Examining how various Economic Factors Influence Migration Rate

DATA-231 Final Project

Tyusha Sarawagi

12/14/2021

Introduction

Humans have been on the move since the earliest humans expanded out of Africa. Even today, more than 258 million people live in countries other than their own. Migration, whether voluntary or forced, has had a significant impact on our globe (Khan, Khalid & Su, Chi-Wei & Tao, 2019). The probability that migration will increase in the future in most countries around the world is extremely high, especially as the digital economy emerges and develops, multinational corporations' increasing role, and the simplification of border crossing (Podra, Kurii, Alkema, Levkiv, & Dorosh, 2020, Quak, 2019). The relationship between the quality of life in countries of origin of potential migrants and the countries of destination play a significant role in migration processes. The chance to earn significant opportunities to achieve one's potential is the primary driving force of economic migration. As a result, we see more people migrating to nations with better levels of development and quality of life and countries with less linguistic and cultural constraints, such as the United States, Canada, and Russia. This paper is devoted to researching how the GDP of a particular country, socio-economic and demographic trends, the population density, literacy rate, and infant mortality rate of a country influence how many migrants move to that country.

My home country, Nepal, is a developing country, and the GDP per capita is below the worldwide average. In Nepal, many folks go to foreign countries, including India, Dubai, the United States of America, Canada, the UK, etc., to earn a living and feed their families back home. So, I am interested in exploring if this is a common trend among other countries in the world, i.e., if people are more likely to leave poorer countries and emigrate to richer ones to have a better lifestyle. I want to know if economic disparities between developing and developed countries encourage laborers to migrate from developing countries to developed countries. This study could also help examine whether countries with higher emigration rates have lower GDP. This makes sense because if all the skilled laborers of a developing country emigrate to developed countries, nobody will be left in the developing countries to help the economy grow, leading to a lower GDP. It's also difficult to find studies of migration's impact on a host country that go beyond effects and include the contribution to the overall economy.

This project aims to explore the factors responsible for individuals migrating to another country. I plan to run multiple linear regression models with net migration as response variable and GDP per capita, death rate, infant mortality, and region as explanatory variables, examining their relationship with net migration. My primary research question for this project is, "are people more likely to emigrate out of a country when it has a lower Gross Domestic Product (GDP)?" I also plan to address questions like, "do countries with higher populations have a greater number of immigrants?", "do countries with higher death rates have higher net migration?", "what is the impact of literacy rate on migration?”, and "do countries with higher infant mortality rates have higher net migration?"

Materials and Methods

The dataset was found from Kaggle and is titled *countries of the world*. The data was uploaded by Fernando Lasso under the CCO: Public Domain license on 26th April 2018. The dataset has been made up of data from the US government website between 1970 - 2017. The specific time of when the data was collected is unclear. It can be used freely by anyone at any time without seeking permission. The sampling method for the data is unclear, too, whether it was a simple random sample, convenience sampling, voluntary response sampling, etc.

Furthermore, the data might look like it is representative of the entire population as it includes all the countries of the world, but that is not the case. There might be biased data collection methods within each country; therefore, we cannot say that the data is representative. The dataset contains statistics on all the countries of the world. The data consists of the country’s population, area, coastline, net migration, GDP, infant mortality, literacy, phones, arable, crops, climate, birth rate, death rate, agriculture, Industry, and Service. Region is a categorical variable and is broken into eleven groups, Asia, Baltics, Commonwealth of independent states, eastern Europe, Latin America & Carib, near east, northern Africa, northern America, Oceania, sub-Saharan Africa, and western Europe. All the other variables are numerical variables with numeric values.

There were 110 missing values in the dataset and getting rid of all the rows with missing values did not seem like a good option because that way, I would lose a lot of information. Therefore, I decided to use two methods to deal with the missing values. Firstly, I deleted the variables Agriculture and Industry from the dataset. I got rid of these two variables because they had a lot of missing values, and I was not going to use them for my modeling. After that, I got rid of the rows, West Sahara and Wallis and Futuna, as they had a lot of missing observations. We still were left with 67 missing values that were randomly scattered across the dataset. I decided to use the mean substitution method of imputing values in all these missing observations. I calculated the mean value of the variables with missing data and used it to replace the missing observations. At this point, we have dealt with all the missing values. Initially, the dataset had 227 observations and 20 variables. However, after data cleaning, we are left with 225 observations and 18 variables.

Data analysis was done by running a multiple linear regression model and integrating various transformations and interaction terms to fit the best model with the available variables. I chose multiple linear regression instead of logistic regression as my dependent variable**,** net migration, is a numerical variable. I created a correlation matrix and various scatterplots between Net Migration and all the other variables to learn more about the variables that I should include in my model and examine their relationships.

Furthermore, I used stepwise regression to find the best model without using any interaction terms or transformations to predict the net migration. For my model, the first model has seven potential predictors for the model, and the final model has four best variables with transformations and interaction terms. My criteria for keeping variables in the model are the p-value < 0.1, an increase in R^2 value, and a decrease in standard residual error. From my final model, I got the value of multiple R^2 to be 0.3232 and the value of adjusted R^2 to be 0.2781. I then calculated the confidence intervals for all my variable and conducted ANOVA F-test to determine if the variables that I got rid of were significant for my model or not. I found out that those variables were not significant, and it helped me eliminate the insignificant variables, thereby reducing the complexity of the model. All the models were also checked for having met all the inference criteria like linearity, normality, independence, etc., before making an inference.

Results

A summary of the initial analysis of the numerical variables net migration, gross domestic product, death rate, and infant mortality can be found in *Table 1* and the summary of the categorical variables Region can be found in *Table 2.* The response variable is net migration which is the difference between the number of immigrants and the number of emigrants, expressed in percentage. GDP per capita measures the economic output of a nation per person. Death rate is a measure of the number of

Variables	Mean	Median	SD	Minimum	Maximum
Net Migration	0.03	0.00	4.87	-20.99	23.06
GDP ($ per capita)	9716.44	5600.00	10063.50	500	55100
Death Rate	9.24	8.10	4.96	2.29	29.74
Infant Mortality (per 1000 births)	35.50	21.03	35.31	2.29	191.19

Table 1: Description of the most relevant numerical variables in the model

deaths in a population, scaled to the size of that population, per year. Infant mortality is the rate of a child born in a specific year or period dying before reaching the age of one. Region is an area, especially part of a country or the world having definable characteristics but not always fixed boundaries. After making the correlation table, some variables are more statistically significant as compared to other variables. The final model contains the variables seen in *Table 1* and *Table 2* as they are statistically significant for my model.

Variable	n	%
Region		
Asia (Near East)	28	12.44%
Commonwealth of Independent States	12	5.33%
Latin America and Caribbean	45	20.00%
Northern Africa	5	2.22%
Oceania	20	8.88%
Western Europe	28	12.44%
Baltics	3	1.33%
Eastern Europe	12	5.33%
Near East	16	7.12%
Northern America	5	2.22%
Sub-Saharan Africa	51	22.67%

Table 2: Summary of the most relevant categorical variable in the model

Table 3 shows us some notable correlations between different variables in our model. Net migration correlates the strongest with GDP per capita amongst all the other variables, having a correlation coefficient of 0.382. As shown in the table, the most correlated predictors are infant mortality and literacy, with a coefficient of determination of -0.750. This strong negative correlation makes sense because educated individuals are known to make informed healthcare decisions. This results in better infant care reflected by lower infant mortality rates. GDP is moderately related to literacy rate with a correlation coefficient of 0.497. This is logical as a country with more literate people is more likely to do better and have a higher GDP than a country

	Net Migration	Infant Mortality	GDP per capita	Death Rate	Region	Literacy
Net Migration	1.000	-0.025	0.382	0.033	0.069	-0.008
Infant Mortality	-0.025	1.000	-0.600	0.655	0.149	-0.750
GDP per capita	0.382	-0.600	1.000	-0.201	0.190	0.497
Death Rate	0.033	0.655	-0.201	1.000	0.330	-0.389
Region	0.069	0.149	0.190	0.330	1.000	-0.193
Literacy	-0.008	-0.750	0.497	-0.389	-0.193	1.000

Table 3: Correlation matrix of all the relevant variables

with fewer literate people. We can see other interesting correlations between infant mortality and death rate; they are strongly correlated with a correlation coefficient of 0.655. Net migration is weakly negatively correlated with literacy rate, having a correlation coefficient of -0.008. These correlations are essential for summarizing a dataset as well as identifying and visualizing patterns in the data.

Coefficients:					
	Estimate	Std. Error	t value	P value	95% CI
(Intercept)	-7.2119	2.7332	-2.639	0.0089	(-12.60, -1.82)
GDP per capita	0.0010	0.0002	4.947	1.55e-06	(0.00065, 0.00151)
R: Baltics	-1.2048	2.6906	-0.448	0.6547	(-6.51, 4.09)
R: C.W of Indep. States	-1.8925	1.4870	-1.273	0.2045	(-4.82, 1.03)
R: Eastern Europe	-0.1553	1.5828	-0.098	0.9218	(-3.27, 2.96)
R: Latín América & Carib.	-1.9029	1.0357	-1.837	0.0675	(-3.94, 0.13)
R: Near East	0.8025	1.3624	0.589	0.5564	(-1.88, 3.48)
R: Northern Africa	-0.2129	2.0615	-0.103	0.9178	(-4.27, 3.85)
R: Northern America	-3.9067	2.2075	-1.724	0.0861	(-8.15, 0.54)
R: Oceania	-2.5250	1.2445	-2.029	0.0437	(-4.97, -0.07)
R: Sub-Saharan Africa	-1.7681	1.1770	-1.502	0.1345	(-4.08, 0.55)
R: Western Europe	1.0549	1.6123	0.654	0.5136	(-2.12, 4.23)
log(Death rate)	2.3028	1.5014	1.534	0.1265	(-0.65, 5.26)
Infant Mortality	0.0323	0.0169	1.906	0.0580	(-0.0011, 0.0657)
GDP*log(Death rate)	-0.0004	0.0001	-3.593	0.0004	(-0.00064, -0.00018)

Table 4: Result from the multiple linear regression where Y is net migration, R implies Region, RSE value= 4.14, F-statistic: 7.162, p-value: 4.623e-12

Table 4 depicts the results of the multiple logistic model that was developed, which reveals the results of the overall efficiency of the significant variables. The final model contains the best variables GDP per capita, region, death rate, and infant mortality with some transformations and interaction terms. It contains interaction term between GDP per capita and Death rate. This interaction was modeled because these events are not independent of each other, and thus one can enhance the other.

Deathrate was log transformed as the variable is right-skewed, and transformation helps with linearity, and it will help achieve a better model fit. The final model shown in *Table 4* produced an R^2 value of 0.2781, indicating that 27.81% of the variation in the model can be explained by these final selected variables. A thousand dollar increase in GDP per capita changes the net migration by (1.08 – 0.42(log(Deathrate))) units, holding all the other variables constant. A hundred-unit increase in deathrate changes the net migration by (2.3 – 0.00042(GDP)) units, holding all the other variables constant. For hundred additional deaths of infants per 1000 births, net migration increases by 3.23 units, holding all the other variables constant. Region is a categorical variable and interpretation is the difference between the response group and the variable group, holding all the other variables constant. The reference group for region is Asia. The difference with Baltics region is 2.52 units lower, Commonwealth of Independent States is 2.15 units lower, eastern Europe is 0.10 units lower, Latin America and the Caribbean is 2.06 units lower, Oceania is 2.52 units lower, and western Europe is 2.02 units lower, holding all the other variables constant.

The 95% confidence interval for all the significant variables are shown in *Table 5*. For GDP per capita, we are 95% confident that a one dollar increase in GDP per capita will increase the net migration between 0.00065 units and 0.0015 units, holding all other variables constant. The 95% confidence intervals for the slope of coefficients of infant mortality contain zero, which is consistent with our findings that they only provide moderate or weak evidence for the model based on the p-value that we saw earlier. For region, we are 95% confident that the difference between Asia and Oceania is a log odds ratio of -4.97 units and -0.07 units.

It is crucial to ensure that the final model satisfies all the assumptions of a multiple linear regression model. The first assumption is constant variance, and it is checked by the residual's vs. fitted values plot of the final model, and it shows reasonably constant variance. Normality of the residuals is satisfied by the standardized residuals vs. theoretical values, residuals lie reasonably close to the theoretical line. Therefore, we can say that the normality condition is met. Another assumption of a multiple linear regression model is the independence of each outcome for net migration. We cannot say that the variables are completely independent of each other. For instance, people living in Nepal are most likely to migrate to India as compared to Russia as India is closer to Nepal and has better facilities than Russia. In this case, we cannot say that it is entirely independent but not to the level that could be concerning or problematic. Another assumption is the influential points, and we do not have any influential points in our data. The sampling method of the data is not provided in the source from where the data was collected, so we are unclear about the random/ representativeness assumption. Finally, linearity conditions were not met in the model, as seen in the scatterplots from the appendix. To deal with the problem of linearity, interaction terms were used, and various transformations were conducted.

Discussion

The final model with the variables GDP per capita, region, death rate, and infant mortality is ultimately the most significant predictor of net migration. Initially, the primary aim of this research paper was to find out if people are more likely to emigrate out of a country when it has a lower Gross Domestic Product (GDP). However, as I was moving forward with my research, I realized that I was asking the wrong question and I could have improved my question. The primary findings of this research were that GDP per capita, region, death rate, and infant mortality are significant predictors of net migration according to the dataset that we are working with. An interesting finding of this paper was that GDP is, in fact, the best predictor of net migration, and this is interesting to me because my primary aim was to explore the relationship between GDP and net migration. From my final model, it might look like GDP does not have much effect on migration, but that is because of the interaction term; but if you look at the previous models, you will see that GDP is the most significant predictor of net migration. I also wanted to explore other relationships like the relationship between population density and net migration, the relationship of death rate and literacy with migration, and the relationship of infant mortality with migration. I was unable to explore all these relationships in-depth because I got rid of some of the variables that were not statistically significant in my model.

One implication from these results is that the death rate was not a significant predictor of net migration in the initial model. However, when we interacted death rate with GDP, the p-value was nearly equal to 0, making this interaction a significant predictor of net migration. Interaction effects indicate that a third variable influences the relationship between an independent and dependent variable. This type of effect makes the model more complex, but if the real world behaves this way, it is critical to incorporate it in your model. Future researchers should explore this in-depth with advanced statistical tools like the interaction plot.

Limitations of this study include that migration is a complex concept and cannot be predicted just through economic factors. There are a lot of other factors such as socioeconomic factors, political factors, and ecological factors that affect net migration. There are a lot of different variables that could predict net migration better, like race, ethnicity, religion, politics, climate change, and many more.

Future researchers in this field could include this variable in their model for a better result. Another limitation can be that the sample size of the data is small with a limited number of variables. It could lead to a higher error of margin and question the reliability of the findings. Furthermore, it is unclear that whether the data is random and representative. We do not have information about the sampling method of the data, whether it was a simple random sample, convenience sample, clustered sample, or voluntary response sample. It is also unclear that the data is representative of the entire population.

Bibliography

Podra, Olha & Levkiv, Halyna & Koval, Ganna & Petryshyn, Nataliia & Bobko, Ulyana. (2020). The impact of migration processes on the economy of Ukraine: Trends, reasons, consequences. Journal of the Geographical Institute Jovan Civic. 70. 171-179. 10.2298/IJGI2002171P.

Khan, Khalid & Su, Chi-Wei & Tao, Ran & Yang, Lin. (2019). Does Remittance Outflow Stimulate or Retard Economic Growth? International Migration. 57. 10.1111/imig.12615.

Kaggle. “countries of the world.” Last modified 2018. https://www.kaggle.com/fernandol/countries-of-the-world