Statistical Analysis of Rainfall With Principal Components Regression Analysis Technique

Nurul Aqilah Zulkipli, Muhammad Fauzee Hamdan*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: mfauzee@utm.my

Abstract
Rainfall is a significant meteorological phenomenon with an impact on local weather and global atmospheric circulation. Meteorological factors, such as temperature, humidity, wind, evaporation, and MSL pressure is the factor that influences rainfall. The study aimed to develop a rainfall model specific to a few regions in Pahang state and investigate the relationship between rainfall and meteorological factors in the regions. Rainfall and meteorological data for 11-year from 2010 to 2020 were used, recorded at four different rainfall stations in Pahang state. PCR analysis was used to process the data and obtain principal component variables that explained variability in the original data. Results showed that the two principal component variables derived from the PCR analysis can explain more than 80% of the variability present in the original data (Azapagic et al, 2018). The regression model analysis provided valuable insights into the dynamics of rainfall in the Pahang region and how meteorological factors contribute to rainfall patterns. The study demonstrated the usefulness of combining PCR analysis and R Studio software to understand rainfall patterns in the region.

Keywords: Rainfall of Pahang, Meteorological Factor, Principal Component Regression Analysis.

1.   Introduction

Freshwater is obtained through rainfall with a distribution that varies randomly over both time and location. According to Wang et al (2016), one type of weather and climate that significantly affects all aspects of human life is rainfall. The climate is a phenomenon that can change over space and time. South East Asia is experiencing climate change because of both global and local warming (Suhaila and Jemain, 2009, Tang et al., 2019). The impact of a changing climate on evapotranspiration and atmospheric water storage had an impact on rainfall quantity, frequency, and intensity. According to Meteorology Malaysia, the regular temperature, high humidity, and heavy rainfall make Malaysia's climate unique. The meteorological factor like Temperature, Humidity, Wind, Evaporation, and Air Pressure is most likely the factors that affect the intensity of rainfall. During the northeast monsoon period, the highest total rainfall received by Pahang Basins is about 40 percent annually (Nadiatul et al, 2016).

A few related past research papers have discussed statistical analysis that is widely used in modeling changes in the field of hydrology. Simple multiple regression was used by Rahman et al. (2013) to build a statistical forecasting strategy for rainfall during the summer monsoon across Bangladesh. The research papers presented suggest that the approaches used were not the best because they not take the multicollinearity issue into account. As a result, the models that were produced had multi-collinearity issues and needed to be modified.

Currently, numerous scientific works use analytical techniques to solve the multicollinearity problem. For instance, Hussein and Hytham (2021) used Ridge Regression to solve the multicollinearity

problem by running a Monte Carlo simulation. Principal Components Analysis had been used to solve the multicollinearity problem of big data collection across many fields, such as genomics and business intelligence (Chan et al., 2022). Jaeyong Lee and Hee-Seok Oh (2013) used Principal Components Analysis to solve the multicollinearity problem in climate prediction problem. The purpose of this work is to use principal component regression to build a rainfall model to find out the rainfall relationship with meteorological variables in the Pahang region with the help of R Studio software.

2.0   Principal Component Regression Analysis's Conceptual Principles

One of the basic algebraic concepts utilized in principal component analysis is the concept of eigenvalues and eigenvectors. According to Johnson & Wichern (2007), matrix A eigenvalues are $n \times n$, denoted by:

$$\mathrm{Det}(\lambda I - A) = 0 \tag{1}$$

$$Av = \lambda v \tag{2}$$

where,

Det   : Determinant,
λ : The eigenvalue,
I : Identity matrix, and
v : The eigenvector.

The concept of correlation should be used to portray the linear relationship between two or more quantitative variables. Standardized covariance is used to measure correlation. Multicollinearity could occur if there is a significant enough correlation between the independent variables. If there is multicollinearity, a variable in the regression model that has a high correlation with other variables may have a variable predictive value that is inconsistent and unstable (Vatcheva,2016).

Correlation measures the strength of a linear relationship between two variables. One numerical measure is the Pearson product-moment correlation coefficient, r (Metsämuuronen, 2022).

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{(n\sum X^2 (\sum X)^2) \times (n\sum Y^2 (Y)^2)}} \tag{3}$$

where,

n : The number of data,
$\sum XY$: The sum of the product of the x-value and y-value for each point in the data set,
$\sum X$: The sum of the x-values in the data set,
$\sum Y$: The sum of the y-values in the data set,
$\sum X^2$: The sum of the squares of the x-values in the data set, and
$\sum Y^2$: The sum of the squares of the y-values in the data set.

Properties of r:
   a) $-1 \le r \le 1$: The correlation coefficient always ranges between -1 and 1, inclusive. A value of -1 represents a perfect negative linear relationship, 1 represents a perfect positive linear relationship, and 0 represents no linear relationship between the variables.
   b)   When r is close to 1, it indicates a strong positive linear relationship between the variables x and y. As x increases, y tends to increase as well.
   c)   When r is close to -1, it indicates a strong negative linear relationship between the variables x and y. As x increases, y tends to decrease.
   d)   When r is close to 0, it suggests little or no linear relationship between the variables x and y. The variables may still have a nonlinear relationship or may be independent of each other.

2.1   Multiple Linear Regression Analysis

A statistical method for examining, analyzing, and forecasting correlations between variables is regression analysis. The connection between a model's dependent variable, Y, and one or more independent variables, X_i is expressed using multiple regressions (Montgomery et al., 2012). The following equation can be used to represent multiple regression models:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon \tag{4}$$

where Y is the rainfall (dependent variable) X_1, X_2, X_3, X_4, and X_5 (independent variables) are temperature, humidity, wind, evaporation, and MSL pressure respectively. $\beta_1, \beta_2, \beta_3, \beta_4$, and β_5 are model coefficients of the five independent variables. β_0 is a constant while ε is the error.

In general, the equation for the regression model can be portrayed in matrix notation as follows:

$$Y = X\beta + \varepsilon \tag{5}$$

If the data does not contain multicollinearity, the Least Square Method can be used to estimate the β parameter. It is possible to compute it as follows:

$$\hat{\beta} = (X'X)^{-1} X'Y \tag{6}$$

3.   Principal Component Analysis

Karl Pearson discovered Principal Component Analysis (PCA) in 1901, and Harold Hotelling named it in 1933 (Martin., 2008). According to (Khan et al., 2021), PCA is particularly beneficial for overcoming collinearity in linear regression by merging explanatory factors into a smaller collection of uncorrelated variables. The basis for PCA can be either a covariance matrix or a correlation matrix. The principal component is described as a combination of the p original independent variables represented as a matrix in the following way if α is an orthogonal matrix of $p \times p$

$$W = \alpha' X \tag{7}$$

$$\begin{pmatrix} W_1 \\ W_2 \\ \vdots \\ W_p \end{pmatrix} = \begin{pmatrix} \alpha_{11} & \alpha_{21} & \cdots & \alpha_{p1} \\ \alpha_{12} & \alpha_{22} & \cdots & \alpha_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1p} & \alpha_{2p} & \cdots & \alpha_{pp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix} \tag{8}$$

where,
α: The eigenvector matrix of $p \times p$, and
X: The original variable vector of $p \times 1$.

In the form of a linear combination, it can be notated as:

$$W_j = \alpha_j' X = \alpha_{1j} X_1 + \alpha_{2j} X_2 + \ldots + \alpha_{pj} X_p ; \qquad j = 1,2,3,\ldots,p \tag{9}$$

The original variable's p is converted into a standard score, also known as standardisation, when it is measured in different units. The original variable X can be standardised into the Z score using the formula below:

$$Z = \frac{X - \mu}{\sigma} \tag{10}$$

where,
Z: is the standardized value (Z-score) of the variable X,
X: is the original value of the variable,
μ: is the mean of the variable X, and
σ: is the standard deviation of the variable X.

Using principal components with more than one eigenvalue is required by the principal component analysis using a correlation matrix requirement $((\lambda_j \geq 1)$. The cumulative percentage variation of the primary component serves as a measure for the overall data variance of the independent variable

3.1   Principal Component Regression Analysis

Principal component regression analysis, according to Abdullah et al. (2016), combines principal component analysis with regression analysis, with the principal component analysis operating as the analytical step. Many principal components are selected to be used as independent variables in the principal component regression analysis by estimating the regression coefficient using the Least Square Method.

There are two techniques to produce principal component regression using principal component regression analysis and a covariance or correlation matrix (Jolliffe, 2010). Depending on the range of observations for the independent variable, one of the two techniques is used.

Let the matrix A is an orthogonal matrix $p \times p$ with $A'A = AA' = I$ where $W = XA$. The Principal Component Regression method is then used to convert the multiple linear regression equations method.

$$Y = X\beta + \varepsilon$$

$$Y = XAA'\beta + \varepsilon \qquad (11)$$

$$Y = W\theta + \varepsilon$$

Moreover, where θ is the regression parameter's vector and $\theta = A'B$. The simplified Principal Component Regression model to k principal components is as described in the following:

$$Y = \theta_0 1 + W_k\theta_k + \varepsilon \qquad (12)$$

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 $k \times 1$, and
ε : Vector of error/residual with the size of $n \times 1$.

A few significant tests on the regression coefficient were run in order to analyze the principal component regression model's dependability. Both the principal component regression coefficient test and the principal component regression coefficient test were included in these tests. One may acquire an estimate of the principal component regression model after identifying the significant principal component variable for the dependent variable. The principal component regression model can then be converted to its original independent variable form to investigate the relationship or connection between the independent and dependent variables.

4.   Software in Principal Component Regression Analysis

By utilising the R studio programme, we may reduce and simplify the calculations for Principal Component Regression Analysis. Software for statistical computing and data processing is called R studio (Chambers, 2010). A statistical and graphical programming language called R. R Studio is an integrated development environment (IDE). The creator of the ColdFusion programming language, JJ Allaire, introduced R Studio. Programming in C++ was used in part to construct R Studio.

4.1   Experimental Results

Dataset
All the data information was gathered from the Malaysia Meteorology Service (MMS). The temperature, humidity, wind, evaporation, and MSL pressure are all considered independent variables in this problem. The dependent variable will be rainfall. This study will be examined based on daily observations of the independent variables in Kuantan, Temerloh, Rompin, and Jerantut, Pahang. The rainfall data for the 11 years between 2010 to 2020 were used to create the model. The locations in issue have been found to be flooded zones.
The stations are:
    i.   Station Batu Embun
    ii.   Station Muadzam Shah
    iii. Station Temerloh
    iv.   Station Kuantan

Principal Component Regression Analysis
The dataset of the dependent variable, namely the rainfall (Y) and the independent variables were air temperature (X_1), humidity (X_2), wind speed (X_3), evaporation (X_4), and air pressure (X_5), was gathered to evaluate the relationship between it using the principal component regression method. Table 1 displays a few average daily data for 365 days from Batu Embun station from 2010-2020 that will be considered:

Table 1: Dataset of daily rainfall and meteorological variables in Batu Embun

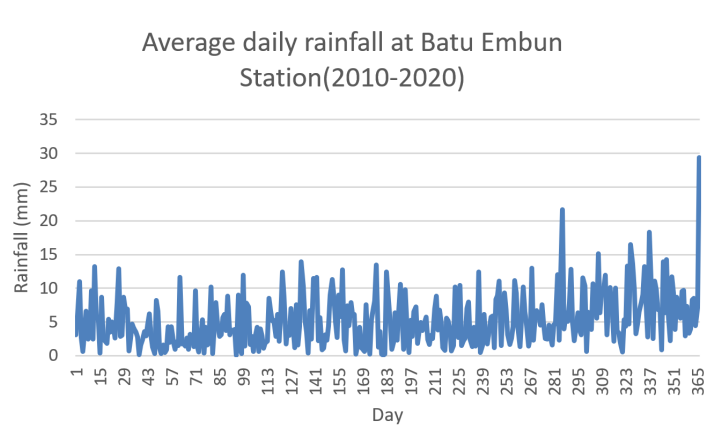| Day	Temperature (° C)	Relative Humidity (%)	Rainfall (mm)	Wind (m/s)	Evaporation (mm)	MSL Pressure (Hpa)
1	25.6	88.2	3.1	0.5	1.8	1011.0
2	25.9	88.5	5.8	0.6	2.2	1011.1
3	25.8	88.8	10.9	0.5	2.1	1011.0
4	26.3	86.5	2.7	0.5	2.2	1010.2
5	26.4	86.0	0.7	0.5	2.5	1010.1
6	26.6	86.0	3.3	0.5	2.8	1010.2
7	26.2	86.5	6.6	0.6	2.5	1010.6
⋮	⋮	⋮	⋮	⋮	⋮	⋮
365	25.6	87.4	7.1	0.6	2.2	1010.9

Figure 1 An average daily rainfall for four stations from 2010-2020

Standardizing the Independent Variables
Since the original independent variables (X) had multiple measurement scales, they were turned into standardised independent variables (Z) in this step. Standardizing the variables is useful in cases where the original variables have different measurement scales or units, as it allows for easier comparison and interpretation of the variables' effects in a statistical analysis. As shown in Table 2, the standardised independent variables may be derived.

Table 2: The Standardized Independent Variables (Batu Embun station)

n	Z_1	Z_2	Z_3	Z_4	Z_5
1	-1.932124459	1.463976709	-1.541806861	-2.725491499	1.475069521
2	-1.653164672	1.568606289	-1.186483994	-1.999650123	1.563734123
3	-1.689550731	1.683361312	-1.778688772	-2.138641025	1.448511105
4	-1.073682991	0.821948605	-1.554966967	-1.999650123	0.457177687
⋮	⋮	⋮	⋮	⋮	⋮

Correlation Matrix Development on Standardized Independent Variables
To determine if the multicollinearity issue was present or not, the correlation matrix between the standardised independent variables (Z) was created in this stage. Equation (3) and data processing in R Studio were used to calculate the correlation between the five standardised independent variables. The information is shown in Table 3.

Table 3: The Correlation between Standardized Independent Variables (Batu Embun station)

	Z_1	Z_2	Z_3	Z_4	Z_5
Z_1	1.0000000	-0.7252289	0.5905187	0.7269760	-0.5762860
Z_2	-0.7252289	1.0000000	-0.7256199	-0.7621025	-0.0299018
Z_3	0.5905187	-0.7256199	1.0000000	0.6998848	-0.0764204
Z_4	0.7269760	-0.7621025	0.6998848	1.0000000	-0.1951773
Z_5	-0.5762860	-0.0299018	-0.0764204	-0.1951773	1.0000000

Based on Table 3, the correlation coefficients indicate that there are some moderate to high correlations between the variables. Specifically, temperature is highly correlated with both humidity and evaporation. Additionally, humidity and evaporation have a high negative correlation. These correlations coefficient suggest the presence of multicollinearity between temperature, humidity, and evaporation. As a result, the dependent variable may not be accurately predicted by the predicted value that is generated.

Multicollinearity can pose challenges in regression analysis, such as inflated standard errors, unstable coefficient estimates, and difficulties in interpreting the individual effects of variables. It is recommended to address multicollinearity before performing regression analysis by applying techniques such as principal component analysis (PCA).

Obtaining Eigenvalues and Eigenvectors
Equations (1) and (2) were used to calculate the eigenvalues and eigenvectors. Following is a list of the eigenvalues and eigenvectors that were acquired by data processing in R Studio:

Table 4: Eigenvalues and Eigenvectors of the Research Variables at Batu Embun Station

	W_1	W_2	W_3	W_4	W_5
Z_1	0.50352	0.30973	0.28732	-0.30730	0.68815
Z_2	-0.49238	0.30967	-0.38818	0.42669	0.57351
Z_3	0.46522	-0.24797	-0.83744	-0.12967	0.06295
Z_4	0.50282	-0.09596	0.17588	0.84054	-0.02281
Z_5	-0.18647	-0.85877	0.18581	-0.01345	0.43937
Eigenvalues	3.19418	1.17475	0.33101	0.24250	0.05756
Proportion	0.63884	0.23495	0.06620	0.04850	0.01151
Cumulative Proportion	0.63884	0.87379	0.93999	0.98849	1.00000

Based on Table 4, the eigenvalues were greater than one ($\lambda \geq 1$) on the first two principal components with eigenvalues of 3.19418 and 1.17475. Both eigenvalues are greater than one, indicating that these principal components contribute significantly to the overall variance in the data. Both principal components were able to explain 87.3% of the diversity of the entire original data. This suggests that these two principal components capture a substantial portion of the dataset's information. Therefore, the principal components used were W_1 and W_2.
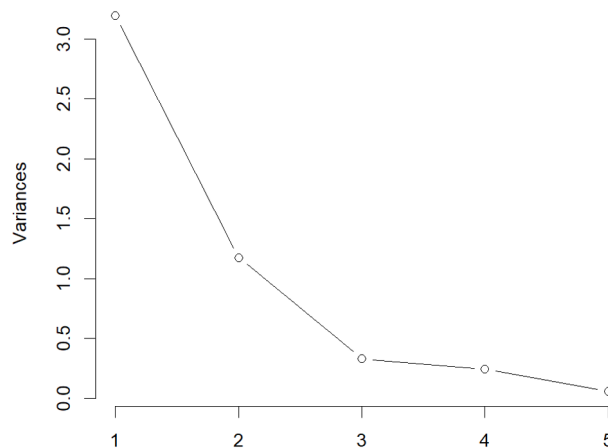

Figure 2   Scree plot for Rainfall Features for four stations from 2010-2020

To create a scree plot, the provided eigenvalues from Table 4 and their corresponding principal components were use (Eigenvalues: [3.19418, 1.17475, 0.33101, 0.24250, 0.05756]). Based on the scree plot from Figure 2, it is common practice to select the number of principal components where the plot levels off or exhibits an "elbow" (Ahmed et al, 2023). In this case, the "elbow" appears at the third PC, therefore the first two PCs should be retained which account for 87.3% of the total variance, as a valuable reduction in dimensionality. Therefore, it is a fact that the first two PCs accounted for 87.3% of total variance of the original variables and simultaneously reduce the data dimension from 5 to 2.

Conduct Principal Component Regression Considering the Dependent Variable
To generate the Principal Component Regression model equation, this step involved estimating the Principal Component Regression coefficients using the Least Square Method (6). Through R Studio data processing software, the Principal Component Regression coefficient (θ) can be obtained as Table 5:

Table 5: Coefficient of Batu Embun (PC1& PC2)

Station	Variables	Coefficients	Standard error	*t*-value	*p*-value
		Coefficients			
Batu Embun	Constant	5.3326	0.2317	23.014	2.00E-16
	W_1	-0.5731	0.1295	-4.424	1.40E-05
	W_2	-0.5392	0.2139	-2.521	0.0123

The data processing programme R Studio can be used to generate the Principal Component Regression coefficient (*θ*), which allows the Principal Component Regression model to be represented as follows:

$$\hat{Y} = 5.3326 - 0.5731W_1 - 0.5392W_2 \text{ (Batu Embun)}$$

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. In the coefficient test, each variable's estimate, standard error, *t*-value, and *p*-value are examined to determine the significance of the coefficients. The null hypothesis for each coefficient test is that the corresponding coefficient is equal to zero.

If the *p*-value for the constant term or any of the variables is statistically significant typically with a significance level of 0.05, it indicates that the variable in the model significantly contributes to the dependent variable. From the Table 5, for the (Constant) term, the estimate is 5.3326 with a standard error of 0.2317. The calculated *t*-value 23.014, and the *p*-value is less than 2e-16, indicating strong evidence to reject the null hypothesis. Therefore, the (Constant) term is statistically significant.

Based on the coefficient test results, both W_1 and W_2 are statistically significant predictors in the linear regression model. Thus, since the principal component variable W_1 and W_2 contributed to the dependent variable Y, the estimated Principal Component Regression model are as follows:

$$\hat{Y} = 5.3326 - 0.5731W_1 - 0.5392W_2 \text{ (Batu Embun)}$$

Because the principal component variable was standardised, the derived principal component regression model can be changed back into the original variable. The variable that represented the independent variable of the data was the principal component variable W_1 and W_2. However, the model only explains 80% amount of the variation in the response variable, which may limit its practical usefulness. The final model of the regression model can be denoted by:

$$\hat{Y} = 355.6126 - 1.7594X_1 + 1.2561X_2 + 5.025X_3 + 5.2091X_4 - 0.4254X_5 \text{ (Batu Embun)}$$

A regression model at Batu Embun station obtained shows the relationship between $\hat{Y}$, X_2, X_3, and X_4 are unidirectional while the relationship between $\hat{Y}$ with X_1 and X_5 is not unidirectional. As the value of X_1 and X_5 increases, the predicted value of $\hat{Y}$ decreases.

Evaluation of Rainfall Prediction Performance.

Table 6: Evaluation of Rainfall Prediction Performance: MSE and RMSE Values

Station	MSE	RMSE
Batu Embun	13.965	3.737

The RMSE is a measure of the difference between the predicted values and the actual values. Lower values of RMSE indicate better predictive performance, as they reflect smaller errors between predictions and actual values. From Table 6, the RMSE values for Batu Embun, Muadzam Shah, and Temerloh stations are 3.737, 3.537, and 3.280, respectively. These values indicate the average difference between the model's predicted values and the actual rainfall values for each station. The

RMSE values for all three stations are relatively low, which suggests that the model's predictions are relatively accurate. The RMSE of 3.737 for Batu Embun station implies that, on average, the model's predictions differ from the actual rainfall values by approximately 3.7 units. Similarly, the RMSE of 3.537 for Muadzam Shah station suggests an average difference of approximately 3.5 units, and the RMSE of 3.280 for Temerloh station suggests an average difference of approximately 3.3 units.

Overall, with these relatively low RMSE values, it can be inferred that the model's predictions are generally close to the actual rainfall values for these stations, indicating relatively good predictive performance.
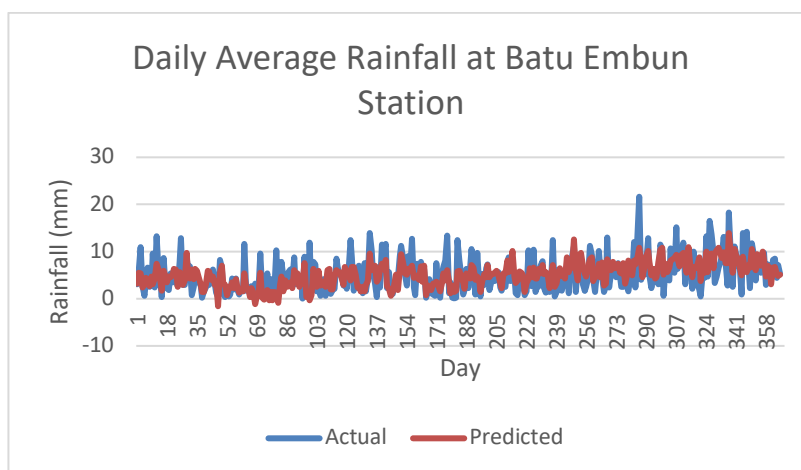

Figure 3   An average daily rainfall for four stations from 2010-2020

Conclusion

It can be concluded that a rainfall model for Pahang can be developed using PCR analysis and meteorological data from the Department of Meteorology Malaysia with the help of R Studio software. The PCR approach reduces the dimensionality of the meteorological variables and extracts the most significant components (PC1 and PC2) that account for more than 80% of the data variability. For Batu Embun and Muadzam Shah stations, all meteorological variables (temperature, humidity, wind, evaporation, and air pressure) showed statistically significant relationships with rainfall, while at Temerloh station, only wind and evaporation had a significant relationship. The regression coefficients for these variables were negative or positive, indicating the direction and magnitude of their influence on rainfall. The regression equations may differ between stations, reflecting the specific characteristics and localized influences present at each location. These differences in the regression equations highlight the importance of considering the unique dynamics of each station when studying the relationship between meteorological variables and rainfall within the state. R Studio can speed up and simplify the calculations for Principal Component Regression analysis.

References

[1]   Abdullah, S., Ismail, M., Fong, S. Y. and Ahmed, A. N. (2016) 'Evaluation for long term PM10 concentration forecasting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) Models', *EnvironmentAsia,* 9(2), pp. 101-110.
[2]   Azapagic, et al. (2018) 'A principal component analysis based three-dimensional sustainability assessment model to evaluate corporate sustainable performance', *Journal of Cleaner Production.*
[3]   Chambers, J. M. (2010) 'Software for Data Analysis: Programming with R'*, Springer Science+Business Media.*

[4] Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W. and Chen, Y.-L. (2022) 'Mitigating the multicollinearity problem and its Machine Learning Approach: A Review', *MDPI*.

[5] Hussein E. and Hytham H. A. (2021) 'A Monte Carlo Study for Dealing with Multicollinearity and Autocorrelation Problems in Linear Regression Using Two Stage Ridge Regression Method', *Mathematics and Statistics*, 9(5), pp. 630–638.

[6] Johnson, R. A. and Wichern, D. W. (2007) 'Applied Multivariate Statistical Analysis', *Prentice-Hall Of India*.

[7] Khan, M. S. I., Islam, N., Uddin, J., Nasir, M. K. and Islam, S. (2021) 'Water quality prediction and classification based on principal component regression and gradient boosting classifier approach', *ScienceGate: Journal of King Saud University - Computer and Information Sciences*.

[8] Lee, J. and Oh, H. -S. (2013) 'Bayesian regression based on principal components for high-dimensional data', *Journal of Multivariate Analysis*.

[9] Metsämuuronen, J. (2022). Directional nature of the product–moment correlation coefficient and some consequences. *Frontiers in Psychology*, 13.

[10] Myers, R. H., Montgomery, D. C., Vining, G. G. and Robinson, T. J. (2012) 'Generalized linear models: with applications in engineering and the sciences', *John Wiley and Sons Inc.*

[11] Nadiatul, A. A. A. G. Mohamad, N. A. and Hui, T. W. (2016) 'Rainfall Analysis to Determine the Potential of Rainwater Harvesting Site in Kuantan, Pahang', *Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences*, 11(11), pp. 7264–7268..

[12] Rahman, M. M., Rafiuddin, M. and Alam, M. M. (2013) 'Seasonal forecasting of Bangladesh summer monsoon rainfall using simple multiple regression model', *Journal of Earth System Science*, 122(2), pp. 551–558.

[13] Suhaila, J., Deni, S.M., Zin, W. Z. W. and Jemain, A. A. (2010) 'Trends in Peninsular Malaysia. Rainfall Data During the Southwest Monsoon and Northeast Monsoon Seasons: 1975–2004', *Sains Malaysiana*, 39(4), pp. 533-54

[14] Tang, K. H. (2019) 'Climate change in Malaysia: Trends, contributors, impacts, mitigation and adaptations', *Science of The Total Environment*, 650, pp. 1858–1871.

[15] Vatcheva, K. P. Lee, M. McCormick, J. B. and Rahbar, M. H. (2016) 'Multicollinearity in regression analyses conducted in epidemiologic studies', *OMICS International*.

[16] Wang, X. -J., Zhang, J. -Y., Ali, M. Shahid, S., He, R. -M., Xia X. -H. and Jiang, Z. (2016) 'Impact of climate change on regional irrigation water demand in Baojixia irrigation district of China', *Mitigation and Adaptation Strategies for Global Change*, 21, pp. 233–247.