PRINCIPAL COMPONENTS TO OVERCOME MULTICOLLINEARITY PROBLEM

Abubakari S.Gwelo
Department of Mathematics and Statistics studies, Mzumbe University, Tanzania
abugwelo@gmail.com

Abstract: *The impact of ignoring collinearity among predictors is well documented in a statistical literature. An attempt has been made in this study to document application of Principal components as remedial solution to this problem. Using a sample of six hundred participants, linear regression model was fitted and collinearity between predictors was detected using Variance Inflation Factor (VIF). After confirming the existence of high relationship between independent variables, the principal components was utilized to find the possible linear combination of variables that can produce large variance without much loss of information. Thus, the set of correlated variables were reduced into new minimum number of variables which are independent on each other but contained linear combination of the related variables. In order to check the presence of relationship between predictors, dependent variables were regressed on these five principal components. The results show that VIF values for each predictor ranged from 1 to 3 which indicates that multicollinearity problem was eliminated. Finally another linear regression model was fitted using Principal components as predictors. The assessment of relationship between predictors indicated that no any symptoms of multicollinearity were observed. The study revealed that principal component analysis is one of the appropriate methods of solving the collinearity among variables. Therefore this technique produces better estimation and prediction than ordinary least squares when predictors are related. The study concludes that principal component analysis is appropriate method of solving this matter.*

Keyword: principal components, multicollinearity, variance inflation factor.

JEL Classification: C01, C02.

1. Introduction

Advancement of science and technology raised attention to make proper decision on several critical matters facing different sectors. Attentions to make right decision have increased a need for utilizing realistic information. With the presence of multidimensional data across the world, managers, planner and policy makers are facing challenges of managing huge data set from different sources that may produce little information due to presence of some redundant information. Accordingly the information obtained from huge numeric data set is important for practical investigation. This is directly linked to the application of linear regression model where by a relationship among predictors can be handled as redundancy. In both social and scientific researches, a need may raise to build a model for studying the relationship between two or more variables so as to measure the internal influence of one variable on others. In such kind of situation it is required to select appropriate technique that fits the interest of the research. There are different statistical techniques that measure relationship, among them includes path analysis, structural equation model, discriminant analysis and linear regression model. However the easiest and popular techniques of studying relationship among variables is linear regression model which focus on treating one variable as “dependent variable” in such way that its change is influenced by the changes of independent variables.

Linear regression model is one of the methods of analyzing relationship between variables which is widely applicable in many researches of different disciplines. Ordinary Least Square procedures of estimating linear regression model assume the predictors are uncorrelated. Violation of this assumption implies that the predictors are correlated, the situation which bring about large standard error, reduce precision of results and weaken statistical power (Gujarati, 2004). The existing literature emphasis that the consequences of the multicollinearity are: large standard error of the estimated coefficients, large probability values with low test statistic that result into unbiased estimates which lastly mislead interpretation and conclusion (Mason ,1987; Mela and Kopalle, 2002; Hoffman, 2010). In some situation, quantitative variable may be predicted based on several independent variables so as to examine the significance influence of each explanatory variable. One of the appropriate approaches in such case is employing multiple linear regressions to assess how well it fits the actual phenomenon. According to Gauss - Markov theorem, when all assumptions of the classical linear regression model are met, the model will produce unbiased estimates that have minimum variance compared to the rest of unbiased linear estimators (Kutner, et al., 2005). When the predictors are correlated their influence on the main model is overlapped. Despite the fact that the multicollinearity does not harm the goodness fit of the model, it results into wrong conclusion when the target is to predict the effect or contribution of each explanatory variable to the model. This hinders the interpretation of coefficient of predictor as a measure of degree of change of predicted variable with a unit change of predictor holding other variables constant. As a result of this problem, the contribution of each predictor can be unrealistic due to overlapping of variables. Since presence of multicollinearity may mislead the analyst and fall in wrong conclusion of the results, it is imperative to present solution to this problem. While other methods of solving multicollinearity are well documented, there is scarce of literature on how principal components methods can be utilized to eliminate multicollinearity. The key objective of this study was to investigate the application of principal components in handling multicollinearity. In the situation where multicollinearity cause problem on estimating parameters of linear regression model, the Principal components can be applied to reduces the dimensions of parameters into small number of parameters while retaining the maximum variance of the original data. Instead of running model with original correlated variables, the Principal components are then used as new independent variables that eliminated the interdependency of variables among predictors.

2. Literature review

Multicollinearity is a problem which is associated with high relationship among explanatory variables where by specific effect of the correlated variables on dependent variable cannot be separated. According to Wooldridge (2010) ,quick judgment of the existence of multicollinearity when running multiple regression model lies on several observations; (1) high value of coefficient of determination (R-square) which measures the proportion of the explained data in the model ; (2) high sensitivity of the estimates where by small change in data bring about large changes to the estimates of the population parameters ;(3) high value of standard error of estimated coefficients with low value of test statistic (t-values) and high p-values ;(4) overall significance of the overall model while individual variables are insignificant. With these impacts, the estimates of the parameters in the classical model will vary from one sample to another when the analysis is repeated with series of samples.

The concept of multicollinearity some time can be explained with a concept of “orthogonality” which means independence among variables. This condition can be attained when the eigen values have length one (1). Presence of dependence among predictors affects the ability of the model to estimate the actual phenomenon. This is due to presence of redundant

information in some explanatory variables where by more than one variable can provide similar information on predicting response variable. In this situation, redundant variables are not realistic in such a way that one coefficient of variables can measure the influence of such variable to the response variables, and the same influence goes to other variable. In actual practice it may be observed that some predictors are correlated and this does not harm the analysis due to fact that mutlicollinearity is a matter of degree of relationship that exist among variables and not the absence of relationship at all. While the relationship between variables may exist, the key point is to observe the degrees of multicollinearity that cannot affect the results of estimated parameters. At population level independent variables constructs are not collinear thus the multicollinearity problem happen due to technical expertise of the researcher particularly in sampling process (Kmenta ,1997).In behavioral research that involves estimation of behavioral constructs it is very rare to find the explanatory variables which are not related with each other. Despite the fact that the multicollinearity is either originated from sampling error or true population, it can be detected during data analysis. There are two types of multicollinearity:

Perfect/exact multicollinearity: the relationship between variable are said to be perfect if the value of correlation coefficient is exactly 1 or -1. When there is a perfect multicollinearity, the data matrix does not exist and thus the inverse of matrix formed as a result of cross product between the data matrix and transposed matrix will not exist. In practice, there is no situation where correlation coefficient between variables can exactly be one (1). In the linear regression model having several explanatory variables (say k), a linear relationship between explanatory variables may exist if the following condition is satisfied:

$$\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} +,.....\alpha_k x_{ik} = 0 \quad (1)$$

Where: $\alpha_0, \alpha_1, ..., \alpha_k$ are constants and cannot all be zero at the same time

x_{ij} = stands for the ith observation on the jth explanatory variable

Equation one (1) can be expressed as:

$$x_{i1} = \frac{1}{\alpha_1}\left(\alpha_0 - \alpha_2 x_{i2} -,.....,-\alpha_k x_{ik}\right) \quad (2)$$

Equation (2) implies that x_{i1} is linearly related with $x_{i2}, x_{13}, ..., x_{ik}$

Near multicollinearity. This happens when the predictors are highly related as indicated by correlation coefficient values being closer to one. The inverse matrix of the cross product of the data matrix and its transposed matrix exist, implies that the determinant matrix is not zero. The common situation is when one predictor is either highly or lowly correlated with another one. There is no uniform definition and agreement on degree of correlation suitable for judging the presence of high correlation. That means at what value the variables can be considered as highly related. In practice the value of correlation coefficient greater than or equal to 0.5 is considered as high correlation. If the explanatory variables are highly related with each other but not perfect relationship, the near multicollinearity can be expressed as follows:

$$\alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} +,.....+\alpha_k x_{ik} + \varepsilon_i = 0 \quad (3)$$

Where: $\alpha_0, \alpha_1, ..., \alpha_k$ are constants and cannot be zero at the same time. ε_i Is the error term.

Equation (3) can be expressed as follows:

$$x_{i1} = \frac{1}{\alpha_1}\left(\alpha_0 - \alpha_2 x_{i2} -,.....,-\alpha_k x_{ik} - \varepsilon_i\right) \quad (4)$$

Equation four (4) implies that x_{i1} is not exactly linearly related with $x_{i2}, x_{13}, ..., x_{ik}$
The principal component technique is one of the remedial solutions to multicollinearity. The technique was presented by Pearson (1901) and Hotelling (1933), which is based on finding the possible combination of merging the correlated variables into new few uncorrelated variables by reducing the original matrix of high dimension into low dimension whose rows and columns are independent of each other. It is multivariate statistical method which is applied in different disciplines where there is a need to reduce multidimensional data set from huge attributes to a reasonable composite attributes. It reduces the complexity of information which is hardy to be interpreted into easier and meaningful interpretation. This procedure is done by transformation of large dimension of interrelated variables into smaller set of uncorrelated variables knows as principal components (PCs). Each Principal component is a linear combination of the original attributes with their coefficients indicate the relative importance to the component. The PCs are listed in order of preference where by only few of them are retained under condition that they account to a maximum variability of the original data set.

3. Material and methods

3.1. Problem formulation

3.1.1. Multiple linear regression

Different social and scientific phenomenons are complex to be understood that need several variables for demonstration. There are different methods of studying relationship based on whether the data are categorical or numeric in nature. Regression analysis refers to a statistical technique for studying relationship among variables and influence of one variable over the others (Montgomery, Peck and Vining, 2012).Thus the multiple linear regression model is a statistical technique which is applicable when someone needs to study the relationship between dependant variable and at least two independent variables. It requires the condition that the dependent variable to be numeric in nature while independent variables may either be numeric or a mix of both categorical and numeric data. The model can be described as follows:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \qquad i = 1,2,.....n \quad (5)$$

Equation (5) can be presented in simplified form as:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \quad (6)$$

Where:

β_0 is a constant coefficient which is also referred to as y- intercept in the first order classical line regression model. It is the value of "y" when the explanatory variables are zero.

β_i is the coefficient of the 'i'th variable which indicates the significant change of y_i for a unit change of x_i keeping other variables constants

P is the number of predictors.

The multiple linear regression models can also be presented in matrix form:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{bmatrix} \quad (7)$$

The main feature in running regression model is to find the goodness fit of the model in a given dataset. The Ordinary Least Square Methods is used to obtain the estimates that minimize the sum of total squared error as described in equation (8).

$$\sum_i \left(y_i - \hat{y}_i\right)^2 = \sum_i \left(y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + ... + \hat{\beta}_p x_{ip}]\right) \quad i = 1,2,.....,n \quad (8)$$

Where by: $\hat{y}_i$ stands for predicted model

Like any other statistical model, the first step in fitting linear regression model is to assess the significance of the overall model, checking on whether model assumptions are met or not.â

3.1.2. Assumptions of the model

It is obviously that statistical techniques are based on several assumptions that need to be tested to guarantee further analysis. If it happens that any of the suggested assumptions have been violated, the results can be characterized with unbiased and inconsistent estimators unlike the robust assumptions which do not bring effects to the results.

Residuals are independent and follow normal distribution. This assumption can be tested by observing histogram or predicted probability (p-p) plot. Violation of this assumption does not lead to bias or inefficiency of the estimators. It only harms calculation of probability (p) values useful for checking the significance However; this assumption can be met by ensuring the sample size is large enough. Central limit theorem implies that as the sample size increases, the sampling distribution of the mean can be approximated by a normal distribution. This assumption is robust and there is no need of testing this assumption when running linear model from reasonable large sample.

Linearity assumption among variables .The model assumes an existence of linear relationship between predictor and predicted variables. This assumption can be checked by using scatter plot where by the distribution of patterns can give a clear picture of linearity. When variables have linear relationship, the model can easily estimate the degree and magnitude of the relationship, but in different situation some variables are not linearly related, for this kind of situation the model may underestimate the true relationship which results into wrong conclusion

Homoscedasticity assumption. This implies that variance of the residual is constant and does not varies across predictors. If the variances of the errors are not the same around regressions, it indicates presence of heteroscedasticity. This assumption can be examined using scatter plot between residuals and predicted variables. Violation of this assumption harm analysis and lead to possibility of omitting type one error, the situation occurs when null hypothesis is wrongly rejected while it is true.

Mutlicollinearity assumption. Multiple linear regression models assume that the independent variables are not highly related with each other. This assumption can be tested using different indicators including correlation matrix, Eigen values and Variance Inflation Factor (VIF).Presence of this problem weakens the precision of estimated coefficients of the model since the standard errors were enlarged which result into low values of test statistic with high probability values(p-value) .Hence reduction of statistical power of the test.

3.2. Diagnostic of multicolinearity

There are different ways of observing collinearity among predictors such as the use correlation matrix, eigen values and Variance Inflation Factor (VIF).How ever only VIF was considered in this study. The justification of using this method is due to fact that the computation and result can be interpreted easily and clearly. The other methods namely correlation matrix and eigen values can give highlights of presence of multicollinearity but cannot give a clear estimate of the degree of multicollinearity (El-Dereny and Rashwan, 2011).

3.2.1. Variance inflation factor

Presence of collinearity among explanatory variables leads to increase in its standard errors as a result of the variance of coefficients of explanatory variables being inflated. The Variance Inflation factor gives the degree to which the variance was inflated.
The variance inflation factor for predictor k is given by:

$$VIF_k = \frac{1}{1-R_k^2} \quad \text{..... (9)}$$

Where R_k^2 stands for the coefficient of determination when Predictor X_k is treated as dependent variable and regressed on other predictors.VIF gives an index which measures the degrees of increase of variance of regression coefficient with the increase of multicollinearity. As a rule of thumb, multicollinearity is considered a problem when $VIF \geq 10$ (Cohen et al, 2004)

3.3. Principal components analysis

Principal component analysis is the multivariate statistical technique that transforms a set of p- variables which were related into new small dimension of independent variables without much loss of information (Johnson and Wichen, 2007). Principal component analysis provide a unique procedure of handling multicollinearity with its procedure of obtaining new independent variables from original data set of interrelated variables. Principal components analysis involve rotation of huge multidimensional data from one point to another point of orthogonal linear axes known as Principal components (PCs).Thus the first linear combination is the principal component that explains the most variability of the original data set. Similarly the second Principal component is orthogonal to the first PC and accounts for the most of the remained variance of the observation.

Suppose a random vector $X' = \lfloor X_1, X_2, ..., X_o \rfloor$ have covariance matrix $\sum$ with eigen values $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ..... \geq \lambda_p \geq 0$, then consider the following linear combination:

$$\begin{aligned} Y_1 &= l_i'X = l_{11}X_1 + l_{21}X_2 + ...+ l_{p1}X_P \\ Y_2 &= l_i'X = l_{12}X_2 + l_{22}X_2 + ...+ l_{p2}X_P \\ &\vdots \\ Y_P &= l_i'X = l_{1p}X_P + l_{2p}X_2 + ...+ l_{pp}X_P \end{aligned} \quad \text{..... (10)}$$

Then

$$\begin{aligned} Var(Y_i) &= l'\sum l \qquad i = 1,2,.....,p \\ Cov(Y_iY_k) &= l_i'\Sigma l_k \qquad i.k = 1,2,.....p \end{aligned} \quad \text{..... (11)}$$

Thus the principal components are the linear combinations in equation (10) whose variances are as large as possible. The ith principal component is the linear combination $l_i'X$ that maximizes $Var(l_i'X)$ subject to $l_i'l = 1$

The procedures of principal components analysis are; (1) checking the key assumptions of the model; (2) exploration of the preliminarily components followed by extraction of the significant components that have maximum variances to be retained in the model;(3) smoothing data through rotation of variance covariance matrix for easy interpretation; (4) the final procedure is interpretation of the outputs.The main issue in principal components analysis relies on decision of which eigen value to be retained for further analysis. The maximum variation can be achieved by selecting the significant eigen values which exceed one while at the same time ignoring insignificant eigen values whose values are lower than one. As documented by Farebrother (1999),the principal components analysis merge related variables into small dimension of variables, then these principal components can be treated as explanatory variables to run regression model by predicting dependent variable on these new variables.

4. Application

4.1. Detecting multicollinearity

In assessing the effectiveness of Principal components in overcoming multicollinearity, determinants of student satisfaction were analyzed with a representative sample of 600 students from one of the public University in Tanzania. Students were requested to rate their perceived satisfaction on different items. The multiple linear regression model of predicting determinants of student satisfaction was formulated as:

$$Y = \alpha_1X_1 + \alpha_2X_2 + ...+, \alpha_{14}X_{14} + \varepsilon \quad \text{..... (12)}$$

Where: Y stands for satisfaction. $X_1, X_2, ...X_{14}$ are predictors, ε is the error term.

The results from table 1 indicate that the overall model is significant (F-value, 51.888, p-value, 0.000). Specifically, an assessment of statistical significance of individual predictors indicates that some predictors are significant while others are insignificant. The major reason is presence of collinearity among predictors. The value of Variance Inflation Factor above 10 indicates presence of multicollinearity except for variable number five and six only whose values are less than 10.

Table 1: Significance of the predictors (before removal of multicollinearity)

Variables	Unstandardized Coefficients		Standardized Coefficients	T	Sig.	Collinearity Statistics	
	B	Std. Error	Beta			Tolerance	VIF
Constant	.782	.167		4.699	.000		
Lighting in class rooms	.515	.124	.405	4.139	.000	.079	12.640
Appearance of the buildings	-.029	.102	-.031	-.281	.779	.064	15.734
Comfortable temperature in class rooms	.594	.083	.812	7.125	.000	.058	17.121
Internet accessibility	-.867	.118	-.784	-7.372	.000	.067	14.882
Efficiency of registration	-.469	.141	-.274	-3.320	.001	.111	8.989
Efficiency of records keeping	.491	.081	.514	6.039	.000	.105	9.550
Availability of the channels for complains	-.748	.159	-.730	-4.715	.000	.032	31.553
Efficiency in dealing with queries	.507	.128	.609	3.972	.000	.032	30.984
Staffs interaction with students	-.771	.236	-.717	-3.273	.001	.016	63.251
Proficiency of Lecturers for teaching and research	.576	.156	.794	3.686	.000	.016	61.148
Availability of personnel to help students	-.057	.069	-.080	-.832	.406	.083	12.050
Availability of Lecturers for consultation and assistance	-.069	.134	-.048	-.514	.607	.085	11.716
Management focus on students	1.654	.208	1.766	7.937	.000	.015	65.180
Availability of private study rooms	-1.080	.152	-1.625	-7.126	.000	.015	68.480
F-value= 51.888, p-value=0.000							

5. Principal component analysis

In principal components method, new artificial variables may either be considered as significant to be included in the model or insignificant to be excluded in the model based on several criteria. There is no agreed limited number of principal components to be retained due to fact that the number varies since different criteria can be employed for such decision. The popular and easiest way of deciding maximum number of principal components to be retained is through observing cumulative variance. Another rule of determining the number of components to be included in the analysis is based on assessing the eigen values as proposed by Kaiser (1960).The rule suggests that significant components are considered as

important when the eigen values are either greater than or equal to one. Another important useful criterion for deciding the number of PCs to be retained is based on observation of visual appearance of scree plot. The arrangement of eigen values are arranged in descending order from largest to lowest values which are presented on y-axis against number of PCs. The break of the plot is the key point which determines the maximum number of PCs to be retained. The significant components are those listed on the left side before the break point of the scree plot, while insignificant components are those appeared on the right hand side of the break point where the plot flattens out. Determining the maximum eigen values based on scree plot is subject to researcher judgment and some time the break point cannot seen clearly. However this criterion is considered useful in providing accurate result when the sample size is large enough and recommended over 200 observations (Stevens, 2012).Despite of the uniqueness of these criteria, it is recommended by Jolliffe (2002) that combining more than one rule is better than sticking on a single rule. This paper utilized the rules of judging significant PCs based on observing the eigen values and cumulative variance indicated by eigen values. Thus interpretation of the principal components results depend on several set of attributes to be studied, variety of entities, degree of relationship among interrelated variables, and the criteria used in judgment

Table 2 presents principal components together with its corresponding eigen values and total variance explained. The principal components in this attributes of satisfaction are uncorrelated attributes in the original data set. The eigen values were listed in descending order from largest to smallest value. The eigen values of the most of principal components decrease downwards and many PCS have small values. Each of the extracted principal components presents the maximum portion to the total variability of the original dataset. Insignificant components with eigen values <1.0 represent relatively low variance than significant components with eigen values ≥ 1.The first principal component (PC) has largest variance that account for 32.8% of the total variance. This PC has comparatively largest eigen value of 4.6 which is equivalent to the eigen values of four variables. The second PC has an eigen value of 3.4 that accounts to 24.5% of the variability of the data. The third PC explains 11.9% of the total variance of the original data. Fourth and fifth PCs explain 9.9% and 9.5% of the total variance respectively. As a rule of thumb, the first five components each have eigen value greater than one and collectively account to 88.5% of variability of the original data set losing only 11.5% of the information .Therefore only five PCs are extracted and retained from fourteenth PCs without much loss of information. This implies that the original information was reduced from 14-dimension of data set into a minimum size (5-diemension) while at the same time maximizing the variability of the original dataset .The rest remaining 9 PCs are considered as insignificant and redundant since they have eigen values <1.0 and contributes to a small portion of the total variance of the original data set. In other word, the Principal components with smallest eigen values are treated as observational error, hence they are removed from analysis.

Table 2: Total variance explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.587	32.762	32.762	4.587	32.762	32.762	3.028	21.629	21.629
2	3.427	24.476	57.239	3.427	24.476	57.239	2.790	19.932	41.561
3	1.658	11.846	69.085	1.658	11.846	69.085	2.202	15.726	57.286
4	1.386	9.898	78.983	1.386	9.898	78.983	2.196	15.685	72.971
5	1.327	9.480	88.463	1.327	9.480	88.463	2.169	15.492	88.463
6	.834	5.956	94.419						
7	.511	3.652	98.071						
8	.098	.698	98.769						
9	.060	.427	99.195						
10	.044	.313	99.508						
11	.035	.252	99.760						
12	.017	.123	99.883						
13	.011	.080	99.963						
14	.005	.037	100.000						

The findings from Table 3 indicate the outputs of varimax methods of rotation which is used to smooth the loadings and hence simplify interpretation. In this study the dimension of data set is large enough, thus observing the high loading of eigen vectors to a particular PC is subject to error and need much attention since none of the vectors are zero .The remedial solution is to rotate the vectors so as to make clear interpretation of the data. After utilizing this method it is therefore seen clearly that each variables map into a particular PC that presents composite variable. The outputs of varimax provide clear interpretation of the PCs in a way that only high loadings are retained to specific components and the low loadings are minimized. This improves impression of output by identifying the variables that are highly related to a corresponding PC.

Table 3: Rotated Component Matrix

Variables	Component				
	1	2	3	4	5
Lighting in class rooms (X_1)		.671			
Appearance of the buildings(X_2)		.666			
Comfortable temperature in class rooms(X_3)		.927			
Internet accessibility (X_4)		.920			
Efficiency of registration(X_5)	.928				
Efficiency of records keeping(X_6)	.920				
Availability of the channels for complains(X_7)					.943
Efficiency in dealing with queries(X_8)					.927
Staffs interaction with students (X_9)				.954	
Proficiency of Lecturers for teaching and research(X_{10})				.960	
Availability of personnel to help students (X_{11})			.933		
Availability of Lecturers for consultation and assistance(X_{12})			.944		
Management focus on students(X_{13})	.786				
Availability of private study rooms (X_{14})	.794				

Thus the five selected principal components are the linear combination of the original variables that contribute much to the total variance. Thus the fitted PCs are:

$$Z_1 = 0.928X_5 + 0.920X_6 + 0.786X_{13} + 0.794X_{14} \quad (13)$$

$$Z_2 = 0.671X_1 + 0.666X_2 + 0.927X_3 + 0.920X_4 \quad (14)$$

$$Z_3 = 0.933X_{11} + 0.944X_{12} \quad (15)$$

$$Z_4 = 0.954X_9 + 0.960X_{10} \quad (16)$$

$$Z_5 = 0.943X_7 + 0.927X_8 \quad (17)$$

The first principal component is the linear combination of four variables which are related namely variable 5, 6, 13 and 14.The second principal components composed of variable 1, 2, 3 and 4 which are highly related. The third PC is formed by variable 11 and 12 while fourth PC consists of variable 9 and 10.The firth PC includes variable 7 and 8.The variables within components are highly related while the group of variables in a particular component are not related with another group of variables loaded to a another component. After removing the principal components which are less important, the modified liner regression model is now:

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \beta_4 Z_4 + \beta_5 Z_5 + \varepsilon \quad (18)$$

Where predictors $Z_1, Z_2, Z_3, \ldots\ldots, Z_5$ are principal components

Table 4: Significance of the predictors (after removal of multicollinearity)

	Unstandardized Coefficients		**Standardized Coefficients**	**T**	**Sig.**	**Collinearity Statistics**	
Variables	B	Std. Error	Beta			Tolerance	VIF
(Constant)	.970	.179		5.414	.000		
Z_1	.470	.041	.428	11.494	.000	.816	1.225
Z_2	.426	.048	.307	8.789	.000	.925	1.081
Z_3	-.143	.033	-.154	-4.348	.000	.896	1.116
Z_4	.156	.030	.198	5.157	.000	.770	1.299
Z_5	-.144	.031	-.173	-4.587	.000	.796	1.256
F-value=56.972		p-value=0.000					

Instead of using Ordinary Least Square (OLS) method of estimating parameters in linear regression model, principal components regression was used. Table 4 indicates results of regressing dependent variables on five explanatory variables (PCs).In comparison this method brought some changes on standard error. In the original regression model where OLS method was employed, standard errors of estimate coefficients were large that weakened the statistical power due to presence of severe multicolinearity. This is contrary to the output of regression model after utilizing principal component, where the standard errors were smaller compared to the OLS methods. In assessing whether the collinearity exist between variables, VIF was computed on each of the variables treating as dependent variable and regress on the rest of the variables. The result indicates that multicollinearity problem was eliminated since VIF values for each of the variables were less than ten (10).

6. Discussion

The key objective was to demonstrate how principal components method can be used to eliminate multicollinearity problem that may exist when running linear regression model. The real application of the techniques was presented in the problem of predicting factors influencing student satisfaction where overall student satisfaction was predicted on several fourteen variables. The results of linear regression model revealed a large standard error of coefficients, the situation which resulted into biasness of the mean estimates of the coefficients. The major reason is the violation of ordinarily least square assumption that requires the predictors to be independent. The Variance Inflation Factor was used as indicator to detect collinearity among predictors. It was observed that VIF values of twelve predictors exceed 10 which indicate presence of multicollinearity problem. Thus ignoring this statistical problem can lead to wrong conclusion.

After confirming the presence of high relationship between independent variables, the principal components was utilized to find the possible linear combination of variables that can produce large variance without much loss of information. The first component contained the variables which were highly related namely variables number 5,6,13 and 14.Similarly, the second component contained variable number 1, 2, 3 and 4.Third component contained variable number 11 and 12 while fourth component contained variable number 9 and 10.The last principal component combined variable number 7 and 8.These original fourteen (14) set of variables were transformed into five (5) variables (Principal components) as a linear combination of related variables, but the new variables are independent to each other.

The last step was to assess the efficiency of Principal component methods in solving multicollienarity. In order to examine the presence of relationship between predictors, dependent variables were regressed on these five principal components. The results show that VIF values for each predictor range from 1 to 3 which indicate that multicollinearity problem was eliminated. Principal components method helps not only in identifying which variables are highly related, but also providing solution for improving results of the estimated coefficients. The method transforms a set of linearly related variables into artificial variable that are not related with each other. If these new variables can be named meaningfully they may be treated as variables for further analysis and considered as a remedial solution to mutlicollinearity. Regardless of the strength of principal components in removing multicollinearity, its application is limited to a large sample size specifically a minimum of 300 observations (Comrey and Lee, 1992).

7. Conclusion

The principal objective of this study was solution to multicollinearity when fitting linear regression model. Multicollinearity was detected using Variance Inflation Factor (VIF) and then principal component analysis as solution to the problem was presented. The study indicated that principal component analysis is one of the appropriate methods of solving this matter. Therefore applying principal components produce better estimation and prediction than ordinary least squares when predictors are related.

References

Cohen, P., West, S.G. and Aiken, L.S., 2014. *Applied multiple regression/correlation analysis for the behavioral sciences.* Psychology Press.

Comrey, A.L. and Lee, H.B., 1992. *A First Course in Factor Analysis*, 2nd Ed., Hillsdale, NJ: Lawrence Elbraum

El-Dereny, M. and Rashwan, N.I., 2011. Solving Multicollinearity Problem Using Ridge Regression Models. *International Journal of Contemporary Mathematical Sciences*, 6(12), pp. 585-600.
Farebrother, R. W., 1999. A class of statistical estimators related to principal components. *Linear algebra and its applications*, 289(1-3), pp. 121-126.
Gujarati, D.N., 2009. *Basic econometrics*. Tata McGraw-Hill Education.
Hoffmann, J.P., 2010. *Linear Regression Analysis: Applications and Assumptions*, 2nd Ed., Wade Jacobson KB, Brigham Young University.
Johnson, R.A and Winchern, D. W., 2007. *Applied Multivariate Statistical Analysis,* 6th Ed., New Jersey: Prentice- Hall
Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology,* 24(7), pp. 498-520.
Jolliffe, I. (2002). *Principal component analysis, 2nd Ed.,* New York: Springer- Verlag
Kaiser, H. F., 1960. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1), pp.141-151.
Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W., 2005. *Applied linear statistical models*, 5th Ed., New York: McGraw-Hill
Kmenta, J., & Rafailzadeh, B., 1997. *Elements of econometrics, 2nd Ed*., University of Michigan Press
Montgomery, D. C., Peck, E. A., and Vining, G. G., 2012. *Introduction to linear regression analysis, 5th* ed., Hoboken, New Jersey: John Wiley & Sons.
Mason, G., 1987. Coping with collinearity. *Canadian Journal of Program Evaluation*, 2(1), pp. 87-93.
Mela, C. F., & Kopalle, P. K., 2002. The impact of collinearity on regression analysis: the asymmetric effect of negative and positive correlations. *Applied Economics*, 34(6), pp. 667-677.
Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), pp.559-572.
Stevens, J. P., 2012. Analysis of covariance. *Applied Multivariate Statistics for the Social Sciences, 5th ed.,* Routledge.
Wooldridge, J.M., 2010. *Econometric analysis of cross section and panel data*. Cambridge, Massachusetts: MIT Press.

Bio-note

Abubakari Gwelo is Assistant Lecturer in the Department of Mathematics and Statistics, Faculty of Science and Technology, Mzumbe University in Tanzania. He holds Master degree in Statistics and Bachelor degree in Statistics of the University of Dar es Salaam. His main research interests are Multivariate statistical methods, time series analysis, sampling design, experimental design, Social science researches. He has published widely in various reputable international journals. The published papers are based on theoretical statistics, and its application in different fields such as economics, social science, business, education, and health science.