schooling*** and the ***real GDP***,not depending the region of the world.

- *Introducing a new regression model,with location dummies for each region,in order to conclude the evolution of GDP for each out of the seven regions of the world.*

$$\boldsymbol{rgdpo = \beta_0 + \beta_1 * yr_{sch} + \alpha_k * r\_x + \delta_j * years\_i + \varepsilon}$$

$\beta_0 = intercept\ for\ base\ year\ \&\ base\ region(average\ GDP\ in\ 1950\ in\ Advanced\ Economies)$
$\beta_1 = coefficient\ of\ independent\ variable(effect\ of\ education\ on\ GDP)$
$\alpha_k = coefficient\ of\ region\ dummy(GDP'S\ difference\ between\ base\ region\ and\ region\ x\ )$
$\delta_j = coefficient\ of\ the\ time\ dummy(GDP's\ difference\ between\ base\ year\ and\ year\ i)$
$\beta_0 + \delta_j = intercept\ for\ the\ year\ i(average\ GDP\ for\ Advanced\ Economic\ Region\ in\ year\ i\ )$
$\beta_0 + \alpha_k + \delta_j = intercept\ of\ the\ year\ i\ for\ region\ x(average\ GDP\ of\ region\ x\ in\ year\ i)$
$\varepsilon = random\ error\ term(captures\ the\ effect\ of\ external\ factors)$
$i = \overline{1955,2015}\ ; j = \overline{1,13}\ *$
$k = \overline{1,6}\ ;\ x = \{all\ names\ regions! = "Advanced\ Economies\}\ **$
*(each j corresponding to one year i -being pairs as follows {(i,j)}={(1955,1),(1960,2)…(2015,13)}
**(each k corresponding to one region x–being pairs as follows {(k,x)}={(1,East Asia and Pacific)…}

```
Call:
lm(formula = rgdpo ~ yr_sch + r_ + years_, data = econ)

Residuals:
     Min       1Q   Median       3Q      Max
-1029195  -334248   -79588   112276 16962484

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                        235264     141202   1.666 0.095890 .
yr_sch                              11712      11800   0.993 0.321078
r_East Asia and the Pacific         -1601      85856  -0.019 0.985121
r_Europe and Central Asia         -530059      85681  -6.186 7.92e-10 ***
r_Latin America and the Caribbean -423950      75114  -5.644 1.98e-08 ***
r_Middle East and North Africa    -449221      85470  -5.256 1.68e-07 ***
r_South Asia                      -107658     120127  -0.896 0.370292
r_Sub-Saharan Africa              -542992      91341  -5.945 3.44e-09 ***
years_1955                          40710     157172   0.259 0.795658
years_1960                         101072     147037   0.687 0.491940
years_1965                         120499     146497   0.823 0.410905
years_1970                         141192     143149   0.986 0.324130
years_1975                         167311     143911   1.163 0.245176
years_1980                         198493     145039   1.369 0.171346
years_1985                         206111     146411   1.408 0.159409
years_1990                         275191     145907   1.886 0.059478 .
years_1995                         314164     147771   2.126 0.033664 *
years_2000                         365997     149588   2.447 0.014531 *
years_2005                         457030     151688   3.013 0.002630 **
years_2010                         598813     153847   3.892 0.000104 ***
years_2015                         687598     155942   4.409 1.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 827200 on 1505 degrees of freedom
  (518 observations deleted due to missingness)
Multiple R-squared:  0.1272,    Adjusted R-squared:  0.1156
F-statistic: 10.96 on 20 and 1505 DF,  p-value: < 2.2e-16
```

In our model,intercept value of 235264 mil USD(=β_0) stands for the ***GDP*** of countries from Advanced Economy region in the base year of 1950.

For example,in the base year,the ***GDP*** level of a country in Sud Saharan Africa region was with 542992 mil USD(=α_6) less than the level of a country in Advanced Economy.

In this example,we can conclude that being from a specific region there might be a particular *influence* of the level of ***GDP***-the more developed region of the world,the greater Economic Growth (see Global North and Global South Economic division).

Although,it doesn't have a statistical significance in this regression model,the value of the years of schooling still has a *positive coefficient*,which may help in increasing the ***GDP*** of a country.

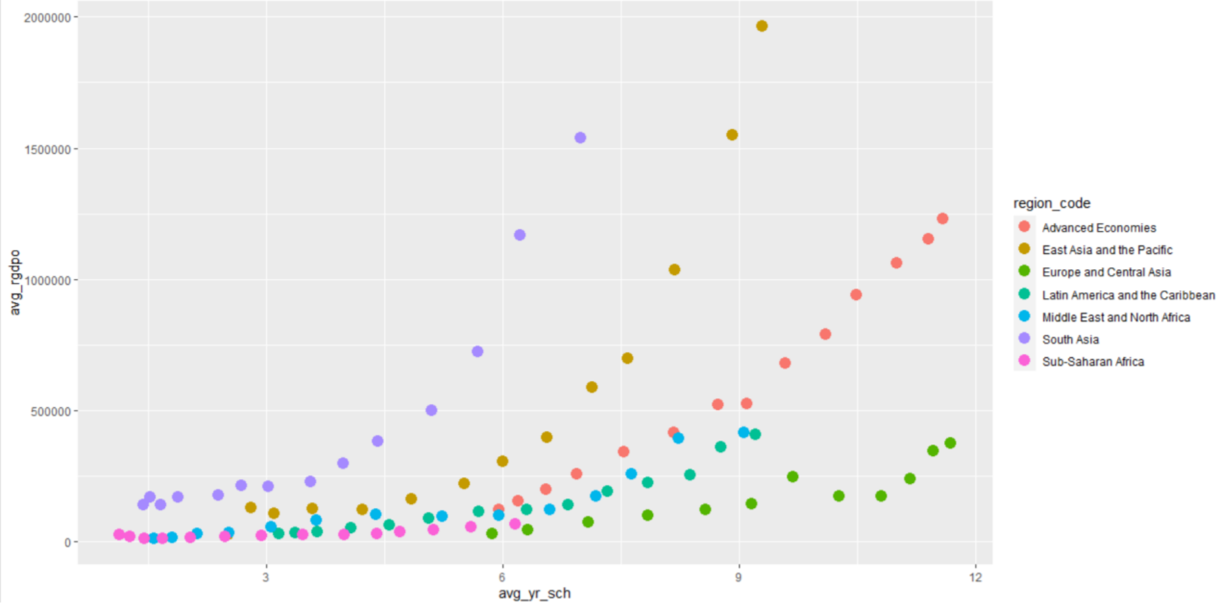
Here, regions variable is *more meaningful* for the result-as checking the *signif. codes symbols*.

Now we might ask the following question:

Why the influence of the region is more significant than the years of schooling?

- *Getting an answer by calculating the average years or schooling and GDP of each world region and then have one single plot to observe the effect of the location.*

```
library(ggplot2) #creating the plot
ggplot(data = econ_regions, aes( avg_yr_sch,avg_rgdpo, colour = region_code)) +
  geom_point(size=3)
```

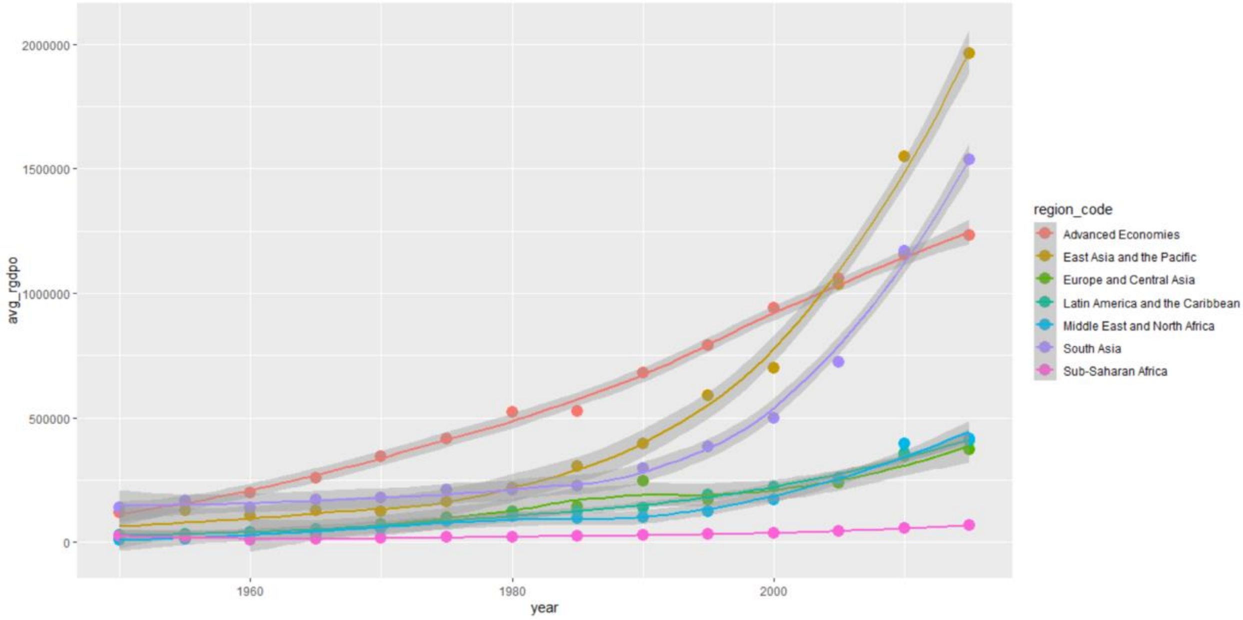


As we can observe on the plot,for the poor regions the ***average of school years*** is between 6 and 9 while in advanced regions the average is over 9 ***years of schooling***-that means when taking into account the *multiple regression* of years of schooling and the origin,the effect of education might have the *same significance level* on the ***GDP***.

So,there are other factors that influence the Economic Growth of a country,apart from Education.

- *Observe the Economic evolution over time and think about what other factors could influence it.*

```
library(ggplot2)
ggplot(data = econ_regions, aes( year,avg_rgdpo, colour = region_code)) +
  geom_point(size=3) + geom_smooth()
```



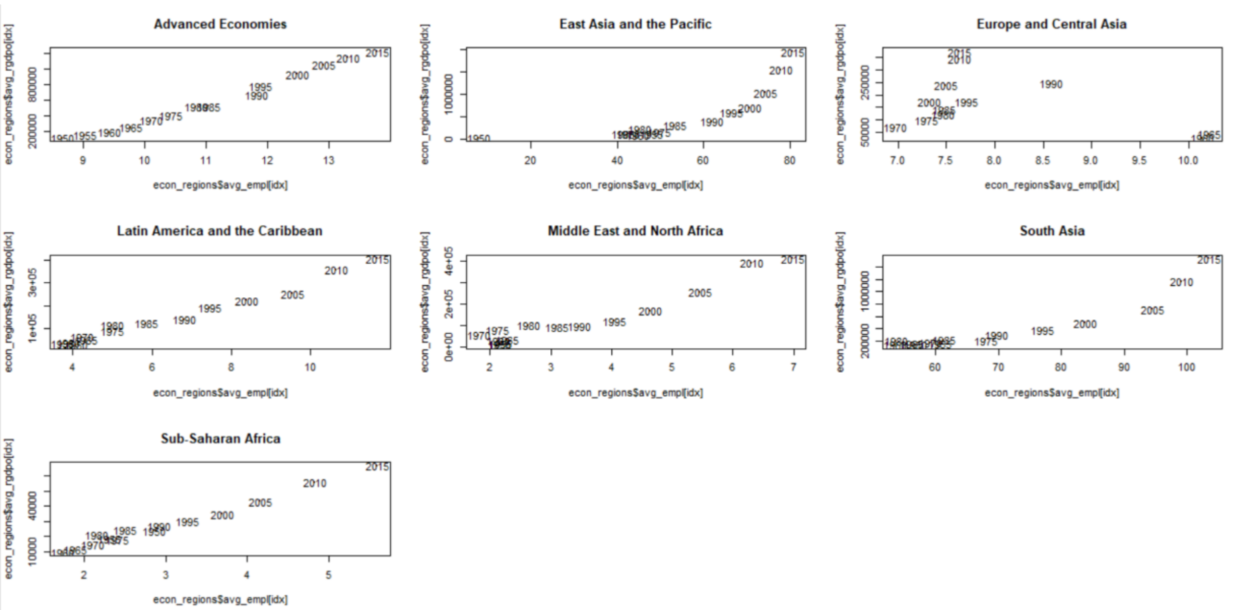
After seeing this plot and the *trendline*,we may observe a great difference between the regions and the ***evolution of GDP*** over the years.So,we may ask if there are other reasons/factors that contribute to this gap?!

II. To answer the previous question,I will introduce the 2nd approach of this report and consider an INTERMEDIATE element linking education and Economic Growth.

In the 2nd approach,we take into consideration the ***employment*** as an intermediate factor to describe the influence of ***education*** on ***Economic Growth*** of a country.

As people get ***further education*** they get also professional skills and the opportunity to perform them on the working market,so they may apply for a job and finally get the status of "employed person". Further,when talking about the Economic situation of a country,we take into consideration the ***factor of employment***. When making predictions-as the level of ***employment*** is higher, the ***GDP*** is higher and so the Economic Growth.

- *General view on the relation between the average of the employment level and gdp and plot the data. (same as done for yr_sch)*



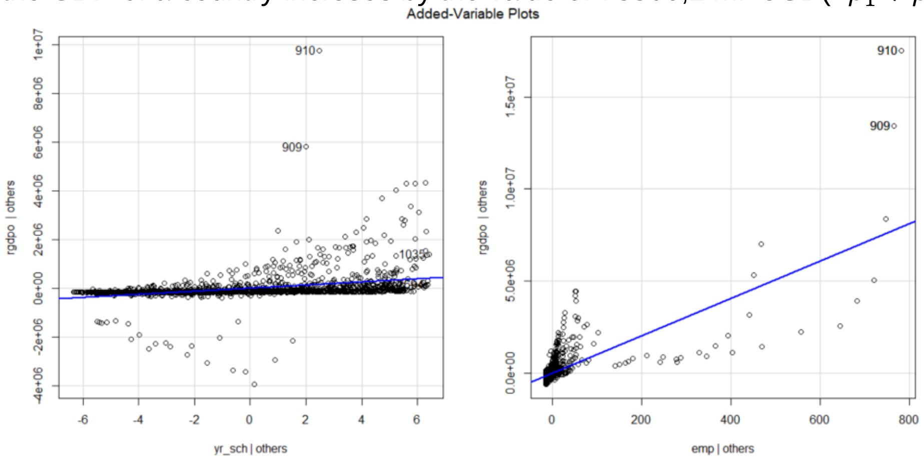
- *multiple linear regression between education and employment on the gdp and plot*

$$\boldsymbol{rgdpo = \beta_0 + \beta_1 * yr_{sch} + \beta_2 * emp + \varepsilon}$$

```
Call:
lm(formula = rgdpo ~ yr_sch + emp, data = econ)

Residuals:
     Min       1Q   Median       3Q      Max
-3974931  -206533   -53416   107364  9624516

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -268024.8    35111.7  -7.633 4.12e-14 ***
yr_sch        63679.2     4766.4  13.360  < 2e-16 ***
emp           10127.0      245.6  41.231  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 594500 on 1450 degrees of freedom
  (591 observations deleted due to missingness)
Multiple R-squared:  0.5633,	Adjusted R-squared:  0.5627
F-statistic: 935.2 on 2 and 1450 DF,  p-value: < 2.2e-16
```

As expected,both years of schooling and employment are *statistically significant* for our model. In other words,the more ***years of schooling*** and higher ***employment***,the higher ***Economic Growth*** of a country.

For every *year of schooling* added(***further education***) and with every 1 mil of people *employed*,the ***GDP*** of a country increses by the value of 73806,2 mil USD(=$\beta_1 + \beta_2$).



We can see the *trend of graphs* which shows as being *correlated*-the tendency is that both ***education*** and ***employment*** has a *significant influence* on a country's ***GDP***.

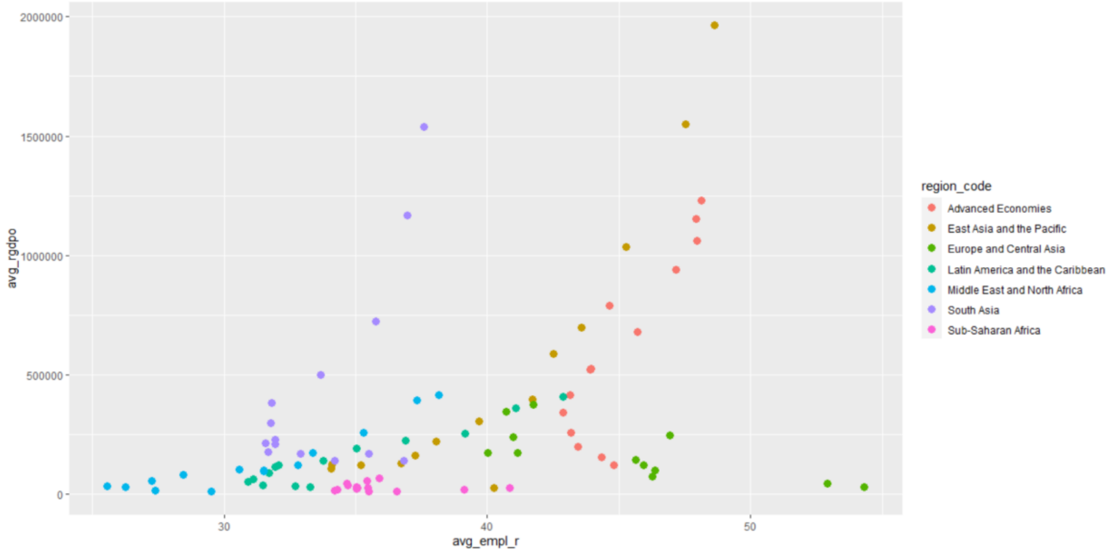
There are also *outsiders*(extreme values) from the 909th and 910th observations. More exactly we are talking about China being one of the strongest and largest Economy of the world-so it is quite normal and understandable to have such a big values.

909	CHN	2010	China	East Asia and the Pacific	13847619.000	781.0355225	1368.810615	51907732.000	8.250
910	CHN	2015	China	East Asia and the Pacific	17985756.000	797.3352661	1406.847870	75872464.000	8.710

- *Designing plots to see the effect of employment on gdp,depending on the region.*

But,to talk about the *real significance* of ***employment*** on the ***GDP*** of a country,we might think about the ***employment rate***(*ratio of employed people out of the total population*).

```
library(ggplot2)
ggplot(data = econ_regions, aes( avg_empl_r,avg_rgdpo, colour = region_code)) + geom_point(size=3)
```

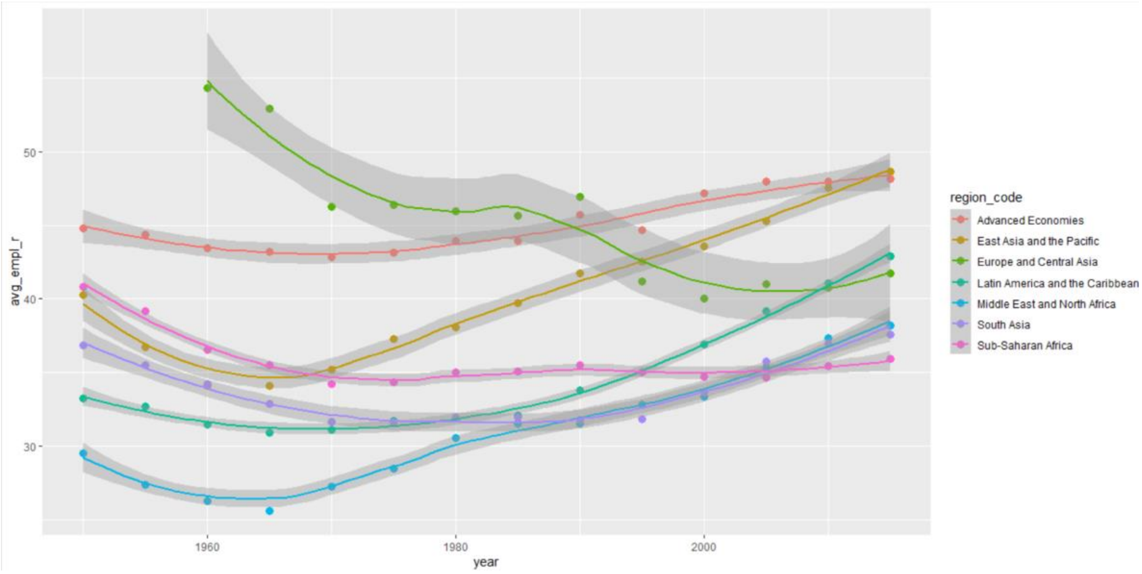


As expected,the higher ***employment rate***,the greater ***Economic Growth*** of a region. As previously said,depending on the region,*the effect is not the same*.

For instance,China being part of the East Asia and The Pacific region,we may observe the extreme yellow point showing a *very strong positive relation* between the average ***employment*** and the ***GDP***. At the same time,countries from not a developed region,have a *lower average* of ***employment rate*** as well as a lower ***GDP*** compared to advanced economies countries.

- *We can also plot the evolution of regions' employment rate over years and conclue their statistical significance for a potential Economic Growth.*

```
ggplot(data = econ_regions, aes( year,avg_empl_r, colour = region_code)) +
  geom_point(size=3) + geom_smooth()
```



As time passes,most of the regions have a quite *constant* ***rate of employment***,there is *not big fluctuations* in the graph. We can admit that there is a *sharp decrease* in the ***employment level*** for the Europe and Central Asia(one potential reason might be the *demographic's aging population*) and we can also see a *sharp increase* for East Asia and Pacific as well as for the Middle East regions(one possible reason might be the *new technologies* that help the indutries be more productive).

- **Conclusions**

To conclude this report,*depending on the region and on the period*,a country's ***Economic Growth*** might be directly influenced by the ***level of education*** of its people(to be more specific,we are talking about number of ***years of schooling***).

These statistics(*regression analysis*) are just *experimentals* and just give us an idea of how events from real life are actually going(here talking about the influence of ***education*** on a country's ***Economic Growth***).

There are also other factors that might be taken into consideration when talking about a country's ***Economic Growth***. For instance,in this report we have also taken into account the ***level of employment***.

In addition,we may also think about a country's or a region's natural resources.As higher and wide range of *resources* a country has,more branches of industry and finally a *positive net balance*(exports-imports) which make think us about a positive ***Economic Growth***.

Let's not forget about *inflation* and *interest rates* and the most known analysis such as *Philips Curve* and *Okun's Law*-that are deep into the analysis of ***Economic Growth***.

To sum up,as initially considered,there is a *positive relation* between ***education*** and the ***Economic Growth*** of a country.

- **References**

1. *United Nations, 2015. World Population Prospects: The 2015 Revision, DVD Edition. United Nations, Department of Economic and Social Affairs, Population Division.*

2. *Barro, R., Lee, J.-W., 2013. A new data set of educational attainment in the world, 1950-2010. Journal of Development Economics 104, 184-198.*

3. *Feenstra, R. C., Inklaar, R., Timmer, M., 2013. The next generation of the penn world table. Tech. Rep. 19255, NBER.*

4. *Data science în mediul R : teorie și aplicații / Stelian Stancu. – Bucureşti : Ed. ASE, 2020*

5. *Statistics for Business and Economics, by Newbold, P., Carlson, W.L., Thorne, B., Pearson.*

6. *Introductory Econometrics: A Modern Approach (7e) by Wooldridge (2020)*