Regression model-based CO2 emission prediction and feature importance analysis between developed and developing countries

Zexiang Li

Faculty of Science, Harbin Institute of Technology, Weihai, Shandong, 264200, China

2200600216@stu.hit.edu.cn

Abstract. With the increasing focus on greenhouse gases, each country is taking its own measures to reduce the release of carbon dioxide (CO2). In order to discuss the CO2 release in developed and developing countries, this paper establishes a regression model by considering gross domestic product (GDP), population, and energy, and obtains prediction equations for developed and developing countries and discusses the influence of each factor on CO2 release. The results are that for developed countries, the impact sort in descending importance is year, population, energy per GDP and GDP, and the impacts of Energy per GDP and population are negative, while the others are positive. For developing countries, it is almost the opposite, with the impact sort in descending importance is GDP, energy per GDP, year, and population, and only Year is negative, while the others are positive. So developed countries need to improve in other areas not discussed, and continue to maintain the economic, energy and population quality dimensions. The developed countries need to improve in other areas that are not discussed, in the economic, energy and population quality levels. And developing countries need to work on these three aspects.

Keywords: machine learning, CO2 emission estimation, regression model.

1. Introduction

Over 150 countries and the European Economic Community signed the United Nations Framework Convention on Climate Change on March 21, 1994, and it was implemented on that day at the global summit [1]. The purpose of the UN Convention is to reduce the concentration of greenhouse gases, stabilize moisture at a level that protects the environment and prevents danger to the Earth's climate system. Because of the principle of "common but differentiated responsibilities", the obligations and implementation procedures of developed and developing countries are distinguished by the convention. Because developed countries are the major emitters of greenhouse gases, the Convention requires them to take specific measures to reduce greenhouse gas emissions and to provide financial resources to developing countries to ensure that they can travel to their obligations under the Convention [2,3]. On the other hand, developing countries also have obligations that need to be assumed by providing national inventories of GHG sources and sinks and by developing and implementing programs that incorporate GHG source and sink measures [4]. On this basis, developed and developing countries have taken more and more measures to reduce carbon emissions, such as the National Plan to Address Climate Change of U.S. and the U.S. First Energy Plan Energy Basic Plan and 2014 Energy White Paper of Japan, and Action Plan and Second National Communication of India. But measures need to be in the right direction. Countries need to know that they are not doing a good job in order to reduce the release of CO2 most quickly without affecting economic development. At the same time, people need to know whether developing countries and developed countries have effectively fulfilled their responsibilities and whether the CO2 emission reduction work has been done well. In today's work, most people only study the CO2 emissions of a country without comparison, so there is a lack of awareness between developed and developing countries about whether they are doing what they should do [5]. By making a comparison, it could be noticed how different countries have different environmental protection efforts relative to each other. For examples, if the contribution of developing countries' GDP to CO2 emissions is smaller or closer to that of developed countries, it means that developing countries are more fulfilling their responsibilities than developed countries [6]. On the contrary, it shows that developing countries have not fulfilled their responsibilities.

Therefore, this paper establishes a regression model, analyses each influencing factor, and mathematically converts the size of the influencing factor into a number for comparison. And the release of CO2 in the future can be predicted.

2. Method

2.1. Dataset

Data used in the paper is gotten on the website, Our World in Data. Raw data is complex and incomplete. It has eighteen features, but most of them do not have relationship with experiment because some of them are about capital but the aim of the experiment is to deserve CO2 produced by countries [7]. And Some of them lack too much data to complete. In case of those, the experiment chooses Australia and Zimbabwe as examples with some of features to compare developed countries with developing countries. These features are shown in Table 1.

Table 1. Features leveraged in the work.

Treasures	Meanings
population	Population of each country or region.
GDP	GDP is calculated in international dollars, using 2011 prices to adjust for price changes over time (inflation) and price differences between countries. It is calculated by multiplying the GDP per capital by the population.
Year	Year of observation.
Energy per GDP	Primary energy consumption per unit of GDP, in kilowatt-hours per international dollar.
CO2	Total annual emissions of carbon dioxide (CO_2) based on production, excluding land use change, in millions of tons. This is based on geographic emissions and does not take into account emissions from traded goods.

And for each feature, they are chosen by different reasons. Because everyone in his whole life produces CO2, it is necessary that population should be chosen. And GPD can represent the production value of the region or country to a certain extent and most of production may releases CO2, so it is important to consider this feature. After that, because every country is changing, there should be a feature to be chosen. Besides that, energy is an important treasure because fossil fuels are used in all most of countries. But GDP may impact energy they consume, so energy per GDP is chosen at last. Finally, CO2 is the research subjects so it is sure to be considered.

Because the four features have different magnitudes and different orders of magnitude, in order to avoid input variables not being used equally and to reduce the complexity of the data, each feature was normalized by equation (1) before the data was used in the experiment, where x is the value of input feature, and min and max denoted its minimal and maximal values.

$$x = \frac{x-min}{max-min} \tag{1}$$

2.2. Linear and ridge regression

Linear regression analysis is a statistical analysis method to determine the interdependence between variables [8]. if 2 or more independent variables are included and the linear relationship between the dependent and independent variables is satisfied, it is called multiple linear regression analysis. the multiple linear regression model is generally expressed as formula (2)

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \varepsilon \tag{2}$$

In equation (2), Y is the dependent variable; X_i is the independent variable; β_i is the regression coefficient. In multiple linear regression analysis, the least squares method is generally used for parameter estimation, and its regression coefficient β can be solved by matrix method through equation (3).

$$\beta = (X^T X)^{-1} \times (X^T Y) \tag{3}$$

However, one of the prerequisites for using multiple linear regression analysis is that there should not be significant correlation between the independent variables X1, X2, ..., Xn. If there is exact correlation or high correlation between the independent variables, it will have a serious impact on the estimation of the regression parameters, making the estimated values very unstable, resulting in poor fit of the regression model and inaccurate model estimation, which is the existence of multicollinearity. If there is multicollinearity, one of the solutions is to use ridge regression analysis instead.

Ridge regression is a modified least squares estimation method. It makes the obtained regression coefficients more realistic and reliable by dropping the unbiased nature of the least squares method, losing some information, and reducing accuracy [9,10]. This is done by adding a small perturbation to the original least squares estimate to make the problem stable and solvable. This can be expressed as Equation (4).

$$\beta = (X^T X + KI)^{-1} \times (X^T Y) \tag{4}$$

In Equation (4), K is the ridge parameter; I is the unit matrix. For ridge regression, when the ridge parameter K varies within 0 to infinity, the ridge regression coefficient β is a function of K. β is a vector consisting of multiple components such as $\beta_1, \beta_2, \dots, \beta_n$ each of which is a function of K. The curve drawn for each component is called a ridge trace. if the instability of the ridge trace is strong, it means that the least squares method cannot reflect the true situation well at this time, and an appropriate value of the ridge parameter K needs to be chosen according to the ridge trace to determine the regression coefficient β .

3. Result

3.1. Experimental setting

The paper will get a ridge trace to determine K-value when the standardized regression coefficients of each independent variable tend to be stable. Generally, the smaller K-value is, the smaller the deviation is. Through analysing F-value, the paper can get whether the model can be meaningful. If the number is significant, there is a regression relationship between the independent and dependent variables. After that the experiment will analyze R squared to judging the model fit. If the number is close to 1, the fit will be well. Then, the significance of the independent variables will be analyzed. If it is significant, it can be used to detect the influence of independent variables to dependent variable. With regression coefficient, the paper will contrast the degree of influence of different independent variables on the dependent variable. At last, the formula of model can be achieved.

3.2. Results of developed country: Australia

Figure 1 visualizes the relationship between the standardized regression coefficients and K-value. K is a parameter K∈[0,1] of the ridge regression, when K=0, the experimental method becomes least squares estimation. The larger k is, the better the effect of eliminating covariance, but it will lead to a greater reduction in the accuracy of the fit. So, when the value of K is chosen, the smallest value of K that makes the standardized regression coefficients of each independent variable tend to be stable should be chosen. This paper determines K=0.01 by the variance expansion factor method.

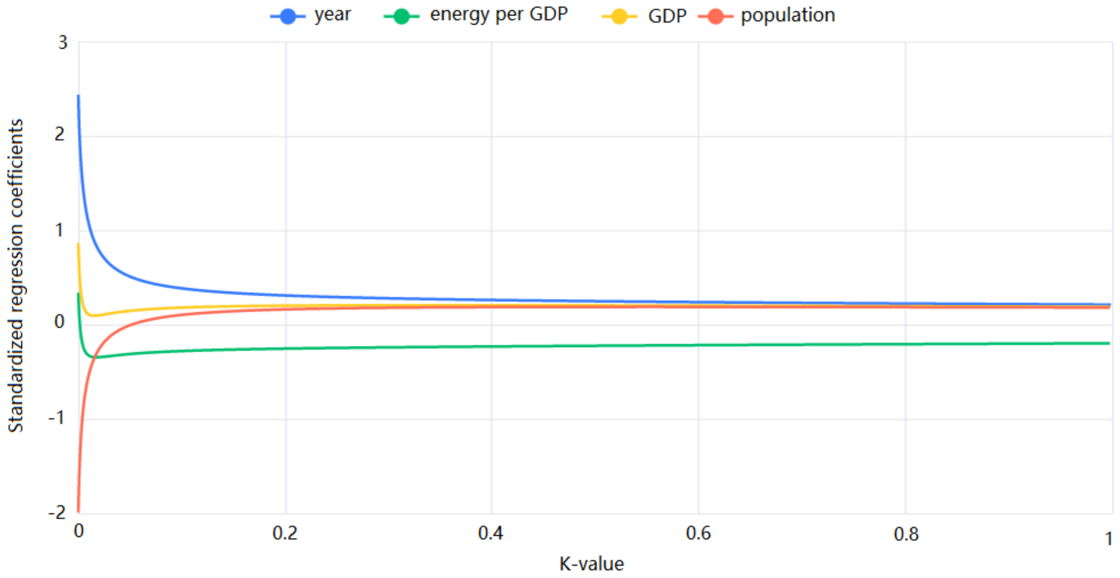

Figure 1. Relationship between the standardized regression coefficients and K-value of Australia (Figure credit: Original).

After the K-value is chosen, the standardized regression coefficients can be chosen. The weights of various model paths, including year, energy per GDP, GDP, and population are 1.077, -0.329, 0.105, and -0.544 respectively.

The number on this figure is standardization factor. So, the degree of the effect of each independent variable on the dependent variable can be obtained in order: year, population, energy per GDP, and GDP.

In addition to that, there are the visualization results of the model fit values on Figure 2. And the model formula is CO2=0.371+1.238*year-0.38*energy_per_GDP+0.114*GDP-0.65*population.

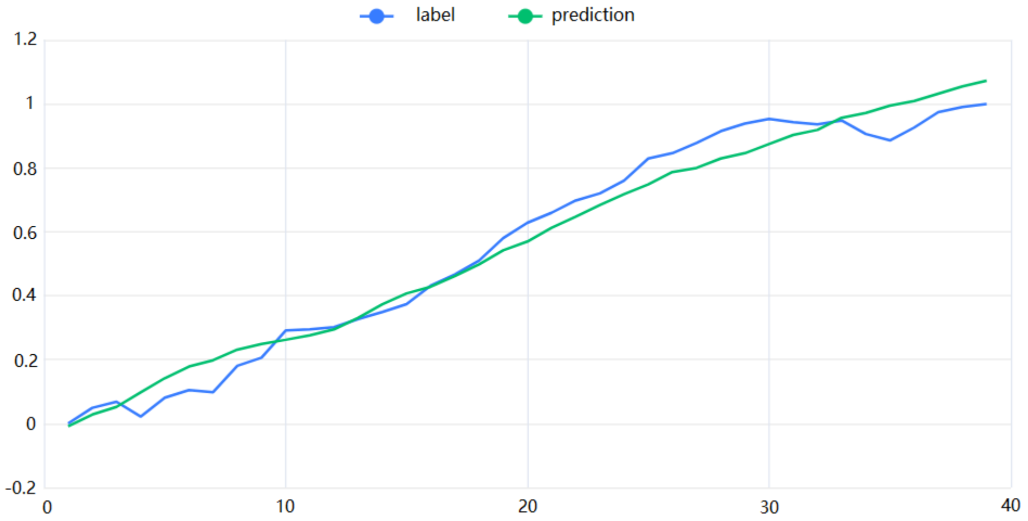

Figure 2. Prediction result of Australia (Figure credit: Original).

Table 1 shows the results of the parameters of this model and the test results, including the standardized coefficients of the model, t-values, the results of the F-test, R², and adjusted R², which are used to test the model and analyze the formula of the model.

Table 1. Model analysis result of Australia, where *** represents 10% significance levels.

K=0.01	Non-standardized coefficient		t	P	R²	Adjusted R²	F
	B	Standard error					
Constants	0.371	0.128	2.896	0.007***	0.973	0.97	307.302 (0.000***)
year	1.238	0.135	9.178	0.000***			
energy	-0.38	0.136	-2.804	0.008***			
gdp	0.114	0.125	0.91	0.369			
population	-0.65	0.137	-4.733	0.000***			

The results of the ridge regression showed that the p-value based on the F significance test is 0.000***, showing significance at the level, so the experiment rejected the original hypothesis, which indicated that there was a regression relationship between the independent and dependent variables. Besides that, the goodness of fit of the model R² was 0.973 so the model performed excellent.

3.3. Results of developing country: Zimbabwe

Figure 3 also visualizes the relationship between the standardized regression coefficients and K-value. Therefore, this paper determines K=0.147 by the variance expansion factor method.

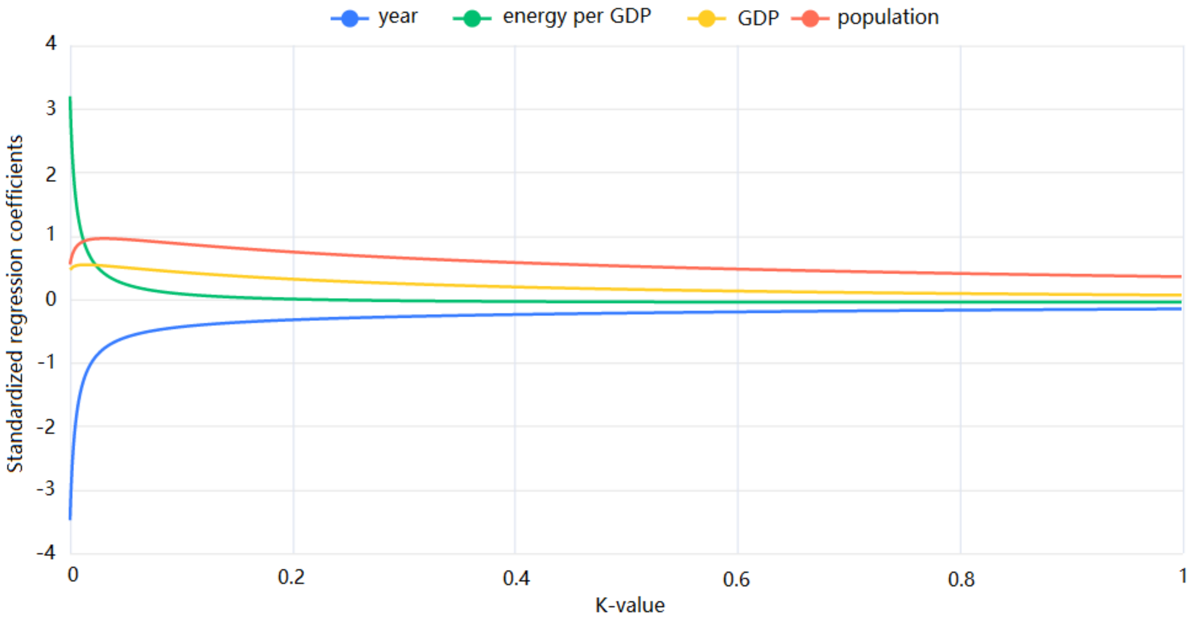

Figure 3. Relationship between the standardized regression coefficients and K-value of Zimbabwe (Figure credit: Original).

Standardization factors of year, population, energy per GDP, and GDP are respectively -0.367, 0.033, 0.37, 0.809. Therefore, the degree of the effect of each independent variable on the dependent variable can be obtained in order: GDP, energy per GDP, year, population.

Also, there are the visualization results of the model fit values in Figure 4. And the model formula is CO2=0.016-0.345*year+0.391*energy_per_GDP+0.847*GDP+0.036*population.

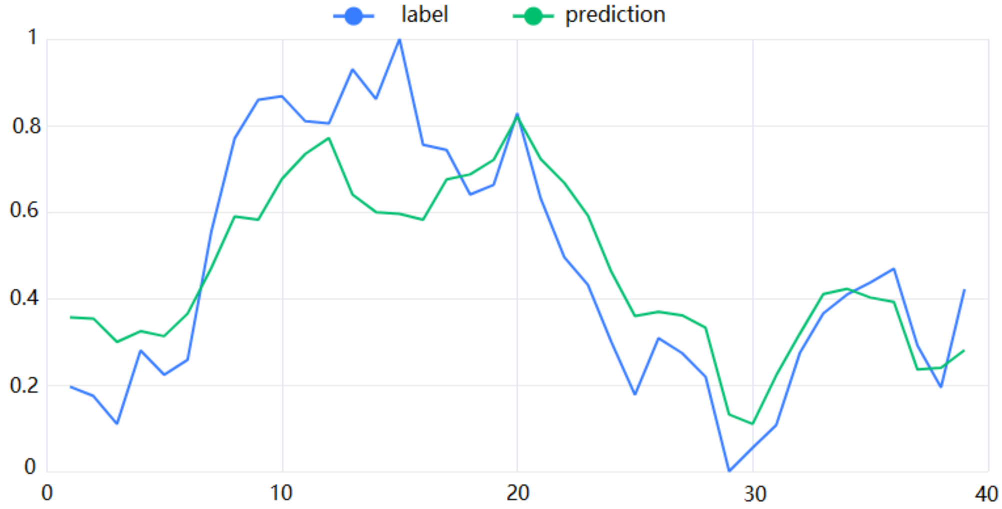

Figure 4. Prediction result of Zimbabwe (Figure credit: Original).

Table 2 also shows the results of the parameters of this model and the test results. Because the p-value based on the F significance test is 0.000***, meaning significance at the level, the experiment rejected the original hypothesis. Therefore, there was a regression relationship between the independent and dependent variables. Besides of that, the goodness of fit of the model R² was 0.179 so the model performed well.

Table 2. Model analysis result of Zimbabwe, where *** represents 10% significance levels.

K=0.01	Non-standardized coefficient		t	P	R^2	Adjusted R^2	F
	B	Standard error					
Constants	0.016	0.095	0.174	0.863	0.719	0.686	21.8 (0.000***)
year	-0.345	0.052	-6.674	0.000***			
energy	0.036	0.057	0.633	0.531			
gdp	0.391	0.095	4.127	0.000***			
population	0.847	0.094	9.029	0.000***			

4. Discussion

Based on the results above, it can be found that for developed countries, energy per GDP and GDP have less effect on the release of CO2. That means that the economic structure of developed countries tends to be stable and the share of environmentally friendly industries is stable, and the effect of energy per GDP on the release of CO2 is negative, so it is reasonable to think that developed countries do not overuse fossil fuels when promoting economic development. In addition to this, the population size of developed countries also has a negative effect on the release of CO2, which means that the more people in developed countries under the same conditions, the less CO2 is released, so it can be thought that the average environmental awareness of people in developed countries is good. But in this case, when the conditions are constant, the release of CO2 still increases with the year, which proves that developed countries are not doing well enough in other factors which the paper is not discussed.

But in developing countries, on the contrary, energy per GDP and GDP have a greater effect on the release of CO2, so the economic structure of developing countries and the use of fossil fuels for economic development play a bad role in reducing the release of CO2. At the same time, population size has a positive effect on the release of CO2, which means that the more people in developing countries, the more CO2 may be released, so developing countries still need to raise awareness of environmental protection. But the release of CO2 decreases with the year when conditions are constant, so developing countries do a good enough job on other factors which the paper is not discussed.

From the discussion, it becomes clear that although the simulation results are more consistent, there are influences that are not included in the model when discussing the influences, and that these influences have a so large effect on the CO2 release that the CO2 release changes with time in the opposite direction to that with the important factors. So, the model still needs to extend the range of influencing factors. At the same time, although this model is discussing developed and developing countries, the data used in the model are only from Australia and Zimbabwe. This inevitably leads to the correctness of the discussed results.

5. Conclusion

In this paper, the author obtained data from the website, Our World in Data, for Australia and Zimbabwe, and modelled the CO2 release for developed and developing countries by ridge regression method, and obtained the regression equation for CO2 release and discussed the influencing factors. The influencing factors are ranked according to the influence under the large influence, for developed countries the importance ranking is year, population, energy per GDP, and GDP, and the influence of Year and GDP is positive, while the influence of population and energy per GDP is negative. For developing countries, the ranking orders is GDP, energy per GDP, year, population, and the effect of year is negative while the effect of other influences is positive. And after discussing each influence, in the future, if the independent variables can be added, it may be possible to obtain what factors make the trend of CO2 release over time opposite to the trend of the important factors. This can be used as a basis for how national policies can be properly adjusted to reduce CO2 emissions. Even governments can improve the environment without affecting economic development.

References

[1] Bodansky, D. (1993). The United Nations framework convention on climate change: a commentary. Yale J. Int'l l., 18, 451.

[2] Carattini, S., Gosnell, G., & Tavoni, A. (2020). How developed countries can learn from developing countries to tackle climate change. World Development, 127, 104829.

[3] Hakimi, A., & Inglesi-Lotz, R. (2020). Examining the differences in the impact of climate change on innovation between developed and developing countries: evidence from a panel system GMM analysis. Applied Economics, 52(22), 2353-2365.

[4] Tan, X., Zhu, K., Meng, X., Gu, B., Wang, Y., Meng, F., ... & Li, H. (2021). Research on the status and priority needs of developing countries to address climate change. Journal of Cleaner Production, 289, 125669.

[5] Ağbulut, Ü. (2022). Forecasting of transportation-related energy demand and CO2 emissions in Turkey with different machine learning algorithms. Sustainable Production and Consumption, 29, 141-157.

[6] Mądziel, M., Jaworski, A., Kuszewski, H., Woś, P., Campisi, T., & Lew, K. (2021). The development of CO2 instantaneous emission model of full hybrid vehicle with the use of machine learning techniques. Energies, 15(1), 142.

[7] Ritchie, H., Roser, M., & Rosado, P. (2020). CO_2 and greenhouse gas emissions. Our world in data. URL: https://ourworldindata.org/co2-emissions?utm_source=tricity%20news&utm_campaign=tricity%20news%3A%20outbound&utm_medium=referral. Last accessed 2023/07/07.

[8] Su, X., Yan, X., & Tsai, C. L. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275-294.

[9] McDonald, G. C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 93-100.

[10] Hoerl, R. W. (2020). Ridge regression: a historical context. Technometrics, 62(4), 420-425.