Partial Least Square (PLS) Method of Addressing Multicollinearity Problems in Multiple Linear Regressions

(Case Studies: Cost of electricity bills and factors affecting it)

D W Wondola*, S N Aulele, F K Lembang
Department of Mathematics and Natural Science Faculty Pattimura University

E-mail: dwiwondola09@gmail.com

Abstract. Multiple regression analysis is a statistical analysis used to predict the effect of several independent variables on the dependent variable. The problem that often occurs in multiple linear regression models is multicollinearity which is a condition of a strong relationship between independent variables. To overcome the problem of multicollinearity, the Partial Least Square method is used. This method reduces independent variables that have no significant effect on the dependent variable, then new variables with smaller dimensions are formed which are linear combinations of the independent variables, therefore the partial significance test (t test) becomes an important part in the formation of PLS components. Furthermore, using the PLS method, we obtain: \( \hat{Y} = 126.220 + 12.034 \text{ (Income)} + 12.437 \text{ (Number of Family Members)} + 12.959 \text{ (House Area)} + 11.919 \text{ (Number of Rooms)} + 12.274 \text{ (Number of Electronic Devices)} \)

1. Introduction

Multiple regression analysis is the development of simple regression analysis in which there are more than one independent variable that affects one dependent variable. To predict the effect of independent variables on the dependent variable, a test is carried out by entering simultaneously various independent variables. In the multiple linear regression equation, produced constants and regression coefficients for each independent variable. In multiple linear regressions known a method used to estimate the regression parameters, namely the Least Squares Method. The purpose of the least squares method is to produce a good estimator by minimizing the number of squares of errors (error). In the least squares' method, there are several assumptions that must be met. Furthermore, based on the Gauss-Markov theorem if the assumptions are met then the resulting parameter estimator has the property of Best Linear Unbiased Estimator (BLUE).

In multiple regression analysis, there is a term called multicollinearity which is a condition of an indication of a strong relationship between independent variables. The term multicollinearity was first introduced by Ragnar Frisch in 1934 which states that multicollinearity occurs if there is a perfect linear relationship or several variables or all variables are independent of multiple linear regression models.
The existence of multicollinearity in the regression model causes major problems. According to Montgomery and Hines in 1990, multicollinearity caused the regression coefficients generated by multiple regression analysis to be very weak in other words it could not produce analytical results that represented the nature or influence of the independent variables. To overcome the case of multicollinearity in the linear regression model, in 1960 Herman O. A. Wold, proposed a procedure known as Partial Least Square (PLS) which is an alternative method to overcome the limitations of the Least Smallest Method when the data experiences multicollinearity.

One application of the Partial Least Square method is to analyse the amount of electricity costs paid by a family. Electricity has become a priority and is needed in people's lives in modern times [2]. In Indonesia, the electricity needs of the community are met by the National Electric Company (PLN) as the holder of electricity concessions. PLN conducts consumer classification based on the electricity costs charged, in 4 (four) groups, namely: households (families), businesses, industry and government or the public. Households as one of the electricity energy concessions, have different usage patterns from one another based on the period of time of use and the amount of power installed in each house. The demand for electricity consumption per household (family) depends on the needs and usage of each household (family) such as the amount of household income (family), the number of people who live, the area of the house, the number of rooms in a house, and the number of electronic devices. These factors have a positive and significant effect on the amount of electricity costs to be paid [4].

The income factor has a significant effect on the number of electronic devices. The greater the income of a household, the greater the ability to buy electronic equipment, meaning that factors that influence the amount of electricity costs of a family can also affect each other (related). In linear regression analysis, this condition is known as multicollinearity [1].

Based on the description above, researchers feel the need to examine the factors that influence the amount of electricity costs in a family and solve the multicollinearity problems that occur. In this study, the SPSS (Statistical Package of Social Science) 2.0.1 version program is used. The problem that will be examined in this study is how to overcome multiple linear regression models that contain multicollinearity problems using the PLS method. The purpose of this study is to overcome the multiple linear regression model that contains multicollinearity problems using the PLS method.

2. Base of Theory

Multiple linear regression analysis is a statistical analysis that aims to investigate the effect of several independent variables namely \( X_1, \ldots, X_k \) upon a dependent variable \( Y \). The multiple regression model with \( k \) independent variables is:

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i .................................................................(1)
\]

In the multiple regression analysis, it is important to note that there are assumptions that must be fulfilled between autocorrelation, heteroscedasticity, normality, as well as significance and multicollinearity tests. If all the assumptions are fulfilled then based on the Gauss-Markov theorem, the resulting model will be BLUE In the analysis of multiple regression known a method that is the least squares method which is a method that aims to minimize squared squares [3].

A multiple linear regression model that involves more than one independent variable, often contains the problem of multicollinearity, a condition where there is a strong correlation between the independent variables \( X \) included in the formation of a linear regression model. It is clear that
multicollinearity is a condition that violates the assumption of linear regression. Of course, multicollinearity is not possible if there are only one independent variable $X$ [5].

The case of multicollinearity in a regression model is partly due to errors in the formation of the regression equation model and the number of observations analysed too little. There are several ways to find out whether there is multicollinearity:

a). Value of correlation (correlation between independent variables)

This procedure is the simplest and easiest detection. A high correlation value between one variable with another shows that there is a linear relationship on these variables.

b). Variance inflation factor (VIF) value and tolerance value

VIF value and Tolerance value are used as criteria to detect the presence of multicollinearity in multiple linear regression involving more than two independent variables. VIF values of more than 10 and Tolerance values of less than 0.1 indicate a serious multicollinearity problem [5].

To overcome the multicollinearity problem in the multiple linear regression model, the PLS method is used. This PLS method can be obtained through multiple regression by drawing conclusions from the significance test. This significance test aims to select the independent variables of PLS component builders and determine the number of PLS components formed. The purpose of PLS is to form components that can capture information from predictor variables to predict response variables.

In forming the PLS component, the dependent variable $Y$ and standardized independent variables $X$ are used. The partial least square regression model with $k$ components can be written as follows:

$$Y = \sum_{h=1}^{k} c_h t_h + e \quad \text{.................................................................(2)}$$

Where $Y$ is the dependent variable, $c_h$ is the regression coefficient for $t_h$, and $t_h$ is the $i$-th major component which is not correlated with each other ($h = 1, 2, .. k$). PLS component requires $t_h = \sum_{i=1}^{p} w(h)iX_i + e$ to be orthogonal.

To form PLS components with $k$ independent variables, the variables that are significant variables are used in explaining $Y$ at $t_1, t_2, ..., t_{h-1}$. Y regression models with respect to $t_1, t_2, ..., t_{h-1}$ and each $X_i$ are as follows:

$$Y = c_1 t_1 + c_2 t_2 + \cdots + c_{h-1} t_{h-1} + a_{hi}X_i + \text{residue} \quad \text{.................................................................(3)}$$

To obtain the orthogonal component of $t_h$ with respect to $t_{h-1}, X_i$ is regressed on the PLS component which is written as follows:

$$Y = p_1 t_1 + p_2 t_2 + \cdots + p_{(h-1)i} t_{h-1} + X_{(h-1)i} \quad \text{.................................................................(4)}$$

with $X_{(h-1)i}$ is the residue that results from the regression every $X_i$ with respect to $t_1, t_2, ..., t_{h-1}$. The $h$-th component is defined as follows:
\[ t_h = \frac{1}{\sqrt{\sum_{i=1}^{p} (a_{hi})^2}} \sum_{i=1}^{p} a_{hi} X_{(h-1)i} \] \hspace{1cm} \text{(5)}

With \( a_{hi} \) is the regression coefficient of \( X_{(h-1)} \) in \( Y \) at \( t_1, t_2, \ldots, t_{h-1} \). If equation (3) is substituted into equation (2) then it is obtained:

\[
\hat{Y} = c_1 t_1 + c_2 t_2 + \cdots + c_{h-1} t_{h-1} + a_{hi}(p_{1i} t_1 + p_{2i} t_2 + \cdots + p_{h-1} t_{h-1} + X_{(h-1)i}) + \text{residue}
\]
\[
= (c_1 a_{hi} p_{1i}) t_1 + (c_2 a_{hi} p_{2i}) t_2 + \cdots + (c_{h-1} a_{hi} p_{(h-1)i}) t_{(h-1)i} + \text{residue}
\]
\[
= c_1' t_1 + c_2' t_2 + \cdots + c_{h-1}' t_{(h-1)i} + \text{residue}
\]

With \( c_{h-1}' = c_{h-1} a_{hi} p_{(h-1)i} \) such as the \( h \)-th PLS component can be written as follows:

\[ t_h = \frac{1}{\sqrt{\sum_{i=1}^{p} \text{cor} (Y, X_{(h-1)i} X_i^*)}} \sum_{i=1}^{p} \text{cor} (Y, X_{(h-1)i}) X_i^* \] \hspace{1cm} \text{(6)}

With \( X_{(h-1)i} \) is the standard residue of each regression and each \( X_i \) with respect to \( t_1, t_2, \ldots, t_{h-1} \). The PLS component calculation stops when there are no more significant independent variables that construct PLS. After the PLS component is obtained, the PLS component that is formed is transformed into the original variable, namely:

\[
Y = \sum_{h=1}^{k} c_h t_h + e
\]
\[
= \sum_{h=1}^{k} c_h \left( \sum_{i=1}^{p} w_{(h)i} X_i \right) + e
\]
\[
= \sum_{i=1}^{p} \sum_{h=1}^{k} c_h w_{(h)i} X + e_i
\]
\[
= \sum_{i=1}^{p} \left( \sum_{k=1}^{k} c_h w_{(h)i} X_i \right) + e
\]
\[
Y = \sum_{h=1}^{k} b_i X_i + e \] \hspace{1cm} \text{(7)}

With:

- \( Y \) : dependent variable
- \( c_h \) : regression coefficient \( Y \) with respect to \( t_h \)
- \( X_i \) : independent variable matrix
- \( t_h = \sum_{i=1}^{p} w_{(h)i} X_i \) : uncorrelated \( h \)-th main component
- \( b_i = \sum_{h=1}^{k} c_h w_{(h)i} X_i \) : weight component for \( X_1 \) variable in the \( h \)-th PLS main component
- \( e \) : error vector
3. Literature References

Supporting literature in conducting this research is in the form of journals, books, writings and articles describing how to use Partial Least Square (PLS) to overcome data containing multicollinearity. The variables used in this study are: Large electricity bill ($Y$), Total income ($X_1$) of the family, Number of family members ($X_2$), Building area of the house ($X_3$), Number of rooms in a building ($X_4$), Number of electronic devices used ($X_5$). The research procedure is explained as Fig. 1:

![Research Procedure Diagram]

**Figure 1.** Research procedure
4. Results and Discussion

4.1. Regression assumption test

4.1.1 Multiple regression model

From the data on the amount of electricity consumption costs and the factors that influence it, using the least squares method based on the SPSS program, obtained:

Table 1. Y variable bounded regression against all independent variables

| Model          | Unstandardized Coefficients | Std. Error |
|----------------|----------------------------|------------|
| 1 (Constant)   | -40.164                    | 15.733     |
| Family Income  | 0.020                      | 0.015      |
| Number of family members | 13.379                   | 6.463      |
| House Size     | 0.662                      | 0.121      |
| Number of Rooms | -5.631                   | 4.689      |
| Number of Electronic Devices | -1.152               | 3.044      |

4.1.2 Normality test

Based on Table 1, regression models are said to be good if they meet the normality assumptions. In this study, the Kolmogorov-Smirnov test with a significance level of 0.05 obtained Asymp. Sig (2-tailed) = 0.487 > 𝛼 = 0.5. Such that 𝐻₀ means that there is no distributed data.

4.1.3 Autocorrelation test

Autocorrelation testing using the Durbin Watson test, resulting in a calculated DW value = 1.808. DW table values are DW Upper = 1.7859 and DW Lower = 1.2305. The calculated DW value = 1.808. Value > DW Upper = 1.7859, it was concluded that autocorrelation did not occur.

4.1.4 Heteroscedasticity test

Heteroscedasticity test is done by looking at the plot of the SPSS output. Heteroscedasticity occurs if there are clear patterns in the output such as wavy. Fig. 2 is a plot of heteroscedasticity:

Figure 2. Heteroscedasticity plot
4.2 Significance Test
In this study, to test the significance used the F test with a significance level of 0.05 obtained the calculated f value = 54.498 > f table value = 2.9. So $H_0$ which means there is no significant effect between the independent variables simultaneously or together on the dependent variable.

4.3 Multicollinearity Test
Multicollinearity test is performed to test whether there is a correlation between independent variables. To detect multicollinearity, it can be seen from the VIF value and the tolerance value of the SPSS output. Following are the VIF values and tolerance of the independent variables:

| Model                  | Collinearity Statistics | VIF |
|------------------------|-------------------------|-----|
|                        | Tolerance               |     |
| Income                 | 0.101                   | 9.918|
| Family members         | 0.100                   | 10.001|
| House Size             | 0.193                   | 5.194|
| Number of Rooms        | 0.059                   | 16.874|
| Number of Electronic Devices | 0.094                  | 10.597|

Based on Table 2, it is obtained that there is a VIF value of more than 10, namely the variables $X_2 = 10.001$, $X_4 = 16.874$ and $X_5 = 10.597$ and the tolerance value less than 0.1, namely the variable $X_4 = 0.059$ and $X_5 = 0.094$, so that it can it was concluded that there was a multicollinearity problem in the independent variables included in the linear regression model. To overcome the multicollinearity problem that occurs, the Partial Least Square method will be used.

4.4 Partial Least Square
4.4.1 Forming the first PLS
Before the first PLS component is formed, it first tests the significance of each variable. Following are the results of the partial significance test based on SPSS output:

| Model                  | T    | Sig. |
|------------------------|------|------|
| Income                 | 8.601| 0.000|
| Number of family members | 11.213| 0.000|
| House Size             | 15.179| 0.000|
| Number of Rooms        | 8.336| 0.000|
| Number of Electronic Devices | 9.133| 0.000|

Based on the amount of data $n = 40$, the number of independent variables = 5, and the significance level of 0.05, obtained the value of $T$ (0.025: 34) = 2.03224. $T$ value calculated $X_1 = 8.601$, $X_2 = 11.213$, $X_3 = 15.17$, $X_4 = 8.336$ and $X_5 = 9.133$ > $T$ table, then reject $H_0$ means that all independent variables have a significant effect on the dependent variable $Y$ so that all variables are used in the formation of the first PLS component.
\[ t_1 = \frac{1}{\sqrt{\sum_{i=1}^{5} \text{cor}(Y, X_i)^2}} \sum_{i=1}^{5} \text{cor}(Y, X_i)X_i^* \] ..........................(8)

We obtain the first PLS component:

| Table 4. First PLS component \((t_1)\) |
|---------------------------------------|
| 2.49438                              | 0.237265 | -1.402549 | -1.218721 |
| -1.43522                             | 0.351694 | 0.679703  | -0.514945 |
| -0.00002                             | -1.463190| 0.992414  | -2.734904 |
| 0.62707                              | -0.100630| 1.283661  | 0.053177  |
| 1.60719                              | -1.690158| -2.444981 | 0.355592  |
| 2.66208                              | -0.332465| 0.335999  | 3.172602  |
| -1.12239                             | 4.778252 | 3.040561  | 0.154573  |
| -0.43554                             | -2.019609| -1.911697 | 0.962651  |
| -1.37551                             | -0.164464| -1.559340 | -2.666519 |
| -0.27103                             | 1.574904 | -1.576816 | 1.076927  |

**4.4.2 Forming the second PLS**

Before forming the second PLS component, the significance test of each independent variable is performed on the dependent variable and the first PLS component \((t_1)\).

| Table 5. T test results for each independent variable and \((t_1)\) |
|---------------------------------------------------------------|
| **Model**          | **T**   | **Sig.**     |
| Income            | -1.740  | 0.000        |
| Number of family members | 0.679  | 0.502        |
| House Size        | 4.476   | 0.000        |
| Number of Rooms   | -4.133  | 0.000        |
| Number of Electronic Devices | -1.721 | 0.0094      |

T value is obtained as in the Table 5, obtained T value calculated \(X_1 = -1.740\), \(X_2 = 0.679\), \(X_4 = -4.133\) and \(X_5 = -1.721 < T\) table, then accept \(H_0\) which means independent variables \(X_1, X_2, X_4\) and \(X_5\) has no significant effect on the dependent variable \(Y\) where as \(X_3 = 4.476 > T\) table = 2.03224 which means \(X_3\) significant effect on the dependent variable \(Y\), so the second PLS component will be calculated by only reviewing the value of the correlation coefficient of \(Y\) to \(X_3\):

\[ t_2 = \frac{1}{\sqrt{\sum_{i=1}^{5} \text{cor}(Y, X_i)^2}} \sum_{i=1}^{5} \text{cor}(Y, X_{13})X_{13}^* \] ..........................(9)

We obtain the second PLS component on Table 6.
4.4.3 Forming the third PLS

To form the third PLS component, first test the significance of all independent variables on the dependent variable and the first PLS component variable and the second PLS component variable. If there are still significant independent variables that build the dependent variable, a third PLS component will be formed, but if there are no more significant independent variables that build the dependent variable, there is no need to form a third PLS component. Using the T test, and a significance level of 0.05 a new calculated T value is obtained as in Table 7.

Table 7. T test results for each independent variable and (t₁) and (t₂)

| Model                  | T      | Sig. |
|------------------------|--------|------|
| Income                 | -1.740 | 0.090|
| Number of family members | 0.679  | 0.502|
| House Size             | -4.550 | 0.000|
| Number of Rooms        | -4.113 | 0.000|
| Number of Electronic Devices | -1.721 | 0.094|

Based on Table 7, the calculated T value is obtained $X_1 = -1.740, X_2 = 0.679, X_3 = -4.550, X_4 = -4.133$ and $X_5 = -1.721< T$ table, , then accept $H_0$ which means independent variables $X_1$, $X_2, X_3, X_4$ and $X_5$ and no significant effect, so the calculation stops at the second PLS component with two new components $t_1$ and $t_2$. After obtaining the PLS component, then the independent variable $Y$ is regressed against the components $t_1$ and $t_2$ with a real level of 0.05, obtained output as in the Table 8, then the new regression model is:

\[
\hat{Y} = 126.220 - 32.098t_1 + 55.119t_2
\]

\[
\hat{Y} = 126.220 - 32.098(0.427X_1^* + 0.460X_2^* + 0.487X_3^* + 0.422X_4^* + 0.435X_5^*) - 55.119
\]

\[
(-0.467X_1^* - 0.503X_2^* - 0.504X_3^* - 0.462X_4^* - 0.476X_5^*)
\]

\[
\hat{Y} = 126.220 - 13.705X_1 - 14.765X_2 - 15.342X_3 - 13.545X_4 + 13.692X_5 + 25.47X_1 + 27.724X_2
\]

\[
+27.779X_3 + 25.467X_4 + 26.236X_5
\]

\[
\hat{Y} = 126.220 + 12.034X_1 + 12.437X_2 + 12.959X_3 + 11.919X_4 + 12.247X_5
\]
\[ \hat{Y} = 126.220 + 12.034 (Income) + 12.437 (Number \ of \ family \ members) \\
+ 12.959 (House \ Size) \\
+ 25.467 (Number \ of \ rooms) + 26.236 (Number \ of \ Electronic \ devices) \]

**Table 8.** Y regression with respect to \( t_1 \) and \( t_2 \)

| Model                  | Unstandardized Coefficients | Std. Error |
|------------------------|-----------------------------|------------|
| (Constant)             | 126.220                    | 4767.505   |
| First PLS Component    | -32.098                    | 2822.896   |
| Second PLS Component   | -55.119                    | 4891.357   |

4.5 *Multicollinearity Test*

Multicollinearity test is performed to test whether there is a correlation between independent variables. To detect multicollinearity, it can be seen from the VIF value and the tolerance value of the SPSS output. Following are the VIF values and tolerance of the independent variables:

**Table 9.** VIF and tolerance values

| Model                  | Collinearity Statistics |
|------------------------|-------------------------|
|                        | Tolerance   | VIF  |
| First PLS Component    | 1.000       | 1.000|
| Second PLS Component   | 1.000       | 1.000|

Based on Table 9, it is obtained that there is no a VIF value more than 10 and no a tolerance value less than 0.1, so that it can it was concluded that there no was a multicollinearity problem independent variables included in the linear regression models.

4.6 *Coefficient of determination*

To test the feasibility of a regression model or in other words, a regression model is said to be good if the coefficient of determination of the formed model is close to 1.

**Table 10.** Coefficient of determination

| Model    | R       | R Square |
|----------|---------|----------|
|          | 0.935\(^a\) | 0.874    |

Based on Table 10, the coefficient of determination value is \( R^2 = 0.874 \). This means that the linear regression model is considered good for estimating the effect of using independent variables on the dependent variable.

5. **Conclusion**

Based on the results of data processing and analysis using the Partial Least Square method to overcome the multicollinearity problem in the amount of electricity bill and the factors that influence
it, this research \( \hat{Y} = 126.220 + 12.034 \text{(Income)} + 12.437 \text{(Number of Family Members)} + 12.959 \text{(House Area)} + 11.919 \text{(Number of Rooms)} + 12.274 \text{(Number of Electronic Devices)} \). Number of Rooms has a greater influence on the amount of electricity costs, so the wider the building of a house, the greater the amount of electricity bills to pay. Other independent variables also have a positive influence on the cost of electricity bills.

**References**

[1] Bachtiar M 2019 Analysis of factors affecting consumer demand for electricity in households in Guntarano Village, Tanantovea District, Donggala Regency *Katalogis* 1 1-14

[2] Hadijah 2014 Analysis of factors affecting household electricity demand in Soppeng Regency Thesis Universitas Islam Negeri Alauddin Makassar.

[3] Indahwati R D, Kusnandar and E Sulistianingsih 2014 Partial least squares method to overcome multicollinearity in double linear regression models *Bimaster* 3 169-174

[4] Kristianto I S 2015 Analysis of household electricity consumption in Tembalang District *Skripsi* Universitas Diponegoro

[5] Putiray Ch M 2016 Stepwise regression method for overcoming multicollinearity problems in multiple linear regression *Skripsi* Universitas Pattimura

[6] Rahmadeni, and D Anggreni 2014 Analysis of the number of workers against the number of patients in arifin achmad pekanbaru hospital using the partial least square method *J. Sains dan Teknologi* 12 48-57

[7] Supriyadi E 2017 Perbandingan Metode *Partial Least Square* (PLS) dan *Principal Component Regression* (PCR) Untuk Mengatasi Multikolinieritas Pada Model Regresi Linier Berganda *UNNES J Mathematics* 6 117-128