Preliminary multiple linear regression model to predict hourly electricity consumption of school buildings

Keovathana Run, Franck Cévaër, Jean-François Dubé

Preliminary multiple linear regression model to predict hourly electricity consumption of school buildings

Keovathana Run[1], Franck Cévaër[1] and Jean-François Dubé[1]

[1] LMGC, University of Montpellier, UMR CNRS 5508, cc048, Place E.Bataillon, 34095 Montpellier cedex 5, France
keovathana.run@umontpellier.fr

Abstract. Energy predicting gains attention for its ability to manage and control energy consumption in a building. The multiple linear regression model is known for its simplicity and effective when dealing with electricity consumption. In this work, the authors have utilized the Multiple Linear Regression (MLR) model to predict the hourly electricity energy consumption, in winter, for school buildings. For the case study, school buildings in the South of France are used. In this model, nine predictor variables are considered, namely, (1) level of indoor CO_2, (2) indoor temperature, (3) indoor humidity, (4) outdoor temperature, (5) outdoor humidity, (6) global solar radiation, (7) day index (weekday/weekend), (8) time index (occupied/non-occupied) and (9) building net floor area. The first order and two-way interaction models are constructed using all predictors. The coefficient of determination (R^2) is a model evaluation metric that assesses the relationship between the vales of the desired outcomes and those that the model predicts. The results show that the two-way interaction model has better R^2 for both training set (R^2 = 74%) and testing set (R^2 = 77%). However, this model gives underestimated results for higher values of electricity consumption starting from 30kWh/hour. It is also not reliable for one of the buildings as the R^2 is only 55% and the inaccuracy rate is 69%. Overall, this model is a starting point for future work to improve its predicting ability by adding other influential explanatory variables.

Keywords: Electricity consumption, School buildings, energy regression model.

1 Introduction

Approximately 40% of EU energy consumption and 46% of energy-related greenhouse gas emissions are attributable to buildings. Nearly 75% of the building stock in the EU is now energy inefficient, and about 35% of structures are older than 50 years [1]. According to data released by the Agency for the Environment and Energy Management (ADEME), the building sector in France is responsible for 25% of CO_2 emissions and 44% of energy usage.

Researchers discovered that energy forecasting techniques that use historically recorded time series energy data have enormous value in energy optimization for existing

buildings [2]. Data-driven models have gained popularity among academics due to their simplicity, ability to handle large data sets, and high prediction accuracy, though this is not true for all types of data-driven models [3].

The objective of this paper is to develop a preliminary MLR model that aims to predict the electric power consumption per hour on school buildings. Two initial models were compared to see their prediction ability namely, the first order model and the two-way interaction model using the forward regression that is described in Section 2.2 and the model evaluation metrics in Section 2.3.

2 Method

2.1 MLR

The method in this study is based on a multivariate regression analysis, which accounts for the variation of the independent variables in the dependent variables synchronically [4]. The multiple linear regression (MLR) model is:

$$Y_i = \beta_0 + x_{i,1}\beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + \mathrm{e}_i$$

Where :
Y is the response (dependent variable);
x is the predictors (independent variables);
β is the unknown regression coefficients
e is unknown errors
$i = 1, \ldots, n$
n is the sample size
e_i is the error to account for the discrepancy between predicted and the observed data

After the models are developed and checked, the predicting is then made. All too often the MLR model seems to fit the *'training data'* well, but when new *'testing data'* is collected, a very different MLR model is needed to fit the new data well. Therefore, it is important to wait until after the MLR model has been showed to make good predictions before claiming that the model gives good predictions [5].

2.2 Predictors selection

The potential predictors are pre-selected. This study employs forward regression, which starts with a model with no predictors, to choose its predictors (the intercept only model). After that, variables are added to the model one at a time until none more can improve it by a particular standard. The variable that significantly improves the model is introduced at each step. A variable stays in the model once it is added.

Given a response vector $Y \in \mathbb{R}^n$, predictor matrix $X \in \mathbb{R}^{np}$, and yields a subset of each size $k = 0, \ldots, \min\{n, p\}$. Formally, the procedure starts with an empty active set $A = \{0\}$, and for $k = 0, \ldots, \min\{n, p\}$, selects the variable indexed by (1) that leads to the lowest squared error when added to A_{k-1}, or equivalently, such that X_{j_k}, achieves the maximum absolute correlation with Y, after we project out the

contributions from $X_{A_{k-1}}$. A note on notation: here we write $X_S \in \mathbb{R}^{n|S|}$ for the sub-matrix of X whose columns are indexed by a set S (and when $S = \{j\}$, we simply use X_j).We also write P_s for the projection matrix onto the column span of X_s , and $PS^{\perp} = IP_s$ for the projection matrix onto the orthocomplement. At the end of step k of the procedure, the active set is updated, $Y_k = A_k - 1 \cup \{j_k\}$, and the forward stepwise estimator of the regression coefficients is defined by the least squares fit onto X_{A_k} [5].

$$j_k = \underset{j \notin A_{k-1}}{argmin} \| Y - P_{A_{k-1} \cup \{j_k\}} Y \|_2^2 = \underset{j \notin A_{k-1}}{argmin} \frac{X_j^T P_{A_{k-1}}^{\perp} Y}{\| P_{A_{k-1}}^{\perp} X_j \|_2} \quad (1)$$

2.3 Model selection

The performance of the models is assessed using 10-fold cross-validation after the best models from the combination of chosen parameters have been developed. Regression analysis uses the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and the Mean Absolute Percentage Error (MAPE) metrics to measure each model's prediction error. The better the model, the lower the MSE, MAE, and RMSE. Coefficient of determination (R^2) denotes the relationship between the values of the desired outcomes and those that the model predicts. The model is better the greater the R^2.

- MAE represents the difference between the original and predicted values extracted by averaged the absolute difference over the data set.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i|$$

- MSE is the average of the squared difference between the original and predicted values in the data set. It measures the variance of the residuals.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

- RMSE is the error rate by the square root of MSE. It measures the standard deviation of residuals.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2}$$

- MAPE is the percentage error calculated in terms of absolute errors, without regards to sign.

$$MAPE = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

- R^2 represents the coefficient of how well the values fit compared to the original values. The value from 0 to 1 interpreted as percentages.

$$R^2 = \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2}$$

where, $\hat{y}$ and $\bar{y}$ are respectively predicted and mean value of y measured value at the i^{th} moment, and N represents the number of predictions.

3 Case study: IUT de Nîmes school buildings

As part of the OEHM project, IUT de Nîmes campus was selected as a case study for the Ph.D. thesis of the first author. This campus is in the south of France, at 43°49'N longitude and 4°19'E latitude. The climate of this region is classified as CSA, with relatively mild winters and hot summers, often referred to as "Mediterranean," according to Koeppen and Geiger [6]. Since 2019, 338 sensors in total of six types (Elsys, Class'Air, CM868LR, IR868LR, BT1-L, and Adeunis) have been placed on the site.

The data from three buildings built in 1969 were collected. They are Civil engineering building (GC), Electrical engineering building (GEII) and Material engineering building (GMP) with total net surface area of $4762m^2$, $3627m^2$ and $6357m^2$, respectively. Naturally ventilated, they also have the same floor plans, a two-story teaching building and a one-story workshop building with a high ceiling. The heating system is hot water radiators, supplied with heat by the urban heating network. The electrical energy is dedicated to the rest of the appliances in the buildings including lighting, the electrical distributor, electrical radiators, and air conditioners (reversible), etc. Therefore, the electricity consumption still depends on weather conditions and the indoor climate.

Table 1. Statistics of collected dataset

Parameters	Building	Min	Median	Mean	Max	SD*	SE**
Consumption (kWh/hour)	GC	5.00	10.00	15.99	80.00	12.06	0.20
	GEII	0.00	5.00	10.06	81.00	10.82	0.18
	GMP	9.00	18.00	25.43	93.00	16.02	0.26
CO_2 (ppm)	GC	378.79	440.27	472.04	1084.26	90.63	1.49
	GEII	368.51	437.32	470.37	1042.86	92.74	1.53
	GMP	373.21	457.32	551.41	1727.71	199.10	3.28
T_{in} (°C)	GC	14.55	19.60	19.47	22.84	1.32	0.02
	GEII	13.47	20.67	19.81	23.33	2.49	0.04
	GMP	17.38	21.74	21.48	24.12	1.09	0.02
HR_{in} (%)	GC	22.53	37.00	37.70	55.67	7.43	0.12
	GEII	18.00	36.00	37.61	69.00	10.96	0.18
	GMP	23.92	36.05	37.25	54.90	6.84	0.11
T_{ext} (°C)		-3.50	8.70	9.04	27.50	4.85	0.08
HR_{ext} (%)		18.00	65.00	64.90	97.00	18.67	0.31
SR (MJ/m^2)		0.00	0.00	0.40	3.28	0.67	0.01

**SD: Standard deviation, **SE: Standard error*

The sensors record and transmit every fifteen-minute for indoor carbon dioxide (CO_2), indoor temperature (T_{in}) and indoor relative humidity (HR_{in}) and every one-hour for real-time electricity consumption. The outdoor temperature (T_{ext}), outdoor relative humidity (HR_{ext}) and global solar radiation (SR) are taken every one-hour from the nearest representative station, the climate data of Nîmes Courbessac from Météo France. The

analysis is done during five months, from November 2021 to April 2022, when all necessary data were available. Each parameter's time basis was reset to every hour using time interpolation.

Table 1. shows the range and variation of each parameter of each building. It is evident that the outdoor weather is between -3.5°C and 27.5°C and highest temperature indoor is between 17.38°C and 24.12°C. Peak value for global solar radiation is 3.28MJ/m^2. From a quick analysis, GMP has the most corresponding variations for all the parameters.

Model development

3.1 Pre-selection variables

The pre-selected explanatory and dependent variables for the models are as follows:

i. Dependent variable: Y = Hourly electricity usage (kWh/hour)
ii. Predictor variable 1: x1 = CO_2 (ppm)
iii. Predictor variable 2: x2 = T_{in} (°C)
iv. Predictor variable 3: x3 = HR_{in} (%)
v. Predictor variable 4: x4 = T_{ext} (°C)
vi. Predictor variable 5: x5 = HR_{ext} (%)
vii. Predictor variable 6: x6 = SR (MJ/m^2)

To get more reliable results, another three proxy variables are added: day index, hour index, and building net floor area.

i. Predictor variable 7: x7 = Day Index (Weekday/ Weekend) \\
ii. Predictor variable 8: x8 = Hour Index (daytime 7h00 - 19h00/ nighttime 19h00 - 6h00)
iii. Predictor variable 9: x9 = Building net floor area (m^2)

3.2 Model selection

The initial model using first order of all nine predictors can be expressed as:

$$Y_1 = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + x_5\beta_5 + x_6\beta_6 + x_7\beta_7 + x_8\beta_8 + x_9\beta_9$$

The analysis is done using Rstudio version 4.0.3. (2020-10-10).The regsubsets() function [leaps package] computes the forward regression; the tuning parameter nvmax specifies the maximum number of predictors to incorporate in the model. It returns a variety of models in sizes ranging from small to large. The performance of the models is then carefully compared to select the best one. Amongst the nine models returned, the best performance is when all nine variables are considered, with RMSE of 0.105 and R2 of 0.60. However, the value of R^2 of the trained model is rather weak, meaning that it can only explain 60% of the variance. Therefore, a two-way interaction of this trained model is carried out for an equation can be written as:

$$Y_2 = \beta_0 + x_1\beta_1 + \cdots + x_9\beta_9 + x_1x_2\beta_{10} + \cdots + x_8x_9\beta_{45}$$

Which $i = 1, \ldots, 8$ and β_0 is the intercept value. Using the same forward regression to find the best combination of predictors. 45 models are returned and the best model of 40 variables has RMSE of 0.08 and R^2 of 0.73. The performance of the trained model increases with the addition of the 2-way interaction between each predictor. To determine whether the interaction is required, an ANOVA test is performed to compare the two trained models. It is extremely statistically significant that the P-Value from the anova is less than 2.2e-16. As a result, the second model is chosen to be applied to the testing set to evaluate its performance.

4 Results and discussion

In this study, the multiple linear regression model is applied on the dataset during winter from November 2021 to April 2022. A training set is made up of 70% of the randomly chosen data from the gathered dataset, while a testing set is made up of 30% of the remaining data. Assessing the performance and correctness of the produced model against already established targets in the collection of predictor variables is the major goal of model testing. The equation of trained model can be written as:

$$\begin{aligned} Y = {} & 0.36 + 0.47x_1 - 0.33x_2 - 0.09x_3 - 0.08x_4 - 0.23x_5 + 0.27x_6 - 0.11x_7 \\ & + 0.17x_8 + 0.17x_9 - 1.4x_1x_2 - 1.46x_1x_3 + 0.6x_1x_4 + 1.25x_1x_5 \\ & - 0.54x_1x_6 - 0.22x_1x_7 + 0.78x_1x_8 + 0.31x_1x_9 + 0.32x_2x_3 \\ & + 0.02x_2x_4 + 0.19x_2x_5 + 0.21x_2x_6 - 0.02x_2x_7 - 0.06x_2x_8 \\ & + 0.4x_2x_9 - 0.15x_3x_4 - 0.05x_3x_5 + 0.4x_3x_6 - 0.14x_3x_7 \\ & - 0.03x_3x_8 + 0.18x_3x_9 + 0.05x_4x_5 - 0.6x_4x_6 + 0.25x_4x_7 \\ & - 0.03x_4x_8 - 0.08x_4x_9 + 0x_5x_6 + 0.17x_5x_7 - 0.08x_5x_8 \\ & - 0.08x_5x_9 - 0.17x_6x_7 - 0.11x_6x_8 + 0.17x_6x_9 - 0.1x_7x_8 \\ & - 0.02x_7x_9 + 0.03x_8x_9 \end{aligned}$$

This regression model is selected for its highest R^2 value of 0.74 while training. The forecasting between the two sets can be seen in Fig. 1. After applying this model on testing set, the R^2 value reached 0.77, higher than the training set.

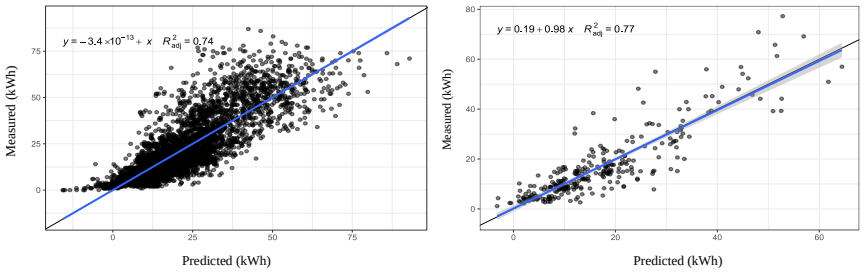

Fig. 1. The correlation between measured electricity consumption and predicted electricity consumption using multiple linear regression of two-way interaction with 40 variables: (right) for training set, (left) for testing set.

The regression beta coefficients are represented by the blue line. When the slope for the training set is equal to 1, the regression has the best fit. The calculated regression line does not, however, exactly fit all the data points. The distance between the points and the regression line increases with increasing electricity use. This demonstrates how poorly our model can anticipate the larger values. The comparison of the scaled output by weekday and weekend is shown in Fig. 2. The figure compares the values that were measured (in red) with those that the developed model predicted (in sky blue). On weekday during the occupied hours, the predictions are coherent with the measured data up until it exceeds the 30kWh/hour when the predictions begin to underestimate the value.

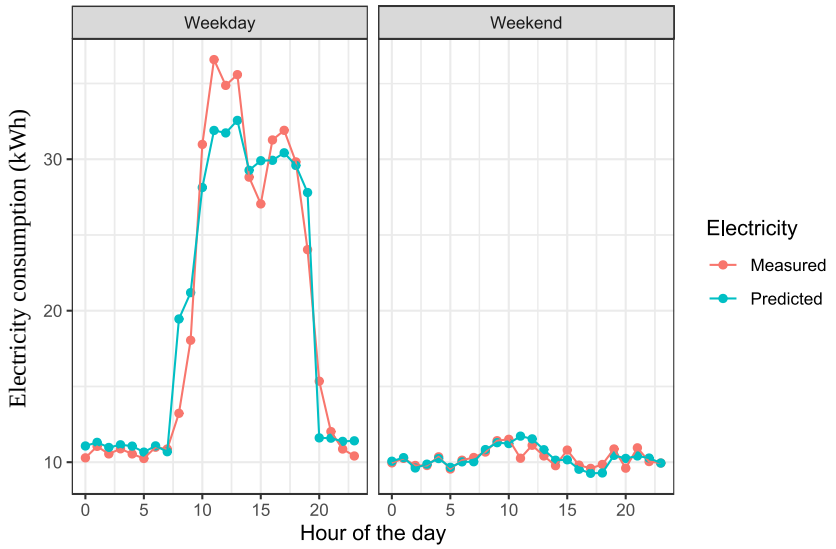

Fig. 2. The comparison between measured and predicted electricity consumption hourly over weekday and weekend.

Table 2 presents four model evaluation matrices, MAE, MSE, MAPE, and RMSE and together with the model performance R^2. The performance of the model is the best on GMP building and the worst on GEII building which has R^2 equals 55% and MAPE of 69%. That means, the interaction between predictors on this building cannot estimate well the energy consumption.

Table 2: The model errors and performance of each building

Building	MAE (kWh)	MSE (kWh)	MAPE (%)	RMSE (kWh)	R^2(%)
GC	4.33	52.48	0.26	7.24	0.64
GEII	4.76	52.33	0.69	7.23	0.55
GMP	5.44	60.86	0.22	7.80	0.76

5 Conclusion

The purpose of this research is to provide a multiple linear regression model that can forecast the hourly electricity usage in educational facilities. To create a one-way and two-way interaction regression model, nine potential explanatory variables were used: CO_2, Tin, HR_{in}, T_{ext}, HR_{ext}, SR, Day index, Hour index, and building area. Better results (R^2 = 73%) are obtained with the combination of two-way interaction models, but the predictor variables also become complex. The model maintains a strong R^2 performance of 74% on the training set and 77% on the testing set. This basic model can be utilized for more research in accordance with section 2.1.

The limits happen on bigger values of electricity consumption starting approximately from 30kWh/hour. Moreover, the prediction on GEII building is not acceptable. The pre-selected predictors might be not the most influential variables. For instance, the outdoor temperature and global solar radiation should be delayed from a few hours to 7 hours for a maximum of effect on the indoor temperatures. As this preliminary model has proved itself to be reliable in most cases, a future study will base on this one but adding more potential predictors such as the delayed in indoor temperature, delayed in global solar radiation and occupancy rate. A validation step should be also included for a further study.

References

[1] B. Anderson, “Energy Performance of Buildings Directive,” p. 169.

[2] C. Deb, F. Zhang, J. Yang, S. E. Lee, and K. W. Shah, “A review on time series forecasting techniques for building energy consumption,” *Renew. Sustain. Energy Rev.*, vol. 74, pp. 902–924, Jul. 2017, doi: 10.1016/j.rser.2017.02.085.

[3] Z. Afroz, G. Shafiullah, T. Urmee, and G. Higgins, “Modeling techniques used in building HVAC control systems: A review,” *Renew. Sustain. Energy Rev.*, vol. 83, pp. 64–84, Mar. 2018, doi: 10.1016/j.rser.2017.10.044.

[4] G. K. Uyanık and N. Güler, “A Study on Multiple Linear Regression Analysis,” *Procedia - Soc. Behav. Sci.*, vol. 106, pp. 234–240, Dec. 2013, doi: 10.1016/j.sbspro.2013.12.027.

[5] D. J. Olive, “Multiple Linear Regression,” in *Linear Regression*, D. J. Olive, Ed. Cham: Springer International Publishing, 2017, pp. 17–83. doi: 10.1007/978-3-319-55252-1_2.

[6] M. C. Peel, B. L. Finlayson, and T. A. McMahon, “Updated world map of the Köppen-Geiger climate classification,” *Hydrol. Earth Syst. Sci.*, vol. 11, no. 5, pp. 1633–1644, Oct. 2007, doi: 10.5194/hess-11-1633-2007.