Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 10, 2022, page 51 - 60

Multiple Linear Regression on Residential Household Air Conditioning Electricity Consumption

Chong Ee Xuan, Muhammad Fauzee Hamdan*
Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia
*Corresponding author: mfauzee@utm.my

Abstract
With the increasing of adopting air-conditioning in the household, the demand for energy is also increasing. The high power of AC consumption will make fluctuations in energy demand and affect the stability of energy demand on the grid. Therefore, this paper focuses on the formation of regression models for predicting air-conditioning electricity consumption by analyzing the relationship between occupant behavior factors and air-conditioning electricity consumption. In this study, the field measurement data was collected in a household from the 10th of January 2021 to 14th of November 2021 while the historical forecast weather was collected from the website of world weather online. Assumption of multilinear regression analysis: normality, linearity and multicollinearity were examined. Next, the process of fitting the model was carried out by using stepwise variable selection method. Stepwise regression consists of a series of steps adding or removing potential independence variables in succession and testing for statistical significance after each iteration can help to construct an effective regression model. The output from the regression model was interpreted and analyzed. Besides, the analysis of the residual was also important to ensure the validation of the model. It is found that the temperature, time of day, and precipitation present a significant contribution to the AC electricity consumption.

Keywords: Multiple Linear Regression; Stepwise Regression; AC electricity consumption

1. Introduction

Due to climate change and increasingly high temperatures, the number of households adopting air conditioning has been increasing to adapt the hot climate conditions [1]. The rising of household air conditioning (AC) consumption from 6.5 % in 1990 to 16.2% in 2000, AC contributes a big portion to household electricity consumption [2]. The average electricity consumption in AC is almost half of the average residential electricity consumption during weekdays [3]. The finding of Randazzo et al. [1] research also showed that households with AC will spend extra 35%–42% on electricity in average than those who do not own AC. The high power of AC consumption will make fluctuations in energy demand and affect the stability of energy demand on the grid. The stability of energy demand and supply is important for cost-effectiveness of resources and reducing the wastage of energy. Therefore, a detailed AC electricity consumption structure is essential for a reliable power system [2].

Determining the factors affecting AC electricity consumption in residential households is very important in predicting load. There are many different factors affecting AC electricity consumptions which include socio-demographic factors, house characteristics, appliance characteristics, and occupant behaviour. Those factors can be classified as monthly income, education level, family composition, total floor area, house type, appliance ownership, use of appliances and so on. Interactions between these factors might also affect the amount of electricity consumption [4]. Besides, there exist a relationship between temperature, income, and air conditioning usage [5]. Thus, the study on the factors in determining AC electricity consumption is required in order to find out which factors are significant contributors to AC electricity consumption in residential households.

Multiple Linear Regression (MLR) analysis is used to predict the AC electricity consumption in this paper. Regression analysis is widely used statistical technique to model the relationship between independent variable and dependent variable. Regression analysis can be used in many applications such as engineering, chemical science, economics and management. Regression model is employed for parameter estimation, prediction or estimation, data description and control. Multiple regression model involves more than one predictor variables [6]. Regression analysis is applicable for independence variable which has linear relationship with dependence variable.

This research aims to (1) determine the factors affecting air conditioning electricity consumption in Malaysian households by using the Pearson correlation method and (2) build a multiple linear regression model on AC electricity consumption model. The amount of MAPE and MSE value is obtained by comparing the forecasting results of AC electricity consumption and testing sample data. The smaller the MAPE and MSE value, the better the predicting results.

2. Literature Review

2.1. Introduction

Electricity is the flow of electrical power or charge. Electricity is also called secondary energy sources because it is generated through the conversion of primary energy sources such as coal, natural gas, oil, nuclear power, and other natural sources. The sources of electricity can be renewable or non-renewable energy sources [7]. Electricity is one of the most important components in our daily lives as it is widely used forms of energy. We use electricity to do many jobs including lighting, heating, cooling our homes and to powering our televisions and computers.

Air conditioner, often abbreviated as AC, uses electricity to remove heat from the interior of the house to the outside environment which is relatively warm and controls the humidity in the house to provide a more comfortable living environment. Climate change and increasing temperature have accelerated the use of AC and electricity consumption in residential households [1]. There was almost 75% of the occupants use the AC for about five to six hours every day at the rates 0.93 kWh /day during day time and 3.43 kWh/day during night time [2]. AC electricity consumption had a big share in the total electricity consumption in residential households. Therefore, it was very important to investigate the factors affecting the AC electricity consumption in residential households to improve forecasting for the energy demand.

The factors affecting electricity consumption in residential households can be categorized into socio-demographic factors, house characteristics [2, 8], appliance characteristics, and occupant behaviour [3,8, 9]. A field questionnaire survey is carried out at Universiti Teknologi Malaysia (UTM), Kuala Lumpur have been analysed by using Pearson Correlation and multiple linear regression (MLR) [10]. Besides, a study by Hisham et al. shows that there exists a positive correlation which is 0.3 to 0.5 between daily total AC consumption and daily mean outdoor temperature [2].

The study on the air-conditioning electricity consumption of office buildings in the Shanghai area was carried out by using the Long Short-Term Memory Method (LSTM) which is recurrent neural network (RNN) algorithms used in the field of deep learning. This method has been widely used in many fields. However, the accuracy of this method in predicting air-conditioning electricity consumption is not so high. The LSTM method has higher accuracy in predicting the hourly electricity consumption data of the office building [11].

Therefore, the study on the AC electricity consumption at a landed house is necessary to be carried out to determine the relationship between the occupants' behaviour and AC electricity consumption. Among the various method of research, MLR is selected as the method to carry out analysis on the AC electricity consumption because the previous study showed that MLR is suitable method to use in investigating the factors affecting the AC electricity consumption.

3. Methodology

3.1. Data

There are two set of data in this study which is the field measurement data of AC electricity consumption and the historical weather forecast is collected from world weather online https://www.worldweatheronline.com/.

3.1.1 Field Measurement

The field measurement has been conducted in a residential household in Semenyih, Selangor from 10th of January 2021 to 14th of November 2021 which lasts for 309 days. The energy monitoring device consists of SDM 230 Modbus Single-Phase Two Module DIN rail meters and a microcontroller unit with SD card storage. The energy-monitoring device was installed to monitor the energy consumption of the AC unit in the master bedroom. The electricity consumption of the AC was measured and recorded hourly. The residential and private premises in Malaysia usually will have voltage supply between 230 to 253 V (Suruhanjaya Tenaga 2008). The recorded data from 10th of January to 31th of October were used to train the model, while the last 133 data were used as testing data to test the accuracy of the model. The data has been pre-processing before starting to analyses. The categorical data which are time of the day is transformed into quantitative data by adding a dummy variable.

Table 1: Description of Dependent and Independent Variables

Labels	Definition
Response Variables	
Y	Energy used by air conditioning
Categorical Explanatory Variables	
X1	The time of the day $X_1 = \begin{cases} 1, & Day\ time \\ 0, & Night\ time \end{cases}$
Continuous Explanatory Variables	
Temp	Average temperature of the hour in ºC
Humidity	Humidity of the hour
Precip	Precipitation in unit millimetre
UV Index	The strength of sun's UV radiation
Heat Index	Heat Index what the temperature feels like to the human body when relative humidity is combined with the air temperature.

3.2. Statistical Method

3.2.1 Correlation Analysis

The Pearson correlation and Spearman Rank order method is the most common method to measure the correlation between the variables. This method is used to measure the degree and direction with respect to the relationship between two variables. In this study, the correlation coefficient is computed for each independent variable with the response variable to check the linear relationship between them. Pearson correlation is used for computing correlation coefficient r of continuous data while Spearman correlation is used for categorical data such as the time of the day which is not normally distributed. The range of r-value is from -1 to 1 where -1 indicates negative correlation and +1 indicates positive correlation. The 0 in the r-value showed that there is no relationship between these variables. The nearer the r-value to -1 or 1, the higher the correlation between these variables [10].

The Pearson correlation formula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}} \tag{1}$$

The Spearman correlation formula:

$$\rho = 1 - \frac{6\,(\sum d_i^2)}{n\,(n^2 - 1)} \tag{2}$$

where

d_i=difference in paired ranks
n=number of cases

3.2.2 Multiple Linear Regression

Multiple linear regression (MLR) was performed to identify and examine the relationship between a dependent or criterion variable and a few independent variables, which are also called predictors. In other words, regression analysis is used to determine the effect of each independent variable on a dependent variable. In this study, this method is used to determine the factors affecting electricity consumption where the independent variables are various factors related to occupants' behaviour and the dependent variable is AC electricity consumption.

The general multiple regression model can be represented by the following equation:

$$y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots\dots + \beta_n X_n + \epsilon_i , i = 1,2,3, \dots , n \quad (3)$$

where
X_1 to X_n are independent variables
$\beta_0, \beta_1, \dots., \beta_n$ are regression coefficients
y is the predicted value of the dependent variable
ϵ is the model error (how much variation there is in predicted y)

This study used statistical software such as Excel and Minitab to analyse the dataset. The four types of variable selection in developing a multiple linear regression model are enter, forward selection, backward elimination, and stepwise selection. The stepwise method was chosen as the best option because it developed multiple linear regression without causing a multicollinearity problem [12].

3.2.2.1 Stepwise Regression

Stepwise regression consists of a series of steps adding or removing potential independence variables in succession and testing for statistical significance after each iteration to construct an effective regression model [13]. The variables will be added or removed solely based on the *t*-statistics of their estimated coefficients. The iteration will stop when there is no more potential independence variable can be entered or removed from our stepwise model, thereby the final model is developed. Stepwise variable selection method can help in identifying the potentially important independent variables and extract the best subset for predicting response variable.

The steps in stepwise regression:

1. Set an entering alpha, $\alpha_e = 0.05$, removing alpha, $\alpha_r = 0.10$
2. Fit each of the independent variable into the one-predictor models ie. y on x_1
3. If *P* value < 0.05, the variable is potential variable to enter the model
   If *P* value ≥ 0.05, the variable is rejected to enter the model
4. The potential variables with the smallest *P* value will enter the model
5. Check the *P* value of each variable in the model
   If there is a *P* value of variable more than 0.05, remove the variable
6. Repeat the step 2 to 5 with two-predictor models and so on
7. Stop the iteration when there is no more *P* value of the variable is less than 0.05

3.2.3 Hypothesis Testing

3.2.3.1 ANOVA Test

Analysis of variance (ANOVA) is used to test the significance of regression in multiple linear regression analysis. The test is used to check if there is a linear relationship exists between the dependence variable and at least one of the predictor variables. The hypotheses showed as below:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \cdots = \beta_n = 0$$

$$H_1: \beta_k \neq 0 \; for\ at\ least\ one\ k$$

The test statistic for the hypothesis of linearity:

$$F\ value = \frac{Mean\ Squared\ Regression}{Mean\ Squared\ Error} \tag{4}$$

H_0 will be rejected if the calculated F value is bigger than $F_{\alpha,k,(n-k+1)}$. Rejection of H_0 implies that the regression coefficient differs from zero. That is at least one predictor variable is significant. The significance F value showed in the ANOVA table can also compare with the significance level, α value. If the significance F value is smaller than α, then the null hypothesis can be rejected.

3.2.3.1 T-Test

The *t* test is used to check the significance of individual regression coefficients in the multiple linear regression model. This test can determine whether an additional predictor variable is significant to a regression model and makes the model more effective [14]. The hypothesis statements to test the significance of a particular regression coefficient, β_i, are:

$$H_0: \beta_j = 0$$

$$H_1: \beta_j \neq 0$$

The test statistic for this test is based on *t* distribution:

$$T_{test} = \frac{\widehat{\beta_J}}{standard\ error\ of\ \widehat{\beta_J}} \tag{5}$$

H_0 will be rejected if the *t*-value lies outside the acceptance region: $-t_{\frac{\alpha}{2},n-2} < T_{test} < t_{\frac{\alpha}{2},n-2}$. Rejection of H_0 implies that the β_j is significant to the model, Besides, we can also draw the conclusion from *p*-value. If the *p*-value is smaller than the significance level, α=0.05, the null hypothesis is rejected. Hence, we have sufficient evidence to conclude that the variable is significantly contribute to the model.

3.2.3 Residual Analysis

Residual for multiple regression model is the difference between the observed value of the dependent variable (y) and the predicted value (ŷ). Each data point has one residual. A residual plot is a graph in which the residuals are displayed on the vertical axis and the fitted value is displayed on the horizontal axis. Residual plot can be used to check the model assumptions. A linear regression model is appropriate for the data if the points in a residual plot are randomly distributed around the horizontal axis; otherwise, a nonlinear model might be more suitable.

$$e = y - \hat{y} \tag{6}$$

3.2.3 Normality Test

Normality test is used to determine if the data is normally distributed by using mathematically method (Razali & Yap, 2010). All the distribution of independent variables and dependent variables need to be investigated before starting regression analysis. The normality of the variable can be determined by using skewness and kurtosis coefficient. The variables may not be said to be skew if the skewness coefficients for all the variables are within the acceptable ±1 range [15].

Besides, the normality of the data can also show by using histogram chart. The data can say to be normally distributed if the histogram look symmetric around the mean of the distribution and has the shape of bell.

Furthermore, the normal probability plot (PP plot) is a graphical technique for determining whether or not a data set is approximately normally distributed. In PP plot, the data are plotted against a theoretical normal distribution thus the points form an approximate straight line indicates the data is normally distributed. Departures from this straight line indicate deviation from normality [16].

3.2.4 Multicollinearity Test

Multicollinearity problem occurs when independent variables in a regression model are correlated. Multicollinearity will affect the *p*-value and estimated coefficients of regression results if the degree of correlation between the variables is high. Therefore, variance inflation factor will be used to investigate the multicollinearity since it is the most commonly used diagnostic method. The variance inflation factor (VIF) can determine the strength of the association between independent variables.

$$VIF = \frac{1}{(1 - R_k^2)} \tag{7}$$

where R_k^2 is the correlation coefficient when the random variable x_k is regressed to all other random variable X.

The value of VIF start at 1 and have no upper limit. If VIF =1 indicates that there is no correlation between these independent variables. If the VIF value between 1 to 5 indicates that there is moderate correlation between these variables, but it is not severe enough to make the measures. For the VIF values more than 10, it suggests that the multicollinearity problem detected and this problem can be overcome by eliminating one of the insignificant variables among the highly correlated pair [17].

3.2.5 Mean Absolute Percentage Error (MAPE)

Mean absolute percentage error (MAPE) is a relative error measure that uses absolute values to keep the positive and negative errors from canceling one another out and uses relative errors to assess the model performance. The smaller the MAPE value, the better the model fit. The formula of MAPE is shown as below.

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{actual - forecast}{actual} \right| \tag{8}$$

3.2.5 Mean Square Error (MSE)

Mean Square Error (MSE) is the mean between actual values and predicted values. It is also known as variance of the residuals. The larger the MSE, the larger the difference between the actual values and predicted value which indicates that the model does not fit the data well. Conversely, the smaller the MSE, the better the model fit. The formula of MSE is shown as below.

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (actual - forecast)^2 \tag{9}$$

4. Results and discussion

4.1. Descriptive Statistic

Initially the correlation between these variables is studied by performing the statistical analysis and the summary is as shown in Table 2. The skewness and kurtosis coefficients of the variables has around ± 1 range which indicates the variables are still acceptable and we may say that the variables are not to be skewed.

Table 2: Summary of Statistics

Variable	N	Min	Max	Mean	Std. deviation	Kurtosis	Skewness
Energy Usage(Y)	1555	0.130	1.400	1.094	0.306	0.570	-1.389
Temp	1555	22.000	36.000	28.633	2.800	-1.123	-0.182
Humidity	1555	37.000	97.000	71.088	12.554	-0.844	-0.065
Precip	1555	0.000	4.400	0.149	0.312	42.220	5.152
Heat Index	1555	25.000	43.000	32.157	3.551	-1.146	-0.179
UV Index	1555	1.000	9.000	4.931	2.965	-1.602	-0.491

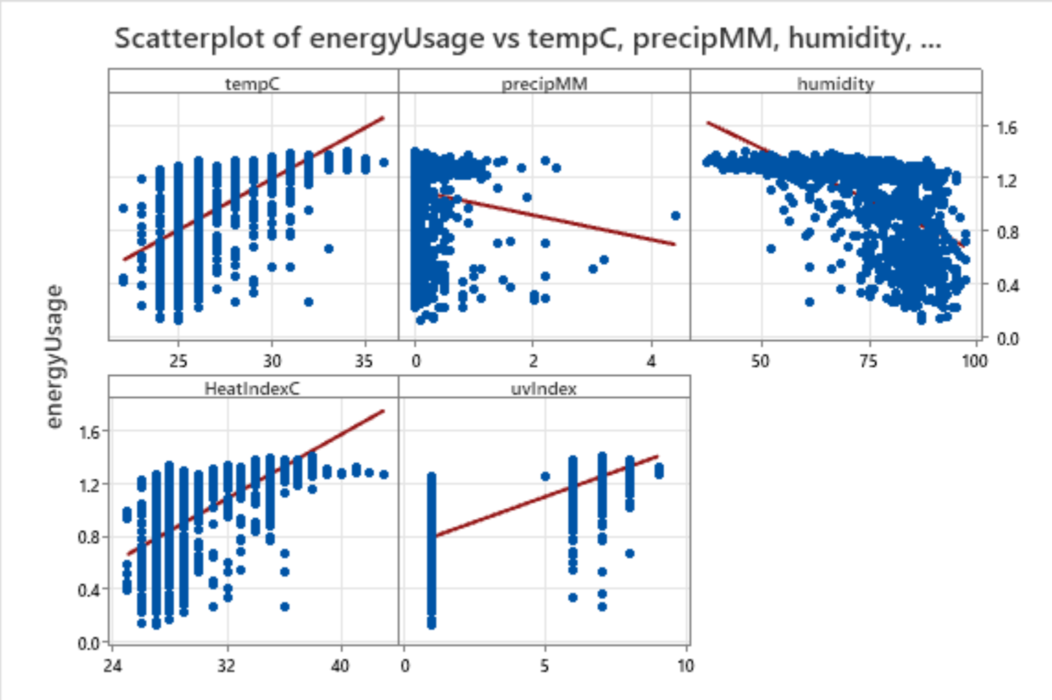


Figure 1 Scatterplot of Continuous Independence Variable and Dependence Variable

Scatterplot in Figure 1 shows there exist linear relationship between the independent variables and dependent variables, the assumption of linearity met. MLR is suitable to be used to build the model to predict the AC electricity consumption.

Table 3: Correlation Matrix

Variable	Y	X1	Temp	Humidity	Precip	Heat Index	UV Index
Y	1	0.780	0.706	-0.642	-0.095	0.705	0.743
X1	0.780	1	0.794	-0.743	0.044	0.799	0.864
Temp	0.706	0.794	1	-0.924	-0.115	0.965	0.874
Humidity	-0.642	-0.743	-0.924	1	0.247	-0.835	-0.805
Precip	-0.095	0.044	-0.115	0.247	1	-0.041	-0.065
Heat Index	0.705	0.799	0.965	-0.835	-0.041	1	0.872
UV Index	0.743	0.864	0.874	-0.805	-0.065	0.872	1

The correlation between each independent variable and AC electricity consumption is listed in Table 3. X1, temperature, heat index and UV index shows strong positive relationship with AC electricity consumption while humidity shows strong negative relationship with AC electricity consumption. The temperature has the correlation coefficient above 0.8 with humidity, heat index and UV index which means that these variables are correlated with each other.

4.2. Model Fitting

The developed model is shown as below:

$$y = 0.078 + 0.341 * X1 + 0.028 * Temperature - 0.060 * Precipitation$$

where X1 is the time of the day

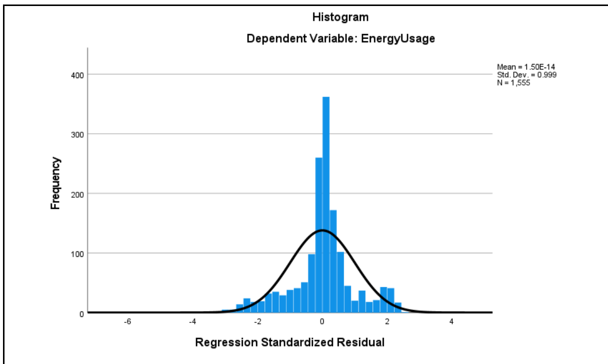
The coefficient of determination, R^2 is 0.586 which indicates that 58.6% of the variation in AC energy consumption can be explained by the time of the day, temperature and precipitation. Since the *p*-value of the regression is less than 0.05 which represents that this model is statistically significantly to predict the AC electricity consumption. Besides, all of the individual *p*-values of variable are also smaller than 0.05 which indicates that all the variables involved in this model are statistically significant contributing to the model. Furthermore, the value of VIF also investigated to ensure the model has no multicollinearity problem. The VIF value of each variable are less than 5 which indicates that there is moderate correlation between these variables but it is not severe enough to require attention. Thus, the model can be accepted.

4.3. Model Validation

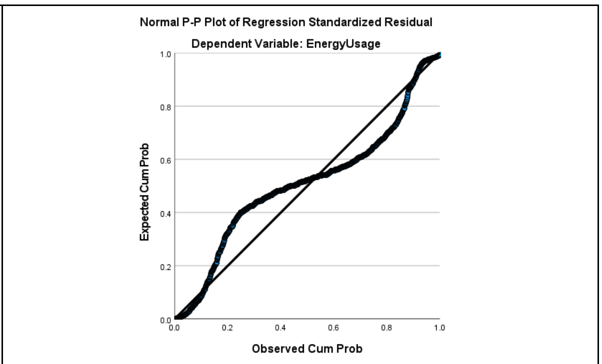
4.3.1. Performance Indicator

The MAPE value recorded as 11.9874% means that the average difference between the forecasted value and the actual value is 11. 9874%. Besides, the MSE value of 0.0307 indicates that the model has average of the square of the errors 0.0307kWh which can be consider the error is relatively small in predicting the AC electricity consumption.

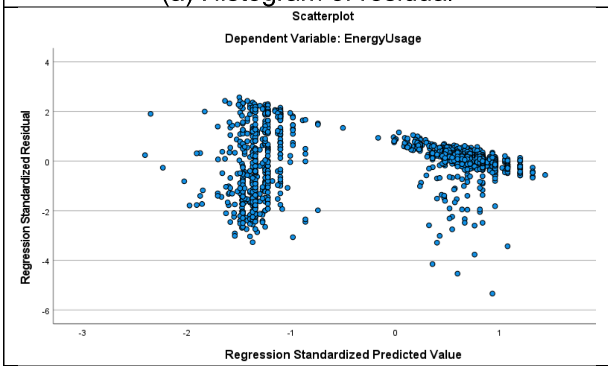
4.3.2. Residual Analysis



(a) Histogram of residual



(b) Normality plot of residual



(c) Standardized residual versus standardized predicted value plot

Figure 2 Residual Analysis

a) The mean of the histogram is $1.5 \times e^{-14}$ which is approximately to zero, we can say that assumption of the mean of error is not violated. The histogram is also showed a bell shaped which indicates that the residual is normal distributed.
b) The normal P-P plot of regression standardized residual and dependent variable showed that the error term follow normal distribution as the residual fall along the line fit.
c) The residuals scatter randomly around the horizontal axis and no particular pattern. This suggest that a linear regression model is appropriate for the data and the model met the linearity assumption. We can also say that the variance of the residual is constant.

Conclusion

The results shows that there exists relationship between the independent variable such as time of the day, temperature and precipitation with the AC electricity consumption dependent variable. These independent variables contribute significantly to the prediction model. This model can explain 58.6% variation of the AC electricity consumption. The performance of the model is evaluated on the separate testing dataset, which gives a MAPE of 11.9874% and a MSE of 0.0307 kWh. It is suggested to do another forecasting method for next research so that the obtained forecasting result could be improved. Besides, researcher also include more affecting factor in building the prediction model as the independent in the dataset now can only explain the 58.6% variation of AC electricity consumption. There might be others factors such as socio-demographic factors, house characteristics and appliance characteristics will affect the AC electricity consumption.

Acknowledgement

The researcher would like to thank a great appreciation to Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia for its support throughout this research. I am thankful to my supervisor, Dr. Fauzee Fauzee Hamdan, for taking time out of his busy schedule to read my paper and provide guidance and suggestion.

References

[1] Randazzo, T., de Cian, E., & Mistry, M. N. (2020). ‘Air conditioning and electricity expenditure: The role of climate in temperate countries’, *Economic Modelling*, 90, 273–287.

[2] Hisham, N. A., Shaikh Salim, S. A. Z., Hagishima, A., Yakub, F., & Saipol, H.F.S. (2021). ‘Statistical analysis of air-conditioning and total load diversity in typical residential buildings’. *Bulletin of Electrical Engineering and Informatics,* 10(1), 1–9.

[3] Maytham, S.A., Azah, M., Raad, Z.H., Hussain, S., & Khairuddin, K. (2017). ‘Awareness On Energy Management in Residential Buildings: A Case Study In Kajang And Putrajaya’, *Journal of Engineering Science and Technology*, Vol. 12(No. 5), 1280–1294.

[4] Perez, K. X., Cole, W. J., Rhodes, J. D., Ondeck, A., Webber, M., Baldea, M., & Edgar, T. F. (2014). ‘Nonintrusive disaggregation of residential air-conditioning loads from sub-hourly smart meter data’, *Energy and Buildings*, 81, 316–325.

[5] Davis, L. W., & Gertler, P. J. (2015) ‘Contribution of air conditioning adoption to future energy use under global warming’. *Proceedings of the National Academy of Sciences*, 112(19), 5962–5967.

[6] Douglas C. Montgomery, Elizabeth A. Peek & G. Geofferey Vining. (2012). “Introduction to Linear Regression Analysis,” 22nd edition, John Wiley and Sons, Canada.

[7] “Malaysia Energy Statistics Handbook 2020,” Suruhanjaya Tenaga (Energy Commission), 2020

[8] Kubota, T.; Jeong, S.; Toe, D.H.C.; Ossen, D.R. (2011). ‘Energy consumption and air-conditioning usage in residential buildings of Malaysia’, *Journal of International Development and Cooperation*, Vol.17, No.3,pp.61-69

[9] Ranjbara, N., Ahmad Zaki, S., Md Yusoff, N., Hagishima, A. (2017). ‘Time Series Data Analysis of Household Electricity Usage during El-Nino in Malaysia’, *Chemical Engineering Transactions.*

[10] Sena, B., Zaki, S.A., Rijal, H.B., Alfredo Ardila-Rey, J., Yusoff, N.M., Yakub, F., Ridwan, M.K. & Muhammad-Sukki, F. (2021). ‘Determinant Factors of Electricity Consumption for a Malaysian Household Based on a Field Survey’. *Sustainability 2021*, 13, 818.

[11] Xiaoyu, L., Hang, Y., Meng, W., Chaoen, L., Zi, W. & Yin, T. (2021). ‘Electricity Consumption Forecast of High-Rise Office Buildings Based on the Long Short-Term Memory Method’, *Energies 2021*, 14(16), 4785

[12] Ghani, I. M. M., & Ahmad, S. (2010). Stepwise Multiple Regression Method to Forecast Fish Landing. *Procedia - Social and Behavioral Sciences*, 8, 549–554.

[13] Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020). Multiple Linear Regression. 2nd edn. Cathie Marsh Institute Working Paper 2020-01.

[14] Hair Jr., J.F., Black, W.C., Babin, B.J. & Anderson, R.E. (2010) Multivariate Data Analysis: A Global Perspective, 7th edn, Pearson Education.

[15] Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. *Procedia - Social and Behavioral Sciences*, 106, 234–240.

[16] Chambers, J., Cleveland,W., Kleiner,B. & Tukey, P. (1983). Graphical Methods for Data Analysis, Wadsworth.

[17] Sinan, Alper & Alkan, Bilal. (2015). A useful approach to identify the multicollinearity in the presence of outliers. *Journal of Applied Statistics*. 42.