International Journal of Analysis and Applications
Volume 19, Number 4 (2021), 576-586
URL: https://doi.org/10.28924/2291-8639
DOI: 10.28924/2291-8639-19-2021-576

A COMPARATIVE STUDY OF THE IMPACT OF DUMMY VARIABLES ON REGRESSION COEFFICIENTS AND CANONICAL CORRELATION INDICES: AN EMPIRICAL PERSPECTIVE

NSISONG EKONG[1,*], IMOH MOFFAT[2], ANTHONY USORO[1], ISEH MATTHEW[1]

[1] *Department of Statistics, Akwa Ibom State University, Ikot Akpaden, Akwa Ibom State, Nigeria*

[2] *Department of Statistics, University of Uyo, Uyo, Akwa Ibom State, Nigeria*

[*] *Corresponding author:* nsisongekong@aksu.edu.ng

Abstract. In this paper, the impact of dummy variables on regression coefficients and canonical correlation indices from an empirical perspective is investigated. To do this, a regression analysis of Crude Oil Prices on US dollars - Naira Exchange Rates is performed, and the extent of the significance of the relationship is noted. Secondly, dummy variables (coded with respect to various economic regimes of interest) is introduced into the regression of the two variables and the impact of such introduction is also noted. And also, a canonical correlation analysis (CCA) of Inflation rate, the dummy variables and Crude Oil Prices and the dummy variables is conducted. Finally, we compare the significant role of the introduction of the dummy variables on the coefficients of the regression and the canonical correlation indices. The results showed that the introduction of dummy variables impact more on the canonical correlation indices than it does on the regression coefficients.

1. Introduction

There are numerous approaches to unravelling the relationships between two or more entities; regression and correlation analyses, amongst others, are some of these approaches. With these myriads of techniques, researchers are often faced with the problems of which to use, when, where and how to use them. Although

Received January 13th, 2021; accepted February 3rd, 2021; published June 17th, 2021.

2010 *Mathematics Subject Classification.* 62P10.

Key words and phrases. regression analysis; canonical correlation analysis (CCA); dummy variables; crude oil prices; exchange rates; Nigeria.

misuses of statistics are inadvertent, it is necessary to put things in check and raise the necessary alarm on its consequences such as erroneous relationships and flawed conclusions. In this paper, focus is on canonical correlation and regression analysis. Some of these consequences are timebound while others span through ages as they are printed and relied upon by others in the field. Therefore, getting to know the proper usability of these techniques is of essence.

In the case of analysing the relationships between two or more variables, we, more often than not, see insignificant relationships whereas there may exist a significant one in actual sense, and vice versa. This issue may be attributed to the approach we use in the analysis. The question now begging for answer is can we checkmate the problem of false results (significance or nonsignificance)? Putting it differently, can a relationship be significant in one approach and insignificant in another? In attempting to answer this, we need to put the methods of interest side by side while comparing their pros and cons.

Also, we sometimes encounter the problem of interrelationship between different approaches. For instance, in ascertaining the relationship between quantitative and qualitative variables, we could be faced with questions like: Would an introduction of dummy variables to a regression line be similar or dissimilar to an introduction of dummy variables to a canonical correlation analysis of these variables? This paper is the at the intersection of the two issues raised above. In the literature, a handful of works have been done on the relationship between regression and canonical correlation analyses. In studying the relationship between canonical correlation analysis and multivariate multiple regression, Lutz and Eckert (1994) posited that the similarities and dissimilarities between multivariate multiple regression and canonical correlation analysis has been inconsistently acknowledged in the literature. In trying to put this in the right perspective, the authors stated that although the two analyses seems to have different objectives, the underlying aspects of the approaches are mathematically equivalent. In their postulation, they put it that a multivariate multiple regression analysis that incorporates discriminant analysis as part of its post hoc investigation will produce identical result as a canonical correlation analysis in terms of omnibus significance testing, variable weighting schemes, and dimension reduction analysis. Similarly, Sun, et al (2008) showed that for a high dimensional data, CCA in multi-label classification can be formulated as least squares problems. The authors did this by proposing several canonical correlation extensions including sparse CCA using what they call 1-norm regularization base on equivalence relationship. Also, Kakade and Foster(2007) in investigating the canonical correlation analysis through multi-view regression approach, provided a semi-supervised algorithm in the multi-view regression problem where the input variable can be partitioned into two different views. This algorithm, according to them, first uses unlabeled data to learn a norm and the uses labeled data in ridge regression algorithm to provide the predictor. By doing this, the authors were able to characterized the intrinsic dimensionality of the subsequent ridge regression problem. In a more succint way, Foucart (1998) says that the results obtained by the selection of canonical variables are better than those given by calssical regession and principal component regression. Apart from these, other pieces of research are carried out on this topic and they could be found in the literature (Everitt (2005); George (1975); and others). Different from these works, we move to ascertain the impact of dummy variables on the regresion coefficients compared to its impact on canonical correlation indices via an empirical perspective. To do this, we first of all regress the dependent variable (in this case, the exchange rates) with the independent variable (crude oil price in this case). Secondly, we introduce dummy variables into the regression model by taking cognizance of pre-covid 19 era, covid 19 era, and post covid 19 era. Thirdly, we also increase the dimensionality of both the dependent and independent variables with a dummy variable and use same to conduct a canonical correlation analysis. We compare the various statistics from both methods to see the signicance induced by the introduction of the dummy variables.

This paper is organised as follows; in section 1, a general introduction is given, section 2 gives an overview of the methodology employed in the data analysis, and section 3 presents the results of the analyses. Final;y, section 4 concludes with a summary.

2. Methodology

In the analysis of data related to this research, we will employ the regression analysis, dummy regression analysis, and canonical correlation analysis.

2.1. **Linear Regression Analysis.**

Regression analysis have become an integral component of any data analysis when it comes to deciphering the relationship between a response variable and explanatory variable(s) given that the dataset are discrete in nature. This is a procedure for fitting a linear model that relates one or more independent variables, $X's$ to a dependent variable Y. The independent variables are also referred to as explanatory variables while the dependent variable is also called response variable. Walpole and Myers (1989), linear regression is a concept of arriving at the best estimate of the relationship between a particular set of variables linearly. Here, the term linearity implies linear in coefficients. The technique typically uncovers the variability by the explanatory variable(s) in the response variable. When the response variable is regress against one explanatory variable, the method is called a Simple Linear Regression. On the other hand, if the number of explanatory variables exceed one, the approach employed in the analysis is called a Multiple Regression Analysis. The general form of the model with n-independent variables is given as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \varepsilon \tag{2.1}$$

where Y is the response variable, X_1, X_2, ...,X_n are the predictor variables with associated parameters β_1 β_2, ..., β_n respectively. ε is the error term associated with the model.

For the variables under consideration, that is US dollars - Naira Exchange Rates and Crude Oil Prices (in dollars), we have a simple linear regression of the form,

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \tag{2.2}$$

where Y is the US dollars - Naira Exchange Rates, X_1 is Crude Oil Prices (in dollars) with associated parameters β_1, that is, the slope of the regression line. β_0 is the regression constant, that is the intercept of the regression line. ε is the error term associated with the model.

2.2. **Dummy Regression Analysis.** A dummy variable is a numerical variable used in regression analysis to represent sub-groups of the sample in the study. It is used in research design to distinguish different treatment groups. According to Draper & Smith (1981), an assignment of some levels to variables in order to account for the fact that the variable may have seperate deterministic effects on the response; variables that result from this sort of assignment are called dummy variables. The variables are limited to two specific values, 0 or 1. Typically, 1 represents the presence of a qualitative attribute, and 0 represents the absence.

In applying the regression technique, we are often presented with explanatory variables which are categorical in nature and/or are in levels. The presence of these levels makes place for such a variable to be coded with dummies. A regression analysis where these dummy variables are used in coding the different levels in the explanatory variables which are categorical in nature is called Dummy Regression. Of course, we should know that if the response variable is also categorical, we make use of the logistic regression technique.

The dummy regression model is equivalent to the regular regression model but for the fact that the there is a dummy variable incorporated into the model to make sense of the levels in the explanatory variables. For every k levels, we need $k-1$ dummy variables.

For the model under consideration, that is, US dollars - Naira Exchange Rates and Crude Oil Prices (in dollars) regression model, we will consider 3 levels. These are, the pre-Covid-19 era and the Covid-19 era. This means we will need 2 dummy variables. The resulting dummy regression model is given as

$$Y = \beta_0 + \beta_1 X_1 + \alpha_1 Z_1 + \varepsilon \tag{2.3}$$

where Y is the US dollars - Naira Exchange Rates, X_1 is Crude Oil Prices (in dollars) with associated parameters β_1, that is, the slope of the regression line. β_0 is the regression constant, that is the intercept of the regression line. Z_1 is the dummy variable coded with 1's for the pre-Covid-19 era and 0's otherwise. Its associated parameter is α_1. ε is the error term associated with the model. Covid-19 era is used in this case as the reference subgroup, hence not coded with dummies.

2.3. **Canonical Correlation Analysis.** Canonical correlation analysis , introduced by Harold Hotelling in 1936, is a way of making sense of cross-covariance matrices. Canonical Correlation Analysis (CCA) is a well-known technique for finding the correlations between two sets of multi-dimensional variables. It projects both sets of variables into a lower-dimensional space in which they are maximally correlated. This nimplies that the canonical correlation indices only provides a measure of linear association between the two variables; when the two are uncorrelated, i.e. where their correlation is zero, this means that there is no linear function that can describe their relationship. Thus, some non-linear relationship could suffice. In partitioning the correlation matrix, we consider the two sets of data $y_1, y_2, ..., y_p$ and $x_1, x_2, ..., x_q$ such that we have p dependent variables and q independent variables. The dimensions of the correlation matrix R are $(p+q)(p+q)$. The matrix can be partitioned so that

$$R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$$

where R_{11} is the matrix of intercorrelations among the p dependent variables, R_{22} is the matrix of intercorrelations among the q independent variables, R_{12} is the matrix of intercorrelations between the p dependent and the q independent variables, and then R_{21} is the transpose of R_{12} The partitioned intercorrelation matrix reveals the basic pattern of association within and between the two sets of variables. To find the pattern, or canonical correlations, we first have to find the latent roots of the canonical equation given as

$$(R_{22}^{-1} R_{21}^{-1} R_{11}^{-1} R_{12} - \lambda I) = 0 \tag{2.4}$$

where λ is the latent root and I is the identity matrix. Eq (2.4) can be written as

$$(M - \lambda I) = 0 \tag{2.5}$$

where

$$M = R_{22}^{-1} R_{21}^{-1} R_{11}^{-1} R_{12}$$

We have to note that the matrix M represents the ways in which the two sets covary with one another and the latent roots of this matrix indicates the size of the common patterns. The canonical roots are the square of the latent roots and is given as

$$\bar{\omega}_i = \sqrt{\lambda}_i, \forall i = 1, 2, ..., min\{p, q\}$$

In order to test the significance of the roots, we use the test statistic known as the Wilk's lambda (Λ), which is described from the canonical correlation roots, and is given as

$$\Lambda = \prod_{i=1}^{min\{p,q\}} (I - \lambda_i) \tag{2.6}$$

To test the hypothesis

H_{0_0}: The variables are not correlated

against

H_{1_0}: The variables are correlated

given that we have n independent observations in a sample and λ_i is the estimated correlation for i=1,2,...,min{p,q}, each row can be tested for significance with the following methods. For the i^{th} row, the test statistic is

$$\chi^2 = -(n-1-\frac{1}{2}(p+q+1))log_e\Lambda \tag{2.7}$$

which is asymptotically distributed as a chi-square with pq degrees of freedom. We reject H_{0_0}, the hypothesis of no relationship if $\chi^2 \leq \chi^2_{pq}(\alpha)$

In this study, the two sets of variable of interest are the Crude Oil Price with the dummy variable (Z_1), and the Exchange rates with a dummy variable (Z_1). The over hypothesis with respect to this study is,

H_{0_1}: The impact of dummy variables on the two approaches are the same

against

H_{1_1}: The impact of dummy variables on the two approaches differs.

The statistic of interest will be the coefficient of determination (the square of the correlation coefficient) R^2 which measures the strength of the relationship between the variables under investiagtion. We compare this statistic from the two approaches and then either accept the H_{0_1} if their differences is negligible, or otherwise reject.

3. Data Analysis and Results

R software is used in the analysis of this work.

3.1. **Data Collection.** Data used for this research work are secondary data obtained from the Central Bank of Nigeria (CBN) website. The exchange rates data can be found on https://www.cbn.gov.ng/rates/exrate.asp while crude oil prices can be found on https://www.cbn.gov.ng/rates/crudeoil.asp. The monthly data are collected from the year 2010 to 2020.

3.2. **Analyses and Results.** The data used in this study were dummy coded to reflect the pre-Covid 19 and Covid 19 era as reflected on **Table 1** of the **Appendix**. Firstly, a regression analysis of crude oil prices on exchange rates was carried out. The output is as given in **Table 2** of the **Appendix**, and the fitted regression line is given as;

$$Y = 426.8999 - 2.5217X_1 \tag{3.1}$$

Its associated statistics are as follows; Residual standard error: 89.29 on 177 degrees of freedom Multiple R-squared: 0.3663, Adjusted R-squared: 0.3627 F-statistic: 102.3 on 1 and 177 DF, p-value: < 2.2e-16.

Secondly, a multiple regression analysis of exchange rates against Crude Oil prices and the dummy variable was done. The output is as given in **Table 3** of the **Appendix**, and the fitted regression line is given below;

$$Y = 513.2538 - 2.0981X_1 - 128.4192Z_1 \tag{3.2}$$

Its associated statistics are s follows; Residual standard error: 83.78 on 176 degrees of freedom Multiple R-squared: 0.4452, Adjusted R-squared: 0.4389 F-statistic: 70.62 on 2 and 176 DF, p-value: $<$ 2.2e-16

Finally, a canonical correlation analysis was carried out on Exchange rates, the dummy variable and Crude oil prices with the dummy variable. The output is as presented in **Table 4** of the **Appendix** Canonical correlation analysis of:Exchange rates, the dummy variable and Crude oil prices with the dummy variable. The correlations are high and the statistic of interest, R^2 is given as 1.0000 and 0.2878 for the two roots respectively.

4. Discussion of Results

From the results we obtain from the first analysis given in **Table 2** of the **Appendix**, although the R^2 is relatively small in value, we found that there is a significant relationship linear between the Exchange Rates (X) and the Crude Oil Prices (Y), as the p-value is very small $<$ 2.2e-16. The result also shows that only about 37% of the variation in Exchange Rates is explained by Crude Oil Prices. In order to ascertain the performance of the Exchange Rates in the two era (pre-Covid 19 and Covid 19), we incorporated the dummy variables, coded with 1's for the pre-covid 19 era and 0's otherwise. This introduction of dummy variables into the model will also give an opportunity to discover the improvability of the model, as the dummy variable is seen to be statistically significant in the model as shown in **Table 3** of the **Appendix**. Also, the R^2 has been noticed to improve from 37% to about 45%. Furthermore, as noticed in **Table 4** of the **Appendix**, the canonical correlation between Exchange rates, the dummy variable and Crude oil prices with the dummy variable revealed a strong correlation between the two sets of variables. The first canonical R^2 has been seen to be a perfect correlation of 100%, far better than that of the multile regression of Exchange Rates with Crude Oil Prices and the Dummy variable. The second canonical R^2 of about 29% is also relatively strong. Comparing the R^2 from the multiple regression analysis and the canonical correlation analysis, we therefore reject the H_{0_1} and conclude that the introduction of dummy variables to the approaches impacts more on the canonical correlation analysis than it does on the multiple regression analysis.

5. Conclusion

In this paper, the impact of dummy variables on canonical correlation indices and regression coefficients was compared empirically. To achieve this, we considered Exchange Rates and Crude Oil prices as dependent and independent variable respectively. First of al, a simple linear regression analysis of Exchange Rates and Crude Oil prices was carried out. At this point, a degree of linear relationship between the two variables was investigated. Having established a significant relationship between the duo, a dummy variable was introduced to the model to examine the improvability of the relationship given a regime-based (pre-Covid 19 and Covid 19 era) approach to studying the relationship. It was uncovered that there is a need to analyse the relationship based on the two era, as the coefficient of the dummy variable was highly significant and the model obviously improved. Finally, the canonical correlation indices from the Exchange Rates and the dummy variable versus the Crude Oil prices and the dummy variable were observed to outperformed the multiple regression of Exchange Rates versus the Crude Oil prices and the dummy variable. The implication of this finding shows that the dummy variable impacts more on the canonical correlation indices than it does on the regression coefficient.

6. Appendix

Table 1. Data on Exchange Rates (X), Crude Oil Prices (Y), and the Dummy Variable (Z),

This is a monthly dataset from January 2006 to November 2020. The dummy variable is coded to reflect the pre-Covid 19 (1's) and Covid 19 (0's) era.

Month	No. of Pregnant Women	No. of Positive	Enrolled	Not Enrolled
January	291	23	16	7
February	48	15	12	3
March	193	12	9	3
April	200	13	11	2
May	89	20	13	7
June	36	12	10	2
July	132	33	16	17
August	96	12	10	2
September	143	9	7	2
October	92	5	0	5
November	250	12	11	1
Total	1570	166	115	51

Table 2. Model Summary for Exchange Rates and Crude Oil Prices

Coefficients:	Estimate	Std. Error	t value	Pr(< $\|t\|$)
(Intercept)	426.8999	20.3925	20.93	< 2e-16 ***
Crude.Oil.Price	-2.5217	0.2493	-10.11	< 2e-16 ***

— Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 89.29 on 177 degrees of freedom Multiple R-squared: 0.3663, Adjusted R-squared: 0.3627 F-statistic: 102.3 on 1 and 177 DF, p-value: ¡ 2.2e-16

Table 3. Model Summary for Exchange Rates and Crude Oil Prices, Dummy Variable

Coefficients:	Estimate	Std. Error	t value	Pr(< $\|t\|$)
(Intercept)	513.2538	25.7657	19.920	¡ 2e-16 ***
Crude.Oil.Price	-2.0981	0.2488	-8.433	1.18e-14 ***
Dummy	-128.4192	25.6611	-5.004	1.35e-06 ***

— Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 83.78 on 176 degrees of freedom Multiple R-squared: 0.4452, Adjusted R-squared: 0.4389 F-statistic: 70.62 on 2 and 176 DF, p-value: ¡ 2.2e-16

Table 4. Canonical correlation analysis of: Exchange rates, the dummy variable and Crude oil prices with the dummy variable

	CanR	CanRSQ	Eigen	percent cum	scree
1	1.0000	1.0000	-2.252e+15	1.000e+02	100**********************
2	0.5365	0.2878	4.041e-01	-1.795e-14	100

	CanR	LR test stat	approx F	numDF	denDF	Pr(> F)
1	1.00000	0.0000	4	350		
2	0.53647	0.7122	71.121	1	176	1.181e-14 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Abbreviations. CCA, Canonical Correlation Analysis; CBN, Central Bank of Nigeria.

Consent for publication. Not Applicable

Ethics approval and consent to participate. Not Applicable

Funding. Not Applicable

Availability of data and materials. All data used for supporting the conclusions of this article are available from the public data repository at the website https://zenodo.org/record/4419710.

Authors' contributions. Nsisong Ekong designed, coordinated the research, and conducted the analysis and drafted the manuscript, Imoh Moffat supervised the research processes, Anthony Usoro oversaw the technical aspect of the research and ensured appropriate research language usage, and Matthew Iseh sourced for data and contributed in the analysis.

Recommendations for further research. There are a number of aspects around the role of dummy variables in canonical correlation analysis (CCA) that follow from our findings which will be of essense for further investigation, including a more robust evaluation of the hypothetical test and approaches respectfully constructed and employeded here:

(1) Factor Loading: Another approach, though similar to CCA, which could be used to tackle this problem is the use of factor loadings for the comparisons. Factor loadings are part of the outcome from factor analysis, which serves as a data reduction method designed to explain the correlations between observed variables using a smaller number of factors. This, acording to our projection, may give a deeper and more robust insight into the role of dummy variable in canonical correlation analysis.

(2) Hypothesis Test: Setting up hypothesis to be ejected based on R value may either be confusing or misleading. A more appropriate and reliable hypothesis testing approach should be developed for questions like this.

Conflicts of Interest: The author(s) declare that there are no conflicts of interest regarding the publication of this paper.

References

[1] J.G. Lutz, T.L. Eckert, The relationship between canonical correlation analysis and multivariate multiple regression, Educ. Psychol. Measure. 54 (1994), 666–675.

[2] N.R. Draper, H. Smith, Applied regression analysis. 2^{nd} ed. Wiley, New York, (1981).

[3] B.S. Everitt, Multiple Regression and Canonical Correlation, in: An R and S-PLUS® Companion to Multivariate Analysis, Springer London, London, 2005: pp. 157–170.

[4] T. Foucart, Paper on multiple linear regression on canonical correlation variable.Biometrical J. 41 (1998), 559-572.

[5] L.K. George, A comparison of canonical correlation and multiple regression in the analysis of change. In: Proceedings of the 1975 Joint Statistical Association, (1975), 262-269.

[6] S.M. Kakade, D.P. Foster, Multi-view Regression via Canonical Correlation Analysis. In: International Conference on Computational Learning Theory, (2007), 82-96.

[7] L. Sun, S. Ji, J. Ye, Least Squares Formulations for Canonical Correlation Analysis. In: International Conference on Machine Learning, (2008), 1024-1031.

[8] R.E. Walpole, R.H. Myers,Probability and Statistics for Engineers and Scientists. 4^{th} ed. Macmillan Publishing Company, New York, (1989).