Project 3: Heating Load Model for Buidlings
Xinyi (Anny) Cui

1 Introduction

How much energy does a house need to keep warm when winter comes, or in a freezing place? The size of the room, the path of the sun's rays, and so on, can all affect the thermal performance of the room. However, what factors affect room warmth (heating load) the most, and how? We can answer these questions by constructing some multiple linear regression models to choose the best one. This paper explores how to construct and choose the most appropriate multiple linear regression model with reasonable explanatory variables to predict the energy required for keeping a building warm.

The background introduces the data information and explain the method and process of constructing the most appropriate multiple linear regression model. The analysis section includes, data exploration, detailed model construction with reasonable explanatory variables, and model selection and evaluation with diagnostic plots. The conclusion section provides the scope of inference, brief results and intervals, and discusses the strengths and weaknesses of the model. The paper will also consider some ways of improvements based on the selected multiple linear regression model and its applicability.

2 Background

The source data was created by Angeliki Xifara (a civil/structural Engineer) and processed by Athanasios Tsanas (a professor in Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK) in 2012. They performed energy analysis using 12 different building shapes simulated in Ecotect. Ecotect is a software used to simulate buildings in the real environment to calculate their energy consumption. They simulated building models with total 8 features such as relative compactness to get 768 samples.

2.1 Data Information

We extract the dataset from Kaggle, *Energy Efficiency Dataset*, to construct reasonable multiple linear regression models and then select the most appropriate one, which heat load is the response. There are two variables in this data set that can be responses, heating load and cooling load. This dataset contains 11 variables, but we only consider 10 of them, heating load as the response, and other variables except cooling load as possible explanatory variables. This paper only explores the multiple linear regression model with heating load as the response variable, so cooling load will not appear in this observation.

2.1.1 Data Content

For each building sample, there are 8 features can be reasonable explanatory variable to predict its heating load, expressed in different units. Tables below provides the description of the data format. **GAD** (see Table 2) represents different distribution scenarios for each glazing area (types). Each number means one of these types.

1 uniform: with 25% glazing on each side,

2 north: 55% on the north side and 15% on each of the other sides,

3 east: 55% on the east side and 15% on each of the other sides,

4 south: 55% on the south side and 15% on each of the other sides, and

5 west: 55% on the west side and 15% on each of the other sides.

0 none of above

RC	SA	WA	RA	OH	O	GA	GAD	HL
0.98	514.5	294.0	110.25	7	2	0	0	15.55
0.98	514.5	294.0	110.25	7	3	0	0	15.55
0.98	514.5	294.0	110.25	7	4	0	0	15.55
0.98	514.5	294.0	110.25	7	5	0	0	15.55

Table 1: The example of the data format with all needed variables in the first four rows.

Data Variables		
Name	Meaning	Description
RC	Relative Compactness (ratio)	Dividing the sum of all surfaces by its total heated volume.
SA	Surface Area (m^2)	Measure of surface area.
WA	Wall Area (m^2)	Measure of total wall areas.
RA	Roof Area (m^2)	Measure of roof area.
OH	Overall Height (m)	Measure of a sample height.
O	Orientation (number)	Number of sun's paths.
GA	Glazing Area (%)	Percentages of the floor area
GAD	Glazing Area Distribution (number)	Different distribution scenarios for each glazing area (types).
HL	Heating Load (kW)	The amount of heating a building needs to maintain the interior temperature at a given level (kilowatts).

Table 2: The name, meaning, and description of the dataset. For GA, there are four levels of percentage $(0\%, 10\%, 25\%, 40\%)$.

2.2 Method

The main purpose of a multiple linear regression is to make predictions with reasonable explanatory variables. Modeling is used to fit the collected data, and we perform a parameter estimation with the most appropriate explanatory variables. The AIC (Akaike Information Criterion) considers the statistical fit of the model and the number of parameters used to fit, and can compare models. A model with a smaller AIC value indicates that a preferred sufficient fitted model with fewer parameters. Then, we can use the most appropriate model with a new data frame of parameters to make predictions.

This paper will construct and select the best multiple linear regression model for estimating the heating load of a building based on the following process:

1. Data exploration. Check if the data needs to be transformed.

2. Correlation. Determine whether there is a correlation between several specific explanatory variables. Selected explanatory variables should be correlated to the response variable but be careful with the multilinearity. The model should be formed according to the linear formula:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_i X_i + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$$

3. Model selection. Fit models with reasonable explanatory variables and use compare them to select the best one.

4. Assess the most appropriate model we selected.

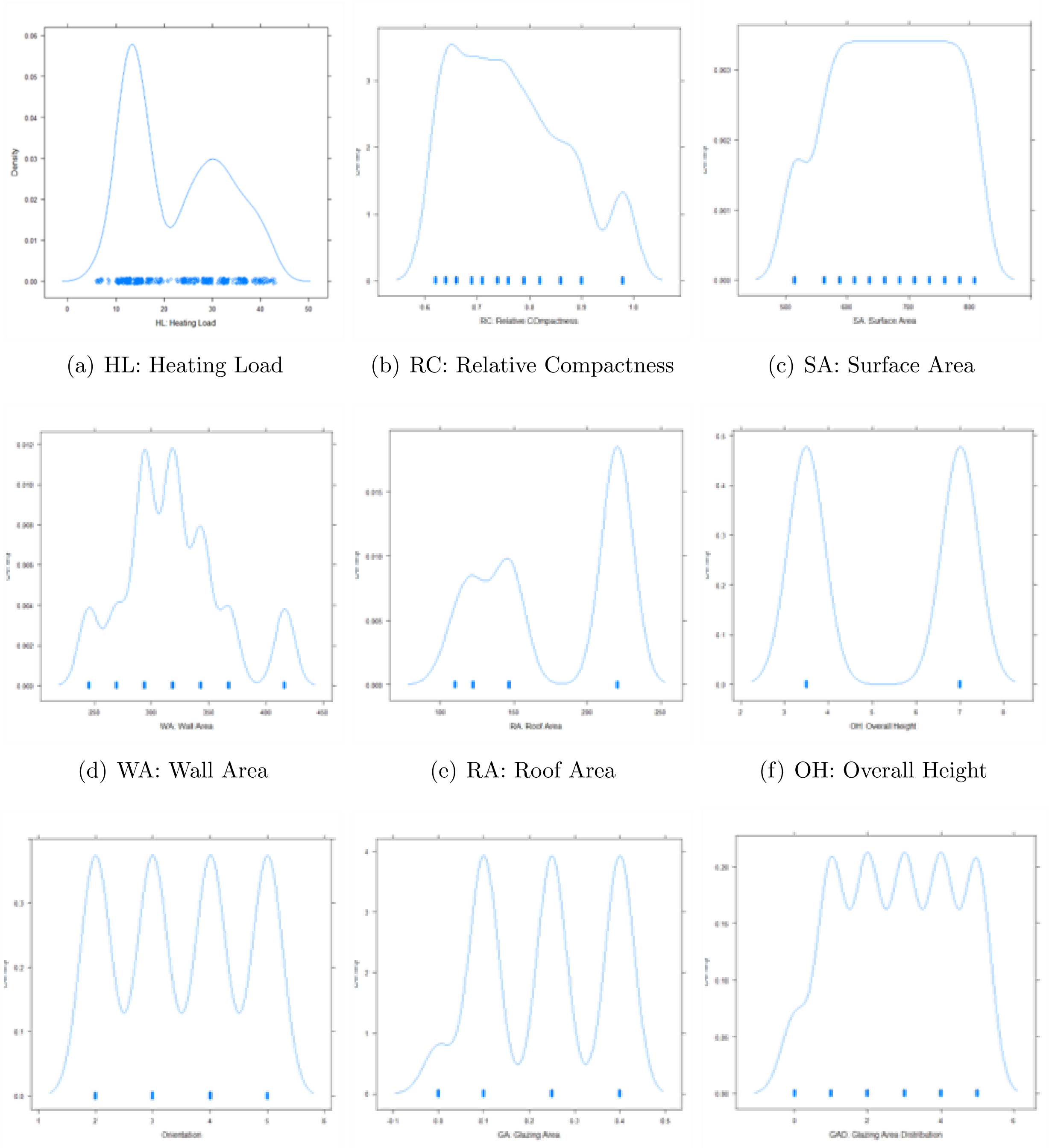
5. Provide a new data frame of parameters and predict the value of the response with the selected model. Know what kind of accuracy this prediction or confidence interval can achieve.

A full MLR model with all possible explanatory variables will reflect the reality better. However, it is the more difficult to interpret and might cause the problem of multilinearity (high correlation of some explanatory variables). We also need to pay attention to whether it supports our hypothesis and if this is the most appropriate model. Therefore, we pursue a model that is as reduced as possible by using AIC and comparisons of R^2 values, linearity and regression diagnostic issues.

3 Analysis

3.1 Data Exploration

Before constructing reasonable models and do the model selection, we need to explore our data to see whether needs a transformation. After doing densityplots (see Figure 1), none of our variables includes skewness or outliers, thus, we do not need to do any transformation.



(a) HL: Heating Load

(b) RC: Relative Compactness

(c) SA: Surface Area

(d) WA: Wall Area

(e) RA: Roof Area

(f) OH: Overall Height

(g) O: Orientation

(h) GA: Glazing Area

(i) GAD: Glazing Area Distribution

Figure 1: Densityplots of all variables. From left to right and then top to bottom are, HL, RC, SA, WA, RA, OH, O, GA, GAD. None of them needs transformation but HL, RA, and OH seem have binary plots.

To see how the response variable HL, kilowatts, is affected by other variables, we create a matrix scatterplot to show their linear relationships (see Figure 2). The matrix of scatterplots includes variables displayed on the diagonal and a variety of fitted lines. This graph shows that the response variable, HL, seems be influenced clearly by RC, SA, WA, RA, OH, and slightly affected by GA and GAD. The variable O has no correlation with any other variables.

We need to be careful with the issue of multilinearity because some pairs of them have clear linear relationship to each other(see Figure 2 and Table 3), such as RC and SA. We need to avoid highly correlated variables appear in the same model. Thus, we need to do more comparisons among our models with selected reasonable explanatory variables and use AIC to select the best one with comparisons.

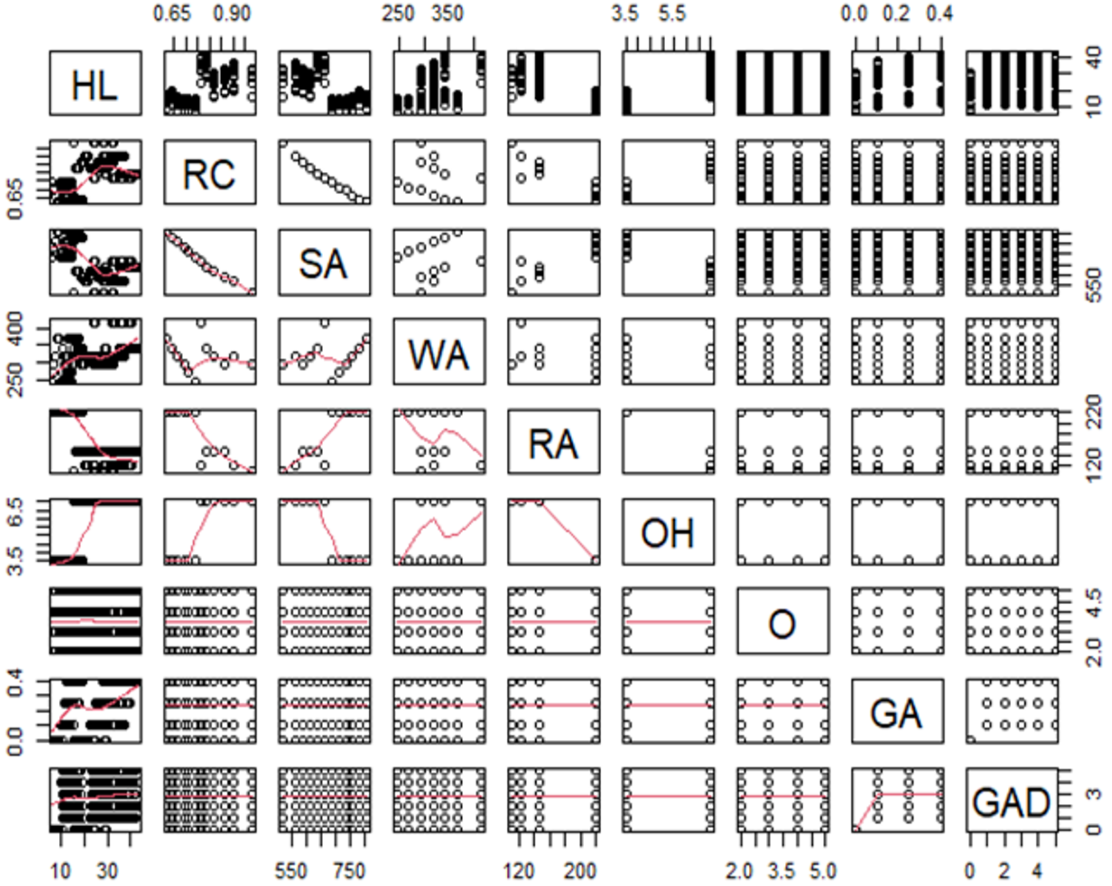


Figure 2: The matrix scatterplot displays each pair of variables. HL is influenced by RC, SA, WA, RA, OH, and slightly affected by GA and GAD, but not O. Possible high correlated explanatory variable pairs are $[RC, SA]$; $[RC, RA]$; $[RC, OH]$; $[SA, RA]$; $[SA, OH]$; $[RA, OH]$

$[RC, SA]$	$[RC, RA]$	$[RC, OH]$	$[SA, RA]$	$[SA, OH]$	$[RA, OH]$
negative	negative	positive	positive	negative	negative

Table 3: Highly correlated pairs. Negative means one variable decrease when the other one increases. Positive otherwise.

3.2 Modeling

3.2.1 Variable Screening

Checking correlation coefficient values for each pair of variables (see Table 4). Selecting reasonable explanatory variables for the response HL by choosing the variables which are highly correlated to HL but not to other selected independent variables.

	HL	RC	SA	WA	RA	OH	O	GA	GAD
HL	1.000								
RC	0.622	1.000							
SA	-0.658	-0.992	1.000						
WA	0.456	-0.204	0.196	1.000					
RA	-0.862	-0.869	0.881	-0.292	1.000				
OH	0.889	0.828	-0.858	0.281	-0.973	1.000			
O	-0.003	0.000	0.000	0.000	0.000	0.000	1.000		
GA	0.270	0.000	0.000	0.000	0.000	0.000	0.000	1.000	
GAD	0.087	0.000	0.000	0.000	0.000	0.000	0.000	0.213	1.000

Table 4: Coefficient values of each pair of variables. Values in the first column close to +1 or −1 means highly correlated to the response HL.

3.2.2 Candidate Models

Constructing reasonable MLR models with following candidates (see Table 5) and use AIC to choose the best two for more detailed comparisons.

Candidate	Form with selected explanatory variables	AIC
Model 1	$HL = \beta_0 + \beta_1 RC + \beta_2 WA + \beta_3 O + \beta_4 GA + \beta_5 GAD$	4447.289
Model 2	$HL = \beta_0 + \beta_1 WA + \beta_2 OH + \beta_3 O + \beta_4 GA + \beta_5 GAD$	3886.792
Model 3	$HL = \beta_0 + \beta_1 WA + \beta_2 OH + \beta_3 GA + \beta_4 GAD$	3884.850
Model 4	$HL = \beta_0 + \beta_1 SA + \beta_2 WA + \beta_3 GA$	4223.834
Model 5	$HL = \beta_0 + \beta_1 WA + \beta_2 OH + \beta_3 GA$	3890.872

Table 5: Five candidate models and their AIC values. Model 2 and Model 3 seems have lower AIC values (Model 3 has the lowest). Model 3 is a reduced form of Model 2, with no variable O.

3.2.3 Model Selection

After checking the AIC values, we find two competitive models, Model 2 with AIC of 3886.792, and Model 3 with AIC of 3884.850. It seems Model 3 is better because it has a lower AIC value. However, since their AIC values are pretty large and close, we cannot easily define which is better as Model 2 has one more explanatory variable, which means it probably reflect the variation in heating load more accurately.

We first check values of variance inflation factor (VIF) for these two models (see Table 6).

Model	Variance Inflation Factor				
Model 2	WA	OH	O	GA	GAD
	1.085714	1.085714	1.000000	1.047508	1.047508
Model 3	WA	OH		GA	GAD
	1.085714	1.085714		1.047508	1.047508

Table 6: VIFs for each explanatory variable in both models. All values of VIF are close to 1, which indicates almost no correlation of selected predictors in models.

Then we can check for analysis of variance for these two models, which provides a p-value of 0.8111. This value is much larger than 0.05, which means it is not necessary for us to keep the complicated form of the model. We can also check the R^2 values of both models, and both provides a value of 0.9108. This means both models can explain 91.08% variation in the response heating load. Thus, the reduced form, Model 3 is better with less explanatory variables.

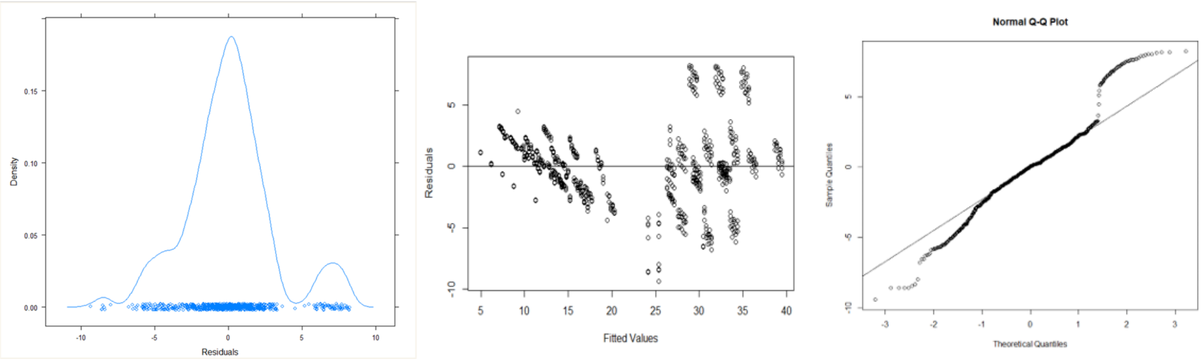
3.3 Fit and Assess

We construct the **Heating Load Model**. This final fitted model (Model 3) is:

$$HL = -24.401478 + 0.051669WA + 4.763278OH + 19.932680GA + 0.203772GAD$$

The F-statistic is 1948 on 4 and 763 DF, which is a good value to show the a statistically significance of coefficients in the linear regression model. The p-value of this model is even less than 2.2×10^{-16}, which is much smaller than 0.05 and indicates that the model is statistically significant.

The diagnostic plots (see Figure 3) indicate that the selected model (Model 3)is well fitted with the selected reasonable explanatory variables. Also, the diagnostic graphs of Densityplot of Residuals (see Figure 3(a)) and the Normal Quantile-Quantile Plot (see Figure 3(c)) clearly displays a binary influence exists in the model. This can be reasonably explained by the variable OH (see Figure 4) since this variable only has two groups of values, 3.5 and 7, and this variable is selected in our final fitted model. OH has the second high coeficient valued of 4.763278 in the model and affect the response HL a lot.

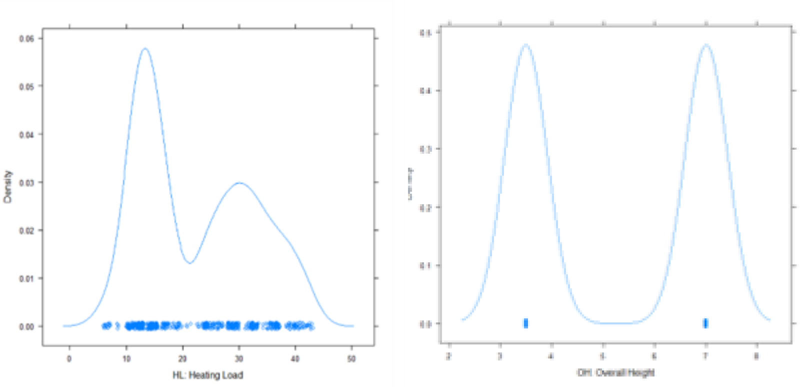


(a) Densityplot of Residuals

(b) Fitted Values v.s. Residuals

(c) Normal Quantile-Quantile Plot

Figure 3: From left to right, graphs are densityplot of model residuals, fitted values v.s. residuals, and the normal quantile plot. The densityplot gives a normal distribution but a kind of binary influence. The fitted values v.s. residuals graph provides a reasonable zero-mean. The normal Q-Q plot gives a linear trend.



(a) Heating Load

(b) Overall Height

Figure 4: Densityplots of variable HL, heating load, and variable OH, overall height. Both variable plots indicates a binary influence.

4 Conclusion

4.1 Scope of Inference

The samples are simulated buildings based on 12 different real building shapes with 8 building parameters. The data is collected partially from a real world (12 kinds of building shapes). Thus, except these 12 sorts of building shapes, we cannot directly use the **Heating Load Model** to predict the energy required for keeping a house warm. All variables are quantitative and the values are relatively random selected, so this model is well fitted with multiple linear regression model. We cannot conclude which cause the heating load, instead, the model can determine the correlation between the response variable HL with selected explanatory variables.

4.2 Inference for Regression

Model 3 wins with the lowest value of AIC (3884.850), and after the test of determination coefficient, its R^2 value is 0.9108 and adjusted R^2 value is 0.9103. These values indicates around 91% variation in HL can be explained by the selected model. The F-statistic of 1948 on 4 and 763 DF is good. Also, we accept the p-value of this model is even less than 2.2×10^{-16} less than 0.05, rejecting the null hypothesis and accept the alternative hypothesis, which means the multiple linear regression model is considered to be statistically significant for us. The relationship between HL and WA, OH, GA, GAD displays a statistically significant linearity with no issues of multilinearity. We can explain that the **Heating Load Model**, *HL* will vary with a negative constant intercept value, and increasing *WA*, *OH*, *GA*, *GAD*, approaching the sample mean. The variable GA will influence HL the most because it has the highest coefficient value of 19.932680.

4.3 Intervals and Interpretation

For buildings with similar shapes provided in the collected dataset, we predict the mean of heating load to be 25.37227 *kW* (95% CI: [24.79152, 25.95303], PI: [19.4127 31.33184]) with randomly selected explanatory variable values of $WA = 318$, $OH = 7$, $GA = 0$, $GAD = 0$.

The 95% confidence interval gives a range of 24.79152 to 25.95303 kW. This means if doing fitted model computations many times with sample values from the same population, we would like to expect 95% of these computations containing the true population mean in the confidence interval, where 24.79152 is the lower limit and 25.95303 is the upper limit.

The prediction interval predicts the range of 19.4127 to 31.33184 kW, where a sample's value of heating load will fall in the future. The prediction interval must consider the uncertainty in estimation of a mean value and the random variation in individual values. Thus, the prediction interval is always wider than the confidence interval, which gives the range of 19.4127 to 31.33184 under the given case of $WA = 318$, $OH = 7$, $GA = 0$, $GAD = 0$.

4.4 Strengths and Weaknesses

Based on our analysis, the selected model can do a good job on predicting the energy required for a building (12 sorts of shapes in 2012) to keep the temperature at the given level. Around 91% variation in heating load can be explained by the model $HL = -24.401478 + 0.051669WA + 4.763278OH + 19.932680GA + 0.203772GAD$. However, we still need to consider that our model is not perfect for predicting a mean value for other kinds of building shapes. Also, since the samples are simulated in Ecotect with the same volume value of $771.75m^3$, we cannot 100% sure there exists no differences compared to the conditions of buildings in the real world. Thus, our results cannot be extended to a broader sorts of buildings or building samples with different values of volume. The causal relationship should not be inferred and we need more data and information to see the correlation between heating load and other possible explanatory variables which are not included in the dataset, such as latitude.

4.5 Way to Improve

Since some explanatory variables are highly correlated and several of them can be divided into different groups with different value levels, such as OH. Thus, we can try to make multiple linear regression models with interactions for better fittness.

5 Bibliography

Data: Energy Efficiency Dataset
https://www.kaggle.com/datasets/elikplim/eergy-efficiency-dataset
From https://archive.ics.uci.edu/ml/datasets/Energy+efficiency

Version: R-4.2.1

Packages: lattice, car

Base Packages: utils, base, graphics, stats

Tsanas, A. and Xifara, A. (2012). *Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools.* Energy and Buildings, Vol.49, pp. 560-567.
https://tarjomefa.com/wp-content/uploads/2017/04/6453-English-TarjomeFa.pdf