# Multiple Nonlinear Regression Modeling of Appliances Energy Use in a Low-Energy House

Yaseen Alwesabi
Binghamton University
**14** PUBLICATIONS **23** CITATIONS
SEE PROFILE

Udara Somarathna
University of Moratuwa
**8** PUBLICATIONS **0** CITATIONS
SEE PROFILE

Soongeol Kwon
Binghamton University
**7** PUBLICATIONS **70** CITATIONS
SEE PROFILE

*Proceedings of the 2019 IISE Annual Conference*
*H.E. Romeijn, A. Schaefer, R. Thomas, eds.*

# Multiple Nonlinear Regression Modeling of Appliances Energy Use in a Low-Energy House

**Y. M. S. Al-Wesabi, K.U.S. Somarathna, Yong Wang, and Soongeol Kwon**

**Department of Systems Science and Industrial Engineering, Binghamton University, NY 13902, USA**

## Abstract

The energy consumption of appliances plays an important role in the aggregated electricity demand of the residential sector. This paper aims to develop a high-performance forecasting model to predict the appliances energy consumption of a low-energy house located in Stambruges, Belgium. The study utilizes multivariate analysis and machine learning techniques to construct linear and nonlinear multiple regression models. The data involves measurements of temperature, humidity, and weather from a nearby airport station and lighting fixtures. Temperature and humidity are recorded every ten minutes using sensors from a wireless network in different rooms of the house. This data includes 27 attributes and 19,735 records. Data prepossessing including log-function, square-root, and box-cox has been conducted to control nonlinearity. Principle Component Analysis (PCA) is used with the regression model to reduce dimensionality and eliminate collinearity. The result shows that the Gradient Boosting Regression (GBR) improves the model to an adjusted R-squared of 41.97%, and the nonlinear third order polynomial regression model raises the percentage to 55.12%, which duplicates the accuracy to third fold compared to published work. The residuals chart shows some patterns that may lead to potentially further improvement based on this regression model.

## Keywords

Multiple nonlinear regression, Gradient Boosting Regressor (GBR), Principal component analysis (PCA), House appliances, Energy use

## 1. Introduction

The energy consumption of appliances is an important factor to consider when trying to understand the energy use in buildings, see [1, 2, 3]. Appliances energy consumption is an ongoing interesting research nowadays, which has been elaborated through several studies. A variety of methodologies have been reported to improve forecasting models of energy consumption. For instance, Ullah et al. applied the hidden Markov model, which has been compared with other forecasting models such as the Support Vector Machine (SVM), Artificial Neural Network (ANN), and Classification and Regression Trees (CART) [4]. Candanedo et al. applied four techniques, i.e., (a) multiple linear regression (MLR), (b) SVM, (c) random forest (RF), and (d) gradient boosting machines (GBM) [5].

The present paper is an extension of Ullah et al. which showed out that the MLR is an inappropriate model for this data since it yields only 18% of the adjusted R-squared [5]. However, they did not consider nonlinear regression models, which may lead to a high-performance prediction model of the energy use of appliances. In this study, we will conduct further and more sophisticated analysis, including nonlinear data transformation, combination MLR with PCA, using machine learning techniques like the Gradient Boosting Regressor (GBR) and eventually applying the nonlinear regression model (NRM) with the third order term.

## 2. Related Work

Several studies were conducted to study the impact of the weather and environment on the energy consumption of appliances in a low-energy house [5]. Candanedo et al. studied the impact of temperature, humidity, pressure, and light on the overall appliances energy consumption by implementing four approaches, MLR, SVM, RF, and GBM [5]. The results showed that the best result was obtained using the GBM model, while MLR failed to predict the data due to no-normality as well as existing patterns in the residual plot with adjusted-R squared of 18%. Nonlinear regression models were not conducted to be a potential gap for the future work.

Several real data imply a nonlinear relationship among variables, which leads to upgrading the analysis to nonlinear models. Besides, preprocessing techniques are so useful to reduce big data sets, PCA is an example to generate new uncorrelated variables. The PCA has been widely used in diverse aspects of applied science and industry (e.g., [6, 7, 8, 9]). Besides, the combination of the PCA and MLR is recommended if the predictors are collinear [10].

The GBR is a widely used optimization technique that can handle both linear and nonlinear regression models. The GBR is an ensemble of weak prediction models based on tree decision learner in a forward style. We have used the GBR as a package in Python. Some good references on the GBR are available in [11, 12].

## 3. Methodology

This paper concentrates on the collected data by [5], which has 19,735 observations. The data consists of 27 independent variables that represent temperature, humidity, pressure, time, and light as well as their influences on the energy consumption. Data collected every 10 minutes for 4.5 months by sensors with M-BUS energy counters. In our study, the MLR model and multi-NRM are conducted to improve the predicted model. The MLR model formula yields.

$$y_i = \sum_{j=0}^{n} \beta_j x_j + \varepsilon, \qquad i = 1, 2, ..., m \tag{1}$$

In this work, there is no linear correlation between the energy and independent variables, so a NRM of third order is chosen as follows.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \varepsilon \tag{2}$$

Since the main objective of this study is to develop a good forecasting model to predict the energy consumption, we consider several approaches such as the basic linear model, the linear model with transform functions, the linear model with PCA, the nonlinear regression models, and the GBR. Initially, we have defined the predictor variables and the response variable.

Boxplots were utilized to check the outliers. Once the outliers were removed, a visual analysis was performed to identify the relationships between the response variable and the predictor variables. We have also calculated the correlation coefficients to identify the linear relationship between the predictor variables and the response variable, as well as the presence of multi-collinearity.

Based on the result of the visual analysis, several linear and nonlinear models were fitted. For each model, residual analysis was performed, and model assumptions were checked. Based on the analysis of the results, several improvements were tried such as fitting different linear models with different ranges of original variables. Thereafter, the GBR machine learning algorithm was applied, and results were analyzed. Finally, the results of different approaches were compared, and conclusions were summarized

## 4. Results and Discussion

The outliers were identified based on a boxplot. According to Figure 1, the dataset contains many noisy observations. These outliers were eliminated from the dataset because there was no background information regarding these observations to conduct any investigations, and there was a large amount of good observations in the original dataset. After eliminating the outliers, 17,175 observations were retained for further analysis.

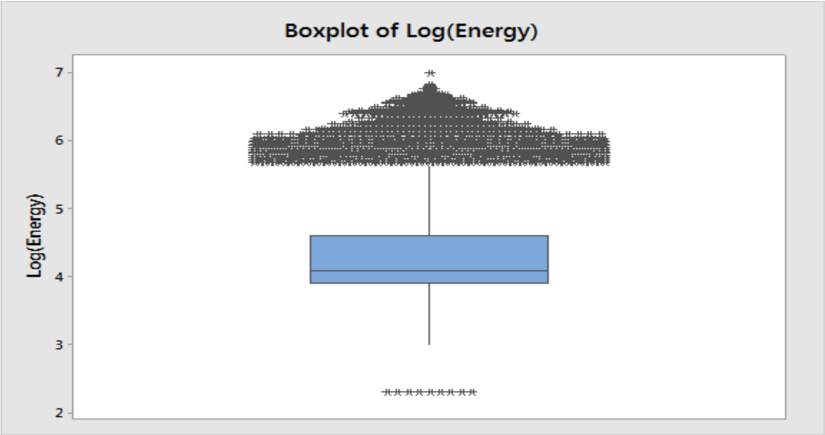

Figure 1: Box plot of the response variable Log (Energy)

A visual analysis was conducted to identify the relationships between variables. We have used both 2D and 3D scatter plots for visualization. Scatter plots (2D) were drawn between the predictor variable and each response variable. 3D scatter plots were drawn between selected pairs of predictor variables in the X, Y axes and the response variable in the Z axis so that we can identify existing interaction or non-linear effects. However, the visualization was difficult, due to a larger number of possible combinations and a large number of observations. To overcome this difficulty, scatter plots were also drawn from a randomly selected smaller subset of observations.

Based on the visual analysis and correlation matrix represented by the scatter matrix in Figure 2, we can make some key observations. Firstly, the linear correlation between the predictor variables and response variable is considerably low. The highest correlation identified between predictor variables and the response variable is 0.123. On the other hand, we have observed higher multi-collinearity among predictor variables. Also, we could not identify any interaction effects between any predictor variables visually. However, it seems that a polynomial relationship exists between the predictor variables and the response variable. It should be noted that, these results are just preliminary visual observations and the statistical significance of such relationships needs to be established.

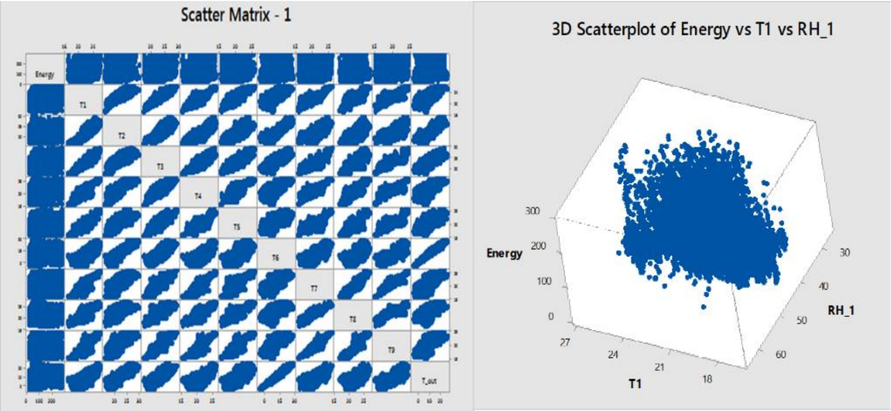

Figure 2: Scatter plots (2D and 3D)

Since we have identified multi-collinearity through the above preliminary investigations, we decided to apply MLR with a stepwise selection method so that we can select a subset of the predictor variables based on their partial correlations. Also, Principle Component Analysis (PCA) was used as a vital reduction technique to create new independent predictor variables, which were then used as new predictor variables.

Another key observation was complex non-linear relationship. Therefore, we decided to use non-linear transformations for the response variable such as log, square root, and box-cox. In addition, interaction terms and higher order terms can be used in MLR with stepwise selection. Finally, machine learning techniques were used with non-linear kernel functions to capture complex non-linear relationships.

### 4.1 MLR for the Original Dataset

As the initial model, the original response variable was fitted with all independent variables, and the key results are shown in Figure 3. According to the residual analysis, it was observed that the normality assumption was violated, and there exists a downward pattern in the residuals. However, the assumption of constant variance is satisfied as the observations are scattered within a constant band and the independence of residuals assumption is satisfied as the residuals are randomly distributed with respect to time (observation order). The adjusted R-squared was only 25.7%. The reduced R-squared could occur due to a higher variation of original data or inaccurate model. The pattern observed in the residual analysis suggests that the model adequacy is a result of the inability of current models to capture the patterns in the original observations. Therefore, it can be concluded that further improvements to the model are possible.

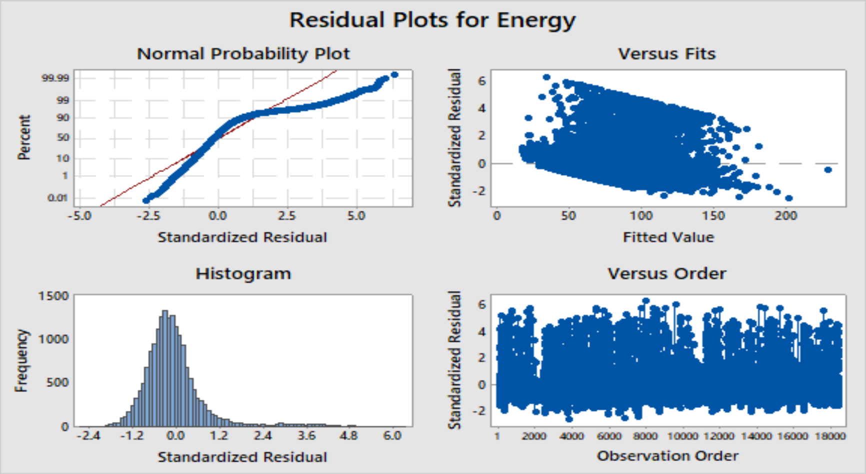

Figure 3: Residual Analysis of the initial model

### 4.2 MLR for Non-Linear Transformations

Since non-linear relationships were observed during visual analysis, several non-linear transformations such as log, square root, and box-cox were applied to the response variable. We assumed that the transformed response would have a linear relationship and MLR was applied with stepwise selection. The stepwise procedure has selected 20 predictor variables out of the 27 original predictors. The only categorical variable has also been removed by the stepwise method. The results of the residual analysis are shown in Figure 4.

The log-transformed model indicates better results as the normality assumption was improved. And the adjusted R-squared was 32.54%, which indicates that the model can explain 32.54% of the variation of the data. However, the residuals still indicate a downward pattern. Therefore, the model can be further improved.

### 4.3 Hybrid MLR and PCA

Since we observed a high level of multi-collinearity, as an alternative approach we have first applied PCA on log-transformed data to create independent new variables and then applied MLR with the new variables. Based on the Scree plot and the Eigenvalues of PCA, five Principal Components were selected which could explain 77.3% of the variation of the data. It was observed that the first principal component was loaded with temperature and the second Principal Component was loaded with relative humidity. Thereafter, the log-transformed response variable was fitted with these five Principal Components. However, the residual analysis still indicates the downward pattern and the adjusted R-squared was reduced to 23.43%. The fact is that the five principal components could only represent 77.3% of the variation may also have affected the adjusted R-squared.

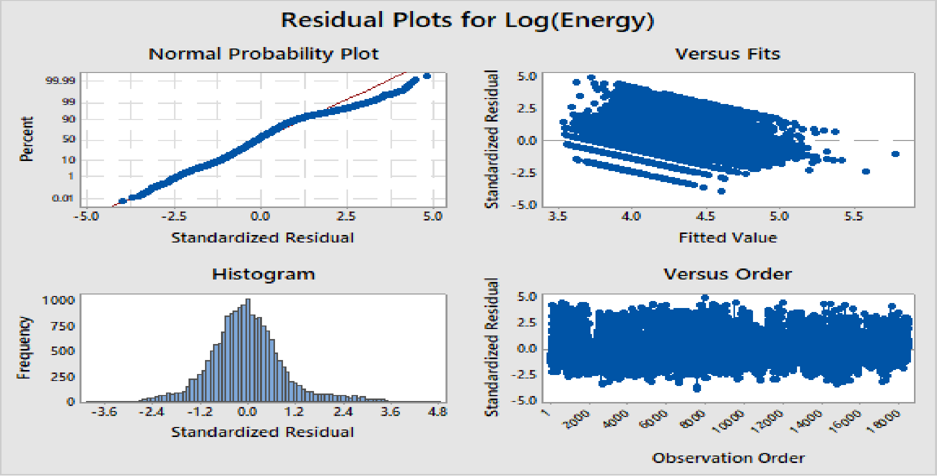

Figure 4: Residual Analysis of the log-transformed model

### 4.4 The GBR

Since there was not much improvement by applying linear models to capture the complex patterns in the observations, we have applied the GBR machine learning algorithm with a non-linear kernel function to the log-transformed response variable. The residual plot and the plot between observed data and fitted values are shown in Figure 5.

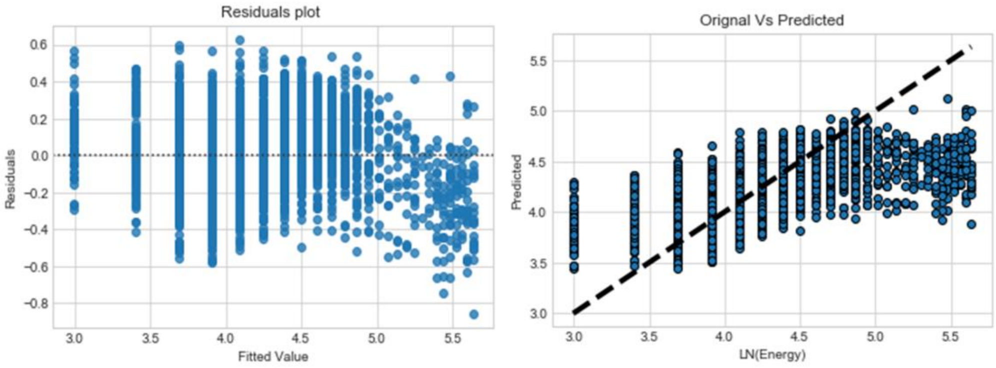

Figure 5: Residual plot and plot of observed data and fitted values

### 4.5 MLR with Higher Order Terms

To capture non-linear polynomial patterns, we have fitted the log-transformed response variable with polynomial terms of predictor variables up to the 3rd degree with stepwise selection. The residual analysis still consists of the downward pattern. However, the adjusted R-squared was increased to 55.12%. The resulting regression equation is very long which cannot be shown here (60 terms). We can see that the regression model includes certain terms to ensure model hierarchy even though those are not significant. By considering the p-value of the hypothesis test for the overall model we can see that the overall regression model is significant at the 0.05 level. Thereafter, we can consider the p-values of the individual regression coefficients. Finally, we can interpret the relative strength and the direction of the relationship of each predictor variable with the response variable by considering the standardized regression coefficients.

The summary of eight different model categories is illustrated in Table 1.

Table 1: Results summary

| **Model category** | **Dependent variable** | **R-Squared (Adj)** |
|---|---|---|
| MLR | Energy | 25.70% |
| MLR (with Non-linear transformation) | Sqrt (Energy) | 29.89% |

| | | |
|---|---|---|
| MLR (with Non-linear transformation) | Lin (Energy) | 32.46% |
| MLR (with Non-linear transformation) | Box-Cox (optimal λ) | 32.73% |
| PCA + MLR | Lin (Energy) | 23.41% |
| MLR (with 3rd order polynomial terms) | Lin (Energy) | **55.12**% |
| Machine Learning (GBR) | Energy | 33.12% |
| Machine Learning (GBR) | Lin (Energy) | 42.79% |

## 5. Conclusions

In many real-world applications, the relationships among variables are complex and nonlinear. In this study, the visual analysis, scatter plots could shed valuable preliminary insights for the study. It can also be concluded that the inclusion of higher order terms may be able to capture non-linear polynomial relationships between the predictor and response variables. The nonlinear third order polynomial regression model raises the percentage to 55.12%, which is the highest percentage compared to other schemes for this dataset. As a future work, time series models can be used for the projection as the data collected over 4.5 months.

## References

1. Arghira, N., Hawarah, L., Ploix, S. and Jacomino, M., 2012, “Prediction of appliances energy use in smart homes,” Energy, 48(1), 128-134.
2. Cetin, K., Tabares-Velasco, P. and Novoselac, A., 2014, “Appliance daily energy use in new residential buildings: Use profiles and variation in time-of-use,” Energy and Buildings, 84, 716-726.
3. Ruellan, M., Park, H. and Bennacer, R., 2016, “Residential building energy demand and thermal comfort: Thermal dynamics of electrical appliances and their impact,” Energy and Buildings, 130, 46-54.
4. Ullah, I., Ahmad, R. and Kim, D., 2018, “A prediction mechanism of energy consumption in residential buildings using hidden markov model,” Energies, 11(2), 358.
5. Candanedo, L.M., Feldheim, V. and Deramaix, D., 2017, “Data driven prediction models of energy use of appliances in a low-energy house,” Energy and Buildings, 140, 81-97.
6. Abdi, H. and Williams, L.J., 2010, “Principal component analysis," Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.
7. Bro, R. and Smilde, A.K., 2014, “Principal component analysis,” Analytical Methods, 6(9), 2812-2831.
8. Dunteman, G.H, 1989, “Principal components analysis, No. (69). Sage.
9. Wold, S., Esbensen, K. and Geladi, P., 1987, “Principal component analysis” Chemometrics and intelligent laboratory systems, 2(1-3), 37-52.
10. Abdul-Wahab, S.A., Bakheit, C.S. and Al-Alawi, S.M., 2005, “Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations,” Environmental Modelling & Software, 20(10), 1263-1271.
11. Friedman, J.H., 2001, “Greedy function approximation: a gradient boosting machine,” Annals of statistics, pages 1189-1232
12. Friedman, J.H., 2002, “Stochastic gradient boosting,” Computational Statistics & Data Analysis, 38, 367-378.