Linear Regression Analysis:
Predicting Energy efficiency of Building

Jiale Wu
Illinois State University

Jiale Wu
Department of Mathematics,
Illinois State University
Normal, Illinois 61761. E-mail: jwu2@ilstu.edu.

Table of Contents

Abstract

Statistical learning plays a key role in many areas of science, finance, and industry. Linear regression is a very useful approach for modeling the relationship between a scalar dependent variable y and one or more independent variables x. In particular, multiple linear regression is very useful and helpful to solve some real issues. Also, classification is very helpful methods of data analysis. In this research, we will use supervised learning with linear methods for regression. We will use some machine-learning techniques, such as multiple linear regressions, stepwise regression, principal component analysis, partial least squares and tree-model. Finally, according to our analysis and results, we will conclude some effective and helpful conclusions.

1. Introduction

Nowadays society, energy efficiency is one of the easiest and most cost effective ways to combat climate change, clean the air we breathe, improve the competitiveness of our businesses and reduce energy costs for consumers. Or rather, energy efficiency develops new, energy efficient technologies while boosting the efficiency of current technologies on the market. Thus, we can define energy efficiency as using less energy to provide the more service. According to some reports, they said that each year, much of the energy the U.S. consumes was wasted through transmission, heat loss and inefficient technology, as well as, leading to cost families and business money and to increase carbon pollution. Present, more and more company need to replace or upgrade machine equipment to make more profit and improve their efficient technology. Sometimes, we can think that the more efficiency equipment, the more accurately results are tested. It is a significant factor for many companies to achieve exactly outcomes. It believes that each company not only would like to have the new equipment with the more service, but also it hopes to use less cost and energy to guarantee the maximum profit. General Speaking, energy-efficient products save money on the energy bill between families and companies, and reduce the amount of greenhouse gases going into the atmosphere. For families, for example, when they remodel a single pane window at home with an energy-efficient, the new one can keep heart in the winter and prevent cool escaping in the summer. Thus, we can easy say that efficient windows are helpful for the air conditioner and make it not run often. Not only can save heart and electric fee, but also people still stay comfortable at home. For companies, energy efficient solutions for buildings and manufacturing supply lines means large-scale energy and cost savings for them. Therefore, we believe that in the future energy efficiency will become to play an important role among the world.

In most cases, data mining can be defined as tools, methodologies, and theories for revealing patterns in data-a critical step in knowledge discovery. Also, it contains supervised learning and

unsupervised learning. Most real problems refer to supervised learning problems. Thus, for supervised learning problems, it has regression and classification analysis. Many statisticians use Machine learning techniques, including regression and classification, to analyze most data set.

In 2012, Athanasios Tsanas and Angeliki Xifara developed a statistical machine-learning framework to accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. They compared a classical linear regression approach against a powerful state of the art nonlinear non-parametric method, random forests, to estimate output variables. Finally, the outcomes of their study support the feasibility of using machine learning tools to estimate building parameters.

In 2013, Mohamed and his group members looked into assessing heating load and cooling load requirements of building as function of building parameters. They use classification method to analyze it. It contained Naïve Bayes, J48, and Bagging approaches. Finally, they conclude that J48 and bagging could give them more confident to say both are good enough make prediction.

An increasing of people would like to use energy-efficient products. As previously mentioned, authors have focused on the parameter estimation of energy performance. In this research, the regression analysis would be applied. We will use machine-learning techniques, including multiple linear regressions, stepwise regression, principal component analysis, partial least squares and tree model.

The remainder of this paper is organized as follows: In section 2 we include a data Collection and Description, followed by section 3 in which we present some methods for analysis with marching machine-learning techniques. Section 4 we draw our conclusion and future work.

2. Data Collection and Description

In this study, data is collected form UCI machine learning repository. Data currently is reported on the below website (https://archive.ics.uci.edu/ml/datasets/Energy+efficiency). In this project, we download Energy efficiency data set online with excel sheet. This resulted in a total of 768 observations that are used in subsequent analysis. In this study, we split the data set into a training set and a test set. We split one half is training set and the other half is testing set. There is no missing value. There are eight input variables (independent variable):

x1= relative compactness,

x2=surface area, x3=wall

area, x4=roof area,

x5=overall height,

x6=orientation, x7=glazing

area, x8=glazing area

distribution.

Also, there are two output variables (dependent variables): y1= heating load y2

= cooling load.

4. Methods for Analysis:

This section wills summary the statistical concepts and machine learning methods that are used to analyze the data. For statistical applications, we will may use statistical software: R to analyze this data and graph plots.

4.1 input variables analysis and data pre-processing:

Before analyzing the data set, in most cases, we will check one main assumption properties of the variables. Normality is a key concept of statistics. Thus, normality test plays an important role for data analysis. It is necessary for scientists and other researches to check the normality of data sets.

In this research, there are 8 depend variables (Xs) and 2 independent variable (Y). So, we need to deal with and manipulate the input data. All variables are summarized and univariate analysis with plots are shown below. Now, we will chick it is normal distribution or not. To analyze the normality for these x variables, normal quantile plot (Q-Q plot), and summary of basic statistical table would be used. From Table 1: basic statically summary, we can find all of input variables' mean and median are slight different. For x4, x5, and x6, they are same. For other variables, their value between mean and median are approximately same. Maybe, we say they are symmetric or normal distribution. Mover ever, to test normality test, normal quantile plot (Q-Q plot) would be plotted. From Figure 1: Q-Q plot, x1, x2, x3 and x 8, they look like marginal normal distributions because most of their points follow the diagonal line. However, for other variables, likely x4, x5, x6 and x7, they look like not very normal since there are some tails and outliers. Therefore, for normality test, we can summary that not all of input variables are marginal normal distribution.

	Minimum	Q1	Median	Mean	Q3	Maximum
X1	0.6200	0.6825	.07500	.07642	0.8300	0.9800
X2	514.5	606.4	673.8	671.7	741.1	808.5
X3	245.0	294.0	318.5	318.5	343.0	416.5
X4	110.2	140.9	183.8	176.6	220.5	220.5
X5	3.50	3.50	5.25	5.25	7.0	7.0
X6	2.00	2.75	3.50	3.50	4.25	5.00
X7	0.0000	0.1000	0.2500	0.2344	0.4000	0.4000
X8	0.000	1.750	3.000	2.812	4.000	5.000

Table 1: basic statically summary

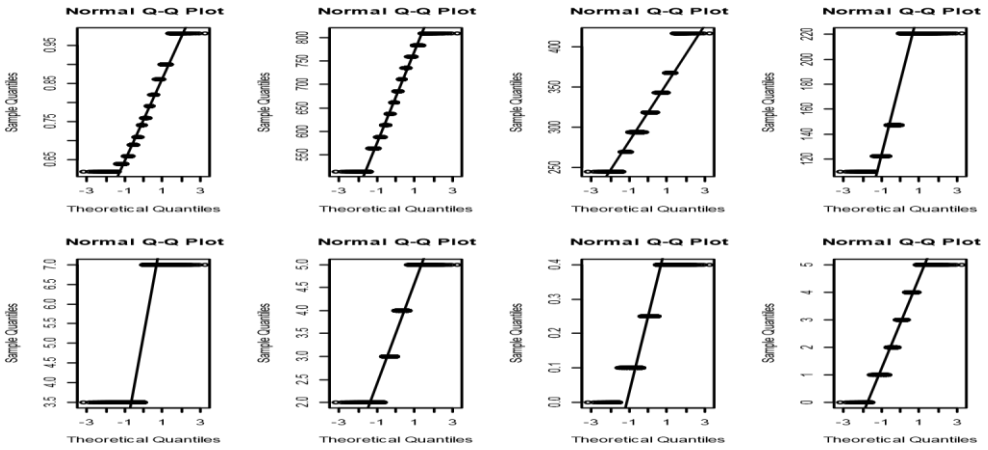


Figure 1: normal quantile plot (Q-Q plot)

4.2.1 Multiple Linear Regression

In our case, there are 8 predictor variables (from x1 to x8) and 2 response variables (y1 and y2), so we use multiple linear regression to analyze it.

There are 8 explanatory variables, and the relationship between the 2 dependent variables (y1 and y2) and the explanatory variables are represented by the following equation: $yi=\beta0+\beta1x1+\ldots \beta8x8+\varepsilon i$ Where β0 is intercept, other βs are parameters, and it assumes errors are identically independent distributions with zero mean and constant variance.

On the one hand, linear regression was performance with the response variable (y1) as heating load and the rest of the variable as predictors. From our output 1, it showed us that x4=roof area does not have any relationship with the significant. Other variables, except x6, are very significant since they have small p-values. Also, we can look at the determination of coefficient and its Rsquare is equal to 0.9162. It means that there are 91.62% variation on the response variable y1 explained by the model. We analyze the model based on hypotheses that: All regression coefficients are zero. So, Null Hypothesis: All regression coefficients are zero. Alternative: at least one coefficients is not zero. From F-test, we know that p-value for the model is equal to 2.2e-16 so that we reject Null Hypothesis. There is at least one coefficient is statistically significant different from others.

On the other hand, linear regression was performance with the response variable (y2) as cooling load and the rest of the variable as predictors. From our output 2, it also showed us that x4=roof area does not have any relationship with the significant. Other variables, except x6, are very significant since they have small p-values. Also, we can look at the determination of coefficient and its R-square is equal to 0.8878. It means that there are 88.78% variation on the response variable y1 explained by the model. We analyze the model based on hypotheses that: All regression coefficients are zero. So, Null Hypothesis: All regression coefficients are zero.

Alternative: at least one coefficients is not zero. From F-test, we know that p-value for the model is equal to 2.2e-16 so that we reject Null Hypothesis. There is at least one coefficient is statistically significant different from others.

In others words, for multiple linear regression, most variable are important variables because there are statistically significant. However, x4 is not relationship with any significant.

For the liner regression method, the core method is the least squares estimates. However, there are some reasons why we are often not satisfied with the least squares estimates. The first is prediction accuracy, and the second reason is interpretation. Due to two main reasons, we use other alternative methods. Below we use subset selection method to analyze it, including forward, and backward selection.

4.2.2 Subset selection method: Forward, backward stepwise

It is known that stepwise selection is a method that allows moves in either direction, dropping or adding variables at the various steps. For forward stepwise, the most important is that the add variable in the model step by step. Opposite, for backward stepwise, the most important is that backward elimination variables from the full model. It means drop one variable from the full model step by step.

For stepwise regression was performance with the response variable (y1) as heating load and the rest of the variable as predictor. First, for forward selection, we can get the forward model and AIC is 1661.42. Second, for backward selection, we can obtain backward model and AIC is 1659.48. Compared backward and forward method, maybe we say backward model is better than forward model because AIC in backward model is smaller than forward model. Then, we use response variable (y2) to analyze it. AIC of forward is equal to 1795.13. AIC of backward is equal to 1792.81. Also, we say backward model is better than forward model because AIC in backward model is smaller than forward model.

4.3 Principal component analysis and Partial least squares

The goals of a principal component analysis are data reduction and interpretation. The objective of my case is to do principal component analysis is to determine the number of principal components. From Table 2: principal components analysis, the fifth principal component explains 97.80% of the total sample variance. So, we can say that sample variation is summarized very well by eight principal components. It can conclude that retaining 5 components would give us enough information.

	Comp1	Comp2	Comp3	Comp4	Comp5
Standard deviation	2.285	1.238	1.104	1.000	0.897
Proportion of Variance	0.522	0.153	0.122	0.100	0.084
Cumulative Proportion	0.522	0.676	0.798	0.898	0.978

Table 2: Principal component analysis

Like using principal components regression (PCR), Partial least squares (PLS) is a dimension reduction method. From Table 3: Partial least squares, the third principal component explains approach 100% of the total sample variance. So, we can say that sample variation is summarized

very well by three principal components. It can conclude that retaining 3 components may give us enough information.

% variance explained	Comp1	Comp2	Comp3	Comp4	Comp5
X	78.10	99.62	99.95	99.97	99.99
Y1	63.49	80.97	99.07	99.97	100
X	78.36	99.62	99.95	99.97	99.99
Y2	63.96	79.92	98.84	99.96	100

Table 3: Partial least squares

4.4 Tree – model approach

In statistics, decision tree is a predictive model, which indicates to map observations about an item to conclusions about the item's target value. It is one of the predictive modeling approaches used between machine learning and data mining. There are three main elements to construct a tree. First, it is the selection of the splits. For example, how to split it? Second, after how to grow tree, how do we decide when to declare a node terminal and stop splitting? Third, we have to assign each terminal node to a class.

In our study, from Figure 4: Regression Tree, we fit a regression tree and it indicates that only three of the variables have been used in construction tree. Also from tree plot, there are seven numbers of terminal nodes. To predict Energy efficiency of Building with a regression tree, it based on x5, x7, and x9. The split of at the top of 5 result in two larger branches. The left branch corresponds to x5< 5.25 and the right branch corresponds to x5>=5.25. Then, for the left subbranch, it also was divided into two branches. Left one: x7 is less than 0.175. Right one: X7 is larger than 0.175. For right sub-branch, at the beginning, it divided into branches, Left one: x2 is less than 624.75. Right one: x2 is larger than 624.75.

Thus, it may say that x5, x7, and x2 are relatively important variables because they are statistically significant.

Next, we used random forest method to analyze this data. The random forest method can improve performance over trees by most situations. The random forest(Breiman 2001) is an ensemble approach that can also be thought of as form nearest neighbor predictor. The random forest starts with a standard maching learing technique called a “decision tree”, which corresponds to our weak learner, In a decision tree, an input is entered at the top ans as it traveraer down the tree the data gets bucketed into smaller and smaller set. From Figure 4: random forest, obviously, x2, x1, x5, and x4 are important variables because they are statistically significant.

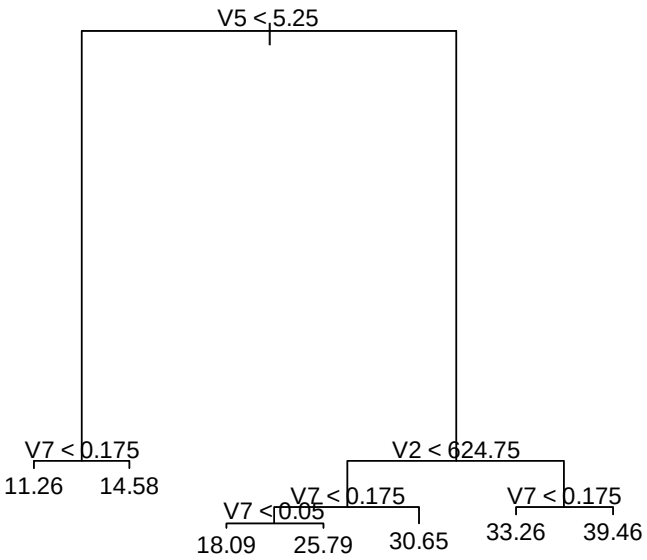


Figure 2: Regression Tree

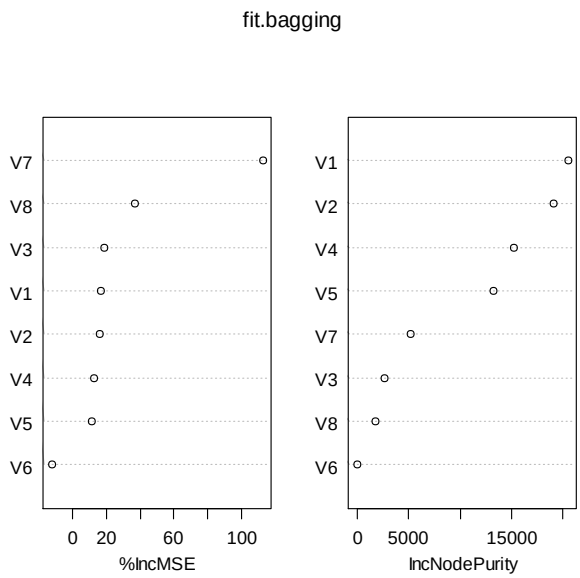


Figure 3: random forest

5. Conclusion and Future work

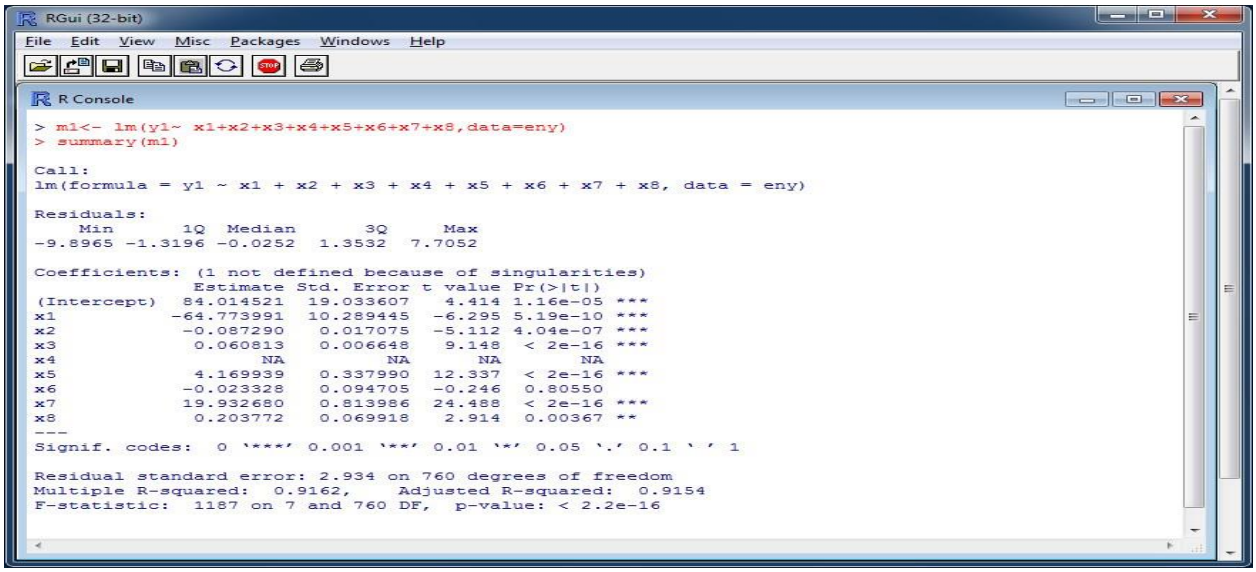
5.1Conclusions

In conclusion, from our analysis, we use regression analysis to analyze the data set. Most variables are important variables because there are statistically significant. Especially, x2=surface area, x5=overall height, and x7=glazing area. In most cases, when other conditions are unchanged, surface decrease will lead to increase the energy efficiency; When other conditions are unchanged, overall height or glazing area increase will lead to increase the energy efficiency. However, x4=roof area is not relationship with any significant. So, this energy efficiency data is no longer complicate for us.

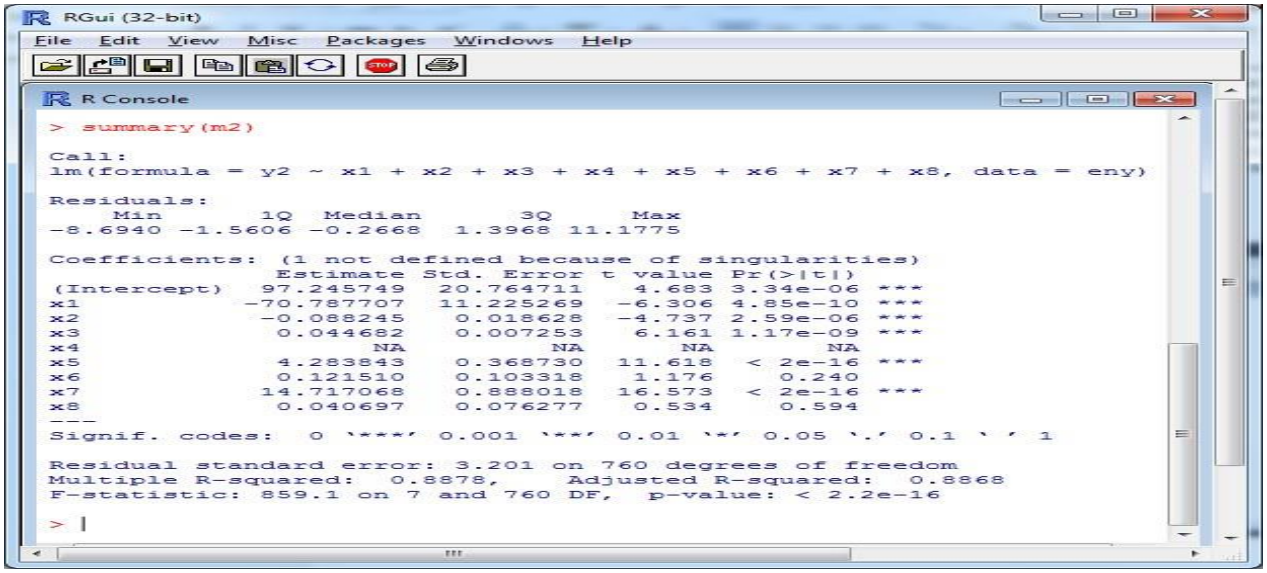
5.2Future work

The more I learn, the more I learn how little I know. For this project, the input variables are not all marginal normal and the response variable is discrete. In the future, we can try to figure them out. And, other methods of support vector machines (SVM) and neutral network will be used in the future.

Appendix



Output 1: model 1



Output 2: model 2

Reference

Tsanas, A. Xifara: 'Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools', Energy and Buildings, Vol. 49, pp. 560-567, 2012

Lee, S., Park, Y., and Kim, C. (2012) Investigating the Set of Parameters Influencing Building Energy Consumption. ICSDC 2011: pp. 211-221.

Richard A. J &Dean W. W, Applied Multivariate statistical analysis: Principal component. Summarizing sample variation by principal components 471-483

Gareth J & Daniel W & Trevor H & Robert T, An Introduction to Statistical Learning with Applications in R:Linear Regression 71-109

Gareth J & Daniel W & Trevor H & Robert T, An Introduction to Statistical Learning with Applications in R:Tree_Based methods

Derksen, Shelley, and H. J. Keselman. "Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables." *British Journal of Mathematical and Statistical Psychology* 45.2 (1992): 265-282.

Hastie, Trevor, et al. "The elements of statistical learning: data mining, inference and prediction." *The Mathematical Intelligencer* 27.2 (2005): 83-85

Iwahashi, Naoto, and Yoshinori Sagisaka. "Statistical modelling of speech segment duration by constrained tree regression." *IEICE TRANSACTIONS on Information and Systems* 83.7 (2000): 1550-1559.