Research Article

Predictive Modelling of Energy Consumption in Malaysia: A Regression Analysis Approach

Suhaila Bahrom[1*], Aisyah Amalina Mohd Noor[2], and Anis Farehan Muhammad Fakihi[3]

1 Universiti Malaysia Pahang Al-Sultan Abdullah; suhaila_b@iium.edu.my; ORCID ID: 0009-0002-2172-9282
2 Universiti Malaysia Pahang Al-Sultan Abdullah; aisyah.amalina@gmail.com
3 Universiti Malaysia Pahang Al-Sultan Abdullah; CSM22016@umpsa.edu.my
* Correspondence: suhaila_b@iium.edu.my; 0148405251.

Abstract: *Global energy consumption is influenced by various human activities, including fossil fuel-based energy generation, household energy usage, and population growth. This case study aims to identify and predict key factors in energy consumption in Malaysia using Regression Analysis. The dataset spans from 2000 to 2020 and includes variables such as access to electricity, renewable energy capacity, electricity from renewables, access to clean cooking fuels, renewable energy share in total consumption, and primary energy consumption per capita. The R software was used to analyse the data. According to the analysis, the predictor variables that are correlated with the primary energy consumption are renewable electricity generating capacity, electricity from renewables, access to clean fuels for cooking, and renewable energy share in total final energy consumption. The findings suggest that increasing the share of renewable energy sources and improving access to clean cooking fuels could potentially reduce overall energy consumption in Malaysia. The regression model developed in this study can be a valuable tool for policymakers and energy planners to forecast future energy demand and formulate strategies to promote sustainable energy usage. Furthermore, the methodology employed can be adapted to analyze energy consumption patterns in other countries or regions, facilitating a deeper understanding of the factors driving global energy consumption.*

Keywords: regression; energy consumption; predictor.

1. INTRODUCTION

In recent years, the energy landscape has undergone significant transformations globally, driven by factors such as technological advancements, economic growth, and a growing awareness of environmental sustainability (Li & Maréchal, 2023). As Malaysia strives to meet its energy demands while addressing environmental concerns, understanding, and predicting energy consumption patterns becomes crucial for effective policymaking and sustainable development (Mahlia, 2002). This case study focuses on employing a Regression Analysis approach to predict energy consumption in Malaysia. The source of our data comes from open-source internet data extracted from the www.kaggle.com website. This dataset focuses exclusively on Malaysia, including key variables such as access to electricity, renewable electricity generating capacity per capita, electricity from renewables, access to clean fuels for cooking, renewable energy share in the total final energy consumption, and primary energy consumption per capita (*Global Data on Sustainable Energy (2000-2020)*). The dataset consists of 126 entries based on selected factors such as access to electricity, renewable electricity generating capacity per capita, electricity from renewables, access to clean fuels for cooking, renewable energy share in the total final energy consumption, and response to primary energy consumption per capita. The period of focus is from 2000 to 2020 in Malaysia.

2. METHOD & MATERIAL

Multiple Linear Regression is a statistical technique that models the relationship between a single dependent variable and two or more independent variables by fitting a linear equation to observed data (Tranmer et al., 2020). The model assumes a linear relationship between the dependent variable and the independent variables, allowing for the estimation of the impact of each independent variable while holding others constant. In general, the model of multiple linear regression is given in equation (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \tag{1}$$

y	:	predicted value of y
β_0	:	estimated value of $y - intercept$
$\beta_1, \beta_2, \ldots, \beta_k$	:	estimated value of regression coefficient

The variables used in this analysis is shown in Table 1.

Table 1. Variables

	Dependent Variable
ENERGY	Primary energy consumption per capita (kWh/person)
	Independent Variables
AE (x_1)	Access to electricity (% of population)
REG (x_2)	Renewable-electricity-generating-capacity-per-capita
ER (x_3)	Electricity from renewables (TWh)
ACC (x_4)	Access to clean fuels for cooking
RES (x_5)	Renewable energy share in the total final energy consumption (%)

In this analysis, we have made the following assumptions:

1. The relationship between dependent and independent variables are linear.
2. All the variables used in this study is Normally Distributed
3. There is no multicollinearities or little multicollinearity exist between independent variables.
4. There should be no significant outliers, high leverage points or highly influential points.
5. The residuals (errors) are approximately normally distributed.
6. Homoscedasticity, which is where the variances along the line of best fit remain similar as we move along the line.

3. FINDINGS

The summary model in Table 2 shows that there is a strong linear relationship between primary energy consumption (ENERGY) and all the independent variables with the adjusted coefficient of determination, $R^2(adjusted) = 0.8681$. We can say that 86.81% of variation in energy consumption can be predicted by access to electricity (AE), renewable-electricity-generating-capacity-per-capita (REG), electricity from renewables (ER) and renewable energy share in the total final energy consumption (RES) while 13.19% may be explained by other factors. The estimated (fitted) of the multiple linear regression model can be written as equation (2):

$$y = -562106.37 + 1030.07x_{AE} \_ 32.98x_{REG} + 810.58x_{ER} + 5009.58x_{ACC} - 3092.53x_{RES} \tag{2}$$

Table 2. Summary model for all factors

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2345.2  -680.9   220.4   642.2  2103.3

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -562106.37  427282.37  -1.316 0.208083
xAE            1030.07    3655.91   0.282 0.781982
xREG             32.98      15.71   2.099 0.053150 .
xER             810.58     220.50   3.676 0.002247 **
xACC           5009.58    2082.24   2.406 0.029485 *
xRES          -3092.53     675.44  -4.579 0.000362 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1302 on 15 degrees of freedom
Multiple R-squared:  0.9011,	Adjusted R-squared:  0.8681
F-statistic: 27.33 on 5 and 15 DF,  p-value: 4.937e-07
```

Among the predictors, x_{REG} exhibits marginal significance $(p - value = 0.0532)$ while x_{ER}, x_{ACC} and x_{RES} are statistically significant, indicating their substantial impact on the dependent variable. However, one predictor does not appear to have a statistically significant impact on the dependent variable, which is x_{AE} $(p - value = 0.7820)$. According to Table 3, there are two independent variables which are AE and REG is not statistically significant with p-value 0.7820 and 0.0532 respectively.

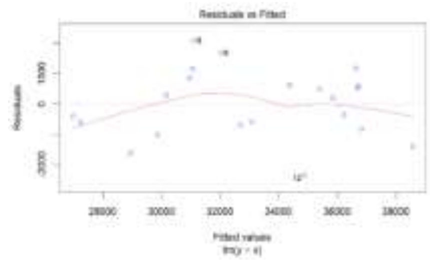

Figure 1. Plot of residual

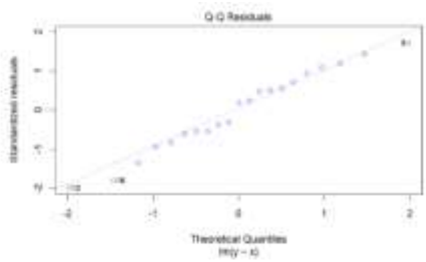

Figure 2. Q-Q plot

Figure 1 illustrates homoscedasticity occur meaning that the residuals are equally distributed across the regression line. Figure 2 indicates that the residuals lie approximately on a straight line. Therefore, the residuals are statistically normally distributed.

Table 3. Regressors Indicating the Best Data Fitting

Independent Variable	Hypothesis Testing
AE (x_1) β_1 : regression coefficient	$H_0 : \beta_1 = 0$ $H_1 : \beta_1 \neq 0$ $(p-value = 0.7820) > (\alpha = 0.05)$. Do not reject H_0. At $\alpha = 0.05$, AE is not a significant predictor for ENERGY.
REG (x_2) β_2 : regression coefficient	$H_0 : \beta_2 = 0$ $H_1 : \beta_2 \neq 0$ $(p-value = 0.0532) > (\alpha = 0.05)$. Do not reject H_0. At $\alpha = 0.05$, REG is not a significant predictor for ENERGY.
ER (x_3) β_3 : regression coefficient	$H_0 : \beta_3 = 0$ $H_1 : \beta_3 \neq 0$ $(p-value = 0.022) < (\alpha = 0.05)$. Reject H_0 At $\alpha = 0.05$, ER is a significant predictor for ENERGY.
ACC (x_4) β_4 : regression coefficient	$H_0 : \beta_4 = 0$ $H_1 : \beta_4 \neq 0$ $(p-value = 0.0295) < (\alpha = 0.05)$. Reject H_0 At $\alpha = 0.05$, ACC is a significant predictor for ENERGY.
RES (x_5) β_5 : regression coefficient	$H_0 : \beta_5 = 0$ $H_1 : \beta_5 \neq 0$ $(p-value = 0.0003) < (\alpha = 0.05)$. Reject H_0 At $\alpha = 0.05$, RES is a significant predictor for ENERGY.

In this study, we are applying best subset regression techniques to identify the best model. Table 4 shows the best subset regression and subset regression summary. The fourth model which contain independent variable REG, ER, ACC and RES have the highest R² (0.9006) and adjusted R² (0.8757) with the lowest Cp (4.0794) and AIC (365.8452). Figure 3 and Fig.4 plot shows the panel of fit criteria for best subset regression.

Table 4. Best Subset Regression

Best Subsets Regression

Model Index	Predictors
1	AE
2	ER RES
3	REG ER RES
4	REG ER ACC RES
5	AE REG ER ACC RES

Subsets Regression Summary

Model	R-Square	Adj. R-Square	Pred R-Square	C(p)	AIC	SBIC	SBC	MSEP	FPE	HSP	APC
1	0.7519	0.7388	0.6975	20.6289	379.0485	317.2940	382.1821	70522772.9830	3676205.9953	188474.2172	0.3003
2	0.8453	0.8281	0.7848	8.4642	371.1302	310.8826	375.3083	46562527.8680	2524921.7664	129959.2086	0.2063
3	0.8628	0.8386	0.7986	7.8061	370.6054	311.0044	375.8280	41868284.6980	2469365.8710	129641.7082	0.2017
4	0.9006	0.8757	0.8064	4.0794	365.8452	309.8677	372.1123	33913485.5690	1977610.5520	106486.7220	0.1616
5	0.9011	0.8681	0.744	6.0000	367.7343	312.6189	375.0460	36144587.2467	2179051.6713	121058.4262	0.1780

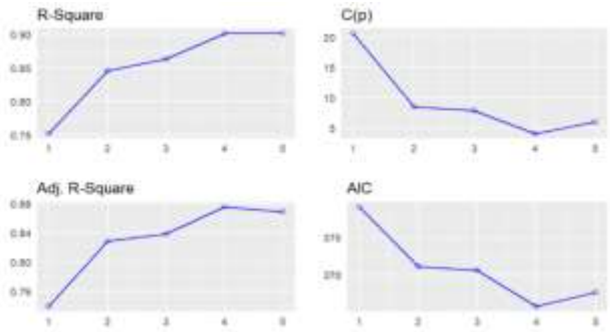

Figure 3. Plot of fit criteria for best subset regression

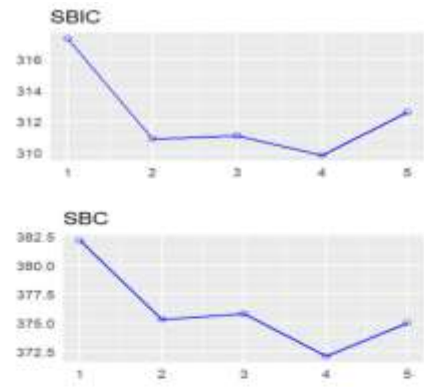

Figure 4. Best subset regression plot

Table 5. Summary model for all factors

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-2363.7  -731.3   164.6   840.5  2051.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -456449.79  198833.59  -2.296 0.035549 *
xREG             35.41      12.73   2.781 0.013351 *
xER             840.43     187.75   4.476 0.000382 ***
xACC           4973.45    2017.61   2.465 0.025397 *
xRES          -3219.19     489.39  -6.578 6.36e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1264 on 16 degrees of freedom
Multiple R-squared:  0.9006,	Adjusted R-squared:  0.8757
F-statistic: 36.23 on 4 and 16 DF,  p-value: 7.834e-08
```

From the analysis, the estimated of the multiple linear regression model is given in equation (3):

$$y_{ENERGY} = -456449.76 + 35.41(x_{REG}) + 840.43(x_{ER}) + 4973.45(x_{ACC}) - 3219.19(x_{RES}) \quad (3)$$

4. DISCUSSION

Table 5 shows the summary for the final model selection. The linear regression output reveals key insights into the relationship between the dependent variable y_{ENERGY} and the predictor variable x_{REG}, x_{ER}, x_{ACC} and x_{RES}. The intercept, $\beta_0 = 456449.79$, represents the estimated value of energy when all predictors are zero. The positive coefficient for $x_{REG}\,(35.41)$ suggests that for each unit increase in x_{REG}, y_{ENERGY} is expected to increase by 35.41 units. Similarly, x_{ER} has a positive coefficient of 840.43, indicating a substantial impact on y_{ENERGY} for each unit increase. The coefficient for x_{ACC} is 4973.45, implying a positive influence on y with increasing x_{ACC} while the negative coefficient for $x_{RES}\,(-3219.19)$ suggests a negative impact on y_{ENERGY} with rising x_{RES}.

The statistical significance of the coefficients is denoted by the p-values. All predictors, have p-values less than 0.05, indicating their statistical significance in predicting primary energy consumption. The coefficient of determination, R^2 is 0.9006 while the adjusted R^2 is 0.8757. We can say that 87.57% of variation in primary energy consumption can be predicted by renewable electricity generating capacity

(REG), electricity from renewables (ER), access to clean fuels for cooking (ACC) and renewable energy share (RES).

5. CONCLUSION

As a conclusion, the statistical data analysis shows that the primary energy consumption in Malaysia can be predicted using multiple linear regression model. According to the analysis, the predictor variables that are correlated with the primary energy consumption are renewable electricity generating capacity, electricity from renewables, access to clean fuels for cooking, and renewable energy share in total final energy consumption.

References

Global Data on Sustainable Energy (2000-2020). Retrieved May 13, 2024, from https://www.kaggle.com/datasets/anshtanwar/global-data-on-sustainable-energy

Li, X., & Maréchal, F. (2023). Deep excavation of the impact from endogenous and exogenous uncertainties on long-term energy planning. *Energy and AI*, 11, 100219. https://doi.org/10.1016/j.egyai.2022.100219

Mahlia, T. M. I. (2002). Emissions from electricity generation in Malaysia. *In Renewable Energy* (27). www.elsevier.com/locate/renene

Tranmer, M., Murphy, J., Elliot, M., & Pampaka, M. (2020). *Multiple Linear Regression* (2nd Edition). https://hummedia.manchester.ac.uk/institutes/cmist/a