Project II

Author: Mohammad Zohaib Saeed

Title: Environmental Impacts of Trade Liberalization in Developing Countries using Logistic Regression

1. Overview

In today's modern world, globalization has enabled countries to adopt and exchange various goods and services beyond their borders with ease. One such example is trade, which is more open nowadays than ever because of removal of trade barriers like tariffs, quotas etc. Due to this, countries are now increasing their trade with a goal to accomplish their economic prosperity. These countries do it by increasing their exports and hence, achieve trade surplus; which is one of the key macroeconomic goals. Trade liberalization or economic openness is thus the total or partial removal of any type of international trade barriers either in the form of tariffs or non-tariffs (Lee, 2005). Trade liberalization enables nations to focus on producing commodities and services in which they have a comparative advantage.

As much as trade liberalization promotes economic growth, increases employment opportunities and makes the countries better-off, it also has some repercussions. One of the major concerns is the adverse impacts on the environment and climate change by releasing greenhouse gas emissions (GHGs), primarily carbon dioxide (CO_2). These gases are harmful to the environment and human health in many ways. It also results in climate change that causes volcano eruption, ozone layer depletion, land degradation, sea level rise, floods and global warming. Most importantly, developing countries are facing severe consequences being vulnerable to climate change. Thus, in order to address whether trade liberalization is adversely impacting the environment in developing countries, this study will empirically analyze the causal relationship between CO_2 emissions and trade liberalization.

1.1 Study Objectives

The objectives of the paper are to find the impacts of trade liberalization on the environment in terms of computing the magnitude of trade liberalization on the primary indicator of environmental quality (CO_2).

2. Empirical Analysis

2.1 Data and Variable Description

To compute the relationship between trade liberalization and environment, the study uses CO_2 as the proxy for environmental quality as a dependent variable. Total trade as a percentage of Gross Domestic Product (GDP) has been used for the trade liberalization. Other control variables included in the model are capital (gross fixed capital formation in US dollars), land and labor (labor force participation rate). Therefore, a panel data has been collected from 1990 to 2020 for all the available developing countries (147) from the World Bank database. The list of developing countries has been prepared by the International Monetary Fund (IMF) based on human development index, GNI per capital and total population. The World Bank (2021) designated nations and territories with a GNI of $12,5353 or more as high-income economies. Anything less than that is classified as a developing country.

2.2 Empirical Specification

Before computation, the data was cleaned and organized in Microsoft Excel. All missing values have been replaced by the mean value to avoid understating the results. The econometric model is presented in equations 1 as it was used by Onwachukwa, Yan, and Tu (2021). CO_2 is used as an environmental quality indicator to check the variation and direction of the impact, while i and t represents year and time, respectively. It is expected that a positive estimate would mean that higher trade results in increasing the CO_2 and the negative estimate would mean otherwise. Due to a large size of a control variable (capital), the dependent variables have also been converted into the log form.

$$Y_{it} = \beta_0 + \beta_1 Tr_{it} + \beta_2 K_{it} + \beta_3 Labr_{it} + \beta_4 Land_{it} + \boldsymbol{\varepsilon}_{it} \qquad (1$$

Table 1: Variable Description **-** World Bank Dataset (https://data.worldbank.org)

Variable	Abbreviation	Unit
Carbon dioxide emissions	Y	Kiloton
Trade openness/ total trade	Tr	As a percentage of GDP
Capital	K	Gross fixed capital formation
Land size	Land	Land area as square kilometers
Labor force participation rate	Labr	Percentage of total population age 15+
Error term	$\mathcal{E}$	
Country and time period	*it*	

2.3 Estimation Technique – Logistic Regression

We assume that our dependent variable (carbon dioxide) is now a Y dependent variable that has a value of either 0 or 1. These values have been assigned by computing the mean of carbon dioxide original values. Subsequently, if the original value is greater than the mean, the value will be 1 and 0 otherwise. Here, 0 represents a lower carbon dioxide emission while 1 represents a higher carbon dioxide emission. A higher carbon dioxide emission indicates a bad environmental quality, but a low carbon dioxide emission would indicate a good environmental quality.

2.3.1 Training and Validation Data

The selected data has been divided randomly to 60 percent as training and 40 percent as validation data under the Data Mining tab of XL Miner. This training data will be used to develop a model and at the same time, the validation data will be used to test the training data to get an overall picture of this analysis. Since the data set is large (30 years), the partition is set as 60 and 40 percent respectively.

Partition	No. of Records
Training	2734
Validation	1823

Table 1: Partition Record for Training and Validation Data

3 Results and Discussion

3.1 Regression Summary and Coefficients

Metric	Value
# Iterations Used	5
Residual DF	2729
Residual Deviance	1409.78728
Multiple R2	0.34508544

Table 2: Regression Summary

Predictor	Estimate	Confidence Interval: Lower	Confidence Interval: Upper	Odds	Standard Error	Chi2-Statistic	P-Value
Intercept	-2.31946638	-3.183076896	-1.45585586	0.098326	0.440625708	27.70999831	1.409E-07
Trade	0.00054602	-0.003787575	0.004879613	1.000546	0.002211058	0.060983883	0.8049476
Capital	2.1187E-11	1.66854E-11	2.56886E-11	1	2.29677E-12	85.09506889	2.844E-20
Labor	-0.01730025	-0.029439506	-0.005160999	0.982849	0.006193611	7.80220003	0.0052183
Land	8.4953E-07	6.57914E-07	1.04115E-06	1.000001	9.77662E-08	75.5061166	3.643E-18

Table 3: Coefficients

Interpretation: Overall, the summary of the model shows that the model is not too good due to the residual degrees of freedom and deviances. Multiple R-squared is also just 34 percent, but it might be due to a large panel data for 147 developing countries and there are other factors to be considered as well.

Based on the coefficients, all explanatory variables are significant at 5 percent level of significance except trade, which is statistically significant at 10 percent level of significance. Additionally, all explanatory variables indicates that any increase in the variables will increase carbon emissions except for labor force participation, which will decrease carbon emissions. All things remains constant, if labor force increases by 1 percent, carbon emissions will also decrease by 1 percent and vice versa. We can further explain that labor force might be used to use environment-friendly production processes that helps in improving the environmental quality. This shows that overall model is good for training and validation data for logistic regression.

3.3 Training: Classification Summary

Confusion Matrix

Actual\Predicted	0	1
0	2342	26
1	212	154

Error Report			
Class	# Cases	# Errors	% Error
0	2368	26	1.097972973
1	366	212	57.92349727
Overall	2734	238	8.705193855

Metrics	
Metric	Value
Accuracy (#correct)	2496
Accuracy (%correct)	91.29481
Specificity	0.98902
Sensitivity (Recall)	0.420765
Precision	0.855556
F1 score	0.564103
Success Class	1

Interpretation: According to the probability of cut-off point as 0.5, for misclassification in the training data, the false negative rate is 57.92% and the false positive rate is 1.09%. The specificity (true negative) is 98% and sensitivity (true positive) is 42%. Overall, the misclassification error rate is 8.71%, which means that 238 out of total 2734 observations are mis-classified. Our classification rule correctly classified 91.3% of the observations. Hence, we say that based on these results, the logistic regression model performs much better in classifying low-carbon emissions (class 0) than high-carbon emission (class 1).

3.4 Validation: Classification Summary

Confusion Matrix		
Actual\Predicted	0	1
0	1573	17
1	140	93

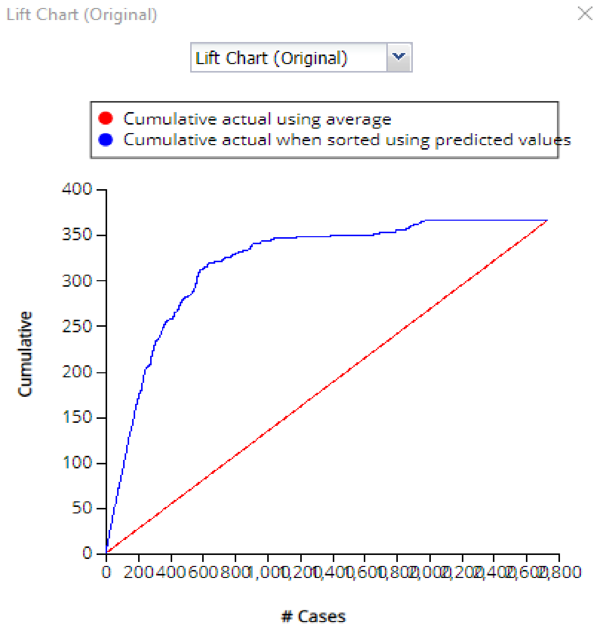
Error Report			
Class	# Cases	# Errors	% Error
0	1590	17	1.06918239
1	233	140	60.08583691
Overall	1823	157	8.612177729

Metrics

Metric	Value
Accuracy (#correct)	1666
Accuracy (%correct)	91.38782
Specificity	0.989308
Sensitivity (Recall)	0.399142
Precision	0.845455
F1 score	0.542274
Success Class	1
Success Probability	0.5

Interpretation: According to the probability of cut-off point as 0.5, for misclassification in the training data, the false negative rate is 60.08% and the false positive rate is 1.06%. The specificity (true negative) is 98% and sensitivity (true positive) is 39%. Overall, the misclassification error rate is 8.61%, which means that 157 out of total 1823 observations are mis-classified. Our classification rule correctly classified 91.3% of the observations. Hence, we say that based on these results, the logistic regression model performs much better in classifying low-carbon emissions (class 0) than high-carbon emission (class 1).

3.5 Lift Chart Comparison between Training and Validation Data



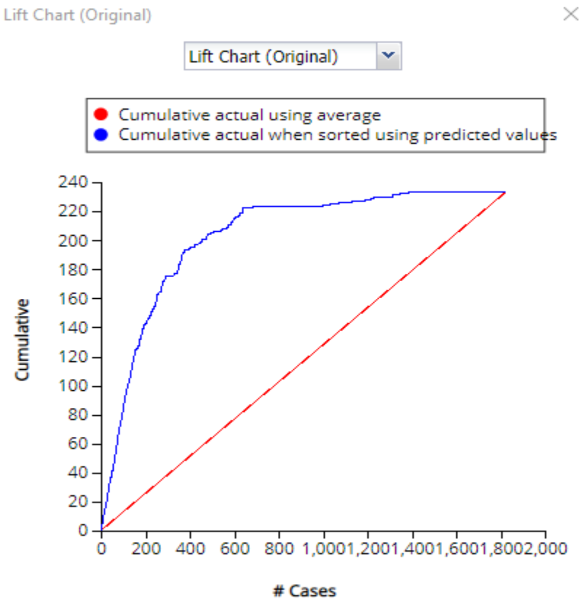


Figure 1: Training Lift Chart

Figure 2: Validation Lift Chart

Interpretation: Based on the error reports from both training and validation data discussed in the previous section. We have 366 cases of "1" out of total 2734 observations and 233 cases of "1" out of total 1823 observation in training and validation data respectively. As seen from the above lift charts, the model performance line in training data shows a slightly higher lift in contrast to the validation data. Therefore, the model performed slightly better in the training phase.

3.6 ROC Curve Comparison between Training and Validation Data

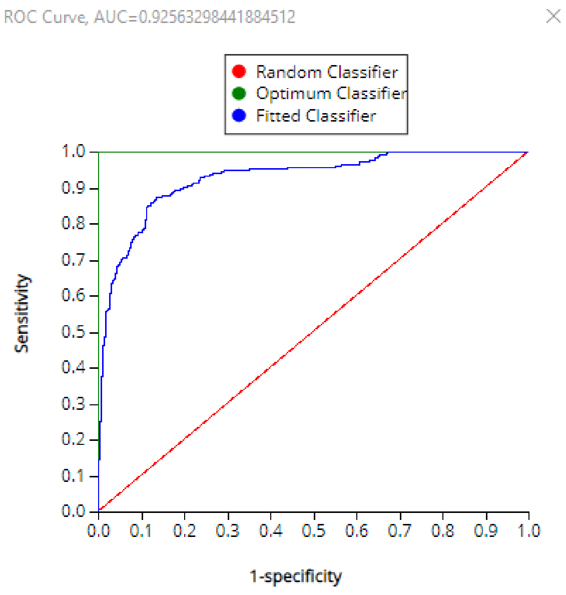


Figure3: ROC Curve for Training Data

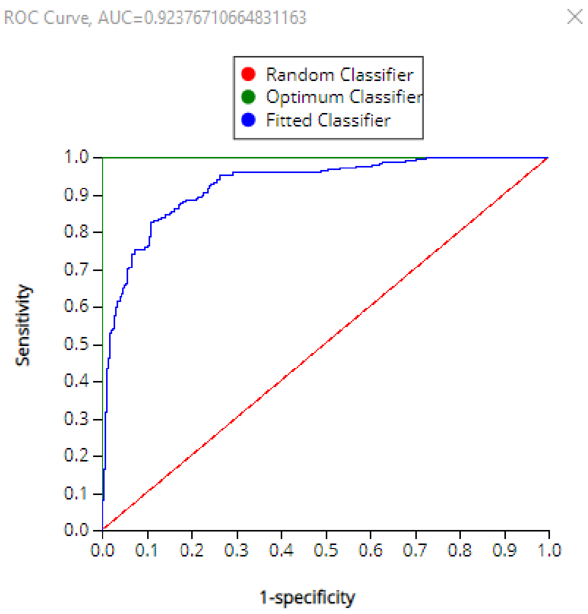


Figure 4: ROC Curve for Validation Data

Interpretation: From the above charts, it is clear that the ROC Curve (area under the curve) for training data is 0.925, which is slightly higher than the AUC of 0.923 of validation data. This shows that the model performs slightly better in the training phase than the validation phase.

Overall, the model shows a consistent performance based on the similar pattern and values in both the lift charts and ROC curves.

4. Conclusion

We have used a logistic regression to test the performance of the model for training and validation data. For this, we created a 60:40 partition for training and validation data. Based on the results, we saw that all the selected explanatory variables were significant at 5 and 10 percent level of significance as per the regression summary. According to the training and validation classification summary results, we found out that false positive, false negative and overall model acccuracy for the training data is slightly better than the validation data. We also witnessed that from the lift charts and ROC curve figures. Thus, the logistic regression performs much better in classifying low scale (class 0) – where the carbon emissions are low than the high scale (class 1) – where the carbon emissions are high in both training and validation data. Overall, the dataset performed well as reflected by a 91% accuracy in both training and validation data.

References

Bank, T. W. (2021). *New World Bank country classifications by income level: 2020-2021*. Retrieved from https://blogs.worldbank.org/opendata/new-world-bank-country-classifications-income-level-2020-2021

Lee, E. (2005). Trade liberalization and employment. *DESA Working Paper No. 5*.

Onwachukwa, C. I., Yan, K.-M. I., & Tu, K. (2021). The Causal Effect of Trade Liberalization on the Environment. *Journal of Cleaner Production*.