International Journal of Innovative Mathematics, Statistics & Energy Policies
12(3):22-42, July-Sept., 2024

 www.seahipublications.org **ISSN: 2467-852X**

Selection Methods And Unbiased Regression Model For Rumuosi Electricity Consumption, Rivers State, Nigeria

Biu, O. Emmanuel[1]., Ligeiaziba, Sylva[2] & Ofili, I. John[3]

[1]Department of Mathematics and Statistics,
University of Port Harcourt, Port Harcourt, Nigeria

[2]Department of Mathematics/Statistics,
Bayelsa Medical University, Yenagoa, Nigeria

[3]Department of Mathematics/Statistics,
Ignatius Ajuru University of Education, Port Harcourt, Nigeria

ABSTRACT
This research work was to provide an appropriate and unbiased regression model for Rumomasi electricity consumption, Rivers State, using model selection methods. Rumuosi average megawatt electricity consumption is the dependent variable (Y_i), and the independent variables are Maximum load electricity consumption (MLEC) "X_1", Average load electricity consumption (ALEC) "X_2", Hours available electricity consumption (HAEC) "X_3" and Hours on outage electricity consumption (HOEC) "X_4". The descriptive statistics of daily hourly consumption to monthly hourly consumption were estimated for each variable, from 2016 to 2020. The four selection methods to estimating suitable regression models (Best Subsets Regression, Stepwise Selection, Forward Selection and Backward Elimination) results were computed. The best model was identified using R-square (R^2), Adjusted R^2 and Variance Inflation Factor (VIF). The purpose in this research was to use model selection methods to build a suitable model for forecasting purpose of the Rumomasi electricity consumption. The study was able to estimate significant parameters for some of the independent variables using R^2, Adjusted R^2 and VIF. It was discovered that Maximum Load Electricity Consumption (MLEC), Average Load Electricity Consumption (ALEC) and Hours Available Electricity Consumption (HAEC) have a statistically significant impact on the Rumuosi average megawatt electricity consumption. Indicating the three predictor's model whose variable are Maximum Load Electricity Consumption (MLEC), Average Load Electricity Consumption (ALEC) and Hours Available Electricity Consumption (HAEC) is the best model for effective and reliable forecasting of electricity consumption.
Keywords: Regression, Electricity, Consumption

INTRODUCTION
Multiple regression analysis, a term first used by Karl Pearson (1908), is an extremely useful extension of simple linear regression in that we use several quantitative (metric) or dichotomous variables in combination rather than just one such variable to predict or explain the value of a quantitatively measured

criterion (outcome/dependent) variable. Most researchers believe that using more than one predictor or potentially explanatory variable can paint a more complete picture of how the world works than is permitted by simple linear regression because behavioral scientists generally believe that behavior, attitudes, feelings, and so forth are determined by multiple variables rather than just one. Using only a single variable as a predictor or explanatory variable as is done in simple linear regression will at best capture only one of those sources. In the words of one author Thompson (1991), multivariate methods such as multiple regression analysis have accrued greater support in part because they "best honor the reality to which the researcher is purportedly trying to generalize". Based on what have already discussed regarding simple linear regression, it may be clear that multiple regression can be used for predictive purposes, such as estimating from a series of entrance tests how job applicants might perform on the job. But the regression technique can also guide researchers toward explicating or explaining the dynamics underlying a particular construct by indicating which variables in combination might be more strongly associated with it. In this sense, the model that emerges from the analysis can serve an explanatory purpose as well as a predictive purpose. As was true for simple linear regression, multiple regression analysis generates two variations of the prediction equation, one in raw score or unstandardized form and the other in standardized form (making it easier for researchers to compare the effects of predictor variables that are assessed on different scales of measurement). These equations are extensions of the simple linear regression models and thus still represent linear regression, that is, they are still linear equations but use multiple variables as predictors. The main work done in multiple regression analysis is to build the prediction equation. This primarily involves generating the weighting coefficients—the b (unstandardized) coefficients for the raw score equation and the beta (standardized) coefficients for the standardized equation. This prediction model informs us that if we weight each of the predictors as the statistical analysis has indicated, then we can minimize our error in predicting the dependent variable.

Eric (2021), multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable. In essence, multiple regression is the extension of ordinary least-squares (OLS) regression because it involves more than one explanatory variable.

The coefficient of determination (R-squared) is a statistical metric that is used to measure how much of the variation in outcome can be explained by the variation in the independent variables. R^2 always increases as more predictors are added to the MLR model, even though the predictors may not be related to the outcome variable.

R^2 by itself can't thus be used to identify which predictors should be included in a model and which should be excluded. R^2 can only be between 0 and 1, where 0 indicates that the outcome cannot be predicted by any of the independent variables and 1 indicates that the outcome can be predicted without error from the independent variables.

Different regression models or several different regression models exist that are optimized depending on the nature of the data being analyzed and the type of question being asked.

The most common example is the ordinary least-squares (OLS) regression, which can be conducted on several types of cross-sectional or time-series data. If you're interested in a binary (yes-no) outcome. For instance, how likely you are to be fired from a job based on your productivity you can use a logistic regression or a probit model. Today, there are hundreds of models that an econometrician has at his disposal.

Econometrics is sometimes criticized for relying too heavily on the interpretation of raw data without linking it to established economic theory or looking for causal mechanisms. It is crucial that the findings revealed in the data are able to be adequately explained by a theory, even if that means developing your own theory of the underlying processes.

Regression analysis also does not prove causation, and just because two data sets show an association, it may be spurious. For example, drowning deaths in swimming pools increase with GDP. Does a growing economy cause people to drown? Of course not, but perhaps more people buy pools when the economy is booming. Econometrics is largely concerned with correlation analysis, and remember, correlation does not equal causation.

Stepwise regression involves selection of independent variables to use in a model based on an iterative process of adding or removing variables. Variance Inflation Factor (VIF) Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. The aim of this study is to be able to provide an appropriate and unbiased regression model for Rumuomasi electricity consumption, Rivers State, using model selection methods and analyze the different factors that contribute to daily electricity consumption and energy demand. The objectives are to compute the descriptive statistics of daily hourly consumption to monthly hourly consumption for each year, use the four selection methods to estimate suitable regression models (Best subsets regression, Stepwise Selection, Forward Selection and Backward Elimination) and determine the best model using R^2, Adjusted R^2 and VIF.

Definition of Operational Definition of Terms

Electricity Consumption - Electric energy consumption is the actual energy demand made on existing electricity supply.

Hour on Outage- Forced Outage Hour means an hour during the whole of which electricity is not produced due to Forced Outage and not Downtime

Daily Maximum Level - Maximum or Peak Load. It represents the maximum power consumed by the load during a specific period of time. It also represents the maximum.

VIF- variance inflation factor is the ratio of the variance of estimating some parameter in a model that includes multiple other terms by the variance of a model constructed using only one term. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

Household- A household consists of one or several persons who live in the same dwelling and share meals. It may also consist of a single family or another group of people. The household is the basic unit of analysis in many social, microeconomic and government models, and is important to economics and inheritance

MATERIALS AND METHODS

This section focused on research design, population of the study, sample and sampling techniques, instrument for data collection, validity and reliability of the instruments, procedure for data collection and method of data collection.

Research Design

The econometric model tends to explore statistical relationship between various factors that determine energy. This study considered a mixed methodology including an econometric analysis using secondary data (10 years of data observation). This study explored the use of the multiple regressions in terms of modeling electricity, estimating the electricity, and forecasting future electricity consumptions in other to achieve this, the electricity demand function for the residential sector with estimated using secondary annual data for the period 2010 – 2020. Energy was considered as a commodity whereby its price acts as a key determinate for energy consumption.

Population of the Study

The population of this study will consist electricity consumption using Port Harcourt Electricity Distribution Company (PHED) focused mainly on six (6) local government in the state for the period of five month each.

Sample and Sampling Techniques

The sample of the study comprised of electricity consumption; the researcher used techniques which are meant for testing the adequacy of regression model.

Data Collection Procedure

The dataset used for this work is a secondary data, collected from Port Harcourt Electricity Distribution Company (PHED).

Models Specification

The following are the variable that the researcher will put into consideration, which are majorly 4 (four) variable called independent variables, that is Maximum Load Electricity Consumption (MLEC) "X_1", Average Load Electricity Consumption (ALEC)"X_2", Hours Available Electricity Consumption (HAEC)"X_3" and Hours on Outage Electricity Consumption (HOEC) "X_4",. While the dependent variable is Daily Average Megawatt; Daily AV MW (Y).

Then, the Actual and Semi-logarithm model are given by

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.1)$$

and

$$Ln(\hat{Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.2)$$

where β_0, β_1, β_2, β_3, β_4 are coefficients of the Models

The four model selection models are best subsets regression, Stepwise Selection, Forward Selection and Backward Elimination.

Best subsets regression: The best subsets regression is **a** model selection approach that consists of testing all possible combination of the predictor variables, and then selecting the best model according to some statistical criteria.

Stepwise Selection: In statistics, stepwise regression is a method of fitting regression models in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some pre specified criterion.

Forward Selection: Forward selection is a type of stepwise regression which begins with an empty model and adds in variables one by one. In each forward step, you add the one variable that gives the single best improvement to your model.

Backward Elimination: Backward elimination is a statistical procedure that removes the least important variables in a step-wise manner, leaving only the most important ones.

Note that the possible models to choose from best subsets regression method are

$$\text{Models} = 2^p - 1 \quad (1.3)$$

$where$, p is number of independent variables

Since we have 4 predictors (or independent variables), then

$$\text{Models} = 2^p - 1 = 15$$

The fifteen model are

$$\hat{Y} = \beta_0 + \beta_1 X_1 + e_i \quad (1.4)$$

$$\hat{Y} = \beta_0 + \beta_2 X_2 + e_i \quad (1.5)$$

$$\hat{Y} = \beta_0 + \beta_3 X_3 + e_i \quad (1.6)$$

$$\hat{Y} = \beta_0 + \beta_4 X_4 + e_i \quad (1.7)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i \quad (1.8)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + e_i \quad (1.9)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_4 X_4 + e_i \quad (1.10)$$

$$\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + e_i \quad (1.11)$$

$$\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_4 X_4 + e_i \quad (1.12)$$

$$\hat{Y} = \beta_0 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.13)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + e_i \quad (1.14)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + e_i \quad (1.15)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.16)$$

$$\hat{Y} = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.17)$$

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + e_i \quad (1.18)$$

The other techniques are

3.6 Testing for Multicollinearity using Variance Inflation Factor (VIF)

Wonsuk et al. (2014) defined variance inflation factor (VIF) as a measure of how much the variance of the estimated regression coefficient b_i is "inflated" by the existence of correlation among the predictor variables in the model. According to the author, a VIF of 1 means that there is no correlation among the i^{th} predictor and the remaining predictor variables, and hence the variance of b_i is not inflated at all. The general rule of thumb is that VIFs exceeding 10 are signs of serious multicollinearity requiring correction. The test of multicollinearity is necessary to ascertain the independence of regression variables.

The variance inflation factor for the ith suspected mediator variable is given by:

$$VIF = \frac{1}{1-R^2} \quad (1.19)$$

where R^2 is the coefficient of determination**.**

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{SSR}{SST} \qquad 0 \le R^2 \le 1 \quad (1.20)$$

where Total sum of squares (Total Variation): $SST = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ (1.21)

Regression sum of squares (Explained Variation): $SSR = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2$ (1.22)

RESULTS

Descriptive Statistics

The data sets collected were arranged from daily hourly consumption to monthly hourly consumption for each year (2016 to 2020). The statistics in terms of mean (average), variance, minimum, maximum etc. are computed in Table 1.1 to 1.5.

Table 1.1: Descriptive Statistics of Monthly Rumuosi Electricity Consumption for 2016.

STATISTICS	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Range	Minimum	Maximum
Daily Maximum	14.2	0.4	15.2	8.0	6.4	40.8	29.3	0.0	29.3
Daily Average	11.2	0.3	11.3	0.0	5.2	26.8	24.3	0.0	24.3
Hours Available	20.9	0.4	24.0	24.0	5.9	34.9	24.0	0.0	24.0
Hours on Outage	3.1	0.4	0.0	0.0	5.9	34.9	24.0	0.0	24.0
Daily Average Megawatt Hour.MWH	249.4	8.0	243.0	0.0	132.3	17508.4	584.1	0.0	584.1

Table 1.2: Descriptive Statistics of Monthly Rumuosi Electricity Consumption for 2017.

STATISTICS	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Range	Minimum	Maximum
Daily Maximum	18.0	0.6	18.1	0.0	10.1	102.4	126.2	0.0	126.2
Daily Average	9.2	0.2	9.5	0.0	3.0	9.1	15.5	0.0	15.5
Hours Available	19.2	0.4	22.0	24.0	7.0	49.2	24.0	0.0	24.0
Hours on Outage	4.8	0.5	2.0	0.0	7.0	49.2	24.0	0.0	24.0
Daily Average Megawatt Hour.MWH	182.3	5.2	194.0	0.0	81.5	6634.3	355.9	0.0	355.9

Table 1.3: Descriptive Statistics of Monthly Rumuosi Electricity Consumption for 2018.

STATISTICS	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Range	Minimum	Maximum
Daily Maximum	20.4	0.2	20.6	19.7	4.0	16.0	29.1	0.9	30.0
Daily Average	9.5	0.2	9.7	8.4	2.6	6.9	15.7	0.7	16.5
Hours Available	22.9	0.1	24.0	24.0	2.3	5.1	15.0	9.0	24.0
Hours on Outage	1.1	0.1	0.0	0.0	2.3	5.1	15.0	0.0	15.0
Daily Average Megawatt Hour.MWH	218.6	3.8	222.3	200.8	65.3	4263.2	388.5	6.6	395.1

Table 1.4: Descriptive Statistics of Monthly Rumuosi Electricity Consumption for 2019.

STATISTICS	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Range	Minimum	Maximum
Daily Maximum	20.2	0.2	20.5	24.2	4.3	18.1	48.0	0.0	48.0
Daily Average	9.4	0.1	9.4	7.7	2.6	6.9	16.4	0.0	16.4
Hours Available	22.6	0.2	24.0	24.0	3.2	10.2	24.0	0.0	24.0
Hours on Outage	1.3	0.2	0.0	0.0	3.1	9.5	24.0	0.0	24.0
Daily Average Megawatt Hour.MWH	215.2	3.5	211.9	166.6	66.6	4435.3	393.0	0.0	393.0

Table 1.5: Descriptive Statistics of Monthly Rumuosi Electricity Consumption for 2020.

STATISTICS	Mean	Standard Error	Median	Mode	Standard Deviation	Sample Variance	Range	Minimum	Maximum
Daily Maximum	19.9	0.4	21.8	10.0	7.5	55.7	52.2	0.0	52.2
Daily Average	11.4	0.2	11.9	0.0	4.3	18.9	26.8	0.0	26.8
Hours Available	21.5	0.2	23.0	24.0	4.7	21.7	24.0	0.0	24.0
Hours on Outage	2.4	0.2	0.0	0.0	4.7	21.7	24.0	0.0	24.0
Daily Average Megawatt Hour.MWH	255.2	5.9	271.3	0.0	113.2	12809.2	644.3	0.0	644.3

Table 1.1 to 1.5 shows the distribution of the Rumuosi electricity consumption during the year 2016 to 2020. Furthermore, the expected (mean) electricity consumption was summarized in Table 1.6 and Figure 1.1 for comparison purpose.

Table 1.6: Summary of the Expected (Mean) Monthly Rumuosi Electricity Consumption from 2016 to 2020

STATISTICS	Mean 2016	Mean 2017	Mean 2018	Mean 2019	Mean 2020
Maximum Load Electricity Consumption	14.2	18	20.4	20.2	19.9
Average Load Electricity consumption	11.2	9.2	9.5	9.4	11.4
Hours Available Electricity Consumption	20.9	19.2	22.9	22.6	21.5
Hours on Outage Electricity Consumption	3.1	4.8	1.1	1.3	2.4
Daily Average Megawatt Hour.MWH	249.4	182.3	218.6	215.2	255.2

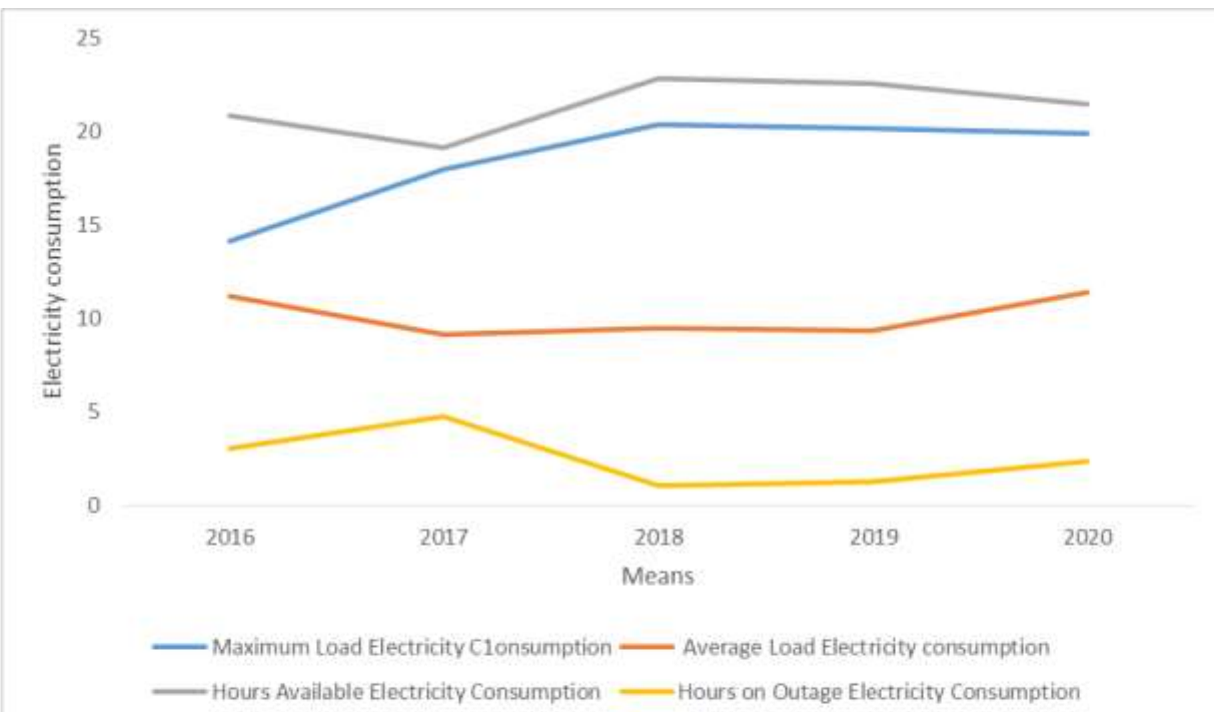

Figure 1.1: Line Graph of Rumuosi Electricity Consumption from 2016 to 2020

From Figure 1.1, the maximum load electricity consumption appreciated from the year 2016 to 2018, then showed a slight depreciation from 2018 to 2020. The hour available of electricity consumption begins to decrease from the year 2016 to 2017, then appreciated from 2017 to 2018 and it has a slight decrease from 2018 to 2020. However, the average load electricity consumption shows a constant behavior, while hour of outage only increases at the year 2017.

Selection Methods and Model Identification

Best subsets regression, methods Stepwise Selection, Forward Selection and Backward Elimination was adopted.

Model Identification for Rumuosi Electricity Consumption 2016

The best subsets regression method results for Actual and Semi-logarithm model of the dependent variable (Daily Average Megawatt; Daily AV MW) on the independent variables [Maximum Load Electricity Consumption (MLEC), Average Load Electricity Consumption (ALEC), Hours Available Electricity Consumption (HAEC), Hours on Outage Electricity Consumption (HOEC)].

Table 1.7: Best Subsets Regression for Actual Model, 2016

Vars	R-Sq	R-Sq(adj)	(pred)	MallowsCp	S	MLEC	ALEC	HAEC	HOEC
1	89.5	89.4	89.3	165.0	43.054		X		
1	83.8	83.7	83.6	398.0	53.376	X			
2	**93.4**	**93.4**	**92.9**	**2.7**	**34.003**		**X**	**X**	
2	93.4	93.4	92.9	2.7	34.007		X		X
3	93.5	93.4	92.8	3.0	33.960	X	X	X	
3	93.5	93.4	92.8	3.1	33.963	X	X		X
4	93.5	93.4	87.8	5.0	34.023	X	X	X	X

Table 1.8: Best Subsets Regression for Semi log Model, 2016

Vars	R-Sq	(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	68.1	68.0	67.7	1487.9	0.31422		X		
1	65.2	65.1	64.7	1647.0	0.32822	X			
2	**95.3**	**95.2**	**94.9**	**2.6**	**0.12117**		**X**	**X**	
2	95.3	95.2	94.9	3.1	0.12127		X		X
3	95.3	95.3	94.9	3.0	0.12102	X	X	X	
3	95.3	95.2	94.9	3.4	0.12112	X	X		X
4	95.3	95.2	92.2	5.0	0.12126	X	X	X	X

where Y_i is Rumuosi average megawatt electricity consumption, X_1 is Maximum load electricity consumption (MLEC), X_2 is Average load electricity consumption (ALEC), X_4 is Hours available electricity consumption (HAEC), X_4 is Hours on outage electricity consumption (HOEC).

Table 1.9: Fitted Regression Models for the different Selection Methods for the Year 2016

Selection Methods	Variables	Coefficients	Actual Model Estimates (p-value)	VIF	Semi-Log Model Estimates (p-value)	VIF
Stepwise	Constant (Intercept)	β_0	-0.970 (0.000**)		2.9024 (0.000**)	
	Average Load (X_2)	β_2	21.403 (0.000**)	1.29	0.08139 (0.000**)	1.06
	Hours Available (X_3)	β_3	5.087 (0.000**)	1.29	0.07216 (0.000**)	1.06
	R^2		93.44%		95.28%	
	$\overline{R}^2$		93.40%		95.25%	
Backward	Constant (Intercept)	β_0	-0.980 (0.000**)		2.9001 (0.000**)	
	Maximum Load (X_1)	β_1	1..530 (0.195)	13.41	0.00612 (0.204)	14.18
	Average Load (X_2)	β_2	21.403 (0.000**)	13.70	0.07411 (0.000**)	14.14
	Hours Available (X_3)	β_3	5.087 (0.000**)	1.29	0.07204 (0.000**)	1.07
	R^2		93.49%		95.31%	
	$\overline{R}^2$		93.41%		95.26%	
Forward Selection	Constant (Intercept)	β_0	-0.970 (0.000**)		2.9024 (0.000**)	
	Average Load (X_2)	β_2	21.403 (0.000**)	1.29	0.08139 (0.000**)	1.06
	Hours Available (X_3)	β_3	5.087 (0.000**)	1.29	0.07216 (0.000**)	1.06
	R^2		93.44%		95.28%	
	$\overline{R}^2$		93.40%		95.25%	

Footnote: **= p-values less than 5%.

Best subsets regression for actual and semi-log model were done in Table 1.10 and 1.11 respectively, it's identified five suitable models with R^2 of 85.9% and 91.5 for actual and semi-log model. Mallow' Cp value is 4.0 for both models (actual and semi-log model). The three predictors' model with MLEC, ALEC and HAEC is identified to be best model.

Table 1.12 showed results of the other three methods (Stepwise Selection, Forward Selection and Backward Elimination) p-values for the estimated parameters of the independent variables. The three methods p-value of the models' parameters are significant at an α level of 0.05ALEC and HAEC are both 0.000 for the two models, indicating that they are significantly related to Rumuosi average megawatt electricity consumption.

In addition, the VIF values for both models after applying stepwise selection, forward Selection and backward elimination methods are all close to 1, which indicates that the predictors are not correlated in both models.

In Table 4.11, the R^2values indicates that the predictors explain 87.1% to 90.5% of the variance in Rumuosi average megawatt. The three methods results indicate that the model with three predictors' of MLEC, ALEC and HAEC fits the data for 2017 very well, just as the result of Best subsets regression. Hence, the three predictor's model has adequate predictive ability. That is

$$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \text{ for both models} \tag{2.2}$$

Model Identification for Rumuosi Electricity Consumption 2018

Table 1.13: Best Subsets Regression for Actual Model, 2018

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	90.2	90.2	90.1	1611.8	20.485	X			
1	34.8	34.6	34.0	12403.8	52.817	X			
2	98.4	98.4	98.1	6.3	8.1745		X	X	
2	90.2	90.1	90.0	1610.1	20.499	X	X		
3	98.5	98.4	98.0	4.0	8.1303	X	X	X	

Table 1.14: Best Subsets Regression for Semi log Model, 2018

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	76.5	76.4	75.3	533.8	0.19236		X		
1	39.4	39.2	36.6	1851.0	0.30907	X			
2	90.9	90.8	89.1	25.9	0.11997		X	X	
2	77.9	77.8	75.8	485.4	0.18675	X	X		
3	91.6	91.5	89.5	4.0	0.11564	X	X	X	

Table 1.15: Fitted Regression Models for the different Selection Methods for the Year 2018

Selection Methods	Variables	Coefficients	Actual Model Estimates (p-value)	VIF	Semi-Log Model Estimates (p-value)	VIF
Stepwise	Constant (Intercept)	β_0	18.25(0.000**)		2.5371(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.306(0.010**)	1.61	0.01029(0.000**)	1.61
	Daily Average Load (X_2)	β_2	22.716(0.000**)	1.60	0.11313(0.000**)	1.61
	Hours Available (X_3)	β_3	-8.433(0.000**)	1.04	0.06576(0.000**)	1.04
	R^2		98.46%		91.57%	
	$\overline{R}^2$		98.45%		91.49%	
Backward	Constant (Intercept)	β_0	18.25(0.000**)		2.5371(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.306(0.010**)	1.61	0.01029(0.000**)	1.61
	Daily Average Load (X_2)	β_2	22.716(0.226)	1.60	0.11313(0.000**)	1.61
	Hours Available (X_3)	β_3	-8.433(0.000**)	1.04	0.06576(0.000**)	1.04
	R^2		98.46%		91.57%	
	$\overline{R}^2$		98.45%		91.49%	
Forward Selection	Constant (Intercept)	β_0	-184.14(0.000**)		2.5371(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.306(0.010**)	1.61	0.01029(0.000**)	1.61
	Daily Average Load (X_2)	β_2	22.716(0.000**)	1.60	0.11313(0.000**)	1.61
	Hours Available (X_3)	β_3	8.433(0.000**)	1.04	0.06576(0.000**)	1.04
	R^2		98.46%		91.57%	
	$\overline{R}^2$		98.45%		91.49%	

Footnote: **= p-values less than 5%.

Similarly, best subsets regression for actual and semi-log model were done in Table 1.13 and 1.14 respectively, which identify four and five suitable models with R^2 of 34.0% to 91.5% for actual and semi-log model. Mallow' Cp value is 4.0 for both models (actual and semi-log model). The three predictors' model with MLEC, ALEC and HAEC is identified to be best model.

Table 1.15 showed results of the other three methods (Stepwise Selection, Forward Selection and Backward Elimination) p-values for the estimated parameters of the independent variables. The three methods p-value of the models' parameters are significant at an α level of 0.05. The p-values MLEC, ALEC and HAEC are all equal to 0.000 for the two models, indicating that they are significantly related to Rumuosi average megawatt electricity consumption.

In addition, the VIF values for both models after applying stepwise selection, forward Selection and backward elimination methods are all close to 1, which indicates that the predictors are not correlated in both models.

In Table 1.14, the R^2 values indicates that the predictors explain 91.5% to 98.5% of the variance in Rumuosi average megawatt. The three methods results indicate that the model with three predictors' of MLEC, ALEC and HAEC fits the data for 2018 very well, just as the result of Best subsets regression. Hence, the three predictor's model has adequate predictive ability. That is

$$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \text{ for both models} \quad (2.3)$$

Model Identification for Rumuosi Electricity Consumption 2019

Table 1.16: Best Subsets Regression for Actual Model, 2019

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	85.5	85.5	85.3	703.3	25.363		X		
1	46.7	46.5	45.1	3564.5	48.709	X			
2	94.8	94.8	93.6	20.4	15.164		X	X	
2	94.7	94.6	93.3	34.3	15.440		X		X
3	95.0	95.0	93.8	10.7	14.948	X	X	X	
3	95.0	94.9	*	12.6	14.987		X	X	X
4	95.1	95.1	*	5.0	14.812	X	X	X	X

Table 1.17: Best Subsets Regression for Semi log Model, 2019

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	55.6	55.4	52.9	988.1	0.35488		X		
1	52.4	52.3	47.1	1083.2	0.36717			X	
2	87.7	87.6	84.6	15.5	0.18722		X	X	
2	85.0	84.9	81.6	96.3	0.20644		X		X
3	88.1	88.0	85.0	3.7	0.18400	X	X	X	
3	87.7	87.6	*	17.3	0.18744		X	X	X
4	88.1	88.0	*	5.0	0.18408	X	X	X	X

Table 1.18: Fitted Regression Models for the different Selection Methods for the Year 2019

Selection Methods	Variables	Coefficients	Actual Model Estimates (p-value)	VIF	Semi-Log Model Estimates (p-value)	VIF
Stepwise	Constant (Intercept)	β_0	-80.0(0.000**)		1.8867(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.832(0.002**)	2.17	0.01232(0.000**)	2.15
	Daily Average Load (X_2)	β_2	22.543(0.000**)	2.00	0.11024(0.000**)	2.00
	Hours Available (X_3)	β_3	4.537(0.000**)	12.14	0.09369(0.000**)	1.14
	Hours on Outage (X_4)	β_4	-2.389(0.006**)	11.72		
	R^2		95.11%		88.12%	
	$\overline{R}^2$		95.05%		88.02%	
Backward	Constant (Intercept)	β_0	-80.0(0.000**)		1.8867(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.832(0.002**)	2.17	0.01232(0.000**)	2.15
	Daily Average Load (X_2)	β_2	22.543(0.000**)	2.00	0.11024(0.000**)	2.00
	Hours Available (X_3)	β_3	4.537(0.000**)	12.14	0.09369(0.000**)	1.14
	Hours on Outage (X_4)	β_4	-2.389(0.006**)	11.72		
	R^2		95.11%		88.12%	
	$\overline{R}^2$		95.05%		88.02%	
Forward Selection	Constant (Intercept)	β_0	-80.0(0.000**)		1.8867(0.000**)	
	Daily Maximum Load (X_1)	β_1	-0.832(0.002**)	2.17	0.01232(0.000**)	2.15
	Daily Average Load (X_2)	β_2	22.543(0.000**)	2.00	0.11024(0.000**)	2.00
	Hours Available (X_3)	β_3	4.537(0.000**)	12.14	0.09369(0.000**)	1.14
	Hours on Outage (X_4)	β_4	-2.389(0.006**)	11.72		
	R^2		95.11%		88.12%	
	$\overline{R}^2$		95.05%		88.02%	

Footnote: **= p-values less than 5%.

Similarly for the year 2019, best subsets regression for actual and semi-log model were done in Table 1.16 and 1.17 respectively, which identify seven suitable models with R^2 of 46.5% to 95.1% for actual and semi-log model. Mallow' Cp value is 5.0 for both models (actual and semi-log model). The four predictors' model with MLEC, ALEC, HAEC and HOEC is identified to be best model.

Table 1.18 showed results of the other three methods (Stepwise Selection, Forward Selection and Backward Elimination) p-values for the estimated parameters of the independent variables. The three methods p-value of the model's parameters are significant at an α level of 0.05. The p-values MLEC, ALEC, HAEC and HOEC are all equal to 0.000 for the two models, indicating that they are significantly related to Rumuosi average megawatt electricity consumption.

However, the VIF values for actual model for predictor are greater than 10(Stepwise selection, forward Selection and backward elimination methods), which indicates that the predictors are correlated. It's suggested that the regression coefficients are poorly estimated due to severe multicollinearity, therefore the suitable model is the three predictors' model MLEC, ALEC and HAEC.

In Table 4.18, the R^2 values indicates that the predictors explain 88.1% to 95.1% of the variance in Rumuosi average megawatt. The three methods results indicate that the model with three predictors' of MLEC, ALEC and HAEC fits the data for 2019 very well, just as the result of Best subsets regression. Hence, the three predictor's model has adequate predictive ability. That is

$$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \qquad (2.5)$$

Model Identification for Rumuosi Electricity Consumption 2020

Table 1.19: Best Subsets Regression for Actual Model, 2020

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	92.4	92.4	92.3	324.3	31.264	X			
1	79.4	79.4	79.2	1493.3	51.403	X			
2	96.0	96.0	95.6	2.6	22.757		X		X
2	95.8	95.8	95.5	17.3	23.214		X	X	
3	96.0	96.0	95.6	3.0	22.738		X	X	X
3	96.0	95.9	95.6	4.5	22.784	X	X		X
4	96.0	96.0	95.6	5.0	22.769	X	X	X	X

Table 1.20: Best Subsets Regression for Semi log Model, 2020

Vars	R-Sq	R-Sq(adj)	(pred)	Mallows Cp	S	MLEC	ALEC	HAEC	HOEC
1	69.1	69.0	67.6	496.5	0.50266				X
1	67.3	67.2	65.8	545.2	0.51672			X	
2	86.9	86.9	85.8	2.4	0.32706		X		X
2	86.6	86.5	85.5	11.7	0.33121		X	X	
3	87.0	86.9	85.8	3.1	0.32689		X	X	X
3	86.9	86.8	85.8	4.4	0.32751	X	X	X	
4	87.0	86.8	85.7	5.0	0.32731	X	X	X	X

Table 1.21: Fitted Regression Models for the different Selection Methods for the Year 2020

Selection Methods	Variables	Coefficients	Actual Model Estimates (p-value)	VIF	Semi-Log Model Estimates (p-value)	VIF
Stepwise	Constant (Intercept)	β_0	12.26(0.010**)		4.4993(0.000*)	
	Daily Average Load (X_2)	β_2	22.462(0.000**)	1.27	0.09919(0.000**)	1.27
	Hours on Outage (X_4)	β_4	-5.193(0.000**)	1.27	-0.11798(0.000**)	1.27
	R^2		95.98%		86.94%	
	$\bar{R}^2$		95.96%		86.86%	
Backward	Constant (Intercept)	β_0	68.2(0.126)		3.754(0.000**)	
	Daily Average Load (X_2)	β_2	22.425(0.000**)	1.29	0.09969(0.000**)	1.29
	Hours Available (X_3)	β_3	-2.32(0.207)	51.93	0.0309(0.243)	51.93
	Hours on Outage (X_4)	β_4	-7.51(0.000**)	52.91	-0.0871(0.101)	52.91
	R^2		96.00		86.99	
	$\bar{R}^2$		95.96		86.88	
Forward Selection	Constant (Intercept)	β_0	12.26(0.010**)		4.4993(0.000**)	
	Daily Average Load (X_2)	β_2	22.462(0.000**)	1.27	0.09919(0.00)**)	1.27
	Hours on Outage (X_4)	β_4	-5.193(0.000**)	1.27	-0.11798(0.000**)	1.27
	R^2		95.98%		86.94%	
	$\bar{R}^2$		95.96%		86.86%	

Footnote: **= p-values less than 5%.

Best subsets regression for actual and semi-log model in Table 1.19 and 1.20 respectively identify seven suitable models with R^2 greater than 0.50, where each line of the output represents a different model. The model with the highest R^2 and adjusted R^2, a low Mallow' Cp value and the lowest Standard error of the regression "S" is identified as the best model. That is the two-predictor model in Table 1.19 and 1.20 with R^2 of 95.6% and 96.0 for actual and semi-log model. Mallow' Cp value of 2.6 and 2.4 for actual and semi-log model. The lowest standard error of the regression "s" of 22.757 and 0.3271 for actual and semi-log model respectively which except MLEC and HOEC. The two predictors' model with ALEC and HAEC is identified to be best model. This fact suggests that the models that include additional predictors may be overfitting the data. Overfit models appear to explain the relationship between the predictor and response variables for the data set used for model calculation but fail to provide valid predictions for new observations.

Table 1.21 showed results of the other three methods (Stepwise Selection, Forward Selection and Backward Elimination) p-values for the estimated parameters of the independent variables. The three methods p-value of the parameters ALEC and HAEC are both 0.000 for the two models, indicating that they are significantly related to Rumuosi average megawatt electricity consumption. The forward selection method p-value parameters for MLE Care 0.195 for actual model and 0.205 for semi log model, indicating that it is not related to Rumuosi average megawatt electricity consumption at an α level of 0.05. This suggests that a model with only ALEC and HAEC may be more appropriate. The VIFs for stepwise selection and backward elimination methods are all close to 1, which indicates that the predictors are not correlated in both models. However, the VIF values for forward selection elimination methods are greater than 10 in both models, suggest that the regression coefficient is poorly estimated due to severe multicollinearity.

Also in Table 1.20, the R^2 values indicates that the predictors explain 86.9% to 96.0% of the variance in Rumuosi average megawatt.

The stepwise selection and backward elimination methods results indicate that the model with only ALEC and HAEC fits the 2016 data set very well, just as the result of best subsets regression. Hence, the two predictor's model with ALEC and HAEC has adequate predictive ability. That is

$$\hat{Y} = \beta + \beta_2 X_2 + \beta_3 X_3 \text{ for both models} \tag{2.5}$$

Table 1.22: Summary of the Model Identification Results

Rumuosi Average Megawatt Electricity Consumption Data Sets	Model	Significant Predictors	Significant Parameters	variables
2016	$\hat{Y} = \beta + \beta_2 X_2 + \beta_3 X_3$	Two	Three	ALEC and HAEC
2017	$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	Three	Four	MLEC, ALEC and HAEC
2018	$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	Three	Four	MLEC, ALEC and HAEC
2019	$\hat{Y} = \beta + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$	Three	Four	MLEC, ALEC and HAEC
2020	$\hat{Y} = \beta + \beta_2 X_2 + \beta_3 X_3$	Two	Three	ALEC and HAEC

From, the above Table 1.22, the best model that is suitable for the Rumomasi Electricity Consumption, Rivers State average megawatt is the three predictor's model.

SUMMARY

This research discussed and estimated power outage in Rumuosi average megawatt electricity consumption. The aim of this research was able to provide an appropriate and unbiased regression model for Rumomasi electricity consumption, Rivers State, using model selection methods. The descriptive statistics of daily hourly consumption to monthly hourly consumption were estimated for each year. The four selection methods to estimating suitable regression models (Best subsets regression, Stepwise Selection, Forward Selection and Backward Elimination) results were computed. The best model was identified using R^2, Adjusted R^2 and VIF.

CONCLUSION

The purpose of this research was to study the use of model selection methods to identify an appropriate and unbiased regression model Rumomasi Electricity Consumption, Rivers State average megawatt. The study was able to estimate significant parameters for some of the independent variables using R^2, Adjusted R^2 and VIF.

REFERENCES

Adom, P., Bekoe, W., & Akoena, S. (2012). Modeling aggregate domestic electricity demand in Ghana: An autoregressive distributed lag bounds co integration approach. *Energy Policy, 42*, 530 – 537.

Babatunde, A., & Shauaibu, I. (2009). The demand for residential electricity in Nigeria: A bound testing approach. https://www.africametrics.org/documents/conference09

Bao, M., Quanline, S., & Jiamei, Z. (2004). Analysis of (gamma) spectra in airborne radioactivity measurements using multiple linear regression. China N.P. Web.

Belsley, D. (1991). *Conditioning diagnostics: Collinearity and weak data in regression*. Wiley. ISBN 978-0-471-52889-0.

Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity, 41*(5), 673 – 690. doi:10.1007/s11135-006-9018-6.

Chatterjee, S., Hadi, A. S., & Price, B. (2000). *Regression analysis by example* (Third ed.). John Wiley and Sons. ISBN 978-0-471-31946-7.

Conklin, P. (2001). Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry, 17*(4), 319 – 330. doi:10.1002/asmb.446.

Dilaver, Z. (2012). Structural times series modelling of energy demand. *PhD Thesis, University of Surrey, Guildford*.

Ebohon, O. J. (1996). Energy, economic growth and causality in developing countries: A case study of Tanzania and Nigeria. *Energy Policy, 24*(5), 447 – 453.

Gujarati, D. (2009). *Multicollinearity: what happens if the regressors are correlated? Basic Econometrics* (4th ed.), 363. McGraw−Hill.

Holtedahl, P., & Joutz, F. L. (2004). Residential electricity demand in Taiwan. *Energy Economics, 26*(2), 201 – 224.

Khanna, M., & Rao, N. D. (2009). Supply and demand of electricity in the developing world. *Annual Review of Resource Economics, 1*, 567 – 596.

Kock, N., & Lynn, G. S. (2012). Lateral collinearity and misleading results in variance-based SEM: An illustration and recommendations. *Journal of the Association for Information Systems, 13*(7), 546 – 580. doi:10.17705/1jais.00302.

Mohamed, H. K., Mostafa, S. M., & Karim, E. S. (2016). *Estimate the contribution of incubation parameters influence egg hatchability using multiple linear regression analysis*. Department of Poultry Production, Faculty of Agriculture, El-Shatby University of Alexandria, Egypt.

Narayan, P. K., & Smyth, R. (2005). The residential demand for electricity in Australia: an application of the bounds testing approach to co integration. *Energy Policy, 33*(4), 467 – 474.

Omri, A., & Kahouli, B. (2014). Causal relationships between energy consumption, foreign direct investment and economic growth: Fresh evidence from dynamic simultaneous-equations models. *Energy Policy, 67*, 913 – 922.

Vincenzo, B., Oronzio, M., & Sergio, N. (2009). Electricity consumption forecasting in Italy using regression models. *Energy, 34*(9), 1413 – 1421.

Treiber, M. (2013). Household energy transition in developing countries: Two alternative frameworks for analysis. *International Energy Agency (IEA), 1*(9), 21. https://file:///C:/Users/farshcm.LSBU/Dropbox/IEW%202013_6A4Treiber.pdf

VanDer Kroon, B., Brouwer, R., & Van Beukering, P. J .H. (2013). The energy ladder: Theoretical myth or empirical truth? Results from a meta-analysis. *Renewable and Sustainable Energy Reviews, 20,* 504 – 513.

VanDen Poel, D., & Larivière, B. (2004). Customer attrition analysis for financial services using proportional hazard models. *European Journal of Operational Research, 157,* 196 – 217. doi:10.1016/S0377-2217(03)00069-9.

Wichers, C. R. (1975). The detection of multicollinearity: A comment. *Review of Economics and Statistics, 57*(3), 366 – 368. doi:10.2307/1923926.

Wolde-Rufael, Y. (2006). Electricity consumption and economic growth; a time series for 17 African countries. *Energy Policy, 34*(10), 1106 – 1114.

Ziramba, E. (2008). The demand for residential electricity in South Africa. *Energy Policy, 36*(8), 3460 – 3466.