Energy Efficiency Data Set Report

Lokesh Sharma, C20973921

Introduction

The aim of the report is to study and analyze the Energy Efficiency Dataset. I have performed multiple linear regression on this dataset using response variable called Heating Load. Along with this variable, I also performed an analysis on other variables in the dataset and studied the significance of these variables to build the linear regression model. I build the model using following two criterions: Ordinary Least squares and Backward Step wise selection method using Akaike Information Criterion(AIC) values to represent the data linearly.

Data Set Information

Energy analysis is performed using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses.

Information about the 10 variables is
1. Relative Compactness 2. Surface Area 3. Wall Area 4. Roof Area 5. Overall Height
6. Orientation 7. Glazing Area 8. Glazing Area Distribution 9. Heating Load 10. Cooling Load
From the above variables, we are considering Heating Load and Cooling Load as response variables.

Exploratory Data Analysis

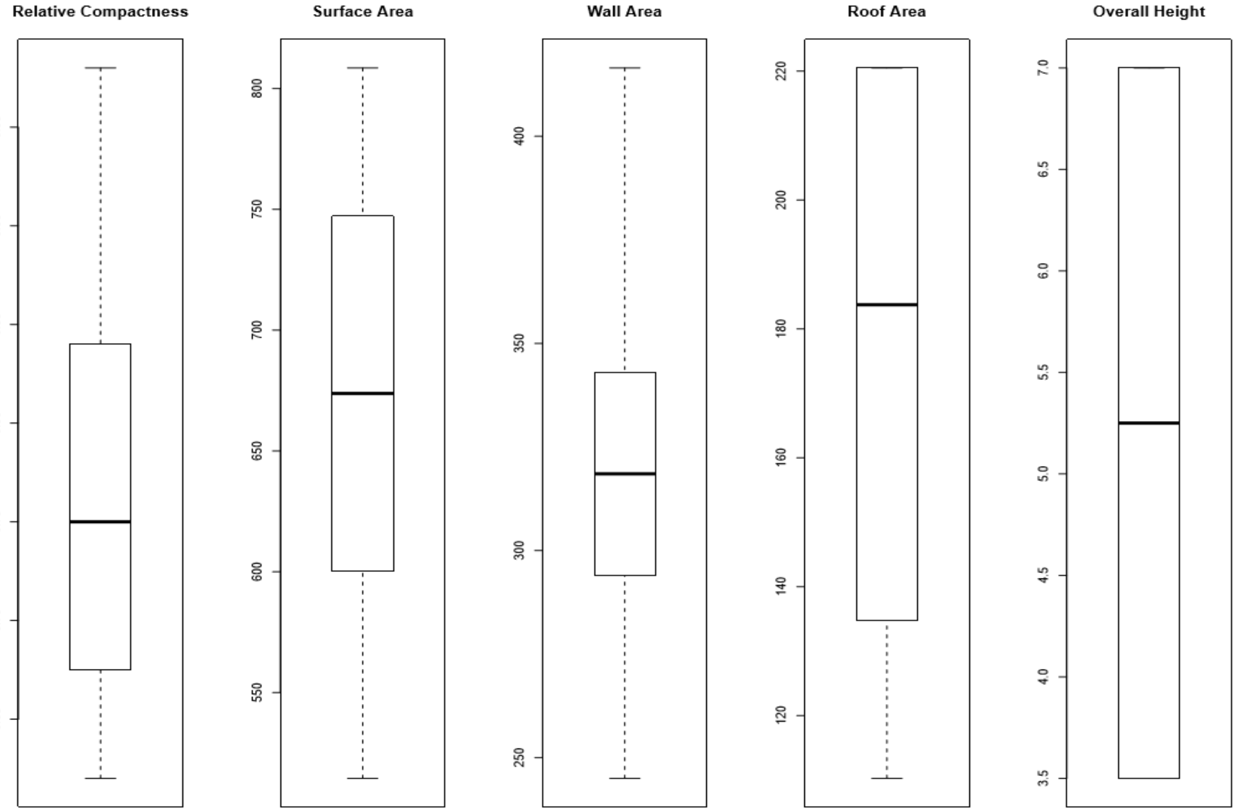
As a first step, I calculated the correlation between these two response variables and other variables and, also studied the correlation among all the variables.

	relative.compactness	surface.area	wall.area	roof.area	overall.height	orientation	glazing.area	glazing.area.distribution	heating.load	cooling.load
relative.compactness	1	-0.9919	-0.20378	-0.86882	0.82775	0	7.62E-20	0	0.62227	0.63434
surface.area	-0.9919	1	0.195502	0.88072	-0.8581	0	4.66E-20	0	-0.65812	-0.673
wall.area	-0.20378	0.1955	1	-0.29232	0.28098	0	0	0	0.45567	0.42712
roof.area	-0.86882	0.88072	-0.29232	1	-0.9725	0	-1.20E-19	0	-0.86183	-0.8625
overall.height	0.827747	-0.8581	0.280976	-0.97251	1	0	0	0	0.88943	0.89579
orientation	0	0	0	0	0	1	0	0	-0.00259	0.01429
glazing.area	7.62E-20	4.66E-20	0	-1.20E-19	0	0	1	0.212964221	0.26984	0.2075
glazing.area.distribution	0	0	0	0	0	0	0.212964221	1	0.08737	0.05053
heating.load	0.622272	-0.6581	0.455671	-0.86183	0.88943	-0.0025865	0.269840996	0.087367594	1	0.97586
cooling.load	0.634339	-0.673	0.427117	-0.86255	0.89579	0.0142896	0.207504991	0.050525119	0.97586	1

As shown in correlation matrix, the correlation between two response variables is very high (i.e. 0.975). Hence by studying one of the response variables, we can predict the value of other response variable. So we have considered Heating Load as the only response variable. Overall Height variable has a very good correlation with the response variable Heating Load (i.e. 0.889), hence we can say that Overall Height is good predictor of response variable Heating Load. Also, Relative Compactness has a good correlation with the response variable, so it is also a good predictor of response variable.

Study of Variables

Before we do the linear regression for the given dataset and variables, we will need to study the variables of the datasets. I have also created boxplots of the quantitative variables to find out any outliers.



It is clear from the Boxplots that there is no outlier in the predictors. Thus, we can go ahead with further analysis

Model Building with Multiple Linear Regression

Linear Regression was performed with the response variable as Heating Load and the rest of the variables as predictors.

During linear regression, we analyze the model based on hypothesis that: All regression coefficients are 0. So, Null Hypothesis=All regression coefficients are zero.
Alternate Hypothesis=At least one coefficient is not zero.
Since the p-value for the model is less than alpha i.e. 2.2e-16 we can reject the Null Hypothesis.

```
# Building the full model using lm() fuction

m1= lm(heating.load~relative.compactness+surface.area+wall.area+roof.area+
       overall.height+factor(orientation)+
       factor(glazing.area)+factor(glazing.area.distribution),data)

summary(m1)
```

The summary of full model shows that Roof Area does not have any relationship with the response variable. It does not predict the value of response variable. The p-value of Orientation and Glazing Area Distribution is also not significant. All other attributes have a significant p-value. Also, the R-squared value is very high. It is 91.62% which implies that it fits the model data. To confirm my initial finding from the full model, I performed Stepwise Linear Regression.

Stepwise Regression
I have performed stepwise regression in the backward direction using step() function in R which is based on **Akaike Information Criterion (AIC).**

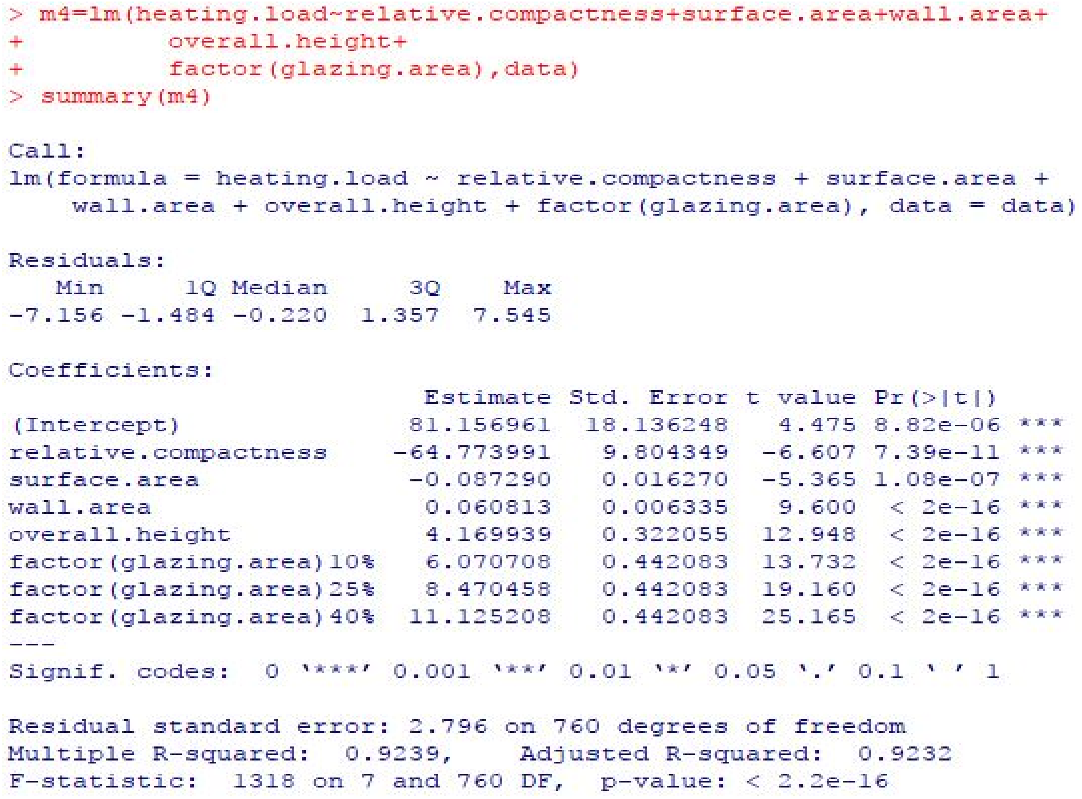
During each step the attribute with the highest AIC value is removed. The AIC of the model increases when the attribute with the lowest p-value is removed.

Model	Number of Predictor	Name of Predictors	R-Square	Adjusted R Square	Residual Standard Error	AIC Value
M1(Full Model)	8	Relative Compactness, Surface Area, Wall Area, **Roof Area**, Overall Height, Orientation, Glazing Area, Glazing Area Distribution	0.9241	0.9227	2.805	1599.24
M2	7	Relative Compactness, Surface Area, Wall Area, Overall Height, Orientation, Glazing Area, **Glazing Area Distribution**	0.9241	0.9227	2.805	1599.24
M3	6	Relative Compactness, Surface Area, Wall Area, Overall Height, **Orientation**, Glazing Area	0.9239	0.9229	2.801	1593.03
M4(Final Model)	5	Relative Compactness, Surface Area, Wall Area, Overall Height, Glazing Area	0.9239	0.9232	2.796	1587.25

Progressing from full model “m1” to final reduced model “m4’, we can see that Adjusted R square value constantly increased and Residual Standard Error constantly decreased leading to a better model.

Also, with respect to the p value obtained from the final model m4, we can see that attributes like Relative Compactness, Surface Area, Wall Area, Overall Height and Glazing Area play a significant part in deciding the value of response variable.

Final Model Results:



```
> m4=lm(heating.load~relative.compactness+surface.area+wall.area+
+          overall.height+
+          factor(glazing.area),data)
> summary(m4)

Call:
lm(formula = heating.load ~ relative.compactness + surface.area + 
    wall.area + overall.height + factor(glazing.area), data = data)

Residuals:
   Min     1Q Median     3Q    Max 
-7.156 -1.484 -0.220  1.357  7.545 

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)    
(Intercept)              81.156961  18.136248   4.475 8.82e-06 ***
relative.compactness    -64.773991   9.804349  -6.607 7.39e-11 ***
surface.area             -0.087290   0.016270  -5.365 1.08e-07 ***
wall.area                 0.060813   0.006335   9.600  < 2e-16 ***
overall.height            4.169939   0.322055  12.948  < 2e-16 ***
factor(glazing.area)10%   6.070708   0.442083  13.732  < 2e-16 ***
factor(glazing.area)25%   8.470458   0.442083  19.160  < 2e-16 ***
factor(glazing.area)40%  11.125208   0.442083  25.165  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.796 on 760 degrees of freedom
Multiple R-squared:  0.9239,    Adjusted R-squared:  0.9232 
F-statistic:  1318 on 7 and 760 DF,  p-value: < 2.2e-16
```

Prediction Equation:

Heating.load.hat= 81.15-(64.77*relative.compactness)-(0.08*surface.area)+(0.06*wall.area)+(4.16*overall.height)+(6.07*(I.glazing.area=10%))+(8.47*(I.glazing.area=25%))+(11.12*(I.glazing.area=40%))

Conclusion

Heating Load and Cooling Load depend on same variables because of high correlation. The variables that plays an important part in predicting their values are Relative Compactness, Surface Area, Wall Area, Overall Height, and Glazing Area.

The paper which cites this dataset is:

A. 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