journals.uoz.edu.krd

گۆڤارا زانستێن مرۆڤایەتی یا زانکۆیا زاخۆ
مجلة العلوم الانسانية لجامعة زاخو
Humanities Journal of University of Zakho (HJUOZ)
*Vol. 10, No. 2, pp. 453– 463, June-2022*

hjuoz.uoz.edu.krd
p-ISSN: 2664-4673
e-ISSN: 2664-4681

# Partial Least Squares Regression Methods with Application of Mas Cement Factory in Sulaymaniyah Governorate

Sherin Youns Mohyaldeen[1*], Muhammad Abduljabar. Alhassawy[2]

[1]Department of Petroleum geology, Technical College of Petroleum and Minerals Science /zakho, Duhok Polytechnic University, Kurdistan Region – Iraq. ( sherinzaxoyi@gmail.com)

[2]Statistics and Informatics, Dohuk education 'high school of commerce, Kurdistan Region – Iraq.( muhammad.ibrahim@uod.ac)

***Received: 04. 2022 / Accepted: 06. 2022 / Published: 06. 2022*** *https://doi.org/10.26436/hjuoz.2022.10.2.847*

**ABSTRACT:**

This paper was dealing with variables for MAS Cement Factory where evince many problems , more than one variable dependent and presence the problem of multicollinearity and so presence the correlation between the predictive variables and the dependent variables and so smallness size the research sample. used the method , Partial Least Squares PLS to Solve the problems above, also considered as one of the methods which dally methodically different in deduction the Components dependent on curing the correlation the presence between the predictive variables and the dependent variables , more over this method is more competence in dealing with the problems above. Through the statistical analysis, the PLS method it has succeeded in establishing the optimal Regression model for all three depended variables for the data of this paper.

**Keywords:** Regression Model, Partial least square method, Multivariate Partial least square Regression, Univariate Partial least square Regression, Components.

## 1. Introduction

In the case of multiple linear regression equation faced multicollinearity, there is a correlation between predictive variables. The estimates produced in this case are influenced by the relationships between the predictive variables and not only by the relationship between the dependent variable and predictive variables. That the columns of the matrix of predictive variables and their classes must be linearly independent with each other (Jeeshim and Kucc ,2002) As well as the problem in the above, there is another problem is the small size of the sample involved also the problem of the existence of more than one dependent variables are supported in the study, To diagnosis the above there is a method of the principal-component analysis (PCA), which aims at creating new orthogonal variables called components instead of predictive variables that are related to each other. This method deals with the correlation between predictive variables without taking correlation between the dependent variables and the predictive variables in the orthogonal component configuration process, Thus, there is a more efficient method than the above method, namely, the Partial Least Squares Analysis (PLSA) method, which aims orthogonal components. The correlation between the dependent variable and the predictive variables, as well as the correlation between predictive variables (Mita and Yan,2008) , It also to form orthogonal components in the case of other dependent variables that depend on the predictive variables are related to those predictive variables. In this case, the method is called the Multivariate PLS. This is the general case, and the specific case is the case of one dependent variable versus (m) of predictive variables called Univariate PLS (Garthwaite, 1994).

The PLSA method is defined as one of the methods of reducing the dimension of the data used in the study. A specific number of orthogonal components was selected and analyzed instead of analyzing a large number of the original variables that were related with complex relationships.

The main aim of this paper is to discuss the method of PLSA in both general (MPLSA) and specific (UPLSA) types, and to identify its different properties in addressing problem of multiple linear relation (multicollinearity) between the dependent variables as well as the predictive variables between them and with the dependent variable, the application of real data from MAS Cement Factory, where the explanatory variables were electric power, black oil, stones and soil. With more than one dependent variables approved, namely Thermal Emission Factor resulting from the production process, the quantity of cement production and the quantity of clinker production for the period (2008 - 2020).

## 2. Literature Review

The statistical studies and research with the partial least square method as well as the statistical studies and research with the problem of multiple linear relations (multicollinearity) are very many and can be referred to some of the following:

The researcher (Garthwaite, 1994) published a study comparing the method of Univariate PLS with five other regression methods, including the method of the lower squares and the main components in terms of their ability to reduce the dimension, ie reducing the number of explanatory variables in the estimated regression model through Three examples include eight (8) explanatory variables, while the second contains twenty (20) explanatory variables, the third contains 50 (50) explanatory variables, and each of the examples above contains one dependent variable and proves that the Univariate PLS method is the best and best at The process of reducing the dimension especially when the error variation is large and the size of the sample used in the search is small.

Sakallioglu and Akdeniz (1998) also published research on the problem multicollinearity and its detection using the factor of distinctive values. Four methods were proposed to address the above problem, including the regression of principle components and the Ridge regression.

(Jeeshim and Kucc ,2002). published a paper on the problem of multiple of linear relationships and their diagnosis through the use of the parameters of the variance amplification factor, the conditional index and the variance ratios. In addition, the researcher (Abdi, 2003) investigated the method of multivariate PLS with an applied example of data consisting of three dependent variables and four predictive variables. The final results of this example were presented using the above method.

* Corresponding Author.

(Maitra and Yan,2008) published a research that dealt with the method of principle components and the method of partial least square method and the comparison between two methods in terms of ability to reduce the dimension by using data of six variables predictive and one variable supported.

## 3. Methodology

### 3.1 Regression analysis and application problems

A regression analysis is a statistical tool used to analyze the relationship between one or more Predictive Variables and a Dependent Variable. Regression analysis is one of the most widely used statistical methods in different sciences because it is used to describe the nature of the relationship between variables through a mathematical model and to know whether they are positive or inverse, linear or nonlinear. (Bluman , 2009)

The regression uses are data description, parameter estimation, prediction as well as control, Regression classified into:

1-Linear regression is divided into:

a. Simple Linear Regression contains only one predictive variable, and its equivalent in general is written as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad (3.1)$$

b. Multiple Linear Regression includes several predictive variables, and its general equation form of predictive variables is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_m X_{im} + \varepsilon_i \quad (3.2)$$

2- Non Linear Regression or Curvilinear Regression is also divided into two parts:

a. Simple Curvilinear Regression also contains only one predictive variable, and its equivalent in general is as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \cdots + \beta_m X_{i1}^m + \varepsilon_i \quad (3.3)$$

b. Multiple Curvilinear Regression includes multiple predictive variables and their equivalent to two predictive variables:

$$Y_i = \beta_{00} + \beta_{10} X_{i1} + \beta_{01} X_{i2} + \beta_{11} X_{i1} X_{i2} + \beta_{20} X_{i1}^2 + \beta_{02} X_{i2}^2 + \cdots + \beta_{mn} X_{i1}^m X_{i2}^n + \varepsilon_i \quad (3.4)$$

In the linear relationship studied there are properties or assumptions related to random error and others related to the adopted variable, and the hypotheses related to random error can be follows: (Poole and Farrell ,1970).

First assumption: $\varepsilon_i$ random variable.

The second assumption:. $E(\varepsilon_i) = 0$

The third assumption:

$$\text{var}(\varepsilon_i) = E[\varepsilon_i - E(\varepsilon_i)]^2 = \sigma_\varepsilon^2$$

Fourth assumption: $\varepsilon_i$ Normal distribution was distributed.

The previous four assumptions can be made in short form

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

Fifth assumption:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E(\varepsilon_i \varepsilon_j) = 0 \qquad \forall i \neq j \quad , \quad (i, j = 1,2,\cdots,n)$$

The sixth assumption: $E(\varepsilon_i X_i) = 0$

The seventh assumption: The predictive variables are not related to each other Non Multicollinearity In fact, the researcher encounters this hypothesis when the model studied includes more than one predictive variable, where there should be no linear multiplicity and thus the effect of each predictive variable can be identified on the separately dependent variable .

As for the properties of the dependent variable, the distribution of this variable should be normal, predicted or mean distribution given in $\overline{Y} = E(Y_i) = \beta_0 + \beta_1 X_i$ the case of simple linear regression. The variance of this variable is given in formula. $\text{var}(Y_i) = E[Y_i - E(Y_i)]^2 = \sigma_\varepsilon^2$ The above properties can be placed in short form $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma_\varepsilon^2)$.

One of the most important application problems faced by researchers is the lack of one or more statistical analysis hypotheses. The estimated model is judged to be an optimal model. The data represent best representation and can be relied on in predicting the future values of the dependent variable. (Jeeshim and Kucc, 2002). But this is not always the case. Often, problems arise that lead to a violation of one or more of the hypotheses of statistical analysis. One of the reasons for this problem is sometimes the limited number of views used in the experiment or research. Choosing the appropriate sample size for the search problem is often an effective way to achieve accurate and reliable estimates for making wise decisions or achieving the scientific goal (Henry, 2013), also the Problem of Multicollinearity when this problem is present, this leads to the lack of all or most of the analysis hypotheses in the ordinary least squares method (OLS) (Jeeshim and Kucc, 2002). As well when studying a particular phenomenon, there is more than one dependent variable dependent or influenced by the explanatory variables themselves, and then there is a correlation between the dependent variables and the explanatory variables, in addition to the correlation between the explanatory variables. In addition, t-tests to test the significance of the coefficients of the explanatory variables in the regression model become suspect (Disatnik and Sivan, 2014).

### 3.2 Diagnosis of Multicollinearity

The problem of Multicollinearity is defined as a high degree of correlation or linear dependence between two or more explanatory variables in the multiple regression models, For the purpose of diagnosing the problem of linear multiplicity in a regression model that contains two or more coefficient variables, we follow the following:

1- Correlation Matrix for explanatory variables:

Simple correlation is the simplest measure of linear interference detection. $[r(X_i, X_j) = \mp 1]$ If this indicates a complete linear relationship between $X_i$ and $X_j$, where $\mathbf{R}$ is the correlation matrix between the explanatory variables, if the linear form can be expressed in a straight line, then there is a linear relationship between $X_i$ and $X_j$, if the coefficient of correlation between $X_i$ and $X_j$ negative, Take the following form: (Bayonne and etc., 2020)

$$X_i = -X_j \qquad \text{or} \qquad X_i + X_j = 0$$

If the simple correlation between $X_i$ and positive $X_j$, the linear relationship takes the following form:

$$X_i = X_j \qquad \text{or} \qquad X_i - X_j = 0$$

If more than two variables have a linear relationship, it is not necessary $r(X_i, X_j)$ to be close to the correct one or even

large. For this reason, the correlation scale alone is insufficient in detecting linear interference.

2- The Eigen Values and Eigen Vectors

The idea of equations, Eigen and vectors is to move a vector from a given field by a given matrix to another area in which the vector is the same multiplied by another numerical value. Where the result of multiplying the Eigen vector $\mathbf{v}$ in the square matrix $\mathbf{A}$ with dimensions m*m produces the same vector after it hits the value of Scalar is $\lambda$ called the Eigen root of the matrix $\mathbf{A}$, that is,

$$\mathbf{Av} = \lambda\mathbf{v}$$

$$\therefore \mathbf{Av} - \lambda\mathbf{v} = 0$$

$$\Rightarrow (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0 \qquad (3.5)$$

According to Rule Cramer, a trivial solution can have this equation in one case if the specified matrix of a matrix $\mathbf{A}$ is zero, ie:

$$\left|\mathbf{A} - \lambda\mathbf{I}\right| = 0 \qquad (3.6)$$

The equation above is called the Characteristic Equation of the matrix $\mathbf{A}$, and its solution gives a distinct formula that takes the following formula:

$$\lambda^{m} + C_{m-1}\lambda^{m-1} + \cdots + C_{1}\lambda + C_{0} = 0 \qquad (3.7)$$

It is a polynomial equation in $\lambda$ a class m, so it has its solutions m or roots $\lambda_1, \lambda_2, \cdots, \lambda_m$. The process of calculating the roots and Eigen vectors of the methods of detection of the multiplicity of linear relationships, if the value of one of the roots equal to zero indicates that there is a linear relationship is complete, and versa when the equal one indicates that the absence of any linear relationship. That is, the closer the value of the characteristic root than zero, the greater the relationship between the predictive variables and the observation, we can infer the most important variables that are significant in comparison with the other values in the characteristic vector.( Assaker and etc.,2014)

3- Condition Number and Condition Index ( $CI$ ):

Belsley, Kuh and Welsch developed the concept of the conditional number scale in 1980 to the CI scale shown in the following (Rawlings, et.al., 1998)

a-Condition Number:

$$\Phi = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}} \qquad (3.8)$$

They represent $\lambda_{max}, \lambda_{min}$ the largest and smallest characteristic of the matrix $\mathbf{X'X}$, respectively. I propose this measure Belsley, Kuh and Welsch in 1980. If the value of this scale is large, this indicates that there is a linear overlap between the explanatory variables and this parameter is weak. It is worth mentioning that the value of this measure is equal to the correct one in the case of the orthogonal matrix.

b- Condition Index ( $CI$ )

$$CI_j = \sqrt{\frac{\lambda_{max}}{\lambda_j}} \qquad (3.9)$$

The conditional guide for measuring linear multiplicity is based on the value of the characteristic root. In 1980, Belsley, Kuh and Welsch suggested that if the values $CI_j$ of the conditional directory of the 10 limits indicated that the degree of linear multiplicity was weak, if ( $30 \le CI_j \le 100$ ) this indicates that the degree of linear multiplicity is moderate( $CI_j$ >100)High linearity.

4- The Variance Proportion

The variance ratios are analyzed in terms $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ of the matrix $\mathbf{X'X}$, which can be expressed $\mathbf{VDV'}$ as a $\mathbf{V}$ orthogonal matrix. Its diagonal columns are characteristic of the matrix $\mathbf{X'X}$ and $\mathbf{D}$ diagonal matrix is the main diameter representing the characteristic values of the matrix $\mathbf{X'X}$,

$$\mathbf{X'X} = \mathbf{V}\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{m+1} \end{bmatrix}\mathbf{V'} = \mathbf{VDV'}$$

(3.10)

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X'X})^{-1} = \hat{\sigma}^2(\mathbf{VDV'})^{-1}$$

(3.11)

Each component of the variance proportion $P_{ji}$ can be found to vary the estimated parameter $\hat{\beta}_i$ by using the following formula:

$$P_{ji} = \frac{\hat{\sigma}^2 v_{ij}^2 / \lambda_j}{\mathrm{Var}(\hat{\beta}_i)}$$

(3.12)

Since it is $\mathrm{Var}(\hat{\beta}_i)$ the variance of the estimated parameter $\hat{\beta}_i$, calculated using the following formula:

$$\mathrm{Var}(\hat{\beta}_i) = \hat{\sigma}^2 \sum_{j=1}^{m+1}\left(\frac{v_{ij}^2}{\lambda_j}\right)$$

(3.13)

The percentages of variance in the identification of any $\hat{\beta}_i$ value are affected by the value when the ratio is greater than 0.50 (Akdeniz, 2000).

5- Variance Inflation Factor (VIF)

This parameter measures the inflation of the variance of the estimated parameters for all the explanatory variables in the model. This measure is based on the examination of the main diagonal elements of the matrix $(\mathbf{X'X})^{-1}$, where ( $VIF_j$ ) is equal $a_{jj}$, which represents $(1-R_j^2)^{-1}$ and $R_j^2$ the coefficient of determination for regression $X_j$ over the rest of the explanatory variables. Some researchers have pointed out that if (

$a_{jj} = VIF_j \geq 10$ ) this amount is sufficient to ignore the variable $X_j$ from the analysis or use another method as a substitute for the OLS in the estimation. (Weaving and etc,2019).

**3.3 Partial Least Squares Analysis (PLSA)**

The components derived from the PLS analysis are specific to the data values of both the explanatory variables and the dependent variable. They also have the potential to analyze a matrix of predictive variables with a matrix of dependent variables to find orthogonal components in cases where more than one life phenomenon or variable is required at the same time So that all dependent variables are dependent on the explanatory variables themselves and linked to them, and the PLS method works to take into account the correlation between the variable or the variables adopted and the predictive variables as well as the correlation between the variables between them, On the construction of the components resulting from the analysis, as in PC, OLS is used for the adopted variable or for each dependent variable against the components derived from the PLS analysis.

The PLS regression method is a new method of adopting regression equations. Recently, it has attracted the attention of many researchers with several modern articles. They adopt new explanatory variables, often called Factors or Component components, where any component is a linear combination of explanatory and variables or dependent variables These components are orthogonal vectors, also called latent vectors (Rosipal and Kramer, 2006).

The PLS regression method is particularly useful in constructing prediction equations when there are a large number of explanatory variables in the experiment under study and the sample data for observations of the variable are few , we have two type of PLSA, The first is called the Univariate PLS. This method is used to predict the values of one dependent variable, which is a comparison of predictive variables. The components of this method are derived from an analysis with predictive variables and then regression OLS for the dependent variable against the resulting components of the Univariate PLS analysis. This case is a special case, the second case is the general case and is called the Multivariate PLSA This method is applied in the case of L of the dependent variables with predictive variables. The components of this method are derived from the L analysis of the variables adopted with predictive variables and then the regression of OLS for each dependent variable against the resulting components of the Multivariate PLS.( Garthwaite , 1994). The PLS method is used in many fields, including organic chemistry, physics, industrial control and social science, and the forefront of its work was in the late 1960s by Wold in 1966 in the field of economics, the discoverer of this method, and in 1975 it was presented under the title Non Linear Iterative Partial Least Squares (NIPALS) and became common in medicine, especially in clinical treatments where the number of observations (number of patients) with the large number of explanatory variables of the patient's symptoms with the number of variables The approved level of health of the patient with the improvement of his health condition. (Haenlein and Kaplan , 2004 ) .

**3.4 Partial Least Squares Algorithm**

On the existence of both array matrix $X$ Independent matrix and matrix $Y$ variables dependent if there is more than one supported variable or vector $Y$ in the case of a single dependent variable, namely in terms of the standard formula. (Geladi and Kowalski ,1986) point out that for any component, $j = 1,2,\cdots,h$; $t_j$ In the form PLS should follow the following steps:

**1 –** Taking random values of random vector $u_{start}$ is a number of values that are taken randomly from the matrix of dependent standard variables $Y$ and the number of these values must be equal, which represents the number of views of all variables in the study in question and also the total of these values should be zero , And if only one dependent variable exists in the analysis used (Univariate PLS), each $u_{start}$ new component calculated is the same standard variable $Y$ .

**2-** Finding the horizontal vector $w'_{old}$ using the formula $w'_{old} = u'_{start} X / u'_{start} u_{start}$ and dimensions of this vector is $1 \times p$ where $p$ the number of explanatory or predictive variables are represented in the experiment under study, noting at this point that there is an overlap between the predictive variables $X's$ and the dependent variable $Y$ or the dependent variables $Y's$ in order to solve the correlation problem existing among them.

**3-** Finding the horizontal vector $w'_{new}$ as standardized through the formula $w'_{new} = w'_{old} / \left\| w'_{old} \right\|$, and the resulting vector $w_j$ of the analysis are orthogonal vectors meaning that the process of addressing the problem of correlation between the variable or variables dependent and predictive variables have been successful, and the dimensions of the matrix $W$ whose columns $w_j$ represent the orthogonal vectors ar $W'W = I$ e, $I$ Identity Matrix.

**4-** Finding the vertical vector $t$ , which is a primary factor extracted from the matrix $X$ using the formula $t = Xw_{new} / w'_{new} w_{new}$, where the sum of the values of that vertical vector $t$ is zero.

The following steps will be for the sector of the dependent variable $Y$ or dependent variables $Y's$:

**5 –** Finding the horizontal vector $q'$ through the formula $q' = t'Y / t't$ , which is one of the loads of the matrix $Y$ and in the case of a single variable $q'$ is a single value is only one.

**6 –** Finding the vertical vector $u$ using the formula $u = Yq / q'q$ and the sum of the values of this vector is zero.

The next step is the convergence test:

**7-** Compare the vector $t$ in step 4 with each of the previous iterations $(w'_{old}, u, q')$. If it is equal to one of them or there is very little difference, then, $t = t_j$ , $w_{new} = w_j$ where $u = u_j$ ، $q = q_j$ $j = 1,2,\cdots,h$ and then move to step 8, otherwise go to step 2 and the vector $u_{start}$ is the vertical vector $u$ that Calculated at step 6 above and so on at each frequency during the calculation of a single component, this

comparison was designed to obtain orthogonal components $t_j$. Thus PLS analysis addresses the correlation problem between predictive variables and variables or dependent variables by calculating orthogonal factors $w_j$ and at the same time, the problem of correlation between predictive variables is addressed by calculating the orthogonal factors $t_j$.

In the next step, the matrix loads $X$ are calculated as follows:

**8** -Calculate the horizontal vector $p'_j$ as in the formula

$$p'_j = t'_j X / t'_j t_j .$$

**9-** To find the regression coefficient $b_j$ due to the regression relation between $u_j$ the dependent variable and $t_j$ represents the explanatory variable where $b_j$ it is only one value for each component and is $t_j$ calculated as:It is $b_j = u'_j t_j / t'_j t_j$ used for the purpose of finding the matrix $F_j$, the matrix $Y$ of multivariate PLS or finding the vector $F_j$ for the only dependent variable The trial under study (Univariate PLS).

**10 –** The matrices of the residues $E_j$, $F_j$ and the matrix $X$, matrix or vector $Y$ respectively are found for the $t_j$ component as follows:

$$F_j = F_{j-1} - b_j t_j q'_j \text{ ، } E_j = E_{j-1} - t_j p'_j \text{ where } X = E_0 \text{ \& } Y = F_0$$

**11-** At the above point, the calculation of the first component $t_j$ and the purpose of calculating the $t_{j+1}$ following component ends with the first step of the algorithm and the same mechanism. Both the vectors $w_j$ ; $p_j$ ; $q_j$ and the regression 457coefficient $b_j$ t must be stored for the purpose of predicting the future values of the variable or the dependent variables $\hat{Y}$.

There are some notes to follow when applying the algorithm above:

**1)** If the private sector of the dependent variables has only one dependent variable then step 7 of the convergence test can be deleted and there is no need for additional repetition.

**2)** After calculating the first component, the matrix $X$ in steps 2, 4 and 8 as well as the matrix or vector $Y$ in steps 5 and 6 are replaced by the corresponding matrix of residues $E_j$ and matrix or vector $F_j$ respectively.

**4-Application part**

In this paper we studied three dependent variables and four independent variables all variables depend on them, the first dependent variable Y1 in environmental pollution is the thermal emission coefficient produced by the cement production process. The remaining variables are the quantity of production for cementY2 and clinkerY3, respectively. Mas Cement Plant for the period 2008-2020,In addition the explanatory variables(independents) were electric power X1, black oil X2, stones X3 and soil X4. The sample size n for each of the above variables is 12 observations along the search period.

Since one of the assumptions of the regression analysis of the dependent variable Y is to follow the normal distribution where it was tested by the statistical testing Kolmkrov Smirnov and found that it does not follow the normal distribution so was taken one of the types of transformation.(Square Root) at significant level 0.05. All the implementations of the study on real data applications are carried out using R version (3.4.4) and Minitab version (17).

**The results of the Multicollinearity are as follows**:

**1- Correlations Coefficients**

**Table 1:Correlation coefficients between predictive and dependent variables.**

**Correlation: sq.Y1, sq.Y2, sq.Y3, x1, x2, x3, x4**

| | sq.Y1 | sq.Y2 | sq.Y3 | x1 | x2 | x3 |
|---|---|---|---|---|---|---|
| sq.Y2 | 0.919 | | | | | |
| | 0.000 | | | | | |
| sq.Y3 | 0.973 | 0.966 | | | | |
| | 0.000 | 0.000 | | | | |
| x1 | -0.325 | -0.370 | -0.314 | | | |
| | 0.065 | 0.034 | 0.075 | | | |
| x2 | 0.970 | 0.950 | 0.977 | -0.304 | | |
| | 0.000 | 0.000 | 0.000 | 0.086 | | |
| x3 | 0.944 | 0.961 | 0.982 | -0.310 | 0.976 | |
| | 0.000 | 0.000 | 0.000 | 0.079 | 0.000 | |
| x4 | 0.688 | 0.693 | 0.693 | -0.418 | 0.673 | 0.673 |
| | 0.000 | 0.000 | 0.000 | 0.015 | 0.000 | 0.000 |

**2-The Condition Number $\phi$ = 1.100909628E+10**

**3-The Condition Index ( $CI_j$ ) and The Eigen Value $\lambda_j$ , $j = 1,2,\cdots,m+1$ .**

**Table 2:Conditional index and Eigen values**

| Number | $CI_j$ | $\lambda_j$ |
|---|---|---|
| 1 | 1 | 1.71308E+13 |
| 2 | 5.174111954 | 6.39891E+11 |
| 3 | 82.31203807 | 2.52843E+09 |
| 4 | 7421.061667 | 3.11061E+05 |
| 5 | 1.100909628E+10 | 1.41343E-02 |

**4-The Variance Proportion $(p_{ji})$**

**Table 3:Values of proportions of parameters**

| NO. | intercept | $X_1$ | $X_2$ | $X_3$ | $X_4$ |
|---|---|---|---|---|---|
| 1 | 8.251041E-33 | 1.068254E-23 | 1.138575E-06 | 0.012903818 | 5.456931E-03 |
| 2 | 0.000000 | 2.952366E-24 | 2.659519E-06 | 0.052316533 | 0.968994355 |
| 3 | 1.615597E-26 | 2.324165E-17 | 0.91556092 | 0.915702814 | 0.017832609 |
| 4 | 1.227287E-17 | 1.682313E-08 | 0.084435505 | 0.01907709 | 7.717045E-03 |
| 5 | 1.000000 | 1.000000 | 0.000000 | 0.000000 | 0.000000 |

**5- The Variance Inflation Factor ( $VIF_j$ )**

**Table 4:Values of estimation parameters $\beta_j$**

| Predictive variable | $VIF_j$ |
|---|---|
| $X_1$ | 1.256 |
| $X_2$ | 20.810 |
| $X_3$ | 20.843 |
| $X_4$ | 2.066 |

**6-The Eigen Vectors ( $V_j$'S )**

**Table 5:Eigen Vectors matrix**

| $v_1$ | $v_2$ | $v_3$ | $v_4$ | $v_5$ |
|---|---|---|---|---|
| -0.000001 | -0.000000 | -0.000017 | -0.005197 | 0.999986 |
| -0.000187 | -0.000019 | -0.003351 | -0.999981 | -0.005197 |
| -0.091373 | 0.026990 | -0.995445 | 0.003353 | 0.000000 |
| -0.927711 | 0.361025 | 0.094944 | -0.000152 | -0.000000 |
| -0.361945 | -0.932166 | 0.007949 | 0.000058 | -0.000000 |

**Interpreting the results of criteria for detecting the Multicollinearity:**

**1-** The value of the conditional number equals 1.100909628E + 10. This value is very large. This indicates that the problem of multiple linear relationships between predictive variables is very high.

**2-** There is a correlation between the two predictive variables of black oil and stones by noting:

The simple correlation coefficient between the two variables in the above correlation matrix is 0.976 and the value of P-Value for that correlation is 0.000 indicating the significance of the correlation between the two variables above at 0.05 = and the correlation of the positive type to a high degree. Also

$VIF_j$ Values for each of the two variables above are equal to 20.8 and this value is too high and greater than the value of 10. This indicates a correlation between the two variables above.

**3 –** The correlation between the two predictive variables electrical power and soil by noting:

The correlation coefficient between the two variables above is -0.418, where the value of P-Value for that correlation is 0.015 and at the significance level of 0.05. This correlation is negative.

It was observed that the characteristic value of $\lambda_{min.} = \lambda_5$ 0.0141343 was close to zero. This indicates the existence of the problem of linear multiplicity. This value corresponds to the fifth characteristic vector $\mathbf{v}_5$. In this vector, the first value among its values was 0.999986. This value corresponds to the first predictive variable, This variable is the cause of the above problem.as well This variable corresponds to the characteristic value $\lambda_5$ and the fifth conditional index $CI_5$, which is equal to 1.100909628E + 10, where it is noted that the last value is very high and larger Of the value of 100 indicates that the degree of linear multiplicity caused by the above variable is very high.

It is possible to say that the problem of linear multiplicity is very high and that all predictive variables cause the above problem. Also, by observing the correlation matrix, there is a significant correlation between each dependent variable.

Now ,the PLSM has two first cases: the general state is called the MPLS method and the second one is the special case called the UPLS method.

**MPLS method**

In this method, components were calculated by analyzing the standard dependent variables matrix together with the matrix of standard explanatory variables at the same time using the micro-squares algorithm. The components that were calculated were four standard and orthogonal components. The first and second components were selected as predictive variables in the OLS regression analysis for each dependent variable The other components are insignificant and their presence weakens the results of the OLS statistical analysis of the estimated model for each dependent variable. The name or description of the first and second components can be given by using the predictive variables in each selected standard component.

**Table 6: $sq.Y_1$ Versus MPLS by OLS**

**Regression Analysis: sq.Y1 versus PCP1, PCP2**

The regression equation is
sq.Y1 = 22.3 + 4.99 PCP1 + 2.13 PCP2

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 22.3178 | 0.4549 | 49.07 | 0.000 | |
| PCP1 | 4.9933 | 0.2809 | 17.78 | 0.000 | 1.000 |
| PCP2 | 2.1288 | 0.5373 | 3.96 | 0.001 | 1.000 |

S = 2.31931 R-Sq = 93.5% R-Sq(adj) = 93.0%
PRESS = 161.255 R-Sq(pred) = 91.55%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 2 | 1784.13 | 892.06 | 165.84 | 0.000 |
| Residual Error | 23 | 123.72 | 5.38 | | |
| Total | 25 | 1907.85 | | | |

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
Durbin-Watson statistic = 1.59791

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|
| 1 | 19.9207 | 0.657308 | (18.5610, 21.2805) | (14.9339, 24.9076) |
| 2 | 18.1575 | 0.761490 | (16.5822, 19.7327) | (13.1076, 23.2073) |
| 3 | 20.1061 | 0.791391 | (18.4690, 21.7433) | (15.0367, 25.1756) |
| 4 | 23.9375 | 0.465798 | (22.9739, 24.9011) | (19.0439, 28.8312) |

Lack of fit test
Overall lack of fit test is significant at P = 0.024

**Table 7: $sq.Y_2$ Versus MPLS by OLS**

**Regression Analysis: sq.Y2 versus PCP1, PCP2**

The regression equation is
sq.Y2 = 685 + 165 PCP1 + 54.1 PCP2

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 684.98 | 16.18 | 42.34 | 0.000 | |

PCP1 164.643 9.992 16.48 0.000 1.000
PCP2 54.09 19.11 2.83 0.009 1.000

S = 82.5024 R-Sq = 92.4% R-Sq(adj) = 91.7%
PRESS = 206724 R-Sq(pred) = 89.96%
Analysis of Variance
Source DF SS MS F P
Regression 2 1902449 951224 139.75 0.000
Residual Error 23 156553 6807
Total 25 2059002

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
Durbin-Watson statistic = 1.34370

Predicted Values for New Observations
New Obs Fit SE Fit 95% CI 95% PI
1 593.469 23.3818 (545.100, 641.838) (416.078, 770.860)
2 532.727 27.0878 (476.691, 588.762) (353.094, 712.360)
3 631.462 28.1514 (573.227, 689.698) (451.131, 811.794)
4 739.175 16.5694 (704.898, 773.451) (565.098, 913.252)
Lack of fit test
Overall lack of fit test is significant at P = 0.016

**Table 8: $sq.Y_3$ Versus MPLS by OLS**

**Regression Analysis: sq.Y3 versus PCP1, PCP2**

The regression equation is
sq.Y3 = 658 + 175 PCP1 + 78.9 PCP2
Predictor Coef SE Coef T P VIF
Constant 657.69 10.33 63.65 0.000
PCP1 175.380 6.381 27.48 0.000 1.000
PCP2 78.86 12.21 6.46 0.000 1.000

S = 52.6882 R-Sq = 97.2% R-Sq(adj) = 97.0%
PRESS = 85054.0 R-Sq(pred) = 96.26%

Analysis of Variance
Source DF SS MS F P
Regression 2 2212681 1106341 398.53 0.000
Residual Error 23 63849 2776
Total 25 2276530
There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
Durbin-Watson statistic = 1.58754

Predicted Values for New Observations
New Obs Fit SE Fit 95% CI 95% PI
1 576.664 14.9322 (545.775, 607.554) (463.378, 689.951)
2 515.392 17.2989 (479.606, 551.177) (400.674, 630.110)
3 575.086 17.9782 (537.895, 612.277) (459.921, 690.250)
4 714.382 10.5816 (692.492, 736.272) (603.212, 825.552)
Lack of fit test
Overall lack of fit test is significant at P = 0.007

Interpreting regression results OLS for each supported variable against MPLS components
Note from the previous results provide all statistical analysis hypotheses for the regression analysis of OLS and each estimated regression model, The explanatory power based on each regression model is estimated to be high, indicating that the estimated model has a high explanatory power in interpreting changes in the dependent variable, The significance of the statistical laboratory F for all estimated regression models indicates that at least one of the regression coefficients is significant (different from zero). The statistical significance of the regression t is also significant for all the regression coefficients and the estimated regression models all at the mean level of significant 0.05. This indicates that each regression parameter is estimated, is significant and differs from zero.

**UPLS method**

In this method, standard components were calculated by analyzing the matrix of standard predictive variables with each standard dependent variable separately. The above analysis was applied three times, ie, the number of dependent variables included in the

research. Each time standard components were obtained for the dependent variable used In the Univariate PLS statistical analysis, the components of each dependent variable differ from the components of any other dependent variable. The components that were calculated at each time the above statistical analysis was performed were four standard and orthogonal components. The other components that were calculated are not significant and their presence weakens the results of the statistical analysis of OLS and all the dependent variables.

**Table 9: $sq.Y_1$ Versus UPLS by OLS**

**Regression Analysis: sq.Y1 versus Comp1, Comp2**

The regression equation is
sq.Y1 = 22.3 + 5.00 Comp1 + 2.09 Comp2

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 22.3178 | 0.4545 | 49.10 | 0.000 | |
| Comp1 | 5.0017 | 0.2809 | 17.81 | 0.000 | 1.000 |
| Comp2 | 2.0867 | 0.5353 | 3.90 | 0.001 | 1.000 |

S = 2.31748 R-Sq = 93.5% R-Sq(adj) = 93.0%
PRESS = 161.383 R-Sq(pred) = 91.54%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 2 | 1784.32 | 892.16 | 166.12 | 0.000 |
| Residual Error | 23 | 123.53 | 5.37 | | |
| Total | 25 | 1907.85 | | | |

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
Durbin-Watson statistic = 1.58383

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|
| 1 | 19.9783 | 0.662285 | (18.6083, 21.3483) | (14.9923, 24.9643) |
| 2 | 18.2246 | 0.767350 | (16.6372, 19.8120) | (13.1746, 23.2747) |
| 3 | 20.0435 | 0.799081 | (18.3905, 21.6965) | (14.9725, 25.1146) |
| 4 | 23.9781 | 0.465206 | (23.0157, 24.9404) | (19.0883, 28.8678) |

Lack of fit test
Overall lack of fit test is significant at P = 0.025

**Table 10: $sq.Y_2$ Versus UPLS by OLS**

**Regression Analysis: sq.Y2 versus Comp1, Comp2**

The regression equation is
sq.Y2 = 685 + 164 Comp1 + 61.6 Comp2

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 684.98 | 15.87 | 43.15 | 0.000 | |
| Comp1 | 164.037 | 9.784 | 16.77 | 0.000 | 1.000 |
| Comp2 | 61.57 | 19.30 | 3.19 | 0.004 | 1.000 |

S = 80.9453 R-Sq = 92.7% R-Sq(adj) = 92.0%
PRESS = 195678 R-Sq(pred) = 90.50%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 2 | 1908302 | 954151 | 145.62 | 0.000 |
| Residual Error | 23 | 150699 | 6552 | | |
| Total | 25 | 2059002 | | | |

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.
Durbin-Watson statistic = 1.33704

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|
| 1 | 589.249 | 22.1415 | (543.445, 635.052) | (415.649, 762.848) |
| 2 | 527.532 | 25.5700 | (474.637, 580.428) | (351.928, 703.136) |
| 3 | 634.188 | 26.6275 | (579.105, 689.271) | (457.913, 810.463) |
| 4 | 736.424 | 16.2958 | (702.713, 770.134) | (565.616, 907.231) |

Lack of fit test

Overall lack of fit test is significant at P = 0.013

**Table 11: $sq.Y_3$ Versus UPLS by OLS**

**Regression Analysis: sq.Y3 versus Comp1, Comp2**

The regression equation is
sq.Y3 = 658 + 176 Comp1 + 75.4 Comp2

| Predictor | Coef | SE Coef | T | P | VIF |
|---|---|---|---|---|---|
| Constant | 657.69 | 10.47 | 62.84 | 0.000 | |
| Comp1 | 175.929 | 6.474 | 27.18 | 0.000 | 1.000 |
| Comp2 | 75.37 | 12.27 | 6.14 | 0.000 | 1.000 |

S = 53.3686 R-Sq = 97.1% R-Sq(adj) = 96.9%

PRESS = 87777.4 R-Sq(pred) = 96.14%

Analysis of Variance

| Source | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| Regression | 2 | 2211022 | 1105511 | 388.14 | 0.000 |
| Residual Error | 23 | 65509 | 2848 | | |
| Total | 25 | 2276530 | | | |

There are no replicates.
Minitab cannot do the lack of fit test based on pure error.

Durbin-Watson statistic = 1.57498

Predicted Values for New Observations

| New Obs | Fit | SE Fit | 95% CI | 95% PI |
|---|---|---|---|---|
| 1 | 578.617 | 15.2661 | (547.037, 610.198) | (463.788, 693.447) |
| 2 | 517.888 | 17.7068 | (481.259, 554.517) | (401.569, 634.208) |
| 3 | 573.910 | 18.3464 | (535.958, 611.863) | (457.167, 690.653) |
| 4 | 715.030 | 10.7170 | (692.861, 737.200) | (602.425, 827.636) |

Lack of fit test
Overall lack of fit test is significant at P = 0.009

From the above Results findings, all statistical analysis hypotheses provide for the analysis of the OLS regression and for each estimated regression model. That we discussed pervious (i.e $R^2$,R adj.,F test…etc).

## 5. Conclusion

The most important conclusions reached by the research through the practical side are:

**1-**The existence of the problem of multiple linear relationships between the predictive variables among them as well as the correlation between the predictive variables and the dependent variables as well as the problem of small size of the sample involved in the research at MAS cement factory. This led to the failure of the OLS method to achieve an optimal regression model.

**2 -** The research revealed that the models extracted in the form of small squares and PLS in both public and private cases MPLS and UPLS respectively were efficient mentioned in the tables 7-11.

**3-** Reducing the dimension of the data, as the PLS method, its general and specific cases, was able to reconcile a regression model in which all the hypotheses of the statistical analysis are available and for all the approved variables included in the research by adopting only two of the orthogonal standard components.

## 6. References

1-Abdi , Herve , (2003), " Partial Least Squares ( PLS ) Regression " , The University of Texas at Dallas , Encyclopedia of Social Sciences Research Methods , E–mail : herve @utdallas . edu , http://www.utdallas . edu/~herve .

2-Akdeniz , Fikri , (2000) , " III-Conditioning and Multicollinearity " , Department of Mathematics , Cukurova University , Adana , Turkey , Linear Algebra and its Applications 321 , pp (295-305) , E-mail: akdeniz@mail.cu.edu.tr .

3- Assaker, G., Hallak, R., Vinzi, V., O'Connor, P. (2014). "An Empirical Operationalization of Countries' Destination Competitiveness Using Partial Least Squares Modeling." Journal of Travel Research, 53 (1): 26-43.

4-Bluman , Allan G. , (2009) , " Elementary Statistics (Astep by step Approach) " , Seventh Edition , Published by MC Graw – Hill , Obusiness Unit of The MC Graw – Hill Companies , New York .

5- Bayonne, Enrique & Marin-Garcia, Juan A. & Alfalla-Luque, Rafaela. (2020). Partial least squares (PLS) in Operations Management research: Insights from a systematic literature review. Journal of Industrial Engineering and Management. 13. 565. 10.3926/jiem.3416.

6-Disatnik , David and Sivan , Liron , (2014) , " The Multicollinearity illusion in Moderated Regression Analysis " , Tel Aviv University , Carnegie Mellon University , USA. , PP ( 1-12 ), E-mail: daviddis@post. Tau. Ac. Il and E-mail: Isivan@andrew. Cmu. Edu.

7-Garthwaite , Paul H. , (1994) , " An Interpretation of Partial Least Squares " , Journal of The American Statistical Association , Theory and Method , Vol . 89 , No . 425 , pp ( 122 – 127 ) .

8-Geladi , Paul and Kowalski , Brucer , (1986) , " Partial Least Squares Regression : Atutorial " , University of

Washington , Seattle , WA 98195 (U . S . A .) , Printed in The Netherlands , pp ( 1 – 17 ) .

9-Haenlein , Michael and Kaplan , Andreas M. , (2004) , " Abeginner,s Guide to Partial Least Squares Analysis " , Copyright©Lawrence Erlbaum Associates , Inc. , E-mail:Michael.Haenlein@Bain.com , PP (283-297) .

10-Henry , Gary T. , (2013),"Practical Sampling",Chapter Title , Sample Size , Publishing Company , SAGE Publications , Inc. , 14 October , City , Thousandoaks

11-Jeeshim and Kucc , (2002) , " Multicollinearity in Regression Models " , No. 625 , pp ( 1-8 ) , http:// php. Indiana. edu /~ kucc 625 .

12-Maitra , Saikat and Yan , Jun , (2008) , " Principle Component Analysis and Partial Least Squares : Two Dimension Reduction Techniques for Regression " , Casualty Actuarial Society , pp (79 – 90) .

13-Poole , Michael A. and Farrell , Patrick N.O. , (1970) , " The Assumptions of the Linear Regression Model " , Received , 10 July , pp (145-157) .

14-Rawlings , John O. , Pantula , Sastry G. and Dickey , David A. , (1998) , " Applied Regression Analysis ", Aresearch Tool , Second Edition , Berlin , New york .

15-Rosipal , Roman and Kramer , Nicole , (2006) , " Overview and Recent Advances in Partial Least Squares " , Copyrighted©Springer-Verlag , Berlin , pp (34 – 51) .

16-Sakallioglu , S. and Akdeniz , F. , (1998) , " Generalized Inverse Estimator and Comparison with Least Squares Estimator " , Cukurova University , Department of Mathematics ,

Copyright©Tubitak , 22 , pp (77 – 84) .

17- 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), e0211776.

**ریکێن جوارگوشەیین بەشەکی دگەل جیبە جیکرن لسەر کارگەها ماس یا چیمە نتووێ ل پاریزکەها سولەیمانیێ**

**پوختە :**

ئەڤ فەکولینە سەرەدەریێ دگەل وان گورانکاریێن ل کارگەها چیمەنتووی (ماس) ل پاریزکەها سولەیمانیێ دکەت,گەلەک ئاریشە دیاردبن زوربەیا وان گورانکاریین دەملدەست دگەل گورانکاریین پیشبینیێ کرینە , وەهروەسا هەبوونا پەیوەندیێ دگەل گورانکاریین پیشبینیێ کری و گورانکاریین دەملدەست (معتمد) ,راستە قەبارەیێ ڤێ ڤەکولینێ نموونەکا بچووکە , هاتییە بکارئینان ب رێکا جوارگوشەیین بەشەکی (PLS) بوچارەسەرکرنا ئاریشەیین رابردوو, ئەڤ فەکولینە دیڤچوونا ئاریشا بکەت یین بەری نهو , ئەڤ ئاریشین دەستکرد ریکێن بەرچاڤ وروون ئانکو بەربەلاف بو چارەسەرکرنا پەیوەندیین راستە راست دنافبەرا گورانکاریین پیشبینیێ کری وەهروەسا پەسنا وان رێکین پروگرامیێ جیاواز ئەڤێن کورت , و ئەڤ ئاراستەیین پیڤە گریدایی ئەڤێن رووبدەن دنافبەرا گورانکاریین پیشبینیێ کری و گورانکاریین راستە راست , سەرەرای وێ چەندێ سەرەدەریەکا باش دگەل ئەوان ئاریشەیا بهیتەکرن ئەوین مەدیارکرین ل سەری دنافبەرا ئامارین شیکاری وپاشان گەهشتن بو گورانکاریین چوارکوشەیین بچووک بەشیێ (PLS) سەرەرای وێ چەندێ دڤێت ئەم رازیبوونا خوّ دەربرین ئەڤین پیشبین کری بو گوبیتکا پاشە روژێ وبوّهەمی گورانکاریین دەستکرد وراستە راست وبتنیێ بشت بەستنیێ سەر ڤان هەردوو خالا بکەین.

**طرق انحدار المربعات الصغرى الجزئية مع التطبيق على مصنع ماس للأسمنت بمحافظة السليمانية**

**الملخص**:

تعامل هذا البحث مع متغيرات معمل إسمنت الماس في محافظة السـليمانية حيث ظهرت مشـاكل عدة ، إذ يوجد أكثر من متغير معتمد واحد مع وجود مشـكلة التعدد الخطي بين المتغيرات التنبؤية وكذلك وجود إرتباط بين المتغيرات التنبؤية والمتغيرات المعتمدة فضـلاً عن صـغر حجم عينة البحث, لقد تم إسـتخدام طريقة المربعات الصـغرى الجزئية PLS لمعالجة المشـاكل أعلاه ، باعتبارها من الطرائق الشائعة في حل مشكلة تعدد العلاقات الخطية بين المتغيرات التنبؤية وكذلك بوصفها من الطرائق التي لها منهجية مختلفة في استخلاص المكونات معتمدة على معالجة الإرتباط الموجود بين المتغيرات التنبؤية والمتغيرات المعتمدة ، فضـلاً عن كونها كفوءة في التعاطي مع المشـاكل أعلاه . ومن خلال التحليل الإحصـائي تم التوصـل إلى أن طريقة المربعات الصـغرى الجزئية PLS تمكنت من توفيق أنموذج إنحدار أمثل ولجميع المتغيرات المعتمدة الثلاثة ، فضلاً عن ذلك تفوقها من حيث القدرة على التنبؤ بالقيم المستقبلية لكل المتغيرات المعتمدة وذلك من خلال الإعتماد على مكونين فقط .

**الكلمات الدالة**: انموذج الانحدار,طريقة المربعات الصغرى الجزئية,انحدار المربعات الصغرى الجزئي المتعدد, انحدار المربعات الصغرى الجزئي الاحادي,المكونات.