MULTIPLE REGRESSION ANALYSIS USING CLIMATE VARIABLES

Nur Azilla Binti Jamainal, Assoc. Prof. Dr. Fadhilah Yusof

1.0 INTRODUCTION

Regression analysis is very useful when it comes to study the relationship between variables. Regression analysis can identify the cause and effect of one variable to another variable. Variables is the main part in regression analysis. There are dependent variable (or criterion variable) and independent variable (or predictor variable). In multiple regression, the independent variables can be added more in the model then explain the cause and effect of dependent variable in more variations. Hence, dependent variable can be predicted by building better models using multiple regression analysis.

The objective of this study comprises of (i) to determine correlation between temperature, humidity, wind, solar radiation and evaporation; and (ii) to build relationships between predictand with predictors using multiple linear regression

1.1 STUDY AREA AND DATA

This research will only focus on the Multiple Regression Analysis. Methods involve are stepwise regression, backward elimination and forward selection to find best model selection. This study will also involve lag of time that can predict an approximate interval of time hence observe the effect of rainfall on the climate variables.

There are 5 variables in this problem which are temperature, humidity, wind, solar radiation and evaporation considered as independent variables. Rainfall will act as the dependent variable. This study will be analysed based on daily and monthly observation of the independent variables. This study will use satellite data from Malaysia Meteorology Service (MMS). The data obtained was from 1985 to June 2004. SPSS 22.0 and Microsoft Excel will be used in analysing those data.

1.2 LITERATURE REVIEW

Multiple regression analysis is widely used in hypotheses generated by researchers. These hypotheses may come from formal theory, previous research, or simply scientific hunches. The following hypotheses chosen from a variety of research areas.

Subana Shanmuganathan and Ajit Narayanan (2012) attempt to model the climate change/variability and its lagged effects on oil palm yield using a small set of yield data. In year 2005 and 2006 monthly temperature anomalies that affected Peninsular Malaysia did not affect Borneo's oil palm monthly yield because the afterward temperature is cooler.

Md Mizanur Rahman et al., (2013) developed a statistical forecasting method for summer monsoon rainfall over Bangladesh using simple multiple regression. Predictors for Bangladesh summer monsoon rainfall were identified which are sea-surface temperature, surface air temperature and sea level pressure. Significant correlations exist between Bangladesh seasonal monsoon rainfall and southwest Indian Ocean sea surface temperature, sea level pressure in the central Pacific region around equator and sea air temperature over Somalia.

Marla C. et al., (2010) developed equations for estimating pollutant loads and event mean concentration (which was used to quantify the washed-off pollutant concentration from non-point sources) as a function variables. They gathered runoff quantity and quality data from a 28-month monitoring conducted on the road and parking lot sites in Korea.

Mohamed E. Yassen (2000) examined the relationship between dust particulate and selective meteorological variables such as rainfall, relative humidity, temperature and wind speed in Kuala Lumpur and Petaling Jaya, Malaysia during 1983-1997. He used correlation, simple regression and multiple regression techniques to model dust concentration as a function of meteorological conditions.

1.3 METHODOLOGY

1.3.1 Correlation

Correlation measures the strength of a linear relationship between twovariables. One numerical measure is the Pearson product moment correlation coefficient,r.

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Properties of r:

1. $-1 \leq r \leq 1$
2. Values of $r$close to 1 implies there is a strong positive linear relationshipbetween $x$and y.
3. Values of $r$close to -1 implies there is a strong negative linear relationshipbetween $x$and y.
4. Values of $r$close to zero implies little or no linear relationship between$x$and y.

The value of$r$has no scale and range between -1 and 1 regardless of theunits of $x$and y.

1.3.2 Multiple Regression

Multiple regression models can be presented by the following equation:

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

Where Y is the rainfall (dependent variable), $X_1, X_2, X_3, X_4$and X_5 (independent variable) are temperature, wind, humidity, solar radiation and evaporation respectively. β_1, β_2, β_3, β_4 and β_5 are model coefficients of the five independent variables. b_0 is a constant while ε is the error.

There are assumptions that should be check before building forecasting model which are normality, multicolinearity, linearity and heteroscedasticity. All of the variables in this research must be normal distribution. The normal distribution can be seen via histogram

graph, plot P-P, plot Q-Q, kurtosis and skewness. If the distribution of the data is not normal, transformation need to be done.

It is important to evaluate the goodness-of-fit and the statistical significance of the estimated parameters of the constructed regression models; the techniques commonly used to verify the goodness-of-fit of regression models are the hypothesis testing, R-squared and analysis of the residuals. For this purpose the F-test is used to verify the statistical significance of the overall fit and the t-test is used to evaluate the significance of the individual parameters; The latter tests the importance of the individual coefficients where the former is used to compare different models to evaluate the model that best fits the population of the sample data.

Verifying the multicollinearity is also an important stage in multiple regression modelling. Multicollinearity occurs when the predictors are highly correlated which will result in dramatic change in parameter estimates in response to small changes in the data or the model. The indicators used to identify multicollinearity among predictors are tolerance (T) and variance inflation factor (VIF):

$$\text{Tolerance} = 1 - R^2$$
$$\text{VIF} = \frac{1}{Tolerance}$$

where R^2 is the coefficient of multiple determination:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

Where SST is the total sum of squares, SSR is the regression sum of squares and SSE is the error sum of squares. According to Lin (2008) a tolerance of less than 0.20 – 0.10 or a VIF greater than 5 – 10 indicates a multicollinearity problem.

To evaluate the independence of the errors of the models the Durbin-Watson test (DW) which tests the serial correlations between errors is applied. The test statistics have range of 0 – 4, according to Field (2009) values less than 1 or greater than 3 are definitely matter of concern.

1.3.3 Variable Selection Procedures

A) Forward Selection

The forward selection procedure starts with an equation containing no predictor variables, only a constant term. The first variable included in the equation is the one which has the highest simple correlation with the response variable y.

B) Backward Elimination

The backward elimination procedure starts with the full equation and successively drops one variable at a time. The variables are dropped on the basis of their contribution to the reduction of error sum of squares. This is equivalent to deleting the variable which has the smallest t-test in the equation.

C) Stepwise Regression

The stepwise method is essentially a forward selection procedure but with the added requirement that at each stage the possibility of deleting a variable, as in backward elimination, is considered. In this procedure a variable that entered in the earlier stages of selection may be eliminated at later stages.

1.3.4 Distributed Lag Analysis

Distributed lag analysis is a specialized technique for examining the relationship between variables that involve some delay.
The simplest way to describe the relationship between dependent and independent variables would be in simple linear relationship:

$$y_t = \sum \beta_i x_{t-i}$$

In this equation, the value of the dependent variable at time t is expressed as a linear function of x measured at times $t, t-1, t-2,$ etc. Thus, the dependent variable is a linear function of x, and x is lagged by 1, 2, etc. time periods. The beta weights (β_i) can be considered as the slope parameters in this equation. This equation recognized as a special case of the general linear regression equation. If the weights for the lagged time periods are statistically significant, y variable is predicted (or explained) with the lag are concluded.

1.4 RESULT AND DISCUSSION

1.4.1 Data Arrangement with Lagged Effect

Lagged effect can be found if there exist relationship of some delay of time between variables involved. It allow us to investigate lag of independent variables that affects the dependent variable. Lag 1 until Lag 5 will be examined for both cases of monthly and daily data. Lag 1 represent one day before for daily data. Table 1 shows the arrangement of data based on lag of times.

Table 1 : Arrangement of data based on lag of time for daily data.

	Dependent Variable	**Independent Variable**
Lag 0	6 Jan 1985	6 Jan 1985
Lag 1	6 Jan 1985	5 Jan 1985
Lag 2	6 Jan 1985	4 Jan 1985
Lag 3	6 Jan 1985	3 Jan 1985
Lag 4	6 Jan 1985	2 Jan 1985
Lag 5	6 Jan 1985	1 Jan 1985

The same arrangement as in daily data is done on monthly data. Lag 1 will represent one month before for monthly data and lag 5 represent five months before. Table 2 shows the arrangement of data based on lag of times in month.

Table 2 : Arrangement of data based on lag of time for monthly data.

	Dependent Variable	**Independent Variable**
Lag 0	June 1985	June 1985
Lag 1	June 1985	May 1985
Lag 2	June 1985	Apr 1985
Lag 3	June 1985	Mar 1985
Lag 4	June 1985	Feb 1985

Lag 5	June 1985	Jan 1985

1.4.2 Multiple Regression Model Using Daily Data

1.4.2.1 Descriptive Statistic

Descriptive statistic basically summarize thousands data of dependent and independent variables and represent the entire data that have been collected. This study will use Y_d as daily rainfall and X_{dn} where n= 1, 2, 3, 4, 5 as climate variables where d1, d2, d3, d4 and d5 are daily temperature, humidity, wind, solar radiation and evaporation respectively. Table 3 shows the descriptive statistic for daily data.

Table 3 : Descriptive statistic of daily data.

Variables	Mean	Standard deviation	Skewness	Kurtosis
Y_d	6.5424	15.6307	4.655	36.064
X_{d1}	27.5271	1.0127	-0.239	0.192
X_{d2}	80.8461	6.4764	-0.845	1.169
X_{d3}	1.7648	0.6815	1.266	2.499
X_{d4}	17.3844	5.1414	-1.003	1.086
X_{d5}	3.8061	1.3498	0.184	0.489

Histogram of the variables are normally distributed and this implies that all predictand and predictors fulfilled the need in building a model (Appendix A) There are also no multicolinearity problem as tolerance of variables are not less than 0.10-0.20 and VIF smaller than 5-10 (Appendix C). Homocedasticity means that the variance of errors are same across all levels of the independent variable. Appendix D shows that residuals are quite randomly scattered around 0 (the horizontal line) providing a relatively even distribution.

Each variables has their own summary. Y_d shows that the data is skewed more to the right compared with other variables. The probability distribution function of rainfall have long tail to the right side. In contrast, X_{d2} shows the data skewed more to the left. The probability distribution function of humidity is long tail to the left side compared to other variables. Y_d also give the highest standard deviation and mean value. Y_d also has the highest value of kurtosis as the peak is sharper and has fatter tails.

1.4.2.2 Correlation

A strong relationship between dependent and independent variables are needed before building a model. Before that, the data need to be arranged according to lagged of time. In this study, Lag 1 until Lag 5 have been used for each independent variables as Lag 1 represent 1 day before and Lag 5 represent 5 days before. Then, each variables with each Lag will have different value of correlation. Using data analysis in Excel, correlation between dependent and independent variables are obtained. Table 4 are the results obtained by selecting the highest correlation between dependent and independent variables.

Table 4 : Correlation of daily lagged independent variables with rainfall

	Without Lag	Lag 1	Lag 2	Lag 3	Lag 4	Lag 5
$\mathbf{X_{d1}}$	**-0.2922**	-0.1130	-0.0785	-0.0544	-0.0512	-0.0356
$\mathbf{X_{d2}}$	**0.3121**	0.1780	0.1455	0.1247	0.1201	0.1027
$\mathbf{X_{d3}}$	-0.0931	-0.0983	**-0.1074**	-0.1012	-0.0980	-0.0953
$\mathbf{X_{d4}}$	**-0.2530**	-0.1054	-0.0662	-0.354	-0.0332	-0.0093
$\mathbf{X_{d5}}$	**-0.3551**	-0.1378	-0.1051	-0.0868	-0.0858	-0.0724

From Table 4, the correlation between rainfall and evaporation is the highest while the lowest is the correlation between rainfall and wind. Only rainfall and temperature have negative correlations, the rest have positive correlations.

1.4.2.3 Variable Selection

A) Backward Elimination

Backward Elimination is a method where it starts with full equation then drop the variables which has the smallest t-test. Below are the results obtained using SPSS 22.0.

Table 5 : Coefficient of daily variables in each model for backward elimination

Model		Coefficient	T-test	Significant
1	(Constant)	40.143	4.844	0.0000
	X_{d1}	-1.644	-7.202	0.0000
	X_{d2}	0.274	6.921	0.0000
	X_{d3Lag2}	0.413	1.428	0.153
	X_{d4}	-0.050	-1.163	0.245
	X_{d5}	-2.709	-15.800	0.000
2	(Constant)	41.049	4.975	0.000
	X_{d1}	-1.717	-7.819	0.000
	X_{d2}	0.279	7.099	0.000
	X_{d3Lag2}	0.452	1.571	0.116
	X_{d5}	-2.779	-17.299	0.000
3	(Constant)	46.205	6.102	0.000
	X_{d1}	-1.805	-8.510	0.000
	X_{d2}	0.254	7.065	0.000
	X_{d5}	-2.751	-17.241	0.000

Based on Table 5, there are 3 models that can get from Backward Elimination.

Model 1:

$$Y_d = 40.143 - -1.644X_{d1} + 0.274X_{d2} + 0.413X_{d3Lag2} - 0.050X_{d4} - 2.709X_{d5}$$

In Model 1, all variables are entered in the equation. Based on Table 5, the smallest t-test value and most insignificant variable is the solar radiation with t-test value, 1.163 and p-value = 0.245, followed by wind with lag of 2 days, with second smallest t-test value, 1.428 and p-value = 0.153. Thus, variable solar radiation will be eliminated first from the equation.

Model 2:

$$Y_d = 41.094 - 1.717X_{d1} + 0.279X_{d2} + 0.452X_{d3Lag2} - 2.779X_{d5}$$

In Model 2, since variable wind has been eliminated, there are only 4 variables in the equation. Based on Table 5, wind with lag of 2 days has the smallest t-test value, 1.571 with p -value = 0.116 which is insignificant. Thus, wind will be eliminated from the equation.

Model 3:

$$Y_d = 46.188 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$$

After eliminating 2 variables, the equation now has only 3 variables which are the temperature, humidity and evaporation. Based on the Table 5, all the variables have higher t-test value and significant with p-value = 0.000.

Table 6 : Model summary of MRA using backward elimination

Model	R^2	Adjusted R^2
1	0.152	0.151
2	0.152	0.151
3	0.151	0.151

From the table, all models have same value of Adjusted R^2. Only Model 1 and Model 2 have changes between R^2 and Adjusted R^2 where Adjusted R^2 became smaller because there exist elimination of variables while there is no elimination in Model 3 because all variables in the equation are significant. Thus, based on both tables 5 and 6, Model 3 is better than Model 2 and Model 1 since Model 1 and Model 2 contain variables that are not significant.

Therefore the selected model using backward elimination method is:
$Y_d = 46.188 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$

B) Forward Selection and Stepwise

Forward selection is a method where it start with zero independent variable in the equation then select the variables which has the highest correlation with rainfall. Stepwise is basically a combination of Backward Elimination and Forward Substitution method. Since the result of Forward Substitution and Stepwise are same, below are the result obtained using SPSS 22.0.

Table 7 : Excluded daily variables using forward selection and stepwise

Model		T-test	Significant	Partial Correlation
1	X_{d1}	-12.658	0.000	-0.148
	X_{d2}	11.730	0.000	0.138
	X_{d3Lag2}	-0.428	0.669	-0.005
	X_{d4}	-6.261	0.000	-0.074
2	X_{d2}	7.066	0.000	0.083
	X_{3Lag2}	-1.417	0.156	-0.017
	X_{d4}	-1.841	0.066	-0.022
3	X_{d3Lag2}	1.571	0.116	0.019
	X_{d4}	-1.335	0.182	-0.016

Based on both Table 7, there are 3 models that can get from Forward Substitution and Stepwise.

Model 1:

$$Y_d = 22.201 - 4.115X_{d5}$$

From Table 7, evaporation has the highest correlation with rainfall. Thus, variable evaporation is entered first in the equation. The variable entered also significant since it has high value of t-test based on Table 4.10.

Model 2:

$$Y_d = 86.069 - 2.442X_{d1} - 3.234X_{d5}$$

In Model 2, variable temperature is entered in the equation. This is because, from Table 7, temperature has highest partial correlation, -0.148, after evaporation entered the equation.

Temperature also has high value of t-test, 12.658, and also a significant variable with p-value = 0.000. Thus, variable temperature is entered in the equation.

Model 3:

$$Y_d = 46.188 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$$

From Table 7, variable humidity has the highest partial correlation, 0.083, after variable temperature and variable evaporation entered the equation. Besides, humidity is the only variable that is significant with p-value = 0.000 and highest t-test value, 7.066. Thus, humidity is the last variable entered in the equation.

Table 8 : Model summary of MRA using forward selection and stepwise

Model	R^2	Adjusted R^2
1	0.126	0.126
2	0.145	0.145
3	0.151	0.151

According to Table 8, each model has different value of R^2 and Adjusted R^2. The value of R^2 increases each time a new variable added thus improves the models. Model 3 has the highest value of Adjusted R^2 which means the model is better than Model 1 and Model 2.

Therefore the selected model using forward selection and stepwise method is:
$Y_d = 46.205 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$

1.4.3 Multiple Regression Model Using Monthly Data

1.4.3.1 Descriptive Statistic

Since the data obtained originally in the form of daily data, subtotal of Excel is used to find the sum of each variable of each month. The study can be proceeded to find best model for monthly data. SPSS 22.0 was used in finding the descriptive statistic of the monthly data. This study will use Y_m as monthly rainfall and X_{mn} where n= 1, 2, 3, 4, 5 as climate variables where m1, m2, m3, m4 and m5 are monthly temperature, humidity, wind, solar radiation and evaporation respectively.

Table 9 : Descriptive statistic of monthly data

Variables	Mean	Standard deviation	Skewness	Kurtosis
Y_m	199.0653	135.3564	1.160	2.254
X_{m1}	837.6940	29.0624	-0.212	0.155
X_{m2}	2460.2782	160.8352	-0.940	0.909
X_{m3}	53.7060	12.6717	0.940	0.364
X_{m4}	529.0346	72.1005	-1.878	12.738
X_{m5}	115.8256	21.9630	0.563	-0.195

The histogram of the variables are normally distributed (Appendix B), thus fulfil the need in building a model. There are also no multicolinearity problem as tolerance of variables are not less than 0.10-0.20 and VIF smaller than 5-10 (Appendix C). Appendix D shows that residuals are randomly scattered around 0 (the horizontal line) providing a relatively even distribution which implies that the scatter plot shows a homoscedasticity pattern.

Table 9 shows that X_{m2} has the highest standard deviation and with highest mean. Y_m skewed more to the right compared with other variables. Probability distribution function of monthly rainfall has long tail to the right, same with the daily rainfall. X_{m4} has the most

negative skewness implies that the probability distribution function of monthly solar radiation has long tail to the lefts side.

It seems that only evaporation has the negative value of kurtosis. This indicates that the distribution of monthly evaporation is too flat or even concave if the value is large enough. In addition there are 2 variables which have quite high kurtosis value, which are rainfall and solar radiation. This indicates that the distribution of both monthly variables have sharp peaks.

1.4.3.2 Correlation

The same procedures as in daily data are repeated here. Firstly, the data need to be arranged according to lagged of time. In this study, Lag 1 until Lag 5 have been used for each independent variables as Lag 1 represent 1 month before and Lag 5 represent 5 months before. Each variables with each Lag will have different value of correlation. Using data analysis in Excel, correlation between dependent and independent variables were obtained. Below are the results obtained by selecting the highest correlation between dependent and independent variables.

Table 10 : Correlation of monthly lagged independent variables with rainfall

	Without Lag	**Lag 1**	**Lag 2**	**Lag 3**	**Lag 4**	**Lag 5**
$\mathbf{X_{m1}}$	-0.2147	-0.0429	0.1094	0.2089	**0.2686**	0.2592
$\mathbf{X_{m2}}$	**0.6342**	0.3386	0.1570	-0.0145	-0.0793	-0.0390
$\mathbf{X_{m3}}$	**-0.5318**	-0.4262	-0.2398	-0.0812	0.0089	0.0145
$\mathbf{X_{m4}}$	**-0.4600**	-0.1121	-0.0056	0.0694	0.1685	0.2123
$\mathbf{X_{m5}}$	**-0.6539**	-0.3142	-0.0989	0.0827	0.1425	0.1595

From Table 10, correlation between rainfall and evaporation is the highest while the lowest is the correlation between rainfall and temperature. Temperature and humidity have positive correlation towards rainfall while wind, solar radiation and evaporation have negative correlation towards rainfall.

1.4.3.3 Variables Selection

A) Backward Elimination

Below are the result obtained using SPSS 22.0.

Table 11 : Coefficient of monthly variables in each model for backward elimination.

Model		**Coefficient**	**T-test**	**Significant**
1	(Constant)	-513.498	-1.838	0.067
	Xm_{1Lag4}	0.322	1.386	0.167
	Xm_2	0.297	5.412	0.000
	Xm_3	-0.771	-1.069	0.286
	Xm_4	0.055	0.362	0.718
	Xm_5	-2.371	-4.419	0.000
2	(Constant)	-482.763	-1.817	0.071
	Xm_{1Lag4}	0.322	1.386	0.167
	Xm_2	0.293	5.466	0.000
	Xm_3	-0.841	-1.213	0.226
	Xm_5	-2.256	-5.216	0.000
3	(Constant)	-543.299	-2.079	0.039

	Xm_{1Lag4}	0.313	1.347	0.179
	Xm_2	0.313	6.133	0.000
	Xm_5	-2.484	-6.371	0.000
4	(Constant)	-261.466	-1.667	0.097
	Xm_2	0.311	6.095	0.000
	Xm_5	-2.627	--6.987	0.000

Based on Table 11, there are 4 models that can get from Backward Elimination method which are:

Model 1:

$$Y_m = -513.498 + 0.322X_{m1Lag4} + 0.297X_{m2} - 0.771X_{m3} + 0.055X_{m4} - 2.371X_{m5}$$

In Model 11, all the variables are entered in the equation. Based on Table 4.16, the smallest t-test value, 0.362, with most insignificant value, p-value = 0.718 is variable solar radiation. Variable wind also has insignificant value and second lowest t-test value which are p-value = 0.286 and 1.069.Thus variable solar radiation will be eliminated first from the equation.

Model 2:

$$Y_m = -482.763 + 0.322X_{m1Lag4} + 0.293X_{m2} - 0.841X_{m3} - 2.256X_{m5}$$

In Model 2, since variable solar radiation has been eliminated, there are 4 variables found in the equation. Based on Table 11, variable wind has the lowest value of t-test, 1.213, also p-value = 0.226 which is insignificant. Variable temperature with lag of 4 months also has low t-test value, 1.386 and p-value = 0.167 which is insignificant. Thus, variable wind will be eliminated.

Model 3:

$$Y_m = -543.299 + 0.313X_{m1Lag4} + 0.313X_{m2} - 2.484X_{m5}$$

Previously, variable solar radiation and variable wind have been eliminated, so the equation now contain 3 variables. From table above, temperature with lag of 4 months has the lowest t-test value and insignificant which are 1.347 and p-value = 0.179 respectively. Thus temperature is eliminated from the equation.

Model 4:

$$Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$$

In this model, there only exist 2 variables which are humidity and evaporation variables since the other variables have been eliminated previously. Based on the table above, all variable in the equation have high value of t-test value and significant with p-value = 0.000

Table 12 : Model summary of MRA using backward elimination

Model	**R^2**	**Adjusted R^2**
1	0.516	0.505
2	0.516	0.507
3	0.512	0.506
4	0.508	0.504

According to Table 12, each model has different value of R^2 and Adjusted R^2. From Model 1 until Model 4, the value of R^2 decreasing since variable that is not qualified will be dropped from the equation to achieve a best model. Model 2 has the highest value of Adjusted R^2, but in Model 2 there exist variables that are not significant, thus it cannot conclude that the model is better than Model 1, Model 2 and Model 4. Since only Model 4 contains only significant variables, therefore Model 4 is chosen.

Therefore the selected model using backward elimination method is:
$Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$

B) Forward Selection and Stepwise

Same goes to daily data, the result of forward substitution and stepwise of monthly data are the same. Below are the result obtained.

Table 13 : Excluded monthly variables using forward selection and stepwise

Model		T-test	Significant	Partial Correlation
1	X_{m1Lag4}	1.134	0.258	0.075
	X_{m2}	6.095	0.000	0.376
	X_{m3}	-2.822	0.005	-0.184
	X_{m4}	-0.106	0.916	-0.007
2	X_{m1Lag4}	1.347	0.179	0.089
	X_{m3}	-1.168	0.244	-0.078
	X_{m4}	0.651	0.515	0.043

Based on both Table 13, there are 2 models that can be obtained from forward substitution and stepwise method.

Model 1:

$$Y_m = 671.492 - 4.079X_{m5}$$

From the Table 10, evaporation has the highest correlation with rainfall. Thus, variable evaporation is entered first in the equation.

Model 2:

$$Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$$

From Table 13, variable humidity has the highest partial correlation with rainfall, 0.376, after variable evaporation entered the equation. Partial correlation of temperature, wind and solar radiation towards rainfall are quite low with value nearly 0. Besides, humidity has high t-test value, 6.095, and p-value = 0.000 which is significant. Thus, variable humidity is the most suitable variable entered to the equation after variable evaporation.

Table 14 : Model summary of MRA using forward selection and stepwise

Model	R^2	Adjusted R^2
1	0.428	0.425
2	0.508	0.504

According to Table 14, each model has different value of R^2 and Adjusted R^2. The value of R^2 increasing along the models since new variable added thus improves the models. Model 2 has the highest value of Adjusted R^2 which means the model is better than Model 1.

Therefore the selected model for forward selection and stepwise method is:

$$Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$$

1.4.4 Summary Analysis

In summary, during analysing daily and monthly data, backward elimination, forward selection and stepwise methods has managed to produce the same result. Model for case I: daily data is $Y_d = 46.188 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$and model for case II: monthly data is $Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$. Both models did not have lagged variables which implies that there are no lagged effect in the models.

R^2 of daily model is 0.151 which indicates that the model explains 15.1% variability of the response data around its mean. It also shows the data are 15.1% close to the fitted regression line. R^2 of monthly model is 0.508 which implies the model explains 50.8% variability of response data around its mean. At the same time, 50.8% of the data are close to the fitted regression line. These mean that the monthly model is better fits compared to daily model.

Generally, it is better to look at Adjusted R^2 rather than R^2 and to look at the standard error of the regression rather than the standard deviation of the errors in order to strengthen the above statement. Adjusted R^2 of daily model is 0.151 and monthly model is 0.504 which indicates that monthly model contains predictors that improves the model more rather than daily model. Since monthly model has high adjusted R^2 value, 0.504 compared to daily model 0.151, monthly model is preferable than daily model.

1.5 CONCLUSION

1.5.1 Conclusion

Multiple regression analysis is usually used to determine the relationship of two or more independent variables. Correlation between rainfall and climate variables plays an important role in order to build a good model for rainfall prediction. There are 3 methods used in multiple regression analysis which are the backward elimination, forward selection and stepwise. The results of these 3 methods are compared to determine which method is the best. The data obtained from Malaysia Meteorological Servicesare analysed by using Microsoft Excel and SPSS 22.0. In this study, the data analysed were in the form of daily data and monthly data.

Generally, both daily and monthly data analysis shows that evaporation has the highest correlation towards rainfall and the lowest correlation goes to solar radiation variable. Based on daily data analysis, only wind that need to be lagged with 2 days data. However, the correlation value is not high compared to other climate variables and not added in the daily model. Unlike monthly data analysis, temperature require a lag of 4 months. However, the correlation value is not high compared to other climate variables and thus not added in the monthly model.

In analysing daily data, the result obtained shows that all 3 methods produced the same result. In this case, the model contain 3 significant variables which are the temperature, humidity and evaporation. In the second case, monthly data are used in the analysis. All 3 methods give the same result, same with analysing daily data. The model contains only 2 variables that are significant which are humidity and evaporation.

Model selected from analysing the daily data is $Y_d = 46.188 - 1.805X_{d1} + 0.254X_{d2} - 2.751X_{d5}$. The model shows that daily variables of temperature, X_{d1}, humidity, X_{d2}, and evaporation, X_{d5}, have the best correlation towards rainfall. These 3 variables are significant and the model have high value of Adjusted R^2 than other models, which makes this model better for predicting rainfall pattern in the future.

For monthly data, model selected is $Y_m = -261.466 + 0.311X_{m2} - 2.627X_{m5}$. The model shows that monthly variables of humidity, X_{m2} and evaporation, X_{m5} have the best correlation towards rainfall. Even though this model has small value of Adjusted R^2, model that contains all significant variables is a priority so that the model would produce smaller error in predicting the rainfall in future.

In overall, from the result obtain, analysing using monthly data is better since the model has higher value of Adjusted R^2 compared to model of daily data analysis. In terms of methods used, forward selection and stepwise always give the same results thus both are the best methods because these methods will strictly picked variables that are significant for the model and produce the best model. Lag of time is also not important in this case because there is no effect of lag based on both models. The independent variables did not need to lag in order to predict the rainfall amount since the result can be obtained on same day or month.

ACKNOWLEDGEMENT

The authors wish to thank Malaysia Meteorological Services for allowing us to do analysis on the climate data.

REFERENCES

Andy,F. (2005). *Discovering statistics using SPSS: (and sex, drugs and rock 'n' roll)*, London: SAGE, 2005.

Jacob Cohen, Patricia Cohen, Stephen G. West, Leona S. Aiken (2003), *Applied Multiple Regression/Correlation Analysis for the Behavioural Sciences Third Edition*, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, 2003.

John O. Rawlings, Sastry G. Pantula, David A. Dickey (1998), *Applied Regression Analysis: A Research Tool, Second Edition*, Springer-Verlag New York, Inc., North Carolina State University Raleigh, 1998.

Samprit Chatterjee, Ali S. Hadi (2006), *Regression Analysis by Example, Fourth Edition*, John Wiley & Sons, Inc., Hobokon, New Jersey, 2006.

Timothy Z. Keith (2006), *Multiple Regression and Beyond*, Pearson Education, University of Texas at Austin, 2006.

Barsugli, J.J., Sardeshmukh, D.D., 2002. The Journal of Climate. *Global atmospheric sensitivity to tropical SST anomalies throughout the Indo-Pacific basin*, 15(23) 3427-3442.

Chattopadhyay, G., Chattopadhyay, S., Jain, P., 2010. *Multivariate forecast of winter monsoon rainfall in India using SST anomaly as a predictor: neorocomputing and statistical approaches*. Comptes Rendus Geoscience 342(100, 755-765.

F.Mekanik, M.A. Imtez, S. Gato-Trinidad, A. Elmahdi. (2013). The Journal of Hydrology. *Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes*, 503, 11-21.

Hartmann, H., Becker, S., King, L, 2008. International Journal of Climatology. *Predicting summer rainfall in the Yangtze River basin with neural networks*, 28(7) 925-936.

Intan Martina Md Ghani, Sabri Ahmad. (2010). The Journal of Procedia Social and Behavioral Sciences. *Stepwise Multiple Regression Method to Forecast Fish Landing*, Vol. 8, 549-554.

Lau, K., Weng, H., 2001. The Journal of Climate *Coherent nodes of global SST and summer rainfall over China: an assessment of the regional impacts of the 1997-98 El Nino*, 14(6) 1294-1308.

Marla C. Maniquaz, Soyoung Lee, Lee Hyung Kim (2010). The Journal of Environmental Sciences. *Multiple linear regression models of urban runoff pollutant load & event mean concentration considering rainfall variables*, 22(6) 946-952.

Mohamed E. Yassen (2000). The Journal of Social Sciences and Humanities. *The relationship between dust particulates & meteorological parameters in Kuala Lumpur and Petaling Jaya, Malaysia*. Universiti Kebangsaan Malaysia.

Shukla, R.P., Tripathi, K.C., Pandey, A.C., Das, I.M.L., 2011. *Prediction of Indian summer monsoon rainfall using Niño indices: a neural network approach*. Atmospheric Research 102 (1–2), 99–109.

Subana Shanmuganathan, Ajit Narayanan (2012). The Journal of Geoinformatics Research Centre and School of Computing and Mathematical Science. *Modelling the climate change effects on Malaysia's oil palm yield.* IEEE Symposium on E-learning, E-management and E-services (IS3e).

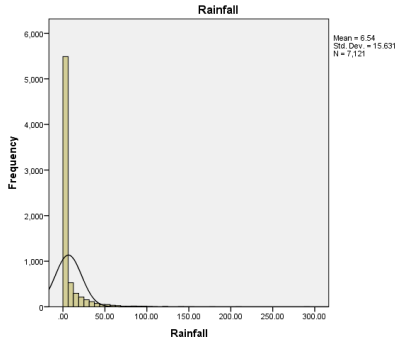
Yufu, G., Yan, Z., Jia, W., 2002. *Numerical simulation of the relationships between the 1998 Yangtze River valley floods and SST anomalies*. Advances in Atmospheric Sciences 19 (3), 391–404.

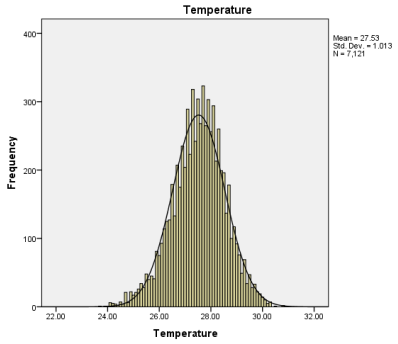
StatSoft, Inc. (2013). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft. WEB: *http://www.statsoft.com/textbook/*

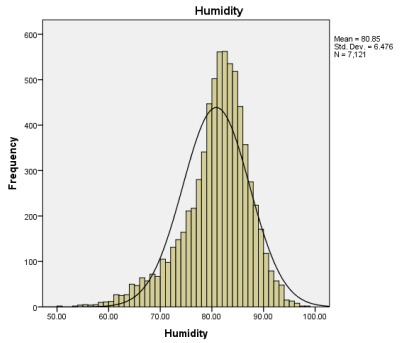
INDEX

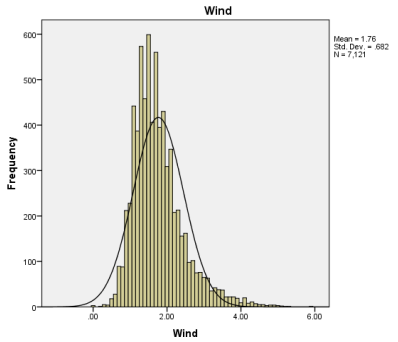
APPENDIX A

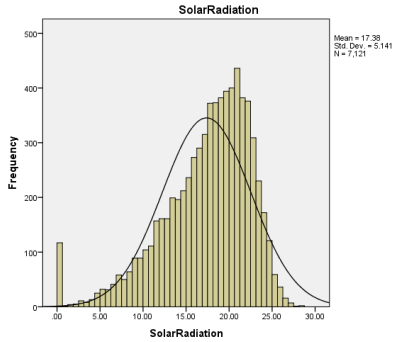
Normality of Variables Based on Histogram Using SPSS 22.0 for Daily Data

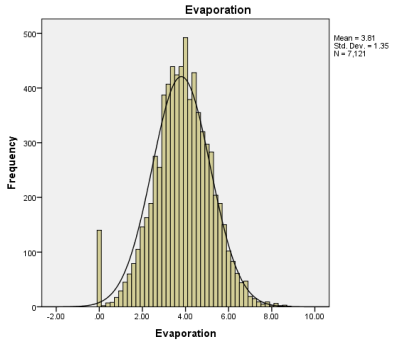






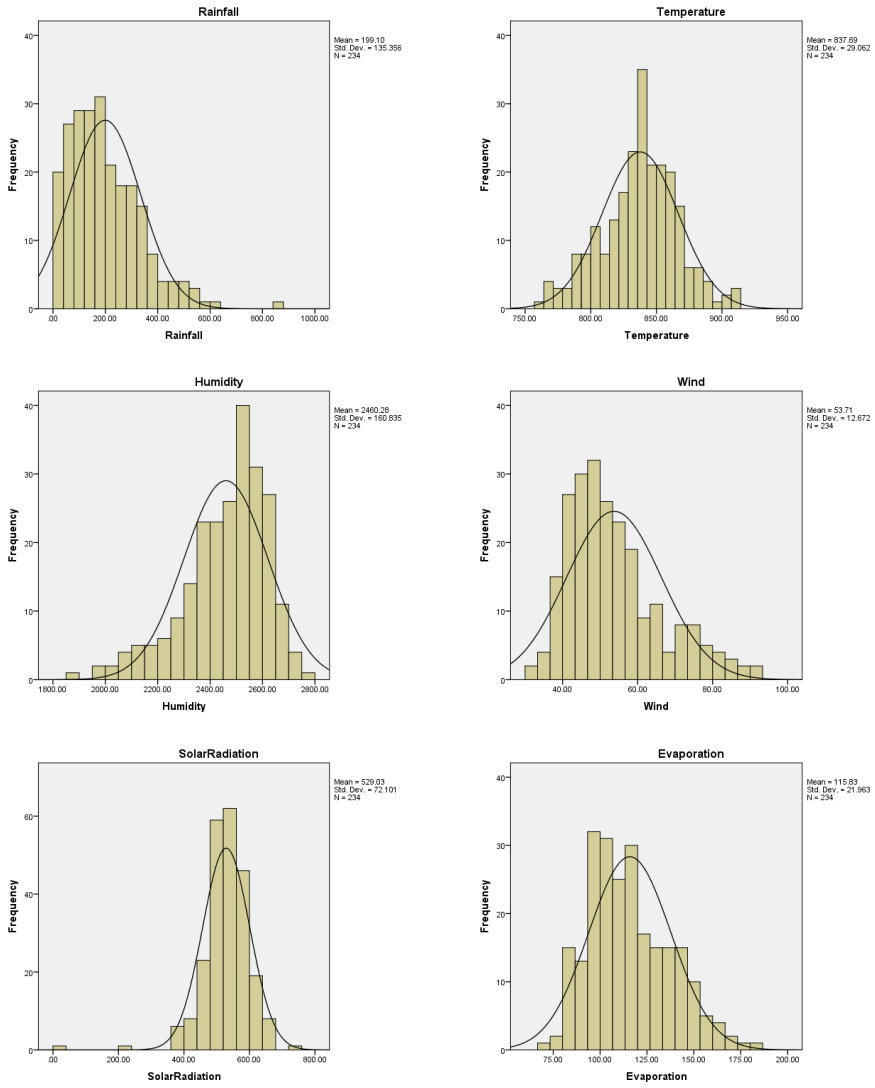






APPENDIX B

Normality of Variables Based on Histogram Using SPSS 22.0 for Monthly Data



APPENDIX C

SPSS 22.0 Output for Multicolinearity and Independence of Error

Backward elimination using daily data

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	40.143	8.288		4.844	.000		
	TempLag0	-1.644	.228	-.107	-7.202	.000	.545	1.834
	HumidLag0	.274	.040	.113	6.921	.000	.445	2.245
	WindLag2	.413	.289	.018	1.428	.153	.751	1.331
	SolarLag0	-.050	.043	-.016	-1.163	.245	.595	1.682
	EvaporLag0	-2.709	.171	-.234	-15.800	.000	.545	1.835
2	(Constant)	41.049	8.251		4.975	.000		
	TempLag0	-1.717	.220	-.111	-7.819	.000	.589	1.696
	HumidLag0	.279	.039	.115	7.099	.000	.451	2.216
	WindLag2	.452	.287	.020	1.571	.116	.761	1.313
	EvaporLag0	-2.779	.161	-.240	-17.299	.000	.621	1.611
3	(Constant)	46.188	7.576		6.097	.000		
	TempLag0	-1.805	.212	-.117	-8.503	.000	.631	1.585
	HumidLag0	.254	.036	.105	7.066	.000	.540	1.854
	EvaporLag0	-2.751	.160	-.237	-17.229	.000	.629	1.591

a. Dependent Variable: Rainfall

Forward selection and stepwise method using daily data

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	22.201	.519		42.819	.000		
	EvaporLag0	-4.115	.128	-.355	-32.034	.000	1.000	1.000
2	(Constant)	86.069	5.071		16.971	.000		
	EvaporLag0	-3.234	.145	-.279	-22.329	.000	.769	1.300
	TempLag0	-2.442	.193	-.158	-12.658	.000	.769	1.300
3	(Constant)	46.188	7.576		6.097	.000		
	EvaporLag0	-2.751	.160	-.237	-17.229	.000	.629	1.591
	TempLag0	-1.805	.212	-.117	-8.503	.000	.631	1.585
	HumidLag0	.254	.036	.105	7.066	.000	.540	1.854

a. Dependent Variable: Rainfall

Backward elimination using monthly data

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-513.498	279.454		-1.838	.067		
	TempLag4	.322	.233	.068	1.386	.167	.891	1.123
	HumidLag0	.297	.055	.350	5.412	.000	.518	1.929
	WindLag0	-.771	.721	-.072	-1.069	.286	.483	2.072
	SolarLag0	.055	.153	.025	.362	.718	.468	2.135
	EvaporLag0	-2.371	.536	-.380	-4.419	.000	.294	3.406
2	(Constant)	-482.763	265.738		-1.817	.071		
	TempLag4	.322	.232	.068	1.386	.167	.891	1.123
	HumidLag0	.293	.054	.345	5.466	.000	.542	1.847

	WindLag0	-.841	.693	-.078	-1.213	.226	.520	1.922
	EvaporLag0	-2.256	.433	-.362	-5.216	.000	.450	2.223
3	(Constant)	-543.299	261.285		-2.079	.039		
	TempLag4	.313	.232	.066	1.347	.179	.892	1.121
	HumidLag0	.313	.051	.369	6.133	.000	.598	1.671
	EvaporLag0	-2.484	.390	-.398	-6.371	.000	.555	1.803
4	(Constant)	-261.466	156.832		-1.667	.097		
	HumidLag0	.311	.051	.367	6.095	.000	.599	1.670
	EvaporLag0	-2.627	.376	-.421	-6.987	.000	.599	1.670

a. Dependent Variable: Rainfall

Forward selection & stepwise using monthly data

Coefficients[a]

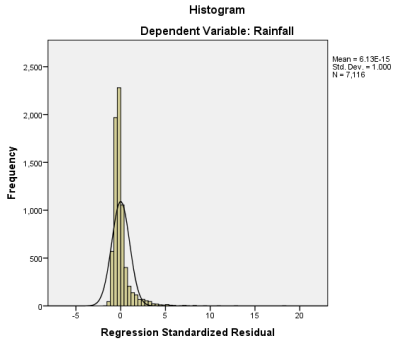
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	671.492	36.815		18.240	.000		
	EvaporLag0	-4.079	.313	-.654	-13.022	.000	1.000	1.000
2	(Constant)	-261.466	156.832		-1.667	.097		
	EvaporLag0	-2.627	.376	-.421	-6.987	.000	.599	1.670
	HumidLag0	.311	.051	.367	6.095	.000	.599	1.670

a. Dependent Variable: Rainfall

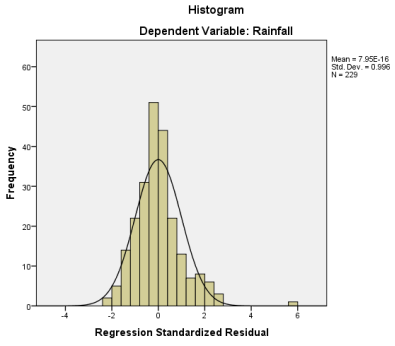
APPENDIX D

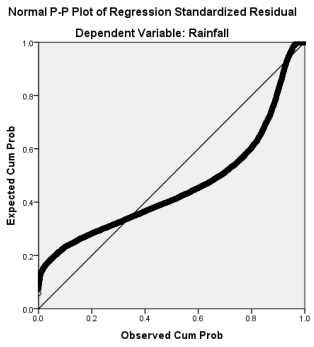
SPSS 22.0 Output for Heterocedasticity

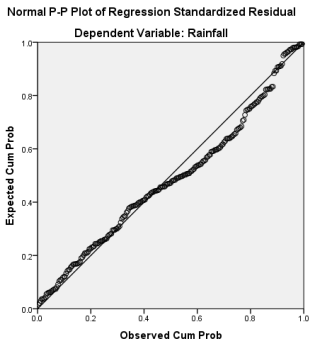
Daily Data Analysis

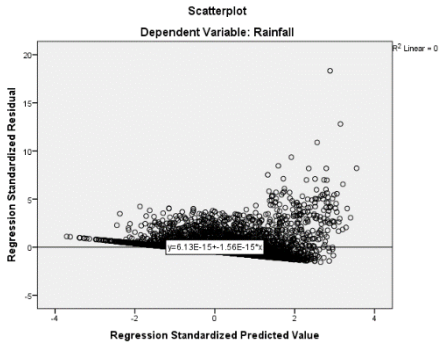


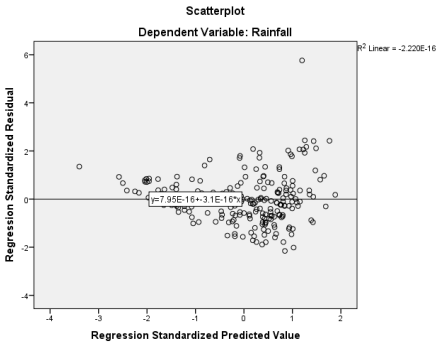
Monthly Data Analysis




Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Rainfall
Expected Cum Prob
Observed Cum Prob


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Rainfall
Expected Cum Prob
Observed Cum Prob


Scatterplot
Dependent Variable: Rainfall
Regression Standardized Residual
Regression Standardized Predicted Value


Scatterplot
Dependent Variable: Rainfall
Regression Standardized Residual
Regression Standardized Predicted Value