Ruoheng Du

Regression and Multivariate Data Analysis

Prof. Jeff Simonoff

May 9, 2023

# Predicting Level of Development for Countries

Economic development is a complicated process that is frequently linked to a variety of benefits, including raised living standards, and less poverty. However, it also has some negative sides, such as increased environmental degradation, increased carbon emissions, and increased public health burdens. Therefore, it is necessary to study the relationship between economic development and healthcare expenditure, energy supply, and carbon dioxide emissions to ensure that the benefits of economic growth are realized sustainably.

The following analysis in the report is based on retrospective data of 30 countries, in which 15 of them are developed countries and another 15 of them are developing countries, from the health expenditure, energy supply per capita, and carbon dioxide emissions per capita statistical table of the UNdata website (http://data.un.org/). The categorization of developed countries versus developing countries is based on World Economic Situation and Prospects 2019 (https://www.un.org/development/desa/dpad/wp-content/uploads/sites/45/WESP2019_BOOK-ANNEX-en.pdf). This report will use level of development (0 for developing countries and 1 for developed countries) as the target variable, health expenditure (% of GDP), energy supply per capita (in ten gigajoules), and emissions per capita (in metric tons of carbon dioxide) as predicting variables in the following logistic regression model.

As for the first predicting variable, health expenditure as a percentage of GDP (i.e., health expenditure in the following analysis), it refers to the extent of which a country spends on

healthcare. This is interesting because it shows the dedication of a country to its healthcare system and the welfare of its population. In comparison to the absolute amount spent on healthcare, which can be misleading due to larger economies often having higher levels of healthcare expenditure without necessarily prioritizing health spending, this indicator provides a more accurate measure.

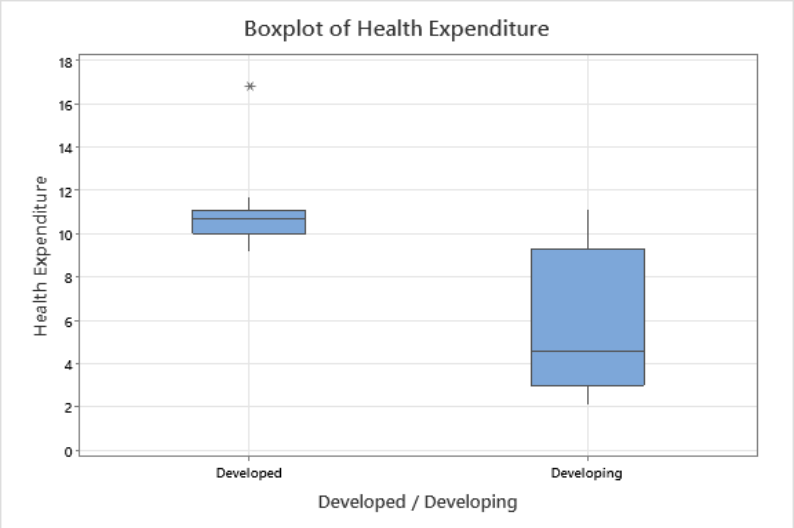
As for the second predicting variable, energy supply per capita in ten gigajoules (i.e., supply per capita in the following analysis), is the amount of energy available for consumption per person in a given country. It takes into account the density of its population, making comparisons of energy production and efficiency between countries more meaningful and providing insights into different levels of economic development for countries.

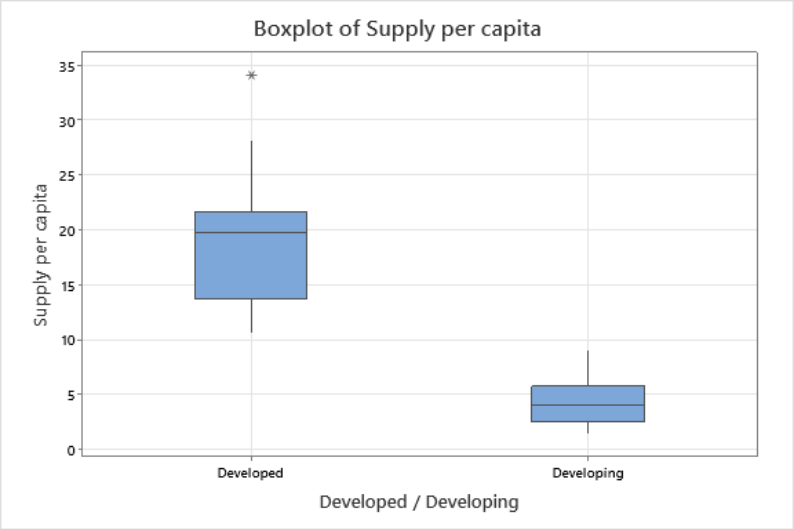
As for the third predicting variable, carbon dioxide emissions per capita in metric tons of carbon dioxide (i.e., emissions per capita in the following analysis), can be defined as the average amount of carbon dioxide emitted by an individual within that country. In general, developed countries would have higher carbon dioxide emissions per capita than developing countries, but it is also noticeable that with more advanced development, developed countries are more likely to have better measures in dealing with it, which makes its relationship with level of development interesting.

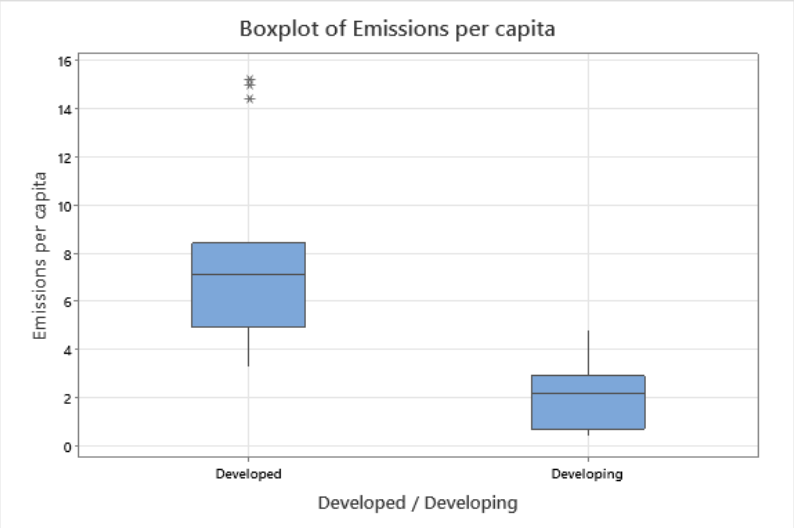
Moreover, the data and categorization standards used in this report are sourced from 2019. This is done with the intention of minimizing the effects of the COVID-19 outbreak and ensuring consistency in the selected variables.

Here is the statistics and the boxplots.

| Variable | Developed / Developing | N | Mean | SE Mean | StDev | Minimum | Q1 | Median | Q3 | Maximum |
|---|---|---|---|---|---|---|---|---|---|---|
| Health Expenditure | Developed | 15 | 10.927 | 0.451 | 1.748 | 9.200 | 10.000 | 10.700 | 11.100 | 16.800 |
| | Developing | 15 | 5.627 | 0.765 | 2.963 | 2.100 | 3.000 | 4.600 | 9.300 | 11.100 |
| Supply per capita | Developed | 15 | 18.89 | 1.71 | 6.64 | 10.60 | 13.70 | 19.80 | 21.60 | 34.20 |
| | Developing | 15 | 4.273 | 0.573 | 2.218 | 1.500 | 2.500 | 4.000 | 5.800 | 9.100 |
| Emissions per capita | Developed | 15 | 7.88 | 1.01 | 3.91 | 3.30 | 4.90 | 7.10 | 8.40 | 15.20 |
| | Developing | 15 | 2.113 | 0.352 | 1.364 | 0.400 | 0.700 | 2.200 | 2.900 | 4.800 |






As we can see, all three variables show clear separation between developed countries and developing countries. With the mean of 10.927, 18.89, and 7.88 respectively, developed countries do certainly have a greater average health expenditure, supply per capita, and emissions per capita. Another distinction is that developed countries have a wider interquartile range of values than developing countries, with the exception of health expenditure.

Here's the logistic regression output with all three variables.

# Binary Logistic Regression: Development versus Health Expenditure, Supply per capita, Emissions per capita

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

* WARNING * The model could not be fit properly. Maximum likelihood estimates of parameters do not exist due to complete separation of data points. The results are not reliable. Please refer to help for more information about complete separation.

## Method

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 30 |

## Response Information

| Variable | Value | Count | |
|---|---|---|---|
| Development | 1 | 15 | (Event) |
| | 0 | 15 | |
| | Total | 30 | |

## Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))

Y' = -81 + 6.0 Health Expenditure + 3.66 Supply per capita - 2.8 Emissions per capita

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -81 | 127 | -0.64 | 0.521 | |
| Health Expenditure | 6.0 | 14.7 | 0.41 | 0.681 | 1.70 |
| Supply per capita | 3.66 | 9.50 | 0.39 | 0.700 | 7.87 |
| Emissions per capita | -2.8 | 15.5 | -0.18 | 0.855 | 6.30 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| Health Expenditure | 422.4281 | (0.0000, 1.40443E+15) |
| Supply per capita | 38.7600 | (0.0000, 4.70876E+09) |
| Emissions per capita | 0.0596 | (0.0000, 8.61676E+11) |

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 99.93% | 92.71% | 8.03 | 9.63 | 13.64 | 1.0000 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 26 | 0.03 | 1.000 |
| Pearson | 26 | 0.02 | 1.000 |
| Hosmer-Lemeshow | 8 | 0.02 | 1.000 |

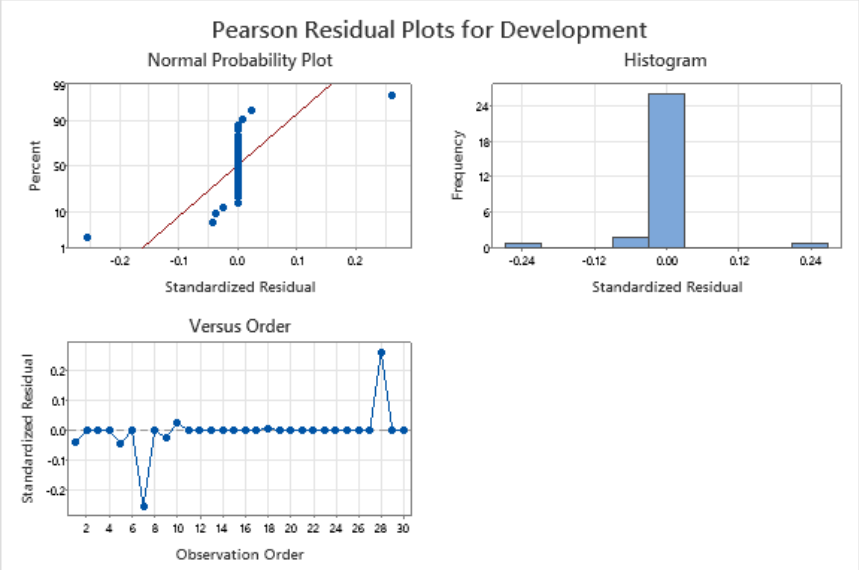
## Analysis of Variance

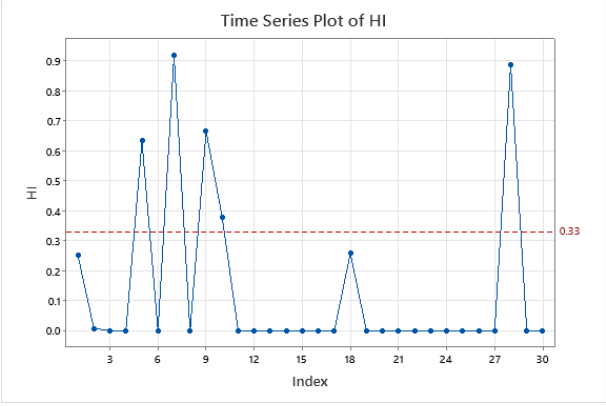
| Source | DF | Adj Dev | Adj Mean | Likelihood Ratio Chi-Square | P-Value |
|---|---|---|---|---|---|
| Regression | 3 | 41.5579 | 13.8526 | 41.56 | 0.000 |
| Health Expenditure | 1 | * | * | * | * |
| Supply per capita | 1 | * | * | * | * |
| Emissions per capita | 1 | * | * | * | * |
| Error | 26 | 0.0309 | 0.0012 | | |
| Total | 29 | 41.5888 | | | |

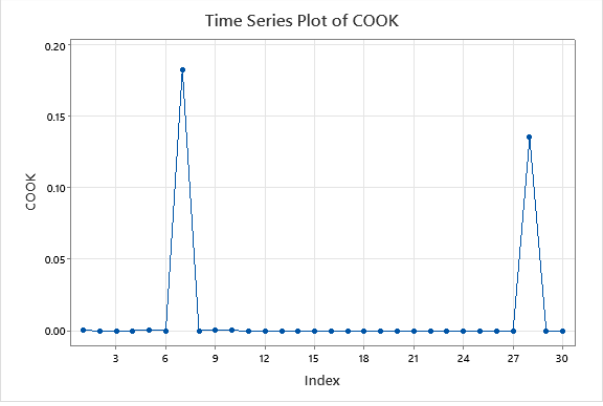
## Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 225 | 100.0 | Somers' D | 1.00 |
| Discordant | 0 | 0.0 | Goodman-Kruskal Gamma | 1.00 |
| Ties | 0 | 0.0 | Kendall's Tau-a | 0.52 |
| Total | 225 | 100.0 | | |

*Association is between the response variable and predicted probabilities*




The complete separation of the data limits the analysis of the variance table. To three decimal places, the p-value for the Likelihood Chi-square statistic of the total regression is

highly statistically significant. As a result, it should be strongly rejected that the slopes of all the three predicting variables are equal to zero. Consequently, despite the variance table's missing data, the three predictors have a high predictive power over the response variable. And the Hosmer-Lemeshow test statistic with a highly statistically significant p-value of 1.00 shows that the data are well fit by the logistic regression model. Additionally, Somers' D and Area under the ROC Curve both exhibit values of 1.00, showing perfect separation. This implies that this model with three predictors is successful in accurately differentiating between developed and developing countries.

As shown in the odds ratios for continuous predictors table above, the health expenditure coefficient says that an increase of one percentage point in the health expenditure is associated with multiplying the odds of being developed countries by 422.4281, holding all other variables in the model fixed; the supply per capita coefficient says that an increase of one ten gigajoules in the supper per capita is associated with multiplying the odds of being developed countries by 38.76, holding all other variables in the model fixed; the emissions per capita coefficient says that an increase of one metric tons of carbon dioxide in the emissions per capita is associated with multiplying the odds of being developed countries by 0.0596, holding all other variables in the model fixed.

Let's try to see if a simpler model would work almost as well.

**Response is Development**

| Vars | R-Sq | R-Sq (adj) | R-Sq (pred) | Mallows Cp | S | Health Expenditure | Supply per capita | Emissions per capita |
|---|---|---|---|---|---|---|---|---|
| 1 | 70.1 | 69.0 | 63.9 | 4.8 | 0.28323 | | X | |
| 1 | 56.0 | 54.4 | 50.5 | 19.3 | 0.34341 | X | | |
| 2 | 74.1 | 72.2 | 65.4 | 2.6 | 0.26801 | X | X | |
| 2 | 70.3 | 68.1 | 62.8 | 6.6 | 0.28721 | | X | X |
| 3 | 74.7 | 71.8 | 64.4 | 4.0 | 0.26990 | X | X | X |

The model selection outputs indicate that we should focus on two logistic regression models for further analysis. The first model includes only the predicting variable of supply per capita, while the second model includes both health expenditure and supply per capita as predictors. These models have adjusted R-squared and predicted R-squared values that are similar to the three-predictor models. Additionally, the one-predictor model has a Mallows Cp value higher than the existing three-predictor model, while the two-predictor model has the lowest value of all the models. However, the best subset result is not conclusive for logistic regression, and as a result, these two models need to be tested further.

**Binary Logistic Regression: Development versus Health Expenditure, Supply per capita**

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

* WARNING * The model could not be fit properly. Maximum likelihood estimates of parameters do not exist due to complete separation of data points. The results are not reliable. Please refer to help for more information about complete separation.

**Method**

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 30 |

**Response Information**

| Variable | Value | Count | |
|---|---|---|---|
| Development | 1 | 15 | (Event) |
| | 0 | 15 | |
| | Total | 30 | |

**Regression Equation**

P(1) = exp(Y')/(1 + exp(Y'))
Y' = -89.9 + 6.7 Health Expenditure + 2.53 Supply per capita

**Coefficients**

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -89.9 | 99.1 | -0.91 | 0.364 | |
| Health Expenditure | 6.7 | 10.6 | 0.63 | 0.531 | 1.11 |
| Supply per capita | 2.53 | 3.52 | 0.72 | 0.472 | 1.11 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| Health Expenditure | 787.8398 | (0.0000, 9.11468E+11) |
| Supply per capita | 12.5704 | (0.0126, 12507.3489) |

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 99.91% | 95.11% | 6.04 | 6.96 | 10.24 | 1.0000 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 27 | 0.04 | 1.000 |
| Pearson | 27 | 0.02 | 1.000 |
| Hosmer-Lemeshow | 8 | 0.02 | 1.000 |

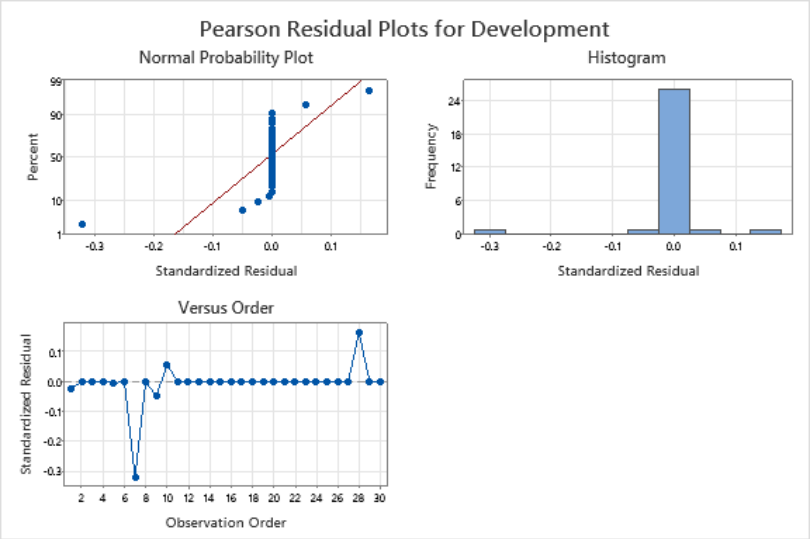
## Analysis of Variance

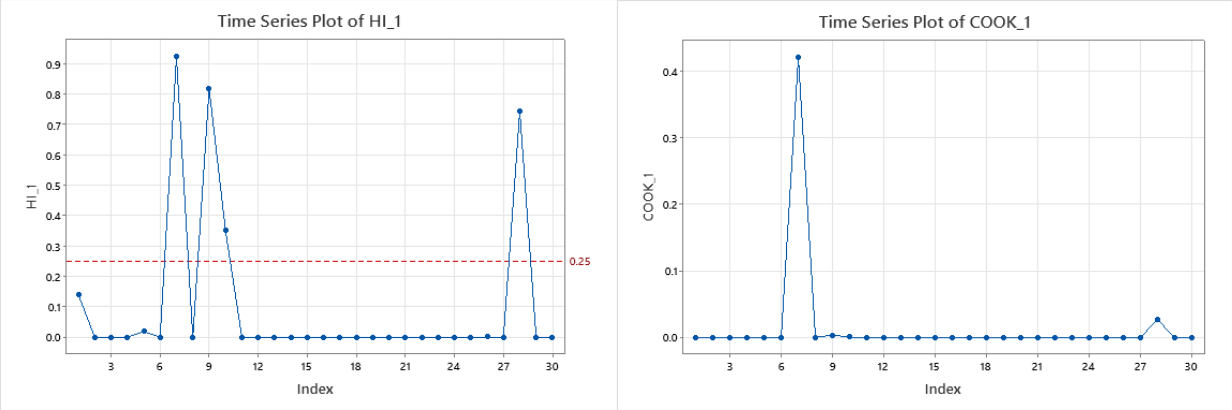
| | | | | Likelihood Ratio | |
|---|---|---|---|---|---|
| Source | DF | Adj Dev | Adj Mean | Chi-Square | P-Value |
| Regression | 2 | 41.5532 | 20.7766 | 41.55 | 0.000 |
| Health Expenditure | 1 | * | * | * | * |
| Supply per capita | 1 | 16.5428 | 16.5428 | 16.54 | 0.000 |
| Error | 27 | 0.0357 | 0.0013 | | |
| Total | 29 | 41.5888 | | | |

## Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 225 | 100.0 | Somers' D | 1.00 |
| Discordant | 0 | 0.0 | Goodman-Kruskal Gamma | 1.00 |
| Ties | 0 | 0.0 | Kendall's Tau-a | 0.52 |
| Total | 225 | 100.0 | | |

*Association is between the response variable and predicted probabilities*






The above is the result of two-predictor model. As shown in the odds ratios for continuous predictors table above, the health expenditure coefficient says that an increase of one percentage point in the health expenditure is associated with multiplying the odds of being developed countries by 787.8398, holding all other variables in the model fixed; the supply per capita coefficient says that an increase of one ten gigajoules in the supper per capita is associated with multiplying the odds of being developed countries by 12.5704, holding all other variables in the model fixed. And the Hosmer-Lemeshow test statistic with a highly statistically significant p-value of 1.00 shows that the data fit the logistic regression model pretty well. What's more, Somers' D and Area under the ROC Curve both exhibit values of 1.00, showing perfect separation and implying that this model with two predictors is successful in accurately differentiating between developed and developing countries. This complete separation of the data still limits the analysis of the variance table. As before, it contains a p-value of 0.000 for the Likelihood Chi-square statistic of the total regression, indicating that we should strongly reject the null hypothesis that the slopes of all the predicting variables are equal to zero. But now, we are able to read the p-value of the supply per capita coefficient, which is 0.000, indicating that this predicting variable contributes to the overall predictive ability of the model. The AIC value

is 6.04 and AICc value is 6.96 in this model, which is smaller than that of 8.03 and 9.63 respectively.

## Binary Logistic Regression: Development versus Supply per capita

* WARNING * When the data are in the Response/Frequency format, the Residuals versus fits plot is unavailable.

* WARNING * The model could not be fit properly. Maximum likelihood estimates of parameters do not exist due to complete separation of data points. The results are not reliable. Please refer to help for more information about complete separation.

### Method

| | |
|---|---|
| Link function | Logit |
| Residuals for diagnostics | Pearson |
| Rows used | 30 |

### Response Information

| Variable | Value | Count |
|---|---|---|
| Development | 1 | 15 (Event) |
| | 0 | 15 |
| | Total | 30 |

### Regression Equation

P(1) = exp(Y')/(1 + exp(Y'))
Y' = -53.8 + 5.46 Supply per capita

## Coefficients

| Term | Coef | SE Coef | Z-Value | P-Value | VIF |
|---|---|---|---|---|---|
| Constant | -53.8 | 44.2 | -1.22 | 0.224 | |
| Supply per capita | 5.46 | 4.47 | 1.22 | 0.222 | 1.00 |

## Odds Ratios for Continuous Predictors

| | Odds Ratio | 95% CI |
|---|---|---|
| Supply per capita | 235.7216 | (0.0369, 1.50668E+06) |

## Model Summary

| Deviance R-Sq | Deviance R-Sq(adj) | AIC | AICc | BIC | Area Under ROC Curve |
|---|---|---|---|---|---|
| 99.84% | 97.44% | 4.07 | 4.51 | 6.87 | 1.0000 |

## Goodness-of-Fit Tests

| Test | DF | Chi-Square | P-Value |
|---|---|---|---|
| Deviance | 28 | 0.07 | 1.000 |
| Pearson | 28 | 0.03 | 1.000 |
| Hosmer-Lemeshow | 8 | 0.00 | 1.000 |

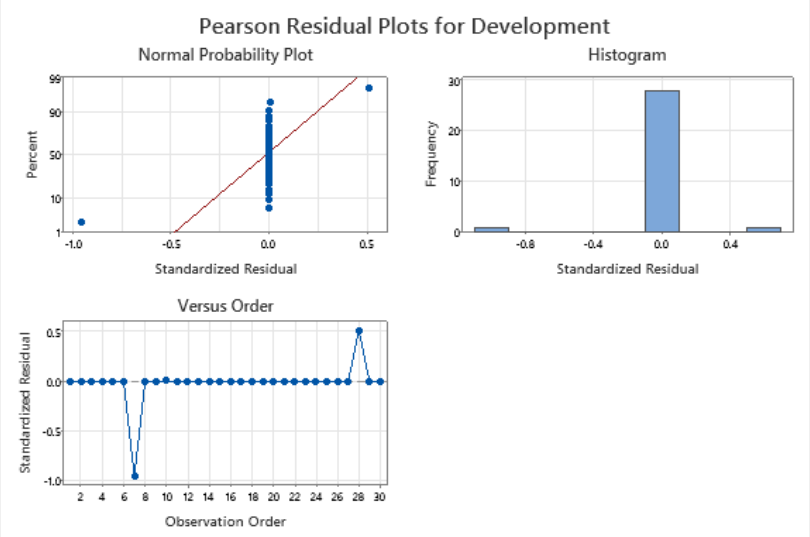
## Analysis of Variance

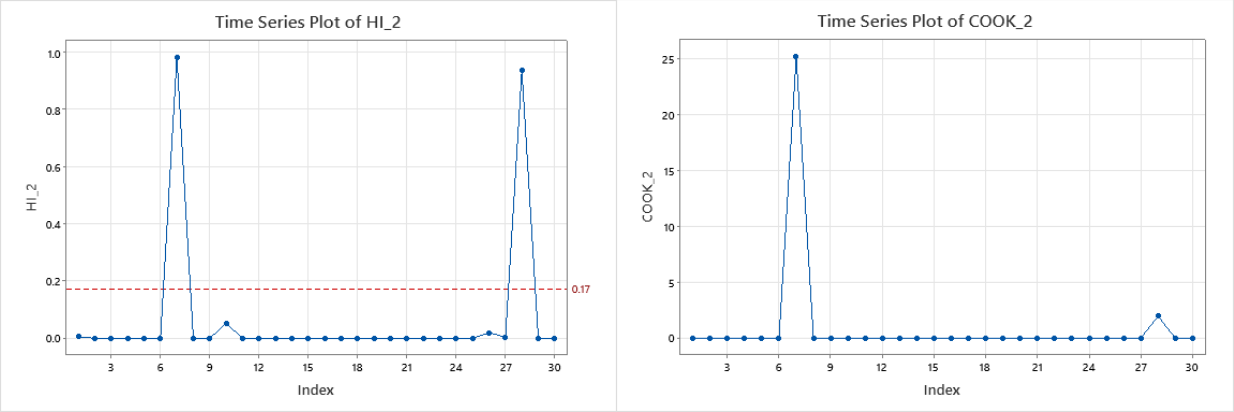
| Source | DF | Adj Dev | Adj Mean | Likelihood Ratio Chi-Square | P-Value |
|---|---|---|---|---|---|
| Regression | 1 | 41.5227 | 41.5227 | 41.52 | 0.000 |
| Supply per capita | 1 | 41.5227 | 41.5227 | 41.52 | 0.000 |
| Error | 28 | 0.0661 | 0.0024 | | |
| Total | 29 | 41.5888 | | | |

## Measures of Association

| Pairs | Number | Percent | Summary Measures | Value |
|---|---|---|---|---|
| Concordant | 225 | 100.0 | Somers' D | 1.00 |
| Discordant | 0 | 0.0 | Goodman-Kruskal Gamma | 1.00 |
| Ties | 0 | 0.0 | Kendall's Tau-a | 0.52 |
| Total | 225 | 100.0 | | |

*Association is between the response variable and predicted probabilities*






The above is the output of the one-predictor model. As shown in the odds ratios for continuous predictors table above, the only coefficient, supply per capita, says that an increase of one ten gigajoules in the supper per capita is associated with multiplying the odds of being developed countries by 235.7216. And the Hosmer-Lemeshow test statistic has a highly statistically significant p-value of 1.00, meaning that the data fit the logistic regression model very well. What's more, Somers' D and Area under the ROC Curve still both exhibit values of 1.00, indicating perfect separation. As a result, this model with only one predictor is still successful in accurately differentiating between developed and developing countries. With only one predictor, complete separation of the data stops limiting the analysis of the variance table. Same as the previous two models, it contains a p-value of 0.000 for the Likelihood Chi-square statistic of the total regression, indicating that we should strongly reject the null hypothesis that the slopes of all the predicting variables are equal to zero. And the p-value of the supply per capita coefficient, which is also 0.000, which is highly statistically significant. The AIC value is 4.07 and AICc value is 4.51 in this model, which is the lowest among all the model tested.

As a result, the one-predictor model with level of development as response variable and supply per capita as predicting variable is our preferred logistic regression model in this report.

Besides, it is unnecessary to exclude unusual points in such a one-predictor model that displays complete separation and perfect description of the pattern in the data.

Lastly, let's take a look at the confusion matrix which is generated based on whether the predicted probability is greater than or less than 0.5.

**Rows: Development  Columns: Predict**

| | 0 | 1 | All |
|---|---|---|---|
| 0 | 15 | 0 | 15 |
| | 50 | 0 | 50 |
| 1 | 0 | 15 | 15 |
| | 0 | 50 | 50 |
| All | 15 | 15 | 30 |
| | 50 | 50 | 100 |

*Cell Contents*
*Count*
*% of Total*

As shown above, 100% of the countries were correctly classified, much higher than Cpro = 1.25 * (0.5 * 0.5 + 0.5 * 0.5) = 62.5% and Cmax = 50%, definitely reinforcing the strength of our one-predictor logistic regression model.

In conclusion, this report explored the association between the level of development of a country and three factors: health expenditure, supply per capita, and emissions per capita. And our logistic regression analysis showed that energy supply per capita is a strong predictor of a country's level of development, while health expenditure and emissions per capita had a weaker association. This is most likely a result of how crucial energy supply is to a country's ability to power infrastructure like transportation, as well as the basic needs of a country's citizen for lighting and cooking, and to ensure national security in the event of energy shocks or supply disruptions. However, there is still much that needs to be investigated further, such as the

inclusion of longitudinal data to account for changes over time and the investigation of other significant variables that could affect the level of development of a country.