FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 9 No. 1, January, 2025, pp 240 - 245
DOI: https://doi.org/10.33003/fjs-2025-0901-2981

COMPARATIVE ANALYSIS OF RIDGE AND PRINCIPAL COMPONENT REGRESSION IN ADDRESSING MULTICOLLINEARITY

***[1]Arum, K. C., [1]Ndukwe, S. C., [1]Oranye, H. E. and [2]Sule, O. B.**

[1]Department of Statistics, University of Nigeria, Nsukka.
[2]Department of Mathematics and Statistics, Confluence University of Science and Technology, Osaro, Kogi State, Nigeria.

*Corresponding authors' email: kingsley.arum@unn.edu.ng

ABSTRACT

Multicollinearity arises when two or more regressors are correlated in multiple linear regression model (MLRM) and in most cases, one regressor variable can be predicted from another. Multicollinearity majorly results in inefficient regression model estimates and poor performance of the regression model. However, multicollinearity problem can easily be handled using various methods such as ridge regression, lasso regression, principal components regression, etc. This study compared the effectiveness of two estimators in handling multicollinearity problem in a given dataset. The estimators being compared are ridge estimator (RE) and principal components estimator (PCE). This research uses secondary data obtained from World Bank database, International Monetary Fund (IMF) database, and the Nigerian Debt Management Office to compare the two approaches of handling multicollinearity problem in MLRM. The presence of multicollinearity in the dataset was established using the correlation matrix of predictors and the Variance Inflation Factors (VIF's). Then ridge regression and principal components regression methods were used to fit models to the dataset respectively and their mean squared errors (MSE) were obtained. The MSE was used as performance evaluation measure for the regression models. Both methods addressed the problem multicollinearity in the datasets but the ridge estimator performed better than PCE by having the smallest mean squared error.

Keywords: Linear regression, Multicollinearity, Ridge, Principal component, Mean squared error

INTRODUCTION

Regression analysis is a method for investigating the functional relationship among variables. It is used to determine the relationship between two or more variable so that one can gain information about one of them through knowing values of the other(s).

Arum et al. (2023) defined "regression" as a generic term that refers to a method or an approach that is used to fit model to data in order to account for the relationship that exist between two or more variables. The application of regression analysis is numerous and occur in almost every field, including engineering, physical and social sciences to mention but a few.

Linear regression model (LRM) helps to show the relationship between a response variable (y), and one or more explanatory variable(s). The coefficients of the linear regression model are usually estimated using least squares estimator. The least square estimator according to Gauss-Markov theorem is the best linear unbiased estimator (BLUE).

One of the assumptions of LRM is that the explanatory variable should be linearly independent, however, this rarely happens because most econometric/financial data do correlate together, leading to linear dependency among the regressor, a situation known as multicollinearity.

Multicollinearity occurs when two or more regressors are highly correlated. It affects the performance of regression models because does not allow for stable estimates that are capable of absorbing changes in model re-specification. When there are linear dependencies among the regressors, the problem of multicollinearity occurs (Kibria and Lukman, 2020). There are various sources of multicollinearity in a dataset; some of them are as follows, the methods of data collection employed by the researcher during the process of data gathering, constraints which may have been defined in the model or in the population from which samples are being drawn, poor specification of models and over defined models. Whatever the source of multicollinearity may be, the effects of multicollinearity are very pronounced. Multicollinearity leads to an increase in the variance of the estimated coefficients, and also, increases the odds of obtaining estimates of the coefficient with high values and also exhibits wrong signs. Multicollinearity causes coefficient of estimates to become very sensitive to changes in model specification, this happens when a researcher tries adding or dropping variables in a model (re-specifying the model). Also, due to the presence of multicollinearity, the parameter estimates are changes drastically showing that the results obtained are not robust. The presence of multicollinearity in a model or dataset can be detected through the following:

i. Pairwise correlation of predictor variables: Checking the correlation matrix of the regressors with each other. Significant correlations of 0.5 and above signal a multicollinearity issue (Montgomery et al., 2012).
ii. The Variance inflation factor (VIF): The VIF is the most frequently used indicator for detecting multicollinearity as it shows how the variance of the estimator is inflected by the presence of multicollinearity. VIFs exceeding 10 indicates severe multicollinearity (Montgomery et al., 2012).
iii. Eigenvalues of the predictor variables: The eigenvalues of the matrix of predictor variables can be used in detecting multicollinearity in the data. Small eigenvalues which are close to zero or small in relation to other eigenvalues are indicators of the presence of multicollinearity (Montgomery et al., 2012).
iv. Condition numbers (CN): This is the ratio of the highest to the lowest eigenvalues. When the CN is less than 100, there is no multicollinearity problem. However, CN that lies within 100 and 1000 signifies moderate to strong multicollinearity, and CN exceeds 1000, it signifies severe multicollinearity (Montgomery et al., 2012).

In this study, correlation matrix of the predictors and VIFs was used to detect multicollinearity in the dataset. In handling the problem of multicollinearity, ridge estimator (RE) and principal components estimator (PCE) can be used.

Multicollinearity over time has been a problem in LRM and as a result, the ordinary least squares estimator (OLSE) performs poorly in the estimation of the model parameters. Fortunately, RE and PCE have been developed by researchers to tackle multicollinearity problem in a multiple linear regression model. In this work, compares the two approaches RE and PCE in addressing multicollinearity problems using fininicial datasets via MSE criterion.

The intent of the researcher in this work is to check which approach to handling the problem of multicollinearity performs better between ridge regression and PCR. Hence, when faced with multicollinear problems, a researcher need not become discouraged or desert the work entirely but can go on to apply any of these approaches as is deemed fit in order to overcome the multicollinear problem. Some research works have been done on various approaches to handling multicollinearity, developing new estimators which are more robust in dealing with the problem of multicollinearity, combining different estimators to come up with better ones and comparing the efficiency of different estimators to handle multicollinearity. Some of these research works as well as their findings are presented below.

El-Dereny and Rashwan (2011) introduced different methods of ridge regression to solve multicollinearity problems were introduced by. These methods include ordinary ridge regression (ORR), generalized ridge regression (GRR), directed ridge regression (DRR). The authors discussed properties of ridge regression estimators and methods of selecting biased estimators. Simulated data was used to make comparison between the ridge regression methods and OLS method. From the results of this study, it was discovered that all the methods of ridge regression are better than the OLS method when multicollinearity exists.

Zhang and Ibrahim (2020) conducted a simulation study using SPSS for Ridge and OLS regression procedures for multicollinear data was carried out. The authors discovered that the performance of the evaluated ridge estimator as well as the performance of any ridge-type estimator, depends on the variance of the random error, the correlations among the explanatory variables and the unknown coefficient vector. Their study indicates that, while ridge regression may be effective when multicollinearity is not serious, it is not effective when the explanatory variables are highly correlated.

Lukman *et al*. (2020) introduced new approach to estimating the model parameters using principal components. This approach requires using the PCs as regressors to predict the dependent variable and further utilizing the predicted variable as a dependent variable on the original explanatory variables in an ordinary least square regression. The authors found that the result of the parameter estimates is the same with the principal components regression estimator. Also, the sampling properties of the new estimators were proved to be same as the existing ones. The approach was applied to real-life data and it was concluded that the principal component regression estimator is more efficient than the OLS estimator in tackling multicollinearity in the regression model.

To further address the issue of multicollinearity, Ayinde et al. (2021) reviewed various known methods and estimators that address the multicollinearity problem in a multiple linear regression model (MLRM), such as ridge regression, partial least squares (PLS), principal component regression estimator, and combined estimators' approach. Ayinde et al. (2021) recommended that estimators should be developed for solving multicollinear problems through principal components and partial least squares techniques employing partitioning and extraction of predictor variables, which is to be tested on many linear regression models with varying degrees of multicollinearity using simulated data.

A new estimator called the robust r-k estimator which circumvents the challenge of multicollinearity and outliers in a multiple linear regression model was developed by Arum and Ugwuowo (2022). The authors combined the following estimators, M-estimator, principal component and ridge estimators to form the robust r-k estimator that contains properties of PCE, Ridge estimator and M-estimators. Simulated and real-life data were used and the new estimator performed better than the other estimators compared with it by having the smallest MSE.

Since principal component estimator of Massy (1965) and ridge estimators of Hoerl and Kennard (1970) mitigates the problem of multicollinearity in LRM. It is of interest to determine which of this regression estimator will be more efficient in addressing the problem of multicollinearity in a financial dataset.

Furthermore, this research is significant because regression plays a crucial role in the world of model building and as such, the occurrence of multicollinearity should not be a surprise to a researcher since the researcher can apply any of the methods described in this work in order to deal with it. It cuts across various sectors where collinear data may surface such as in health, economics and finance, agriculture, biological and even in the academia. This study aimed at comparing the methods of ridge regression and principal component regression approaches in addressing multicollinearity problem in a secondary dataset using the mean squared error (MSE) as performance evaluation criterion. This will be achieved through the following objectives; to compare the two methods in other to select the suitable estimator with minimum MSE, that will be used to fit the appropriate model to the dataset. To obtain parameter estimates for the identified regression model in (i) above for prediction.

MATERIALS AND METHODS

Situations may arise where two more regressors can concurrently affect the response variable, thereby leading to a multiple linear regression model (LRM) given in equation (1) below as,

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon} \qquad (1)$$

$\underline{Y}$ presents an $n \times 1$ vector of response variable, $\underline{X}$ is $n \times p$ matrix of regressors, $\underline{\beta}$ is a $p \times 1$ vector of unknown coefficients in the LRM, and $\underline{\varepsilon}$ is an $n \times 1$ vector of random errors such that $E(\underline{\varepsilon}) = 0$ and $Var(\underline{\varepsilon}) = \sigma^2 I_n$, I_n is an $n \times n$ identity matrix. The unknown coefficients $\underline{\beta}$ is usually estimated using least squares estimator (LSE). LSE of $\underline{\beta}$ in (1) is given as,

$$\hat{\beta} = \left(\underline{X}'\underline{X}\right)^{-1}\underline{X}'\underline{Y}, \qquad (2)$$

where $(\underline{X}'\underline{X})$ the matrix of the regressors. The least squares estimator suffer setback when the explanatory variables are affected by multicollinearity. Based on this, ridge and principal component regression estimation methods were developed to address this shortcoming of LSE in the presence of multicollinear explanatory variables.

This research work compares the ridge and principal component regression estimation method in analyzing financial data

Data Collection, Method and Analysis

The datasets used in in this research are secondary data collected from various online sources which are the World

Bank database, International Monetary Fund (IMF) database and the Nigerian debt management office. The data was collected over a period of 51 years i.e., from 1972 – 2022. Since ridge regression and principal components regression require that the data be standardized, the raw data collected online was compiled and standardized (standardization is a process in which the average is removed from each predictor variable and result is divided by its standard deviation (scaling)) to ensure that all predictors are on the same scale to avoid pulling the mean to values with higher magnitude since the mean tends to be affected by bigger values and outliers.

The analysis in this research was carried out using two statistical software which are Statistical Package for Social Sciences (SPSS) and the R programming language. The SPSS is a statistical package owned and managed by IBM corporation used for data management, multivariate analysis, advanced analytics etc. SPSS is used in this work to perform principal component regression (PCR). R is an open-source programming language used by data miners and statisticians for data analysis and to develop statistical software. R is used in this work to perform the Ridge Regression analysis.

Table 1: Descriptive Statistics of the dataset

Variables	N	Range	Min	Max	Mean	Std. Dev.	Variance
Aggregate Investment	51	137753500000	9246500000	147000000000	42640010386.39215000	32360930285.398502	10472298089364221000
Real GDP	51	3.2559	-.9148	2.3411	-.000004	.9999906	1.000
Net export	51	4.1233	-.8383	3.2850	.000002	1.0000028	1.000
Interest rate	51	5.9587	-4.7591	1.1996	-.178890	1.0000027	1.000
Money supply	51	3.6181	-.5757	3.0424	-.000002	.9999951	1.000
Inflation rate	51	4.4827	-.9521	3.5306	.000004	.9999978	1.000
External reserve	51	3.0367	-.8711	2.1656	-.000002	.9999930	1.000
Exchange rate	51	3.5710	-.7701	2.8009	.000006	1.0000058	1.000
Debt	51	3.8439	-1.5978	2.2461	.000000	.9999935	1.000

The response variable (y) is aggregate investment while the remaining eight variables are the predictor variables. All predictor variables were standardized to be on the same scale. The response variable may or may not be standardized and it does not affect the regression results. Hence, the response variable is not standardized.

Detecting multicollinearity in the data

As stated earlier in this research, the correlation matrix of predictors and the VIF would be used for the detection of multicollinearity in the datasets.

Table 2: Correlation matrix of predictor variables

	Real_Gdp	Net_exp	Int_rate	Money_sup	Inf_rate	Extl_Resv	Exch_rate	Debt
Real_Gdp	1	0.8635	0.3548	0.8838	-0.3077	0.8872	0.8215	0.3725
Net_exp	0.8635	1	0.3451	0.6332	-0.2869	0.8761	0.6199	0.2183
Int_rate	0.3548	0.3451	1	0.3448	-0.4494	0.4021	0.3844	0.3074
Money_sup	0.8838	0.6332	0.3448	1	-0.2259	0.7503	0.9317	0.4957
Inf_rate	-0.3077	-0.2869	-0.4494	-0.2259	1	-0.3361	-0.2472	0.151
Extl_Resv	0.8872	0.8761	0.4021	0.7503	-0.3361	1	0.7767	0.2731
Exch_rate	0.8215	0.6199	0.3844	0.9317	-0.2472	0.7767	1	0.6241
Debt	0.3725	0.2183	0.3074	0.4957	0.151	0.2731	0.6241	1

From the result of the correlation matrix in table two above, it can be seen that Real GDP has a high correlation of 0.8635 with net export, 0.8838 with money supply, 0.8872 with external reserve, and 0.8215 with exchange rate. Net export has a high correlation of 0.8761with external reserve, 0.6332 with money supply and 0.6199 with exchange rate. Money supply has a high correlation of 0.9317 with exchange rate, 0.7503 with external reserve. Exchange rate has a high correlation of 0.7767 with external reserve. These high correlations are indicators of multicollinearity.

Table 3: Variance inflation factors of predictor variables

Predictor Variables	VIF
Real GDP	19.818
Net export	7.355
Interest rate	1.719
Money supply	20.743
Inflation rate	1.856
External reserve	10.025
Exchange rate	19.586
Debt	3.339

From the table 3 above, real GDP, money supply, external reserve and exchange rate all indicate multicollinearity given that their VIF's are greater than 10.

The VIF values that ranges between 5 to 10 indicate mild multicollinearity while VIF values greater than 10 indicates the presence of severe multicollinearity see Lukman et al. (2020) and Jegede et al. (2022). Also, Net export with a VIF of 7.355 indicates a somewhat likely multicollinearity problem as it lies in the range of 5-10. Interest rate, Inflation rate are not affected by multicollinearity since their VIF's lie between 1- 4.

Hence, the results from the correlation matrix are in tandem with the results of the VIF table, it shows that variables like real GDP, money supply, external reserve, exchange rate and net export are affected by multicollinearity. In other to handle this issue of multicollinearity present in the data, the two methods that is considered in this work are ridge regression and principal components regression. First, we apply ridge regression on the dataset after which PCR will be applied on the dataset as well and the estimator with the least MSE among RE and PCE will be used to fit our ideal regression model to the dataset.

Ridge Regression

Ridge regression is a procedure for obtaining biased estimator of regression model originally proposed by Hoerl and Kennard (1970). The ridge regression introduces a small bias, k, so that the variance can be substantially reduced, which leads to a smaller MSE. Ridge regression is applied to centered and scaled predictor variables and the response may or may not be scaled but must be centered. The ridge estimator penalizes the estimates for being too large but in so doing introduces a bias in the model. The ridge estimator is given as

$$\hat{\beta}_R = \left(\underline{X}'\underline{X} + kI\right)^{-1}\underline{X}'\underline{Y} \qquad (3)$$

where, $\hat{\beta}_R$ is a vector of the ridge regression estimates obtained after performing ridge regression on the set of regressors. Where X is the matrix of explanatory variables of the regression model, and k is the shrinkage parameter and I is a (p x p) identity matrix. When k = 0, ridge estimator collapses to the least squares estimator. The ridge estimator $\hat{\beta}_R$ contains a shrinkage parameter k which controls the amount of regularization.

According to Arum and Ugwuowo (2022), shrinkage is the reduction in the effects of sampling variation and a shrinkage estimator is an estimator that either implicitly or explicitly incorporates the 'shrinkage' effect. A shrinkage parameter can be selected either using a ridge trace which shows the ridge coefficients as a function of k. The shrinkage (biasing) parameter k, is chosen as the smallest value of k (that introduces the smallest bias) after which the regression coefficients when the model has stabilized, Hoerl and Kennard (1970).

Another method which has become widely accepted as the method of choosing k is the cross validation (CV). The CV technique involves randomly dividing the set of observations into different groups or folds, of equal sizes. The first group is the validation set, and the method is fitted on the remaining n -1 folds. Then the MSE is computed on the observations in the held-out fold. The procedure is repeated n times and each time, a different group of observations is treated as a validation set. Cross validation is done using computer programs/software like R programming language. In this study, the n-fold cross validation will be used to select the shrinkage parameter.

Principal components regression (PCR)

Principal components (PC) regression is a dimension reduction technique that involves projecting each data point onto only the first few principal components to obtain lower dimensions of the data while preserving as much of the data's variation as possible Massy (1965). In PC regression, instead of regressing the response variables on the regressors directly, the principal components of the regressors (explanatory variables) are used. A researcher normally uses only a subset of all the PC for the regression, making PC regression a kind of regularized method and also a type of shrinkage estimator which handle problem of multicollinearity in the dataset. The PC regression estimator introduced by Massy (1965) is given by,

$$\hat{\beta}_{pcr} = \left(\underline{Z}^T\underline{Z}\right)^{-1}\underline{Z}^T\underline{Y} \qquad (4)$$

where $\hat{\beta}_{pcr}$ is the vector of the PC regression estimates obtained after performing PC regression on the set of standardized variables, Z is the matrix of standardized variables of the regression model, Y is the vector of values of the response variable.

It is desired in PCR that the proportion of variation explained by regressors is maximum whilst the PC associated with the smallest eigenvalues are deleted. The principal components to regress on are selected either by examining the scree plot or using the Kaiser-Guttmann criterion. The scree plot is a graph of the eigenvalues against the corresponding PC number. While examining the scree plot, we find a point at which the proportion of variance explained by each subsequent principal component drops off (where lesser variance is explained). Then we can select those PC's that explain more variation than the rest. The scree plot can be obtained using statistical software like SPSS, R etc. According to the Kaiser rule, the only principal components to be retained are those whose eigenvalues are approximately 1 and above. The idea behind the Kaiser rule is that any principal components with eigenvalues less than 1 has lesser information than one of the original variables and should not be retained. In this study, the Kaiser rule will be used in selecting the principal components to be regressed. Hence, in other to determine which principal components to retain and which ones to drop, the amount of variation which each component explains will be computed using their eigenvalues and those that are approximately one and above are selected according to the Kaiser rule.

Table 4: PCR Result of Explained Variations for the Independent Variables

Variables	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.776	59.694	59.694	4.776	59.694	59.694
2	1.294	16.178	75.872	1.294	16.178	75.872
3	.955	11.940	87.811	.955	11.940	87.811
4	.496	6.203	94.014			
5	.271	3.384	97.398			
6	.131	1.633	99.032			
7	.058	.727	99.759			
8	.019	.241	100.000			

From Table 4, three out of eight variables (components) contributed to the regression model
using PC regression approach. These three components accounted for 59.694%, 16.178%, 11.940% of the total variations respectively. Cumulatively, these components accounted for
approximately 88% of the variations in the independent variables. The remaining components
accounted variation of 12%.

Component Matrix

The elements of the component matrix represent the various correlations that exist between each of the original variables and the extracted components. Various rotation methods can be used to obtain the component matrix but the most common is the varimax rotation will be used. Varimax represents variable maximization and it redistribute factor loadings in a way that each variable measures strictly one factor (component); it will help in understanding our factors and the variables that measure them. Also, correlations of 0.3 and below were removed as they are considered insignificant (not related). The rotated component matrix for our PC regression is given in table 5 below.

Table 5: Rotated Component Matrix

	Component		
	1	**2**	**3**
Real GDP	.940		
Net export	.897		
Interest rate		.374	.824
Money supply	.813	.431	
Inflation rate			-.829
External reserve	.918		
Exchange rate	.764	.538	
Debt		.932	

Analyzing the component matrix in table 5 above, component one is measured by real GDP, net export, money supply, external reserve and exchange rate and will be termed as growth indices. Component two is measured by interest rate, money supply, exchange rate, debt and will be termed as Monetary policies. Component three is measured by interest rate and inflation rate and will be termed as rates.

Using the 3 components extracted above, PC regression is performed and the estimate is obtained. The MSE of the PC estimator is obtained as 0.571.

Having obtained the mean squared errors for both the ridge and PC estimators, it is easily seen that the ridge estimator performed better in handling problem of multicollinearity in the dataset by having the smallest mean squared error of 0.3777 value. It should however be noted that both estimators performed well in addressing multicollinearity in the dataset but the ridge regressor outperformed the principal components regressor with respect to the dataset used.

Thus, the ridge estimator will be used to fit an appropriate model to the dataset which will mitigate the effect of multicollinearity present in the data. The coefficients of the model fit by the ridge estimator is given in table 6 below.

Table 6: Ridge Regression Estimates for the Variables

S/no	Variables	Estimate
1	Intercept	-0.0748
2	Real GDP	0.7084
3	Net export	-0.1454
4	Interest rate	-0.4180
5	Money supply	0.1234
6	Inflation rate	-0.2569
7	External Reserve	0.0330
8	Exchange Rate	-0.0361
9	Debt	0.1031

The appropriate regression model is the ridge regression model given as

$$Y_i = -0.0747 + 0.7080X_1 - 0.1450X_2 - 0.4180X_3 + 0.1237X_4 - 0.2569X_5 + 0.0328X_6 - 0.0362 + 0.1030X_8$$

CONCLUSION

Multicollinearity is always a challenge to most researchers when dealing with econometric and financial data. When it is present in the data, the ordinary least squares estimators are affected because the matrix of predictors (X^TX matrix) becomes ill-conditioned and suffers break down. The effects of multicollinearity, its indicators, types and possible solutions have been addressed in this research work. In this study, secondary data was collected online from World Bank database, International Monetary Fund (IMF) database and the Nigerian Debt Management Office. The data was compiled into one dataset and used for the analysis. The analysis was carried out using ridge regression and principal components regression methods which have been shown to address the issue of multicollinearity from previous research works. The major finding of this work is that the ridge estimator which was applied to the multicollinear data performed better than the principal component estimator (PCE) by having the smallest mean squared error (MSE) for the financial data used in this study. Though, both methods that was used yielded the good results of handling multicollinearity in the data, but ridge estimator performed better than PCE from our judging criterion which is the MSE.

REFERENCES

Arum, K. C., & Ugwuowo, F. I. (2022). Combining principal component and robust ridge estimators in linear regression model with multicollinearity. Concurrency and Computation Practice and Experience, 34 (10), 12 – 20. https://doi.org/10.1002/cpe.6803

Ayinde, K., Ugochinyere N., & Olusegun O. (2021). Solving multicollinearity problem in linear regression model: The review suggests new idea of partitioning and extraction of the explanatory variables. *Journal of Mathematics and Statistics Studies*, 2 (1), 12 – 20.

El-Dereny, M., & Rashwan, N. I. (2011): Solving multicollinearity problem using ridge regression models. *International Journal of Contemporary Mathematical Sciences*, 6(12), 585-600.

Hoerl A. E., & Kennard R.W (1970). Ridge Regression: Applications to Non-Orthogonal Problems. Journal of *Technometrics*, 12 (1), 69-82.

Jegede, S. L., Lukman, A. F., Ayinde, K., & Odeniyi K. A. (2022). Jackknife Kibria-Lukman M-Estimator: Simulation and Application.

Kibria, G. B. M., & Lukman, A. F. (2020). A new ridge-type estimator for the linear regression model: Simulations and applications. *Scientifica*, 2020, 1–16. https://doi.org/10.1155/2020/9758378

Lukman, A.F., Ayinde K., Olusegun O., & Hamidu A. (2020). A new approach of principal component regression with applications to collinear data. *International Journal of Engineering Research and Technology*, 13 (7), 1616 – 1622.

Massy, W.F. (1965). Principal Components Regression in exploratory statistical research. *Journal of the American. Statistical. Association*. 60, 234-256.

Montgomery, D. C. Peck E. A., & Vining, G.G. (2012). Introduction to linear regression analysis. (3rd edition). Wiley and Sons Inc. New York.

Zhang, J., & Ibrahim M. (2020). A simulation study on SPSS ridge regression and ordinary least squares regression procedures for multicollinearity data. *Journal of Applied Statistics*, 32 (6), 571 – 588.