Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.7, No.10, 2017

A Comparative Study on Bias Regression Methods in the Presence of Multicollinearity Based on Gamma and Chi Square Distributions

Usman, U. Yau, S.A. Zakari, Y.
Department of Mathematics, Usmanu Danfodiyo University, Sokoto. Nigeria

Abstract
The aim of this study is to compare some regression methods in the presence of multicollinearity problem. This problem makes the estimated regression coefficients by least squares method to be conditional upon the correlated predictor variables in the model. It is also a condition in a set of regression data that have two or more regressors which are redundant and have the same information. Therefore, some regression methods that handle with multicollinearity such as partial least square regression (PLSR), ridge regression (RR) and lasso regression (LR) had reported. In this paper, the methods were compared using simulated data that follows gamma and chi square distributions with P=4 and 10, and n=60 and 90. All results were compared with each other through Mean Square Log Error (MSLE), Mean Absolute Error (MAE) and R^2 of their estimated values for different methods. The results show that when P=4 and n=60 RR is better methods with gamma distribution, but with chi square distribution PLRS is better methods. Also, when P=4 and n=90, RR shows better results with gamma distribution but with chi square distribution all methods have equal predictive ability. However, at P=10 and n=60 RR performed better with both gamma and chi square distributions while RR shows better results at both gamma and chi square distributions when P=10 and n=90.
Keywords: Multicollinearity, Partial Least Square Regression, Ridge Regression, Principal Component Regression

1. Introduction

In regression, the objective is to explain the variation in one or more response variables, by associating this variation with proportional variation in one or more explanatory variables. In regression, one frequent obstacle is that several of the explanatory variables will vary in rather similar ways. As a result, their collective power of explanation is considerably less than the sum of their individual powers. This process is called multicollinearity, is a common problem in regression analysis. Handling multicollinearity in Regression analysis is important because least squares estimations assume that predictor variables are not correlated with each other.

Econometrics were the first people who paid attention to the multicollinearity, after the international economic crisis in 1928, some of them discussed it as a problem, others took it as a study case specially in preparing budgets because the relation between the variables were very complicated. No precise definition of multicollinearity has been firmly established in the literature [1]. Literally, two variables are collinear if the data vectors representing them lie on the same line. More generally, k variables are collinear if the vectors that represent them lie in a subspace of dimension less than k, i.e. if one of the vectors is a linear combination of others. In practice, such "exact multicollinearity" rarely occurs. A broader notation of multicollinearity is therefore needed to deal with the problem as it affects statistical estimation. Thus, two variables are collinear if they lie almost on the same line, that is, if the angle between them is small. In the event that one of the variables is not constant, this is equivalent to saying that they have a high correlation between them. Also, multicollinearity refers to the situation where there is either an exact or approximately exact linear relationship among the explanatory variables [2]. It is a problem that always occurs when two or more predictor variables are correlated with each other in the applications of regression analysis. This problem makes the estimated regression coefficients by least squares method to be conditional upon the correlated predictor variables in the model. It is also a condition in a set of regression data that have two or more regressors which are redundant and have the same information. Redundant information means what one variable explains about Y is exactly what the other variable explains. In this case, the two or more redundant predictor variables would be completely unreliable since the $\beta_j$would measure the same effect of those $X_i$and the same goes for the other β. Furthermore, $(X'X)^{-1}$ would not exist because the denominator $(1-r^2)$ is zero. As a result, the estimates of β cannot be found since the elements of the inverse matrix and coefficients become quite large [3]. [4] compared three regularized regression methods by Root Mean Square Error(RMSE) and Root Mean Square Error Cross Validation(RMSECV) on real data. The data consists of the Gross Domestic Product Per Capita (GDPPC) in Turkey. Ordinary Least Squares regression (OLS) and Ridge Regression (RR) were found to be the best among other because the value of RMSE is minimum while partial least squares regression (PLSR) is to be the best among other because RMSECV is minimum. Finally, they concluded in their study that PLSR was the superior method in terms of the prediction ability as compared to the other regularized models.

[5] compared partial least squares regression, principal component regression, ridge regression and multiple linear regression methods in modeling and predicting daily mean PM10 concentrations on the base of various meteorological parameters obtained for the city of Ankara, in Turkey. The analysed period is February 2007. Their results show that while multiple linear regression and ridge regression yield somewhat better results for fitting to this dataset, principal component regression and partial least squares regression are better than both of them in terms of prediction of PM10 values for future datasets. In addition, partial least squares regression was the remarkable method in terms of predictive ability as it had a close performance with principal component regression even with less number of factors. [6] compared the performance of robust biased Robust Ridge Regression (RRR), Robust Principal Component Regression (RPCR) and RSIMPLS methods with each other and their classical versions known as RR, PCR and PLSR in terms of predictive ability by using trimmed Root Mean Squared Error (TRMSE) statistic in case of both of multicollinearity and outliers existence in an unemployment data set of Turkey. Their analysis results show that RRR model is chosen as the best model for determining unemployment rate in Turkey for the period of 1985-2012. Robust biased RRR method showed that the most important independent variable affecting the unemployment rate is Purchasing Power Parities (PPP). The least important variables affecting the unemployment rate are Import Growth Rate (IMP) and Export Growth Rate (EXP). Hence, any increment in PPP cause an important increment in unemployment rate, however, any increment in IMP causes an unimportant increase in unemployment rate. Also, any increment in EXP causes an unimportant decrease in unemployment rate.

2.0 Regression Methods for Multicollinearity Problem

The biased regression methods and latent variables can handle the problem of multicollinearity in linear regression modelling. There are lots of proposed methods to shrinkage or select subset of independent variables [7, 8, 9]. Thus, several methods have been established to overcome the deficiencies of multicollinearity. Among such methods are continuum regression of [10, 11] Partial Least Square, Principal Component Ridge Regression [12]. In this research, we considered ridge, lasso and partial least square regression methods.

2.1 Partial Least Squares Regression

The PLSR searches for a set of components (called latent vectors) that performs a simultaneous decomposition of X and Y with the constraints that this components explain as much as possible the covariance between X and Y. In this method, the component is extracted from which the rest of the components are extracted in such a way that they are uncorrelated (orthogonal). How this algorithm functions will now be described to show how the PLS method works. The first is defined as:

$$t_i = W_{11}X_1 + W_{12}X_2 + \cdots + W_{ip} = \sum W_{ij}\, X_j \qquad (1)$$

Where, X_j are the explanatory variables, Y is the dependent variables.

The W_{ij} is the coefficient:

$$W_{ij} = \frac{cov(X_j,Y)}{\sqrt{\sum_j^p cov(X_j,Y)^2}},\ \text{j=1, 2, 3…}p \qquad (2)$$

From which it can be deduced that in order to obtain W_{ij} the scalar product (X_j, Y) must be calculated for each$J = 1, 2 \ldots P$.

Calculating the second component is justified when the single component model is inadequate i.e. when the explanatory power of regression is small and another component is necessary. The second component is denoted by t_2 and it will be a linear combination of the regression residues of X_j variables on components t_1 instead of the original variables. In this way, component orthogonality is assured. To do this, the residual for the single component regression must be calculated which will be,

$e_1 = Y - \hat{Y} = Y - \beta_1 t_1$with

$$\beta_1 = \frac{cov(y_i,t_i)}{\|t_1\|^2} \qquad (3)$$

The second component is obtained as:

$$t_2 = W_{21}e_{11} + W_{22}e_{12} + \cdots + W_{2p}e_{1p} \qquad (4)$$

With $$W_{2j} = \frac{cov(e_{ij},e_1)}{\sqrt{\sum_j^p cov^2(e_{ij},e_1)}},\ \text{j=1, 2, 3…}p \qquad (5)$$

The residuals e_{ij} are calculated by computing the simple regression of $x_j$ont_1,

$X^*{}_j = \alpha_j t_j,\ j = 1, 2, \ldots, p$ therefore,

$$e_{ij} = X_j - X^*{}_j = X_j - \alpha_j t_j \qquad (6)$$

Where, the estimators of the regression coefficients have been calculated thus:

$$\alpha_j = \frac{cov(x_j,t_1)}{\|t_1\|^2} \qquad (7)$$

Now with e_i and e_{ij}, only the scalar products have to be computed $cov(e_i, e_{ij})$,
for j = 1 . . .P, to be able to compute t_2.
To construct subsequent components, the same steps are performed as for the two previous components. This iterative procedure is continued until the number of components to be retained is significant.

2.2 Ridge Regression

When multicollinearity exists, the matrix X′X, where X consists of the original regressors, becomes nearly singular. Since $Var(\beta) = \delta^2(X'X),^{-1}$ and the diagonal elements of $(X'X)^{-1}$ become quite large, this makes the variance of β to be large. This leads to an unstable estimate of β when OLS is used.

Steps in Performing Ridge Regression

STEP I**:**

Consider the following regression model:

$$Y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \varepsilon \quad (8)$$

where, $\beta_1, \beta_2, \beta_3$ etc. are the parameters of the model and ε are random terms.

STEP II:

Standardize data by subtracting each x observation from its corresponding mean and dividing by its standard deviation i.e. $\frac{x_i - \mu_i}{\sqrt{\delta_i}}$

STEP III:

Arrange the predictors into convenient matrix. Suppose we have n observations of k predictors, this will be a $n \times k$ matrix x. And arrange the key parameters into a β**.** So that viewing the response variable as an n-vector, our model becomes:

$$y = x\beta + \varepsilon \quad (9)$$

where, ε is now a vector of the random noise in the observed data vector Y**.**

Note: the least square parameter β_{Ls} can be estimated by finding the parameter values which minimized the sum square residuals i.e.

$SSR = \sum(Y - X\beta)'(Y - X\beta)$. The solution turns out to be a matrix equation,

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y \quad (10)$$

where, X' is the transpose of the matrix X**.**

According to [13], the potential instability in using the least squares estimator could be improved by adding a small constant λ to the diagonal entries of the $X'X$ marix before taking its inverse.

The result is the Ridge regression estimator

$$\hat{\beta}_{RIDGE} = (X'X + \lambda I)^{-1}X'Y \quad (11)$$

where I is the $p \times p$ identity matrix and $X'X$ is the correlation matrix of independent variables values of lambda lie in the range (0 and 1). When $\lambda = 0$, $\hat{\beta}_{RIDGE}$ becomes $\hat{\beta}_{OLS}$.Obviously, a key aspect of ridge regression is determining what the best value of the constant that is added to the main diagonal of the matrix $X'X$ should be to maximize prediction. There are many procedures for determining the best value. The simplest way is to plot the values of each $\hat{\beta}_{RIDGE}$versus λ. The smallest value for which each ridge trace plot shows stability in the coefficient is adopted [14].

2.3 LASSO Regression

The LASSO (Least Absolute Shrinkage and Selection Operator), [8] is another regularization method, but here the penalty is applied to the sum of the absolute values of the regression coefficients, the L_1 norm. The penalty function is given by Pen(β) = $\lambda_i \sum_{j=1}^{p} |\beta_j|$

The objective is to minimize

$$\hat{\beta}_{LASSO} = argmin_{BERP}(Y - X\hat{\beta})^T(Y - X\hat{\beta}) + \lambda_i \sum_{j=1}^{p} |\beta_j| \quad (12)$$

Where λ is a non-negative regularization parameter.

Since the penalty term is no longer quadratic, there is no explicit formula for the mean squared error of the Lasso estimator.

Generally, the $Bias(\hat{\beta}_{LASSO})$ also increases as the tuning parameter λ increases. While the variance, $Var(\hat{\beta}_{LASSO})$ decreases.

2.4 Simulation Study

In this research work, Monte Carlos was performed with different levels of multicollinearity with [15]. By

following [16, 17, 18, 19, 20, 21, 22].

Also, four sets of correlations were considered corresponding to ρ = 0.7, 0.8, 0.9 and 0.99 as used by [23]. Using the condition number, $CN = \frac{\lambda_{max}}{\lambda_{min}}$, it can be shown that these values of ρ will include a wide range of low, moderate and high correlations between variables. The n observations for the dependent variable Y are determined by:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \varepsilon_i, n = 1,2, \ldots, p \quad (13)$$

We simulated data that follows gamma and chi square distributions with P=4, 10 predictors variables [24] for each observations (n = 60 and 90) respectively. The goal is to develop a linear equation that relates all the predictor variables to a response variable. For the purpose of comparing the three methods under multicollinearity, the analysis was done using R software. Also, simulation was replicated 100 times in order to obtain the approximate distribution considered in the study in real life situation.

2.5 Evaluation of the results

Table 1: MSLE of LR, RR and PLSR when P=4 and n=60

Regression	Gamma	Chi Square
LR	0.54113	2.167892
RR	0.53982*	1.990815
PLSR	0.59923	1.98776*

The Table 1 shows the results of MSLE of the biased regression methods. It revealed that RR has minimum error under gamma distribution while PLSR has higher predictive power over both LR and RR when data follows chi square distribution. In summary, RR performed better under gamma distribution. And PLSR is better in chi square distribution.

Table 2: MSLE of LR, RR and PLSR when P=4 and n=90

Regression	Gamma	Chi Square
LR	4.15214	2.479848
RR	1.13257*	2.482757
PLSR	6.61317	2.479466*

From Table 2, it shows that RR has the least values in gamma distribution than both LR and PLSR. Also, under chi square distribution PLSR has the minimum error then both LR and RR.

Table 3: MAE of LR, RR and PLSR when P=4 and n=60

Regression	Gamma	
LR	0.91037	13.99083
RR	0.50010*	13.62415
PLSR	5.50016	13.62365*

The result obtained from Table 3 revealed that RR has minimum error under gamma distribution. Also, in chi square distribution PLSR is the more efficient. However, it shows that only RR performed better in gamma distribution and PLSR also performed better in chi square distribution.

Table 4: MAE of LR, RR and PLSR when P=4 and n=90

Regression	Gamma	
LR	4.31415	1.32001*
RR	0.53827*	9.93310
PLSR	6.03829	18.50231

From the results presented in Table 4 RR performed better than both LR and PLSR under gamma distribution while LR performed better in chi square distribution than RR and PLSR. Thus, RR is better under gamma distribution and also, LR is better when data follows chi square distribution.

Table 5: R^2 of LR, RR and PLSR when P=4 and n=60

Regression	Gamma	
LR	0.02819*	0.02477*
RR	0.03052	0.06667
PLSR	0.09180	0.07746

Table 5 presents R^2 LR, PLSR and RR. The results revealed that LR has minimum error under gamma and chi square distributions, This means that, there is higher predictive power observed compared to PLSR and RR under these distributions. However, LR performed better under gamma and chi square distributions.

Table 6: R^2 of LR, RR and PLSR when P=4 and n=90

Regression	Gamma	
LR	0.00201*	0.13471
RR	0.01638	0.08811*
PLSR	0.08807	1.98122

Table 6 shows the values of R^2 when P=4 and n=90. It revealed that LR has minimum error under gamma distribution. This means that, there is higher predictive power observed compared to PLSR and RR under gamma distribution, while RR has minimum error under chi square distribution than LR and PLSR. However, LR performed best under gamma distribution while RR is the best under chi square distribution.

Table 7: MSLE of LR, RR and PLSR when P=10 and n=60

Regression	Gamma	
LR	15.46718	5.66001
RR	9.46836*	0.55281*
PLSR	19.46713	11.87111

Table 7 present MSLE of LR, PLSR and RR. The results revealed that RR has minimum error under gamma and chi square distributions, This means that, there is higher predictive power observed compared to PLSR and LR under those distributions. However, RR performed better under gamma and chi square distributions.

Table 8: MSLE of LR, RR and PLSR when P=10 and n=90

Regression	Gamma	
LR	21.03623	9.11401
RR	13.03634	9.08934*
PLSR	4.96546*	1.91891

Table 8 shows that PLSR has minimum error under gamma distribution, This means that, there is higher predictive power observed compared to LR and RR gamma distribution, while RR has minimum error under chi square distribution.

Table 9: MAE 0f LR, RR and PLSR when P=10 and n=60

Regression	Gamma	
LR	0.51204	1.76542*
RR	0.51144	2.89200
PLSR	0.51142*	1.90408

From Table 9, the results shows that PLSR has minimum error under gamma distribution, this means that, there is higher predictive power observed compared to LR and RR while LR has minimum error under chi square distribution.

Table 10: MAE of LR, RR and PLSR when P=10 and n=90

Regression	Gamma	
LR	0.66478	6.23015
RR	0.66392*	0.31900*
PLSR	0.66409	1.97061

Table 10 shows that RR has minimum error under both gamma and chi square distributions, this signifies that, there is higher predictive power observed compared to PLSR and LR.

Table 11: R^2 of LR, RR and PLSR when P=10 and n=60

Regression	Gamma	
LR	2.18105	0.98990
RR	0.18048*	0.30562*
PLSR	1.62261	1.08611

Table 11 shows that RR has predictive ability in both gamma and chi square distributions than both PLSR and LR.

Table 12: R^2 of LR, RR and PLSR when P=10 and n=90

Regression	Gamma	
LR	1.00923	1.09812*
RR	0.09321*	1.10101
PLSR	0.83872	1.96711

From Table 12 which presents the results of R^2 of LR, PLSR and RR. The results shows that RR has minimum error under gamma distribution. This means that, there is higher predictive power observed compared to PLSR and LR. Also, when data follows chi square distributions LR is the most efficient.

3. Conclusion

In this paper, we compared three biased regression methods namely: LR, RR and PLSR when there exist multicollinearity problem in the independent variables. Also, we used MSLE, MAE and R^2 as a predictive ability. As such, based on the discussions of the results, we concluded that at P=4 and n=60, RR is the most efficient when data follows gamma distribution, when data follows chi square distribution PLSR is the most efficient. At P=4 and n=90, RR shows better result in gamma while in the chi square distribution all methods have equal predictive ability. Furthermore, at P=10 and n=60, RR performed better in both gamma and chi square distributions. Lastly, at P=10 and n=90, RR shows result when data follows gamma and chi square distributions.

Reference

1 Balsely, H.L. (1970). Quantitative Research Method Business and Economics. Random House Publisher, New York.
2. Gujarati, D.N. (2003). *Basic Econometrics.* 4th ed., McGraw Hill, New York.
3. Younger, M.S. (1979). *A Handbook for Linear Regression,* DUXBURY Press. USA.
4. Yeniay, O. and Goktas, A. (2002). A Comparison of Partial Least Squares Regression with other Prediction Methods. *Hacettepe Journal of Mathematics and Statistics*. 31: 99-111.
5. Esra, P. and Suleyman, G. (2015). The Comparison of Partial Least Squares Regression, Principal Component Regression and Ridge Regression with Multiple Linear Regression for Predicting PM10 Concentration Level Based on Meteorological Parameters. *Journal of Data Science*. 13(4): 663-691. 29.
6. Esra, P. and Semra, T. (2016). The Comparison of Classical and Robust Biased Regression Methods for Determining Unemployment Rate in Turkey: Period of 1985-2012. *Journal of Data Science.* 14(4): 739-768.
7. Kejian, L. (2004). More on Liu-Type Estimator in Linear Regression, Communications in Statistics – *Theory and Methods,* 33(11): 2723-2733.
8. Tibshirani, R. (1996). Regression Shrinage and Selection via the Lasso. *Journal of the Royal Statistical Society, SeriesB58,* 267-288.
9. Toker, S. and Kaçiranlar, S. (2013), On the Performance of Two Parameter Ridge Estimator under the Mean Square Error Criterion, *Applıed Mathematıcs and Computatıon*, 219, 4718-4728.
10. Stone, M. and Brooks, R.J. (1990). Continuum Regression. Cross-Validated Sequentially Constructed Prediction Embracing OLS, PLSR, PCR. *Journal of royal Statistics Society.* 52:237-269.
11. Sundberg, R. (1993). Continum Regression and Ridge Regression. *Journal of Royal Statistics*. 55: 653-659.
12. Bjoksilon, A. and Sundberg, R. (1999). A Generalized view on Continuum Regression. *Scandinavian Journal of Statistics.* 25: 17-30.
13. Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression Applications to Non-orthogonal Problems, *Tecnometrics.* 12(1): 69-82.
14. Myers, R.H. (1990). *Classical and Modern Regression with Applications*. 2nd edition, Duxbury Press.
15. R Core Team (2016). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL.https://www.R-project.org.
16. Mc Donald, G.C. and Galarneau, A. (1975). A Monte Carlo Evaluation of some Ridge Type Estimators. *Journal of American Statistical Association,* 70, 407-416. http;//.dx.doi.org/10.1080/01621459.1975.10479882.

17. Wichern, D. and Churchill, G. (1978). A Comparison of Ridge Estimators. *Technometrics*, 20, 301-311.
18. Gibbons, D.G. (1981). A Simulation Study of some Ridge Estimators. *Journal of the American Statistical Association*. 76(373): 131-139.
19. Kibria, B.M.G. (2003). Performance of some new ridge regression estimators. *Communications in Statistics – Theory and Methods*, 32, 419-435.
20. Arumairajan, S. and Wijekoon, P. (2014). Improvement of Ridge Estimator when Stochastic Restrictions are Available in the Linear Regression Model. *Journal of Statistical and Econometric Methods,* 3. 35-48.
21. Arumairajan, S. and Wijekoon, P. (2015). Optimal Generalized Biased Estimator in Linear Regression Model. *Open Journal of Statistics*, 5, 403-411.
22. Usman, U., Zakari, Y. and Musa, Y. (2017). A Comparative Study on the Prediction Performance Methods of Handling Multicollineariy Based on Normal and Uniform Distributions. *Journal of Basic and Applied Research International.* 22(3): 111-117.
23. Khalaf, G. (2012). Improved Estimator in the Presence of Multicollinearity. *Journal of Modern Applied Statistical Methods.* 11(1), 152-157.
24. Toka, O. (2016). A Comparative Study on Regression Methods in the presence of Multicollinearity. *Journal of Statisticians: Statistics and Actuarial Sciences IDIA*. 9(2): 47-53.