International Journal of Innovative Mathematics, Statistics & Energy Policies 5(4):19-29, Oct.-Dec., 2017

© SEAHI PUBLICATIONS, 2017 www.seahipaj.org ISSN: 2467-852X

Component Regression Models To Predict The Student's Final Grade Point Average (GPA) in the Polytechnic

Alabi Mudasiru Adebayo[1] & Issa Suleman[2]

[1]Department of Mathematics and Statistics, Akanu Ibiam Federal Polytechnic Unwana, Afikpo, Ebonyi, Nigeria

[2]Department of Statistics, University of Ilorin, P.M.B. 1515, Ilorin, Nigeria
Email: [1]alabi_ma@yahoo.com, [2]isstatisticsman01@gmail.com

ABSTRACT

The purpose of this study is to investigate an empirical determination of the predictors (the courses) variables that has an effect on student cumulative grade point in the Polytechnic. In Addition, the paper also attempts to improve the predictive power of multiple linear regression models using principal components as input for predicting the CGPA of incoming freshers. Performance indicator such as Coefficient of Determination (R^2), Normalized Absolute Error (NAE), Root Mean Square Error (RMSE), and Coefficient of Variation (CV) were used to measure the accuracy of the models. The result of this paper shows that, Principal component regression (PCR) model performs better than multiple linear regression (MLR) based on the performance indicators, because PCR had minimum NAE, RMSE, CV and the coefficient of determination has higher predicted accuracy than MLR. PCR also reducing their complexity and eliminating data co-linearity.

Keyword: Student Courses and Scores, Principal Component Regression, Multiple Linear Regression, and Performance Indicators.

1.0 INTRODUCTION

Multiple linear regression is defined as a multivariate technique for determining the correlation between a response variable *Y* and some combination of two or more predictor variables, *X*, (see, for example, Montgomery and Peck (1982), Draper and Smith (1998), and McClave and Sincich (2006), among others, for details). Multiple linear regression is one of the most widely used statistical techniques in educational research. It is regarded as the "Mother of All Statistical Techniques." For example, many colleges and universities develop regression models for predicting the GPA of incoming freshmen. The predicted GPA can then be used to make admission decisions. In addition, many researchers have studied the use of multiple linear regression in the field of educational research. The use of multiple linear regression has been studied by Shepard (1979) to determine the predictive validity of the California Entry Level Test (ELT). In Draper and Smith (1998), the use of multiple linear regression is illustrated in a prediction study of the candidate's aggregate performance in the G. C. E. examination. The use of multiple regression is also illustrated in a partial credit study of the student's final examination score in a mathematics class at Florida International University conducted by Rosenthal (1994). In Shakil (2001), the use of a multiple linear regression model has been examined in predicting the college GPA of matriculating freshmen based on their college entrance verbal and mathematics test scores.

PCA is mostly used for reducing the multiple dimensions associated to multiple linear regressions which create new variables called the principal component (PCs) that are orthogonal and uncorrelated to each other. The first PC explains the largest fraction of the original data variability and second PC explains larger fraction than third PC and so on (AbdulWahab *et al*., 2005; Wang and Xiao, 2004; Sousa *et al*., 2007). Varimax rotation is mostly used to obtain the rotated factor loadings that represent the contribution of each variable to a specific principal component. Principal component regression (PCR) is a method that combines linear regression and PCA. PCR establishes a relationship between the output variable (y) and the selected PC of the input variables (*x*i). In addition, many previous researchers such as Kothai *et al*., (2008) performed principal component analysis using varimax rotation to identify five major sources contributing to coarse and fine particulate mass. Morandi *et al.* (1991) compared between factor analysis/multiple regression and principal component analysis /regression models, the results indicated that the number and type of the sources resolved by the two approaches were similar. Pires *et al.*(2008) focused on the determination of the parameters that influence the concentration of tropospheric ozone using PCA.

The aim of this study is to compare the predictive power of multiple linear regression and principal component regression models for examine the nature of relationship between Cumulative grade point average (CGPA) and all the courses of the students' score (statistical and non-statistical courses) in the polytechnic.

2.0 METHODOLOGY

2.1 Data Description

The sample of 51 students were selected from record 2013/2014, 2014/2015 and 2015/2016 session of Mathematics and Statistics department, AkanuIbiam federal polytechnic, Unwana Afikpo, Nigeria and holds 33 variables altogether with 32 independent variables (input variables) and a dependent variable (output variable).

2.2 Multiple Linear Regression

Multiple linear regression is one of the modeling techniques to investigate the relationship between a dependent variable and several independent variables. This is a generalisation of the simple linear regression model. In the multiple linear regression model, the error term denoted by ε is assumed to be normally distributed with mean 0 and variance δ^2 (which is a constant). ε is also assumed to be uncorrelated. We assume that the multiple linear regression models have *k* independent variables and there n are observations. Thus the regression model can be written as (Kovac-Andric *et al*., 2009).

The regression equation used was:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \text{ With } i = 1, 2, ..., n$$

Where b_i are the regression coefficients, X_i are independent variables and ε is error associated with the regression. To estimate the value of the parameters, the least squares method was used.

2.3 Principal Component Analysis

Principal component analysis was used to find a small set of linear combinations of the covariates which are uncorrelated with each other. This will avoid the multicollinearity problem. Besides, it can ensure that the linear combinations chosen have maximal variance. Application of principal component analysis (PCA) in regression has long been introduced by Kendall (1957) in his book on Multivariate Analysis. Jeffers (1967), suggested for regression model to achieve an easier and more stable computation, a whole new set of uncorrelated ordered variables that is the principal components (PCs) be introduced (Lam *et. al*., 2010). We then use the factors identified from factor analysis to generate the principal components Z_i as follows.

For the 32 variable $X_1, X_2, \cdots, X_{32}$ measured on the 51 students, the i^{th}principal component, Z_i can be written as a linear combination of the original variables. Thus

$$Z_i = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{i32}X_{32}$$

The principal components are chosen such that the first one, $Z_1 = a_{11}X_1 + a_{12}X_2 + \cdots + a_{132}X_{32}$ accounts for as much of the variation in the in the original variables as possible subject to the constraint that

$$a_{11}^2 + a_{12}^2 + \cdots + a_{132}^2 = 1$$

Then the second principal component $Z_2 = a_{21}X_1 + a_{22}X_2 + \cdots + a_{232}X_{32}$ is chosen such that its variance is as high as possible with similar constraint that

$$a_{21}^2 + a_{22}^2 + \cdots + a_{232}^2 = 1.$$

Another useful constrain is that the second component is chosen such that it is uncorrelated with the first component. The remaining principal components are chosen in the same way.

2.4 Keiser Meyer Olkin's and Bartlett's test of Sampling Adequacy and measuring the Homogeneity of variance across variables for Students' Data.

We wish to determine the hidden factors behind the variables (courses offered by students) in order to determine the natural groupings (factors that are highly correlated with each other and those that are weakly correlated with others) of students' performance. We will use Keiser Meyer Olkins (KMO) measure of sampling adequacy to determine whether the sample is adequate for the analysis. Rejection of the null hypothesis will imply that the sample is not adequate for the analysis. It is also important that we check for sphericity of the data set using Bartlett's test of sphericity. Rejection of the null hypothesis will imply that the data set is good for the analysis.

H_{01}: The sampled data is adequate for the study
H_1: The sampled data is not adequate for the study.
Or
H_{02}: $\delta_1 = \delta_2 = \cdots = \delta_k$
H1: $\delta_i \neq \delta_k$ *for at least one pair* (i, j)
Test Statistics: KMO
Decision Rule: Reject H_0 in favor of H_1 at 0.05 level of significance if p-value $\lesssim$ 0.05 otherwise do not reject H_0

3.0 RESULT AND DISCUSSION

This section deals with Regression analysis and Principal Component Regression techniques, and interpretation. The specific variables discussed in this section include: Cumulative Grade Point Average (CGPA), Statistical courses and Non-Statistical courses.

3.1. Residual Plots for Cumulative Grade Point Average (CGPA)

The study of residuals (or error) is very important in deciding the adequacy of the statistical model. If the error shows any kind of pattern, then it is considered that the model is not taking care of all the systematic information. The residual analysis shows that the residuals are distributed normally with zero mean and constant variance. The plots of fitted values with residuals for Cumulative Grade Point Average (CGPA) model indicated that the residuals are uncorrelated i.e. the residuals are contained in a horizontal band and hence obviously that variance is constant.

The regression computer program outputs for residual plots of cumulative grade point average are given in Figure 1 below. The paragraphs that follow examine the goodness of fit model based on residual plots.

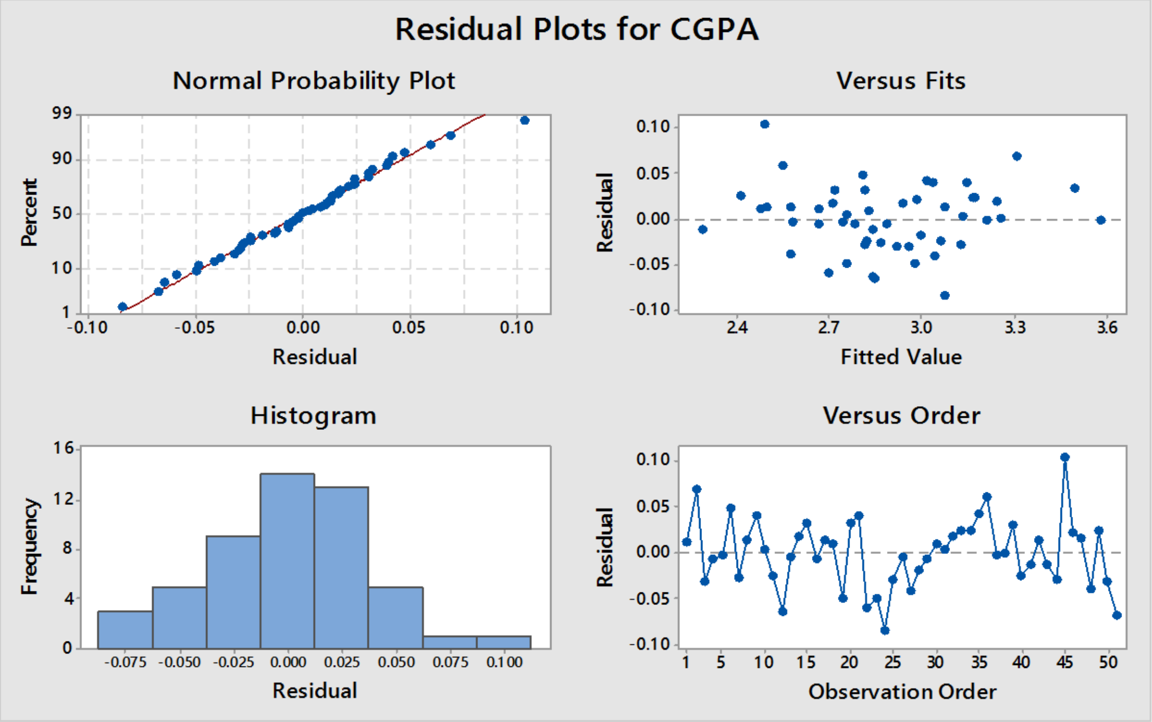

Figure 1 Residual plots for Consumer price Index

Interpreting the Graphs (Figure 1)

(5) From the normal probability plot, we observe that there exists an approximately linear pattern. This indicates the consistency of the data with a normal distribution. The outliers are indicated by the points in the upper-right corner of the plot.

(6) From the plot of residuals versus the fitted values, it is evident that the residuals get smaller, that is, closer to the reference line, as the fitted values increase. This may indicate that the residuals have non-constant variance, (see, Draper and Smith (1998), among others, for details).

(7) The histogram of the residuals indicates that no outliers exist in the data.

(8) The plot for residuals versus order is also provided in Figure 1. It is defined as a plot of all residuals in the order that the data was collected. It is used to find non-random error, especially of time-related

effects. A clustering of residuals with the same sign indicates a positive correlation, whereas a negative correlation is indicated by rapid changes in the signs of consecutive residuals.

3.2 Multiple Linear Regression

The Multiple linear regression models were developed with the highest R^2 (0.988) is obtained. The range of values for Variance Inflation Factor (VIF) for the independent variables is between 1.53 until 3.06. The value is lower than 10 indicating that there is no multicollinearity between the independent variables. Durbin Watson statistic shows that the model does not have any first order autocorrelation problem (DW=1.611). From the Analysis of Variance, we observe that the F-statistic 117.674 with (prob-value is 0.000). This implies that that the model estimated by the regression procedure is significant at a α-level of 0.05. Thus at least one of the regression coefficients is different from zero. (See, Table 3.1)

In Multiple regression modeling techniques such as the one employed in this study, predictions and evaluation of models are mainly based only on the function of the significant predictor variables. Therefore, for us to generate a reduced form of the model that contains only the significant variables at a respectable alpha-value, the backward elimination procedure was applied to arrive at the final CGPA model. In this present paper, variables were retained and/or eliminated at the 0.10 significance level. After seventeen backward elimination processes, sixteen statistically significant courses (Subjects) were retained in the model. The interest rates cover: STA225, STA311, STA312, MTH 314, COM122 etc. The result of the regression is summarized in Table 3.1 below.

The courses of STA225, STA312, STA411, STA412, STA416, STA421, STA424, STA427, MTH314, COM312, COM322 and GNS162 were found to be statistically significant at the 1 percent α-level with p-values of 0.001 each, 0.008, and 0.005 respectively while other courses such as STA311, STA414, STA421, and STA313 were found to be statistically significant at the 5 percent and 10 percent α-level with p-values of 0.048, 0.025, 0.034 and 0.066 respectively.

In terms of statistical courses, twelve courses (STA225, STA311, STA312, STA313, STA411, STA412, STA414, STA416, STA421, STA422, STA424, and STA427) entered into the model were retained. The coefficients associated with the statistical courses are 0.003, 0.002, 0.003, 0.003, -0.004, 0.005, 0.002, 0.005, 0.008, 0.002, 0.005, and 0.006 respectively, signifying that a 1-unit increases in the student score (grade), as result of increases or decreases in CGPA holding other variables constant while Non-statistical courses, four (MTH314, COM312, COM322, and GNS162) entered into the model were retained. The coefficients estimated with non-statistical courses are 0.003, 0.005, 0.004, and 0.003 respectively which signifying that a 1-unit increases in the students score (grade), as result of increases in students CGPA holding other variables constant.

In order to ascertain the fit of the model, the coefficient of Determinant (R-square), Coefficient of variation (C.V), mean square error (MSE), Root mean square error (RMSE) and Ave. Abs pct. Error. A look at the Coefficient of Determinant (R-square), Coefficient of Variation (C.V), mean square error, Root mean square error, and Ave. Abs. pct Error values in Table 3.1 reveals that the model recorded some values of 0.988, 0.0172, 0.0025, 0.0497 and 0.811 respectively (see, Table 3.1).From final step (step 17) using back ward elimination, the regression coefficients for the dependent variable were used to derive the equation for CGPA as given by equation below.

$$CGPA = -0.226 + 0.0.003STA224 + 0.002STA311 + 0.003STA312 + ... + 0.004COM322 + 0.003GNS162$$

Table 3.1 multiple linear regression based on raw data (Estimation Result)

Predictors variables	Coefficients	Std. Error	t-value	p-value	VIF
(Constant)	-0.226	0.114	-1.991	0.055	
STA225	0.003	0.001	2.795	0.008	1.946
STA311	0.002	0.001	2.055	0.048	2.683
STA312	0.003	0.001	3.619	0.001	3.062
STA313	0.003	0.001	1.901	0.066	2.458
STA411	-0.004	0.001	-3.030	0.005	1.890
STA412	0.005	0.001	4.055	0.000	1.925
STA414	0.002	0.001	2.350	0.025	2.322
STA416	0.005	0.001	5.476	0.000	1.762
STA421	0.008	0.001	7.273	0.000	2.036
STA422	0.002	0.001	2.211	0.034	1.680
STA424	0.005	0.001	4.693	0.000	2.378
STA427	0.006	0.001	4.805	0.000	1.581
MTH314	0.003	0.001	3.863	0.000	2.846
COM312	0.005	0.001	4.146	0.000	2.183
COM322	0.004	0.001	4.158	0.000	2.019
GNS162	0.003	0.001	4.223	0.000	1.534

R-squared	0.988	Mean square Error	0.0025
Adjusted R-squared	0.967	Root Mean Square Error	0.0497
F-statistic	117.674	Coefficient of Variation	0.0172(1.72%)
Prob(F-statistic)	0.000	Normalization Abs. Error	0.811

STA225 represent small scale biz, STA311 represent statistical theory, STA312 represent Applied Gen. statistics, STA313 represent Statistical inference, STA314 represent Operation research I, STA315 represent Technical English, STA321 is Statistical theory II, STA322 represent Sampling techniques, STA323 represent Design &Analysis Expt., STA324 represent Statistical mgt., STA325 represent Biometrics, STA411 represent Operation research II, STA418 represent Small biz mgt. STA422 represent Demography II, COM122 represent Computer Operator, COM312 represent Data base Design, MTH314 represent Mathematical method, MTH322 represent Mathematical method II,

3.3 Principal Component Regression

We wish to determine the hidden factors behind the variables (courses offered by students) in order to determine the natural groupings (factors that are highly correlated with each other and those that are weakly correlated with others) of students' performance. The correlations between the independent variables are in the range of -200 to 0.760. Another important test for PCA is the Kaiser-Meyer-Olkin (KMO) of sampling adequacy and Bartlett's test of sphericity. Kaiser (1974) recommends accepting values greater than 0.5 that means the result for this research is acceptant with the value of KMO is 0.671. Barlett's test is highly significant ($p < 0.001$) and therefore factor analysis is appropriate for this data.

Table 3.2: KMO Statistics for Sampling Adequate and Bartlett's test for Homogeneity

Test	DF	Approx. Chi-Square	P-value
Keiser-Meyer-Olkin Measure of Sampling Adequate	-	-	.671
Bartlett's Test of Sphericity	696	1071.634	0.000

Table 3.3: Total Variance Explained

Component	Initial Eigenvalue			Extraction sums of Squared loadings			Rotation sums of Squared loadings		
	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %	Total	% of variance	Cumulative %
1	10.521	32.879	32.879	10.521	32.879	32.879	4.167	13.021	13.021
2	3.256	10.175	43.054	3.256	10.175	43.054	4.004	12.513	25.534
3	2.346	7.332	50.386	2.346	7.332	50.386	3.719	11.622	37.157
4	1.668	5.212	55.598	1.668	5.212	55.598	2.465	7.703	44.860
5	1.606	5.019	60.617	1.606	5.019	60.617	2.414	7.543	52.403
6	1.330	4.158	64.775	1.330	4.158	64.775	2.255	7.046	59.449
7	1.265	3.953	68.727	1.265	3.953	68.727	1.970	6.157	65.607
8	1.173	3.667	72.395	1.173	3.667	72.395	1.694	5.295	70.901
9	1.023	3.197	75.591	1.023	3.197	75.591	1.501	4.690	75.591
10	.886	2.769	78.361						
11	.860	2.686	81.047						
12	.775	2.422	83.469						
13	.658	2.057	85.525						
14	.577	1.802	87.327						
15	.540	1.686	89.013						
16	.506	1.580	90.593						
17	.452	1.413	92.006						
18	.418	1.305	93.311						
19	.331	1.036	94.347						
20	.286	.892	95.240						
21	.272	.851	96.090						
22	.240	.751	96.841						
23	.185	.578	97.419						
24	.155	.485	97.904						
25	.140	.438	98.342						
26	.127	.397	98.739						
27	.110	.344	99.083						
28	.090	.282	99.365						
29	.072	.224	99.590						
30	.057	.179	99.768						
31	.048	.149	99.918						
32	.026	.082	100.00						

Table 3.3 lists the eigenvalues associated with each linear component (factor) before extraction, after extraction and after rotation. Before extraction, SPSS has identified thirty-two (32) linear components within the data set. The eigenvalues associated with each factor represent the variance explained by the particular linear component and also displays their eigenvalue in term of the percentage of variance explained (so, Factor 1 explains 32.879% of total variance). PCA extracts all Factors with eigenvalues greater than 1; the cumulative variance explained by nine principal components is 75.591%. The eigenvalues associated with these factors are again displayed in the label extraction sums of squared loading. In the final part of the Table 3.3, the eigenvalues of the factors after rotation are displayed. Rotation has the effect of optimizing the factor structure and one consequence for these data is that the relative importance of the nine factors is equalized.

Before rotation, Factor 1 accounted for considerably more variance than the remaining other factors (32.879% compare to 10.175%, 7.332%, 5.212%, 5.019%, 4.158%, 3.953%, 3.667%, and 3.197%), however after extraction it accounts for only 13.021% compared to 12.513%, 11.622%, 7.703%, 7.543%, 7.046%, 6.157%, 5.295%, and 4.690%.

Table 3.4: Rotated Component Matrix

	Components								
	1	2	3	4	5	6	7	8	9
STA225							0.796		
STA311	0.515	0.409		0.436					
STA312	0.790								
STA313			0.421	0.488					
STA314		0.692		0.476					
STA315					0.687				
STA321	0.711								
STA322				0.674					
STA323			0.405					0.638	
STA324									0.866
STA325	0.579								
STA411				0.733					
STA412	0.445	0.433							
STA413			0.783						
STA414			0.681						
STA415						0.693			
STA416			0.701						
STA417		0.588							
STA418							0.613		
STA421		0.443							
STA422						0.527	0.412		
STA423	0.442		.644						
STA424	0.435	0.431							0.403
STA425		0.764							
STA426						0.803			
STA427								-.831	
MTH314	0.670								
MTH322		0.743							
COM122		0.687							
COM312					0.854				
COM322			0.531		0.537				
GNS162	0.520								

Rotated matrix rotation using varimax rotation with Kaiser Normalization is shown in Table 3.4. This matrix contains the loading of each variable onto each factor where values less than 0.4 are suppressed from the output. . Based on these factor loadings, the factors represent

- STA311, STA312, STA321, STA325, STA412, STA423, STA424, MTH314, and GNS162 are loaded strongly on Factor 1.
- STA311, STA314, STA412, STA417, STA421, STA424, STA425, MTH322, and COM122 are loaded strongly on Factor 2.
- STA313, STA323, STA413, STA414, STA416, STA423, and COM322 are loaded strongly on Factor 3
- STA311, STA313, STA314, STA322, and STA411 are loaded strongly on Factor 4.

- STA315, COM312, and COM322 are loaded strongly on Factor 5.
- STA415, STA422, and STA426 are loaded strongly on Factor 6.
- STA225, STA418, and STA422 are loaded strongly on Factor 7.
- STA323 and STA427 are loaded strongly on Factor 8.
- STA324 and STA424 are loaded strongly on Factor 9.

Table 3.5: The Coefficient of Principal Component Score of Variables

Variables	PC1	PC 2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
STA 225	-0.055	-0.082	-0.021	0.047	0.005	-0.055	0.449	0.039	0.180
STA 311	0.090	0.088	-0.024	0.142	-0.045	-0.039	-0.021	-0.032	-0.162
STA 312	0.294	-0.120	-0.078	-0.018	0.094	0.009	-0.049	0.043	-0.008
STA 313	-0.003	-0.159	0.064	0.191	0.076	0.004	0.088	0.230	0.055
STA 314	-0.126	0.264	0.002	0.206	-0.021	-0.028	-0.193	-0.169	-0.023
STA 315	-0.055	0.038	-0.068	0.072	0.296	-0.036	0.010	-0.098	-0.037
STA 321	0.203	0.001	-0.031	0.048	0.012	-0.065	-0.060	-0.105	0.043
STA 322	-0.044	0.030	-0.002	0.314	-0.109	0.017	-0.084	-0.076	0.010
STA 323	0.052	-0.018	0.191	-0.097	0.060	-0.182	-0.056	0.387	0.072
STA 324	0.000	-0.074	-0.073	0.024	0.008	-0.022	0.058	-0.003	0.630
STA 325	0.265	-0.138	-0.018	-0.110	0.191	-0.004	0.122	0.009	-0.306
STA 411	0.110	-0.126	-0.043	0.350	0.140	-0.178	0.007	-0.005	0.035
STA 412	0.111	0.069	-0.197	-0.041	-0.091	0.176	-0.001	0.150	0.067
STA 413	0.012	-0.059	0.278	-0.007	-0.058	-0.100	-0.008	0.031	0.030
STA 414	-0.070	-0.047	0.224	0.069	-0.146	0.005	0.153	-0.019	-0.151
STA 415	-0.037	-0.088	-0.014	-0.051	0.179	0.440	-0.176	0.041	-0.088
STA 416	-0.141	0.077	0.301	0.007	0.005	-0.043	-0.072	0.008	-0.131
STA 417	-0.047	0.144	0.033	-0.043	0.000	0.044	0.038	0.131	-0.097
STA 418	0.060	0.115	0.055	-0.246	-0.070	-0.076	0.323	-0.073	-0.116
STA 421	-0.278	0.126	-0.067	0.248	0.003	0.154	0.170	0.014	-0.044
STA 422	-0.059	-0.045	-0.111	0.084	-0.085	0.332	0.240	0.044	-0.093
STA 423	0.050	0.010	0.188	-0.079	-0.045	-0.074	-0.003	-0.048	0.117
STA424	0.030	0.113	0.015	-0.162	-0.059	0.128	-0.084	-0.167	0.243
STA 425	-0.051	0.244	0.067	-0.055	0.034	-0.085	-0.056	0.075	-0.118
STA 426	-0.098	-0.028	-0.085	-0.037	-0.086	0.535	0.028	-0.063	0.043
STA 427	0.025	0.011	0.064	0.004	0.100	-0.058	-0.058	-0.515	0.063
MTH 314	0.211	-0.003	0.040	-0.015	-0.041	-0.132	-0.072	0.113	0.040
MTH 322	-0.026	0.258	0.088	-0.061	-0.106	-0.109	-0.054	-0.034	0.067
COM 122	0.070	0.207	-0.171	-0.090	0.013	-0.049	0.164	-0.066	0.022
COM 312	0.142	-0.126	-0.066	-0.022	0.432	-0.034	-0.045	0.021	0.027
COM 322	-0.156	0.026	0.226	-0.100	0.229	0.092	-0.187	-0.113	0.066
GST 162	0.164	0.065	-0.155	0.050	-0.088	-0.057	0.163	-0.056	0.010

Table 3.5: the first nine principal component's scores are computed from the original data using the coefficients listed under PC1, PC2 and PC9 respectively:

$$PC1 = -0.055STA225 + 0.090STA\ 311 + 0.294\ STA\ 312 += \cdots = +0.164GST\ 162$$

$$PC2 = -0.082\ STA\ 225 + 0.088STA\ 311 - 0.120\ STA\ 312 += \cdots = +0.065\ GST\ 162$$

. . . . .

. . . . .

. . . . .

$$PC9 = 0.180STA\ 225 - 0.162\ STA\ 311 - 0.008STA\ 312 += \cdots = +0.010\ GST\ 162$$

Table 3.6: Multiple linear Regression based on Principal Component scores Coefficient

Predictor	Coefficient	Std. Error	t	p-value	VIF
Constant	2.889	0.007	388.731	0.000	1.000
PC1	0.036	0.002	15.619	0.000	1.000
PC2	0.070	0.004	16.865	0.000	1.000
PC3	0.083	0.005	16.888	0.000	1.000
PC4	0.052	0.006	8.987	0.000	1.000
PC5	0.060	0.006	10.060	0.000	1.000
PC6	0.084	0.007	12.920	0.000	1.000
PC7	0.065	0.007	9.805	0.000	1.000
PC8	-0.006	0.007	-0.874	0.387	1.000
PC9	0.043	0.007	5.777	0.000	1.000
R-squares(R^2)	0.990	Mean Square Error	0.0024		
Adj. R-Square(R^2)	0.989	Root Mean Square Error	0.0485		
F-statistic	162.411	Coefficient of variation	0.0170(1.70%)		
Prob(F-statistic)	0.000	Normalization Abs..Error	0.715		
Dubin-Watson(DW)	1.772				

Multiple linear regression analysis was repeated by using principal component analysis as inputs. Here coefficient of determination (R^2) is 0.990. The Value for variance Inflation Factor (VIF) for the independent variables is 1 indicating no multicollinearity problem. Durbin Watson statistic shows that the model does not have any first order autocorrelation problem (DW=1.772). The residual analysis shows that the residuals are distributed normally with zero mean and constant variance. Nine main factors from PCA were used as independent variables and the following model was obtained. The principal component scores of selected PCs (PC1-PC9) are used as predictor variables for MLR analysis. The results revealed that multicollinearity was removed and PC1, PC2, PC3, PC4, PC5, PC6, PC7, and PC9 were found to be statistically significant while PC8 is not significant, as shown in Table 3.6.

The final model can be written as:-

$$CGPA = 2.889 + 0.036 * PC1 + 0.070 * PC2 + 0.083 * PC3 + 0.052 * PC4 + ... + 0.043 * PC9$$

The coefficients of first nine principal components and constant are $\beta_1 = 0.036$, $\beta_2 = 0.070$, $\beta_3 = 0.083$, $\beta_4 = 0.052$, $\beta_5 = 0.060$, $\beta_6 = 0.084$, $\beta_7 = 0.065$, $\beta_8 = -0.006$, $\beta_9 = 0.043$ and $\beta_0 = 2.889$ respectively. Since the p-values (i.e. 0.000) for all the coefficients' are less than 0.05 except the coefficient of principal component 8 (PC8) which p-value (i.e. 0.387) is greater than 0.05 therefore we reject null hypothesis, Ho at all other coefficient except PC8 and conclude that PC1, PC2, PC3, PC4, PC5, PC6, PC7, and PC9 have significant impact on CGPA on final students'.

3.4 Comparison of performance Between PCR and MLR

Performance indicators were used to compare between MLR and PCR for students' final cumulative grade point average. Table 3.7 shows the performance indicator values. The value of the accuracy measure is Coefficient of Determination. The accuracy measure for PCR is higher than for MLR. The values of the error measures namely Normalized Absolute Error, Root Mean Square Error and Coefficient of variation are smaller for PCR than for MLR. This shows PCR gives better result than MLR based on accuracy measures and error measures. So, PCR should provide a better prediction than MLR.

Table 3.7: Performance Indicator between MLR and PCR models

Performance indicators	MLR	PCR
Coefficient of Determinant (R^2)	0.988	**0.990**
Normalized Absolute Error	0.811	**0.715**
Root Mean Square Error	0.0497	**0.0485**
Coefficient of variation	0.0183	**0.0170**

4.0 CONCLUSIONS

In this paper, multiple linear regression was used to predict students' final cumulative grade point average using as predictors (all the courses of the students score) variables. Two different approaches were used, considering original data and principal component as inputs. The result showed that the used of principal component as input provides a more accurate result than original data because it reduction of the number of predictor variables, decrease of the model complexity and better interpretation of MLR models by removing indirect effects related to predictor.

The quality and reliability of the developed models were evaluated through performance indicators (Coefficient of determinant (R^2), Normalized Absolute Error (NAE), Root Mean Square Error (RMSE), and Coefficient of Variation (CV)). Assessment of model performance indicated that principal component regression can predict particulate matter better than multiple regressions. Similar conclusions were found by previous studies (Ul-Saufie *et. al.*, 2011; Sousa *et al*., 2007; Ozbay *et al.*, 2011). However models adequacy checked by various statistical methods showed that the developed multiple regression models can also be used for prediction of cumulative grade point average.

REFRENCES

AbdulWahab S.A., Bakheit C.S. and AlAlawi S.M., (2005). *Principal component and multiple regression analysis in modelling of groundlevel ozone and factors affecting its concentrations*,
Environmental Modelling& Software, **Vol.** 20, No. 10, p. 1263–1271.

Draper, N. R., and Harry S. (1998). *Applied Regression Analysis (3rd edition).* New York: John Wiley & Sons, INC.

Jeffers J. N.R (1967).*Two case studies in the applicationof principal component analysis*. Applied Statistics, No. 16, p. 225-236.

Kaiser, H.F., (1974). Index of factorial simplicity. Psychometrika, 39, pp 31–36

Kendall M.G (1957). A course in multivariate Analysis, London, Griffin

Kovac-Andric, E., Brana, J., Gvozdic V., (2009).*Impact of meteorological dactors on ozone concentrations modeled by btime series and multivariate statistical methods.Ecological*.Infomatics, No. 4, p. 117-122

Kothai P.,Saradhi, Prathibha1 P., Philip K. Hopke, Pandit G. G., Puranik V.D. (2008). Source Apportionment of Coarse and Fine Particulate Matter at Navi Mumbai, India . Aerosol and Air Quality Research, **Vol.** 8, No. 4, p. 423-436.

Lam, K.C., Tau, T., M.C.K., (2010). *A Material supplier selection model for property developers using fuzzy principal component analysis*. Automation in Construction, No. 19, p. 608-618

Maria T. Morandi, Paul J. Lioy, Joan M. Daisey, (1991). *Comparison of two multivariate modeling approaches for the source apportionment of inhalable particulate matter in Newark*, NJ Original Research Article Atmospheric Environment. Part A. General Topics, **Vol.**25, No. 56, p. 927-937

McClave, J. T., and Sincich, T. (2006). *Statistics (10th edition).* Upper Saddle River, NJ: Pearson Prentice Hall.

Montgomery, D. C., and Peck, E. A. (1982). *Introduction to Linear Regression Analysis*. New York: John Wiley & Sons, INC.

NasriHarb and Ahmed El-Shaarawi, (2006).*Factors Affecting Students' Performance*. MPRA Paper No. 13621, posted 26. February 2009 04:55 UTC. Online at http://mpra.ub.uniM muenchen.de/13621/

Ozbay, Bilge, Keskin, Gulsen Aydin, Dogruparnak, Senay Cetin, Ayberk, Savas (2011*). Multivariate methods for ground-level ozone modeling*, Atmospheric research.

Piresa J.C.M., Sousaa S.I.V., Pereiraa M.C., AlvimFerraza M.C.M. and Martins F.G. (2008). *Management of air quality monitoring using principal component and cluster analysis—Part I: SO2 and PM10 Atmospheric Environment*, **Vol.**42, No. 6, February 2008, p. 1249-1260

Rosenthal, M. (1994). *Partial Credit Study*. University Park, Florida: Department of Mathematics, Florida International University.

Shakil, M. (2001). *Fitting of a linear model to predict the college GPA of matriculating freshmen* *based on their college entrance verbal and mathematics test scores, A Data Analysis I Computer Project.* University Park, Florida: Department of Statistics, Florida International University.

Shepard, L. (1979). *Construct and Predictive Validity of the California Entry Level Test.* Educational and Psychological Measurement, 39: 867 – 77.

Sousa, S.I.V., Martins, F.G., AlvimFerraz M.C.M. and Pereira M.C. (2007). *Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations,* Environmental Modelling& Software, No. 22, p. 97–103.

S. Wang and F. Xiao, (2004). *AHU sensor fault diagnosis using principal component analysis method,* Energy and Buildings, **Vol.**36, No. 2, p. 147–160.

Ul-Saufie A.Z, Yahya A.S, Ramli N.A (2011). *Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang,* International Journal of Environmental Sciences Vol. 2 No.2. pp 415-422