Statistics for Data Analytics

Statistical Analysis using Multiple regression and Time series models

Submitted by:

Victory Chimamaka Uwaoma
MSc. Data Analysis
National College of Ireland
Dublin, Ireland
x19210931

Submitted to: Tony Delaney

1. AIM AND OBJECTIVES

This research project aims at performing two statistical analysis namely, the Multiple regression and Time series on two distinctive data. The software used to perform these analyses are Statistical Package for the Social Sciences (SPSS) version 25 and R statistics software, the data on the gross domestic product per capita of several countries was gotten from the World Development Indicators (WDI) and the second data on the average daily concentration of ozone from the National Institute of Water and Atmospheric Research (NIWA) a research center in New Zealand.

2. INTRODUCTION

Gross Domestic Product (GDP) per capita is one of the determinants of the economic performance of a country which accounts for its number of citizens. It is the ratio of the gross domestic product of the country by the population at large. It is also the ideal indicator for the standard of living in a country, which shows you how successful a nation feels with its citizens.

The GDP data will be analyzed using the multiple linear regression analysis that was collated by the World Development Indicators (WDI), a global standard for accessing economic growth of a country. The WDI offers a detailed analysis of progress benefiting from World Bank data with more than 30 collaborators. The data used was chosen from the year 2016 which has several countries showing the GDP per capita for that year alone. This will be used as the dependent variable and the Cost of export (border compliance), Cost of import (border compliance) and the Total tax rate (% of commercial profit) as the independent variables.

According to National Aeronautics and Space Administration (NASA) officials (Newman & Nash, 2018), the ozone layer is a layer in the earth's stratosphere which has an altitude of approximately 10 km (6.2 miles) and comprises of three atoms of oxygen (O_3) that helps prevent most of the ultraviolet radiation from the sun touching the earth. Ozone is formed when chemicals react with each other in the atmosphere, for instance, air pollutants from car exhaust, causing an increase in the concentration of the ozone, thereby, making the layer depreciate and become vulnerable.

In the earth today, due to the numerous chemical reactions from our environment, the high concentration of ozone has become harmful to living organisms which includes humans, plants and animals. The National Institute of Water and Atmospheric Research (NIWA) a research center in New Zealand carried out a research on the ozone layer of the average daily concentration of ozone in two forms. The first sample was collected over a period of time in 1987 utilizing Dobson spectrophotometer (number 72) as the measuring instrument. The second sample was from a satellite observation against the Dobson global network. Both samples were collected at Lauder in Otago in good weather conditions and direct sunlight, this was done from 1978. Both sample of data was collected over a long period of time between 1979 – 2016 owing to the fact that some days had cloud, rain, or too much wind.

3. RESEARCH QUESTIONS

a. The Multiple regression analysis will answer the research question of to what extend does the Cost of export (border compliance), Cost of import (border compliance) and Total tax rate affect the GDP per capital.

b. The Time series analysis will answer the research question of forcasting the high Ozone concentration for the next 5 year.

4. DATA SOURCES

The first data being used on Multiple linear regression analysis was downloaded from 4 different related data sources and merged together to give complete information on World Development Indicator of several countries in the year 2016 (http://data.un.org/Explorer.aspx). The data comprises of 6 columns (country_or_area, year, gdp_per_capita, cost_of_export_border_compliance, cost _of_import_border_compliance, total_tax_rate_of_commercial_profit) on the World Development Indicator with a sample size of 226.

The second data which will be done on Time series is gotten from the National Institute of Water and Atmospheric Research (NIWA) in New Zealand within a time period of 1979-2016 but has in total 3 years of data. (https://catalogue.data.govt.nz/dataset/average-daily-ozone-concentrations-19792016/resource/47c1c889-807e-4e43-a760-731230ef530d). The data downloaded contains 4 columns/variables (month, day_of_year, statistics, and average_column_ozone_concentration) on the average daily ozone concentrations, with a total of 1098 samples. Both datasets are in .csv an acceptable format which SPSS and R software supports.

4.1 Data Cleaning and Transformation

In this phase, the data downloaded was in .csv file which is an acceptable format in SPSS. The cleaning and transformation process were handled manually by selecting the appropriate fields to make the data fit the multiple regression model. The time series dataset did not require any form of cleaning. The procedure were as follows:

- Before downloading the files for each of the dataset of the multiple regression, I filtered every dataset using the year (2016) as the common value and then downloaded the 4 datasets in .csv format.
- I opened all the datasets using excel and cleaned it in such a way that they were of the same length (uniformity), eliminating the countries that were not common to all.
- I cleaned the data until they were uniform, then merged them together with each value matching their country into a new excel sheet.
- The new World Development Indicator dataset has Country or Area, Year, GDP per capita, Cost of export (border compliance), Cost of import (border compliance) and Total tax rate (% of commercial products) as columns.

4.2 Variables

The table below shows the summary of the variables of multiple regression and time series analysis to be used.

Multiple Regression

Variable Name	Type	SPSS Measure	Unit of Measurement	Class (Variable)
country_or_area	String	Nominal	Countries	—
year	Numeric	Scale	Year (2016)	Numeric
gdp_per_capita	Numeric	Scale	Current Int. $	Dependent (Continuous)
cost_of_export_border_compliance	Numeric	Scale	US $	Independent (Continuous)
cost_of_import_border_compliance	Numeric	Scale	US $	Independent (Continuous)
total_tax_rate_of_commercial_profit	Numeric	Scale	% of commercial profit	Independent (Continuous)

Fig 1: Summary of the variables used in the Multiple Regression analysis

Time Series

Variable Name	Type	SPSS Measure	Unit of Measurement	Class (Variable)
Month	Date	Nominal	In months	—
Day_of_year	Numeric	Scale	In days	Independent (Numeric)
Statistic	String	Nominal	Average, Max, Min	Ordinal
Average_column_ozone_concentration	Numeric	Scale	Dobson unit (DU)	Dependent (Continuous)

Fig 2: Summary of the variables used in the Times series analysis

5. METHODOLOGY

During this phase, the analysis was carried out using multiple regression and time series. The results were recorded and interpreted accordingly.

5.1 MULTIPLE REGRESSION

Multiple linear regression is an extension of the simple linear regression. It is a statistical technique used to build a model that predicts or estimates one quantitative variable which is the dependent or response variable (y) by using at least two or more other quantitative variable known as the independent variables (x). Unlike the simple linear regression that uses exactly one x (independent) variable to predict the y (dependent) variable i.e. $x \rightarrow y$, the multiple regression on the other hand uses two or more x variables in predicting the y value i.e. $x_1, x_2, \ldots, x_k \rightarrow y$.

There are basically three types of regression; standard or simultaneous, hierarchical or sequential and the stepwise regression, but in this project, I will be making use of the standard or simultaneous regression which is the most common (Pallant, 2016).

The standard regression also known as the simultaneous regression is the process whereby all the independent variables are entered simultaneously into the equation. Each independent variable is being analyzed in respect of its predictive power, in comparison to all other independent variables (Pallant, 2016).

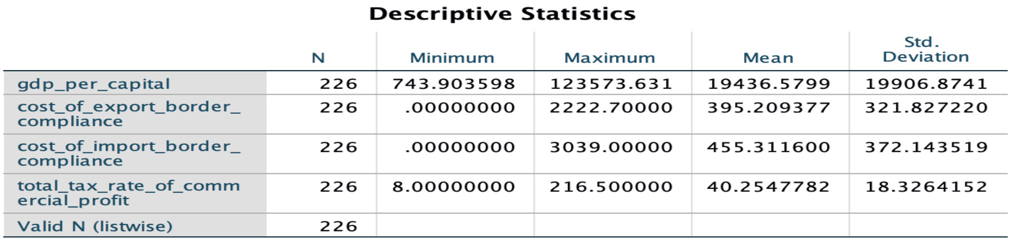
A certain number of criteria (assumptions) have to be met before Multiple linear regression is conducted to ensure that the data is suitable for analysis. These criteria include; sample size, outliers, normality, linearity, homoscedasticity, independence of residuals.

a. **Sample size:** When carrying out a multiple regression analysis, it is advisable not to use a small sample size of data to avoid the issue of generalizability where the result gotten from the model cannot be used to generalize (cannot be repeated) other samples.

(Tabachnick & Fidell, 2013) provide a formula for estimating sample size, taking into consideration the number of independent variables you hope to use: $N > 50 + 8m$ (where N is the sample size and m is the number of independent variables). In this case, the sample data being used has a total of 226 with 3 independent variables. Placing the following into the formula

$$226 > 50 + 8(3)$$
$$226 > 74$$

From the above, we can infer that the data satisfies the assumptions of the sample size. The table below from SPSS shows the descriptive statistics of the data.

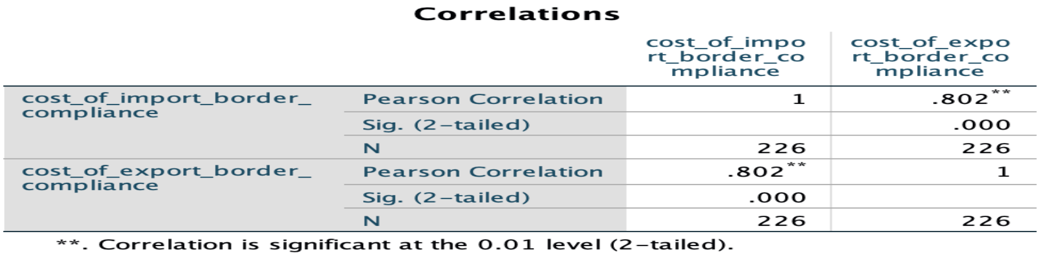


Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
gdp_per_capital	226	743.903598	123573.631	19436.5799	19906.8741
cost_of_export_border_compliance	226	.00000000	2222.70000	395.209377	321.827220
cost_of_import_border_compliance	226	.00000000	3039.00000	455.311600	372.143519
total_tax_rate_of_commercial_profit	226	8.00000000	216.500000	40.2547782	18.3264152
Valid N (listwise)	226				

Fig 3: Table showing Descriptive Statistics

b. **Multicollinearity and singularity:** Multicollinearity and singularity is another assumption to watch out for before performing a multiple regression. It is the relation between the independent variables only. When the independent variables are highly correlated (r = .9 and above) with each other, multicollinearity is said to exist. Singularity on the other hand exists when one independent variable is a combination of other independent variables. In other to determine if the independent variables have the above-mentioned assumption, the Pearson correlation test, Variance Inflation Factor (VIF) and scattered plot analysis will be done on the independent variables; cost of import (x_1), cost of export (x_2) and total tax rate (x_3) to test for multicollinearity.

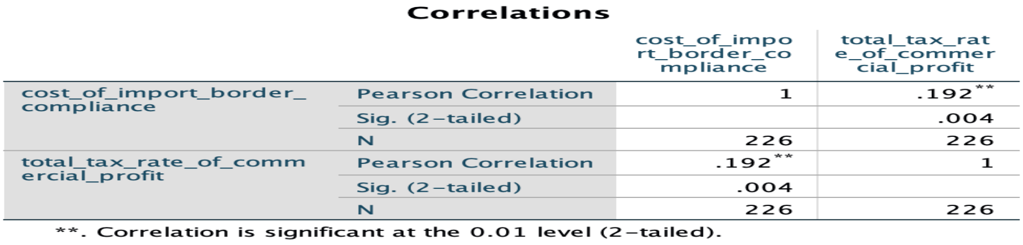


Correlations

		cost_of_import_border_compliance	cost_of_export_border_compliance
cost_of_import_border_compliance	Pearson Correlation	1	.802**
	Sig. (2-tailed)		.000
	N	226	226
cost_of_export_border_compliance	Pearson Correlation	.802**	1
	Sig. (2-tailed)	.000	
	N	226	226

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 4: Pearson correlation between x_1 and x_2

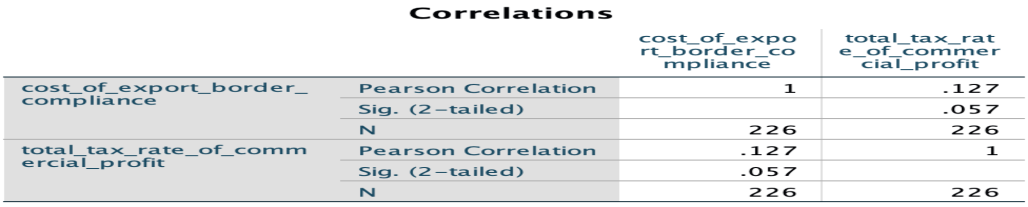


Correlations

		cost_of_import_border_compliance	total_tax_rate_of_commercial_profit
cost_of_import_border_compliance	Pearson Correlation	1	.192**
	Sig. (2-tailed)		.004
	N	226	226
total_tax_rate_of_commercial_profit	Pearson Correlation	.192**	1
	Sig. (2-tailed)	.004	
	N	226	226

**. Correlation is significant at the 0.01 level (2-tailed).

Fig 5: Pearson correlation between x_1 and x_3



Correlations

		cost_of_export_border_compliance	total_tax_rate_of_commercial_profit
cost_of_export_border_compliance	Pearson Correlation	1	.127
	Sig. (2-tailed)		.057
	N	226	226
total_tax_rate_of_commercial_profit	Pearson Correlation	.127	1
	Sig. (2-tailed)	.057	
	N	226	226

Fig 6: Pearson correlation between x_2 and x_3

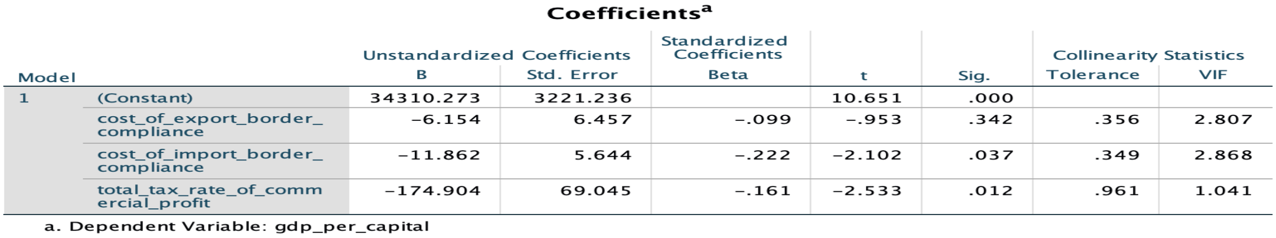
Interpretation for Multicollinearity (using Correlation)

At the end of the assumption test carried out, Pearson correlation test output in Figures above shows:

- The relationship between x1 and x2 is r = .802 less than r = .9 and above, meaning that multicollinearity might be suspected.
- The relationship between x1 and x3 is r = .192 less than r = .9 and above, again, it shows no multicollinearity and lastly,
- The relationship between x2 and x3 is r = .127 less than r = .9 and above, which shows no multicollinearity.

Summary for Multicollinearity (using Correlation)

All in all, all 3 independent variables fulfill this assumption.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	34310.273	3221.236		10.651	.000		
	cost_of_export_border_compliance	-6.154	6.457	-.099	-.953	.342	.356	2.807
	cost_of_import_border_compliance	-11.862	5.644	-.222	-2.102	.037	.349	2.868
	total_tax_rate_of_commercial_profit	-174.904	69.045	-.161	-2.533	.012	.961	1.041

a. Dependent Variable: gdp_per_capital

Fig 7: Showing the Coefficients table

Interpretation for Multicollinearity (using Variance Inflation Factor VIF)

In this case, we want a VIF way less than 10 and from the output in Fig 7 above, we can see that the cost of export has a VIF of 2.807, while the cost of import has a VIF of 2.868 and the total tax rate has a VIF of 1.041

Summary for Multicollinearity (using VIF)

All in all, all 3 independent variables fulfill this assumption.

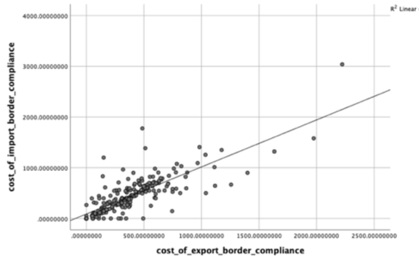


Fig 8: Scattered plot of cost of import(x_1) against cost of export(x_2)

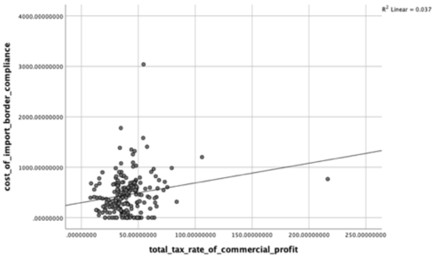


Fig 9: Scattered plot of cost of import(x_1) against total tax rate(x_3)

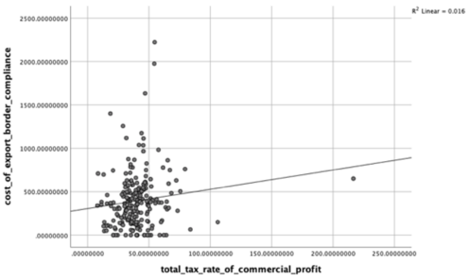


Fig 10: Scattered plot of cost of export(x_2) against total tax rate(x_3)

Interpretation for Multicollinearity (using Scattered plot)

At the end of the multicollinearity test carried out using scattered plot, the output in Fig 9, 10 and 11 shows that:

- In Fig 8, there is 64.3% high correlation between cost of import (x_1) and cost of export (x_2).
- In Fig 9, there is a 3.7% correlation between cost of import (x_1) and total tax rate (x_3), which does not appear highly correlated.
- In Fig 10, there is a 1.6% correlation between cost of export (x_2) and total tax rate (x_3), which does not appear highly correlated with each other.

Summary for Multicollinearity (using Scattered plot)

At the end of the multicollinearity assumption, since cost of import is highly correlated with cost of export, I will not use both in the multiple linear regression because they are redundant and will not be suitable for the model. Therefore, I will be taking out the cost of export (x_2). This will leave the model with only 2 independent variables.

c. **Outliers:** are values that lie at abnormal distances from the spread of values in the data affect the multiple linear regression model. outliers will be checked on all variables that will be used in the regression analysis, the dependent and independent variables.

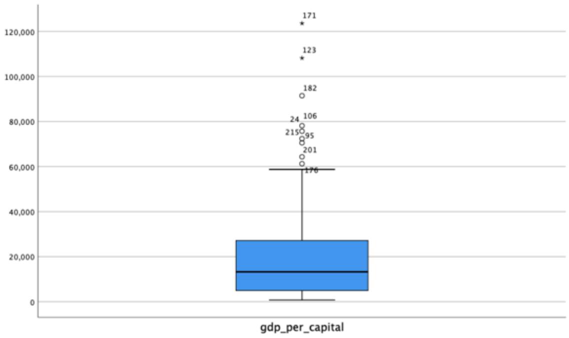


Fig 11: Showing outliers for GDP per capital (y)

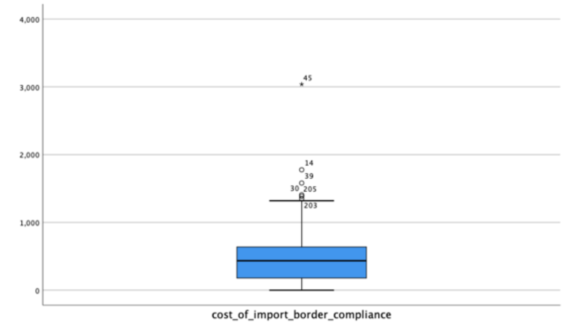


Fig 12: Showing outliers for Cost of import (x_1)

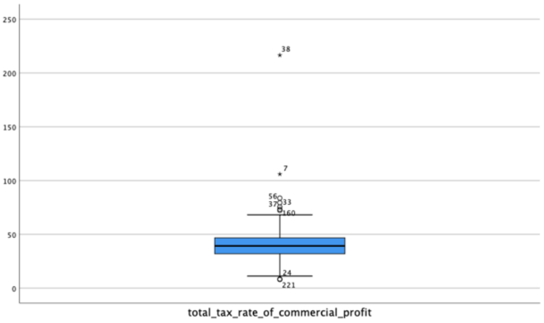


Fig 13: Showing outliers for Total tax rate (x_3)

Interpretation for Outliers (using Boxplot)

The figures above show the test assumption for outliers using boxplot, the symbol (o) identifies a normal outlier and (★) shows extreme outliers.

- Fig 11 showing the outliers in GDP per capita has 9 outliers with 2 extreme cases (123 and 171) in the data
- Fig 12 showing the outliers in cost of import has 6 outliers with only 1 extreme case (45) in the data and finally,
- Fig 13 showing the outliers in the total tax rate has 8 outliers with 2 extreme cases (7, 38).

Summary for Outliers (using Boxplot)

In total, there are 23 outliers in the dataset. Since we cannot adjust the values of the outliers, we will have to get rid of the extreme cases of outliers only (throw it out) by deleting them from the sample in order to satisfy our model. The five extreme cases include (123, 171, 45, 7, 38).

d. **Normality, Linearity, Homoscedasticity, Independence of residuals:** The normality, linearity, homoscedasticity and independence of residuals all talk of the various parts of the scores distribution and the nature that the relationship of the variables take (Pallant, 2016). The normal probability plot is an effective way to check the above-mentioned assumptions.

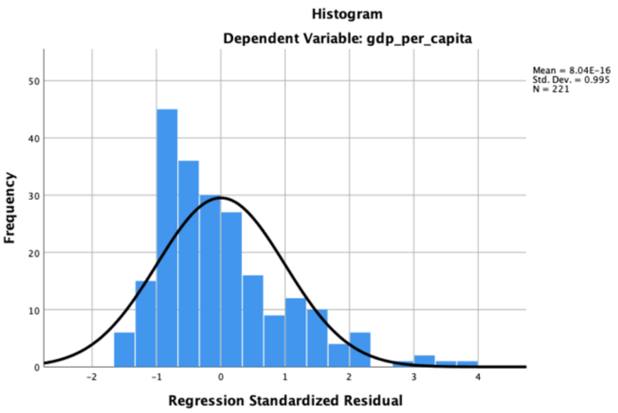


Fig 14: Histogram

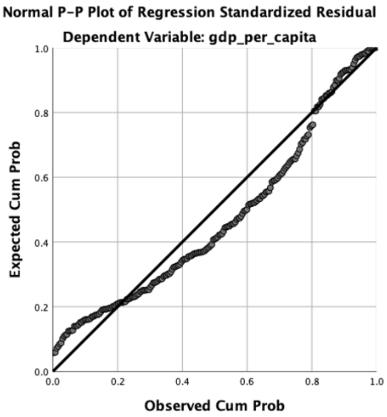


Fig 15: Probability plot

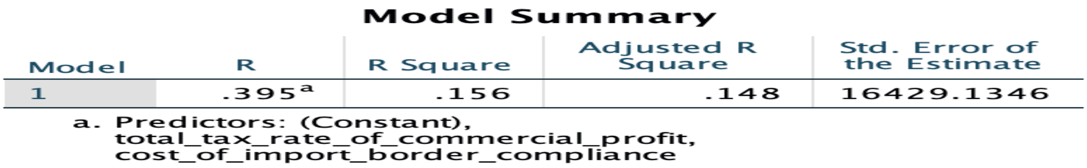
Interpretation for Normality, Linearity, Homoscedasticity, Independence of residuals (using Histogram and Probability plot)

- The bell shape in the histogram in Fig 14 shows that there is a normal distribution in the data.
- The Normal P-P plot of regression standardized residual in Fig 15 shows that they are no major deviations from the regression line drawn across the data points.

Summary for Normality, Linearity, Homoscedasticity, Independence of residuals (using Histogram and Probability plot)
The above interpretations show that the independent variables fulfill this assumption mentioned above.

INTERPRETATION OF THE MULTIPLE REGRESSION ANALYSIS

After all the above assumptions have being met, the multiple regression was done using a single dependent variable, GDP per capita (y) and two independent variables, Cost of import (x_1) and total tax rate (x_2). The model summary table shown below in Fig 16 is required to evaluate how well the algorithm used fits the data and also to interpret the results properly.



Model Summary

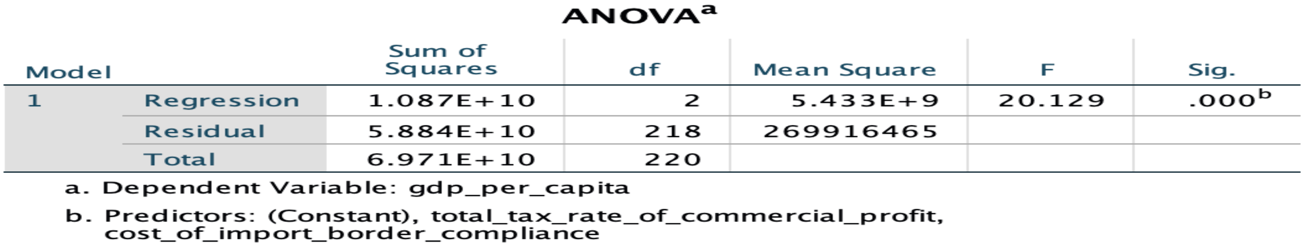
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.395[a]	.156	.148	16429.1346

a. Predictors: (Constant), total_tax_rate_of_commercial_profit, cost_of_import_border_compliance

Fig 16: Model Summary

Interpretation of Model Summary

From the model summary table given above, the R square takes a value of .156 which interprets the independent variables are only accounting for 15.6% of variation in the dependent variable. This indicates that the two independent variables are not correctly able to predict the outcome of the dependent variable because it has a low R square percentage. It has a standard error of 16429.1346 which shows the average distance of the data point from the regression line. It is measured in the unit of the dependent variable (current Int. $).



ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.087E+10	2	5.433E+9	20.129	.000[b]
	Residual	5.884E+10	218	269916465		
	Total	6.971E+10	220			

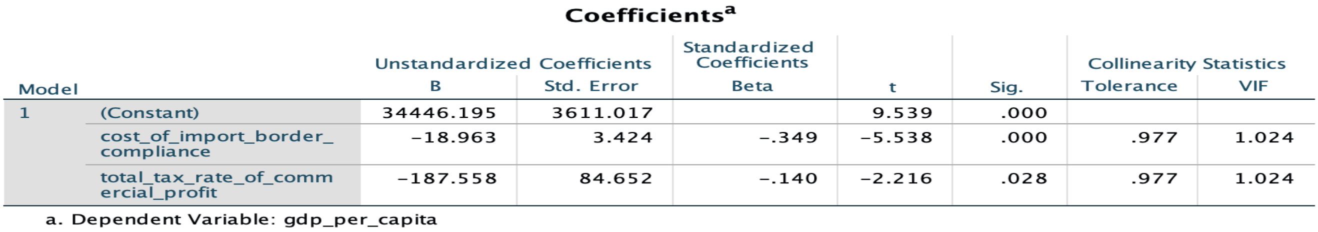
a. Dependent Variable: gdp_per_capita

b. Predictors: (Constant), total_tax_rate_of_commercial_profit, cost_of_import_border_compliance

Fig 17: ANOVA

Interpretation of ANOVA

ANOVA gives the significance of the overall model. The ANOVA table above in Fig 17 shows that the model is statistically significant with ($p < 0.05$) and F = 20.129. Therefore, the relationship between two independent variables are statistically significant in predicting the GDP per capital of the different countries.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	34446.195	3611.017		9.539	.000		
	cost_of_import_border_compliance	-18.963	3.424	-.349	-5.538	.000	.977	1.024
	total_tax_rate_of_commercial_profit	-187.558	84.652	-.140	-2.216	.028	.977	1.024

a. Dependent Variable: gdp_per_capita

Fig 18: Coefficients

Given that the ANOVA table has validated the significance of the model, hence, the multiple linear regression equation of the model can be derived from the coefficients table in Fig 18. From the coefficients table, $\beta_0 = 34446.195$, $\beta_1 = -18.963$, $\beta_2 = -187.558$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where, $\hat{y}$ is the response/dependent variable, β_0 is the y-intercept where X equal zero (constant) and x is the independent variable

By substituting the values, we get the equation to be:

$$\hat{y} = 34446.195 - 18.963x_1 - 187.558x_2$$

$$\hat{y} = 34446.195 - 18.963(\text{cost of import}) - 187.558(\text{total tax rate})$$

The equation tells us that if the total tax rate is held constant, then the cost of import will decrease by $18.963 US dollar for each additional cost of import made. Likewise, if cost of import is held constant, then the total tax rate will decrease by 187.558% of commercial profit for each additional tax paid.

5.2 Time Series

The time series is a sequence of data points that measures the same thing over a long period of time or it can also be seen as data measured at successive period of time at uniform time interval.

The time series analysis will be carried out using R. Before I choose among a variety of time-series models, the following procedures will be done.

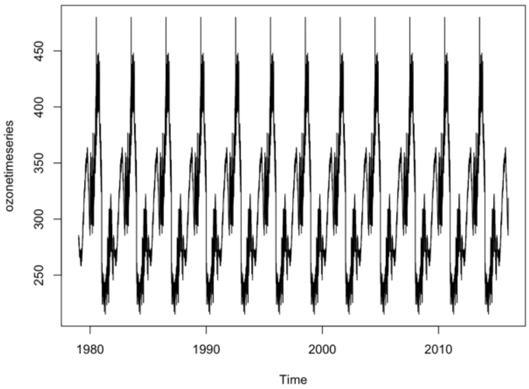


Fig 19: Time series plot

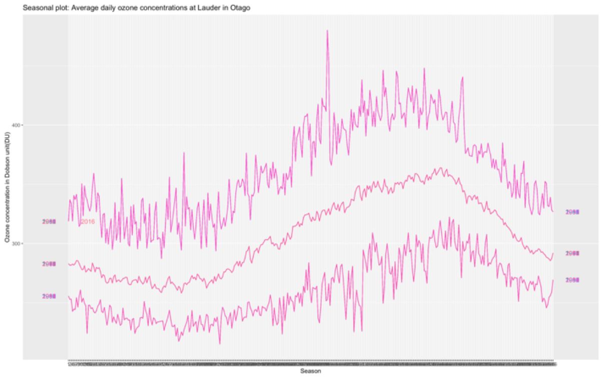


Fig 20: Seasonal plot

The time plot above shows that the time series could perhaps be explained using an additive model, since the data has constant fluctuations which are roughly constant in size over a period of time. The seasonal plot in *Fig 20* is similar to a time plot in that the observations was collected against the particular "seasons" in which the data were recorded. It shows that the data has a seasonal pattern from the previous years before 2016.

For Additive time series model, the formula is written as:

$$Y_t = S_t + T_t + \epsilon_t$$

Where *Yt* = each data point at time t, S_t = *Seasonality*, T_t = *Trend* and ϵ_t = *Error* (white noise)

A seasonal time series is comprised of a trend pattern, a seasonal component and an irregular component. The decomposition of the time series involves the classification of the time series into these three components; the trend, seasonal and irregular components using the decompose() function.

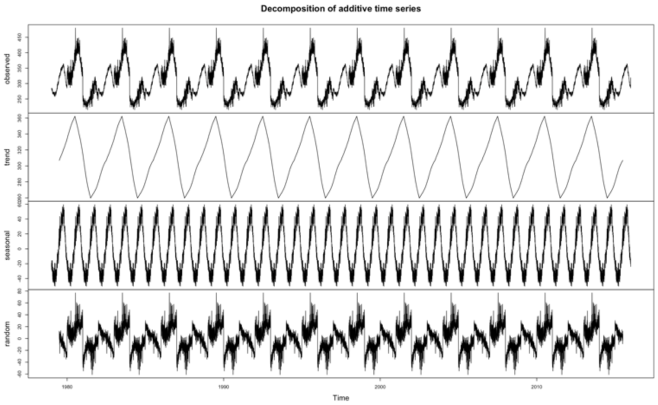


Fig 21: Decomposition of additive time series

The figure above shows the decomposition of an additive seasonal time series. It shows the original time series (top), following closely is the estimated trend component (second from top), thereafter, the estimated seasonal component (third from top), and the estimated irregular component (bottom). Here we see a constant trend and a linear seasonality with the same frequency.

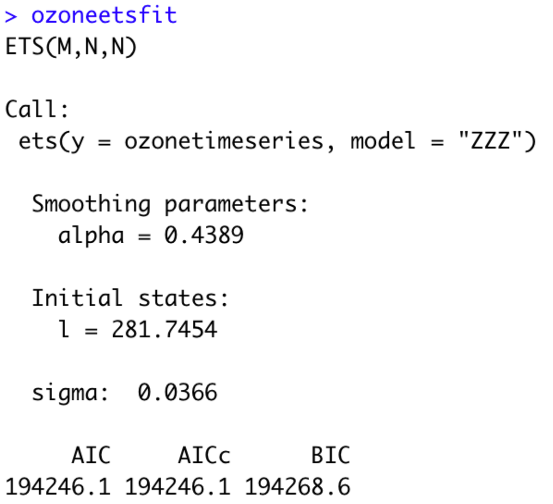
Forecasts using Exponential Smoothing

Exponential smoothing for time series data is used to make short-term forecasts.

The ETS model also provides an automated way to select the right method. The function is written as:

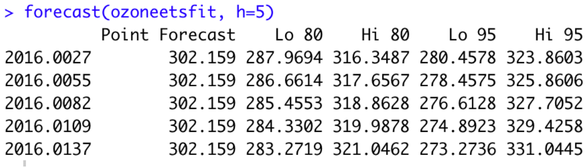
ets(ts, model = "ZZZ")

where ts = time series and the model is defined by three letters. The first letter depicts the error type, the second letter depicts trend type, and the last letter depicts the seasonal type. Acceptable letters are A for additive, M for multiplicative, N for none, and Z for automatically selected).



```
> ozoneetsfit
ETS(M,N,N)

Call:
 ets(y = ozonetimeseries, model = "ZZZ")

  Smoothing parameters:
    alpha = 0.4389

  Initial states:
    l = 281.7454

  sigma:  0.0366

     AIC     AICc      BIC 
194246.1 194246.1 194268.6 
```

Fig 22: Fitting the model using ets()



```
> forecast(ozoneetsfit, h=5)
          Point Forecast    Lo 80    Hi 80    Lo 95    Hi 95
2016.0027        302.159 287.9694 316.3487 280.4578 323.8603
2016.0055        302.159 286.6614 317.6567 278.4575 325.8606
2016.0082        302.159 285.4553 318.8628 276.6128 327.7052
2016.0109        302.159 284.3302 319.9878 274.8923 329.4258
2016.0137        302.159 283.2719 321.0462 273.2736 331.0445
```

Fig 23: Forecasting with the model

6. Final Conclusion

In conclusion, the Multiple regression analysis which was used to answer the question of to what extend does the Cost of export (border compliance), Cost of import (border compliance) and Total tax rate affect the GDP per capital, shows that the regression model was not a good fit because of the low R square value of .156 which accounts for only 15.6% of variation in the dependent variable.

On the other hand, the goal of time series was not to fit the best possible forecasting model for Ozone concentration, but to forecast the high Ozone concentration for the next 5 year which it did and thus a good model for forecasting.

In a real-world application, lots of time should be spent on preprocessing, feature engineering and feature selection. The multiple regression will not be a good fit in further analysis but the time series model can be used further for forecasting.

References

1. Newman, P. A. & Nash, E. R., 2018. *NASA Ozone Watch.* [Online]Available at: https://ozonewatch.gsfc.nasa.gov/facts/SH.html [Accessed 16 March 2020].
2. Pallant, J., 2016. *SPSS SURVIVAL MANUAL A step by step guide to data analysis using SPSS.* 6th Edition ed. Sydney: Allen & Unwin Academic.
3. Tabachnick, B. G. & Fidell, L. S., 2013. *Using Multivariate Statistics.* 5th Edition ed. California: Pearson Education Limited.