Analsying ease of doing business and GDP

Zsombor Hegedus

2020 december 18

Executive summary

In this paper I am going to examine the pattern of association between the time required to start a business and the income a country generates. In order to analyse the link between these two variables, I used OLS regressions on data downloaded from WDI in the year 2019. My results show that there is a negative pattern of association between the variables, and in my data GDP per capita is expected to be 0.25 lower for every observation where the time required to start a business was one day longer. Starting a business can be a complex and bureaucratic process, but as my results imply it is beneficial for governments to strive for creating an environment that nurtures small businesses instead of making it harder for them to enter their respective markets.

The analysis was done as homework for chapter 8 exercise 4 - files for the current analysis are stored in this github repo

Introduction

I am going to analyise the relationship between the days needed to start a business and the GDP per capita ratio of a given country (the GDP per capita ratio is based on purchasing power parity and expressed in international dollars - for the sake of simplicity I will just refer to this as GDP per capita going forward). My assumption is that if there are hard limits for a small company to enter a market, it will negatively impact the economy. I believe this to be the case because if the market is dominated by a couple of big players only, the level of competition will be low. The more parties will bring forward price competition which is good for the consumers, and it also incentivises companies to come up with new innovative ideas, to reduce their costs or improve their products. In such an environment, the parties that stand out, might decide to expand internationally increasing the global competititveness of the country. To prove my theory, and to quantify this assumed impact, I will use linear regression models on 2019 data for all countries, downloaded from the Wold Development Indicators database.

First of all, the two variables that I'm going to examine are *gdppc* (GDP per capita) and *time2business* (time required to start a business). Unfortunately data was not available for all countries but in the end I still ended up with 178 observations, where each observation is a country. Some summary statistics can be seen in Table 1 and the distribution of the variables are also visualised in Figure 1.

variable	mean	median	std	iq_range	min	max	skew	numObs
GDP per capita	20.66	13.37	20.70	24.75	0.75	114.48	1.64	178
Days needed to start business	17.89	11.75	20.13	13.38	0.50	173.00	3.81	178

Table 1: Summary statistics of examined variables

We can see that both variables are skewed with a long right tail. I treat both as ratio variables, altough one could argue that when we talk about days to start a new business, we rather do comparison on the absolute differences and not in relative terms (e.g. in country A, it takes 2 more days to start a business than in country B). The mean GDP per capita was at 20.66, with a few extreme values such as the GDP per capita of the Luxemburg and Singapore, but naturally I kept all observations as they are valid records.

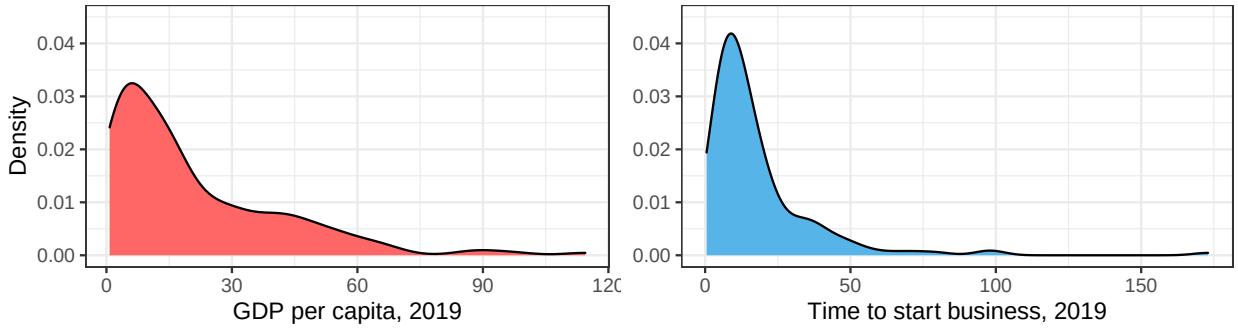

Figure 1: Distribution of days needed to start a business and GDP per capita for examined countries 2019

When looking at time to start a business (methodology on how data was collected is available in this link), it was apparent that most of my observations are concentrated between 0 and 30 days with mean being ca.18 days, but a non-negligible amount of countries had higher values than that with one ultimately high value of 153 days that belonged to Laos PDR. Since I'm looking for an average association in the data, and since I have no means to prove that this value is incorrect I decided not to exclude this from my population. But it is worth to note that this is an influential variable that has a relatively big impact on every slope coefficient that I have, so it might be worth investigating this further (with two other observations that have close to 100 days).

Regression analysis

To get an understanding on the relationship between my variables I created four scatter plots with a lowess function that can be seen in Figure 2. The first scatter plot in the top left shows the level-level case, which is the two variables without any transformation. We can see two very important phenomena (1) there is a substantial impact coming from the influential variables and (2) the lowess shows that the association is non-linear. Since I have ratio variables, I also visualised three other scenarios, where I used log transformation (log-level, level-log, log-log case). In my view the most interesting results are from the log-log case, where we can see that the lowess almost shows a linear trend in the data.

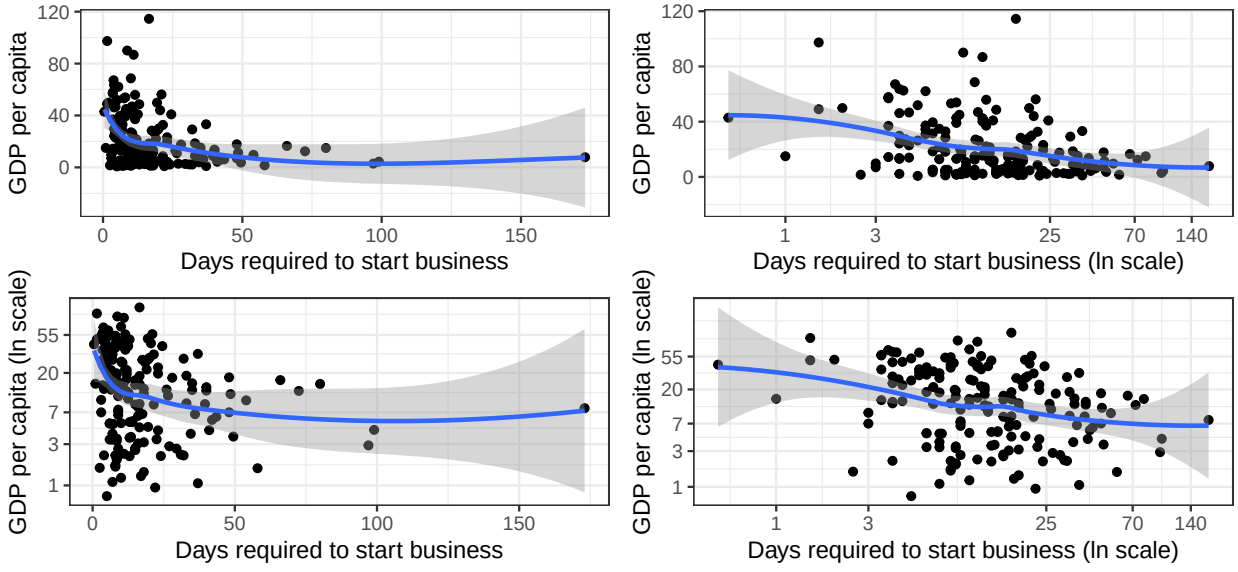

Figure 2: Relationship between days needed to start a business and GDP per capita for examined countries in 4 different ways 2019

Based on these findings I thought that it is best I transform the variables. When it comes to the left hand side cariable, GDP per capita, since it is a ratio variable with a distribution with a long right tail, I decided to take it's natural logarithm. For the x which is a variable with similar characteristics to y, in order to capture the non-linear, I decided to experiment to improve the fit of the baseline, level-level case:

1. I take it's natural logarithm as well
2. I use it's quadratic form
3. I use a piecewise linear spline model, since it was quite interesting to see from the lowess, that the pattern is the steepest between days 0 and 10, and then the slope changed slightly again after day 50. I one knots on day 10, since when I experimented with multiple knots I couldn't say with 95% confidence that they were not the same.

These experiments can be seen in Figure 3.

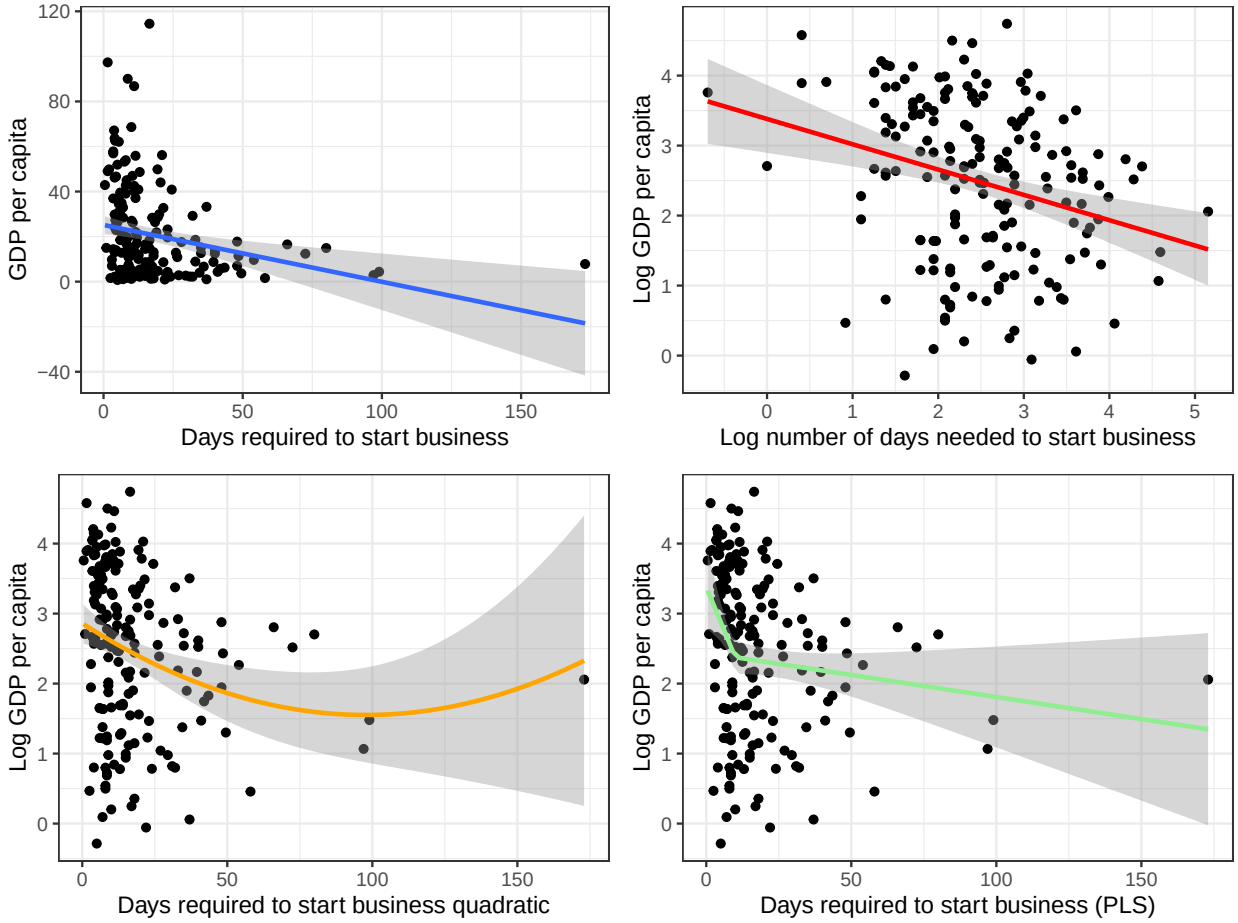

Figure 3: Regression models visualised for days needed to start a business vs GDP per capita for examined countries in 4 different ways 2019

We can see all regression lines capturing the non-linearity, so deciding between them is going to come down to inspecting their fit. In order to do so I look at the R^2 for each case and see how much of the variation of y is explained by the above models. I used the simple level-level regression as a base case and see how much I can improve this with the rest of the models. I also added an extra model, a weighted regression, to check if it makes a difference if we use population weights. The below model summary shows these results:

	level-level	log-log	quadratic	PLS	weighted
(Intercept)	25.18*	2.69*	2.86*	3.38*	2.77*
	[20.63; 29.72]	[2.45; 2.93]	[2.58; 3.14]	[2.79; 3.97]	[2.34; 3.20]
time2business	−0.25*	−0.01*	−0.03*		−0.02*
	[−0.41; −0.09]	[−0.02; −0.00]	[−0.04; −0.01]		[−0.04; −0.00]
t2b_sq			0.00*		
			[0.00; 0.00]		
lspline(time2business, cutoff)1				−0.10*	
				[−0.17; −0.03]	
lspline(time2business, cutoff)2				−0.01	
				[−0.01; 0.00]	
R^2	0.06	0.04	0.06	0.07	0.10
Adj. R^2	0.05	0.03	0.05	0.06	0.10
Num. obs.	178	178	178	178	178
RMSE	20.12	1.12	1.12	1.11	5.71

* Null hypothesis value outside the confidence interval.

It can be seen from the above that all of the models have a pretty weak fit ranging from 4%-10%, so the model is not performing particularly well when explaining the variation in y. From the 5 models, the weighted regression does the best job, but I wouldn't say it makes sense to use it as GDP per capita variable which is already normalised by the population of a country. The log-log case has the worst fit, so log transformation didn't really improve the model, and neither did the quadratic form. However there is a little improvement coming from the PLS model, with results that are also interesting to look at, so I would say that the two best models from the lot are the level-level and the PLS.

The level-level case: This model states that we can expect the GDP per capita variable to be 0.25 units lower when time to start business is 1 day longer. To be more precise with a one day longer time to start a business with 95% confidence we can say that the unit by which gdp pc is expected to be lower is between 0.41 and 0.09. The intercept coefficient was also significant at 5%, which basically means that we can expect GDP per capita measure to be at 25.18 in case a business can be started instantly, without any waiting time.

PLS: In my view the findings in the PLS are noteworthy because they highlight an interesting behaviour in our data - the regression line is significantly steeper for observations *time2business* is between 0 and 10 days. I can say that it is significantly steeper, because 95% CIs are not overlapping between β_1 and β_2 for the regression lines that are separated by the knots. The interpretation for β_1 is important here: for observations where *time2business* was below 10 days, we can expect GDP per capita to be 10% lower for observations having 1 day longer time to start a business. This is also significant at 5%.

Even though none of the models proved to have a good fit in the data, all of them showed a slope with significant negative impact at 5%. Subsetting the population by groups with similar geographical location, or economic structure might improve how we can analyse these patterns, but the above is already indicative that it is in the best interest of any country to reduce the number of days to start a business in order to achieve a higher economic output.

Summary

In this paper I analyised the pattern of association between times required to start a business and GDP per capita. My intuition was that there exist a negative relationship between this two - with more days required to start a business we can expect a country's generated income to be lower. I downloaded GDP per capita and time required to start a business variables from 2019 from the World Bank's database. I ran linear regression models with a few transformations in variables to accomodate the non linear patterns in my data. None of my models provided a good fit but all of them implied a significantly negative slope between the two variables. One of the models I used was a PLS that implied that the behaviour of countries where *time2business* was below 10 days was diffenert. Overall the main message of this paper is to highlight that it is worth considering for governments to improve their processes with regards to the foundation of companies.