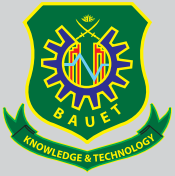
BAUET JOURNAL
Published by
Bangladesh Army University of Engineering & Technology
Journal Homepage: http://journal.bauet.ac.bd



The Utilization of Statistical Machine Learning Techniques in the Prediction of Economic Growth in Ghana

Mohammed Awal Osman[1], and A.H.M. Rahmatullah Imon[2*]

[1]Department of Mathematical Sciences, Research Student, Ball State University, Muncie, IN 47306, USA
[2]Department of Mathematics, Ball State University, Muncie, IN 47306, USA

Abstract: Economic growth has been an area of a great deal of attention to any government and policymakers of a country. The results of many empirical studies in Ghana have proven the causal linkage between public debt and economic growth using traditional statistical methods such as regression and times series. The linear regression model has the most popular choice for designing this type of relationship. Linear models suffer some set back such as the absence of normality and other standard assumptions and some constrained where economic complexity is concerned. Machine learning approaches have gained more popularity now to model this type of data. We may consider a linear regression model itself as a supervised machine learning approach. The main objective of this study is to find a suitable learning algorithm to model the causal linkage between economic growth and its associated economic variables. We have considered three different types of models, the supervised multiple linear regression model, the semi-supervised diagnostic-robust regression model and unsupervised principal component regression model. The effectiveness of the models is evaluated by several diagnostic and goodness of fit tests and cross validation. Numerical results obtained from both model fitting and cross validation show that the diagnostic-robust regression performs best followed by the principal component regression and multiple linear regression.

Keywords: *Supervised learning; Unsupervised learning; Semi-supervised learning; Robust regression; Cross validation.*

Introduction: Over the last century, improvement in the well-being of every society has been the responsibility of any government in the world through investment into infrastructures, education, health, and other sectors of the economy for growth acceleration. Spencer and Orley [1] described growth as an expansion of a country's potential GDP or national output. According to [2] growth is concerned with the expansion of an economy's ability to produce (potential GDP) over time. Government plays a vital role in the financing of social goods and other development programs that would improve the livelihood of its citizen. However, issues in developing countries, unable to provide such developments are a result of shortfalls of revenue and therefore unable to mobilize sufficient public revenue from taxation to meet their expenditure. Therefore, high public expenditures against low revenue push the deficit beyond targets that in turn, policymakers seek to find alternative ways to acquire revenues to fulfill these deficits. Theoretically, financing development projects through debt can help a country to build its production capacity and facilities economic growth. The stated policy of debt in Ghana is to carry out practical macroeconomic programs to improve the socio-economic conditions and wellbeing of its people through sound economic policies such as government spending in the areas of health, education, infrastructure, and other social. This allows the government to borrow in both domestic and external capital markets also through the World Bank, International Monetary Fund, and floating of bonds respectively to support its social services and developmental projects.

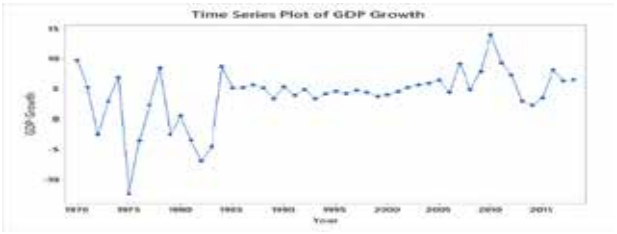
The results of many empirical studies in Ghana and some heavily indebted poor countries (HIPC) nations (see [3-8]) have proven the causal linkage between debt and economic growth using conventional econometric and statistical techniques such as times series and regression. The other variables that emerge as important in explaining growth are Gross Capital Formation (% of GDP), Total Debt Service (% of GNI), Foreign Direct Investment, Trade Openness etc. However, what is lacking in all studies so far is linear regression model is a supervised learning technique and it suffers some set back in the absence of conventional regression assumptions such as normality, homoscedasticity, linear independence among the explanatory variables, absence of outliers and influential

Article history:
Received 01 July, 2021
Received in revised form 2 October, 2021
Accepted 15 November, 2021
Available online 15 November, 2021

Corresponding author details: Rahmatullah Imon
E-mail address: rimon@bsu.edu
Tel: +1 765 285 8650

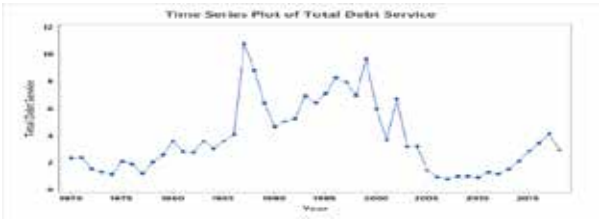
observations etc. Regression diagnostic techniques for modeling GDP for Bangladesh have been studied by [9]. We employ an unsupervised machine learning technique, called the principal component regression, next. This method is designed for a high dimensional data to combat the multicollinearity issue. Finally, we employ a semi-supervised machine learning technique, called the diagnostic-robust method. This method was proposed by [10] extending the idea of [11].

Data and Methodology: In this study, we have used a secondary data obtained from the World Bank's World Development Indicator for Ghana covering the period from 1970 to 2019.



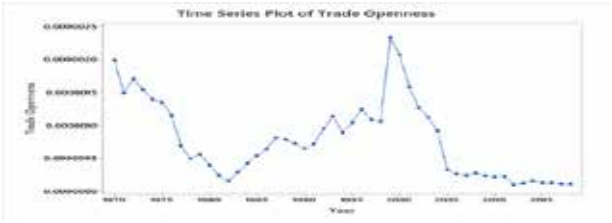
GDP Growth (Annual %)

Gross Capital Formation (% of GDP)

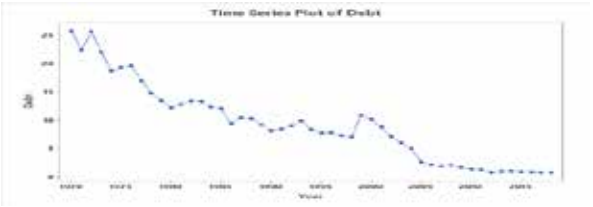


Total Debt Service (% of GNI)

Foreign Direct Investment



Trade Openness



Debt (% of GDP)

Fig.1: Time Series Plot of Different Economic Variables in Ghana.

The dataset contains fifty observations each with six features which are GDP Growth (Annual %), Gross Capital Formation (% of GDP), Total Debt Service (% of GNI), Foreign Direct Investment, Trade Openness, and Debt (% of GDP). Time series plots of different economic variables in Ghana are presented in Figure 1. These plots show that there were some irregular pattern in the GDP growth between 1970 and 1985. Since then there is a slightly increasing trend in the GDP growth over the years. The gross capital formation shows a slightly increasing pattern over the years. The total debt service shows an increasing pattern between 1970 and 2000. Since then it shows a decreasing pattern. Foreign direct investment was steady between 1970 and 2005 and since then it shows a decreasing pattern. Trade Openness decreases between 1970 and 1982, increases between 1983 and 2000 and dicreases since then. The good thing about Ghana's economy is its debt tends to decrease over the entire period of time.

Methodologies: In our study we have used several methods for explaining the development of economic growth in Ghana.

Supervised Learning Multiple Regression Method: Linear regression is a supervised learning method that involves the dependence of one variable (response) on one or more variables (features). It is a traditional method for investigating and modeling the relationship between variables. Let us consider a linear regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i,\ i = 1, 2, \ldots, n \qquad \text{eq. 1}$$

where Y is the dependent variable, representing GDP growth rate, the X's are the independent variables, $\boldsymbol{\beta}$'s are the effects of the independent variables and $\boldsymbol{\varepsilon}$ is the error term. We can express the multiple regression model given in eq. 1 in matrix notation as

$$Y = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{eq. 2}$$

The ordinary least squares estimate of the unknown parameter are obtained by minimizing the sum of squares errors. We can adopt several measures to check or test the goodness of fit of models. Among them R-square, adjusted R-square, and ANOVA F are very simple and commonly used.

The multiple regression used the method of ordinary least square for estimating model parameters under certain assumptions, constituting the estimators as best to make inferences and predicting. The violation of these assumptions may cause non-normality, heteroscedasticity, multicollinearity, and the existence of influential observations and/or outliers. A comparatively simple graphical display for checking the normality assumption in regression analysis is to prepare a normal probability plot of residuals, which makes use of normal probability paper. This method is based on the fact that if each residual is plotted against its expected value under normality, the resulting points should lie approximately on a straight line. In recent years the Jarque-Bera [12] test for normality is a large sample test combining the coefficient of skewness (S) and kurtosis (K) measures of the least square residuals in one test statistic has become very popular and uses the following test statistic:

$$\text{JB} = (n / 6)\ [S^2 + (K - 3)^2/4] \qquad \text{eq. 3}$$

A slight modification of the JB test was done by [13]. His proposed statistic based on rescaled moments (RM) of ordinary least squares residuals is defined as

$$\text{RM} = (nc^3/ 6)\ [S^2 + c(K - 3)^2/4] \qquad \text{eq. 4}$$

where $c = n/(n - p)$, $p = k + 1$ is the number of independent variables in a regression model including the constant. Both the JB and the RM statistic follow a chi square distribution with 2 degrees of freedom.

For heteroscedasticity a plot of the squared residuals against the fitted values of the response variable could be useful. If it shows a systematic pattern or a funnel shape, it indicates the unequal variance of the error term. Several analytical tests are available in the literature for the same. The White [14] test would be employed in the study because it is very easy to understand and has a wide range of applications. Here we use the squared regression residuals to run the regression on X from which we calculate R^2. The White test is based on the fact that under homoscedasticity,

$$n\boldsymbol{R^2} \sim \boldsymbol{\chi^2}(p) \qquad \text{eq. 5}$$

follows the chi-square with p df, p is the number of explanatory variables in the model including the constant.

Another important violation of standard assumptions is the collinearity among the explanatory variables. This problem is commonly known as a multicollinearity problem. Inspection of the correlation coefficients among the explanatory variables could be a very simple and easy way of detection of multicollinearity. In a standard situations these correlations should be insignificant. Significant correlations indicate the presence of multicollinearity. Another popular technique is the variance inflation factor (VIF). It measures the magnitude of the increase of variance caused by multicollinearity. The VIF quantity bigger than 10 declares the presence of severe multicollinearity.

In Statistics, we often observe that the values of descriptive measures are often much influenced by few extreme observations which are commonly known as outliers. Different aspects of outliers with their consequences are discussed by [15] and [16]. In a regression problem, observations are judged as outliers based on how unsuccessful the fitted regression equation is in accommodating them and that is why observations corresponding to excessively large residuals are treated as outliers. Usually the Studentized residuals are employed for the detection of outliers.

We call an observation an outlier when its corresponding standardized residual value exceeds 3 in absolute value. To understand the effect of an unusual observation to the fit of the model we often use some influence measures. Among them the Cook's distance and difference in fits (DFFITS) have become very popular with the statisticians [17]. In our study we would like to consider DFFITS. Belsley *et al*. [18] recommended considering observations as influential if $|\text{DFFITS}i| \geq 3\sqrt{p/n}$.

Unsupervised Learning Principal Component Analysis Method: Unsupervised learning is a process for gaining insights by summarizing data in innovative ways. Unlike in the supervised learning methods, in unsupervised learning, there is no simple data analysis goal such as prediction or classification of a target of interest. Unsupervised learning is carried out as a part of exploratory data analysis for pattern discovery. The goal is usually to discover subgroups among the *p*-variables or *n* observations. A principal component analysis is a well-established technique for dimensionality reduction and multivariate analysis. PCA summarizes the variation in correlated multivariate attributes to a set of uncorrelated components. The extracted uncorrelated components are called principal components using either eigen decomposition or singular value decomposition to estimate the eigenvectors of the covariance matrix of the original variables. The eigenvector and its elements referred to as loadings define a direction in feature space along which the data vary the most. The objective of PCA is to achieve parsimony and reduce dimensionality by extracting the smallest number of components that account for most of the variation in the original multivariate data and to summarize the data with little loss of information. Montgomery *et al*. [19] suggested using this method in regression to combat the multicollinearity problem. In a regression set up we can write

$$Y = Z\alpha + \varepsilon \qquad \text{eq. 6}$$

where $Z = XT$, $\alpha = T'\beta$, $T'X'XT = Z'Z = \Lambda$, and $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_p)$ is a $p \times p$ diagonal matrix of the eigenvalues of $X'X$ and T is a $p \times p$ orthogonal matrix whose columns are the eigenvectors associated with $\lambda_1, \lambda_2, ..., \lambda_p$. The columns of *Z*, which define a new set of orthogonal regressors, are referred to as principal components. For the transformed model, we obtain the OLS estimators:

$$\hat{\alpha} = (Z'Z)^{-1}Z'y = \Lambda^{-1}Z'y \qquad \text{eq. 7}$$

$$Var(\hat{\alpha}) = \sigma^2(Z'Z)^{-1} = \sigma^2\Lambda^{-1} \qquad \text{eq. 8}$$

The principal components regression approach combats multicollinarity by using less than full set of principal components in the model. To obtain the principal components estimators, assume that the regressors are arranged in order of decreasing eigenvalues,

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p > 0$$

Suppose that the last *s* of these eigenvalues approximately equal to zero. In principal components regression the principal components corresponding to near-zero eigenvalues are removed from the analysis and least squares applied to the remaining components. That is:

$$\hat{\alpha}_{pc} = B\hat{\alpha} \qquad \text{eq. 9}$$

where $b_1 = b_2 = ... = b_{p-s} = 1$ and $b_{p-s+1} = b_{p-s+2} = ... = b_p = 0$. Thus the principal components estimator is:

$$\hat{\alpha}_{pc} = \begin{bmatrix} \hat{\alpha}_{p-s} \\ --- \\ 0_s \end{bmatrix}$$

Or in terms of the standardized regressors:

$$\hat{\beta}_{pc} = T\hat{\alpha}_{pc} \qquad \text{eq. 10}$$

Mason and Gunst [20] showed by simulation study that principal components regression offers considerable improvement over least squares when the data are ill-conditioned.

Semi-Supervised Learning Diagnostic-Robust Method: This method is first proposed by [10] and further studied by [13, 21, 22] and many others. The identification of outliers and influential observations by Studentized residual and Cook's distance or DFFITS are designed for a single case, but in reality the unusual observations may occur in groups and these methods become very ineffective in that case. Chatterjee and Hadi [23] pointed out that unusual observations can also generate multicollinearity problem. In the standard literature the detection of outliers and the detection of multicollinearity are dealt separately. A holistic approach is suggested by [23] to deal these two issues together. Later [24] proposed the block adaptive computationally efficient outlier nominators (BACON) to fit a model after taking care of these two problems. Following their work, [21] proposed generalized Studentized residuals

$$r_{si} = \begin{cases} \dfrac{y_i - \hat{y}_i^{(-D)}}{\sqrt{\sigma_i^{(-D)}\left(1 - h_i^{(-D)}\right)}} & i \in R \\ \dfrac{y_i - \hat{y}_i^{(-D)}}{\sqrt{\sigma_i^{(-D)}\left(1 + h_i^{(-D)}\right)}} & i \in D \end{cases} \qquad \text{eq. 11}$$

when D is the outlier set and R is the clean set based on BACON. Since BACON is a classification technique, the diagnostic-robust regression can be considered as a semi-supervised learning method. Observations are called outliers if their corresponding generalized Studentized values exceed 3 in absolute value. *Generalized difference in fits* (GDFFITS) are proposed by [21] as

$$GDFFITS_i = \begin{cases} \dfrac{\hat{y}_i^{(-D)} - \hat{y}_i^{(-D-i)}}{\sqrt{\sigma_i^{(-D-i)} h_{ii}^{(-D)}}} & for \quad i \in R \\ \dfrac{\hat{y}_i^{(-D+i)} - \hat{y}_i^{(-D)}}{\sqrt{\sigma_i^{(-D)} h_{ii}^{(-D+i)}}} & for \quad i \in D \end{cases} \qquad \text{eq. 12}$$

Observations satisfying the cut-off point $\left|GDFFITS_i\right| \geq 3\sqrt{k/(n-d)}$ are declared as influential observations where d is the number of unusual cases.

Cross Validation: One of the objectives of the study is to find an optimal algorithm in assessing the linkage and the impact of a nation's debt on economic growth. Cross-validation is a common technique for evaluating the performance of predicting model. Is a technique for assessing how the results of a statistical analysis will generalize to a new independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in a new and unknown dataset. However, a new dataset is often not available in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). There are several cross-validation techniques for evaluating the performance of the fitted model, the study used a validation set approach. A validation set approach estimates the test error by randomly holding out a subset of the original data for the fitting process, as a training set, and the fitted model is used to predict the response for the observations in the validation set to obtain the resulting validation set error. We may determine the accuracy of the prediction model by computing several popular accuracy measures (see [25]) such as Mean Absolute Error (MAE), Root Mean Squares Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Analysis of Results: This section contains results of our study. We have considered the Annual GDP Growth as an indicator of economic progress and consider the other variables as the determining factors of GDP growth. Thus the variables under study are classified as follows. Here the response variable is GDP Growth (Annual %), and the explanatory variables are Year, Gross Capital Formation (% of GDP), Total Debt Service (% of GNI), Foreign Direct Investment, Trade Openness, and Debt (% of GDP).

Multiple Regression: At first we fit a linear regression model based on the method of least squares. It is worth mentioning that the value of R^2 for this fit is 32.35% with the adjusted R^2 equals 22.69% indicating a poor fit. The summary results of this fit are presented in Table 1. The table of the coefficients shows that only one (Trade Openness) out of 6 explanatory variables is significant at a 5% level. We know from the literature review that the variable Debt should be one of the most significant predictor for GDP growth but for this data this variable is significant at the 10% level, but not at the 5% level.

Table 1: Summary Results of Multiple Regression Model.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	398	504	0.79	0.434	
Year	-0.194	0.252	-0.77	0.445	40.65
Gross Capital Formation	-0.195	0.122	-1.60	0.117	1.87
Total Debt Service	-0.138	0.316	-0.44	0.665	1.99
Foreign Direct Investment	-0.0000003	0.0000002	-1.23	0.226	6.87
Trade Openness	5387766	2431356	2.22	0.032	6.00
Debt	-0.865	0.506	-1.71	0.095	36.24

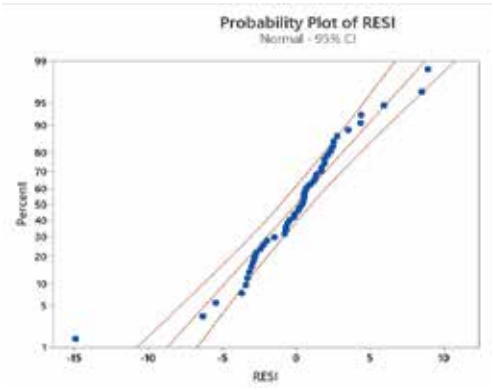


Fig. 2: Normal probability plot of residuals.

The normal probability plot as shown in Figure 2 indicates non-normal behavior of errors. But for an analytical confirmation we perform the Jarque-Bera, and rescaled moment test of residuals. The value of the Jarque-Bera statistic is 51.91. The value of the rescaled moment statistic is 94.41. We know that both the Jarque-Bera and the rescaled moment tests follow a Chi-square distribution with 2 degrees of freedom. At the 5% level, the cut-off value for a Chi-square (2) is 5.99. Hence there exists a very strong evidence against the normality of the errors.

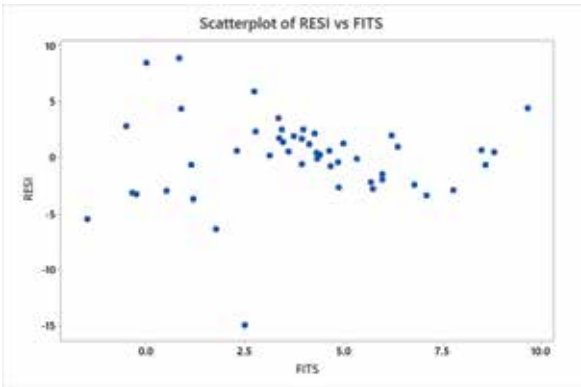


Fig. 3: Residuals Vs Fits Plot.

In addition to normality test, we check for possible heteroscedasticity. We present residuals vs fits plot for this model in Figure 3. Usually a funnel shape indicates the existence of heteroscedasticity. We do not see any clear pattern in this plot. For a formal test we compute the residuals first and then fit the squared residuals on the same set of explanatory variables. The resulting value of the White test statistic is 9.996 which should follow a Chi-square distribution with 7 degrees of freedom. At the 5% level, the calculated value of Chi-square (5) is 14.067, so we fail to reject the null hypothesis of homoscedasticity and declare that there is not enough evidence of heteroscedasticity in this data.

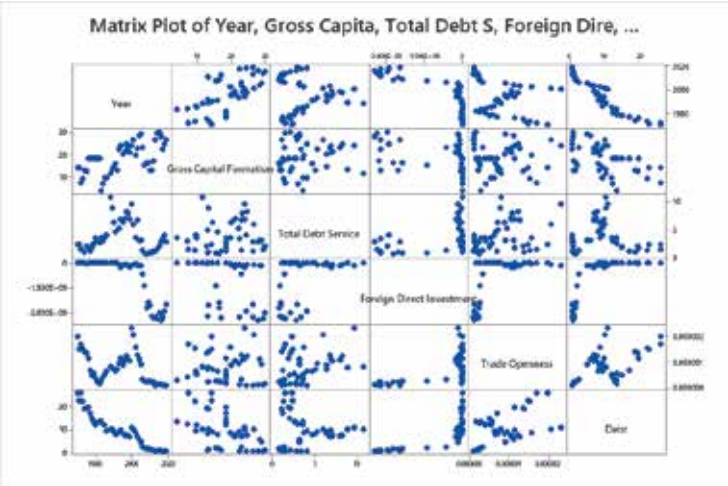


Fig. 4: Matrix Plot of Explanatory Variables.

Table 1 shows that the VIF values corresponding to Year and Debt are excessively large. Figure 2 presents a matrix plot of explanatory variables. We observe from this plot that a very strong linear relationship exists between Year and Debt. The correlation table for the explanatory variables as given in Table 2 also shows a very high correlation between Year and Debt. Thus we can conclude that this fit is severely affected by multicollinearity.

Since all the variables considered here are time series data, there is a possibility that the series may be affected by autocorrelation. We have employed the Durbin-Watson (DW) statistic for checking autocorrelation. The value of the DW statistic for this data is 1.688. At the 5% level the critical region is DW < 1.291. Since the calculated DW value does not fall into the critical region we don't have any strong evidence of autocorrelation for this data.

Table 2: Correlation Matrix for Multiple Regression Model.

	Year	Gross Capital Formation	Total Debt Service	Foreign Direct Investment	Trade Openness
Gross Capital Formation	0.616	-	--	--	--
Total Debt Service	-0.005	0.133	--	--	--
Foreign Direct Investment	-0.785	-0.352	0.413	--	--
Trade Openness	-0.456	-0.059	0.385	0.612	--
Debt	-0.947	-0.527	-0.016	0.722	0.628

Finally we check for outliers and influential observations. We compute Studentized residuals and DFFITS for each observation and found that the observation of 1970 is influential and the observation of 1975 is outlier and at the same time is influential.

Principal Component Regression: Next we fit the model by the principal component regression (PCR). Here we need to compute the eigen values and the eigen vectors associated with the explanatory variables. The eigen values are

Eigenvalue	1.63463E+18	119	23	4	3	-0

Table 3: Eigen Vectors Associated With the Explanatory Variables.

Variable	PC1	PC2	PC3	PC4	PC5	PC6
Year	-0.000	0.823	0.280	0.139	-0.474	-0.000
Gross Capital Formation	-0.000	0.395	-0.907	-0.077	0.128	0.000
Total Debt Service	0.000	0.124	0.051	0.857	0.497	0.000
Foreign Direct Investment	1.000	0.000	0.000	-0.000	-0.000	0.000
Trade Openness	0.000	-0.000	-0.000	0.000	0.000	-1.000
Debt	0.000	-0.389	-0.312	0.490	-0.715	-0.000

and the eigen vectors are given in Table 3.

This leads to a principal component regression fit. The main purpose of a PCR is to fix the multicollinearity problem and the results presented in Table 4 show that it does its job. All the VIF values are less than 10. The maximum VIF is 6.87 which was 40.65 for multiple regression. Two explanatory variables, Total Debt Service, and Debt become statistically significant at a 5% level. But this model has some other issues as well and we suspect the PCR cannot fix those problems. The value of R^2 and adjusted R^2 for this fit are same as the multiple regression model.

Table 4: Summary Results of Principal Component Regression Model.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	398	504	0.79	0.434	
Year	-0.000000	0.000000	-1.23	0.226	6.87
Gross Capital Formation	0.0824	0.0533	1.55	0.130	2.26
Total Debt Service	0.385	0.143	2.69	0.010	1.81
Foreign Direct Investment	-0.554	0.438	-1.26	0.213	4.35
Trade Openness	0.617	0.468	1.32	0.194	5.13
Debt	-5387766	2431356	-2.22	0.032	6.00

The normal probability plot as shown in Figure 5 indicates non-normal behavior of errors. The value of the Jarque-Bera statistic is 42.14. Hence the value of the rescaled moment statistic is 75.57. Both of them reject the hypothesis of normality. Likewise the multiple regression model the Studentized residuals and DFFITS for the PCR model identifies the observation of 1970 as influential and the observation of 1975 as outlier and influential at the same time.

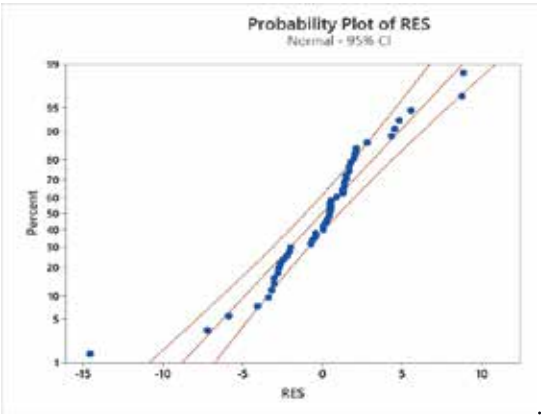


Fig. 5: Normal Probability Plot of Residuals for PCR.

Diagnostic-Robust Regression: We have just experienced that a multiple regression model based on the least squares cannot adequately fit the data. It suffers a huge set back in every respect. The adjusted R^2 is poor (only 22.69%), only one out of six explanatory variables is statistically significant. It also suffers from non-normality and multicollinearity. We have strong evidence of the presence of outliers and influential observations. This is an ideal situation of employing the Diagnostic-Robust approach. The BACON algorithm identifies the explanatory variable Year to be responsible for causing multicollinearity. It also identifies several outliers and influential observations for this data as given in Table 6. We find altogether five unusual cases. Observations corresponding to year 1975 and 1983 are outliers and influential at the same time, observation of 1978 is an outlier and observations of 1970 and 1972 are influential.

Table 5: Diagnostic-Robust Outliers and Influential Observations.

Year	Generalized Studentized Residuals (3.00)	Generalized DFFITS (1.1832)
1970	2.52300	1.42823
1972	-2.05824	-1.32615
1975	-4.61093	-1.37133
1978	3.25223	1.10105
1983	-3.21940	-1.41126

Now we fit the data by the diagnostic-robust approach. We notice a huge improvement in the goodness of fit. The value of R^2 for this fit increases to 60.00% from 32.35% obtained by the multiple regression model. We observe a significant increase in the adjusted R^2 which becomes 54.73% in comparison to 22.69% obtained by the multiple regression model. The summary results of the diagnostic-robust fit are presented in Table 6. For multiple regression we observed that only one (Trade Openness) out of 6 explanatory variables is significant at a 5% level. For the diagnostic-robust model two additional explanatory variables Gross Capital Formation and Debt emerge as highly significant since their corresponding p-values are 0.000. One more variable, that is, Foreign Direct Investment becomes significant at the 10% level.

Table 6: Summary Results of Diagnostic-Robust Regression.

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	12.51	2.33	5.36	0.000	
Gross Capital Formation	-0.3798	0.0736	-5.16	0.000	1.50
Total Debt Service	-0.138	0.191	-0.72	0.475	1.91
Foreign Direct Investment	-0.0000004	0.0000002	-1.73	0.092	3.78
Trade Openness	4596986	1034868	4.44	0.000	2.44
Debt	-0.544	0.122	-4.45	0.000	3.92

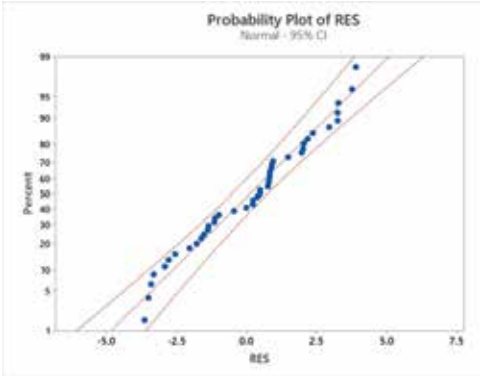


Fig. 6: Normal Probability Plot of Residuals for Diagnostic-Robust Model.

The normal probability plot as shown in Figure 6 shows much normal behavior of errors than before. The value of the Jarque-Bera statistic is 1.479 and the value of the rescaled moment statistic is 2.644. Both of them are less than 5.99, hence there is no evidence of nonnormality of the errors. Finally we check for multicollinearity. All the VIF

values as shown in Table 6 are less than 10. The maximum VIF is only 3.92. Hence there is not enough evidence of multicollinearity for this model.

Now we offer a quick comparison between the three different methods used to fit the GDP growth of Ghana. This comparison is summarized in Table 7.

Table 7: Summary Results of Principal Component Regression Model.

Measures	Multiple Linear	Principal Component	Diagnostic-Robust
R^2	32.35%	32.35%	60.00%
Adjusted R^2	22.69%	22.69%	54.73%
Nonnormality	Yes	Yes	No
Multicollinearity	Yes	No	No
Heteroscedasticity	No	No	No
Affected by Outliers	Yes	Yes	No

Among these three methods the multiple linear regression model based on the least squares method performs worst. It suffered from all major problems except heteroscedasticity. The principal component regression performs marginally better than linear regression, at least it is free from multicollinearity but its performance is not entirely satisfactory. However, the diagnostic-robust approach performs best overall. It outperforms the other two competitors in every respect and produce a fit which is very satisfactory.

Cross Validation Results: Here we report a cross validation study which is designed to evaluate forecasts of GDP growth generated by different models. We have used roughly 80% observations as a training set and the remaining 20% as a test set for the three models considered in our study. In other words, we have used data from 1970 to 2009 as a training set and from 2010 to 2019 as a test set. We offer a numerical comparison to evaluate how different models preform in cross validation to forecast the GDP growth and the summary results are presented in Table 8.

Table 8: Accuracy Measures of Different Forecasts for GDP Growth.

Models	MAE	RMSE	MAPE
Multiple Linear Regression	2.001	2.300	42.77
Principal Component Regression	1.829	2.202	39.00
Diagnostic-Robust Regression	1.771	2.144	32.93

Table 8 clearly shows the merit of using the diagnostic-robust regression in forecasting the GDP growth. It outperforms the other two competitors in every respect. Diagnostic-robust regression produces forecasts that yield the minimum mean absolute error (MAE), the root mean (sum of) squares (of) errors (RMSE), and the mean absolute percentage error (MAPE). This method is followed by the principal component regression. Its MAE, RMSE, and MAPE are smaller than those of the linear regression method. The multiple linear regression performs least in this study.

Conclusions: The variables and methods that can generate the most desirable forecasts of GDP growth in Ghana. We used annual economic data during the period 1970-2019. Since 1985 there is a slightly increasing trend in the GDP growth over the years. The gross capital formation shows a slightly increasing pattern over the years as well. The total debt service shows an increasing pattern between 1970 and 2000, since then it shows a decreasing pattern. Foreign direct investment was steady between 1970 and 2005 and since then it shows a decreasing pattern. Debt tends to decrease over the entire period of time. We employed supervised linear regression, unsupervised principal component regression and semi-supervised diagnostic-robust regression models. The diagnostic-robust regression produces the best fit in terms of R^2 and adjusted R^2. Trade Openness, Gross Capital Formation, and Debt have highly significant impact on the GDP growth. None of these models suffers from heteroscedasticity, but both the linear regression and the principal component regression suffer from nonnormality and poor fit due to outliers and influential observations. The principal component regression does not suffer from the multicollinearity problem. The

diagnostic-robust regression does not suffer from multicollinearity or outlier problem and emerges as the most effective method for fitting the data. A cross validation study supports the fact that the methods which perform better in fitting the model also perform better in generating forecasts. The diagnostic-robust regression performs best in forecasting the GDP growth. It outperforms the other two competitors in every respect, i.e., in terms of possessing the minimum MAE, the minimum RMSE, and the minimum MAPE. This method is followed by the principal component regression. The multiple linear regression model performs worst in this study.

Acknowledgements: The authors express their thanks and gratitude to the anonymous reviewers for giving some useful suggestions that led to considerable improvement in the methodology and presentation of the results.

References:

[1] M. H. Spencer, M. A. Orley, Contemporary Macroeconomics, 8th ed., Worth Publishers, New York, 1993.
[2] O. J. Blanchard, S. Fischer, Lectures on Macroeconomics, MIT Press, Cambridge, Massachusetts, 2014.
[3] J. C. Anyanwu, A. E. Erhijakpor, Do international remittances affect poverty in Africa?, African Development Review, 22 (2010) 51–91.
[4] H. Djulius, Foreign direct investment or external debt and domestic saving: Which has greater impact on growth, ETIKONOMI, 17 (2018) 37–44.
[5] D. A. Ejigayehu, The effect of external debt on economic growth: A panel data analysis on the relationship between external debt and economic growth, Technical Report, Department of Economics, Södertörns University, 2013.
[6] I. A. Elbadawi, J. Benno, C. Ndulu, N. Njuguna, Debt overhang and economic growth in Sub-Saharan Africa, In Zubair Iqbal and Ravi Kanbur (Eds.), External Finance for Low-Income Countries, IMF, 49–76, 1996.
[7] V. Owusu-Nantwi, C. A. Erickson, Public debt and economic growth in Ghana, African Development Review, 28 (2016) 116–126
[8] A. Siddique, E. A. Selvanathan, S. Selvanathan, The impact of external debt on economic growth: Empirical evidence from highly indebted poor countries. Technical Report, Department of Economics, University of Western Australia, 2015.
[9] M. N. Hasan, S. Rana, M. B. Malek, K. R. Das, and N. Sultana, Modeling Bangladesh's gross domestic product using regression approach, Malaysian Journal of Mathematical Sciences, 10 (2016) 233-246.
[10] A. H. M. R. Imon, Subsample Methods in Regression Residual Prediction and Diagnostics, Unpublished PhD Thesis, University of Birmingham, U. K., 1996.
[11] A. S. Hadi, J. S. Simonoff, Procedures for the identification of outliers in linear models, Journal of the American Statistical Association, 88 (1993) 1264–1272.
[12] C. M. Jarque, A. K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, Economics Letters, 6 (1980) 255–259.
[13] M. Habshah, R. Norazan, A. H. M. R. Imon, The performance of diagnostic-robust generalized potentials for the identification of multiple high leverage points in linear regression, Journal of Applied Statistics, 36 (2009) 507–520.
[14] H. White, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, Econometrica, 48 (1980) 817–838.
[15] V. Barnett, T. B. Lewis, Outliers in Statistical Data, 3rd edition, Wiley, New York, 1994.
[16] A. S. Hadi, A. H. M. R. Imon, M. Werner, Detection of outliers, Wiley Interdisciplinary Reviews: Computational Statistics, 1 (2009) 57–70.
[17] S. Chatterjee, A. S. Hadi, Influential observations, high leverage points, and outliers in linear regression, Statistical Science, 1 (1986) 379–393.
[18] D. A. Belsley, E. Kuh, R. E. Welsch, Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, Wiley, New York, 1980.
[19] D. Montgomery, E. Peck, G. Vining, Introduction to Linear Regression Analysis, 4th ed., Wiley, New York, 2006.
[20] R. L. Mason, F. Gunst, Selecting principal components in regression, Statistics and Probability Letters, 3 (1985) 299–301.
[21] A. H. M. R. Imon, Identifying multiple influential observations in linear regression, Journal of Applied Statistics, 32 (2005) 929–946.
[22] A. Bagheri, H. Midi, and A.H.M.R. Imon, The effect of collinearity-influential observations on collinear data set: A Monte Carlo simulation study, Journal of Applied Sciences, 10 (2010) 2086–2093.
[23] S. Chatterjee, A. S. Hadi, Sensitivity Analysis in Linear Regression, Wiley, New York, 1988.
[24] N. Billor, A. S. Hadi, P. F. Velleman, BACON: Blocked adaptive computationally efficient outlier nominators, Computational Statistics and Data Analysis, 34 (2000) 279–298.
[25] A. H. M. R. Imon, Introduction to Regression, Time Series, and Forecasting, Nandita Prokash, Dhaka, 2017.