

Term Project

Zsiros, Gabriella

2022-12-20

Introduction

As the purpose of my Analysis, I wanted to look at migration patterns of various economies over the world. The main source I used is part of the Worldbank Databank, where I selected some of the indicators I was interested in. Source: <https://databank.worldbank.org/indicator/SP.DYN.LE00.IN/1ff4a498/Popular-Indicators>

Each observation represents a country, and I had several year's data at my disposal. As the main indicator is migration data, I wanted to find out if there is a relationship between migration numbers and "how well" country's economy is doing. For this purpose I chose one of the basic indicators of an economy: GDP per capita. In the next paragraphs I will introduce the dataset I specifically selected from Worldbank and will try to determine a relationship between the variables by running a linear regression.

About the data:

The dependent variable (y) in question is migration, which is called 'Net migration' as indicator. This "the number of immigrants minus the number of emigrants" that is observed during a five year period. This means that data of 2017 actually covers the period of 2012-2017; the data series before that, from 2007-2012.. etc. Considering the net migration data, it is an important conclusion that observations (countries) with negative net migration value experience more residents moving from the country than moving into. Positive numbers mean the migration pattern just the other way around, with more people immigrating to a country than emigration from it.

The independent variable in my dataset is GDP per Capita, in US dollar currency.

Since the migration numbers are not really conclusive in themselves, it is important to consider the Population of the country as well.

Aside from these variables above, and the country names, I have included Life expectancy at birth and Human Capital index in the dataset. Life expectancy in this context means "*Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.*" while Human Capital index (HCI) is a calculated value between 0 and 1 which indicates the contributions of health and education to worker productivity.

To make the indicators easier to interpret, I have introduced the following naming conventions:

- GDP per Capita is denoted as **GDPPC**
- Life expectancy at birth denoted as **LEaB**
- Human Capital Index as **HCI**
- Migration with regards to the population is added as a new calculated variable and denoted as **Mig_rate**



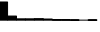



For the purpose of the analysis, I selected only **one period of time: 2017**.

Models

Concentrating on the renamed and relevant columns, I have decided to remove the missing variables and keep extreme values. I have also added a calculated column of Migration numbers divided by population, thus creation a net migration ratio, which fares better with the intended independent variable, GDPPC.

An overview of the data shows the GDPPC having a log-normal or possibly pareto distribution, with a long right tail.

Table 1: Statistucal features of the dataset

	Unique (#)	Missing (%)	Mean	SD	Min	Median	Max	
Migration	136	0	47 446.9	622 335.1	-2 663 434.0	-4630.5	4 774 029.0	
Population	154	0	47 380 139.1	159 851 168.6	95 843.0	10 553 774.5	1 396 215 000.0	
GDPPC	154	0	14 584.3	20 104.2	286.4	5189.9	107 142.1	
LEaB	154	0	72.5	7.8	52.9	74.3	84.7	
HCI	134	0	0.6	0.2	0.3	0.6	0.9	
Mig_rate	154	0	0.0	0.0	-0.1	0.0	0.2	

Lowess

Comparing several models and plotting the dependent and independent variables, a sensible choice was to take the logarithmic value of the GDPPC (x) as the change/increase is better represented by approximate percentages, rather than actual dollar values. With level x, the observations are clustered around relatively small numbers. Following the curve of the lowess model with level variables, the slope of the curve intuitively changes at specific values the x is taking, namely 5000 and 25000 \$ GDP per capita which will be used late on as nodes in a spline transformation in the analysis. (*see table in appendix*)

Taking a logarithmic value of the dependent variable (y) is not preferred here, as the Net Migration rate takes on values that are not positive as well.

After a log transformation of the x, Locally Weighted Scatterplot smoothing (LOWESS) seems to capture a good curved fit, which is depicted by default with a 95% percent confidence interval. Standard error is visibly better where we have more data, where x takes on small values and increases as the datapoint become more scarce.

Linear models

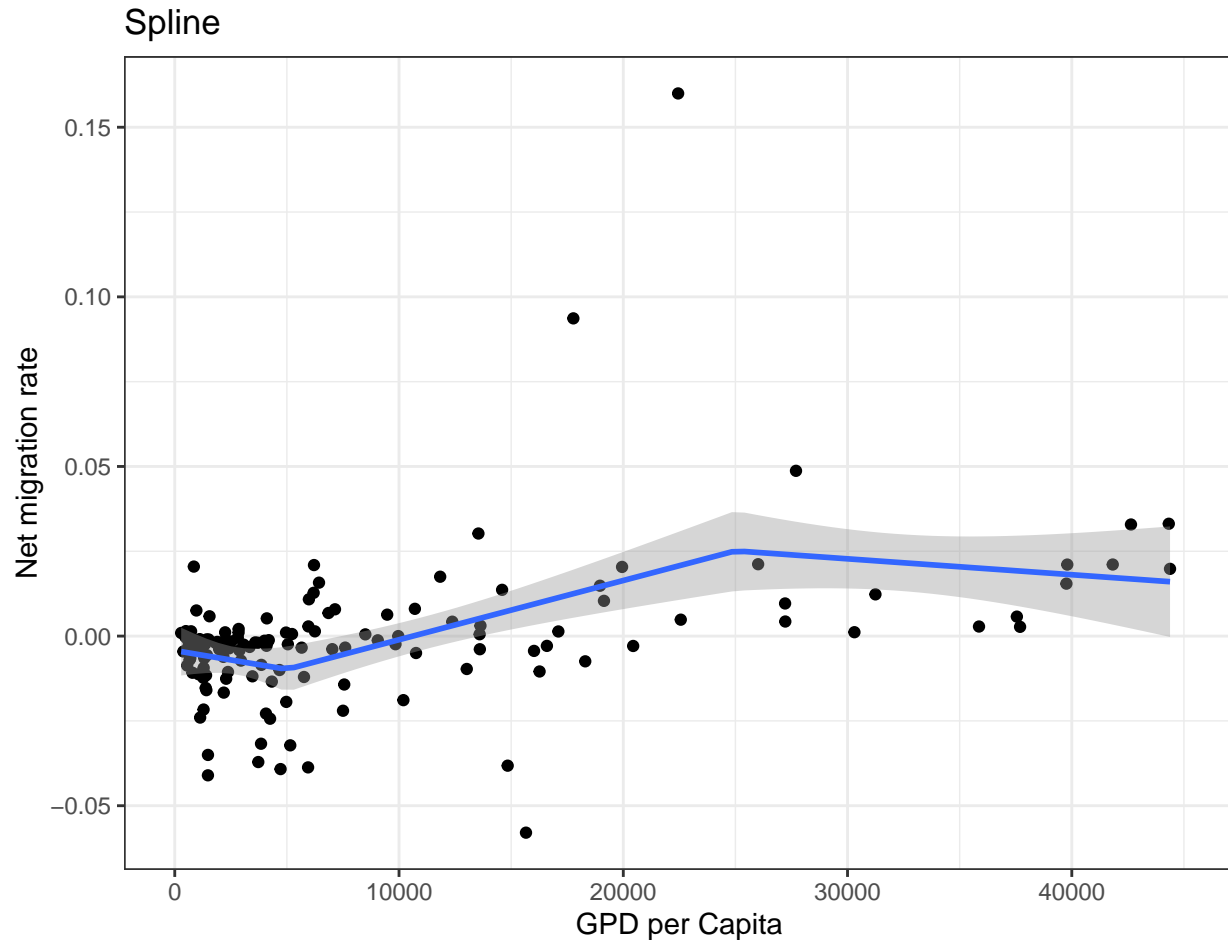
After comparing more linear models the best fit probed to be the Spline model with nodes at 5000,25000 and 45000 respectively.

Table 2: Spline model

	(Intercept)	<5k	>5k,<25k	>25k,<45k	>45k
Model 1	-0.004 (0.002)	-0.000 001 (0.000 000 9)	0.000 002 (0.000 000 7)	-0.000 000 5 (0.000 000 8)	0.000 000 8 (0.000 000 3)

The intercept shows that the Net migration rate where the x takes on the value of 0 is -0.005 (meaning 0.5% of the population emigrates from the origin country). However, since the GDP per capita's minimum value is 286.4, this information has little relevance. Under 5000, the slope coefficient has a value of - 0.000001,

as the GDP per capita increases by 1 unit (dollar), the net migration rate decreases by 0.000001% (as the net migration is already denoted in percentage). This can also be expressed by 0.01% change for every 10 thousand dollar increase in GDP. After 5000, the slope changes however, a 0.000002% increase in Net migration rate until 25000 then 0.0000005 decrease (almost a flat line, this is one decimal lower than the other coefficients), and finally a 0.0000008 increase. Also can be interpreted as 0.01% decrease, 0.02% increase, 0.005% decrease, then 0.08 increase by every 10 thousand dollars increase in the countries GDP per Capita



Alternatives

As alternative models, I have analyzed two alternative models: one not transformation the variables, only considering the Migration as ratio, while the other took the logarithmic value of GDP per capita. Using robust standard errors, both the R^2 and adjusted R^2 supports the original Spline.

The second model (level-level) shows a 0.0000006 increase in net migration rate by increasing the GDP Per capita on average, while the third model (level-log) shows a 0.00007 increase in the migration rate when the GDP per capita is increased by 1%.

Table 3: Model comparison

	Model 1	Model 2	Model 3
(Intercept)	−0.004 (0.002)	−0.007 (0.002)	−0.062 (0.010)
<5k	−0.000 001 (0.000 000 9)		
>5k,<25k	0.000 002 (0.000 000 7)		
>25k,<45k	−0.000 000 5 (0.000 000 8)		
>45k	0.000 000 8 (0.000 000 3)		
GDPPC		0.000 000 6 (7×10^{-8})	
log(GDPPC)			0.007 (0.001)
Num.Obs.	154	154	154
R2	0.338	0.297	0.210
R2 Adj.	0.321	0.292	0.205
Std.Errors	HC3	HC3	HC3

Multivariate regression

Using additional variable such as life Expectancy at Birth and Human Capital Index introduces more granularity to the data. Taking the nominal value of GDP per capita, the model shows that in countires with the same GDP per capita level, the migration increases be 0.0005% with a higher life expectancy, while the rate decreases by 0.042% the the HCI is increased and all aohter variables remain the same.

Table 4: Multivariate model

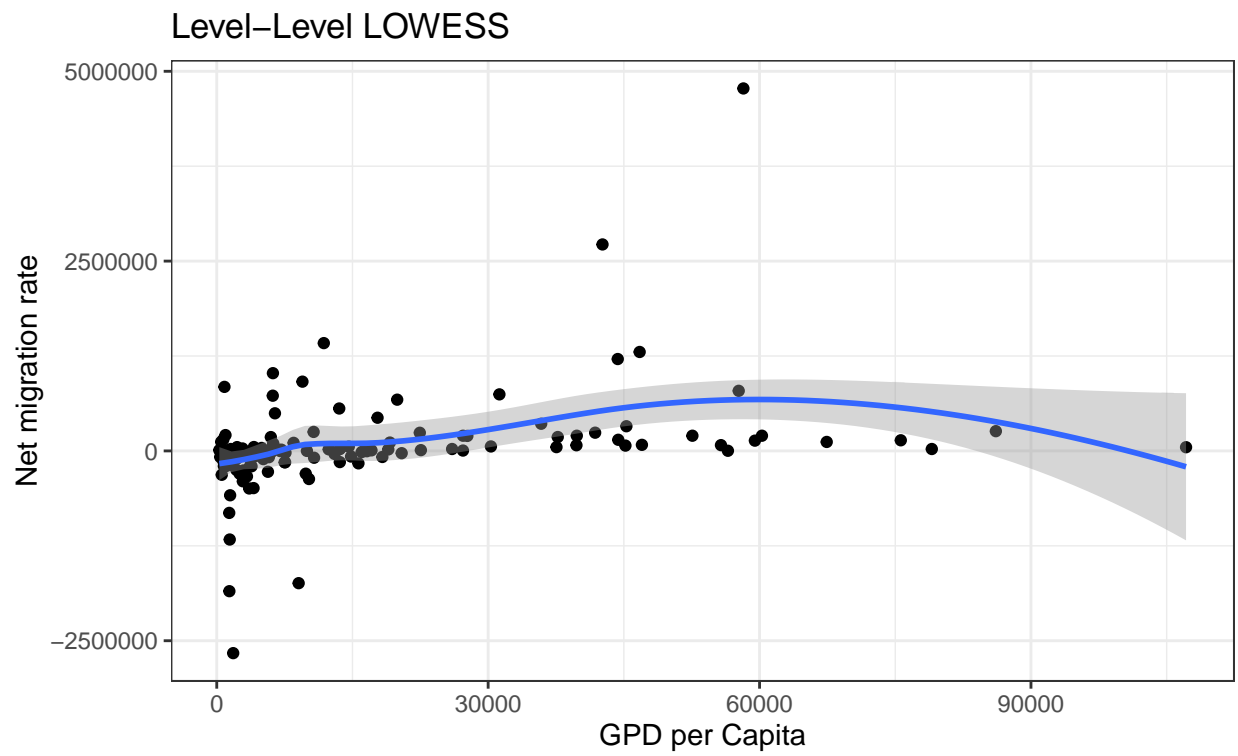
	(Intercept)	GDPPC	LEaB	HCI
Model 1	−0.024 (0.018)	0.000 000 7 (0.000 000 1)	0.0005 (0.0004)	−0.041 (0.021)

Summary and conclusion

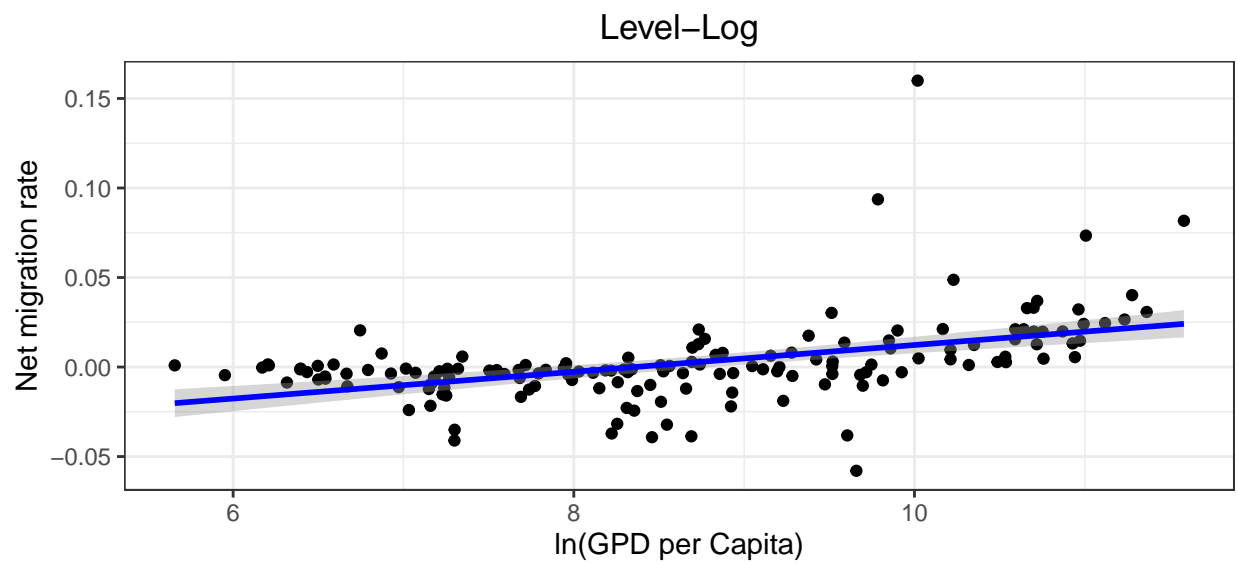
The regression of the dataset could be captured with various methods, of which the spline model proved to be the most accurate. Depending on the the GDP per capita range, the change in net migration rate varies, at some point more significant than others, hence the fit of the spline model. Introducing more variable from the several indicator available at WorldBank, we can introduce some granularity which opens to topic for further discussions on what makes a country a migration target, on top the the GDP economic indicator.

Appendix

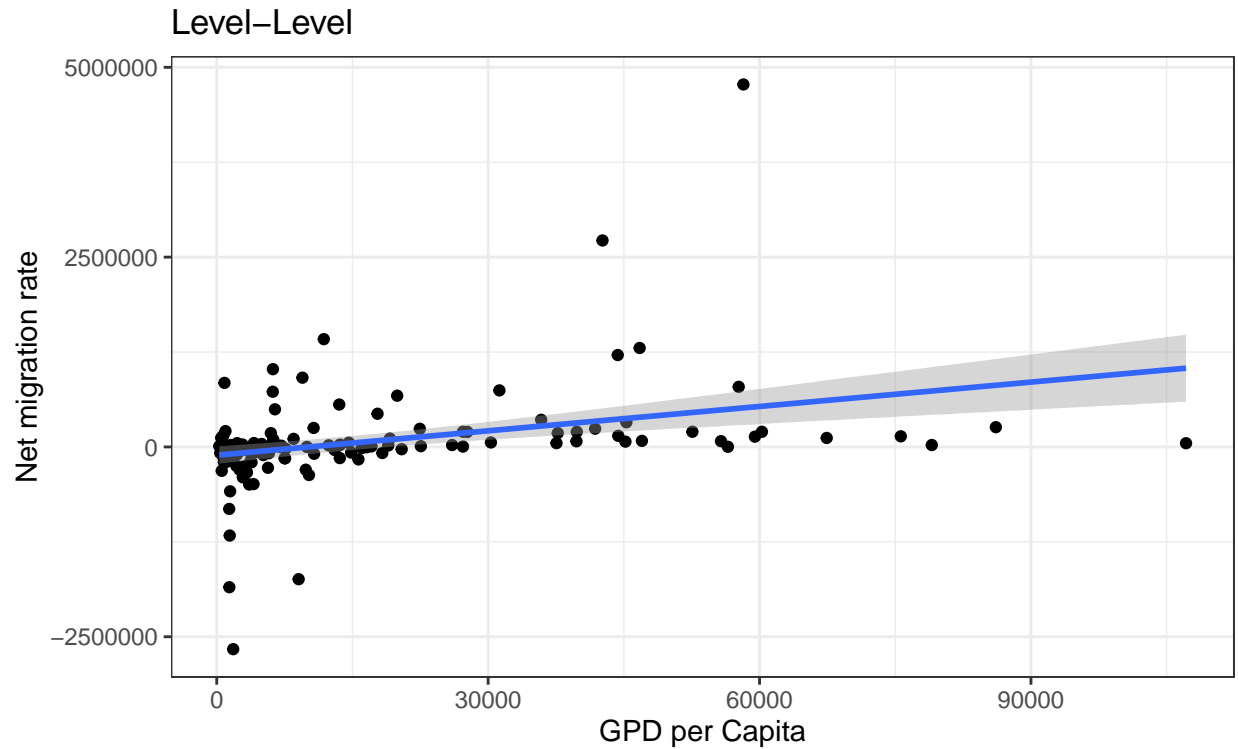
Level-level plot with lowess



Level-Log plot



Level-level plot



Datatable

##	Country.Name	Migration	Population	GDPPC	LEaB	HCI
## 1	Afghanistan	-314602	36296111	553.3551	64.13000	0.389
## 2	Albania	-69998	2873457	4249.8037	78.33300	0.621
## 3	Algeria	-50002	41389174	4192.3377	76.49900	0.523
## 6	Angola	32066	29816769	2845.4317	60.37900	0.361
## 8	Argentina	24000	44044811	13595.0374	76.37200	0.611
## 9	Armenia	-24989	2944789	3860.2181	74.79700	0.572
## 11	Australia	791229	24601860	57695.5713	82.50000	0.803
## 12	Austria	324998	8797566	45281.7234	81.64390	0.793
## 13	Azerbaijan	6002	9854033	5229.5261	72.69300	0.597
## 15	Bahrain	239000	1494077	22445.4526	77.03200	0.668
## 16	Bangladesh	-1847503	159685421	1394.7814	72.05200	0.479
## 19	Belgium	240000	11375158	41825.7628	81.49268	0.757
## 21	Benin	-10000	11175192	1112.8171	61.17400	0.406
## 25	Bosnia and Herzegovina	-107926	3351534	5150.2733	77.12800	0.618
## 26	Botswana	14999	2205076	6855.1603	68.81200	0.424
## 27	Brazil	106000	207833825	8498.2939	75.45600	0.560
## 30	Bulgaria	-24001	7075947	7599.1250	74.81463	0.676
## 31	Burkina Faso	-125000	19193236	693.7264	60.76800	0.369
## 32	Burundi	10003	10827010	286.3955	60.89800	0.380
## 34	Cambodia	-149999	16009413	1289.9858	69.28900	0.493