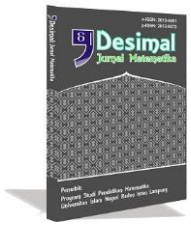
Desimal: Jurnal Matematika Vol 3 No 3 (2020) 211-218

Contents lists available at DJM

DESIMAL: JURNAL MATEMATIKA

p-ISSN: 2613-9073 (print), e-ISSN: 2613-9081 (online), DOI 10.24042/djm

http://ejournal.radenintan.ac.id/index.php/desimal/index



Rainfall Model Using Principal Component Regression Analysis with R Software in Sulawesi

Annisa Alma Yunia[1], Dianne Amor Kusuma[1], Bambang Suhandi[2], Budi Nurani Ruchjana[1,*]

[1] Universitas Padjadjaran, Indonesia
[2] Balai Observatorium Nasional Kupang, Indonesia

ARTICLE INFO

Article History
Received : 25-03-2020
Revised : 30-07-2020
Accepted : 29-08-2020
Published : 20-09-2020

Keywords:
Rainfall of Sulawesi, Local Scale Factor, Principal Component Regression Analysis.

*Correspondence: E-mail:
budi.nurani@unpad.ac.id

Doi:
10.24042/djm.v3i3.6108

ABSTRACT

Indonesia is a tropical country that has two seasons, rainy and dry. Nowadays, the earth is experiencing the climate change phenomenon which causes erratic rainfall. The rainfall is influenced by several factors, one of which is the local scale factor. This research was aimed to build a rainfall model in Sulawesi to find out how the rainfall relationship with local scale factor in Sulawesi. In this research, the data used were secondary data which consisted of 15 samples with 6 variables from Badan Pusat Statistik (BPS). The limitation of the sample size in this study was due to the limited secondary data available in the field. The data was processed using Principal Component Regression Analysis. The first step was reducing local scale factor variables so that the principal component variable could be obtained that can explain variability from the original data which then that variable was analyzed using principal regression analysis. The data were analyzed by utilizing R Studio software. The results show that two principal component variables can explain 75.2% of the variability of original data and only one principal component variable that was significant to the rainfall variable. The regression model explained that the relationship between rainfall, humidity, air temperature, air pressure, and solar radiation was in the same direction while the relationship between rainfall and wind velocity was not in the same direction. Overall, the results of the study provided an overview of the application of the Principal Component Regression analysis to model the rainfall phenomenon in the Sulawesi region using the R program.

http://ejournal.radenintan.ac.id/index.php/desimal/index

INTRODUCTION

Rain is a common natural phenomenon that occurs around the world. The climate change phenomenon caused by the greenhouse gas effect, the rainfall has been erratic. Meanwhile, rainfall is needed as an estimator of water availability for local living things which determines the boundaries of the rainy season and dry season and controls flood and drought disasters. The intensity of rainfall is usually influenced by several

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

factors, one of which is the local scale factor. The local scale factor refers to the air humidity, air temperature, air pressure, wind speed, solar radiation, and so on. Sulawesi is one of the areas in Indonesia that is considered to have quite low rainfall because the rainfall is around 1000-2000 mm/year.

Several scientific papers have discussed rainfall models, one of which is the sea surface temperatures rainfall model in West Kalimantan using the Stepwise Regression method (Handiana et al., 2016) and rainfall estimation model with climatic factors in Bangladesh using Multiple Regression Analysis (Navid & Niloy, 2018). Based on the mentioned scientific works, the methods employed were less than optimal because they did not consider the multicollinearity problem. As a result, the models obtained contained multi-collinearity problems so that the models should be improved.

At present, there have been many scientific works that employ analytical methods to overcome the multicollinearity problem, for example, overcoming the problem of multicollinearity using the Ridge Regression (Gorgees & Ali, 2017), overcoming the multicollinearity problem on factors that affect the human development index in East Java using Principal Components Analysis (Sudrajat, 2016), and overcoming the multicollinearity problem in the factors that affect the JCI on Indonesia Stock Exchange using the Latent Root Regression. The analysis methods used in the mentioned scientific papers were generally assisted by Microsoft Excel, SAS, and SPSS (Untari & Susanti, 2017).

In this research, a rainfall model was built based on multivariate data in the Sulawesi Region using Principal Component Regression Analysis to overcome the multicollinearity problem, so that the right model could be obtained for prediction studies with the help of R Studio software.

METHODS

The method used in this research was a literature study for theoretical studies and experimental studies through simulation and data processing with the Principal Component Regression model.

The Basic Concepts of Principal Component Regression Analysis

One of the basic concepts of algebra used is the eigenvalues and eigenvectors associated with Principal Component Analysis. The eigenvalues of matrix $\mathbf{A}$ is $n \times n$, notated by:

$$\det(\lambda\mathbf{I} - \mathbf{A}) = 0 \quad (1)$$

(Johnson & Wichern, 2007)

The concept of correlation is also needed to describe the linear relationship between two or more quantitative variables. The correlation value is a standardized covariance. If the correlation value between the independent variables is strong enough, it can cause multicollinearity. If there is multicollinearity, a variable that has a strong correlation with other variables in the regression model might have an unreliable and unstable power of prediction (Rencher, 2002).

Regression Analysis

Regression analysis is a statistical method for examining, modeling, and predicting relationships between variables. The relationship of a model can be expressed in an equation that connects the independent variable (X) with the dependent variable (Y) (Montgomery et al., 2012). In general, a regression model with p independent variables and n observation can be written as follows:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i;\ i = 1,2,\dots,n \quad (2)$$

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

The regression equation model, in general, can also be written in matrix notation as follow:

$$Y = X\beta + \varepsilon \tag{3}$$

In estimating the β parameter, the Least Square Method can be performed if the data does not contain multicollinearity. It can be calculated as follows:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{4}$$

where $\hat{\beta}$ is an unbiased estimate for the parameter β, such that $E(\hat{\beta}) = \beta$.

Principal Component Analysis

Principal Component Analysis (PCA) was first discovered by Karl Pearson in 1901 and named PCA by Harold Hotelling in 1933. According to (Sudrajat, 2016), Principal Component Analysis is the best method to solve the problem because it can overcome the multicollinearity (correlation is zero) in all research data conditions. PCA can be formed based on a covariance matrix or a correlation matrix.

If α is an orthogonal matrix of $p \times p$, the principal component is defined as a combination of p original independent variable that can be expressed in the form of a matrix as follows:

$$\mathbf{W} = \boldsymbol{\alpha}'\mathbf{X} \tag{5}$$

$$\begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \tag{6}$$

Description,

α: The eigenvector matrix of $p \times p$

X: The original variable vector of $p \times 1$

in the form of a linear combination, ut can be notated as:

$$W_j = \alpha_j'X = \alpha_{1j}X_1 + \alpha_{2j}X_2 + \cdots + \alpha_{pj}X_p \; ; \; j = 1,2,3,\ldots,p \tag{7}$$

If $p\,of$ the original variable is measured with different units of measurement, the variable is transformed into a standard score (standardization). Standardization of the original variable X into the Z score can be done using the following formula:

$$W_j = \boldsymbol{\alpha}_\mathbf{j}'\mathbf{X} = \alpha_{1j}X_1 + \alpha_{2j}X_2 + \cdots + \alpha_{pj}X_p \; ; \; j = 1,2,3,\ldots,p \tag{8}$$

The criteria of the Principal Component Analysis with a correlation matrix is to use principal components with more than one eigenvalues ($\lambda_j \geq 1$). The cumulative percentage variance of the principal component representing the total data variance (information) of the independent variables is approximately 75%.

Principal Component Regression Analysis

According to (Mariana, 2013), Principal Component Regression Analysis is a principal component analysis technique that is combined with regression analysis where the principal component analysis is used as the analysis stage. The principle of the Principal Component Regression analysis is to select several principal components that will be used as independent variables in

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

regression by estimating the regression coefficient using the Least Square Method.

There are two ways of forming Principal Component Regression through principal component analysis, namely using a covariance matrix or a correlation matrix (Jolliffe, 2010). Both methods are used depending on the condition of the observation range of the independent variable.

If matrix A is an orthogonal matrix $p \times p$ with $A'A = AA' = I$ where $W = XA$, then the multiple linear regression equations process becomes the Principal Component Regression as follows:

$$Y = X\beta + \varepsilon$$
$$Y = XAA'\beta + \varepsilon$$
$$Y = W\theta + \varepsilon \quad (9)$$

where θ denotes the vector of the regression parameter and $\theta = A'B$.

The Principal Component Regression model that has been reduced to k principal components is stated as follows:

$$Y = \theta_0 1 + W_k\theta_k + \varepsilon \quad (10)$$

Where:

Y: Dependent variable vector with the size of $n \times 1$

X: Independent variable matrix with the size of $n \times p$

β: Regression coefficient vector with the size of $p \times 1$

1: Vector with all elements equal 1 with the size of $n \times 1$

W_k: Variable matrix principal component with the size of $n \times k$

θ_k: Vector of Principal Component Regression coefficient with the size of $k \times 1$

ε: Vector of error/residual with the size of $n \times 1$

In measuring the accuracy of the Principal Component Regression model, several significant tests had been performed on the regression coefficient, namely the Principal Component Regression coefficient test as a whole with the F-test and the Principal Component Regression coefficient test individually with the t-test. After obtaining the significant principal component variable to the dependent variable, an estimate of the Principal Component Regression model could be obtained. Then, the Principal Component Regression model could be transformed back into the original independent variable form to see the relationship or influence between the independent and dependent variables.

R Studio Software in Principal Component Regression Analysis

R studio is software related to computing and data processing for statistics (Chambers, 2008). R Studio is an integrated development environment (IDE) for R software which is a programming language for statistics and graphics. R Studio was founded by JJ Allaire, the creator of the ColdFusion programming language. R Studio is partly written in the C++ programming language. The development of R Studio was started in December 2010 and version 1.0 was released on November 1, 2016.

The steps for using R Studio in Principal Component Regression Analysis refer to the PCR blog at http://www.milanor.net. The summary of the steps is as follows:

1. Insert the packages to be used
2. Input the research data
   > data = read.csv ("data.csv")
3. Standardize the independent variables into standard form
   > Z = scale (data $ X)
4. Create a correlation matrix between standard independent variables
   > R = cast (Z)
5. Calculate the eigenvalues of the correlation matrix
   > eigen (R) $ values

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

6. Determine the principal component scores and principal components to be used
   > PCA = princomp (Z, scores = TRUE)

   > summary (PCA)
   > PCA $ loadings
   > Score = PCA $ scores [1: n,]
7. Perform the Principal Component Regression variables with the dependent variables
   > reg = cbind (Y, score)
   > regression = data.frame (reg)

   > regression

   > PCR = lm (Y ~ Score1 + Score2, regression)
   > summary (PCR)
   (Alice, 2016)

Research Data

The data used in this research were the data of rainfall, humidity, air temperature, air pressure, wind speed, and solar radiation in 15 districts/cities of Sulawesi in 2018. The data were obtained from the Badan Pusat Statistik (BPS) Sulawesi as presented in Appendix 1.

Based on the research variables in Appendix 1, the dependent variable could be determined, namely the rainfall(Y). The independent variables were humidity(X_1), air temperature(X_2), air pressure(X_3), wind speed(X_4), and solar radiation(X_5).

RESULTS AND DISCUSSION

Building the Principal Component Regression Model

In this section, a rainfall model had been built in Sulawesi, especially at the 15 studied points using the Principal Component Regression Analysis by performing the following steps:

Standardizing the Independent Variables

In this step, the original independent variables (X) were transformed into standardized independent variables (Z) using equation (10) because they had different measurement scales. Through the R Studio data processing, the standardized independent variables could be obtained as displayed in table 1.

Table 1. The Standardized Independent Variables

n	Z_1	Z_2	Z_3	Z_4	Z_5
1	0.78	-1.63	-1.52	-0.61	-0.32
2	-0.81	0.41	-0.16	-0.61	-0.01
3	0.78	-1.63	-1.52	-0.61	-0.32
⋮					
14	-1.35	2.10	1.20	0.69	-0.73
15	-0.81	0.41	0.68	0.69	-0.73

Establishing a Correlation Matrix between Standardized Independent Variables

In this step, the correlation matrix between the standardized independent variables(Z) was formed to see whether the multicollinearity problem present or not. The correlation between the five standardized independent variables was calculated using equation (3) and through R Studio data processing. The data can be seen in Table 2.

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

Table 2. The Correlation between Standardized Independent Variables

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	1.00				
Z_2	-0.36	1.00			
Z_3	-0.19	0.83	1.00		
Z_4	-0.46	0.44	0.47	1.00	
Z_5	0.22	0.18	0.35	0.08	1.00

Based on Table 2, it can be seen that there are a pair of variables, namely Z_2 and Z_3 with a correlation value of 0.83, Thus, it can be concluded that the data contained multicollinearity problem. This can cause the predicted value generated to be unable to predict the dependent variable precisely.

Determining Eigenvalues and Eigenvectors

In this step, the eigenvalues and eigenvectors were calculated as in equations (1) and (2). Through R Studio data processing, the obtained eigenvalues and eigenvectors are as follows:

Table 3. Eigenvalues and Eigenvectors of the Research Variables

	W_1	W_2	W_3	W_4	W_5
Z_1	-0.33	0.61	-0.21	0.67	0.16
Z_2	0.57	0.06	-0.48	-0.09	0.66
Z_3	0.56	0.26	-0.31	0.11	-0.71
Z_4	0.47	-0.23	0.59	0.60	0.13
Z_5	0.18	0.71	0.53	-0.41	0.10
Eigenvalues	2.449	1.312	0.635	0.465	0.138
Proportion	0.490	0.262	0.127	0.093	0.028
Commulative Proportion	0.490	0.752	0.879	0.972	1.000

Determining the Score of Principal Component Variables and Principal Component Variables to be Used

In this step, the principal component variable scores were calculated using the equation (9). Through R Studio data processing software, the scores of the principal component variables were obtained.

Table 4. The Principal Component Variable Scores

n	W_1	W_2	W_3	W_4	W_5
1	-2.38	-0.11	0.55	0.26	0.02
2	0.13	-0.38	-0.33	-0.96	0.18
3	-2.38	-0.11	0.55	0.26	0.02
⋮					
14	2.51	-1.05	-1.06	-0.23	0.34
15	1.08	-0.97	-0.21	0.22	-0.33

Based on Table 4, it can be seen that the eigenvalues were greater than one ($\lambda \geq 1$) on the first two principal components with eigenvalues of 2,449 and 1,312. Both principal components were able to explain 75.2% of the diversity of the entire original data. Therefore, the principal components used were W_1 and W_2.

Performing Principal Component Regression Variables Determined by the Dependent Variable

In this step, the Principal Component Regression model was built by estimating the Principal Component Regression

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

coefficients using the Least Square Method (6). Through R Studio data processing software, the Principal Component Regression coefficient (θ) can be obtained so that the Principal Component Regression model can be written as follows:

$$\hat{Y} = 1678.37 + 16.84\, W_1 + 66.36\, W_2$$

To find out whether there is a principal component variable (W) contributed to the dependent variable (Y), the overall regression coefficient test was carried out using the F test. Based on the Anova table and the R Studio data processing software, the results are as follows:

Table 6. Test Results of Regression Coefficient Variable for Principal Components W_1 and W_2 with Dependent Variables Y

$F_{observed}$ (F_h)	$F_{critical}$ (F_t)
4.04	3.89

Based on Table 6, it can be seen that $F_h > F_t$, then H_0 was rejected. It can be inferred that there was at least one principal component variable (W) contributed to the dependent variable (Y).

Then, to determine whether there was a contribution from each principal component variable (W) to the dependent variable (Y), an individual regression coefficient test was performed using the t-test. The results of the test are as follows:

Table 7. The Result of t-test for the Regression Coefficients Variableof Principal Components W_1 and W_2 on the Dependent Variables Y

	$t_{observed}$ (t_h)	$t_{critical}$ (t_b)
W_1	0.93	2.18
W_2	2.69	

Based on Table 7 it can be seen that $|(t_h)_1| < t_b$, then H_0 was accepted. Thus, it can be concluded that the principal component of the variable W_1 did not significantly influence the dependent variable Y. Then, it can be seen that $|(t_h)_2| > t_b$, then H_0 was rejected.it can be concluded that W_2 significantly influenced the dependent variable Y. Thus, since only the principal component variable W_2 contributed to the dependent variable Y, the estimated Principal Component Regression model are as follows:

$$\hat{Y} = 1678.37 + 66.36\, W_2$$

The principal component variable W_2 was the variable that represented the independent variable of the original data, so, the obtained Principal Component Regression model can be transformed back into the original variable because the principal component variable was standardized. The final model of the regression model can be denoted by:

$$\hat{Y} = -9977.02 + 21.53\, X_1 + 6.75\, X_2 + 9.03\, X_3 - 19.81\, X_4 + 10.61\, X_5$$

CONCLUSION

Based on the description, it can be concluded that a rainfall model in the Sulawesi region with local-scale factors on secondary data obtained from Badan Pusat Statistik (BPS) can be built using Principal Component Regression analysis assisted by R Studio software. The five original independent variables were reduced to two principal component variables which can explain 75.2% of the

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081

original data diversity and only one principal component variable that was significant to the dependent variable. Thus, a regression model has been obtained which shows the relationship between rainfall, air humidity, air temperature, air pressure, and solar radiation is unidirectional while the relationship between rainfall and wind speed is not unidirectional. The use of R Studio can simplify and speed up the calculations of the Principal Component Regression Analysis.

ACKNOWLEDGMENTS

The researchers would like to thank the Rector of Universitas Padjadjaran who has provided research funding through the Academic Leadership Grant of 2020 with contract number 1427/UN6.3.1/LT/ 2020 for the dissemination of the lecturers and students' research results.

REFERENCES

Alice, M. (2016). *Performing principal components regression (PCR) in R*. R User Group of Milano.

Chambers, J. (2008). *Software for data analysis: programming with R*. Springer.

Gorgees, H. M., & Ali, B. A. (2017). Employing ridge regression procedure to remedy the multicollinearity problem. *Ibn AL-Haitham Journal For Pure and Applied Science*, *26*(1), 320–327.

Handiana, D., Wahyono, S. C., & Susanti, D. S. (2016). Perancangan model prediksi curah hujan bulanan berdasarkan suhu permukaan laut di kalimantan selatan. *jurnal fisika flux: jurnal ilmiah fisika FMIPA Universitas Lambung Mangkurat*, *10*(1), 1–12.

Johnson, R. A., & Wichern, D. W. (2007). *Applied multivariate statistical analysis (6th Edition)*. Prentice Hall.

Jolliffe, I. T. (2010). *Principal component analysis*. springer.

Mariana. (2013). Analisis komponen utama. *matematika dan pembelajaran*, *1*(2), 189–204.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis*. Wiley.

Navid, M. A. I., & Niloy, N. H. (2018). Multiple linear regressions for predicting rainfall for Bangladesh. *Communications*, *6*(1), 1–4.

Rencher, A. C. (2002). *Methods of multivariate analysis (2nd Edition)*. John Wiley & Sons.

Sudrajat, A. (2016). Metode principal component analysis untuk mengatasi multikolinearitas pada regresi linear berganda (studi kasus faktor yang mempengaruhi indek pembangunan manusia di Jawa Timur). *Jurnal Penelitian Kesehatan*, *14*(4).

Untari, D. P., & Susanti, M. (2017). Latent root regression dalam mengatasi multikolinearitas. *Pythagoras: Jurnal Pendidikan Matematika*, *12*(1), 23–32.

Copyright © 2020, Desimal, Print ISSN: 2613-9073, Online ISSN: 2613-9081