Final Project R Outline

1. Determine an economic question of interest.
    a. Is spending on public education correlated with a higher GDP growth in the US?
2. Identify existing data sets that can be used to answer the question.
    a. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/SE.XPD.TOTL.GD.ZS
        i. Government expenditure on Public Education (as a % of GDP)
    b. https://databank.worldbank.org/metadataglossary/world-development-indicators/series/NY.GDP.MKTP.KD.ZG
        i. Annual GDP Growth (%)
    c. https://databank.worldbank.org/source/world-development-indicators/Series/NY.GDP.MKTP.KD.ZG
        i. Federal Direct Investment (net inflow as % of GDP)
3. Design a regression analysis to answer the question
    a. GDP-Growth = EDU-Expenditures + FDI-Investments
4. Write out the implied model used, and any assumptions made.
    a. GDP-Growth = EDU-Expenditures + FDI-Investments
        i. Attempted to control for Year
        ii. Attempted to lag variables (account for time series/does not assume GDP-growth is independent)
5. Create a data summary table, a results summary table, and graphs to explain the data (and/or results).

Chris Kioschos
Professor Gebben
Comp. Econ.
5 November 2023

Public Education Spending and GDP-Growth

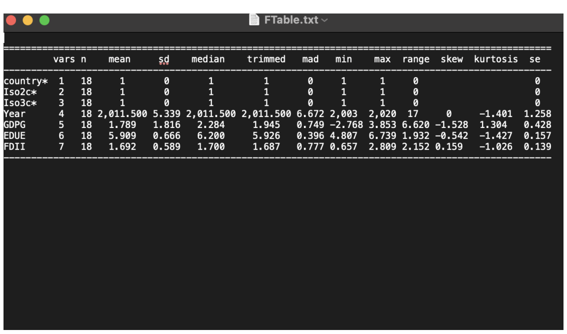
Introduction

How beneficial is public education to an economy? Additionally, are there diminishing returns to investing in said education? These were the questions that ultimately inspired my regression of GDP-Growth with investments into Public Education. I also chose Foreign Direct Investment inflows as an additional independent variable. FDI investments were chosen due to their supposed effect on GDP-Growth and for comparison with education. Ultimately, by teasing out the correlation between public education and GDP-Growth, we can determine whether future investments (into public education) are worth it for an economy's future growth. Lastly, I chose this model because I thought the analysis would be simple and the results meaningful.

Data

To begin my analysis, I used WDI tools to download and combine the three data-sets {WDI("USA",indicator=c("NY.GDP.MKTP.KD.ZG","SE.XPD.TOTL.GD.ZS","BX.KLT.DINV.WD.GD.ZS"),2003, 2020)}. I then renamed the columns and also cleaned up the data by setting any NA values to zero {colnames(OG) <-
c("Country","iso2c","iso3c","Year","GDP-Growth","EDU-Exp","FDI-Invest")}.

I also saved and named this data as a csv under the name 'FOG'. Additionally, to account for my data being a timeseries, I attempted to lag my three variables by five years. After this, I created a function that I could plug different time lags into. I then saved this dataset as a csv, and named it 'LOG'. Although the operation was successful, NA values were also introduced into my data. This was likely due to a class error, character rather than numeric, but I was unable to find the solution. Ultimately, I decided to regress my original data-set due to its lack of NA values. Despite not fully accounting for all the problems of regressing a time series, I decided that this was still better than using data that was corrupted. Towards the end of my paper, in the analysis section, I will discuss the issues with regressing a time series that my model failed to account for. I then created a summary table using my original data, and saved it as a text file under the name ‘FTable’. The mean, median, maximum, and minimum can all be found with this table. It is also important to note that all of my tables were saved as text files using stargazer. After this table was saved, I then performed a multiple linear regression. To accomplish this, I created a mock formula for the regression by setting GDP-Growth = EDU-Expenditures + FDI-Investments. Using my original data and this created formula, I then teased out important information by creating several different files and tables. The first file, named ‘REG.txt’, is a text file containing a table of the regressed data. The second file, named ‘MOD.txt’, contains additional details for the data that was regressed. The third file, named ‘csum.txt’ contains the summary of the coefficients for the regressions analysis. Finally, the last file is named ‘Sum’ and contains an additional summary of the regression although not as a table. The data it lists is included as a screenshot on the previous page (below the plots). Finally, I created four different scatter plots from the regressed data. Using the rds file ‘MOD’, I plotted the regression and then saved each of the images as PNG’s.



FTable.txt

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	18	1	0	1	1	0	1	1	0			0
Iso2c*	2	18	1	0	1	1	0	1	1	0			0
Iso3c*	3	18	1	0	1	1	0	1	1	0			0
Year	4	18	2,011.500	5.339	2,011.500	2,011.500	6.672	2,003	2,020	17	0	−1.401	1.258
GDPG	5	18	1.789	1.816	2.284	1.945	0.749	−2.768	3.853	6.620	−1.528	1.304	0.428
EDUE	6	18	5.909	0.666	6.200	5.926	0.396	4.807	6.739	1.932	−0.542	−1.427	0.157
FDII	7	18	1.692	0.589	1.700	1.687	0.777	0.657	2.809	2.152	0.159	−1.026	0.139

Model

I created my model by combining all three of my datasets, lagging the variables, and by performing a multiple regression of said data. Ideally, by regressing expenditures on education and FDI investments with GDP-Growth, I should be able to find the effect of education on GDP. Ultimately, by comparing the correlated effects on GDP-Growth, we can tease out the effect of spending on public education by comparing it to the effect FDI has. Below can be seen different summaries of the regression, with each comparing education and foreign investments with GDP-Growth.

Regression Code

{MOD <- lm(GDPG ~ EDUE + FDII, FOG)}

E-Table Code (Regression)(Controlled for Year)

REG_F <- feols(log(GDPG)~EDUE+FDII|Year,FOG)

etable(REG_BASIC,REG_FDI)

SUM	list [11] (S3: summary.lm)	List of length 11
call	language	lm(formula = GDPG ~ EDUE + FDII, data = FOG)
terms	formula	GDPG ~ EDUE + FDII
residuals	double [18]	1.659 2.072 2.259 0.630 -0.332 -2.114 ...
coefficients	double [3 x 4]	1.316 -0.168 0.868 4.473 0.683 0.771 0.294 -0.247 1.125 0.773 0.808 0. ...
aliased	logical [3]	FALSE FALSE FALSE
sigma	double [1]	1.84456
df	integer [3]	3 15 3
r.squared	double [1]	0.08922287
adj.r.squared	double [1]	-0.03221408
fstatistic	double [3]	0.735 2.000 15.000
cov.unscaled	double [3 x 3]	5.8797 -0.8551 -0.4558 -0.8551 0.1369 0.0271 -0.4558 0.0271 0.1748 ...

Mod.txt

	Dependent variable:
	GDPG
EDUE	-0.168 (0.683)
FDII	0.868 (0.771)
Constant	1.316 (4.473)
Observations	18
R2	0.089
Adjusted R2	-0.032
Residual Std. Error	1.845 (df = 15)
F Statistic	0.735 (df = 2; 15)
Note:	*p<0.1; **p<0.05; ***p<0.01

REG.txt

	.1	REG_BASIC	REG_FDI
1	Dependent Var.:	log(GDPG)	log(GDPG)
2			
3	Constant	1.965 (1.843)	1.607* (0.6847)
4	EDUE	-0.2152 (0.3113)	
5	FDII		-0.5070 (0.3668)
6			
7	S.E. type	IID	IID
8	Observations	16	16
9	R2	0.03302	0.12007
10	Adj. R2	-0.03605	0.05722

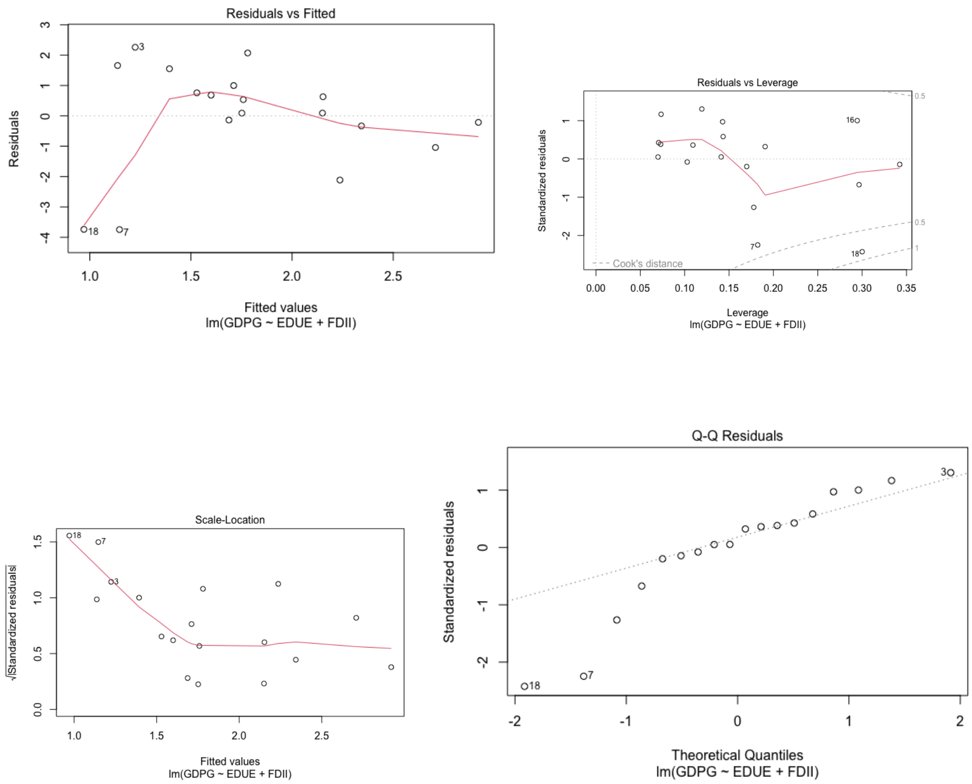
csum.txt

	Estimate	Std. Error	t value	Pr(> | t|)
(Intercept)	1.316	4.473	0.294	0.773
EDUE	-0.168	0.683	-0.247	0.808
FDII	0.868	0.771	1.125	0.278

Analysis

Being completely forthcoming, I am not completely sure what these scatter-plots exactly mean. Having never taken econometrics, or any coding previously, I will do my best to interpret. Residuals constitute the difference between the actual values, and the values predicted by the model. The first graph's line, titled "Residual vs Fitted", illustrates the idea that higher fitted values are not associated with higher residuals. This would infer the idea that when predicting higher investments in education or foreign investment, they are not alway associated with higher values for GDP-growth. Interestingly, there is a positive correlation in the short-term between fitted values and residuals. Fitted values from 1 to 5 are associated with negative residuals, and are also positively correlated. This may infer the previously discussed theory of diminishing returns, or that those with low values experience growth while those already developed do not. The only other graph that also grabs my attention is the fourth graph titled "Q-Q Residuals". This graph, unlike the previous graph, shows a clean and positive correlation between my independent and dependent variables. In this fourth graph, as theoretical quantities rise the standard residuals also rise. This may infer the idea that despite the relationship not holding in the short-term, the relationship is theoretically (in the longer term) positively correlated.

Moving on to my tables, I ultimately found that neither of my independent variables were statistically significant in their effect of GDP-Growth. Looking at my CSum table, one can see this phenomenon from the given P-Values. A P-Value is significant when it is less than .05, which none of my independent variables are. Boht of the P-Values given, for my independent variables, are substantially greater than .05. With EDUE at .808 and FDII at .278, both were found to be statistically insignificant in determining GDP-Growth. While there may be an overall weak positive correlation (as illustrated in the fourth graph) between my variables, it is not significant enough to conclude causation.



{plot(MOD)} (for above graphs)

csum.txt

	Estimate	Std. Error	t value	Pr(> | t|)
(Intercept)	1.316	4.473	0.294	0.773
EDUE	−0.168	0.683	−0.247	0.808
FDII	0.868	0.771	1.125	0.278

{CSUM <- summary(MOD)$coefficient} (For table above)

Text Tables

stargazer(CSUM, summary = FALSE, type = "text", out = "csum.txt")

(how they were saved)

Conclusion

Ultimately, there seems to be at least a weak association between education expenditures and GDP-Growth. Despite not being statistically significant, the graphs may infer some positive relationship. Additionally, the weak correlation found may be the result of time series tendency to overestimate. This is a known issue for time series, where the correlated result is greater than what it actually is. This unaccounted for issue may negate the weak relationship found within my created scatter-plots. Furthermore, my model failed to account for my independent variables being affected by my dependent variables. My model fails to account for GDP-Growths effect on public education spending or Foreign Direct Investments. I attempted to account for this issue with a time-lag, however, this introduced errors into my model in addition to not fully accounting for the issue. All in all, my model found a weak correlation that is likely less than listed due to issues with regressing a time series.