Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/47/20241101

Exploring correlations between economic indicators with natural and societal factors based on linear regression model

Shangxuan Zhong

The Department of Statistics, University of Toronto, Toronto, M5S 1A1, Canada

wind.zhong@mail.utoronto.ca

Abstract. Existing research on the determinants of a nation's economic development has predominantly centered on individual factors, including energy, land resources, education, taxes, employment, and healthcare. Regrettably, there is a paucity of studies that holistically examine these factors collectively and assess their respective contributions to economic development. Therefore, the primary objective of this study is to investigate the interrelationships between economic indicators and various natural and societal factors. The article firstly uses the Pearson's correlation coefficient to filter out a portion of the higher degree of correlation from factors that may have an impact on the country's economic development for further analysis. For the selected factors, using two linear regression models: Ordinary Least Square (OLS) method for preliminary modeling for the extent of affects between each factors and economy; and Fully Modified Ordinary Least Squares (FMOLS) method, as an optimization model, further eliminating the less influential variables. After obtaining the final impact model of the linear correlation, the data is screened based on the variables within the model. A portion of the selected data is used as a training set for training the model and the remaining data is used as a test set for testing the performance. The results of the study show that factors including land area, army size, CO2 emissions, population, minimum wage, would have varying degrees of integrated impact on the economic development of the country.

Keywords: Economy, Data Analysis, Linear Model, Machine Learning.

1. Introduction

Economic development, a fundamental property of a country, can provide more employment opportunities and increase income levels, thereby improving people's quality of life, reducing the poverty rate and promoting social stability, which plays a crucial part in maintaining social harmony and promoting the improvement of people's livelihood [1, 2]. Previous research has established that the development of economy can encourage scientific research and technological innovation, and thus facilitating industrial upgrading and increasing productivity. This facilitates the country in maintaining its position as a global leader in science and technology [3].

In the realm of economic analysis, understanding the intricate relationships between key economic indicators and various socio-economic factors holds immense significance. Economic indicators provide valuable insights into a country's financial health, while factors like health outcomes, job market conditions, tax policies, and energy prices can significantly influence its economic trajectory.

DOI: 10.54254/2755-2721/47/20241101

Understanding these interconnections is vital for policymakers, researchers, and business stakeholders alike, as it enables informed decision-making and effective policy formulation. As the world grapples with the far-reaching implications of global events such as pandemics and economic crises, a comprehensive analysis of these relationships becomes even more crucial. Nations with great economy would have higher global competitiveness and prominence on the international stage [4]. By studying these intricate connections, this article tries to get the possible impact of various factors on national economic development and the extent of the impact, so as to provide a reference for relevant researchers in this field.

Previous studies have extensively explored individual relationships between economic indicators and societal factors, e.g. education, health and employment. The contribution of existing studies lies in establishing key cause-and-effect relationships, serving as building blocks for economic analysis. The existing body of research on immigrants suggests that a substantial and statistically significant impact of immigrant numbers on the domestic Gross Domestic Product (GDP) of destination countries [5]. Factors including renewable and non-renewable energy sources, and also natural resources, are found to be influencing economic growth have been explored in studies of Usman et al., which shows that all these considerations would contribute to promoting economic growth [6]. Also, the influence of education to countries' economic development have been proposed by Dumciuviene [7]. However, these studies tend to focus on a specific factor or a limited set of variables, overlooking the intricate web of interactions that characterize real-world economies. While past research has provided valuable insights into isolated linkages, there is a notable limitation in the holistic exploration of multi-dimensional correlations between economic indicators and multiple interrelated natural or societal aspects.

Since much uncertainty still exists about the relationship between the economic development and socio-economic factors such as taxes, energy, employment, and education. The motivation behind this study stems from the need to address the limitations of previous research and to provide a more encompassing perspective on the relationships between economic indicators and natural factors or societal determinants. The intricacies of modern economies demand a nuanced approach that considers the multifaceted nature of interactions. Through an extensive analysis that incorporates a wide array of different factors such as natural energy, health care coverage, income and taxes, educational resources, etc., with the methods including Pearson Correlation, normal Ordinary Least Square (OLS), as well as the Fully Modified Ordinary Least Squares (FMOLS). By examining the interplay of multiple variables, this study strives to provide a more comprehensive and insightful framework for analyzing the complex dynamics that shape economies around the world, which also provides a possible reference for researchers in related fields.

2. Method

2.1. Dataset description and preprocessing

In this paper, a comprehensive dataset named "Global Country Information Dataset 2023" from Kaggle [8] was employed. This dataset provides a great deal of information on various global countries, covering abundant and various indicators and attributes, including demographics, environmental, economic, healthcare and education etc. There are 195 countries with 34 different features for each. However, there exists some blank or missing values in the dataset, and some of the indicators, like the name of a country's largest city, hold no relevance for this study.

To preprocess the original dataset, the first step is to convert each column into "float" data type, and delete all the signs such as comma (,), percent sign (%), or dollar sign ($). Next is to filter the needed data columns, including GDP, Carbon Dioxide (CO2) emissions, total tax rate, and more. While the data that is useless to this study, such as names of capitals, codes of currencies used, etc., will be excluded. Furtheremore, for rows with blank data cells, they are similarly eliminated, and only the data from 121 countries will be left in the end.

DOI: 10.54254/2755-2721/47/20241101

2.2. Methods for analyzing the correlation

Since this study used different methods, the following part will describe the rationale and role of each model and method, and how to implement these methods or apply these models in the processed dataset.

2.2.1. Pearson Correlation Coefficient [9]. To find the correlation between two sets of data, the Pearson correlation coefficient is one of the most widely used method, which can measure the extent of their association. The calculation formula is as shown in equation:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (1)$$

Thereinto, x and y represent the observed values of the two sets of variables, while $\bar{x}$ and $\bar{y}$ denote the means of x and y, respectively. And the i is the index.

The range of correlation coefficient is from -1 to 1. A stronger positive correlation is reflected by a coefficient close to 1, that is, an increase of one variable will cause the trend of increase of other variables. While a value approach -1 implies strong negative correlation, and a value close to 0 means that there is almost no linear relationship between the two variables [10].

When calculating the correlation coefficient, the p-value is used to assess whether the observed correlation coefficient is statistically significant. Under the null hypothesis (usually means no correlation between the two sets of variables), the p-value represents the probability of observing the correlation coefficient within a certain range. Generally, a threshold such as 0.05 (a 5% significance level) will be used as a criterion for judging the p-value. If the computed p-value is less than the threshold, the observed correlation coefficient would be statistically significant. While a p-value greater than the threshold suggests lacking statistical significance [11].

In this study, GDP and GDP per capita are used as indicators that respond to a country's level of economic development. In terms of each variable in the dataset, the degree of correlation between GDP and GDP per capita will be calculated, so that the variables that are highly correlated with these two are filtered out to be used in the following study of predicting models.

2.2.2. Ordinary Least Square (OLS) regression model. OLS is a common method used to fit linear regression models. In OLS, by making the squared difference between actual observed values and model predictions smallest, the regression coefficients are estimated, and it would give an optimal fitting line [12].

The main idea of OLS is to find a straight line such that the sum of the distances of all data points to this line is minimized. And in this study, it will be completed by the following steps. Based on the results of the Pearson correlation coefficient analysis, the variables and data needed in the data set were selected to build a linear model. The usual form is $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$, where x, y are the independent and dependent variables separately, s are the regression coefficients, and is the error term. Next is to minimize the sum of squares of errors, that is, find a set of s such that the sum of squares of the error term is minimized. And finally, the fit of this regression model needs to be evaluated by statistical indicators such as R-squared, adjusted R-squared, standard error, etc.

2.2.3. Fully Modified Ordinary Least Squares (FMOLS) regression model

FMOLS is an econometric method used to address endogeneity, heteroskedasticity and autocorrelation in data. Compared to the OLS method, FMOLS introduces a correction term in the estimation of the model to deal with possible problems in the data, resulting in more robust estimates [13].

The FMOLS algorithm adopted in this study is an improvement based on OLS algorithm to simulate achieving the effect of the FMOLS algorithm. Specifically, FMOLS introduces lagged terms to facilitate solving the endogeneity relationship between the dependent and explanatory variables. Also, it corrects heteroskedasticity and autocorrelation by estimating the coefficients through the

DOI: 10.54254/2755-2721/47/20241101

introduction of correction terms. Meanwhile, FMOLS typically employs differencing variables to address the unit root issue, thus enhancing the credibility of the variable estimations.

3. Results and discussion

3.1. Data results

Table 1. Correlation Coefficient Between GDP and Other Indicators with p-value.

Indicators	Correlation Coefficient	p-value
Population Density	0.006	0.946
Land Area (Km2)	0.545	$1.026*10^{-10}$
Agricultural Land (%)	0.042	0.646
CPI	-0.062	0.499
Birth Rate	-0.206	0.023
CO2-Emissions	0.917	$2.670*10^{-49}$
Armed Forces size	0.637	$4.257*10^{-15}$
Forested Area (%)	0.020	0.830
Gasoline Price	-0.025	0.787
Gross tertiary education enrollment (%)	0.228	0.012
Infant mortality	-0.171	0.060
Life expectancy	0.194	0.033
Minimum wage	0.217	0.017
Out of pocket health expenditure	-0.147	0.109
Physicians per thousand	0.097	0.290
Population	0.626	$1.547*10^{-14}$

DOI: 10.54254/2755-2721/47/20241101

Table 1. (continued).

Population: Labor force participation (%)	0.018	0.847
Tax revenue (%)	-0.130	0.155
Total tax rate	0.097	0.289
Unemployment rate	0.047	0.612
Urban_population	0.781	$4.355*10^{-26}$
Agricultural Land (Km2)	0.746	$8.937*10^{-23}$
Forested Area (Km2)	0.385	$1.296*10^{-05}$
Armed Forces per capita	-0.050	0.586
Urban_population ratio	0.162	0.076
CO2-Emissions per capita	0.256	0.005
Labour Force	0.686	$3.968*10^{-18}$

Table 2. Correlation Coefficient Between GDP per capita and Other Indicators with p-value.

Indicators	Correlation Coefficient	p-value
Population Density	0.151	0.099
Land Area (Km2)	0.165	0.071
Out of pocket health expenditure	-0.456	$1.439*10^{-07}$
Agricultural Land (%)	-0.033	0.719
CPI	-0.201	0.027
Birth Rate	-0.532	$3.297*10^{-10}$
CO2-Emissions	0.167	0.067
Armed Forces size	0.022	0.813
Forested Area (%)	-0.007	0.943
Gross tertiary education enrollment (%)	0.539	$1.876*10^{-10}$
Population: Labor force participation (%)	-0.033	0.716
Tax revenue (%)	0.276	0.002
Minimum wage	0.915	$7.703*10^{-49}$

DOI: 10.54254/2755-2721/47/20241101

Table 2. (continued).

Labour Force	-0.004	0.964
Life expectancy	0.621	$3.007*10^{-14}$
Population	-0.012	0.897
Urban_population ratio	0.552	$5.348*10^{-11}$
Gasoline Price	0.326	$2.670*10^{-4}$
Total tax rate	-0.147	0.108
Unemployment rate	-0.069	0.454
Urban_population	0.061	0.503
Agricultural Land (Km2)	0.162	0.076
Infant mortality	-0.522	$8.404*10^{-10}$
Forested Area (Km2)	0.114	0.213
Armed Forces per capita	0.044	0.635
Physicians per thousand	0.520	$1.004*10^{-09}$
CO2-Emissions per capita	0.611	$9.566*10^{-14}$

Table 3. OLS Regression Results of GDP and Selected Indicators.

Indicators	Coefficient	Standard Error	t-value	P>|t|
constant	$2.325*10^{11}$	$1.640*10^{12}$	0.142	0.887
Land Area (Km2)	$-5.831*10^{5}$	$1.490*10^{5}$	-3.920	0.000
Armed Forces size	$-1.991*10^{5}$	$5.380*10^{5}$	-0.370	0.712
Birth Rate	$1.035*10^{10}$	$1.430*10^{10}$	0.723	0.472
CO2-Emissions	$3.716*10^{6}$	$2.520*10^{5}$	14.749	0.000
Gross tertiary education enrollment (%)	$-5.170*10^{11}$	$4.000*10^{11}$	-1.291	0.199
Life expectancy	$-3.570*10^{9}$	$2.020*10^{10}$	-0.176	0.860
Minimum wage	$1.037*10^{11}$	$2.860*10^{10}$	3.620	0.000
Population	$1.910*10^{4}$	3947.079	4.838	0.000
Urban_population	$1.872*10^{4}$	6704.564	2.792	0.006
Agricultural Land (Km2)	$8.024*10^{5}$	$2.280*10^{5}$	3.516	0.001
Forested Area (Km2)	$7.688*10^{5}$	$2.690*10^{5}$	2.859	0.005
Labour Force	$-6.062*10^{4}$	8229.871	-7.365	0.000
R2	Adjusted R2	F-statistic	AIC	BIC
0.941	0.935	144.200	6954	6990

DOI: 10.54254/2755-2721/47/20241101

Table 4. OLS Regression Results of GDP per capita and Selected Indicators.

Indicators	Coefficient	Standard Error	t-value	P>|t|
constant	$1.642*10^4$	$2.400*10^4$	0.685	0.495
Birth Rate	-152.852	141.913	-1.077	0.284
CPI	-2.623	5.133	-0.511	0.610
Gasoline Price	1098.009	2706.036	0.406	0.686
Gross tertiary education enrollment (%)	-7466.882	4000.638	-1.866	0.065
Infant mortality	-44.257	106.121	-0.417	0.677
Life expectancy	-151.190	298.989	-0.506	0.614
Minimum wage	4698.402	340.980	13.779	0.000
Out of pocket health expenditure	-1163.880	4103.542	-0.284	0.777
Physicians per thousand	632.625	726.349	0.871	0.386
Tax revenue (%)	$-1.072*10^4$	$1.170*10^4$	-0.920	0.360
Urban_population ratio	2347.668	4106.697	0.572	0.569
CO2-Emissions per capita	$6.013*10^5$	$1.910*10^5$	3.140	0.002
R2	Adjusted R2	F-statistic	AIC	BIC
0.874	0.860	62.280	2485	2522

Table 5. FMOLS Regression Results of GDP and Fewer Indicators.

Indicators	Coefficient	Standard Error	t-value	P>|t|
constant	$2.794*10^{11}$	$4.830*10^{10}$	5.782	0.000
Land Area (Km2)	$-2.038*10^5$	$7.680*10^4$	-2.653	0.009
Armed Forces size	$3.606*10^5$	$2.690*10^5$	1.342	0.182
CO2-Emissions	$1.383*10^6$	$1.780*10^5$	7.761	0.000
Minimum wage	$8.693*10^{10}$	$1.160*10^{10}$	7.487	0.000
Population	6511.362	2110.631	3.085	0.003
Urban_population	9888.097	3378.125	2.927	0.004
Agricultural Land (Km2)	$2.962*10^5$	$1.180*10^5$	2.508	0.014
Forested Area (Km2)	$2.363*10^5$	$1.360*10^5$	1.734	0.086
Labour Force	$-2.541*10^4$	4523.667	-5.616	0.000
Residuals	0.631	0.034	18.550	0.000
R2	Adjusted R2	F-statistic	AIC	BIC
0.985	0.983	698.900	6789	6819
Mean Squared Error of Predict GDP			R2 of Predict GDP	
$2.602*10^5$			0.724	

DOI: 10.54254/2755-2721/47/20241101

Table 6. FMOLS Regression Results of GDP per capita and Fewer Indicators.

Indicators	Coefficient	Standard Error	t-value	P>|t|
constant	-16.280	805.175	-0.020	0.984
Gross tertiary education enrollment (%))	3223.321	2893.050	1.114	0.268
Minimum wage	3613.235	230.626	15.667	0.000
Physicians per thousand	913.873	487.120	1.876	0.063
Co2-Emissions per capita	$3.847*10^5$	$1.220*10^5$	3.146	0.002
Residuals	0.436	0.051	8.603	0.000
R2	Adjusted R2	F-statistic	AIC	BIC
0.921	0.917	267.000	2415	2432
Mean Squared Error of Predict GDP			R2 of Predict GDP	
$1.385*10^6$			0.859	

3.2. Discussion

From Table 1, considering the values of correlation coefficients, and comparing the p-value with a 0.05 threshold, it shows that “Land Area”, “Armed Forces size”, “Birth Rate”, “CO2 Emissions”, “Gross tertiary education enrollment”, “Life expectancy”, “Minimum wage”, “Population”, “Urban population”, “Agricultural Land”, “Forested Area”, and “Labour Force”, these features have high correlations with the “GDP”. Thus, they will be used to build a linear model with GDP.

Similarly, from Table 2, it reflects that indicators including “Birth Rate”, “CPI”, “Gasoline Price”, “Minimum wage”, “Infant mortality”, “CO2 Emissions per capit”,”Physicians per thousand”,”Life expectancy”, “Gross tertiary education enrollment”, “Tax revenue”, “Out of pocket health expenditure”and “Urban population ratio”, are having high correlations with the “GDP per capita”, and would be used to build linear model with GDP per capita.

According to the models fitted by the OLS approach (show in Table 3 and Table 4), the problem is the same. That is, although most of these selected variables show a good linear relationship with GDP or GDP per capita (since the majority p-values are relatively small), there are also some variables demonstrate a small weight in the model, such as “Birth Rate” and “GDP”, or “Life expectancy” and “GDP per capita”. Their p-values are generally large, and do not show a high degree of linear correlation.

From Table 5 and Table 6, by further excluding less relevant variables and using the optimized FMOLS method, the performance of the two linear models became better. The value of R2 has increased (from 0.941 to 0.985, and 0.874 to 0.921) which means the new models can explain more of the variance. Since the new models have smaller values of Bayesian Information Criterion (BIC) and Akaike Information Criterion(AIC), these reflect that they are more interpretable. The standard errors in the two new models are all smaller than before, so that the estimated coefficients seem to be more reliable. And this time, the p-values are also smaller, thus the coefficients are all significant.

Meanwhile, after training and testing the original data with linear regression based on the new model, it yields small Mean Square Error (MSE), indicating that the model has a lower prediction error and is closer to the actual data. While the value of R2 is close to 1, reflecting that the model is able to explain most of the variance in the target variable.

4. Conclusion

This study was designed to explore the comprehensive impact of natural and societal factors such as energy, land area, education, and employment to economic indicators. In this investigation, the correlation coefficient and the corresponding p-value are used to judge the degree of correlation

DOI: 10.54254/2755-2721/47/20241101

between the variables and economic indicators, and screen the variables needed to build the model. The OLS method is used to build preliminary linear models of GDP and GDP per capita with other influencing factors, while the FMOLS method optimizes the performance of the models. The results of this investigation show that factors which have a relatively large impact on GDP include the land area, size of army, emissions of CO2, minimum wage, population, and labour force. And the minimum wage for residents, number of doctors per thousand, CO2 emissions per capita and the total enrolment rate in higher education, these factors seem to have a significant influence of GDP per capita. This demonstrates the need for countries to focus on the impact of these factors when developing their economies. However, being limited to scope of the data set, this study lacks a comprehensive view of long-term economic development for these countries. Meanwhile, only some possible factors given in the dataset that may have an impact on economic development were considered, and the methods used can only express linear relationships, while the current situation would be much more complex than this. Thus, future research needs to analyze more comprehensive, time-spanning data by using more non-linear models.

References

[1] Dabla-Norris M E and Kochhar M K and Suphaphiphat M N and Ricka M F and Tsounta M E 2015 Causes and consequences of income inequality: A global perspective (IMF Staff Discussion Note No. 15/13) International Monetary Fund
[2] Easterlin R A 2003 Explaining happiness Proceedings of the National Academy of Sciences 100(19) p 11176-11183
[3] Acemoglu D and Robinson J A 2013 Why nations fail: The origins of power, prosperity, and poverty Currency
[4] Hausmann R and Hidalgo C A 2011 The network structure of economic output Journal of Economic Growth 16(4) p 309-342
[5] McKenzie D and Theoharides C and Yang D 2014 Distortions in the international migrant labor market: evidence from Filipino migration and wage responses to destination country economic shocks American Economic Journal: Applied Economics 6(2) p 49-75
[6] Usman M and Jahanger A and Makhdum M S A and Balsalobre-Lorente D and Bashir A 2022 How do financial development energy consumption natural resources and globalization affect Arctic countries' economic growth and environmental quality An advanced panel data simulation Energy 24 p 122515
[7] Daiva Dumciuviene 2015 The Impact of Education Policy to Country Economic Development Procedia - Social and Behavioral Sciences 191 p 2427-2436
[8] Kaggle 2023 Countries of world https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023
[9] Analytics Vidhya 2021 Beginner's Guide to Pearson's Correlation Coefficient https://www.analyticsvidhya.com/blog/2021/01/beginners-guide-to-pearsons-correlation-coefficient/
[10] Cohen I and Huang Y and Chen J and Benesty J 2009 Pearson correlation coefficient Noise reduction in speech processing p 1-4
[11] Thisted R A 1998 What is a P-value Departments of Statistics and Health Studies
[12] Craven B D and Islam S M N 2011 Ordinary least-squares regression The SAGE dictionary of quantitative management research p 224-228
[13] Khan M W A Panigrahi S K Almuniri K S N Soomro M I Mirjat N H and Alqaydi E S 2019 Investigating the dynamic impact of CO2 emissions and economic growth on renewable energy production: Evidence from FMOLS and DOLS tests Processes 7(8) p 496