Energy Analysis

Dillirajan Sankar, Faheem Kamaludeen Mohideen, Nishitha Chidipothu

Table of contents

1 INTRODUCTION

2 GOAL

3 DATA DESCRIPTION

4 MODELING

5 RESULTS

6 CONCLUSIONS

Introduction

The increasing demand for sustainable and energy-efficient building has led to to a growing interest in the prediction and estimation of energy performance in residential buildings.

Heating load is a measure of the amount of heat required to maintain a comfortable indoor temperature in a building during the heating season.

By accurately predicting the amount of heat required to maintain a comfortable indoor temperature, they can adjust their heating systems accordingly, potentially saving on energy costs and reducing their carbon footprint.

Additionally, predicting heating load can be helpful in the design and construction of new buildings.

Goals

- ## Objectives

1. Perform energy analysis of residential buildings.
2. Perform Residual Analysis to detect and remove outliers.
3. Determine the best fitting model using selection process.

- ## Questions

1. **Does a multiple linear regression model require all regressors for the best model?**
2. **What are the most significant predictors that affect heating load?**

Data Description

Energy Efficiency Dataset

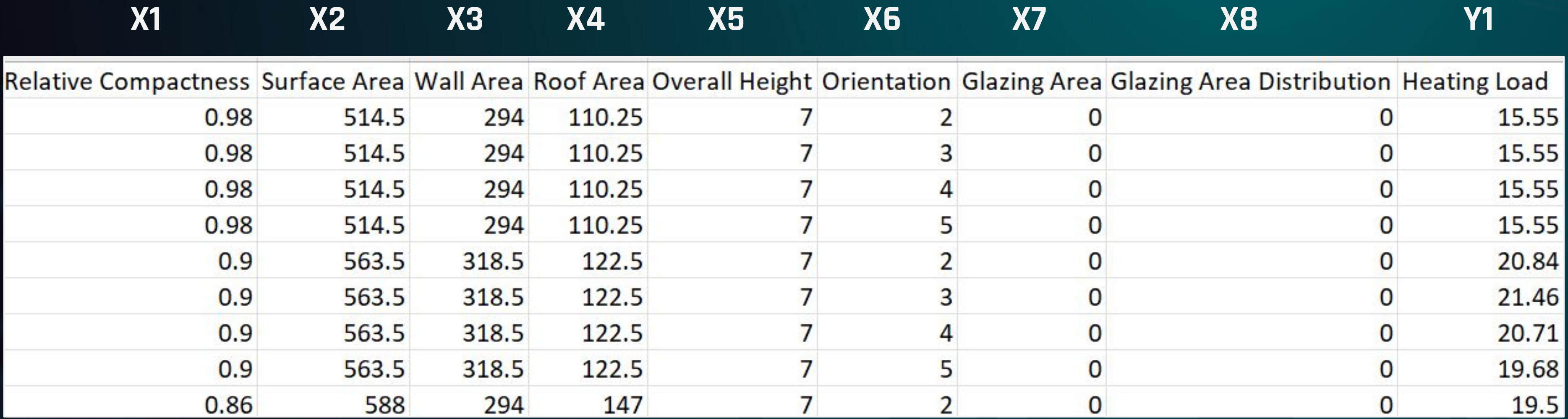
- Data Source: Energy Efficiency Dataset

- The dataset used in this analysis was created by Angeliki Xifara and processed by Athanasios Tsanas at the University of Oxford.

- This dataset contains 768 instances (residential buildings) with 8 features and targets Heating load (Y1) & Cooling load (Y2).

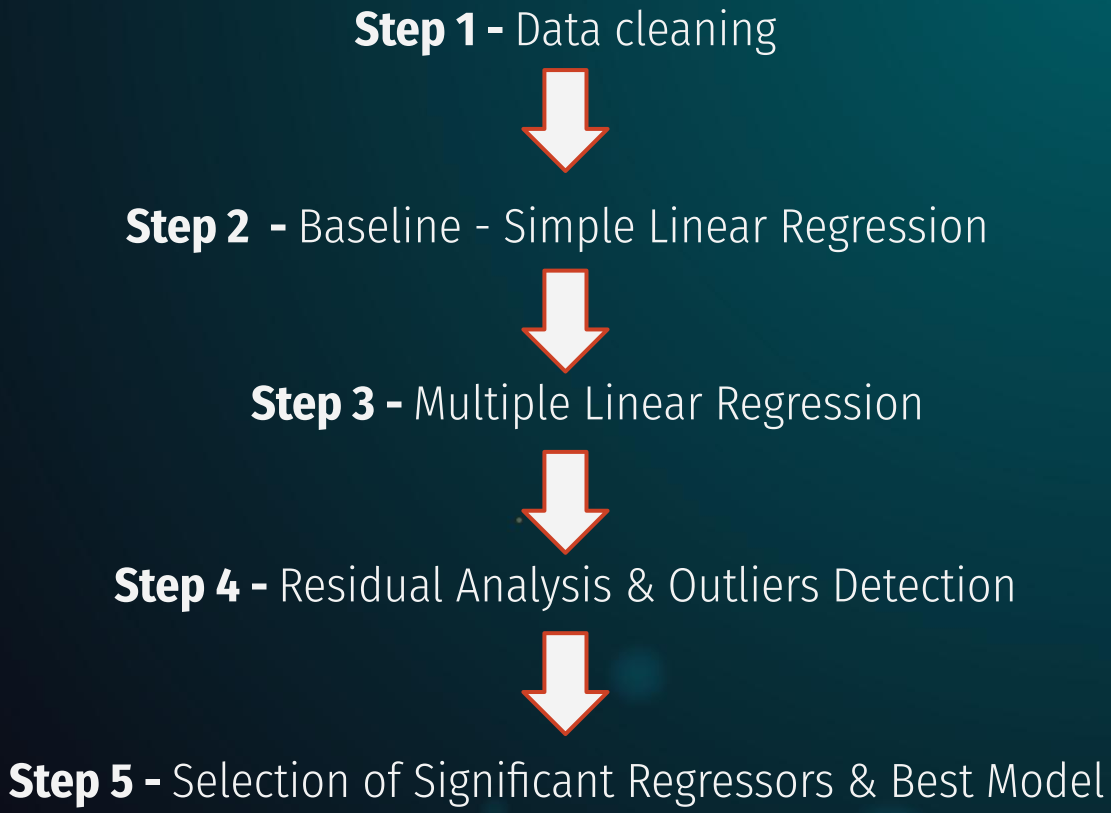
- The 8 regressors include- Relative compactness (X1), Surface Area (X2), Wall Area (X3), Roof Area (X4), Overall height (X5), Orientation (X6), Glazing Area (X7), Glazing Area Distribution (X8).

Data Sample



X1	X2	X3	X4	X5	X6	X7	X8	Y1
Relative Compactness	Surface Area	Wall Area	Roof Area	Overall Height	Orientation	Glazing Area	Glazing Area Distribution	Heating Load
0.98	514.5	294	110.25	7	2	0	0	15.55
0.98	514.5	294	110.25	7	3	0	0	15.55
0.98	514.5	294	110.25	7	4	0	0	15.55
0.98	514.5	294	110.25	7	5	0	0	15.55
0.9	563.5	318.5	122.5	7	2	0	0	20.84
0.9	563.5	318.5	122.5	7	3	0	0	21.46
0.9	563.5	318.5	122.5	7	4	0	0	20.71
0.9	563.5	318.5	122.5	7	5	0	0	19.68
0.86	588	294	147	7	2	0	0	19.5

Project Approach



Step 1 - Data cleaning

Step 2 - Baseline - Simple Linear Regression

Step 3 - Multiple Linear Regression

Step 4 - Residual Analysis & Outliers Detection

Step 5 - Selection of Significant Regressors & Best Model

Data Cleaning & Descriptive Statistics

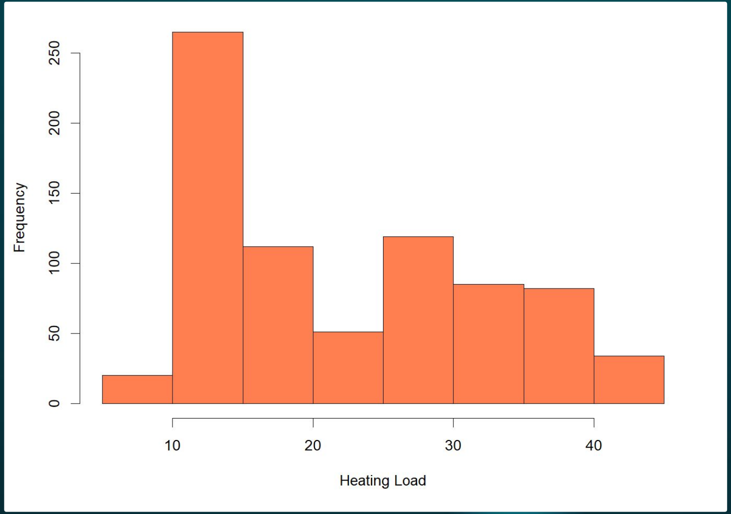
Only Heating Load is used as target

Checking for Nulls:

- No null values in the observations

Descriptive Statistics:

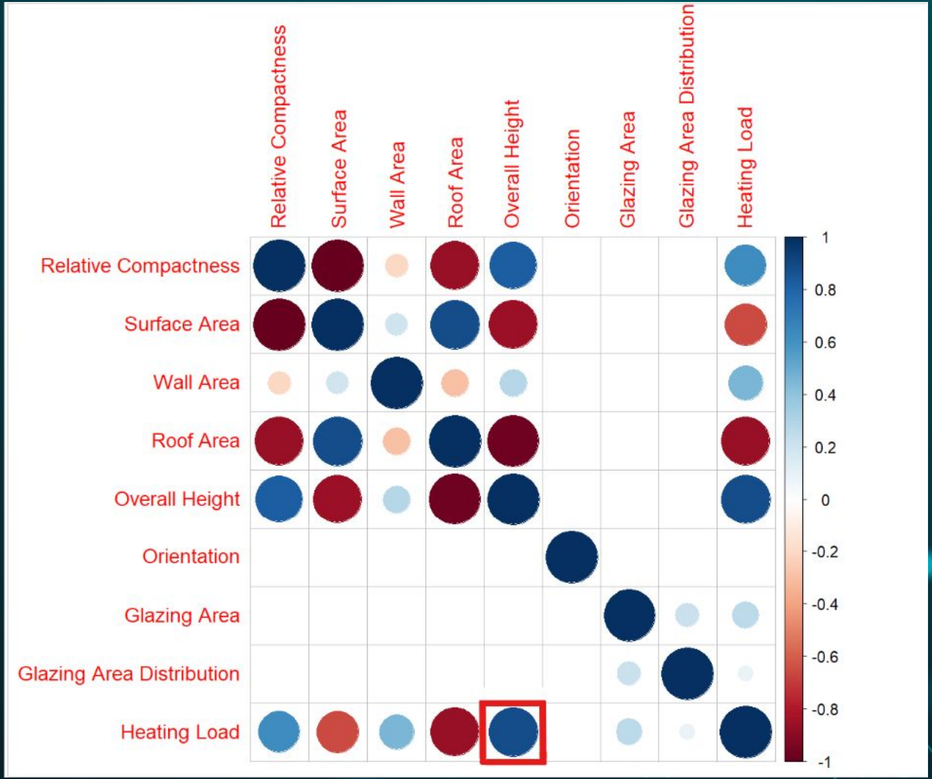
- Maximum heating load: 43.10 kilowatts
- Minimum heating load: 6.01 kilowatts
- Mean heating load: 22.31 kilowatts
- Median heating load: 18.95 kilowatts



Correlation between features and target

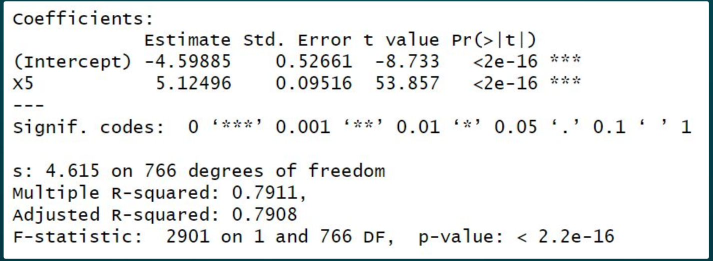
Strong correlation between:

Overall Height & Heating Load



Baseline Model

- Simple Linear Regression
- The single regressor: Overall Height
- Due to strong correlation with the target
- Correlation of 0.889



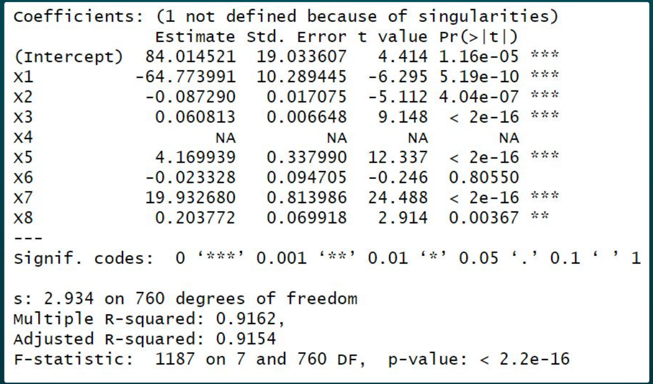
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.59885    0.52661  -8.733   <2e-16 ***
X5           5.12496    0.09516  53.857   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 4.615 on 766 degrees of freedom
Multiple R-squared: 0.7911,
Adjusted R-squared: 0.7908
F-statistic:  2901 on 1 and 766 DF,  p-value: < 2.2e-16
```

Baseline Model Results:

Multiple R-squared	Adjusted R-squared
0.7911	0.7908

Multiple Linear Regression Model

- All 8 regressors
- N/A for ‘Roof Area’



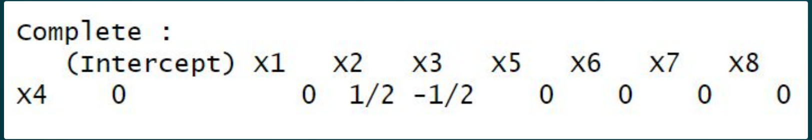
```
Coefficients: (1 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  84.014521  19.033607   4.414 1.16e-05 ***
X1          -64.773991  10.289445  -6.295 5.19e-10 ***
X2           -0.087290   0.017075  -5.112 4.04e-07 ***
X3            0.060813   0.006648   9.148  < 2e-16 ***
X4                  NA         NA      NA       NA    
X5            4.169939   0.337990  12.337  < 2e-16 ***
X6           -0.023328   0.094705  -0.246  0.80550    
X7           19.932680   0.813986  24.488  < 2e-16 ***
X8            0.203772   0.069918   2.914  0.00367 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 2.934 on 760 degrees of freedom
Multiple R-squared: 0.9162,
Adjusted R-squared: 0.9154 
F-statistic:  1187 on 7 and 760 DF,  p-value: < 2.2e-16
```

Model Results:

Multiple R-squared	Adjusted R-squared
0.9162	0.9154

Updated Multiple Linear Regression Model

- Roof area = ½ (Surface Area) - ½ (Wall Area)
- Perfect collinearity exists
- Remove ‘Roof Area’

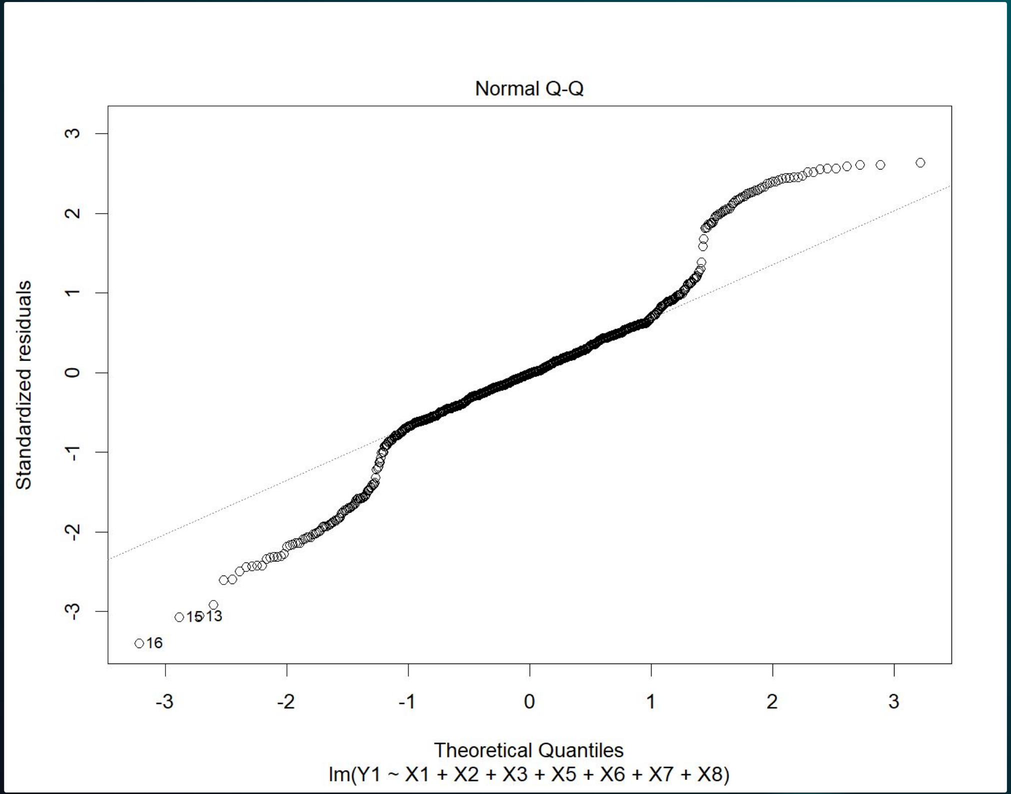


```
Complete :
   (Intercept) x1   x2   x3   x5   x6   x7   x8
x4    0           0  1/2 -1/2    0    0    0    0
```

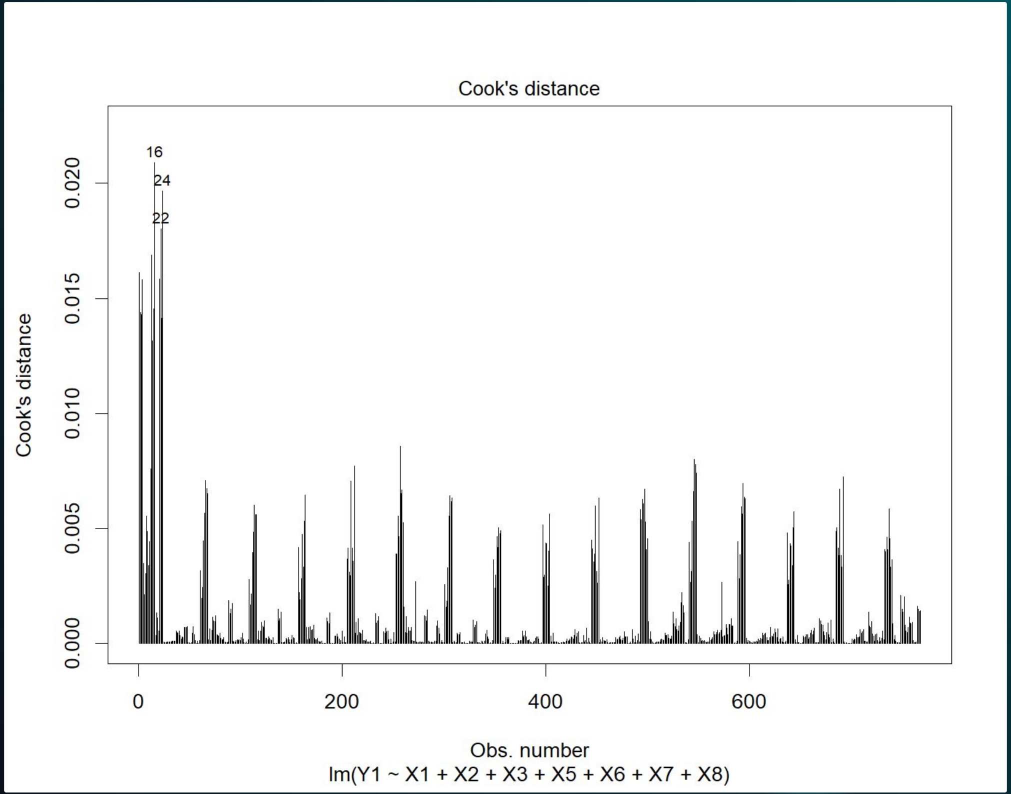
Model Results:

Multiple R-squared	Adjusted R-squared
0.9162	0.9154

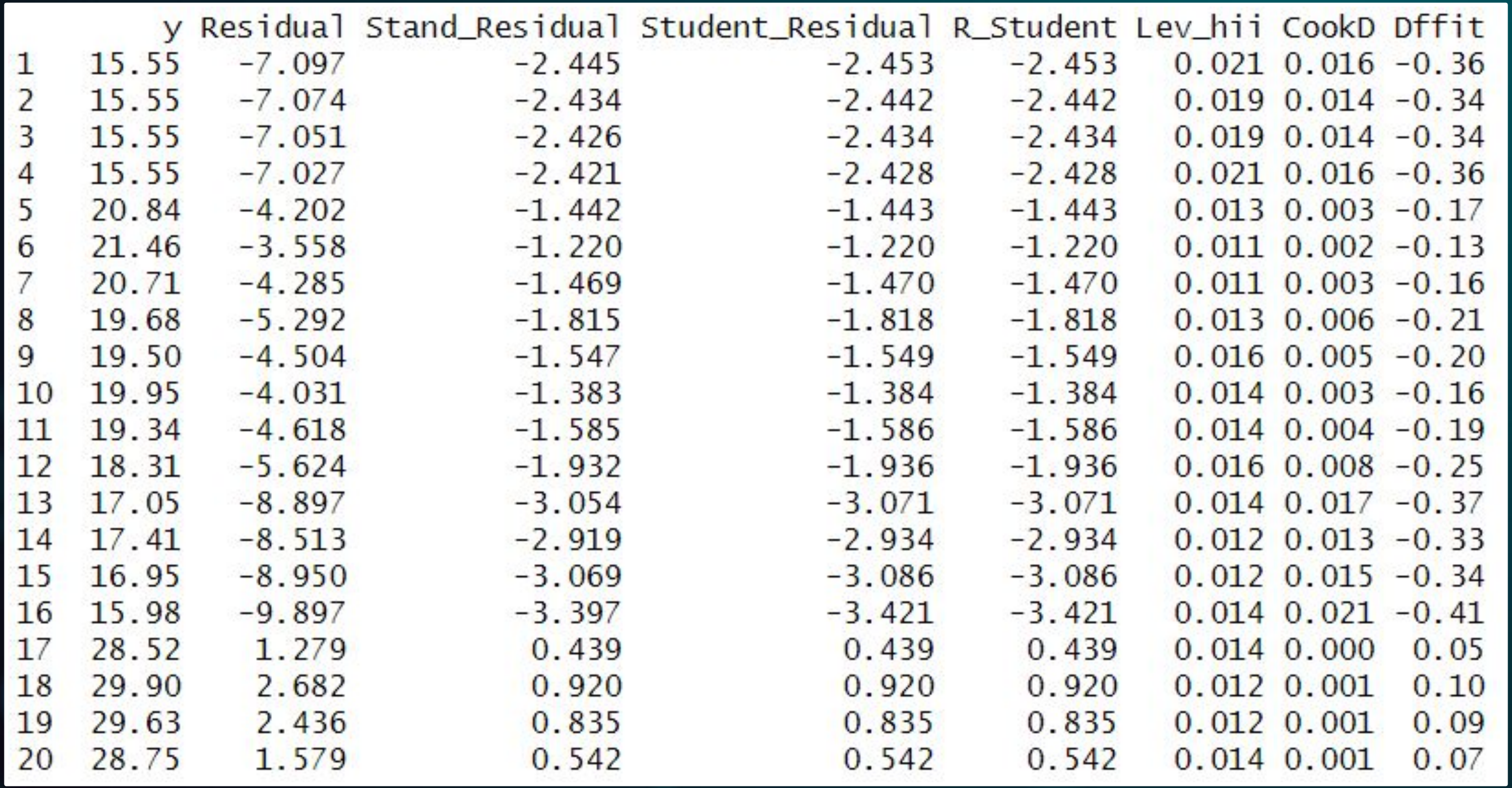
QQ-Plot


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-3
16
15
13
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Y1 ~ X1 + X2 + X3 + X5 + X6 + X7 + X8)

Cook’s Distance Plot

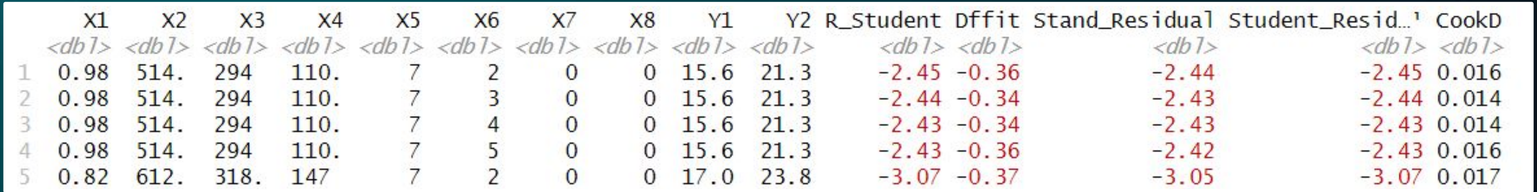


Residual Analysis



```
       y Residual Stand_Residual Student_Residual R_Student Lev_hii CookD Dffit
1  15.55   -7.097         -2.445           -2.453    -2.453   0.021 0.016 -0.36
2  15.55   -7.074         -2.434           -2.442    -2.442   0.019 0.014 -0.34
3  15.55   -7.051         -2.426           -2.434    -2.434   0.019 0.014 -0.34
4  15.55   -7.027         -2.421           -2.428    -2.428   0.021 0.016 -0.36
5  20.84   -4.202         -1.442           -1.443    -1.443   0.013 0.003 -0.17
6  21.46   -3.558         -1.220           -1.220    -1.220   0.011 0.002 -0.13
7  20.71   -4.285         -1.469           -1.470    -1.470   0.011 0.003 -0.16
8  19.68   -5.292         -1.815           -1.818    -1.818   0.013 0.006 -0.21
9  19.50   -4.504         -1.547           -1.549    -1.549   0.016 0.005 -0.20
10 19.95   -4.031         -1.383           -1.384    -1.384   0.014 0.003 -0.16
11 19.34   -4.618         -1.585           -1.586    -1.586   0.014 0.004 -0.19
12 18.31   -5.624         -1.932           -1.936    -1.936   0.016 0.008 -0.25
13 17.05   -8.897         -3.054           -3.071    -3.071   0.014 0.017 -0.37
14 17.41   -8.513         -2.919           -2.934    -2.934   0.012 0.013 -0.33
15 16.95   -8.950         -3.069           -3.086    -3.086   0.012 0.015 -0.34
16 15.98   -9.897         -3.397           -3.421    -3.421   0.014 0.021 -0.41
17 28.52    1.279          0.439            0.439     0.439   0.014 0.000  0.05
18 29.90    2.682          0.920            0.920     0.920   0.012 0.001  0.10
19 29.63    2.436          0.835            0.835     0.835   0.012 0.001  0.09
20 28.75    1.579          0.542            0.542     0.542   0.014 0.001  0.07
```

Outliers



```
      X1    X2    X3    X4    X5    X6    X7    X8    Y1    Y2 R_Student Dffit Stand_Residual Student_Resid…¹ CookD
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>     <dbl> <dbl>          <dbl>           <dbl> <dbl>
1   0.98  514.  294   110.     7     2     0     0  15.6  21.3     -2.45 -0.36          -2.44           -2.45 0.016
2   0.98  514.  294   110.     7     3     0     0  15.6  21.3     -2.44 -0.34          -2.43           -2.44 0.014
3   0.98  514.  294   110.     7     4     0     0  15.6  21.3     -2.43 -0.34          -2.43           -2.43 0.014
4   0.98  514.  294   110.     7     5     0     0  15.6  21.3     -2.43 -0.36          -2.42           -2.43 0.016
5   0.82  612.  318.  147      7     2     0     0  17.0  23.8     -3.07 -0.37          -3.05           -3.07 0.017
```

R-Student Outliers

$$|t_i| > t_{\alpha/2,n-p-1}$$

$$t_{\alpha/2,n-p-1} = 1.96309$$

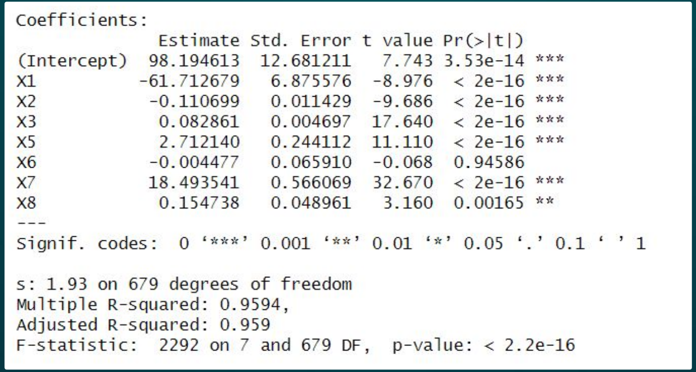
81 Observations were out of the threshold and removed

Other Residuals

- Standardized Residuals all within 3
- Studentized Residuals less than 3

Multiple Linear Regression Model without Outliers

- 7 Regressors
- Roof Area removed due to collinearity



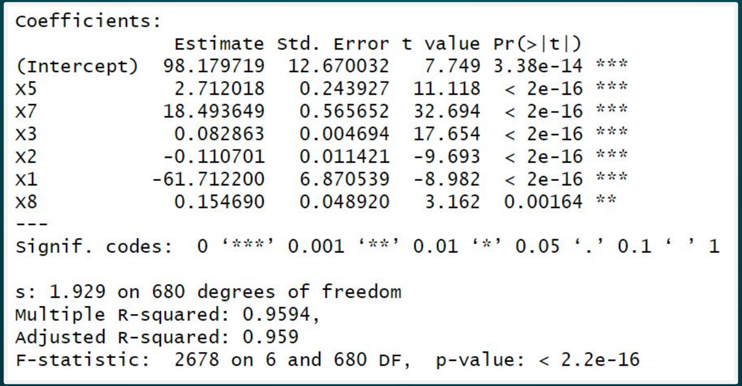
```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  98.194613  12.681211   7.743 3.53e-14 ***
X1          -61.712679   6.875576  -8.976  < 2e-16 ***
X2           -0.110699   0.011429  -9.686  < 2e-16 ***
X3            0.082861   0.004697  17.640  < 2e-16 ***
X5            2.712140   0.244112  11.110  < 2e-16 ***
X6           -0.004477   0.065910  -0.068  0.94586
X7           18.493541   0.566069  32.670  < 2e-16 ***
X8            0.154738   0.048961   3.160  0.00165 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 1.93 on 679 degrees of freedom
Multiple R-squared: 0.9594,
Adjusted R-squared: 0.959
F-statistic:  2292 on 7 and 679 DF,  p-value: < 2.2e-16
```

Model Results:

Multiple R-squared	Adjusted R-squared
0.9594	0.959

Forward, Backward and Stepwise Selection

- All Forward, Backward and Stepwise Selections had same model
- Stepwise built model is chosen
- 6 Regressors left - Orientation is removed
- Same R-Squared values

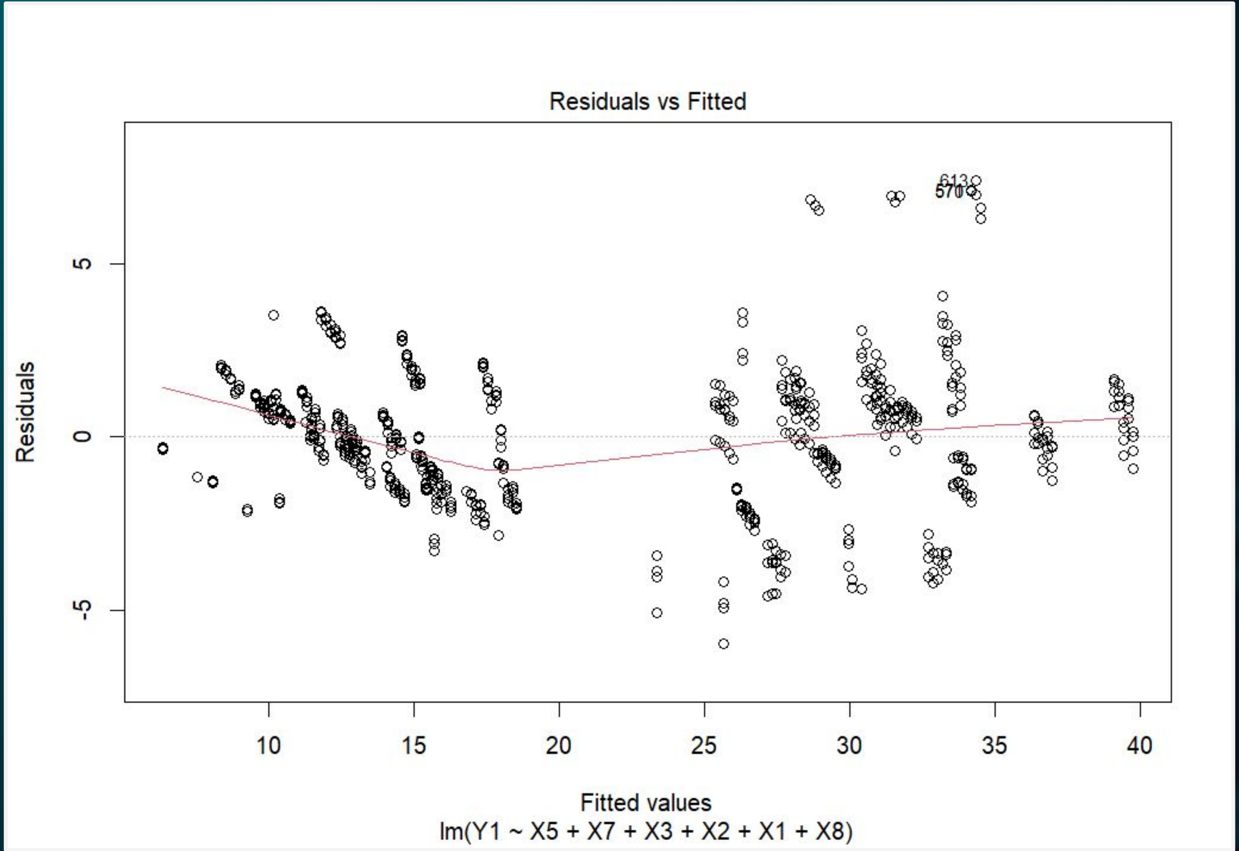


```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  98.179719  12.670032   7.749 3.38e-14 ***
x5            2.712018   0.243927  11.118  < 2e-16 ***
x7           18.493649   0.565652  32.694  < 2e-16 ***
x3            0.082863   0.004694  17.654  < 2e-16 ***
x2           -0.110701   0.011421  -9.693  < 2e-16 ***
x1          -61.712200   6.870539  -8.982  < 2e-16 ***
x8            0.154690   0.048920   3.162  0.00164 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

s: 1.929 on 680 degrees of freedom
Multiple R-squared: 0.9594,
Adjusted R-squared: 0.959 
F-statistic:  2678 on 6 and 680 DF,  p-value: < 2.2e-16
```

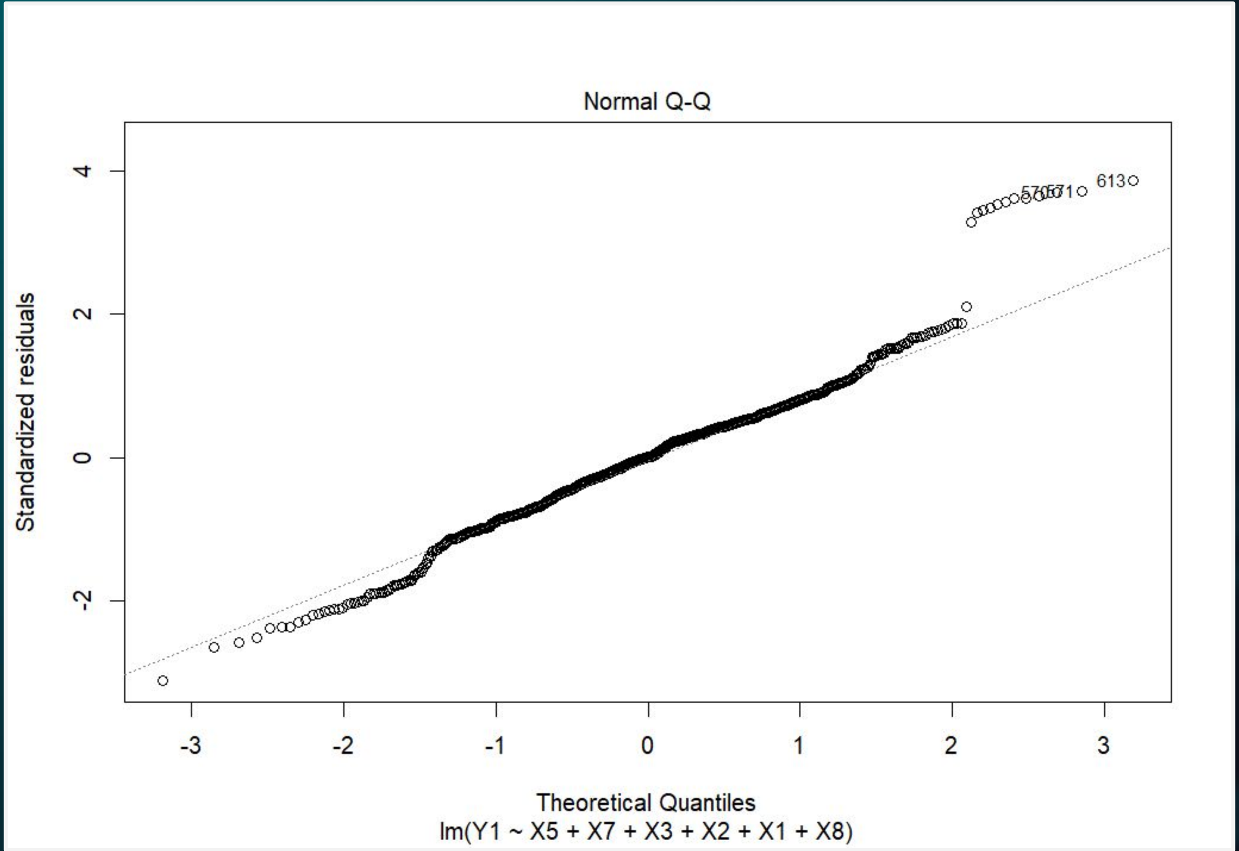
Model Results:

Multiple R-squared	Adjusted R-squared
0.9594	0.959

Residual Vs Fitted Plot



Updated QQ-Plot



Inference

Model	Multiple R-squared	Adjusted R-squared
SLR (Baseline)	0.7911	0.7908
MLR	0.9162	0.9154
MLR (No Outliers)	0.9594	0.959
MLR(Stepwise)	0.9594	0.959

Conclusion

Our analysis provides valuable insights into the energy performance of residential buildings and can help inform decisions related to building design and energy efficiency.

The MLR model we developed had a high adjusted R-squared value of 0.959, indicating that it was a good fit for the data.

1. **Does a multiple linear regression model require all regressors for the best model?**

The best model does not need all the regressors for better predictability. Roof Area is perfectly collinear, so is not required and Orientation does not affect the heating load at all.

2. **What are the most significant predictors that affect heating load?**

Our analysis revealed that the overall height (X5), glazing area (X7), and wall area (X3) were the most significant factors affecting the heating load of the buildings.

Thank You