Journal of Modern Mathematics and Statistics 7 (4): 41-46, 2013
ISSN: 1994-5388

Statistical Modeling of Global Warming

Igwenagu Chinelo Mercy
Department of Industrial Mathematics, Applied Statistics,
Enugu State University of Science and Technology, Nigeria

Abstract: The problems associated with global warming, ranging from increase in global temperature change in agricultural yields, glacier retreat, species extinction, increase in the ranges of diseases and disease vectors were reviewed. These underscore the need to reduce emission which causes global warming. The proposed method of emission reduction is by emission trading according to the Kyoto protocol. If this proposal holds for countries to participate actively, it is important to build a model for estimating their level of CO_2 emission. The aim of this study is to develop an exploratory model of global warming, using CO_2 emission as a surrogate. This was done using regression analysis and principal component analysis to explore some possible factors that could cause global warming and to know their actual contributions. The regression analysis result with a $p<0.001$ indicates that CO_2 emission is related to some of the input variables used. However due to the effect of multicollinearity among the variables used, supervised principal component regression analysis was used and the result of the analysis shows that model built on this method gave a good fit.

Key words: CO_2 emission, global warming, multicollinearity, modeling, principal component

INTRODUCTION

An increase in the emission of greenhouse gases, such as carbon dioxide (CO_2), methane (CH_4) and nitrous oxide (N_2O) from the soil surface to the atmosphere has been of worldwide concern over the last several decades. Carbon dioxide is recognized as a significant contributor to global warming and climatic change, accounting for 60% of global warming or total greenhouse effect. The accumulation of Greenhouse Gas (GHG) emissions in the atmosphere is arguably the most serious environmental threat of the time; recently carbon dioxide (CO_2) emissions are the largest GHG, accounting for over 80% of the emissions in the US environmental impact assessment, 2003. CO_2 emissions arise from the combustion of carbon fuels, such as gasoline in vehicles and coal in power plants. Energy-related carbon emissions are a global problem and the US produces more emission than any other country, accounting for 24% of the world's energy-related emission, EIA (op.cit). There is not yet complete agreement as to the extent and effects of global warming. It has been documented that the mean temperature of the earth has increased by 1.6°C since 1860 (Talaro and Talaro, 2003). If this rate of increase continues by 2020, a rise in the average temperature of 4-5°C will begin to melt the polar ice caps and raise the levels of the ocean 2-3 ft. Some experts predict more serious effects, including massive flooding of coastal regions, changes in rainfall pattern, expansion of deserts and long term climatic disruptions. Earthly warning signs of global warming are appearing in the Antarctic where the land mass is breaking up at an increased rate and in the mass melting of glaciers in many other parts of the world (Fig. 1). The greenhouse effect has recently been a matter of concern because greenhouse gases appear to be increasing at a rate that could disrupt the temperature balance. In effect, a denser insulation layer will trap more heat energy and gradually heat up the earth.

The debate centers on how the strength of the greenhouse effect is changed when human activities increase the atmospheric concentrations of some greenhouse gases. Bola (2010) in his study stated that global warming is caused by increase in greenhouse gases whose major effects are rise in temperature and sea level. Sozen *et al.* (2007) looked at greenhouse gas emission from sectoral energy consumption, comprising Industry (I), Transport (T), Household (H), Agriculture (A), Service (S) and Other (O). They used the Artificial Neural Network (ANN) approach to develop an equation for estimating greenhouse gas emission in Turkey.

$$E_i = C_{1i}I + C_{2i}T + C_{3i}H + C_{4i}A + C_{5i}S + C_{6i}O + G_i$$

Where:

E_i = The energy consumption i = 1, 2, ..., 6
C = Constant
I, T, H, A, S and O = Various energy sectors

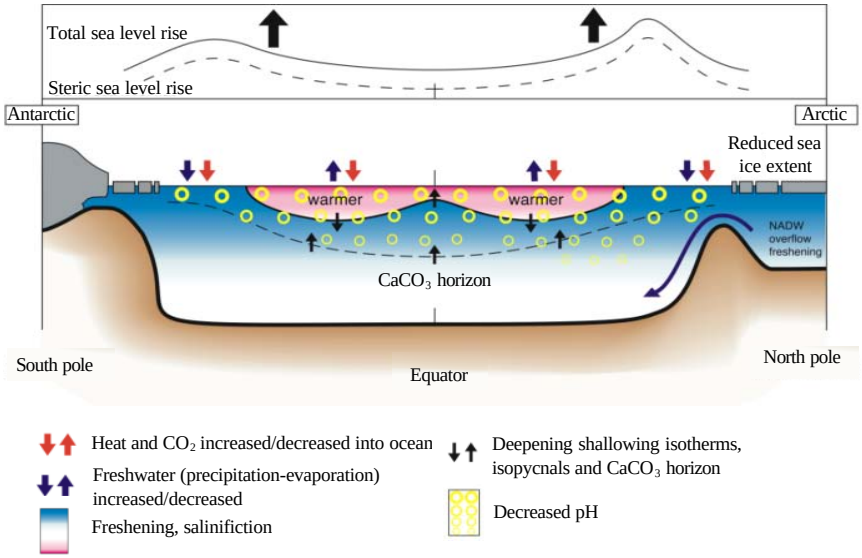

Fig. 1: Example of the rise in sea level due to global warming

Fekedulegn *et al.* (2002) in their study used principal components regression in Dendroecology, they concluded that the method has been recognized as good tool for developing response functions.

MATERIALS AND METHODS

Researchers assume there are p features measured on N observations. Let, X be an N×p matrix of feature measurements and y the N vector of outcome measurements. Researchers assume that the outcome is a quantitative variable. The supervised principal component procedures are as outlined:

- Compute (univariate) standard regression coefficients for each feature
- Form a reduced data matrix consisting of only those features whose univariate coefficient exceeds a threshold θ in absolute value (θ is estimated by cross-validation)
- Compute the first (or first few) principal components of the reduced data matrix
- Use these principal component (s) in a regression model to predict the outcome

The details of the method are as follows, assume that the columns of X (variables) have been standardized to have mean zero. The singular value decomposition of X is written as:

$$X = UDV^T \quad (1)$$

Where:
U = N×m
D = m×m
V = m×p

where m = min (N-1, p) is the rank of X. D is a diagonal matrix containing the singular values d_j, the columns of U are the principal components u_1, u_2, ..., u_m, these are assumed to be ordered so that $d_1 \geq$, $d_2 \geq$, ..., $d_m \geq 0$.

Let, s be the p-vector of standardized regression coefficients for measuring the univariate effect of each variable separately on y:

$$S_j = \frac{x_j^T y}{\|x_j\|} \quad (2)$$

With $\|x_j\| = \sqrt{x_j^T x_j}$. Actually, a scale estimate $\hat{\sigma}$ is missing in each of the S_j but since it is common to all, researchers can omit it. Let, C_θ be the collection of indices such that $|s_j| > \theta$, researchers denote by X_θ the matrix consisting of the columns of X corresponding to C_θ. The SVD of X_θ is:

$$X_\theta = U_\theta D_\theta V_\theta^T \quad (3)$$

Letting $U_\theta = (u_{\theta,1}, u_{\theta,2}, \ldots u_{\theta,m})$, researchers call $u_{\theta,1}$, the first supervised principal component of X and so on. Researchers now fit a univariate linear regression model with response y and predictor $u_{\theta,1}$.

$$\hat{y}^{spc,\theta} = \bar{y} + \hat{\gamma}.u_{\theta,1} \quad (4)$$

Note that since $u_{\theta,1}$ is a left singular vector of X_θ it has mean zero and unit norm. Hence, $\hat{\gamma} = u_{\theta,1}^T y$ and the intercept is $\bar{y}$ the mean of y. Researchers use cross-validation to estimate the best value of θ (Bair *et al.*, 2004).

In this study, researchers consider only the first supervised principal component. Note that Eq. 3:

$$U_\theta = X_\theta V_\theta D_\theta^{-1} = X_\theta W_\theta \quad (5)$$

So for example, $u_{\theta,1}$ is a linear combination of the columns of X_θ: $u_{\theta,1} = X_\theta w_{\theta,1}$. Hence, the linear regression model estimate can be viewed as a restricted linear model estimate using all the predictors in X_θ.

$$\hat{y}^{spc,\theta} = \overline{y} + \hat{\gamma}.X_\theta w_{\theta,1} \quad (6)$$

$$\hat{y}^{spc,\theta} = \overline{y} + \hat{\gamma}.X_\theta \hat{\beta}_\theta \quad (7)$$

Where, $\hat{\beta}_\theta = \hat{\gamma} w_{\theta,1}$. In fact by padding $w_{\theta,1}$ with zero (corresponding to the variables excluded by C_θ) the estimate is linear in all p variables. Given a test feature vector x^*, researchers can make predictions from the regression model as follows:

Researchers center each component of x^* using the means, researchers derived from the data:

$$x_j^* - x_j^* - \overline{x}_j,$$
$$\hat{y}^* = \overline{y} + \hat{\gamma} x_\theta^{*T} w_{\theta,1} = \overline{y} + \hat{\gamma} x_\theta^{*T} \hat{\beta}_\theta$$

Where, x^*_θ is the appropriate sub-vector of x^*. In the case of uncorrelated predictors, it is easy to verify that the supervised principal components procedure has the desired behavior: Bair (opcit), it yields all predictors whose standardized univariate coefficients exceed θ in absolute value. Considering the model below:

$$Y = bo + b_1 xi_1 + b_2 xi_2 + ... + b_p x_p + e_I \quad (8)$$

Suppose the independent variables x_1, x_2, ..., x_k are standardized as follows: x_{ji} is transformed into x_{ji}^s using:

$$x_{ji}^s = \frac{x_{ji} - x_j}{s_{xj}} \quad (9)$$

Where:

s_{xj} = The standard deviation of the independent variable x_j

Superscript s = The independent variables are standardized

The process of standardizing the independent variables allows for an alternative formulation of Eq. 8 as follow:

$$y_1 = \beta_0^s + \beta_1^s \left[\frac{x_{1i} - \overline{x}_1}{s_{x1}}\right] + \beta_2^s \left[\frac{x_{2i} - \overline{x}_2}{s_{x2}}\right] + ... + \beta_k^s \left[\frac{x_{ki} - \overline{x}_k}{s_{xk}}\right] + \varepsilon_i \quad (10)$$

Let, $b^s = (b_1^s, b_2^s, ..., b_k^s)$ be the least squares estimator of β^s. If a data set is used to fit the standardized model in Eq. 10, then the estimate of the coefficients of the model of Eq. 8 can be obtained from the estimates of the coefficients for the standardized variables using the following transformations:

$$b_j = \left[\frac{b_j^s}{S_{x_j}}\right] j = 1,2,...,k \quad (11)$$

And:

$$b_0 = b_{0^s} - \left[\frac{b_1^s \overline{x}_1}{s_{x_1}}\right] - \left[\frac{b_2^s \overline{x}_2}{s_{x_2}}\right] - ... \left[\frac{b_k^s \overline{x}_k}{s_{x_k}}\right] \quad (12)$$

The earlier review indicates that it is always possible to move from one model formulation to another regardless of which model was used for the analysis. The variables considered to be possible correlates of CO_2 emission for 50 selected countries.

Where:
X_2 = Gross Domestic Product (GDP)
X_3 = Industrial output
X_4 = Export output
X_5 = Energy consumption
X_6 = Manufacturing output

Where, the dependent variable Y (CO_2 emission) is the surrogate of global warming.

RESULTS

Using SPSS package, the regression analysis of the variables in Table 1 gave the result with R^2 value of 0.985 indicates that the variables used accounts for 98.5% of the total variation in the response variable that is accounted for by the fitted regression model. The ANOVA table show that the test is significant with p-value of 0.000. It is important to note that the F-statistics value of 575.087 indicates that CO_2 emission (Y) is related to at least some of the input variables. Considering the coefficients of β_i individually, GDP (X_2) has a p-value of 0.042, industrial output (X_3) has a p-value of 0.205 and export output (X_4) has a p-value of 0.249, energy consumption (X_5) is highly significant with a p-value of 0.000 while manufacturing output (X_6) has a p-value of 0.005. Manufacturing output and energy consumption have p<10%.

Table 1: Standardized CO_2 emission data and possible correlates

Countries	CO_2 emi.	GDP	Ind.	Exp.	Enc.	Man.
USA	5.075033	6.11	5.41	-0.04	5.59	5.08
China	4.124944	0.64	1.80	-0.12	3.23	2.75
Russia	0.938820	-0.11	-0.05	-0.17	1.15	0.00
India	0.772408	-0.05	-0.08	-0.18	0.92	-0.14
Japan	0.694703	2.15	2.89	-0.12	0.82	2.77
Germany	0.284051	1.09	1.36	-0.07	0.36	1.31
Canada	0.129219	0.11	0.22	-0.15	0.13	0.14
UK	0.081551	0.75	0.77	-0.11	0.05	0.66
S. Korea	-0.029630	-0.05	0.12	-0.16	-0.02	0.13
Italy	-0.043980	0.50	0.56	-0.14	-0.09	0.57
Mexico	-0.054880	-0.06	-0.11	-0.17	-0.15	-0.10
S. Africa	-0.055790	-0.32	-0.37	-0.19	-0.26	-0.37
Iran	-0.058950	-0.35	-0.35	-0.19	-0.21	-0.44
Indonesia	-0.109520	-0.30	-0.23	-0.18	-0.14	-0.24
France	-0.113690	0.71	0.51	-0.13	0.15	0.43
Brazil	-0.151990	-0.10	0.02	-0.18	-0.06	-0.30
Spain	-0.153180	0.14	0.19	-0.16	-0.21	0.05
N. Zealand	-0.426460	-0.38	-0.46	-0.19	-0.31	-0.42
Australia	-0.156600	-0.08	-0.20	-0.18	-0.27	-0.30
S. Arabia	-0.173390	-0.30	-0.14	6.77	-0.21	-0.42
Poland	-0.174440	-0.30	-0.35	-0.18	-0.33	-0.36
Thailand	-0.210240	-0.35	-0.35	-0.18	-0.34	-0.30
Turkey	-0.248600	-0.27	-0.38	-0.18	-0.37	-0.38
Algeria	-0.277960	-0.39	-0.41	-0.19	-0.43	-0.42
Malaysia	-0.292970	-0.37	-0.37	1.32	-0.43	-0.37
Venezuela	-0.297510	-0.38	-0.42	-0.19	-0.44	-0.44
Egypt	-0.310660	-0.40	-0.46	-0.19	-0.24	-0.42
UAE	-0.318930	-0.38	-0.38	-0.18	-0.39	-0.30
Netherlands	-0.325450	-0.12	-0.18	-0.15	-0.36	-0.26
Argentina	-0.325700	-0.35	-0.40	-0.19	-0.42	-0.38
Pakistan	-0.340430	-0.39	-0.38	-0.18	-0.39	-0.28
Czech	-0.348370	-0.38	-0.43	-0.19	-0.42	-0.42
Nigeria	-0.351080	-0.41	-0.42	-0.19	-0.31	-0.36
Belgium	-0.363240	-0.24	-0.32	-0.16	-0.32	-0.33
Greece	-0.366920	-0.32	-0.42	-0.19	-0.50	-0.44
Israel	-0.390190	-0.37	-0.37	-0.19	-0.33	-0.45
Austria	-0.391470	-0.28	-0.32	-0.18	-0.29	-0.34
Chile	-0.398260	-0.39	-0.43	-0.15	-0.34	-0.30
Hungary	-0.403040	-0.38	-0.44	-0.19	-0.44	-0.43
Colombia	-0.406290	-0.38	-0.46	-0.18	-0.15	-0.30
Sweden	-0.406840	-0.25	-0.30	-0.17	-0.30	-0.35
Denmark	-0.406910	-0.30	-0.40	-0.18	-0.29	-0.40
Singapore	-0.407550	-0.38	-0.44	-0.18	-0.31	-0.40
Switzerland	-0.418330	-0.24	-0.33	-0.17	-0.29	-0.31
Hong Kong	-0.421120	-0.35	-0.32	-0.18	-0.31	-0.42
Norway	-0.375230	-0.30	-0.30	-0.18	-0.21	-0.44
Philippines	-0.381720	-0.39	-0.46	-0.19	-0.43	-0.43
Finland	-0.395170	-0.34	-0.40	-0.19	-0.30	-0.39
Portugal	-0.401470	-0.35	-0.43	-0.19	-0.51	-0.43
Ireland	-0.416600	-0.34	-0.38	-0.17	-0.30	-0.35

Calculated from list of countries by their CO_2 emission while other variables where obtained from The Economist Fact Book (2007) and CDIAC (2006)

The significance of the parameters of the model is tested using the hypothesis: H_0: $\beta_i = 0$ vs. H_1: $\beta_i \neq 0$. From the regression analysis output, the p-value of β_3 is 0.205 and the value of this t is given as -1.285.

The p-value is, therefore p-value = $2\times p\ (X \geq 1.285)$ where the random variablex has a t-distribution with n-k-1 = 43 degrees of freedom. From the computer output, the p-values of 0.205 and 0.249 are >10% indicating that the corresponding input variables, Industrial output (X_3) and export output (X_4), respectively do not give a good fit, therefore can be dropped from the model. Using backward elimination method, new result with k-1 input variable was obtained. The input variables left becomes the GDP (X_2), energy consumption (X_5) and manufacturing output (X_6). Analysis show that all these variables are significant with p-value of 0.000. The model generated is:

$$\hat{Y} = 0.001 - 0.390X_2 + 1.014X_5 + 0.336X_6$$

Using a statistical package (MINITAB), the principal component analysis yielded the following results:

$$Y_1 = 0.470x_2 + 0.487x_3 - 0.019x_4 + 0.481x_5 + 0.492x_6$$
$$Y_2 = -0.070x_2 - 0.069x_3 - 0.976x_4 + 0.006x_5 - 0.010x_6$$
$$Y_3 = -0.323x_2 - 0.195x_3 + 0.214x_4 + 0.133x_5 - 0.090x_6$$
$$Y_4 = 0.087x_2 - 0.373x_3 - 0.012x_4 + 0.809x_5 - 0.393x_6$$
$$Y_5 = 0.738x_2 - 0.005x_3 + 0.001x_4 - 0.286x_5 - 0.556x_6$$
$$Y_6 = -0.342x_2 + 0.762x_3 - 0.027x_4 + 0.122x_5 - 0.535x_6$$

Where, Y_i (i = 1, 2, ..., 6) the linear combination of the original variables are the principal components. The variances (var Y_i) = λ_i are as follows:

$$\lambda_1 = 3.9973,\ \lambda_2 = 1.0059,\ \lambda_3 = 0.8715$$
$$\lambda_4 = 0.0890,\ \lambda_5 = 0.0343,\ \lambda_6 = 0.0019$$

Components 1-3 demonstrated eigen-values of 3.9973, 1.0059 and 0.8715, respectively. It can be seen that these three components demonstrated eigen values >1. This means that three components will be retained by the eigen value one criterion method.

From the principal component analysis result, it can be seen that the proportion of the total variance explained by the first principal component is 66.6%, second principal component is 16.8% and third is 14.5%. The cumulative proportion of the total variance explained by the first to third principal component is 97.9%.

This indicates that the three principal components should be retained. The selected principal components are:

$$PC_1(Y_1) = 0.470x_2 + 0.487x_3 - 0.019x_4 + 0.481x_5 + 0.492x_6$$
$$PC_2(Y_2) = -0.070x_2 - 0.069x_3 - 0.976x_4 + 0.006x_5$$
$$PC_3(Y_3) = -0.323x_2 - 0.195x_3 + 0.2146x_4 + 0.133x_5$$

Considering the coefficients of the original variables in the selected principal components above, it was observed that the coefficients of the original variable in

the first selected principal component are all positive, manufacturing output has the highest weight, though the coefficient of export output (X_4) is negative and negligible. The coefficients of the original variables in the second selected component are all negative except energy consumption (X_5) but export output (X_4) has the highest weights. In the third selected component, GDP (X_2) has the highest coefficient.

Where, PC_1-PC_3 are the selected components. Principal component regression often utilizes explanatory variables that are standardized so that X′X is proportional to the correlation matrix (Marx and Smith, 1990). The regression analysis result of these selected components show R^2 value of 0.973 indicating that the reduced variables accounts for 97.3% of the total variation in the response variable that is accounted for by the fitted regression model. The principal component regression model generated is:

$$\hat{Y} = -0.584 + 0.003PC_1 + 0.001PC_2 + 0.005PC_3$$

The principal components regression built in the earlier model involved the linear combination of all the original variables.

However, the mention earlier method has no direct relationship with the original variables, hence researchers compute supervised principal component regression. Using the supervised principal component, considering only the variables that are significant based on the correlation coefficient and the simple regression analysis of each variable with CO_2 emission, the selected variables are GDP, industrial output, energy consumption and manufacturing output with β coefficients of 0,799 for GDP 0.847 for industrial output, 0.983 for energy consumption and 0.892 for manufacturing output.

The supervised principal component analysis result selected only one component. The selected principal component is:

$$Y_1 = 0.0511x_2 + 0.6715x_3 + 0.0742x_5 + 0.7355x_6$$

The supervised principal component regression analysis yielded the following result:

$$\hat{y}^{spc,\Theta} = \bar{y} + \hat{\gamma}.X_\Theta\hat{\beta}_\Theta$$

$$\hat{y}^{spc,\Theta} = 498055.2 + 0.0018G - 0.023I - 0.0026E + 0.0254M$$

Where:

$\bar{y}$ = The intercept (the mean of CO_2 emission Y)
G = GDP
I = Industrial output
E = Energy consumption
M = Manufacturing output

DISCUSSION

Global warming is a topical issue that needs attention of people from various fields of study. Its measurement seems to be a bit challenging, since no one can actually measure the thin air.

An alternative is to explore some other variables that seem to have some degree of relationship with global warming and build a model with those variables. The model built could serve as a guide in measuring global warming. In this study, regression analysis and principal component analysis have been explored in building a model for measuring global warming. The regression analyses result with R^2 value of 0.985 indicates that the variables used account for 98.5% of the total variation in the value of the response variable that is accounted for by the fitted regression model. The high R^2 value is a classical symptom of multicollinearity. Due to the noticed effect of multicollinearity among these variables, the principal component analyses introduced reduced the high dimension variable to small dimension which were used to replace the correlated variables with uncorrelated linear functions of the original variables. The regression analysis carried out with these principal components so formed yielded better results. The principal component regression analysis gave R^2 value of 0.973 indicating that the reduced variables accounts for 97.3% of the total variation in the value of the response variable that is accounted for by the fitted regression model.

CONCLUSION

This study has shown that models built using supervised principal component regression gave a good fit. As such any prediction made using this model is likely to yield a good result. Based on this study in measuring global warming, the selected variables used in this study could be considered, since the R^2 value has shown that they gave a good fit and the model developed could be used to estimate the quantity of CO_2 emission.

ACKNOWLEDGEMENT

Researcher is grateful to Dr. Evelyn Okeke for her assistance in some aspects of the computing.

REFERENCES

Bair, E., T. Hastie, D. Paul and R. Tibshirani, 2004. Prediction by supervised principal components. http://www-stat.stanford.edu/~tibs/ftp/spca.pdf

Bola, B., 2010. Cause and effect for global warming in Nigeria. http://www.goodlife.com.ng.

CDIAC, 2006. List of countries by carbon dioxide emission. Data Collected by the US Department of Energy's Carbon Dioxide Analysis Center.

Fekedulegn, B.D. J.J. Colbert, R.R. Hicks Jr. and M.E. Schuckers, 2002. Coping with multicollinearity: An example on application of principal components regression in dendroecology. USDA Forest Service. http://www.fs.fed.us/ne/morgantown/4557/dendrochron/rpne721.pdf.

Marx, B.D. and E.P. Smith, 1990. Principal component estimation for generalized linear regression. Biometrika, 77: 23-31.

Sozen, A., Z. Gulseven and E. Arcaklioglu, 2007. Forecasting based on sectoral energy consumption of GHGs in Turkey and mitigation policies. Energy Policy, 35: 6491-6505.

Talaro, K.P. and A. Talaro, 2003. Foundations in Microbiology. 2nd Edn., McGraw-Hill Education, New York, USA., pp: 231-235.

The Economist Fact Book, 2007. List of countries by their GDP. Export Output, Energy Consumption, Industrial Output and Manufacturing Output.