Supplementary data 2

Testing assumptions for regression modelling

1. Existence of a linear relationship between dependent and independent variables

This is discussed in the main article in section 3.2.

2. Variables (dependent and independent) are normally distributed

Taking the dependent variable as the logarithm of electricity demand (log*ED*), a normality test was conducted using SPSS. Altogether, there were 55 valid cases while conducting the normality test on the dependent variable. In view of Table S1, both the Shapiro-Wilk and Kolmogorov-Smirnov tests have p-values greater than 0.05, indicating that the outcome variable is normally distributed.

Table S1 Tests of normality for the dependent variable and independent variables.

Tests of normality for dependent variable						
	Kolmogorov-Smirnov			Shapiro-Wilk		
log*ED*	Statistic	df	Sig.	Statistic	df	Sig.
	0.080	55	0.200	0.979	55	0.426

Tests of normality for all variables						
	Kolmogorov-Smirnov			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
log*ED*	.080	55	.200	.979	55	.426
log*FA*	.094	55	.200	.963	55	.084
log*SN*	.073	55	.200	.971	55	.198
log*LGT*	.104	75	.200	.975	55	.295
log*AC*	.189	55	.000	.904	55	.000

While testing the normality for all variables simultaneously, it is seen from Table S1 that all variables are normally distributed except for log*AC*. Li et al. [1] reports that the validation of normality can sometimes be ignored in linear regression models because the errors of the linear regression model, and not the dependent variable, are to be normally distributed. Also, if one has a large data set with records of more than 20, which is the case in this study, then validation of normality can be ignored.

3. Absence of outlier in the regression model

Significant outliers and influential data points can place undue influence on a model, making it less representative of the data as a whole [2]. To test outliers in the overall regression model, Cook's distance was saved in the regression model. In the "Enter" method, the maximum Cook's distance ranged from 0.142 to 0.191 for different regression models as seen in Table 4. Also, the maximum Cook's distance was 0.233 for Stepwise and Backward regression. Because the maximum Cook's distance was less than 1, it indicates that there were no influential data points in the regression model.

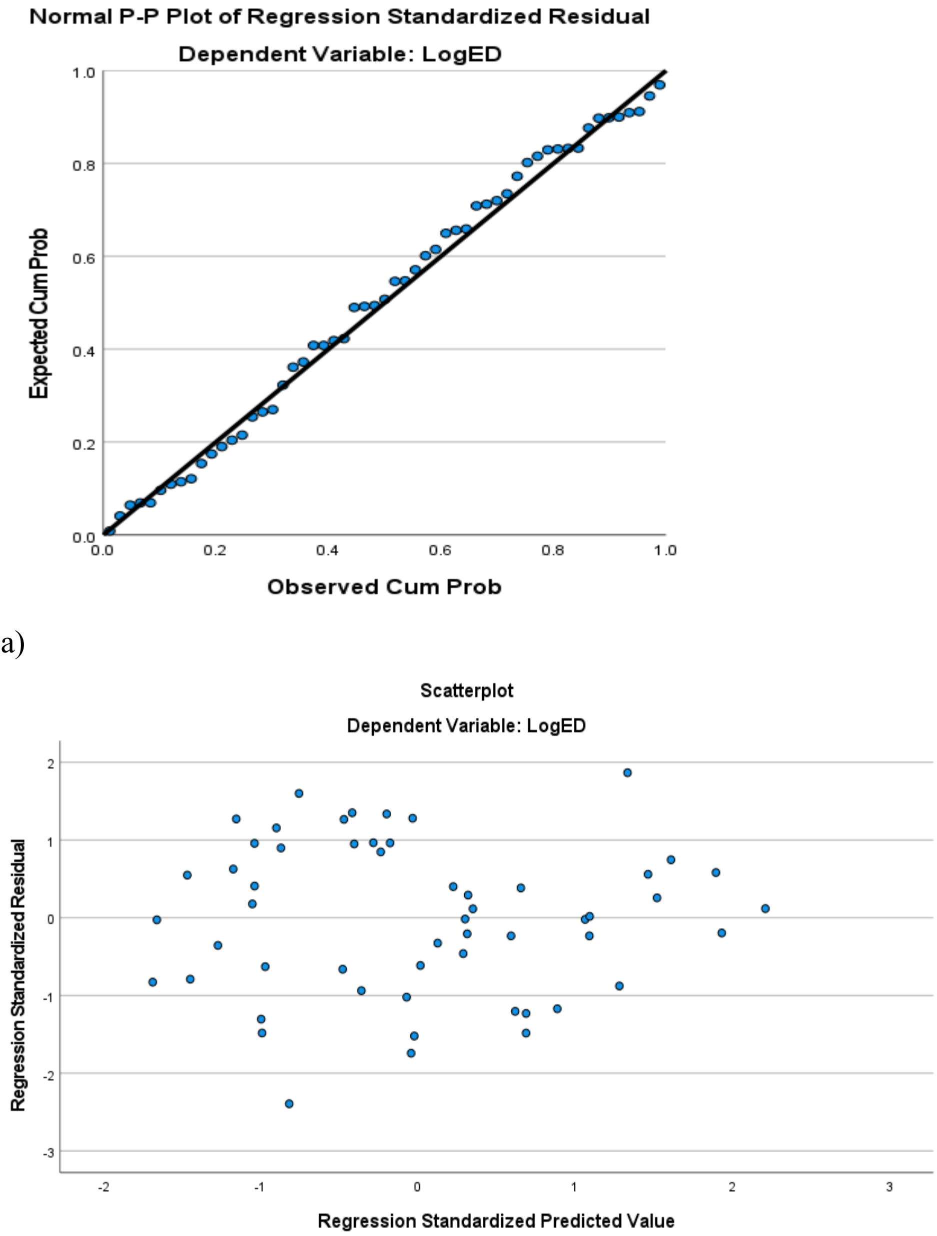
4. Absence of multi-collinearity between independent variables

Multi-collinearity is when two or more independent variables are highly correlated. If multi-collinearity exists, then changes in one independent variable will affect another independent variable and this will give rise to errors in one's regression model. If Pearson's coefficient is more than 0.8 between independent variables, the independent variables are collinear. Analysis shows that Pearson's coefficient is less than 0.8 for all independent variables considered for MLR, indicating that they are not collinear. This is also confirmed in the collinearity statistics. The variance inflation factor (VIF) for all independent variables was less than 10, indicating that multi-collinearity does not exist.

In addition, the independent variable must have a high correlation with the dependent variable. Hence, the independent variables considered for MLR have a high positive Pearson’s correlation coefficient with dependent variable $\log ED$; the total number of lights ($\log LGT$) has the highest correlation coefficient of 0.790, followed by student number ($\log SN$) with 0.710 of Pearson’s coefficient and 0.692 and 0.686 for $\log FA$ and $\log AC$ respectively.

5. Residuals are normally distributed and homoscedastic

Residual is the difference between the predicted value from the regression model and the observed value of the outcome (or dependent variable) in the data set. The probability-probability (P-P) plot of the standardised residuals for the dependent variables is used to test if the residuals are normally distributed. From Figure S1a, it is seen that the residuals are lying close to the diagonal normal line in the P-P plot indicating that the residuals are normally distributed. This means that the regression model is able to explain all trends in the dataset. Similar figures are also obtained for other types of regression methods (Enter and Backward) used in this study.



(a)

(b)

Figure S1. (a) Normal P-P Plot of Regression Standardized Residual and (b) scatter plot of standardised residual again standardised predicted value for stepwise linear regression.

Homoscedasticity means that variance of errors is the same across all levels of the independent variables [3]. From the scatter plot shown in Figure S1b, the spread of the residuals is fairly constant because, for each point of the linear regression model, the standard residuals are between -2.393 and 1.936, which is within 3 standard deviations. This indicates that there is the presence of homoscedasticity, that is, the variance of errors is the same across all independent variables.

In addition, considering the Durbin-Watson statistic of the "Enter" method for various independent variables taken, the statistic ranges from 1.502 to 1.840 which is close to value 2, indicating that the residuals are uncorrelated, that is, individual data points are independent of each other. Similarly, the Durbin-Watson statistic for stepwise regression was 1.676 and for backward regression the Durbin-Watson statistic was also 1.676.

References

[1] Li X, Wong W, Lamoureux EL, Wong TY. Are linear regression techniques appropriate f-or analysis when the dependent (outcome) variable is not normally distributed?Investigative ophthalmology & visual science. 2012;53(6):3082-3.
[2] OU. Assumptions of Multiple Regression. https://www.open.ac.uk/socialsciences/spsstutorial/files/tutorials/assumptions.pdf [Accessed on 3rd October 2022]: The Open University; 2022.
[3] Osborne JW, Waters E. Four assumptions of multiple regression that researchers always test. Practical assessment, research, and evaluation.