

Comparative Study of Least Squares, Ridge, Lasso and Principal Component Regression Algorithms

¹Agu S. C., ²Ugah J. O., ³Ajaero G. N. and ⁴Ibegbulem C. A.

¹Madonna University Nigeria, Elele Campus. ²Ebonyi State University Abakiliki, Nigeria. ³Imo State Polytechnic, Omuma Oru-East, Nigeria. ⁴Madonna University Nigeria, Elele Campus.

Corresponding author: sndyaguu@gmail.com

Abstract

Least squares and other traditional statistical methods are widely used for fitting predictive analytic models. However, the methods have a number of limitations that can lower the predictive accuracy of the models. The study considered a typical dataset with two main objectives. First, to fit four predictive models by applying least squares and three trending machine learning algorithms - ridge regression, lasso regression and principal component regression and second, compare their performance. The methods were assessed using 5-fold cross validation technique. The result revealed that principal component regression outperformed the others with minimal mean square error of $-7.55E+21$. This suggested that principal component regression would perform optimally for building predictive model for the related dataset.

Keywords: Machine learning, Least squares, Ridge regression, Lasso regression, Principal component regression, 5-fold cross validation, Gross domestic product.

Introduction

The existing work in India used least squares algorithm to predict GDP based on some selected macroeconomic variables, and applied coefficient of determination R^2 , correlation and ANOVA in analysing the measurement of the model's fit and the impact of these variables on the GDP (Divya & Rama, 2014). Whereas the relationship between GDP, exports, imports, foreign exchange reserve and foreign debt in Indonesia was determined using Vector Auto regression (VAR) (Syrkri, 2020), a recent study utilized granger causality technique to test the significant effects of financial structure in bank credit to the private sector ratio (BC), market capitalization ratio (MC), liquid liability ratio (LLR), turnover ratio (TR) and value of traded share (VTS) on GDP of Nigerian economy (Yua *et al*, 2017).

Least squares, VAR and granger causality technique are traditional statistical methods which disappointments over measured GDP in recent years, have spurred widespread concerns about whether the statistical systems are efficiently capturing economic variables (Karen and Louise, 2018). According to Gareth *et al* (2017), the disappointments are attributed by the weaknesses inherent in least squares and the traditional statistical method as (1) high assumption of linear relationship between predictor variables and the response variable which in most cases are violated. This has the effect of reducing the predictive accuracy of the model. (2) high dimensional data, that is, when the number of observations are not much more than the number of predictors $n \gg p$ as it is in this study,

the variability of least squares approach increases which tends to overestimate the model leading to inaccurate predictions and when the number of observations are lesser than the number of predictors $n < p$, the least squares will not have a unique solution (3) Using R^2 statistics and correlation as the metrics for measuring the fitness of the model will perform well on the training data set but will perform poorly on a test data set which will lead to poor prediction on unseen data for future predictions. (4) No support for tuning parameter - cost function thus; bias-variance trade-off cannot be applied to minimize the test mean square error.

In the light of the limitations, the study aims at building three machine learning predictive models by applying ridge regression, lasso regression and principal component regression and one statistical model using ordinary least squares, revealing strategies which perform better than others. The notion is that data scientists should accept with responsiveness and espouse machine learning as a potent tool to predicting related projects.

Method and Techniques

The data for the study was secondary data collected from World Bank (World Bank, 2021). The population size of the study covered the period from 1970 to 2020. But, due to the fact that the Nigeria's macroeconomic indicators and the GDP data were not captured by the World Bank within the period between 1970 and 1980, the sample size of the study is between the periods of 1981-2019. The selected independent macroeconomic (predictor) variables for the study were the Nigeria's export rate, import rate, official exchange rate, total population, foreign direct investment, government expenditure and Oil revenue, while the dependent (response) variable is the Nigeria's GDP.

The study applied Cross Industry Standard Process for Data Mining (CRISP-DM) methodology phases:

1. Data preparation: the seven macroeconomic indicators and the GDP data were individually extracted from World Bank, merged into one dataset, removed of the missing values from the dataset, prepared the independent and the dependent variables, and then split the training and testing datasets.
2. Modelling: the dataset was trained, and the models for the dataset were individually fitted using the Least squares, Lasso regression, Ridge regression, and Principal component regression algorithms.
3. Evaluation: the assessment metrics used 5-fold Cross Validation to determine the Mean Square Error (MSE) of the 4 different model estimation approaches.

Theory/Calculation

The general form of representing the relationship between GDP and the macroeconomic factors dataset of the study is expressed in equation 1 (Wikipedia, 2021).

$$gdp = f(imr, exr, xcr, pop, fdi, fge, oil) + e \quad \text{equation 1}$$

where the fixed but unknown function f represents the information that the predictors provide about the GDP and the random variable e represents the error term which is independent of the predictors and has mean zero.

Two-step model-based approach

The Linear model and the estimation of coefficients were the two step methods adopted by the study.

Linear model

The study assumes that the function f in equation 1 is linear in $(imr, exr, xcr, pop, fdi, fge, oil)$ as shown in Equation 2.

$$gdp = \beta_0 + \beta_1 imp + \beta_2 exp + \beta_3 xcr + \beta_4 pop + \beta_5 fdi + \beta_6 fge + \beta_7 oil \quad equ 2$$

The assumption simplifies the problem of estimating f (Gareth *et al*, 2017). Instead of having to estimate an entirely arbitrary 7-dimensional function $f(imr, exr, xcr, pop, fdi, fge, oil)$, only 8 coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7$ are needed to be estimated.

β_0 represents the equation's constant (intercept) and $\beta_1 \dots \beta_7$ represent the coefficients of the selected macroeconomic variables that define the relationships between the variables and the GDP.

Specifically, β_0 is interpreted as the expected value of the gdp when all the macroeconomic variables are zero. β_1 is interpreted as the average effect on gdp of a one unit increase in import rate, holding all other predictors fixed. The same interpretation of β_1 is given to $\beta_2 \dots \beta_7$ for each of their respective macroeconomic variables.

Estimation of the coefficients

The coefficients of equation 2 need to be estimated to allow for establishment of the relationship between the macroeconomic variables and the GDP as shown in equation 3.

$$\widehat{gdp} = \hat{\beta}_0 + \hat{\beta}_1 imp + \hat{\beta}_2 exp + \hat{\beta}_3 xcr + \hat{\beta}_4 pop + \hat{\beta}_5 fdi + \hat{\beta}_6 fge + \hat{\beta}_7 oil \quad equ 3$$

where $\widehat{gdp}$ represents the predicted GDP, while $\hat{\beta}_0, \hat{\beta}_1 \dots \hat{\beta}_7$ represent the estimated coefficients.

Analysis of the Algorithms for Estimating the Coefficients

Least squares

Least squares compute the estimated coefficients $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_7$ in equation 3 by utilizing some calculus in order to minimize the Residual Sum of Squares (RSS) given in equation 4 (Wikipedia, 2021).

$$RSS = \sum_{i=1}^n (gdp_i - \widehat{gdp}_i)^2$$

$$= \sum_{i=1}^n (gdp_i - \hat{\beta}_0 - \hat{\beta}_1 imp_i - \hat{\beta}_2 exp_i - \hat{\beta}_3 xcr_i - \hat{\beta}_4 pop_i - \hat{\beta}_5 fdi_i - \hat{\beta}_6 fge_i - \hat{\beta}_7 oil_i)^2 \quad 4$$

where n is the total number of years, and i represents individual yearly observations.

Ridge regression

The ridge regression coefficient estimates represented by $\hat{\beta}_\lambda^R$ are computed by minimizing two terms: (a) The least square Residual Sum of Squares (RSS) term and (b) The regularization penalty term as shown in equation 5 (Saptashwa, 2018).

$$RSS + \lambda \sum_{i=1}^n \hat{\beta}_i^2$$

$$= RSS + \lambda (\hat{\beta}_1^2 + \hat{\beta}_2^2 + \hat{\beta}_3^2 + \hat{\beta}_4^2 + \hat{\beta}_5^2 + \hat{\beta}_6^2 + \hat{\beta}_7^2) \quad 5$$

where lambda, λ is a tuning parameter, the second term, $\lambda(\hat{\beta}_1^2 + \hat{\beta}_2^2 + \hat{\beta}_3^2 + \hat{\beta}_4^2 + \hat{\beta}_5^2 + \hat{\beta}_6^2 + \hat{\beta}_7^2)$, known as the regularization penalty or cost function is the value that minimizes the expected test mean square error by simultaneously yielding low bias and low variance (Gareth *et al*, 2017). The equation can also be specified as below:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(GDP_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right) \right\} \text{ subject to } \sum_{j=1}^p \beta_j^2 \leq s$$

Lasso regression method

The Lasso regression coefficient estimates represented by $\hat{\beta}_\lambda^L$ are the values that minimize two terms as shown in equation 6 (Saptashwa, 2018).

$$RSS + \lambda \sum_{i=1}^n |\hat{\beta}_i|$$

$$= RSS + \lambda (|\hat{\beta}_1| + |\hat{\beta}_2| + |\hat{\beta}_3| + |\hat{\beta}_4| + |\hat{\beta}_5| + |\hat{\beta}_6| + |\hat{\beta}_7|) \quad 6$$

where the regularization penalty, $\lambda \sum_{i=1}^n |\beta_j|$ has the effect of forcing some of the coefficient estimates to zero when the tuning parameter λ is sufficiently large (Gareth *et al*, 2017). The equation can also be specified as below:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^n \left(GDP_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right) \right\} \text{ subject to } \sum_{j=1}^p |\beta_j| \leq s$$

Principal component regression

First, we obtain the transformed predictors $Z_1, Z_2, \dots, Z_m$ representing $M < P$ linear combinations of principal components from the original predictors p (Gareth *et al*, 2017), (Sabharwal & Anjum, 2016) given as in equation 7

$$Z_m = \sum_{j=1}^p \phi_{jm} X_j \quad 7$$

for some constants $\phi_{1m}, \phi_{2m}, \dots, \phi_{pm}, m = 1, \dots, M$.

Second, we apply least squares to fit the linear regression model using the M predictors given as in equation 8

$$gdp_i = \theta_0 + \sum_{m=1}^M \theta_m Z_{im} + \epsilon_i, \quad i = 1, \dots, n \quad 8$$

where the regression coefficients are given by $\theta_0, \theta_1, \dots, \theta_M$

5-fold cross-validation (5-fold CV) technique

5-fold CV metrics was used for evaluating the mean square error of the least squares, ridge, lasso and principal component regressions. It randomly divides the set of observations in the Nigeria's dataset into 5 groups, or folds, of approximately equal size. The first fold is treated as a validation set, and the method is fit on the remaining 5 – 1 folds. The mean squared error (MSE), MSE_1 , is then computed on the observations in the held-out fold. This procedure is repeated 5 times; each time, a different group of observations is treated as a validation set. This process results in 5 estimates of the test error, $MSE_1, MSE_2, \dots, MSE_5$. The 5-fold CV estimate is computed by averaging these values in equation 9 (Gareth *et al*, 2017).

$$CV_{(5)} = \frac{1}{5} \sum_{i=1}^5 MSE_i \quad \text{equation 9}$$

Besides non-computational advantages to performing 5-fold CV, it also offers bias-variance trade-off to enable us select the optimal cost function.

Results and Discussion

It is important to establish the models that would impact predictive accuracy to aid data scientists in related research projects. The contribution of this study is to use machine

learning methods to build predictive models. In this regard, the study applies PCR, RR, LR, and OLS methods in building four predictive models. We intend to apply the tuning parameter (λ) for RR and LR that will accommodate the bias-variance trade-off limitation of the traditional OLS statistical method, obtain $M < P$ transformed predictors of PCR and then evaluate the models' predictive accuracy by utilizing 5-fold CV technique to estimate the Mean Square Errors of the corresponding models.

The lambda, λ (cost function) values of $1e-15$, $1e-10$, $1e-8$, $1e-3$, $1e-2$, 1, 5, 10, 20, 30, 35, 40, 50, 55, and 100 were used for tuning both the Ridge and Lasso regressions respectively, The result revealed that the best λ for Ridge and Lasso were 100 and 0.001 respectively and four (4) principal components were used for the transformation of the linear combination of the original predictors of the Principal Component Regression approach.

Predictive accuracy

The best turning parameters 100 and 0.001 and 2 principal components were used in the processing of the 5-fold CVs for the Ridge, Lasso and principal component regression methods respectively. The line in Figure 1 displayed the 5-fold CV MSE of the 39 observations of Nigeria's dataset. The small squares on the line indicate the MSE associated with the corresponding methods on the x-axis. Consequently, the OLS and lasso regression methods have highest MSE followed by ridge regression method. The principal component regression (PCR) minimized the MSE. The last small square on the line indicate $\text{Var}(e)$, the irreducible error which corresponds to the lowest achievable MSE among all possible methods. Hence, the MSE represented by the PCR is closest to optimal and is suggested for use in building a predictive model for Nigeria's GDP.

Conversely, the findings agree with the study conducted by Karen and Louise, (2018) revealing that machine learning methods capture predictive ability and perform better than traditional OLS but in contrast to the work of Divya and Rama (2014) which revealed otherwise.

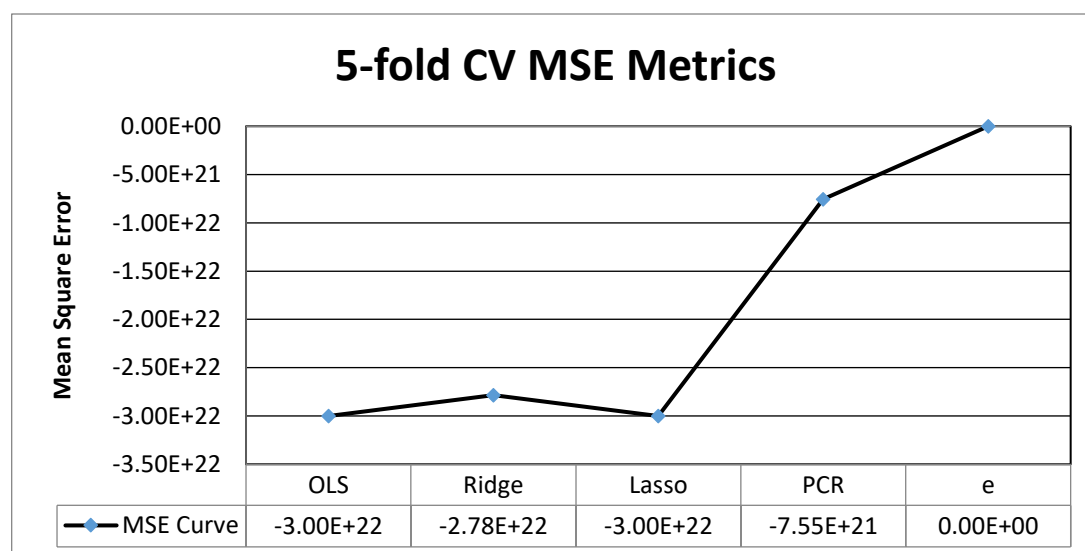


Figure 1: 5-fold Cross validation MSE plot for OLS, Ridge, Lasso, PCR and e

Conclusion

Due to inflexible and high linear relationship assumption of least squares method between the predictor variables and the response variable, adopting principal component regression algorithm that allows for transformation of the original macroeconomic predictors into less number and tuning parameters in building model yields optimal prediction accuracy.

References

- Divya, K. H., & Rama, D. V. (2014). A Study on Predictors of GDP: Early Signals. *Journal of Procedia Economics and Finance*, 11, 375 – 382. [https://doi.org/10.1016/S2212-5671\(14\)00205-6](https://doi.org/10.1016/S2212-5671(14)00205-6)
- Gareth, J., Daniela, W., Trevor, H., & Robert T. (2017). *An Introduction to Statistical Learning with Application in R*. Springer (Chapters 3,5, & 6)
- Karen, D., & Louise, S. (2018). *GDP as a Measure of Economic Well-being*. Retrieved from <https://www.brookings.edu/research/gdp-as-a-measure-of-economic-well-being/>. Accessed 6th March, 2021
- Pamanalytics. (2020). *The CRISP-DM Methodology*. Retrieved from <http://www.pamanalytics.com/downloads/The%20CRISP-DM%20Methodology.pdf> Accessed 4th October, 2021.
- Sabharwal, C., and Anjum, B. (2016). *Data Reduction and Regression Using Principal Component Analysis in Qualitative Spatial Reasoning and Health Informatics*. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1870-90442016000100031. Accessed 11th November, 2021
- Saptashwa, B. (2018). *Ridge and Lasso Regression: L1 and L2 Regularization*. Retrieved from <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>. Accessed 11th November, 2021
- Syrkri, A. U. (2020). The Relationship between Gross Domestic Products with International Balance of Payment: Empirical Evidence from Indonesia. *Journal of Development Economics*, 5, 103-119. <http://dx.doi.org/10.20473/jde.v5i2.18275>
- Wikipedia. (2021). *Residual sum of squares*. Retrieved from https://en.wikipedia.org/wiki/Residual_sum_of_squares. Accessed 11th November, 2021
- World Bank. (2021). *Indicators*. Retrieved from <https://data.worldbank.org/indicator/>. Accessed 26th April, 2021
- Yua, H., Adoms, F. U., Okaro, C. S., & Ogbonna, K. S. (2017). Causal Relationship between Financial Structure and Economic Growth in Contemporary African Economy: A Case Study of Nigeria from 1990-2018. *European – American Journal*, 8(3), 59-68