Bayero Journal of Pure and Applied Sciences, 6(1): 27 – 31
Received: November 2012
Accepted: March 2013
ISSN 2006 – 6996

PRINCIPAL COMPONENT REGRESSION ANALYSIS OF CO_2 EMISSION

Okonkwo, E.N., Okeke, J. U. and Nwabueze, Joy Chioma.
[1]Nnamdi Azikiwe University, Awka, Nigeria
[2]Anambra State University, Uli, Nigeria
[3]Micheal Okpala University of Agriculture, Umudike Nigeria
Corresponding author: evelyn70ng@yahoo.com

ABSTRACT
Principal component regression (PCR) model is developed, in this study, for predicting and forecasting the abundance of CO_2 emission which is the most important greenhouse gas in the atmosphere that contributes to global warming. The model was compared with supervised principal component regression (SPCR) model and was found to have more predictive power than it using the values of Akaike information criterion (AIC) and Swartz information criterion (SIC) of the models.
Keywords: Global warming, CO_2, Principal component regression (PCR), Supervised principal component regression (SPCR), Akaike information criterion (AIC) and Swartz information criterion (SIC)

INTRODUCTION

Global warming refers to the increase in the earth's temperature as measured by the average temperature of the earth's near-surface air and ocean. The gradual rise in the temperature of the earth's atmosphere is caused by an increase of green house gas in the air. On earth, the major greenhouse gases are vapour, which causes about 36-70% of the greenhouse effect (not including clouds), CO_2 which causes 9-26%, CH_4 which causes 4-9%, and O_3 which causes 3-7%. Some other naturally occurring gases contribute in a very small fraction (Wikipedia, the free encyclopedia). The most important greenhouse gas (GHG) is CO_2; it is the most abundant in the atmosphere (360 ppmv(part per million volume)) apart from vapour, has a high calorific power and is easily generated by human activities, essentially by the burning of fossil, fuels, and wood. It accounts for 60% of global warming or total greenhouse effect. CO_2 is the greenhouse gas of reference, and the other gases are stated in units of CO_2. The amount of CO_2 released collectively by respiration, anaerobic microbial activities, fuel combustion, and volcanic activity, has increased to more than 30% since the beginning of the industrial era Environmental Impact Assessment (EIA) (2003).

Talaro and Talaro (2003) documented that the mean temperature of the earth has increased by 1.6 ^oC since 1860. If this rate of increase according to them continues, by 2020 a rise in the average temperature of 4^oC to 5^oC will begin to melt the polar ice caps and raise the levels of the oceans by 2 to 3 feet. Some expert predict more serious effects, including massive flooding of costal regions, changes in rainfall patterns, expansion of deserts, and long term climatic disruptions. "The UN climate panel issued its first report in February 2007 (UN 2007), saying that it was at least 90% certain that mankind was to blame for global warming. Their second report on April 2007, warned of more hunger, drought, heat-waves, and rising seas" (Guardian May 1, 2007). EEA (2013) stated that average concentration of various greenhouse gases in the atmosphere have reached the highest levels ever recorded and concentrations are increasing. According to the agency the combustion of fuel from human activities and land-use changes are largely responsible for the increase.

Tucker (1995) considered carbon dioxide emissions and global gross domestic product (GDP) of 137 countries and stated that a positive relationship exists between them. Song et al (2007) stated that CO_2 emission from fossil fuel is a major cause of global warming effect, but it is hard to remove completely in actuality. Moreover, according to them energy consumption is bound to increase for the continuous economic development of a country that has an industrial formation requiring high energy demand.

Because of the ever rising rate of green house gases on air, this study centered on developing principal component regression (PCR) model which will be use in monitoring and forecasting the rate at which CO_2 emission is growing. Principal component regression model is considered appropriate for this study because we observed the presence of multicollinearity among the predictor of CO_2 emission. Also, in this study, PCR model generated will be compared with the result from supervised principal regression using Akaike information criterion (AIC) and Swartz information criterion (SIC) for model selection.

Description of Principal Component Regression

Consider the standard regression model,

$$Y = \alpha + X\beta + \varepsilon, \qquad (1)$$

where Y is an n-dimensional response vector, X is an $n \times p$ predictor matrix, β is a p-dimensional vector of unknown regression parameters, ε is a random vector satisfying $E(\varepsilon) = 0$ and $var(\varepsilon) = \sigma^2 I$ for some unknown $\sigma^2 > 0$, and I represents the n-dimensional column vector of ones. Let the centered Y, be Y_c and centered X, be X_c, hence the sum of Y_c and X_c is zero. In matrix form we may write it as $Y_c = X_c\beta + \varepsilon_c$ where ε_c is the centered ε . The goal of the analysis is to estimate β. Let r denotes the rank of X_c. The full-column rank case $(r = p)$ is most relevant because all components of β are estimable only when $r = p$ .This can also be used when the matrix is not of full rank, that is, when $r < p$.

Let Z denotes a $n \times r$ matrix whose columns are an orthonormal basis for the column space of X_c. Rather than projecting Y_c onto the column space of X as in ordinary least squares (OLS) regression, we wish to project Y_c onto a subspace of the column space of X spanned by a subset of the columns of Z . Let c consists of k distinct integers chosen from 1,…,r that represent the indices of selected columns of Z . Let C denotes the $r \times k$ matrix consisting of columns of the $r \times r$ identity matrix corresponding to $j \in c$. An estimator of β based on the orthonormal basis Z and the chosen set c is given by

$$\hat{\beta}_{pcr} = U(U'U)^{-1}CC'Z'Y_c \qquad (2)$$

Where $U = [w_1, \ldots, w_r]$ the unique $p \times r$ matrix of rank r such that $X_c = Zw'(i.e., W = X_c'Z)$

Note that

$$X_c\hat{\beta}_{pcr} = ZW'\hat{\beta}_{pcr} = ZCC'Z'$$

$$= (ZC)[(ZC)(ZC)]^{-1}(ZC)'Y \qquad (3)$$

Thus, when $c = \{1, \ldots, r\}$, $X_c\hat{\beta}_{pcr}$ is the unique projection of Y on the column space of Z or equivalently, the unique projection of Y on the column space of X. It follows that, for $c = \{1, \ldots, r\}$, $\hat{\beta}_{pcr}$ is a least squares estimator of β . In general, $\hat{\beta}_{pcr}$ is the unique and unbiased least squares estimator of β when the rank of X is p. When $C \subset \{1, \ldots, r\}$, $X_c\hat{\beta}_{pcr}$ is the projection of Y onto the subspace spanned by the column of Z indexed by C.

General expressions for the expectation and variance of $\hat{\beta}_{pcr}$ are given by

$$E(\hat{\beta}_{pcr}) = ACC'W'\beta \text{ and } var(\hat{\beta}_{pcr}) = \sigma^2 ACC'A' = \sum_{j \in c} a_j a_j' \qquad (4)$$

where $A = [a_1, \ldots, a_r] = W(W'W)^{-1}$. When $C \subset \{1, \ldots, r\}$, expression (4) shows that $\hat{\beta}_{pcr}$ is a potentially biased estimator of β . It can be shown that $\hat{\beta}_{pcr}$ will be biased for β if and only if β is in the column space of W and $w'_j = 0$ for any j$\notin C$. Expression (4) shows that the variance of any component of $\hat{\beta}_{pcr}$ will be no larger than the variance of the corresponding component of the least squares estimator. To balance the virtue of lower variance with the cost of higher bias, we seek the set C that will yield the estimator with the lowest total mean squared error (MSE), that is, we wish to find C so that

$$MSE \equiv E\|\hat{\beta}_{pcr} - \beta\|^2 = E\|\hat{\beta}_{pcr} - E(\hat{\beta}_{pcr})\|^2 + \|E(\hat{\beta}_{pcr}) - \beta\|^2 \qquad (5)$$

is minimized. MSE defined in (5) is the trace of the MSE matrix $E\{(\hat{\beta}_{pcr} - \beta)((\hat{\beta}_{pcr} - \beta))'\}$. A variety of other performance measures based on the MSE matrix can be used to judge the equality of an estimator. We choose to consider MSE as defined in (5) because it is intuitively appealing to find the estimator that will minimize the expected squared Euclidean distance of the estimator from the estimand.

Note that the first term on the right hand side of (5) is equal to

$$\text{Trace}\{var(\hat{\beta}_{pcr})\} = \sigma^2 Trace(ACC'A') = \sigma^2 Trace(C'A'AC)$$

$$= \sigma^2 \sum_{j \in c} a'_j a_j \qquad (6)$$

Using the identity $WA'A = A$, we obtain that the second term on the right hand side of (5) is equal to

$$\|ACC'W'\beta - \beta\|^2 = \beta'\beta + \beta'WCC'A'ACC'W'\beta - 2\beta'WA'ACC'W'\beta$$

$$= \beta'\beta + \sigma^2 \sum_{j\in C}\left(\theta_j \sum_{i\in C}\theta_i a_i - 2\theta_j \sum_{k=1}^{r}\theta_k a_k\right)' a_j \quad , \qquad (7)$$

where $\theta = (\theta_1, \ldots, \theta_r) = \frac{W'\beta}{\sigma}$. By (6) and (7), minimizing $MSE(C|\beta, \sigma^2)$ with respect to c is equivalent to minimizing

$$g(c\backslash\theta) = \sum_{j\in C}\left(a_j + \theta_j \sum_{i\in C}\theta_i a_i - 2\theta_j \sum_{k=1}^{r}\theta_k a_k\right)' a_j \qquad (8)$$

with respect to c.

Application of Principal Component Regression

Let $d_1 > d_2 > \cdots > d_r > 0$ where d_j^2 denote the jth non-zero eigenvalue of $X'X$. Let $v_1, \ldots, v_r$ denote the corresponding eigenvectors of $X'X$. The sample principal components corresponding to the non-zere eigenvalues of $X'X$ are $Xv_1, \ldots, Xv_r$. In principal component regression, Y is regressed on a subset of the sample principal components. The estimated regression coefficients for the principal components in the chosen subset are used to obtain regression coefficients for the original column of X. For example, suppose Y is regressed against the first, second and fourth sample principal components to obtain

$$\hat{Y} = \hat{\tau}_1 Xv_1 + \hat{\tau}_2 Xv_2 + \hat{\tau}_4 Xv_4 = [X(\hat{\tau}]_1 v_1 + \hat{\tau}_2 v_2 + \hat{\tau}_4 v_4), \qquad (9)$$

where τ_j denote the ordinary least squares estimate of the regression coefficient for the jth principal component. Then a principal components estimator of β is given by $\hat{\tau}_1 v_1 + \hat{\tau}_2 v_2 + \hat{\tau}_4 v_4$. The principal component regression estimators like $\hat{\tau}_1 v_1 + \hat{\tau}_2 v_2 + \hat{\tau}_4 v_4$, are of the form $\hat{\beta}_{pcr}$ describe in section 2.0. The singular value decomposition of X implies $X = SDV'$, where

$S \equiv \left[\frac{Xv_1}{d_1}, \ldots, \frac{Xv_r}{d_r}\right]$, $D \equiv Diag(d_1, \ldots, d_r)$, and V=$[v_1, \ldots, v_r]$.

Note that the orthonormal columns of S are the sample principal components of X, scaled to unit length. We may equate S and VD with Z and W respectively. Further more $\theta_j = \frac{d_j v_j \beta}{\sigma}$. The estimator $\hat{\beta}_{pcr}$ simplifies to $VD^{-1}CC'S'Y$. In the example considered previously, C consist of the first, second, and fourth column of $r \times r$ identity matrix, and we have

$$\hat{\beta}_{pcr} = VD^{-1}CC'S'Y = VD^{-2}DCC'CC'S'Y = VCC'D^{-2}CC'S'Y$$

$$= VC(C'D^2C)^{-1}C'DS'Y = VC\left[(SDC)'(SDC)\right]^{-1}(SDC)'Y$$

$$= [v_1, v_2, v_4]([Xv_1, Xv_1, Xv_4]'[Xv_1, Xv_1, Xv_4])^{-1}[Xv_1, Xv_1, Xv_4]'Y$$

$$= \tau_1 v_1 + \tau_2 v_2 + \tau_4 v_4 \qquad (10)$$

Corollary 1.The principal component regression of β with smallest total mean square error is obtained by regressing Y against the sample principal components indexed by the set $\left\{j: |\theta_j| = \frac{d_j|v'_j\beta|}{\sigma} > 1\right\}$.

Interpretation of the corollary 1: Often in principal components regression, the principal components corresponding to the smallest eigenvalues are discarded. The variance of $\hat{\beta}_{pcr}$ in principal components regression is $\sum_{j\in C} d_j^{-2} v_j \, [\, v_j$, so excluding the principal components with the smallest eigenvalues from C can greatly reduce the variances of the components of $\hat{\beta}_{pcr}$. Although the strategy makes sense when variance reduction is a priority, several authors have an unimportant relationship with response variable Y. Jollife (1982) provided examples where the principal components corresponding to small eigenvalues have high correlation with Y. Hadi and Ling (1998) provided an example where only the principal component associate with the smallest eigenvalue was correlated with Y. The criterion suggested by Corollary 1 clearly discourages the use of principal component with small eigenvalues. However eigenvalue size is not the only thing we consider for the

selection of principal component. Even when d_j is small, the jth component will be selected when $|v'_j\beta|$ is sufficiently large. The quantity $d_j|v'_j\beta|$ is a measure of the strength of relationship between the jth component and the expected value Y given X. Note that the angle between the jth principal component and $E(Y|X)$ is

$$\cos^{-1}\left(\frac{v'_jX'X\beta}{d_j\|X\beta\|}\right) = \cos^{-1}\left(\frac{d_jv'_j\beta}{\|X\beta\|}\right). \quad (11)$$

Thus, when $d_jv'_j\beta$ is large, the angle between the jth principal component and $E(Y|X)$ is close to zero or to 180 degrees. In either case, the jth principal component will be useful when forming a linear combination of principal components to estimate $E(Y|X)$. On the other hand, when $d_jv'_j\beta$ is close to zero, the jth component is nearly orthogonal to $E(Y|X)$ and will be of little use when estimating $E(Y|X)$ with linear combination of principal components.

MATERIALS AND METHODS

The data for this study were obtained from Economic Fact book (2007). The data is on CO_2 emission and their possible correlates from randomly selected 50 countries whose annual CO_2 emissions in thousands of metric tons are at least 4. The possible correlates of CO_2 studied include gross domestic product (GDP) (X_1), industrial output (X_2), export output (X_3), energy consumption (X_4), manufacturing output (X_5), and population (X6).

RESULTS AND DISCUSSION

The application of principal component regression (PCR) on the CO_2 data gave the model for predicting CO_2 as

$$CO_2 = 0.0107X_1 + 0.0022X_2 - 0.0002X_3 + 0.0017X_4 + 0.0015X_5 + 0.0012X_6$$

The application of SPCR on the CO_2 data gave the model for predicting CO_2 as

$$CO_2 = 0.0018X_1 - 0.0232X_2 - 0.0026X_4 + 0.0254X_5$$

To select the best model for predicting CO_2 emission the four models were used in computing the residual analysis using the logarithm transform of Akaike Information criterion (AIC) and Swartz Information criterion (SIC).

$$AIC = e^{\frac{2k}{n}}\left(\frac{e'e}{n}\right)$$

$$SIC = n^{\frac{k}{n}}\left(\frac{e'e}{n}\right)$$

From the two functions we obtained the table of residual analysis as

Table 1: Residual Result of CO_2 Emission

	Residual Analysis		
Analytical Method	**Average Error**	**AIC**	**SIC**
Classical PCR	384369.3	27.8449	28.0722
SPCR	0.33	28.0102	28.2696

Table 1 shows that PCR model has the minimum estimated error with AIC and SIC values of 27.8449 and 28.0722 respectively.

This shows that PCR model is more adequate than SPCR model in predicting CO_2 emission in the atmosphere. The AIC and SIC values of PCR and classical SPCR are very close to each other. Based on this, one may say as well that both methods performed equally well.

CONCLUSION

The results show that PCR model outperformed SPCR model in predicting CO_2 emission in the atmosphere. The closeness of the AIC and SIC values of PCR and SPCR indicates that both models can be used in predicting CO_2 emission in the atmosphere.

From our findings, we therefore, conclude that the best model for predicting CO_2 emission in the atmosphere is PCR model.

REFERENCES

European Environmental Agency (EEA) (2013). European Union. www.eea.europa.eu, Jan 24, 2013 10:54 am.

Environmental Impact Assessment (EIA) (2003). U.S. Energy Information Administration pg 35-36.

Hadi, A.S. and Ling, R.F. (1998). Some cautionary notes on the use of principal component regression, The American Statistician, **52,** 15-19.

Jollife, I. T. (1982). A note on the use of principal components in regression, Applied Statistician. **31,** 300-303.

Song H., Lee S., Maken S. S., and Part J. (2007). Environmental and Economic Assessment of the Chemical Absorption Process in Korea, **35** , Issue 10, October 2007, 5109-5116.

Talaro K. P., and Talaro A. (2003). Green house gases in the atmosphere, Foundation in Microbiology, 2nd ed. 231-235.

The Economic Fact book (2007). List of countries by their GDP, Industrial output, manufacturing output, energy consumption, and export output, pg 10.

The Guardian Newspaper (2007). **24,** 10360, April 6.

Tucker M. (1995). Carbon dioxide emission, and global GDP. Ecological Economics, **15,** Issue 3, December 1995, 215-223.

United Nation (2007). Carbon dioxide emission from the generation of electric power.