Heating Load Modeling for Energy Efficient Buildings

DATA 603: Statistical Modeling with Data
Project Report

Prepared by:
Asiah Zibrila, Deysy Londono and Yao Zong

Contents

List of Figures

Introduction

As climate change due to excessive greenhouse gas emission has become a global issue, the world is exerting more and more efforts to save energy and to reduce CO2 emission. Such efforts are also being applied to buildings, particularly in reducing CO2 emission by lowering the energy consumption of buildings through energy performance improvement.

Keeping a comfortable temperature, accounts for a significant portion of the energy used in the average home, which in turn contributes towards global energy consumption. The global contribution from buildings towards energy consumption, both residential and commercial has steadily increased reaching figures between 20% and 40% in developed countries [1], which raises concerns mainly over exhaustion of energy resources and environmental impact. Also, Growth in population, increasing demand for building services and comfort levels, together with the rise in time spent inside buildings, assure the upward trend in energy demand will continue in the future.

For this reason, energy efficiency in buildings has become a key objective for policies and regulations, many governments impose legal constraints in residential building energy performance [2], making this topic of interest for modelling and analysis. In general, heating load (HL) and cooling load (CL) are two of the most important modes of the energy consumption in buildings, our goal is to model these variables as functions of the parameters of the structure, which include areas of the walls and roof, windows characteristics, and position in relation to the sunlight.

Objective

This report aims to explore how the Heating Load of a building is affected by attributes such as wall and roof areas, compactness of the building, glazing area and its distribution and orientation of the building. This variables will be used to fit a suitable model for prediction if possible.

Methodology

Dataset

The data was downloaded from the UC Irvine Machine Learning Repository in CSV format[3]. This website maintains data sets as a service to the machine learning community.

We performed this energy analysis using 12 different building shapes. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, among other parameters. The data set comprises 768 samples and 8 features, that we used with the goal of modelling the relationship between these attributes and the heating load.

Below you will find the complete list of variables used:

- Y1: Heating Load (KWh/m^2)
- Y2: Cooling Load (KWh/m^2)
- X1: Relative Compactness
- X2: Surface Area (m^2)
- X3: Wall Area (m^2)
- X4: Roof Area (m^2)
- X5: Overall Height (m)
- X6: Orientation (2, 3, 4, 5 stand for “North”, “East”, “South”, “West”)
- X7: Glazing Area (0.0, 0.1, 0.25, 0.4 stand for 0%, 10%, 25%, 40% of floor area)
- X8: Glazing Area Distribution (1,2, 3, 4,5 correspond to “Uniform”, “North”, “East”, “South”, “West”)

Modelling Plan

We will proceed by applying the methods taught in DATA603. The first step will be to check if there is a correlation between the two response variables. If there is a correlation, we will make a single model, otherwise a model for each response variable will be fit.

The second step will be to fit a linear regression model using all the predictors and test for multicollinearity. If we find that one or more of the independent variables are correlated, we can remove one by one the highly correlated variables while retaining most of the information. Once this process has been done, stepwise regression will be performed to find a model of main effects (additive model).

The third step will be using individual t-test to determine significant high-order terms and interactions. Once we have an estimated model with all the significant high-order and interaction terms, we will proceed with the diagnosis, that is, we will check if there are outliers and if the assumptions of linearity, homoscedasticity, normality, and independence are satisfied. We aim to do this diagnosis by checking the Cook's distance and leverage, using residual plots, histograms, Q-Q plots, Shapiro-Wilktest and Breusch-Pagan test.

If the results suggest the model does not satisfy these assumptions, we will make transformations in an attempt to improve the model.

Workload Distribution

It's agreed upon that each group member will put in equal effort to find and fit their own model to the dataset. After the most suitable selection procedure, final model and diagnostic is determined through group disicussion, the workload will be distributed in presentation slides preparation, presentation and report writing. The workload of the aforementioned tasks is distributed as follow:
Deysy - Introduction, Methodology, Conclusions and Recommendation;
Yao - Model Selection Procedure;
Asiah - Assumption Checking and Diagnostic;
All members are responsible for the administrative tasks for presentation and the project report, which includes proof-reading, suggesting changes/edits when they see fit.

Process and Results

Checking Correlation between dependent variables:

Reading the data set and checking for correlation between response variables:

```
data =read_excel("ENB2012_data.xlsx")
cor(data$Y1,data$Y2)
```

```
## [1] 0.9758617
```

This result shows that the Heating Load (Y1) and Cooling Load (Y2) are highly correlated, which means typically the ability for a building to remain warm or cool is the same. Therefore we are going to fit a single model using Y1.

Checking Multicollinearity

Before proceeding with selection of independent variables, we want to eliminate any potential multicollinearity, with which the coefficient estimates can swing dramatically and become very sensitive to small changes

in the model. VIF is used to determine if there's correlation between two independent variables. It can be computed using the formula:

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_j}}$$

```
model1 = lm(Y1~X1+X2+X3+X4+X5+factor(X6)+X7+factor(X8),data=data)
imcdiag(model1, method="VIF")
```

```
##
## Call:
## imcdiag(mod = model1, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                     VIF detection
## X1            105.5241         1
## X2                 Inf         1
## X3                 Inf         1
## X4                 Inf         1
## X5             31.2055         1
## factor(X6)3     1.5000         0
## factor(X6)4     1.5000         0
## factor(X6)5     1.5000         0
## X7              1.2604         0
## factor(X8)1     3.9271         0
## factor(X8)2     3.9271         0
## factor(X8)3     3.9271         0
## factor(X8)4     3.9271         0
## factor(X8)5     3.9271         0
##
## Multicollinearity may be due to X1 X2 X3 X4 X5 regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

The result shows an extremely large VIF for X2, X3 and X4. We will start by dropping X2:

```
model2 = lm(Y1~X1+X3+X4+X5+factor(X6)+X7+factor(X8),data=data)
imcdiag(model2, method="VIF")
```

```
##
## Call:
## imcdiag(mod = model2, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                     VIF detection
## X1            105.5241         1
## X3             27.6627         1
```

```
## X4          211.9383         1
## X5           31.2055         1
## factor(X6)3   1.5000         0
## factor(X6)4   1.5000         0
## factor(X6)5   1.5000         0
## X7            1.2604         0
## factor(X8)1   3.9271         0
## factor(X8)2   3.9271         0
## factor(X8)3   3.9271         0
## factor(X8)4   3.9271         0
## factor(X8)5   3.9271         0
##
## Multicollinearity may be due to X1 X3 X4 X5 regressors
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

The second variable with a large VIF is X4, After dropping X4, we test the model again:

```
model3 = lm(Y1~X1+X3+X5+factor(X6)+X7+factor(X8),data=data)
imcdiag(model3, method="VIF")
```

```
##
## Call:
## imcdiag(mod = model3, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                VIF detection
## X1          9.2503         0
## X3          3.1619         0
## X5          9.6261         0
## factor(X6)3 1.5000         0
## factor(X6)4 1.5000         0
## factor(X6)5 1.5000         0
## X7          1.2604         0
## factor(X8)1 3.9271         0
## factor(X8)2 3.9271         0
## factor(X8)3 3.9271         0
## factor(X8)4 3.9271         0
## factor(X8)5 3.9271         0
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

After this two variables are dropped, no predictor has a VIF higher than 10. We will use the remaining variables to run a stepwise regression to fit a first-order model.

Stepwise Model Selection

We use the Stepwise Regression Procedure to determine the first order model. At $\alpha = 0.05$, P-value for entering the model is set to 0.1 so that estimators that are barely significant wouldn't be eliminated yet, in case potential significant estimating power can be found in high order terms and interaction terms.

```
ols_step_both_p(model3, pent=0.1,prem=0.3, details=FALSE)
```

```
##
##                                 Stepwise Selection Summary
## ----------------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step     Variable     Removed     R-Square    R-Square      C(p)          AIC         RMSE
## ----------------------------------------------------------------------------------------------
##    1        X5        addition       0.791       0.791    1237.6840    4532.4914    4.6149
##    2        X7        addition       0.864       0.864     542.0180    4205.3741    3.7273
##    3        X3        addition       0.910       0.910     103.5860    3890.8720    3.0352
##    4    factor(X8)    addition       0.919       0.918      21.2450    3821.9456    2.8927
##    5        X1        addition       0.921       0.920      -0.7950    3799.8793    2.8496
## ----------------------------------------------------------------------------------------------
```

```
model4 = lm(Y1~X1+X3+X5+X7+factor(X8),data=data)
summary(model4)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X5 + X7 + factor(X8), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3068 -1.5588  0.0232  1.4189  7.3450
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -15.183790   2.589482  -5.864 6.76e-09 ***
## X1          -14.532402   2.958472  -4.912 1.10e-06 ***
## X3            0.034976   0.004194   8.340 3.49e-16 ***
## X5            5.606753   0.182300  30.756  < 2e-16 ***
## X7           16.848333   0.867099  19.431  < 2e-16 ***
## factor(X8)1   4.527653   0.522063   8.673  < 2e-16 ***
## factor(X8)2   4.435986   0.522063   8.497  < 2e-16 ***
## factor(X8)3   4.183000   0.522063   8.012 4.24e-15 ***
## factor(X8)4   4.388208   0.522063   8.406  < 2e-16 ***
## factor(X8)5   4.182444   0.522063   8.011 4.28e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.85 on 758 degrees of freedom
## Multiple R-squared:  0.9212, Adjusted R-squared:  0.9202
## F-statistic: 984.3 on 9 and 758 DF,  p-value: < 2.2e-16
```

The result suggests that the remaining 5 predictors (X1,X3,X5,X7 and X8) are significant. At this point, the model has 0.9202 adjusted R-squared and 2.85 RMSE.

Higher Order Model

Now we will determine if high order terms need to be added to the model by examining the following plots:

```
red_data = data[,c(1,3,5,7,8,9)]
ggpairs(red_data,lower = list(continuous = "smooth_loess", combo = "facethist",
                              discrete = "facetbar", na = "na"))
```

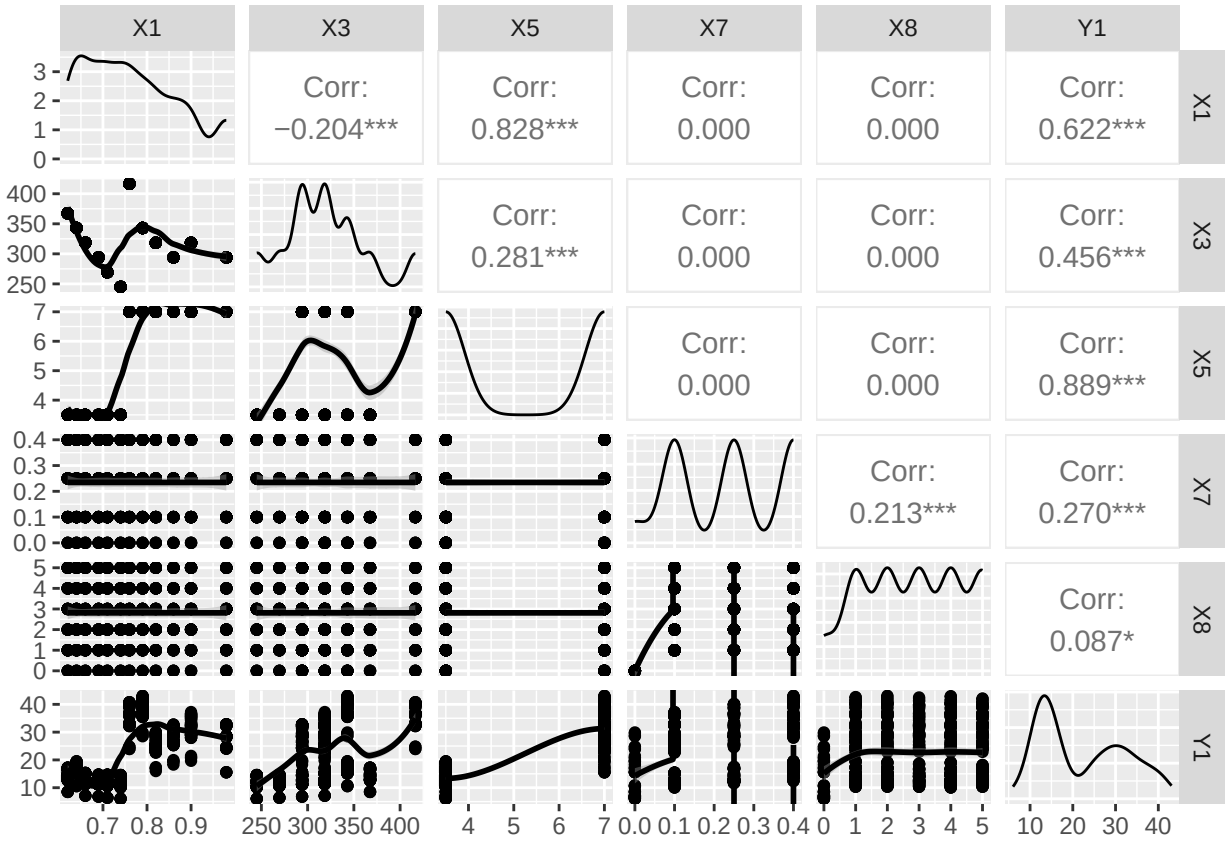


Figure 1: Pairwise Correlation Plots between Independent Variables

From the pairwise correlation plot, a non-linear pattern can be observed between (X1,Y1) and (X3,Y1). Thus, higher order terms of X1, X3 will be added to the first order model until they become insignificant.

```
high4 = lm(Y1~X1+X3+X5+X7+factor(X8)+I(X1^2)+I(X1^3)+
             I(X3^2)+I(X3^3)+I(X1^4)+I(X3^4)+I(X1^5)+I(X3^5),data=data)
summary(high4)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X5 + X7 + factor(X8) + I(X1^2) +
##     I(X1^3) + I(X3^2) + I(X3^3) + I(X1^4) + I(X3^4) + I(X1^5) +
##     I(X3^5), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.0754 -0.6167 -0.1070  0.6063  3.8034
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.141e+05  1.363e+04 -15.702   <2e-16 ***
## X1           1.524e+06  9.481e+04  16.077   <2e-16 ***
## X3          -4.994e+02  3.143e+01 -15.890   <2e-16 ***
## X5           1.540e+00  1.324e+00   1.163    0.245
## X7           1.685e+01  3.113e-01  54.118   <2e-16 ***
## factor(X8)1  4.528e+00  1.874e-01  24.155   <2e-16 ***
## factor(X8)2  4.436e+00  1.874e-01  23.666   <2e-16 ***
## factor(X8)3  4.183e+00  1.874e-01  22.316   <2e-16 ***
## factor(X8)4  4.388e+00  1.874e-01  23.411   <2e-16 ***
## factor(X8)5  4.182e+00  1.874e-01  22.313   <2e-16 ***
## I(X1^2)     -3.802e+06  2.438e+05 -15.598   <2e-16 ***
## I(X1^3)      4.716e+06  3.116e+05  15.135   <2e-16 ***
## I(X3^2)      3.419e+00  2.063e-01  16.571   <2e-16 ***
## I(X3^3)     -1.162e-02  6.730e-04 -17.267   <2e-16 ***
## I(X1^4)     -2.908e+06  1.979e+05 -14.694   <2e-16 ***
## I(X3^4)      1.958e-05  1.090e-06  17.956   <2e-16 ***
## I(X1^5)      7.134e+05  4.996e+04  14.280   <2e-16 ***
## I(X3^5)     -1.306e-08  7.014e-10 -18.617   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.023 on 750 degrees of freedom
## Multiple R-squared:  0.9899, Adjusted R-squared:  0.9897
## F-statistic:  4344 on 17 and 750 DF,  p-value: < 2.2e-16
```

After reaching the fifth order, the estimator X5 starts to lose it's significance. Thus the model will include higher order terms of X1 and X3 up to the forth order.

```
summary(high3)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X5 + X7 + factor(X8) + I(X1^2) +
##     I(X1^3) + I(X3^2) + I(X3^3) + I(X1^4) + I(X3^4), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1708 -0.7470  0.0030  0.7912  4.0091
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.374e+04  7.989e+02  -42.23   <2e-16 ***
## X1           1.447e+05  3.590e+03   40.32   <2e-16 ***
## X3           7.225e+01  2.295e+00   31.48   <2e-16 ***
## X5           1.697e+01  4.816e-01   35.23   <2e-16 ***
## X7           1.685e+01  3.763e-01   44.77   <2e-16 ***
## factor(X8)1  4.528e+00  2.266e-01   19.98   <2e-16 ***
## factor(X8)2  4.436e+00  2.266e-01   19.58   <2e-16 ***
## factor(X8)3  4.183e+00  2.266e-01   18.46   <2e-16 ***
```

```
## factor(X8)4  4.388e+00  2.266e-01   19.37   <2e-16 ***
## factor(X8)5  4.182e+00  2.266e-01   18.46   <2e-16 ***
## I(X1^2)     -2.727e+05  6.824e+03  -39.96   <2e-16 ***
## I(X1^3)      2.258e+05  5.692e+03   39.66   <2e-16 ***
## I(X3^2)     -3.672e-01  1.108e-02  -33.15   <2e-16 ***
## I(X3^3)      8.109e-04  2.344e-05   34.59   <2e-16 ***
## I(X1^4)     -6.942e+04  1.761e+03  -39.42   <2e-16 ***
## I(X3^4)     -6.580e-07  1.836e-08  -35.84   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.237 on 752 degrees of freedom
## Multiple R-squared:  0.9853, Adjusted R-squared:  0.985
## F-statistic:  3354 on 15 and 752 DF,  p-value: < 2.2e-16
```

From the result above, the higher order model has drastically improved the adjusted R-squared from 0.9202 to 0.985.

Interaction Model

Interaction terms will be determined by individual t-tests. We will test all possible interaction terms and remove the insignificant terms.

```
##
## Call:
## lm(formula = Y1 ~ (X1 + X3 + X5 + X7 + factor(X8))^2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9072 -1.0133 -0.0366  0.7431  6.7424
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.262e+02  9.722e+01  -2.327  0.02026 *
## X1              1.797e+02  9.565e+01   1.878  0.06074 .
## X3              4.528e-02  8.874e-02   0.510  0.61008
## X5              5.030e+01  1.524e+01   3.301  0.00101 **
## X7             -5.459e+01  1.962e+01  -2.783  0.00553 **
## factor(X8)1    -6.515e+00  1.178e+01  -0.553  0.58039
## factor(X8)2    -7.146e+00  1.178e+01  -0.607  0.54430
## factor(X8)3    -6.397e+00  1.178e+01  -0.543  0.58729
## factor(X8)4    -7.515e+00  1.178e+01  -0.638  0.52372
## factor(X8)5    -6.151e+00  1.178e+01  -0.522  0.60173
## X1:X3           3.971e-01  8.421e-02   4.715 2.89e-06 ***
## X1:X5          -4.695e+01  1.524e+01  -3.080  0.00215 **
## X1:X7           6.505e+01  2.261e+01   2.877  0.00413 **
## X1:factor(X8)1  8.047e+00  1.361e+01   0.591  0.55464
## X1:factor(X8)2  8.591e+00  1.361e+01   0.631  0.52818
## X1:factor(X8)3  8.408e+00  1.361e+01   0.618  0.53699
## X1:factor(X8)4  9.016e+00  1.361e+01   0.662  0.50800
## X1:factor(X8)5  8.180e+00  1.361e+01   0.601  0.54809
## X3:X5          -4.491e-02  1.629e-02  -2.757  0.00597 **
## X3:X7           8.088e-02  3.205e-02   2.523  0.01183 *
```

```
## X3:factor(X8)1  1.169e-02  1.930e-02   0.606  0.54497
## X3:factor(X8)2  1.272e-02  1.930e-02   0.659  0.50991
## X3:factor(X8)3  1.062e-02  1.930e-02   0.551  0.58211
## X3:factor(X8)4  1.282e-02  1.930e-02   0.665  0.50652
## X3:factor(X8)5  9.812e-03  1.930e-02   0.508  0.61128
## X5:X7          -1.021e+00  1.393e+00  -0.733  0.46400
## X5:factor(X8)1  1.892e-01  8.388e-01   0.226  0.82162
## X5:factor(X8)2  1.507e-01  8.388e-01   0.180  0.85750
## X5:factor(X8)3  1.848e-01  8.388e-01   0.220  0.82571
## X5:factor(X8)4  1.425e-01  8.388e-01   0.170  0.86520
## X5:factor(X8)5  2.458e-01  8.388e-01   0.293  0.76956
## X7:factor(X8)1  2.048e+00  2.485e+00   0.824  0.41015
## X7:factor(X8)2  2.029e+00  2.485e+00   0.817  0.41445
## X7:factor(X8)3  5.368e-01  2.485e+00   0.216  0.82904
## X7:factor(X8)4  2.058e+00  2.485e+00   0.828  0.40777
## X7:factor(X8)5         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.583 on 733 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9345
## F-statistic: 322.8 on 34 and 733 DF,  p-value: < 2.2e-16
```

```
int2 = lm(Y1~X1+X3+X5+X7+factor(X8)+X1:X3+X1:X5+X1:X7+X3:X5+X3:X7,data=data)
summary(int2)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X5 + X7 + factor(X8) + X1:X3 + X1:X5 +
##     X1:X7 + X3:X5 + X3:X7, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7845 -1.0662 -0.0875  0.7896  6.8684
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -236.96103   96.19470  -2.463 0.013987 *
## X1           189.06644   94.48693   2.001 0.045753 *
## X3             0.05769    0.08703   0.663 0.507595
## X5            50.23088   15.12360   3.321 0.000939 ***
## X7           -51.61230    8.04777  -6.413 2.51e-10 ***
## factor(X8)1    4.52765    0.47007   9.632  < 2e-16 ***
## factor(X8)2    4.43599    0.47007   9.437  < 2e-16 ***
## factor(X8)3    4.18300    0.47007   8.899  < 2e-16 ***
## factor(X8)4    4.38821    0.47007   9.335  < 2e-16 ***
## factor(X8)5    4.18244    0.47007   8.897  < 2e-16 ***
## X1:X3          0.39709    0.08367   4.746 2.48e-06 ***
## X1:X5        -46.94903   15.14544  -3.100 0.002008 **
## X1:X7         58.73302    6.71975   8.740  < 2e-16 ***
## X3:X5         -0.04490    0.01618  -2.775 0.005653 **
## X3:X7          0.07403    0.01629   4.544 6.43e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.566 on 753 degrees of freedom
## Multiple R-squared:  0.9365, Adjusted R-squared:  0.9353
## F-statistic: 793.5 on 14 and 753 DF,  p-value: < 2.2e-16
```

By keeping X1:x3, X1:X5, X1:X7, X3:X5, X3:X7 in the interaction model, the adjusted R-squared has been increase to 0.9353 compared to 0.9202 in the first order model.

Final Model

Interaction terms and higher order terms will be combined to form a final model. The interaction term X3:X5 is removed from the model since it's not significant anymore after combining the interaction terms and higher order terms.

```
model = lm(Y1~X1 + X3 + X5 + X7 + factor(X8)+
             X1:X3 + X1:X5 + X1:X7 + X3:X7 +
             I(X1^2) + I(X1^3) + I(X3^2) + I(X3^3) + I(X1^4) + I(X3^4),
           data=data)
summary(model)
```

```
##
## Call:
## lm(formula = Y1 ~ X1 + X3 + X5 + X7 + factor(X8) + X1:X3 + X1:X5 +
##     X1:X7 + X3:X7 + I(X1^2) + I(X1^3) + I(X3^2) + I(X3^3) + I(X1^4) +
##     I(X3^4), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07036 -0.30361 -0.02345  0.35724  2.88851
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.090e+04  4.619e+02 -88.547  < 2e-16 ***
## X1           1.752e+05  2.346e+03  74.672  < 2e-16 ***
## X3           7.682e+01  1.129e+00  68.065  < 2e-16 ***
## X5           8.193e+01  2.123e+01   3.859 0.000124 ***
## X7          -5.161e+01  1.738e+00 -29.692  < 2e-16 ***
## factor(X8)1  4.528e+00  1.015e-01  44.593  < 2e-16 ***
## factor(X8)2  4.436e+00  1.015e-01  43.690  < 2e-16 ***
## factor(X8)3  4.183e+00  1.015e-01  41.198  < 2e-16 ***
## factor(X8)4  4.388e+00  1.015e-01  43.220  < 2e-16 ***
## factor(X8)5  4.182e+00  1.015e-01  41.193  < 2e-16 ***
## I(X1^2)     -3.273e+05  5.090e+03 -64.311  < 2e-16 ***
## I(X1^3)      2.714e+05  4.910e+03  55.276  < 2e-16 ***
## I(X3^2)     -3.732e-01  6.017e-03 -62.034  < 2e-16 ***
## I(X3^3)      8.161e-04  1.432e-05  56.986  < 2e-16 ***
## I(X1^4)     -8.366e+04  1.694e+03 -49.380  < 2e-16 ***
## I(X3^4)     -6.593e-07  1.238e-08 -53.244  < 2e-16 ***
## X1:X3       -2.550e+00  8.220e-02 -31.025  < 2e-16 ***
## X1:X5       -9.856e+01  2.782e+01  -3.543 0.000420 ***
## X1:X7        5.873e+01  1.451e+00  40.466  < 2e-16 ***
## X3:X7        7.403e-02  3.519e-03  21.037  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5542 on 748 degrees of freedom
## Multiple R-squared:  0.9971, Adjusted R-squared:  0.997
## F-statistic: 1.334e+04 on 19 and 748 DF,  p-value: < 2.2e-16
```

At this point we have the following model:

$$\widehat{Y_1} = -4090 + 17520X1 + 76.82X3 + 81.93X5 - 51.61X7 + 4.52X8_1$$
$$+4.436X8_2 + 4.18X8_3 + 4.38X8_4 + 4.182X8_5 - 32730X1^2 + 27140X1^3$$
$$-0.373X_2^2 + 0.0081X3^3 - 83660X1^4 - 0.000006X3^4$$
$$-2.5(X1 \times X3) - 98.56(X1 \times X5) + 58.73(X1 \times X7) + 740.3(X3 \times X7)$$

Checking Model Assumptions

At this point we perform the six basic assumptions checking for MLR to ensure that our model is reliable and trustworthy. The basic assumptions are:

Linearity Assumption

The linearity assumption assumes that there is a linear relationship between the response variable and the predictors in our model.To verify this, We use residual plots to identify non-linearity in the data. From the residual plot below, we can see that there is no discernible pattern in the residuals. Hence, we conclude that the residuals are linear. It means that linearity assumption is met for our final model.

Independence Assumption

Next we check for correlation among the residuals. Refer to the residual versus fitted plot above to confirm the indpendence. We can see that the residuals are not clumped and there are no visible trends, we conclude that the errors are independent. Most importantly our data set is not time related, hence we do not expect errors to be correlated.

Equal Variance Check

It is very important for the error terms in a multiple linear regression to have a constant variance. Next we test for equal variance by using the spread of residual over the predicted values, the scale location plot and the Breusch-Pagan test. From the plots shown below, the residuals appear to show a cone shaped pattern which indicates that the variances of the error terms increases with the value of their response. Therefore, this is an indication that the data Has nonconstant variance or heteroscedasticity.

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```

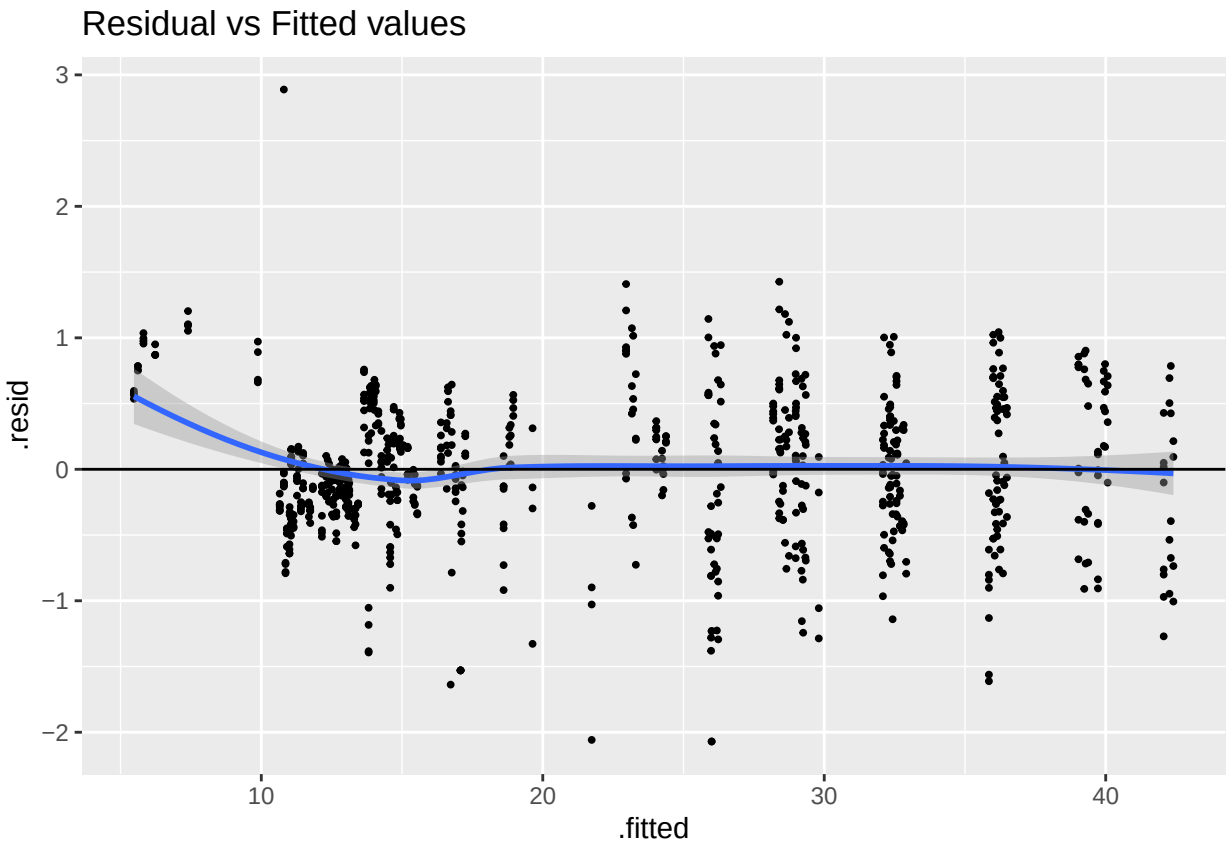


Figure 2: Residual vs Fitted Plot

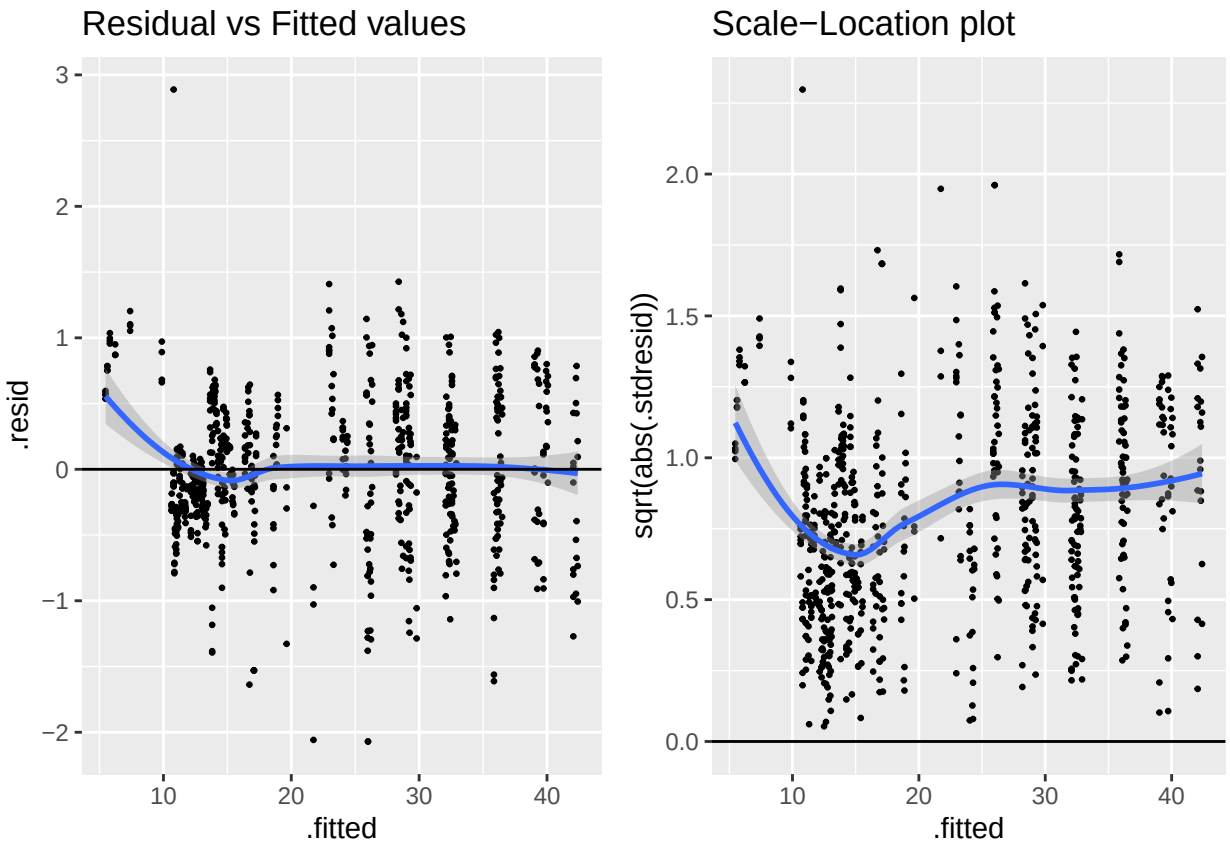


Figure 3: Residual vs Fitted Plot, Scale-Location Plot

To statistically confirm this assumption, we will perform the Breusch-Pagan test of equal variance. The hypothesis is formulated as:

$$H_0 : \text{heteroscedasticity is not present (homoscedasticity)}$$
$$H_a : \text{heteroscedasticity is present}$$

From the results of the Breusch-Pagan test (BP = 169.24, p-value < 2.2e-16), we would reject the null hypothesis and conclude that heteroscedasticity is present.

Normality Assumption

Here we employ a histogram and Q-Q plot to visually access the normality of the residuals. The histogram plot looks quite normal according the binwidth used, but however this is not reliable. The Q-Qplot on the other hand has as many points fitted on the diagonal but however the residuals are heavily deviated from the diagonal in the lower tail, this is an indication of non normality. To confirm this we proceed to statistivcally test normality by applying the Shapiro-Wilk test. We formulate the hypothesis for this test as:

$$H_0 : \text{the sample data are significantly normally distributed}$$
$$H_A : \text{the sample data are not significantly normally distributed}$$

The Shapiro-Wilk test yields a value of W = 0.98323 and a corresponding p-value = 1.061e-07. We reject the null hypothesis and conclude that there is enough evidence to support that the residuals are not normally distributed.Therfore our model fails the normality test.

Multicollinearity

The reults of the multicollinearity checked at the first stage of our model selection indicated that the variables in the final model have less to moderate VIF values. Therefore multicollinearity is passed. Results of VIF is shown below:

```
imcdiag(model3, method="VIF")
```

```
##
## Call:
## imcdiag(mod = model3, method = "VIF")
##
##
##  VIF Multicollinearity Diagnostics
##
##                 VIF detection
## X1           9.2503         0
## X3           3.1619         0
## X5           9.6261         0
## factor(X6)3 1.5000         0
## factor(X6)4 1.5000         0
## factor(X6)5 1.5000         0
## X7           1.2604         0
## factor(X8)1 3.9271         0
## factor(X8)2 3.9271         0
## factor(X8)3 3.9271         0
## factor(X8)4 3.9271         0
## factor(X8)5 3.9271         0
```

```
##
## NOTE:  VIF Method Failed to detect multicollinearity
##
##
## 0 --> COLLINEARITY is not detected by the test
##
## ===================================
```

Outliers

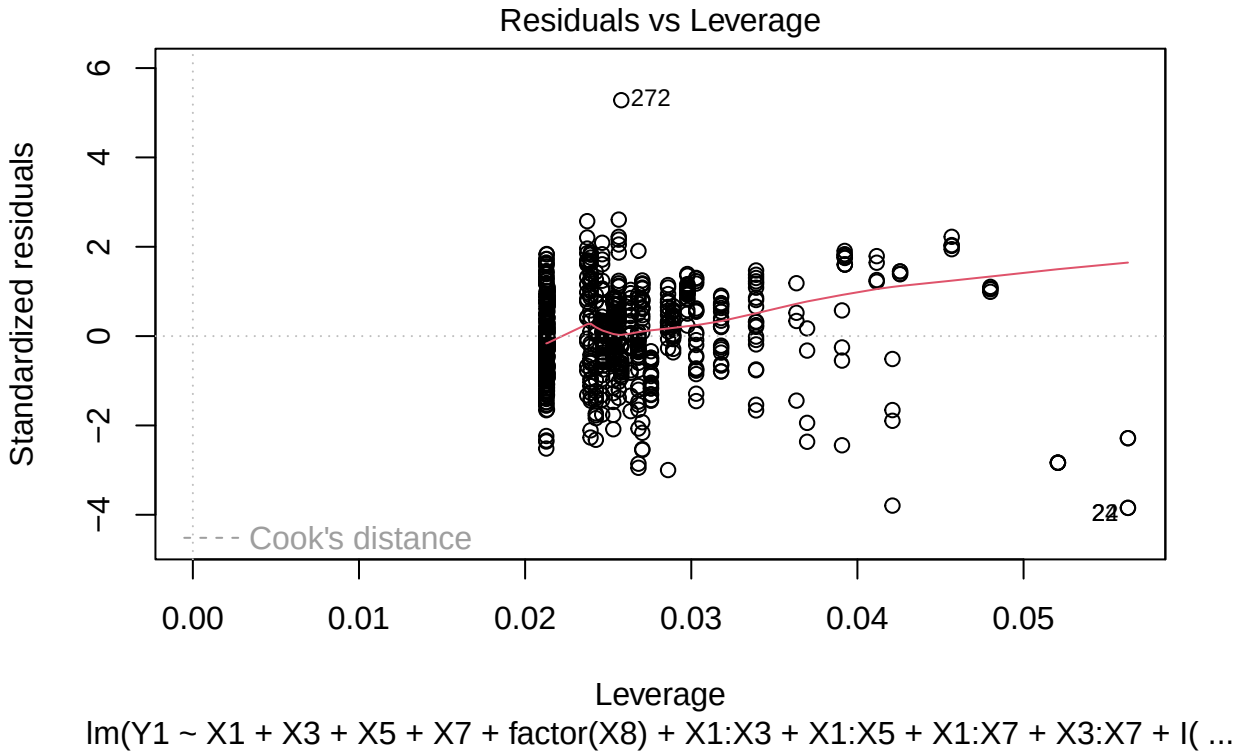


Figure 4: Standardized Residuals vs Leverage Plot

Below is a graph of the Cook's distance plot, overall we can see that point 8 and 272 have high influence but however the value is less than 1 and we therfore consider them as not influencial or minimal.

With reference to the above model diagnostics, our results concluded that our model meets the linearity assumption, the independence assumption and the multicolinearity assumption as well as no outlears detected. But however the model failed the equal variance and normality assumption.In attempt to rectify this failure, we proceed to perform the Box-Cox transformation in the next section.

BoxCox Transformation

The results below show the BoxCox transformation applied to our model. One can see in the graph that the best lambda is between a range of 0.7 and 0.9. The best λ computed is 0.7878. This value is applied to transform the response variable Y1 and the new model is formed.

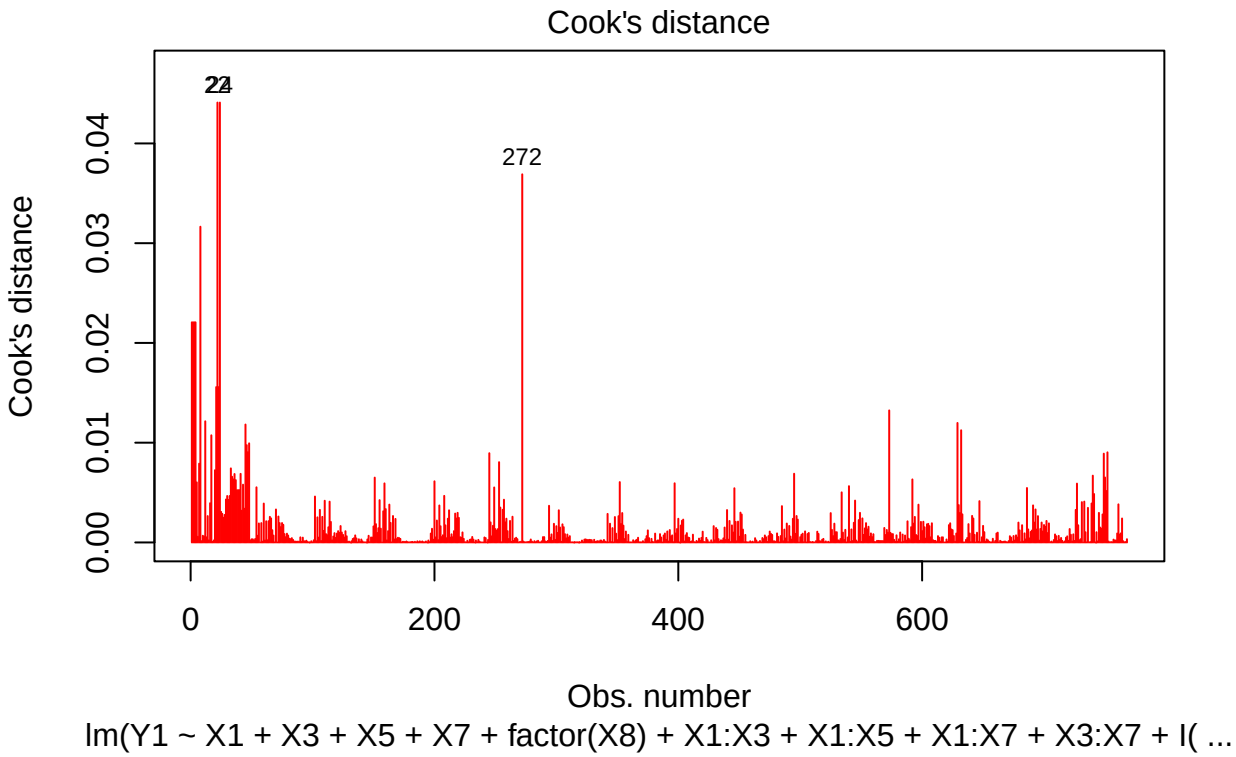


Figure 5: Cook's Distance Plot

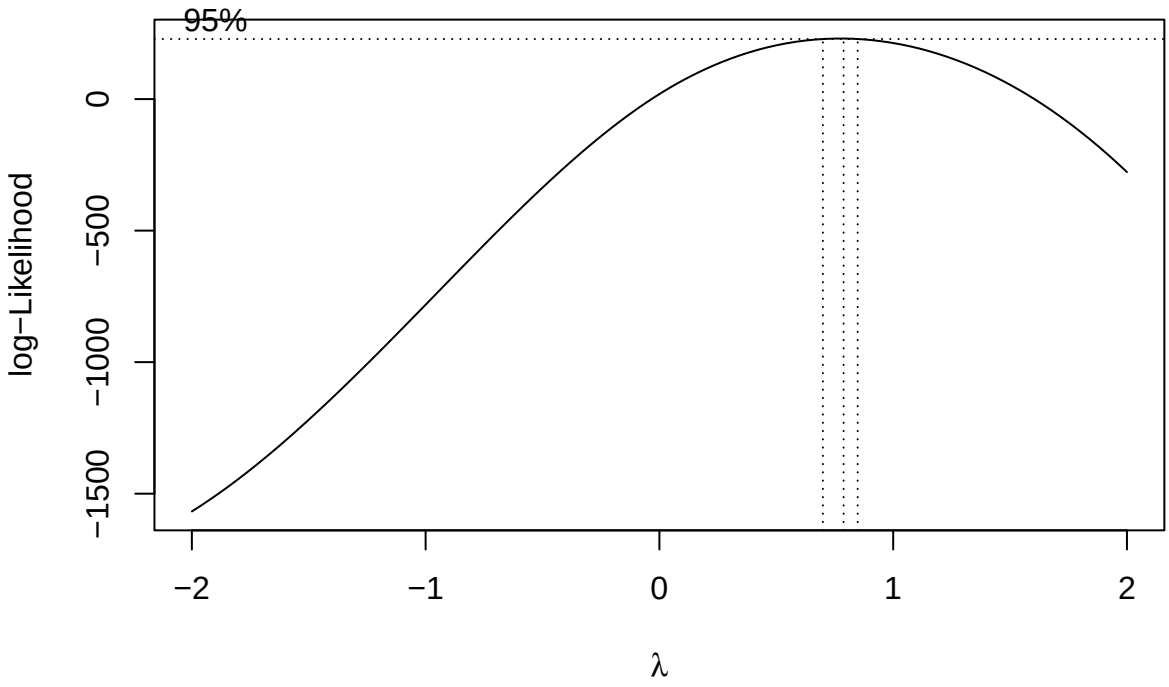


Figure 6: Box-Cox plot

```
##
## Call:
## lm(formula = (((Y1^0.7878) - 1)/0.7878) ~ X1 + X3 + X5 + X7 +
##     factor(X8) + X1:X3 + X1:X5 + X1:X7 + X3:X7 + I(X1^2) + I(X1^3) +
##     I(X3^2) + I(X3^3) + I(X1^4) + I(X3^4), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.92754 -0.14168 -0.00488  0.18178  1.75502
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.954e+04  2.277e+02 -85.818   <2e-16 ***
## X1           8.255e+04  1.157e+03  71.368   <2e-16 ***
## X3           3.795e+01  5.565e-01  68.188   <2e-16 ***
## X5          -1.697e+01  1.047e+01  -1.621   0.1055
## X7          -1.662e+01  8.571e-01 -19.392   <2e-16 ***
## factor(X8)1  2.544e+00  5.006e-02  50.809   <2e-16 ***
## factor(X8)2  2.495e+00  5.006e-02  49.838   <2e-16 ***
## factor(X8)3  2.364e+00  5.006e-02  47.215   <2e-16 ***
## factor(X8)4  2.469e+00  5.006e-02  49.319   <2e-16 ***
## factor(X8)5  2.360e+00  5.006e-02  47.148   <2e-16 ***
## I(X1^2)     -1.516e+05  2.510e+03 -60.389   <2e-16 ***
## I(X1^3)      1.233e+05  2.421e+03  50.944   <2e-16 ***
## I(X3^2)     -1.878e-01  2.967e-03 -63.292   <2e-16 ***
## I(X3^3)      4.173e-04  7.061e-06  59.097   <2e-16 ***
## I(X1^4)     -3.745e+04  8.354e+02 -44.830   <2e-16 ***
## I(X3^4)     -3.413e-07  6.106e-09 -55.892   <2e-16 ***
## X1:X3       -1.366e+00  4.053e-02 -33.705   <2e-16 ***
## X1:X5        2.675e+01  1.372e+01   1.950   0.0515 .
## X1:X7        2.320e+01  7.157e-01  32.420   <2e-16 ***
## X3:X7        2.371e-02  1.735e-03  13.662   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2733 on 748 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9972
## F-statistic: 1.463e+04 on 19 and 748 DF,  p-value: < 2.2e-16
```

The final model equation after the Box-Cox transformation is:

$$\widehat{Y_1} = -19540 + 82550X1 + 37.95X3 - 16.97X5 - 16.62X7 + 2.544X8_1$$
$$+2.495X8_2 + 2.364X8_3 + 2.469X8_4 + 2.36X8_5 - 151600X1^2 + 123300X1^3$$
$$-0.1878X_3^2 + 0.0004X3^3 - 37450X1^4 - 0.0000003X3^4$$
$$-1.366(X1 \times X3) - 26.75(X1 \times X5) + 23.2(X1 \times X7) + 0.02371(X3 \times X7)$$

The above transformed model yielded a RMSE value of 0.2733 and R^2_{adj} of 0.9972 approximately 99.72% as compared to the initial final model which hada RMSE of 0.5542 and R^2_{adj} of 0.997,approximately 99.7%. We can observe that the adjusted r squared of both models are similar but the transformed model has an improved RMSE value.

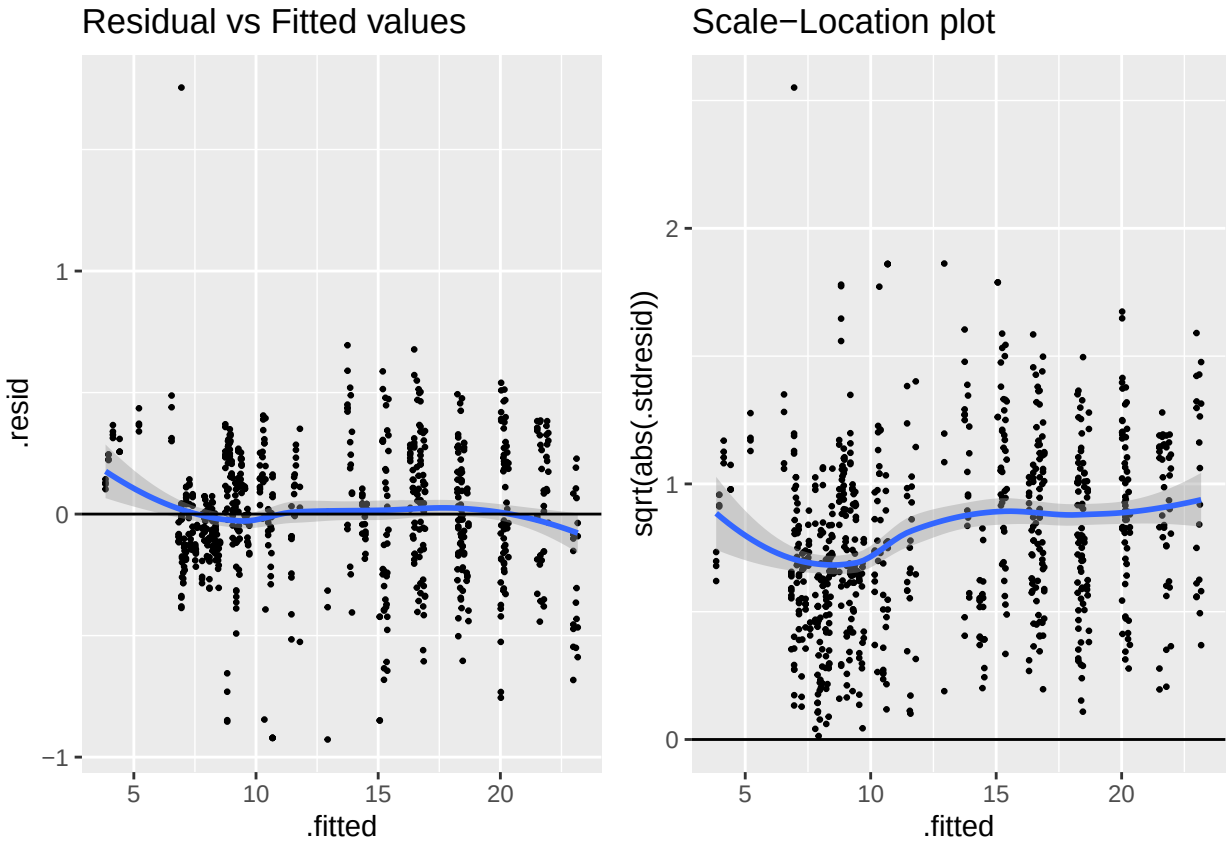
Having found the new model we then perform the diagnostics checks again to verify if there are improvements in th new model.

Equal Variance check for the Box-Cox Model:

From Figure 6, we are still able to see a cone-shaped patter from fiited vs residual plot, and the blue line in scale-location plot is not perfectly horizontal, which indicates that heteroscedasticity is still present.

```
library(gridExtra)
e=ggplot(bcmodel, aes(x=.fitted, y=.resid)) +
geom_point(size=0.5) + geom_smooth()+
geom_hline(yintercept = 0)+ggtitle("Residual vs Fitted values")

#a scale location plot
f=ggplot(bcmodel, aes(x=.fitted, y=sqrt(abs(.stdresid)))) +
geom_point(size=0.5) + geom_hline(yintercept = 0) +
geom_smooth()+ ggtitle("Scale-Location plot")
#grid
grid.arrange(e,f,ncol=2)
```



From the Breusch-Pagan test, with p-value still smaller than $\alpha = 0.05$, we can conclude that the equal variance assumption still doesn't hold after transformation.

```
bptest(bcmodel)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  bcmodel
## BP = 97.787, df = 19, p-value = 1.345e-12
```

Normality check for the Box-Cox Model:

From Figure 7, we can see the histogram of residual looks normal, but points are not perfectly aligned along the reference line in the Q-Q plot.

```
c = ggplot(bcmodel, aes(x=.resid)) + geom_histogram(binwidth = 0.1)
d = ggplot(bcmodel,aes(sample=.resid)) + stat_qq() + stat_qq_line()
grid.arrange(c,d)
```

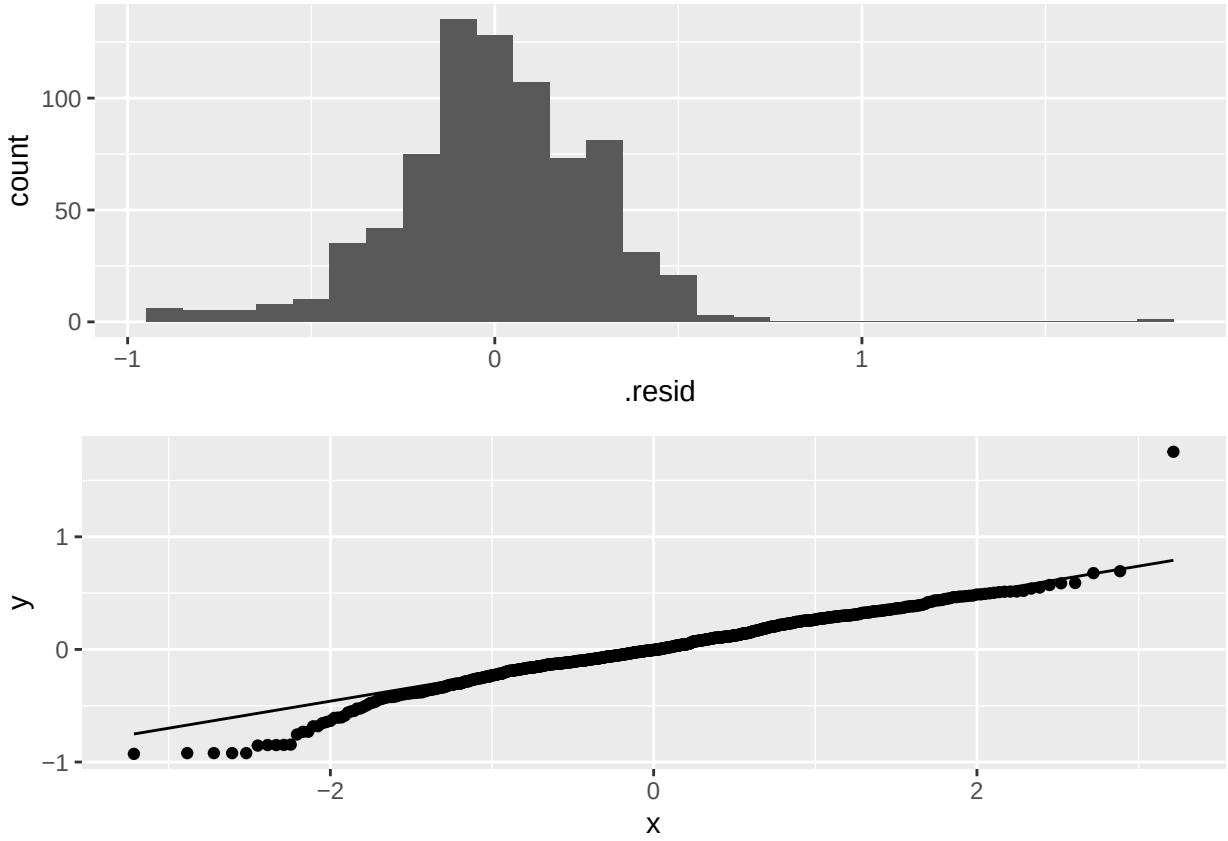


Figure 7: Histogram and Q-Q Plot of Residuals for Box-Cox Model

From the Shapiro-Wilk test, with p-value still smaller than $\alpha = 0.05$, we can conclude that the normality assumption still doesn't hold after transformation.

```
shapiro.test(residuals(bcmodel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(bcmodel)
## W = 0.9687, p-value = 9.264e-12
```

From the above checks, the Box-Cox model's linearity and independence assumptions hold. From the BP test results the p-value is = 1.345e-12 and the Shapiro test gives a p-value = 9.264e-12. In both cases the p-values are still less than 0.05, indicating that the residuals are not normally distributed and heteroscedasticity is still present. This implies that the Box-Cox transformation did not improve the model.

Conclusions and Recommendation

Heating Load and Cooling load, our two response variables are highly correlated. Consequently, finding a model that fits one of them will lead to the model for the second. In terms of the predictors, the compactness of a building, the wall area, the overall high, the percentage of windows (glazing area), and the distribution of the windows have a higher impact on the Heating Load (response variable) than the roof area and orientation of the building.

Even though the glazing area and its distribution are slightly correlated to each other, they are uncorrelated with the other attributes of the building, which made them good predictors for the model. In regards to the roofing area, it has some impact in the Heating Load. But since it is highly correlated with other features, it is safe to leave this term out of the model.

It could be observed that the first order model wouldn't be the best model for prediction and that the impact of the predictors wall area(X3) and Compactness(X1) was better described with high order terms. With respect to the interactions, it was observed that the interactions of the compactness of the building with some other variables like the glazing area and overall height have a moderate effect on the heating load.

The main challenge we encountered was that after finding a model using the methods learnt in this course, the model did not meet the assumptions for linear regression. Even after transformations were performed, the model still presented normality and heteroscedasticity issues. Therefore, the forecast accuracy may be distorted when using this model.

There are several ways in which this data can be modeled for predictive purposes. Some of them involve exploring alternative regression models such as robust regression or PLS, which might improve the model and potentially allow for a more practical prediction of the Heating Load.

References

[1] L. Pérez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption information, Energy Build. 40 (3) (2008) 394–398.

[2] K. Kavyalola, Robust modeling of heating and cooling loads using partial least squares towards efficient residential building design. Journal of Building Engineering 18 (2018) 467–475.

[3] UC Irvine Machine learning repository. Energy Efficiency Dataset. https://archive.ics.uci.edu/dataset/242/energy+efficiency

[4] A. Tsana's, A. Xiara, Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools, Energy Build. 49 (2012) 560–567.