Journal of Engineering and Science Research 1 (2): 164-170, 2017
e-ISSN: 2289-7127
© RMP Publications, 2017
DOI: 10.26666/rmp.jesr.2017.2.25

eISSN (online) : 2289-7127
Journal of Engineering and Science Research

Modeling of Annual Air Temperature in Malaysia Using Multiple Regression Model

Nur Hanim Mohd Salleh and Husna Hasan
School of Mathematical Sciences, Universiti Sains Malaysia,
11800 USM, Pulau Pinang, Malaysia.

Abstract: Annual air temperature data obtained from twenty-two meteorological stations across Malaysia are modeled using multiple regression. A correlation test was conducted to find statistical relationship between each of the dependent variables: annual maximum and annual average air temperature and predictor variables: longitude, latitude, elevation and wind speed. Regression models using least square estimation method were developed relating the dependent variables to independent variables and the adequacy of the models is determined by the coefficient of determination. The result shows that the longitude and wind speed factors have a significant influence on the annual air temperature in Malaysia.

Key words: *air temperature, correlation, multiple regression, least square method, climatic variable, climate modeling*

INTRODUCTION

Spatial information on climatological data has been widely used by many researchers in various disciplines as a basis for understanding the process of their study. The information on climatic variable such as air temperature and total precipitation are usually observed at the local meteorological stations. One of the methods used to compute the distribution of climatic variable from discrete data is multiple regression analysis (see [1] and [2]). Multiple regression fits a relationship between a dependent variable and several independent variables. The parameters in multiple regression models can be estimated using the least square method [3] and maximum likelihood estimator [4]. The building of multiple regression involves selection of predictor variables and also the use of diagnostic tools to assess the validity of the model.

The motivation of this study arises from the study by Lado et. al [1] in which they assessed the influence of geographical factors such as nominal altitude (ALT), latitude (LAT) and longitude (LON) of the meteorological stations to the air temperature in State of São Paulo (Brazil) using multiple regression analysis. From the analysis, they found that the multiple regression analysis was a suitable method for modeling of the air temperature for the State of São Paulo with the coefficients of determination varied from 0.924 to 0.953. Only two significant geographical variables contributed to the prediction of the values of air temperature in the studied area which were ALT and LAT. The influence of geographical factors on the air temperature was also investigated by Ninyerola et al. [2] using the multiple regression analysis. In their study, three additional factors were included that are continentality (CON), solar radiation (RAD) and cloudiness factor (CLO). The result of the study indicated that geographical factors that are ALT and LAT did influence the air temperature in Catalonia, northeast of Spain.

Thus, in this study, we aim to identify the influence of three geographical factors such as Longitude (LON), Elevation (ELE), Latitude (LAT) and an additional factor of wind speed (WDSP) factor on the air temperature in Malaysia. Malaysia is divided into two parts, West Malaysia and East Malaysia. West Malaysia is a peninsular with Thailand sitting at the

Corresponding Author: Nur Hanim Mohd Salleh, School of Mathematical Sciences, Universiti Sains Malaysia, 11800 USM, Pulau Pinang, Malaysia, E-mail: nurhanim_sh@yahoo.com, Tel: +60174005636

North and Singapore as its southern neighbor while East Malaysia consists of Sabah and Sarawak. Throughout the year, the climate in Malaysia is warm and humid with the air temperatures are around 30°C during the day and 22°C at night [5]. Malaysia is located near the Equator between latitude 1° and 7° North and the longitude 100° and 119° East [6]. This country has two monsoon seasons; south-west and north-east which originate from Indian Ocean and South China Sea.

DATA AND VARIABLES

Annual maximum daily temperatures data ranging from 1981 to 2012, for twenty-two meteorological stations in Malaysia is obtained from the National Climatic Data Center website. The Bayan Lepas, Butterworth, KLIA, Kota Bharu, Kuantan, Langkawi, Malacca, Mersing, SAAS, SAS, Senai, Sitiawan, and Subang stations are located in Peninsular Malaysia while Bintulu, Kota Kinabalu, Kuching, Kudat, Labuan, Miri, Sandakan, Sibu, and Tawau stations are located in Sabah and Sarawak. The five variables considered in this study are temperature, LON, ELE, LAT and WDSP as shown in Table 1. LON is the angular distance between the Prime Meridian and points east or west of it on the surface of the Earth while LAT is the angular distance between the Equator and points north or south of it on the surface of the Earth [7]. Elevation is defined as the distance above sea level and usually measured in meters or feet.

Table 1: List of Variables

Temperature	Unit
Temperature	Fahrenheit
Longitude (LON)	Thousandths of decimal degrees
Latitude (LAT)	Thousandths of decimal degrees
Elevation (ELE)	Tenths of meters
Wind speed (WDSP)	Knots

RESEARCH METHODOLOGY

A relationship between the dependent variable Y and some predictor variables $X_{i1}, X_{i2}, \ldots, X_{i,p-1}$ in the multiple linear regression model can be presented by a function [8] as follow:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i \qquad i = 1, 2, \ldots, n \tag{1}$$

where $\beta_0, \beta_1, \beta_2, ..., \beta_{p-1}$ are unknown parameter, $X_{i1}, X_{i2}, \ldots, X_{i,p-1}$ are predictor variables, ε_i is the independent errors which is normally distributed with min value zero and variance σ^2. The response function for the equation (1) can be written as follow:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \ldots + \beta_{p-1} X_{p-1} \tag{2}$$

The estimation methods that can be used to estimate the parameter in a regression model are least square estimation (LSE) method. LSE method considers a deviation of Y_i from an expected value which is the error, ε_i.

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i + \ldots + \beta_{p-1} X_{p-1}) \tag{3}$$

where p is the number of parameters in the model. Each of deviations is squared and the sum of squared of all deviations can be represented by Q:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i - \ldots - \beta_{p-1} X_{p-1})^2 \tag{4}$$

The value of β_0 and β_1 are estimated by minimizing the value of Q.

The predictor variable selection procedures are carried out to identify the best subset of the predictor variables to be fitted in the regression model [9] by using automated procedure such as the enter method. Diagnostic analysis is conducted to determine nonlinearity, examine multicollinearity [10], and identify non-constant error variance and normality of residuals [11]. The scatter plot and the residual plot are used to determine the existence of nonlinearity in the dataset. If the plots are linearly distributed, it indicates that the model is linear. Multicollinearity among the predictor variables is examined using the variance inflation factor (VIF) which indicating the existence of multicollinearity when the VIF value is greater than 10.

A residual plot is plotted to identify the non-constant variance. The fan-shaped residual plot specifies that the regression model has non-constant variance. Normality of residuals is determined by using a normality test such as *Shapiro-Wilk* and *Kolmogorov-Smirnov.* The adequacy of the regression model is determined by the coefficient of determination, R^2:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \qquad (5)$$

where SSR is the sum of squared regression and SST is the total sum of squares. The SSE is sum of squared error and R^2 assumes values between 0 and 1. A higher value of R^2 indicates the appropriateness of a regression model to represent the data.

DATA ANALYSIS AND RESEARCH FINDINGS

The analysis of this study was divided into two parts. The first part of the study considers the mean values of the climatic variables that are annual average temperatures and annual maximum temperatures as the dependent variables in the multiple regression analysis while the mean value of annual average wind speed (WDSP) and three geographical factors: longitude (LON), latitude (LAT) and elevation (ELE) of the meteorological stations are considered as predictor variables. In the second part of the study, the analysis is focused on the exact value of annual maximum temperature as the dependent variable. For the predictor variables, the three geographical factors and the value of wind speed correspond to the value of annual maximum temperature are considered as the predictor variables.

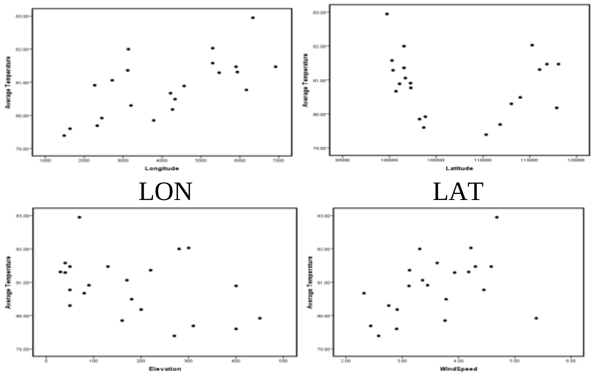
Part I: Analysis for Mean of Annual Average Temperature

A correlation test is conducted to determine the correlation between the mean of annual average temperature and LON, LAT, ELE and WDSP respectively. From the correlation test in Table 2, it is found that LON, ELE and WDSP have significant correlation with the mean of annual average temperature while LAT does not correlate with the mean of annual average temperature. LON and WDSP have positive correlation with the mean of annual average temperature. Thus, when LON and WDSP are increasing, the mean of annual average air temperature also will increase. In contrast, ELE has a negative correlation with the mean of annual average temperature.

The relationships can also be observed from the scatter plots of mean of annual average temperature as in Figure 1 for each of the independent variables. Based on the plots, the relationship between the mean of annual average temperature and LON, ELE, and WDSP form a straight line which can be represented by a linear distribution. LON and WDSP show a positive linear relationship while ELE shows a negative linear relationship.

Table 2: Correlation Test for Mean of Annual Average Temperature

Variables	Correlation Coefficient, r	p-value
LON	0.657	0.001
LAT	-0.140	0.535
ELE	-0.401	0.065
WDSP	0.442	0.040



LON LAT

ELE WDSP

Fig 1 Scatter Plot for Mean of Annual Average Temperature

Consequently, a multiple regression analysis is conducted to investigate the relationship between the mean of annual average temperature and all four predictor variables. A variable selection procedure using enter method is performed to determine the important predictor variables which influence the annual average temperature. The selection procedure shows that the only one significant variable is LON as stated in Table 3. Next, the linear regression model with one significant predictor variable, LON. From Table 4, the results show that LON variable is significant.

The suitability of the linear regression model with one significant variable is examined by performing a diagnostic on the model. It is found that the standardized residual is normally distributed as stated in Table 5 and the residual plot as shown in Figure 2 shows constant error variance. The multicollinearity among the variables does not exist (VIF value less than 10) since there is only one variable fitted in the regression model. The coefficient of determination value, R^2 of the model is 0.432. Although the scatter plot of LON versus mean of annual average temperature demonstrates a linear relationship, based the diagnostic results, it can be concluded that a linear regression model is not suitable to explain the relationship between the mean of annual average temperature and LON since the R^2 value is low. However, it is important to note that the outcomes of correlation test and regression analysis indicate that the LON variable does influence the mean of annual average temperature.

Table 3: Multiple Regression Analysis for Mean of Annual Average Temperature with Four Variables

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.026	2.580
LAT	-0.0004	0.000	0.121	1.068
ELE	0.000	0.001	0.937	1.706
WDSP	0.092	0.240	0.707	1.637

Table 4: Linear Regression Analysis for Mean of Annual Average Temperature with One Variable

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.001	1.000

Table 5: Normality Test for Linear Regression Model with One Variable

Variables	Statistic value	df	p-value
Standardized Residual	0.908	22	0.044

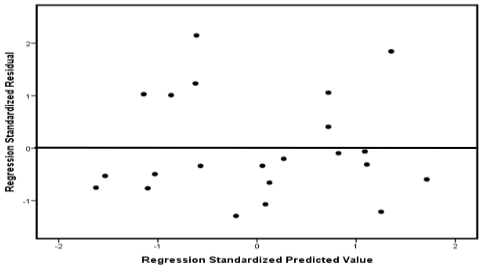


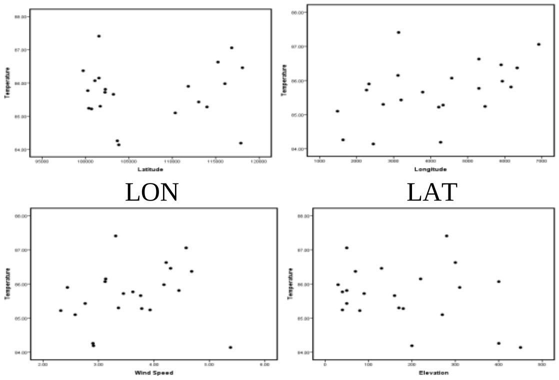
Fig 2 Residual Plot for Linear Regression Model with One Variable

Part I: Analysis for Mean of Annual Maximum Temperature

With regard to the importance of comparing the result of average temperature and extreme temperature, the above procedure is repeated again to see the existence of relationship between all predictor variables and mean of annual maximum temperature. The result of correlation test between LON, LAT, ELE, and WDSP and mean of annual maximum temperature are shown in Table 6. Only the LON variable appears to have a positive correlation with the mean of annual maximum temperature. So, when the LON increases, the mean of annual maximum temperature also increases. To observe visibly the relationship between the LON, LAT, ELE, and WDSP and mean of annual maximum temperature, the scatter plot for each of the independent variables are plotted as shown in Figure 3. Only LON shows a positive linear relationship with the mean of annual maximum temperature.

Table 6: Correlation Test for Mean of Annual Maximum Temperature

Variables	Correlation Coefficient, r	p-value
LON	0.475	0.025
LAT	0.011	0.960
ELE	-0.140	0.534
WDSP	0.201	0.371



LON LAT

ELE WDSP

Fig 3 Scatter Plot for Mean of Annual Maximum Temperature

Subsequently, to examine the influence of all predictor variables towards mean of annual maximum temperature, a multiple regression analysis is conducted. The result of regression analysis in Table 7 indicates that only LON variable has a significant influence to the mean of annual maximum temperature. Next, the LON variable is fitted again into a regression model and the result of regression analysis shows that the LON variable is significant as stated in Table 8. The model diagnostics are conducted to examine the adequacy of a linear regression model with one significant variable to explain the relationship between predictor variable and dependent variable. From the result in Table 9, the normality assumption is met as the standardized residual is normally distributed. The error variance is constant as shown in Figure 4 and the value of the coefficient of determination, R^2 of the model is 0.226.

Table 7: Multiple Regression Analysis for Mean of Annual Maximum Temperature with Four Variables

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.098	2.580
LAT	0.000	0.000	0.827	1.068
ELE	0.000	0.002	0.805	1.706
WDSP	-0.131	0.285	0.653	1.637

Table 8: Linear Regression Analysis for Mean of Annual Maximum Temperature with One Variable

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.025	1.000

Table 9: Normality Test for Linear Regression Model with One Variable

Variables	Statistic value	df	p-value
Standardized Residual	0.966	22	0.627

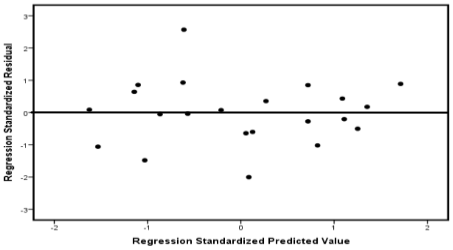


Fig 4 Residual Plot for Regression Model with One Variable

Part II: Analysis for Annual Maximum Temperature

Due to unsatisfactory result of analysis for mean of annual maximum temperature, the daily temperatures at each meteorological station are blocked into annual maximum temperature and the values of wind speed correspond to the value of annual maximum are identified. Table 10 shows the result of correlation test between LON, LAT, ELE, and WDSP and annual maximum temperature. There are three variables which show a significant correlation with the annual maximum temperature. LON and WDSP correlate weakly and positively with the annual maximum temperature. As LON and WDSP increase, the annual maximum temperature also increases. On the other hand, ELE demonstrates a weak and negative correlation with the annual maximum temperature.

Table 10: Correlation Test for Annual Maximum Temperature

Variables	Correlation Coefficient, r	p-value
LON	0.203	0.000
LAT	0.024	0.583
ELE	-0.087	0.048
WDSP	0.146	0.001

A multiple regression analysis is conducted to examine the relationship between the annual maximum temperature and all predictor variables. The two variables, LON and WDSP exhibit a significant influence to the annual maximum temperature resulting from the result of the multiple regression analysis as stated in Table 11. These both significant variables are fitted again into a multiple regression model.

Table 11: Multiple Regression Analysis for Annual Maximum Temperature with Four Variables

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.001	1.730
LAT	0.000	0.000	0.835	1.104
ELE	0.000	0.001	0.498	1.525
WDSP	0.071	0.086	0.063	1.145

The result in Table 12 indicates that the variables are significant at $\alpha = 0.10$. The diagnostics of the regression model are conducted to assess the adequacy of the model. Figure 5 shows that the error variance is constant but the normality assumption is not met as the standardized residual is not normally distributed as in Table 13. The value of the coefficient of determination, R^2 of the model is low which is 0.048. Comparing the result between analyses for mean of annual maximum temperature in Part I and analysis for annual maximum temperature in Part II, it appears that the WDSP factor also affects the annual maximum temperature other than the LON factor.

Table 12: Linear Regression Analysis for Annual Maximum Temperature with One Variable

Variables	β-value	Standard error	p-value	VIF
LON	0.000	0.000	0.000	1.128
WDSP	0.073	0.038	0.055	1.128

Table 13: Normality Test for Linear Regression Model with Two Variables

Variables	Statistic value	df	p-value
Standardized Residual	0.888	21	0.000

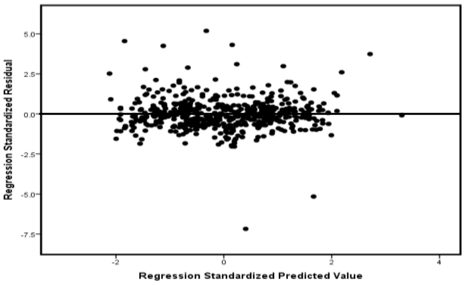


Fig 5 Residual Plot for Regression Model with Two Variables

CONCLUSION

Finding the factors which influence the air temperature is essential to develop understanding in climate studies. This study focused on investigating the relationship between the geographical factors: LON, LAT, ELE and the wind speed (WDSP) factor and the annual air temperature in Malaysia. For the first part of analysis, the result of correlation test shows that LON and WDSP correlate positively with the annual average air temperature while LAT correlates negatively with the annual average air temperature. For the case of annual maximum air temperature, only LON variable shows a significant positive correlation while the other three factors do not demonstrate any significant correlation.

Although the scatter plots of some predictor variables form a straight line which indicate the existence of linear relationship with the dependent variable, but when they are fitted into a multiple regression model, only LON variable is significant. Therefore, the result obtained from the first part of this study is not as expected and contradict with the study done by Lado et al. [1] and Ninyerola et al. [2]. The difference in the result may be due to the geographical location and climate. For the second part of analysis, the correlation test indicates that LON, ELE and WDSP factor have a significant correlation with the annual maximum of temperature. However, only LON and WDSP factors show a significant influence towards the annual maximum temperature in the multiple regression analysis. For this reason, the future research would be more comprehensive if the influential factors (e.g. LON and WDSP) are included as the variables in the air temperature modeling using other approaches.

ACKNOWLEDGMENTS

This work is funded by School of Mathematical Sciences, Universiti Sains Malaysia under Short Term Grant 304/PMATHS/6311029.

REFERENCES

[1] Lado L.R., Sparovek G., Torrado P.V., Neto D.D., Vazquez F. M. 2007. Modelling Air Temperature for The State of São Paulo, Brazil. Sci.Agric 64(5), 460-467.

[2] Ninyerola M., Pons X., Roure J.M. 2000. A Methodological Approach of Climatological Modelling of Air Temperature and Precipitation Through Gis Techniques. Int. J. Climatol 20, 1823-1841.

[3] Stimson J.A., Cramines E.G., Zeller R.A. 1978. Interpreting Polynomial Regression. Sociological Methods & Research 6(4), 515-524.

[4] Myung I.J. 2003. Tutorial on Maximum Likelihood Estimation. Journal of mathematical Psychology 47.

[5] Ministry of Higher Education. 2009. Malaysian Education Glimpse of Malaysia. Retrieved January 30, 2014, from http://www.mohe.gov.my/educationmsia/education.php?article=glimpse

[6] Framji K.K. and Garg B.C., Luthra S. D. L. 1982. Irrigation and Drainage in the World: A Global Review. Volume II, New Delhi: International Commission on Irrigation & Drainage.

[7] National Atlas of the United States. 2013. Latitude and Longitude. Retrieved January 30, 2014, from http://nationalatlas.gov/articles/mapping/a_latlong.html

[8] Neter F.J., Kutner M.H., Nachtsheim C.J., Wasserman W. 1996. Applied Linear Statistical Models. 4th Edition, Prentice Hall, Chicago, Irwin.

[9] Bruce R. 2010. Journal of Targeting, Measurement and Analysis for Marketing. Variable Selection

Methods in Regression: Ignorable Problem, Outing Notable Solution 18(1), 65-75.

[10] Lin F. 2008. Solving Multicollinearity in the Process of Fitting Regression Model Using the Nested Estimate Procedure. Quality & Quantity 42, 417-426.

[11] Osborne J.W., Waters E. 2002. Four Assumptions of Multiple Regression That Researchers Should Always Test. Practical Assessment, Research, and Evaluation 8(2), 1-5.