International Journal of Mathematics and Statistics Studies
Vol.11, No.1, pp.30-39, 2023
Print ISSN: 2053-2229 (Print),
Online ISSN: 2053-2210 (Online)
Website: https://www.eajournals.org/
Publication of the European Centre for Research Training and Development -UK

# The Application of the Least Squares Method to Multicollinear Data

**Amin Otoni Harefa[1], Yulisman Zega[2], Ratna Natalia Mendrofa[3]**
[1,2,3] Department of Mathematics Education, Universitas_Nias, North Sumatera, Indonesia

doi: https://doi.org/10.37745/ijmss.13/vol11n13039 Published May 31, 2023

**Citation**: Harefa A.O., Zega Y., Mendrofa R.N. (2023) The Application of the Least Squares Method to Multicollinear Data, *International Journal of Mathematics and Statistics Studies*, Vol.11, No.1, pp.30-39

**ABSTRACT:** *Regression analysis is an analysis that aims to determine whether there is a statistically dependent relationship between two variables, namely the predictor variable and the response variable. One of the methods for estimating multiple linear regression parameters is the Least Squares Method. Therefore, careful and meticulous analysis and selection of appropriate techniques are required to overcome the multicollinearity problem and ensure accurate and meaningful regression analysis results. Descriptive statistical table of response variables and predictor variables, where the average results are rounded. The regression equation using the OLS method is as follows:* $\hat{Y} = 2{,}037 + 0.302X_1 + 0.206X_2 + 0.172X_3 + 0.342X_4$*. Therefore, it is important to use special techniques such as regularization or PCA to overcome the multicollinearity problem in the data before applying the least squares method. Thus, we can obtain more stable and accurate regression coefficient estimates and a more reliable linear regression model.*

**KEYWORDS:** application, multiple linear regression, analysis, multicollinearity, OLS method

## INTRODUCTION

Multiple linear regression analysis is an analysis with multiple independent variables (Janah & Kartini, 2022). Regression analysis is an analysis that aims to determine whether there is a statistically dependent relationship between two variables, namely the predictor variable and the response variable (Hermawan, 2021). In addition, another purpose of multiple regression analysis is to predict the effect or functional relationship between two or more predictor variables on one response variable (Sartika et al., 2020). Regression analysis is an equation expressed in mathematical form between the independent variable X and the response variable Y (Ningsih et al., 2019). A problem that often arises in multiple regressions is multicollinearity. Multicollinearity problems occur when predictor variables and other predictor variables are correlated with each other (Sartika et al., 2020).

In the research data, sometimes multicollinear data is accidentally found where the independent variables are related to each other, so it becomes a serious problem in analyzing research data (Senaviratna & Cooray, 2019). If there is multicollinearity among the independent variables,

the estimated regression model parameters using the least squares method will produce an unbiased estimator with a large variance. Multicollinearity detection can be done informally, one of which is by the linear correlation coefficient between independent variables, or in a formal way with the variance inflation factor (Bayman & Dexter, 2021).

In a linear regression analysis, "multicollinearity" refers to the presence of a significant correlation between two or more independent variables. This can make it difficult to interpret the predicted regression coefficients effectively when employing the least squares method since they become unstable (Lavery et al., 2019). The presence of multicollinearity cases causes the standard error (SE) of the regression parameters to increase, so an alternative method is needed for estimating multiple linear regression parameters (Auqino et al., 2019). According to (Adeboye et al., 2014) multicollinearity affects the standard error of the regression coefficient estimate, so the estimation results may not be accurate. The effect of multicollinearity obscures the interpretation of the structure of the model equation. To overcome this, eliminate correlated independent variables from the model (Nyrhinen & Leskinen, 2014). In addition, multicollinearity can also cause differences in conclusions between the F statistical test and the t statistical test.

The least squares method is a method in statistics used to find the best line or curve that can represent data. The goal is to find the mathematical model that best fits the data (Ningsih et al., 2019). The least squares method can be applied to various types of mathematical models, such as linear, polynomial, and exponential models (Klau, 2019). This method is also very useful in processing data that contains noise or uncertainty because it can reduce the effect of data variability (Janah & Kartini, 2022). The least squares method is used to determine the coefficient values in the straight line equation by minimizing the sum of the squared differences between the observed values and the values predicted by the mathematical model (Sartika et al., 2020). In multiple linear regression, the least squares method is used to determine the coefficient value of each independent variable (Razali et al., 2021).

In the multiple linear regression analysis method, there is a classic assumption test, namely the heteroscedasticity test, which aims to test whether the inequality of the residual variance from one observation to another in the regression model is fixed (Janah & Kartini, 2022). In the least squares method, the main objective is to find the regression coefficient value that best estimates the value of the dependent variable from the independent variables. However, when there is multicollinearity between the independent variables, the least squares method cannot produce accurate coefficient values (Obite et al., 2020). To overcome the multicollinearity problem, several techniques can be used, such as removing independent variables that are strongly correlated, transforming the data, or using other methods such as ridge regression or lasso regression. In this way, the results of linear regression analysis can be more accurate and meaningful in explaining the relationship between the dependent and independent variables (Weaving et al., 2019).

Applying the least squares method to multicollinearity data can cause problems in estimating accurate regression coefficients. Therefore, several techniques can be used to overcome the multicollinearity problem in the least squares method, among others (Hair et al., 2014):

1. Using the ridge regression technique: This technique involves adding a "shrinkage parameter" to the regression equation to reduce the variance of the regression coefficient estimates. This technique is suitable for data with many independent variables.
2. Strongly correlated independent variables can be eliminated from the regression model, either one or all of them if there are two or more independent variables that exhibit this correlation
3. Performing data transformation: This technique involves transforming the data used in the regression model, such as logarithms or square roots, so as to minimize the correlation between independent variables
4. Using the lasso regression method: This method overcomes multicollinearity by selecting the independent variables that contribute most to the regression model and ignoring those that are not significant

In applying the least squares method to multicollinearity data, it is important to choose the technique that is most appropriate and relevant to the data used. In addition, careful and thorough analysis is also required to avoid misinterpretation of the regression coefficient estimates (Rencher, 2005). However, it is important to understand that the least squares method cannot completely eliminate multicollinearity problems. Therefore, careful and meticulous analysis and selection of appropriate techniques are required to overcome the multicollinearity problem and ensure accurate and meaningful regression analysis results.

## METHODS

X and the response variable Y, there is a possibility of a different linear relationship for each interval X. This type of research uses a quantitative approach and a literature review. The sampling technique is purposive sampling, with the entire population being sampled according to research needs. The data were analyzed using multiple linear regression solved by the least squares method. Before determining the multiple linear regression equation, the classical assumption test, especially data multicollinearity, will be carried out first.

### Least Squares Method

One of the methods for estimating multiple linear regression parameters is the Least Squares Method (Auqino et al., 2019). In general, the multiple linear regression model is:

$$Y = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_j X_{ij} + e \text{ for i = 1, 2, …, j}$$

Where:

$Y$ : observation of the response variable
$X_{i1}, X_{i2}, X_{ij}$ : observations of the 1st, 2nd,... predictor variable
$\beta_0$ : constant

$\beta_1, \beta_2, \ldots, \beta_j$ : regression parameters
$e$ : error for the observation

Ordinary least square is used to estimate the regression coefficients by minimizing the sum of squared errors. The ordinary least square estimator of β is obtained using:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_j X_{ij} + e_i$$

which can be written briefly in matrix notation as follows:

$$Y = X\hat{\beta} + e \quad (1)$$

Where:
Y : vektor n × 1 pengamatan pada variabel respon
X : an n × k matrix giving n observations over k predictor variables
$\hat{\beta}$ : a k-element column vector of OLS estimators of regression coefficients
$e$ : an N × 1 column vector of N residuals
from Equation 1, obtained

$$e = Y - X\hat{\beta}$$

Therefore

$$e^T e = (Y - X\hat{\beta})^T (Y - X\hat{\beta})$$
$$e^T e = Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\beta \quad (2)$$

from Equation 2 can be derived $e^T e$ against $\hat{\beta}$, namely:

$$\frac{\partial e^T e}{\partial \hat{\beta}} = \frac{\partial (Y^T Y - 2\hat{\beta}^T X^T Y + \hat{\beta}^T X^T X\beta)}{\partial \hat{\beta}}$$

Thus obtained

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

**Multicollinearity**

Multicollinearity was first introduced by Ragnar Frisch in 1934. It is the presence of a definite or exact linear relationship between some or all of the predictor variables in a multiple linear regression model (Auqino et al., 2019).

The presence of multicollinearity will have the following consequences for the results:

a. Although BLUE, the least squares parameter estimator, has a large variance and covariance, the estimate is less accurate.
b. The confidence interval of the estimate of the regression parameter will be wide due to the effects of the following (Bayman & Dexter, 2021).
c. The t-test statistic of one or more regression coefficients will tend to become insignificant due to the effects of (Lavery et al., 2019).
d. Although the t-test statistic tends to be insignificant, the modeled value can be high.

In multiple regression, tolerance is used as an indicator of multicollinearity. Tolerance is estimated by 1-$R^2$, where $R^2$ is calculated by regressing the independent variable of interest on

the other independent variables included in the multiple regression analysis. Researchers desire higher tolerance levels, as low levels are known to affect the results associated with multiple regression analyses. The tolerance level is the value of 1-$R^2$ when each independent variable is regressed against the other independent variables (Adeboye et al., 2014).

Multicollinearity can be detected by the variance inflation factor (VIF). The VIF can be calculated using the following (Lavery et al., 2019):

$$VIF_j = \frac{1}{1 - R_k^2}$$

where $R_k^2$ is the coefficient of determination of variable *Xj* as the response variable against (*k-1*) other predictor variables. If the VIF value is > 10, then there is a case of multicollinearity between the predictor variables (Weaving et al., 2019). In multiple regression, VIF is used as an indicator of multicollinearity. Researchers desire lower VIF levels, as higher VIF levels are known to affect the results associated with multiple regression analysis. In fact, the utility of VIF, in contrast to tolerance, is that VIF specifically indicates the magnitude of inflation in standard errors associated with a given beta weight caused by multicollinearity. A VIF of more than 10 begins to indicate a relatively high level of multicollinearity (Adeboye et al., 2014).

## RESULTS AND DISCUSSION

The data used is secondary data, namely data obtained from the Academic Information System of Nias University. The initial work done is to calculate an overview of the research data summarized in descriptive statistics. The following is an overview of the research data summarized in descriptive statistics in Table 1, as follows:

Table 1 Descriptive Statistics of Response Variables and Predictor Variables

| Variable | Minimum | Maximum | Mean | Standard deviation | Variance |
|---|---|---|---|---|---|
| Y | 30 | 120 | 106.64 | 16.645 | 277.043 |
| $X_1$ | 9 | 36 | 32.13 | 5.023 | 25.227 |
| $X_2$ | 6 | 24 | 21.30 | 3.423 | 11.714 |
| $X_3$ | 5 | 20 | 17.73 | 2.867 | 8.221 |
| $X_4$ | 10 | 40 | 35.48 | 5.687 | 32.338 |

Table 1 is a descriptive statistical table of response variables and predictor variables, where the average results are rounded. Next is to obtain regression equations using the OLS method and determine which OLS parameters are significant to the response variable. The regression equation using the OLS method is as follows:

$$\hat{Y} = 2{,}037 + 0.302X_1 + 0.206X_2 + 0.172X_3 + 0.342X_4$$

After obtaining the multiple regression equation, the comparison between the correlation in each variable is described in table 2 below:

Table 2 Correlations

| | | Y | X1 | X2 | X3 | X4 |
|---|---|---|---|---|---|---|
| Pearson Correlation | Y | 1.000 | .977 | .977 | .974 | .985 |
| | X1 | .977 | 1.000 | .940 | .932 | .941 |
| | X2 | .977 | .940 | 1.000 | .949 | .947 |
| | X3 | .974 | .932 | .949 | 1.000 | .954 |
| | X4 | .985 | .941 | .947 | .954 | 1.000 |

Table 2 above shows that the level of correlation between variables is very high, exceeding the previously set requirement of 0.7, so the data is tested for multicollinearity. After the correlation test, the next step is to detect multicollinearity. Multicollinearity detection is done by looking at the VIF value. The multicollinearity test results are presented in Table 3.

Table 3 Multicollinearity Test Results

| Variable | VIF |
|---|---|
| $X_1$ | 11.275 |
| $X_2$ | 14.394 |
| $X_3$ | 14.650 |
| $X_4$ | 15.551 |

Based on Table 3, variables X1 to X4 have VIF values greater than 10 so it can be concluded that the data used has multicollinearity problems. If the data has multicollinearity after the least squares method, the resulting regression coefficients will be unstable and difficult to interpret. This can reduce the quality and accuracy of the constructed linear regression model. Therefore, it is important to use special techniques such as regularization or PCA to overcome the multicollinearity problem in the data before applying the least squares method. Thus, we can obtain more stable and accurate regression coefficient estimates and a more reliable linear regression model.

Table 4 KMO and Bartlett's Test

| | | |
|---|---|---|
| Kaiser-Meyer-Olkin Measure of Sampling Adequacy. | | .886 |
| Bartlett's Test of Sphericity | Approx. Chi-Square | 193909.542 |
| | df | 6 |
| | Sig. | .000 |

Based on Table 4, the KMO value = 0.886 so with this value, factor analysis can be continued because the value of KMO > 0.5. can be continued because the value of KMO is > 0.5. Meanwhile, the Bartlet Test is used to test whether it is true that the variables involved are correlated.

Hypothesis:
H0 = there is no correlation between independent variables.
Ha = there is a correlation between independent variables.

test criteria by looking at the p-value (significance). Accept if Sig. > 0,05. The KMO and Bartlett's Test table shows that the chi-square value = 193909.542 with 6 degrees of freedom, and a p-value (0.000) <0.05, so it is rejected. This means that there is a correlation between the independent variables.

Table 5 Anti-image Matrices

| | | X1 | X2 | X3 | X4 |
|---|---|---|---|---|---|
| Anti-image Covariance | X1 | .089 | -.028 | -.012 | -.026 |
| | X2 | -.028 | .069 | -.027 | -.018 |
| | X3 | -.012 | -.027 | .068 | -.030 |
| | X4 | -.026 | -.018 | -.030 | .064 |
| Anti-image Correlation | X1 | .910[a] | -.351 | -.152 | -.340 |
| | X2 | -.351 | .885[a] | -.388 | -.271 |
| | X3 | -.152 | -.388 | .877[a] | -.450 |
| | X4 | -.340 | -.271 | -.450 | .873[a] |

a. Measures of Sampling Adequacy(MSA)

Based on the MSA number criteria, the anti-image matrix table shows that all MSA numbers have values above 0.5. MSA numbers have values above 0.5. This means that the analysis can continue.

Table 6 Communalities

| | Initial | Extraction |
|---|---|---|
| X1 | 1.000 | .949 |
| X2 | 1.000 | .961 |
| X3 | 1.000 | .959 |
| X4 | 1.000 | .963 |

Extraction Method: Principal Component Analysis.

The Communalities table shows that for variable $X_1$, a value of 0.949 = 94.9% is obtained. This means that 94.9% of the $X_1$ variable can be explained by the factors formed. Likewise for variables $X_2$, $X_3$ and $X_4$.

Table 7 Total Variance Explained

| | Initial Eigenvalues | | | Extraction Sums of Squared Loadings | | |
|---|---|---|---|---|---|---|
| Component | Total | % of Variance | Cumulative % | Total | % of Variance | Cumulative % |
| 1 | 3.832 | 95.798 | 95.798 | 3.832 | 95.798 | 95.798 |
| 2 | .071 | 1.769 | 97.568 | | | |
| 3 | .053 | 1.331 | 98.898 | | | |
| 4 | .044 | 1.102 | 100.000 | | | |

Extraction Method: Principal Component Analysis.

Total Variance Explained shows that eigenvalues below 1 cannot be used in calculating the number of factors formed, so the factoring process stops at just one factor. Factor one has eigenvalues of 3.832, meaning that with one factor, it can explain 3.832, or 95.79%, of the total variance explained. This factor can explain 3.832 or 95.79% of the total diversity of the original variable.

Table 8 Component Matrix

| | Component |
|---|---|
| | 1 |
| X1 | .974 |
| X2 | .980 |
| X3 | .980 |
| X4 | .981 |

The component matrix table shows that only one factor is formed from the three variables. This means that one factor is the most optimal number to reduce the three independent variables. By using the Component Score Coefficient Matrix table, the equation for the new factor formed is as follows:

$$F_1 = 0.974X_1 + 0.980X_2 + 0.980X_3 + 0.981X_4$$

The resulting factor scores can be used to replace the scores on the original independent variables. Once the multicollinearity-free components of the PCA results are obtained, they are regressed or analyzed for their effect on the independent variables using linear regression analysis.

After obtaining a new independent variable (F1) that is free of multicollinearity through the PCA technique, the next step is to regress the new independent variable (F1) on the independent variable (Y). Because the new independent variable (F1) formed is only one, then the model is used simple linear regression analysis as follows:

$$Y = \beta_0 + \beta_1 F_1 + e_i$$

which is

$$F_1 = 0.974X_1 + 0.980X_2 + 0.980X_3 + 0.981X_4$$

Based on the Coefficients table 9, the regression model is obtained as follows:

$$Y = 106.639 + 16.636F_1$$

Table 9 Coefficients

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | | Collinearity Statistics | |
|---|---|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. | Tolerance | VIF |
| 1 | (Constant) | 106.639 | .003 | | 32271.702 | .000 | | |
| | REGR factor score 1 for analysis 1 | 16.636 | .003 | .999 | 5034.352 | .000 | 1.000 | 1.000 |

a. Dependent Variable: Y

## CONCLUSION

Multicollinearity can cause problems in estimating accurate regression coefficients. The least squares method is used to determine the coefficient values in a straight line equation by minimizing the sum of the squared differences between the observed values and the values estimated by the mathematical model. However, when data is subject to multicollinearity,

several techniques can be used to overcome this problem in the least squares method, including using PCA or ridge regression. PCA can be used to overcome multicollinearity problems in data analysis. In the case of multicollinearity, highly correlated variables can be identified and combined to form new factors that are more independent of each other. The new factors can then be used as input variables in regression analysis or least squares methods, resulting in more accurate regression coefficient estimates. Using PCA to overcome multicollinearity will result in a simpler model that is easier to interpret. However, keep in mind that the interpretation of the new factors should be based on knowledge of the original variables that make up the factor.

## REFERENCES

Adeboye, N. ., Fagoyinbo, I. S., & Olatayo, T. . (2014). Estimation of the Effect of Multicollinearity on the Standard Error for Regression Coefficients. *IOSR Journal of Mathematics*, *10*(4), 16–20. https://doi.org/10.9790/5728-10411620

Auqino, S., Maiyastri, M., & Diana, R. (2019). Perbandingan Metode Kuadrat Terkecil Dan Metode Bayes Pada Model Regresi Linier Berganda Yang Mengandung Multikolinieiritas. *Jurnal Matematika UNAND*, *8*(1), 307. https://doi.org/10.25077/jmu.8.1.307-312.2019

Bayman, E. O., & Dexter, F. (2021). Multicollinearity in Logistic Regression Models. *Anesthesia and Analgesia*, *133*(2), 362–365. https://doi.org/10.1213/ANE.0000000000005593

Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). Multivariate Data Analysis. *British Library Cataloguing-in-Publication Data*, 201–225.

Hermawan, T. (2021). Estimasi Kurva Regresi Spline Pada Data Longitudinal Dengan Metode Kuadrat Terkecil. *Intersections*, *5*(2), 17–25. https://doi.org/10.47200/intersections.v5i2.583

Janah, M., & Kartini, A. Y. (2022). Penerapan Metode Regresi Linier Berganda Pada Kasus Balita Gizi Buruk Di Kabupaten Bojonegoro. *Jurnal Statistika Dan Komputasi*, *1*(2), 74–82. https://doi.org/10.32665/statkom.v1i2.1170

Klau, K. Y. (2019). Penggunaan Regresi Linear Multipel Dan Metode Kuadrat Terkecil Untuk Menganalisisfaktor-Faktor Yang Mempengaruhi Hasil Produksi Jagung Di Kabupaten Belu. *RANGE: Jurnal Pendidikan Matematika*, *1*(1), 21–32. https://doi.org/10.32938/jpm.v1i1.185

Lavery, M. R., Acharya, P., Sivo, S. A., & Xu, L. (2019). Number of predictors and multicollinearity: What are their effects on error and bias in regression? *Communications in Statistics: Simulation and Computation*, *48*(1), 27–38. https://doi.org/10.1080/03610918.2017.1371750

Ningsih, T., Herrhyanto, N., & Rachmatin, D. (2019). Analisis Regresi Linear Piecewise Dua Segmen dengan Menggunakan Metode Kuadrat Terkecil. *Jurnal Eureka Matika*, *7*(2), 63–82.

Nyrhinen, J. N., & Leskinen, E. (2014). Multicollinearity in marketing models: Notes on the application of ridge trace estimation in structural equation modelling. *Electronic Journal*

*of Business Research Methods*, *12*(1), 3–15.

Obite, C. P., Olewuezi, N. P., Ugwuanyim, G. U., & Bartholomew, D. C. (2020). Multicollinearity Effect in Regression Analysis: A Feed Forward Artificial Neural Network Approach. *Asian Journal of Probability and Statistics*, *6*(1), 22–33. https://doi.org/10.9734/ajpas/2020/v6i130151

Razali, M., Elazhari, & Tampubolon, K. (2021). Curve Matching Using Least Square Method and Gaussian Method. *AFoSJ-LAS*, *1*(4), 170–183. http://j-las.lemkomindo.org/index.php/AFOSJ-LAS

Rencher, A. C. (2005). A Review Of "Methods of Multivariate Analysis, Second Edition." *IIE Transactions*, *37*(11), 1083–1085. https://doi.org/10.1080/07408170500232784

Sartika, I., Debataraja, N. N., & Imro'ah, N. (2020). Analisis Regresi Dengan Metode Least Absolute Shrinkage and Selection Operator (Lasso) Dalam Mengatasi Multikolinearitas. *Bimaster : Buletin Ilmiah Matematika, Statistika Dan Terapannya*, *9*(1), 31–38. https://doi.org/10.26418/bbimst.v9i1.38029

Senaviratna, N. A. M. R., & Cooray, T. M. J. A. (2019). Diagnosing Multicollinearity of Logistic Regression Model. *Asian Journal of Probability and Statistics*, *5*(2), 1–9. https://doi.org/10.9734/ajpas/2019/v5i230132

Weaving, D., Jones, B., Ireton, M., Whitehead, S., Till, K., & Beggs, C. B. (2019). Overcoming the problem of multicollinearity in sports performance data: A novel application of partial least squares correlation analysis. *PLoS ONE*, *14*(2), 1–16. https://doi.org/10.1371/journal.pone.0211776