Multiple Linear Regression, its Statistical Analysis and Application in Energy Efficiency

Fahad Mostafa

Texas Tech University, Lubbock, TX 79409

Abstract

In this project, we use a statistical multiple regression to study the impact of eight various predictors (relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area, glazing area distribution) to estimate the cooling load energy efficiency of residential buildings. We try to analyze and visualize the effect of each predictor with each of the response variable using different classical statistical analysis tools used in describing linear models, in such a way so that we can find out the most strongly related predictor variables. Before starting all of this, we use the idea of model selection by stepwise regression technique and compare the AIC of these models and identified a better model between all of them. Then, we compare a classical linear regression approach by simulations on 768 diverse residential buildings show that we can predict CL with low mean absolute error. By using ANOVA we determine variation in the different residuals. Also, we use non constant variance test to verify it. Furthermore, we check leverage and influence points as well as outliers as well as determined cook distance for influential points. By taking box cox transformation and weights, we also introduce WLS technique to fit the model for better results and did all type of important analysis to understand the energy efficiency. Finally, we show 5-fold cross validation to verify our model.

Keywords: exploratory analysis; model selection; MLR; K fold cross validation

Introduction:

Efficient building design is one of the important research interests in current times, the perfect calculation of the heating load (HL) and the cooling load (CL) is required to determine the specifications of the heating and cooling equipment needed to maintain comfortable indoor air conditions. In order to find the required cooling and heating capacities, architects and building designers need information about the characteristics of the buildings. The ultimate goal of this paper is to achieve minimum energy consumption and improve system efficiency. Meng et al (2014) studied multi-zone variable air volume and variable water volume air-conditioning systems. The dynamic models of HVAC sub-systems were built by the adaptive directional forgetting method. 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. Lombard et al. (2008) analyzed information concerning energy consumption in buildings, and particularly related to HVAC systems and comparisons between different countries are presented specially for commercial buildings. In this study we only work with cooling load of buildings.

Source of Data:

Applicable data is the primary criteria of any sort of regression model, because we use this data to actually make the model. If the data is flawed, the model will be flawed. Thus, the first step in regression modeling is to ensure that your data is reliable. There is no universal approach to verifying the quality of your data, unfortunately. Our job then becomes verifying your source's reliability and correctness as much as possible. Here, the data we collected, is from reliable source.

The dataset was created by Angeliki Xifara (angxifara '@' gmail.com, Civil/Structural Engineer) and was processed by Athanasios Tsanas (tsanasthanasis '@' gmail.com, Oxford Centre for Industrial and Applied Mathematics, University of Oxford, UK).

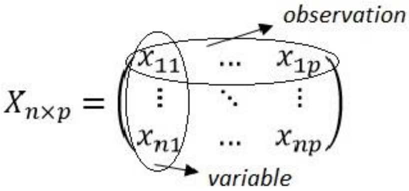
They perform energy analysis using 12 different building shapes simulated in Ecotect. The buildings differ with respect to the glazing area, the glazing area distribution, and the orientation, amongst other parameters. They simulate various settings as functions of the afore-mentioned characteristics to obtain 768 building shapes. The dataset comprises 768 samples and 8 features, aiming to predict two real valued responses. It can also be used as a multi-class classification problem if the response is rounded to the nearest integer. Mathematical representation of the predictors and response to facilitate the presentation of the subsequent regression analysis and results.

Table 01: Nomenclature of predictors and response

Variables	Descriptions
X_1	Relative compactness
X_2	Surface area
X_3	Wall area
X_4	Roof area
X_5	Overall Hight
X_6	Orientation
X_7	Grazing area
X_8	Grazing area distribution
Y_{CL}	Cooling Load

Methodology:

Let us take the data matrix X as,


$X_{n\times p} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$
observation
variable

We want to apply multiple linear regression of X with Y_i. The multiple linear regression model is

$$Y_i \sim N(\beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p, \sigma^2) \qquad (1)$$

independently across the $i = 1, \ldots, p$ observations.

In classical linear regression are similar if $n >> p$ and the priors are uninformative. However, the results can be different for challenging problems, and the interpretation is different in all cases.

The Pearson correlation coefficient is used to measure the strength of a linear association between two variables, where the value r = 1 means a perfect positive correlation and the value r = -1 means a perfect negative correlation. So, for example, you could use this test to find out whether people's height and weight are correlated (they will be - the taller people are, the heavier they're likely to be). There are few requirements for Pearson's correlation coefficient, firstly, scale of measurement should be interval or ratio, variables should be approximately normally distributed, association should be linear and there should be no outliers in the data. We can use heat plot for understanding the frequency of correlation between predictors and with response.

Firstly, we need to start with model selection. Model selection method based on p values or adjusted R square by stepwise regression. These models allow you to assess the relationship between variables in a data set and a continuous response variable. Sometimes we pick variable based on expert's opinion. We can check AIC too. Lowest AIC means better model. However, there are many ways to select a model for MLR.

One of the mostly used model selection method is backward elimination. In this case, we use full model and then we will drop predictor variable until we get our parsimonious model. With backward elimination technique we start with full model and record adjusted R square of each model and its reduced one and pick the model which has highest adjusted R square and minimum AIC. After selecting the number of predictors that we would like to use in the model we will estimate LSEs for each of the input variables by using following rules

The least squares estimate of $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is

$$\widehat{\beta_{ols}} = \operatorname{argmin}_{\beta} \sum_{i=1}^{n} (y_i - \mu_i)^2$$

Where $\mu_i = \beta_0 + X_{i1}\beta_1 + \ldots + X_{ip}\beta_p$, $\widehat{\beta_{ols}}$ is unbiased even if the errors are non-Gaussian. If the errors are Gaussian then the likelihood is proportional to

$$\exp\left[-\frac{\sum_{i=1}^{n}(y_i - \mu_i)^2}{2\sigma^2}\right]$$

Therefore, if the errors are Gaussian $\widehat{\beta_{ols}}$ is also the MLE. Linear regression is often simpler to describe using linear models using matrix form.

Let $Y = (Y_1, \ldots, Y_n)^T$ be the response vector and X be the $n \times (p + 1)$ matrix of covariates

$$\widehat{\beta_{ols}} = (X^TX)^{-1} X^TY$$

If the errors are Gaussian then the sampling distribution is

$$\widehat{\beta_{ols}} = N[\beta, \quad \sigma^2(X^TX)^{-1}]$$

If the variance σ^2 is estimated using the mean squared residual error then the sampling distribution is multivariate t. As with a least squares' analysis, it is crucial to verify this is appropriate using QQ-plots, added variable plots. In regression analysis we make certain assumptions about the conditional distributions of the dependent variable which we try to predict. The following three assumptions that are quite similar to the assumptions we made in ANOVA.

Normality: All conditional distributions are normally distributed (e.g. the distribution of sale volumes in all months in which advertising has been or will ever be some fixed level is normal).

Homoscedasticity: All conditional (normal) distributions have the same variance σ^2 , For checking this we can see the plot of fitted versus standardized residuals and also, we can do the equal variance hypothesis tests.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 October 2020 doi:10.20944/preprints202010.0082.v1

Linearity: The means of the conditional distributions are linearly related to the value of the independent variable. In statistics, the variance inflation factor (VIF) is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares' regression analysis. Then, calculate the VIF factor for $\widehat{\beta_i}$ by the following formula

$$VIF_i = \frac{1}{1 - R_i^2}$$

Where, R_i^2 is the coefficient of determination of the regression equation. Analyze the magnitude of multicollinearity by considering the size of theVIF_i. A rule of thumb is that if $VIF_i > 10$.

Again, the matrix form of

$$\widehat{\beta_{ols}} = (X^T X)^{-1} X^T y$$

the predicted responses can be alternatively written as:

$$\hat{y} = X(X^T X)^{-1} X^T y$$

$$\hat{y} = Hy$$

Where the hat matrix is given by

$$H = X(X^T X)^{-1} X^T$$

H involves the weights $h_{ii}; i = 1 \dots n$ depends on predictors.

The leverage, h_{ii}, quantifies the influence that the observed response variable has on its predicted value $\hat{y}$ That is, if h_{ii} is small, then the observed response y plays only a small role in the value of the predicted response $\hat{y}$. On the other hand, if h_{ii} is large, then the observed response y plays a large role in the value of the predicted response $\hat{y}$. It's for this reason that the $h_{ii}$are called the "leverages" which lies between zero and one.

Cook's distance D_i of observation $i = 1 \dots n$ can be expressed in terms of leverage

$$D_i = \frac{e_i^2}{ps^2}\left[\frac{h_{ii}}{(1 - h_{ii})^2}\right]$$

Where

$$e_i = y - \hat{y}$$

and

$$s^2 = \frac{e^T e}{n - p}$$

Before investigating all of the extreme cases discussed above, if We use this transformation in case of non-normal residuals popped up in QQ plots. The Box-Cox method considers a family of transformations on strictly positive response variables,

$$g_\lambda(y) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}; \; ; \lambda \neq 0 \\ \log(\lambda) \, ; \lambda = 0 \end{cases} \tag{2}$$

The λ parameter is chosen by numerically maximizing the log-likelihood L(λ).

The method of ordinary least squares assumes that there is constant variance in the errors (which is called homoscedasticity). The method of weighted least squares can be used when the ordinary least squares assumption of constant variance in the errors is violated (which is called heteroscedasticity). Error is assumed to be (multivariate) normally distributed with mean vector **0** and nonconstant variance-covariance matrix

$$\begin{pmatrix} \sigma_1^2 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_n^2 \end{pmatrix}$$

If we define the reciprocal of each variance, σ_i^2, as the weight, $w_i = \frac{1}{\sigma_i^2}$, then let matrix W be a diagonal matrix containing these weights:

$$\mathbf{W} = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{pmatrix}$$

The weighted least squares estimate is then

$$\widehat{\beta_{WLS}} = (X^TWX)^{-1} X^TWy$$

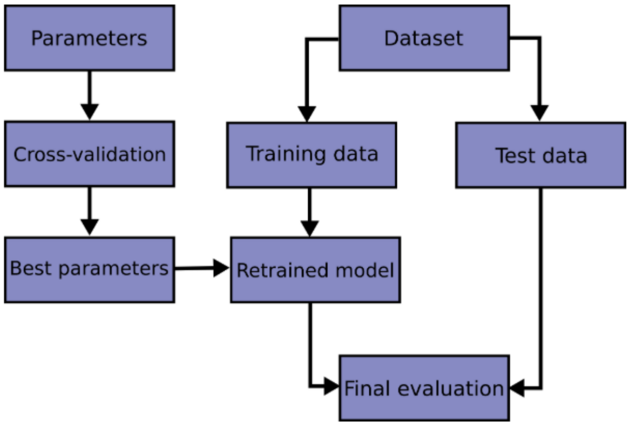
Moreover, $100(1-\alpha)\%$ confidence interval for the regression coefficient, β_j is given by

$$P\left[-t_{\alpha/2,n-k-1} \leq \frac{\hat{\beta}_j - \beta_j}{s\sqrt{g_{jj}}} \leq t_{\alpha/2,n-k-1}\right] = 1 - \alpha.$$

Solving the inequality for β_j gives

$$P(\hat{\beta}_j - t_{\alpha/2,n-k-1}s\sqrt{g_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{\alpha/2,n-k-1}s\sqrt{g_{jj}}) = 1 - \alpha.$$

$$\widehat{\beta_j} \pm t_{\alpha/2,n-k-1} s\sqrt{g_{jj}}$$



Flow chart 01: Cross Validation methodology

Cross validation of Model: K-fold cross validation is one way to improve over the holdout method. The data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set (k-1) times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set *k* different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

Results Discussion:

Exploratory analysis of data: It is always good to see the plot of data, to get a basic sense of its shape and ensure that nothing looks out of place. For instance, we may expect to see a somewhat linear relationship between two parameters. If we see something else, such as a horizontal line, you should investigate further. Our assumption about a linear relationship could be wrong, or the data may be corrupted (see figure 01, below). Or perhaps something completely unexpected is going on. Regardless, one must understand what might be happening before one begins developing the model.

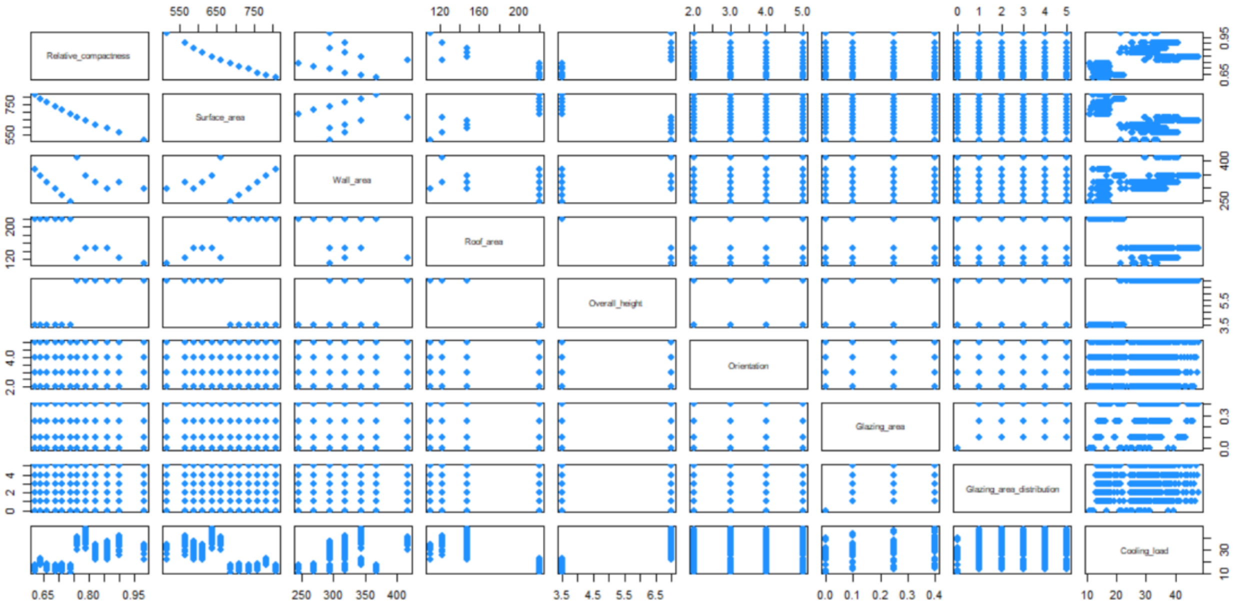


Figure 01: Representation of data set

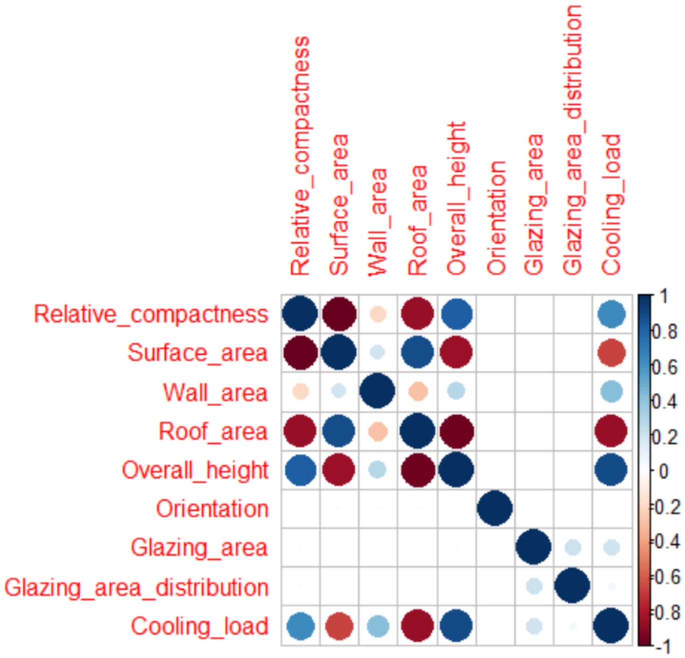


Figure 02: Heat plot of correlation between predictors and response

After watching the data exploration, we can investigate the response (cooling load) of the data set. Most importantly we need to verify whether the output is abiding the assumption of our model in equation (1).

Table 02: Model selection based on adjusted R square and AIC manually

step	Removed predictors	Adjusted R square	AIC
Full model	0	0.8868	3976.62
Model 1	Roof area	0.8868	3976.62
Model 2	Roof area, Glazing area distribution, Orientation	0.8869	3974.303
Model 3	Roof area, Glazing area distribution.	0.8876	3974.908
Model 4	Roof area, Glazing area distribution, Wall area	0.8814	4010.323
Model 5	Roof area, Glazing area distribution, Wall area, Surface area	0.8815	4008.329
Model 6	Roof area, Glazing area distribution, Relative compactness	0.8811	4011.2
Model 7	Roof area, Relative compactness	0.881	4013.789

According to Table 02 we observe that Most importantly model 2 has the lowest AIC and which suggests we should choose model 2. From table 03 given below, we can see the rate of change of cooling load with respect to five different predictors. I deliberately used orientation to show it has no effect to predict cooling load because we can reject it with 5% level of significance. In the model selection part, we also noticed the same idea about orientation. Without orientation, our model gave lowest AIC.

Table 03: summary statistics of the model 3

Coefficients	Estimates	Std. Error	Pr(> |t|)
Intercept	97.336561	20.754252	3.24e-06
Relative compactness	-70.787707	11.219992	4.76e-10
Surface area	-0.088245	0.018620	< 2e-16
Overall height	4.283843	0.368557	1.15e-09
Wall area	0.044682	0.007249	1.15e-09
Orientation	0.121510	0.103269	0.24
Glazing area	14.817971	0.867239	< 2e-16

According to table 03 we realize that orientation is not an important factor for cooling load of building however we can not rely on p value always. According to the model section part of this study we also saw that orientation is not a very important predictor for our response. That is true for our data set. Although, orientation may play an important factor to maintain energy efficiency of buildings. Thus, we want to see the residual plots of model 2 and analyze them. So, it is better to select model 2 for our analysis.

So, the summary statistics of the model 2 can be like below

Table 04: Model 2 summary

Coefficients	Estimates	Std. Error	Pr(>|t|)
Intercept	97.761848	20.756339	$2.94e^{-06}$
Relative compactness	-70.787707	11.222822	$4.8e^{-10}$
Surface area	-0.088245	0.018624	$2.56e^{-06}$
Overall height	4.283843	0.368650	$< 2e^{-16}$
Wall area	0.044682	0.007251	$1.16e^{-09}$
Glazing area	14.817971	0.867458	$< 2e^{-16}$

Also, we need to note the following calculations from the table 05. Residual standard error: 3.2 on 762 degrees of freedom. Multiple R-squared: 0.8876 and Adjusted R-squared: 0.8868. F-statistic: 1203 on 5 and 762 DF, and p value $< 2.2e^{-16}$.

Now the equation of the most appropriate model is

$$y_i = 97.761848 - 70.787707 * Relative\ compactness - 0.088245 * Surface\ area + 4.283843 * Overall\ height + 0.044682 * \text{Wall area} + 14.817971 * \text{Glazing area} \quad (3)$$

From the regression model we can observe that rate of change with respect to relative compactness, surface area has negative effect on cooling load of buildings. On the other hand, rate of change with respect to wall area, overall height and grazing area have positive effect, however it is numerically small for wall area. Which means that small wall area may have less effect on cooling load according to our observation.

There is a problem with the R^2 for multiple regression. Yes, it is still the percent of the total variation that can be explained by the regression equation, but the largest value of R^2 will always occur when all of the predictor variables are included, even if those predictor variables don't significantly contribute to the model. R^2 will only go down (or stay the same) as variables are removed, but never increase.

The Adjusted-R^2 uses the variances instead of the variations. That means that it takes into consideration the sample size and the number of predictor variables. The value of the adjusted-R^2 can actually increase with fewer variables or smaller sample sizes. We should always look at the adjusted-R^2 when comparing models with different sample sizes or number of predictor variables, not the R^2. If you have a tie for two models that have the same adjusted-R^2, then take the one with the fewer variables as it's a simpler model.

Table 05: Analysis of Variance Table (ANOVA)

Source	DF	Sum of Square	Mean Square	F value	Pr(>F)
Relative compactness	1	27931.9	27931.9	2726.885	$< 2.2e^{-16}$
Surface area	1	8254.2	8254.2	805.823	$< 2.2e^{-16}$
Overall height	1	22046.5	22046.5	2152.309	$< 2.2e^{-16}$
Wall area	1	389.0	389.0	37.973	$1.162e^{-10}$
Glazing area	1	2988.9	2988.9	291.797	$< 2.2e^{-16}$
Residuals	762	7805.3	10.2		
Total	767				

Here's a summary of the table of coefficients. We're making our decision at an α = 0.05 level of significance, so if the p-value < 0.05, we'll reject the null hypothesis and retain it otherwise. For this multiple linear regression, we have

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1: At\ least\ one\ \beta\ is\ not\ zero$$

The null hypothesis claims that there is no significant correlation at all. That is, all of the coefficients are zero and none of the variables belong in the model.

The alternative hypothesis is not that every variable belongs in the model but that at least one of the variables belongs in the model. If you remember back to probability, the complement of "none" is "at least one" and that's what we're seeing here. In this case, because our p-value is very small (almost close to zero), we reject that there is no correlation at all and say that we do have a good model for prediction.

Test of heteroscedasticity for the model can be done by the following test

$$H_0: \sigma^{2'}s\ are\ equal$$

Non-constant Variance Score Test has been conducted and Variance formula: ~ fitted.values.

Chi-square = 187.5252, Df = 1, p = 2.22 e^{-16}. At 5% level of significance we can say that we do have sufficient evidence to reject null hypothesis. So, variances are not equal. So, it has a problem with heteroscedasticity.

The method of weighted least squares can be used when the ordinary least squares assumption of constant variance in the errors is violated (which is called heteroscedasticity).

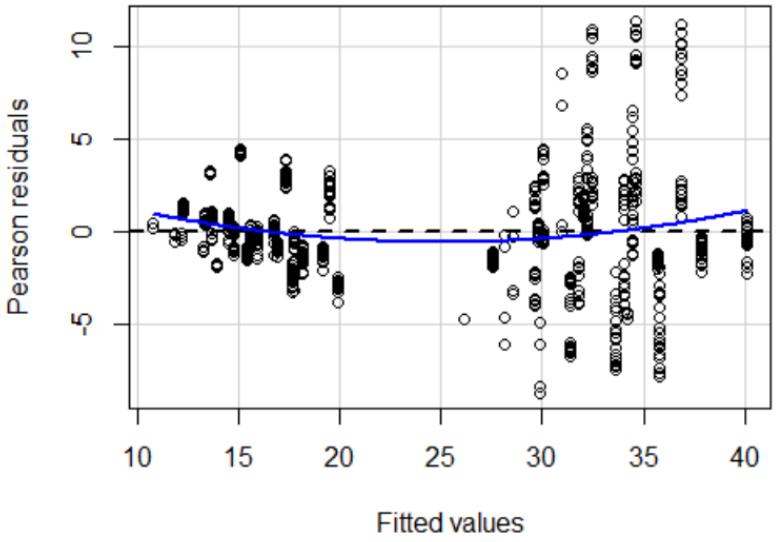


Figure 03: Fitted vs. residual plots for model-2

When conducting a residual analysis, a "**residuals versus fits plot**" is the most frequently created plot. It is a scatter plot of residuals on the *y* axis and fitted values (estimated responses) on the *x* axis. The plot is used to detect non-linearity, unequal error variances, and outliers. In this plot, we see that the residuals tend to increase as we move to the right. Additionally, the residuals are uniformly scattered above and below zero.

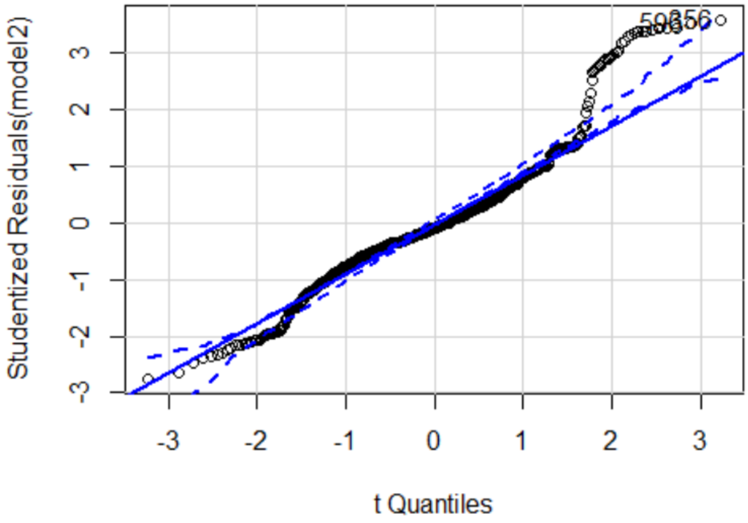


Figure 04: Pearson residual plots for fitted values and other assumptions

From the figure 09 we obtained 356 and 596 are the outlier and residuals are not normal. We can go to data set and delete these values to get rid of outliers. An outlier is a data point whose response y does not follow the general trend of the rest of the data. So, we can cut it off. QQ plot also tell us, our residuals are not normal approximately. It has a fat tail at the end. It is better to fit Variance Gamma or Normal inverse gaussian distribution for better assumption. In our study we do not need to consider such extreme stuffs.

To show it we used the histogram of the response variable and check the normally. We also showed the QQ plot of the response variable to see the actual condition of data beside straight line. We found that the output variable cooling load is not behave like perfectly normal.

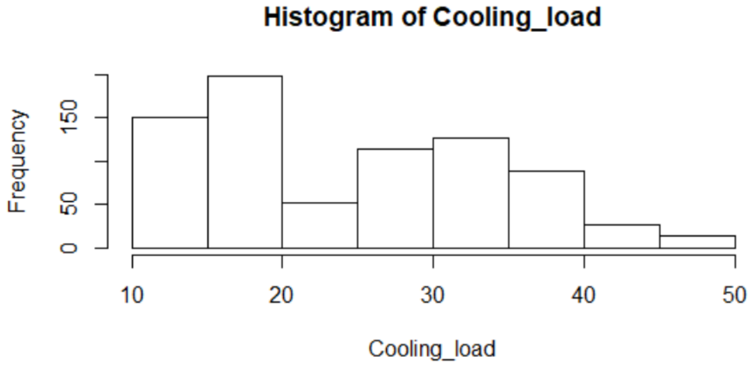


Figure 05: Histogram of response variable

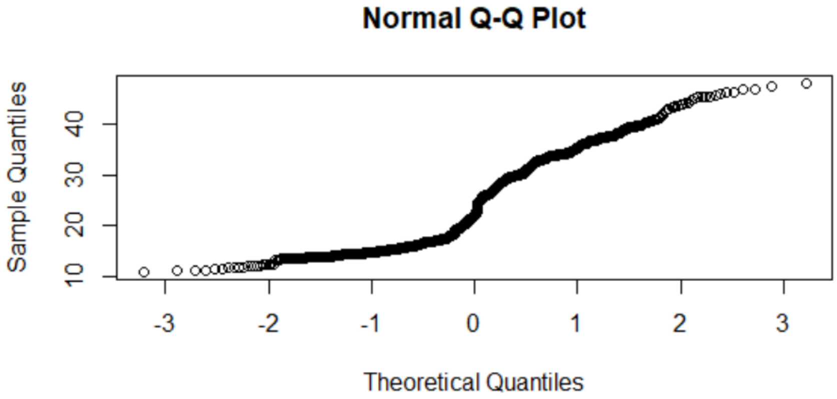


Figure 06: QQ plot of Cooling load

The quantile-quantile (Q-Q) plot, shows the distribution of the data against the expected normal distribution. For normally distributed data, observations should lie approximately on a straight line. However out data does not look like this. So, our data is not quite normal. To make it normal we can use box cox transformation described in equation (2) of the data set. We get $\lambda = -0.8$ from the following box cox plot given by R codes.

we see that λ=-0.8 and is extremely close to the maximum, which suggests a transformation of the form

$$\frac{Cooling\_Load^{\lambda} - 1}{\lambda} = \frac{Cooling\_Load^{-0.8} - 1}{-0.8}$$

The following figure estimate the lambda having max log likelihood.

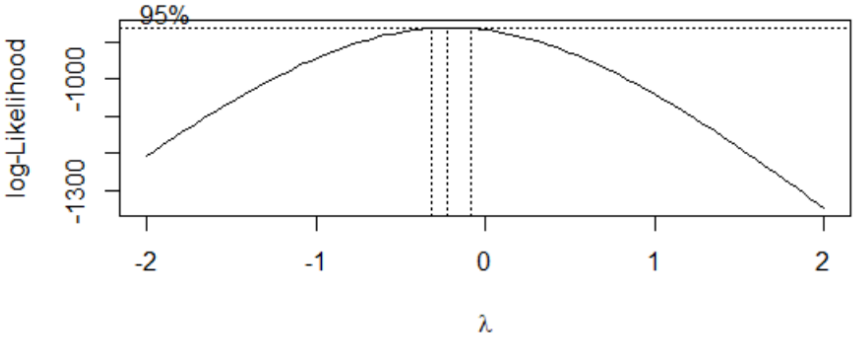


Figure 07: Box-Cox plot to identify λ with 95% Confidence interval.

After taking box cox transformation we check the QQ plot again and want to see the normality condition of residuals.

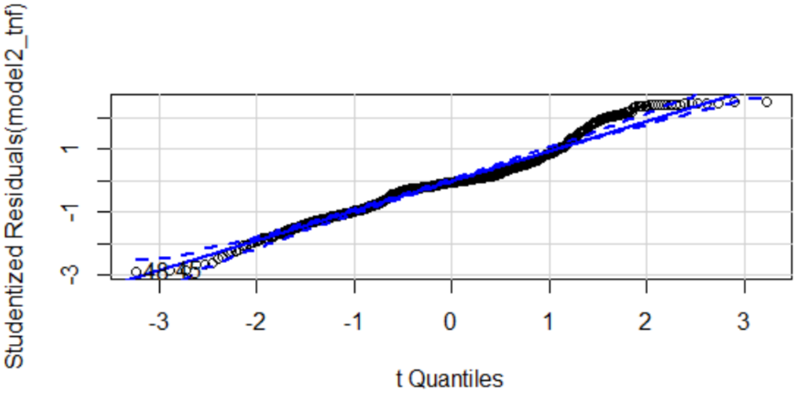


Figure 06: QQ plot of Cooling load after box-cox transform

Here, we get better fit with the straight line and it is approximately normal now. In this case we have 45 and 48 are the outliers. Also, we should remember, we failed in constant variance test. We found that our model is heteroscedastic. Thus, we should introduce WLS.

Table 06: summary for WLS with box cox transformed response

Coefficients	Estimates	Std. Error	Pr(>|t|)
Intercept	$9.424e^{-01}$	$6.312e^{-02}$	$< 2e^{-16}$
Relative compactness	$1.902e^{-02}$	$3.416e^{-02}$	0.5777
Surface area	$7.122e^{-05}$	$5.674e^{-05}$	0.20977
Overall height	$1.878e^{-02}$	$1.13e^{-03}$	$< 2e^{-16}$
Wall area	$7.904e^{-05}$	$2.236e^{-05}$	$1.16e^{-09}$
Glazing area	$5.863e^{-02}$	$2.638e^{-03}$	$< 2e^{-16}$

From the summary statistics we can see that the model becomes

$$y_i = 9.424\mathrm{e}^{-01} + 1.902\mathrm{e}^{-02} * Relative\ compactness + 7.122\mathrm{e}^{-05} * Surface\ area + 1.878\mathrm{e}^{-02} * Overall\ height + 7.904\mathrm{e}^{-05} * \text{Wall area} + 5.863\mathrm{e}^{-02} * \text{Glazing area} \qquad (4)$$

In the WLS model the Residual standard error: 0.009118 on 762 degrees of freedom, Multiple R-squared: 0.913, Adjusted R-squared: 0.913. F-statistic: 1613 on 5 and 762 DF, p $< 2.2\mathrm{e}^{-16}$

Collinearity is found here. VIF tests has been shown and identified values greaten than 10 gives multi collinearity

Table 05: VIF test values for WLS model

Relative compactness	Surface area	Overall height	Wall area	Glazing area
105.096669	201.616276	31.640175	7.692107	1.000008

Table 06: Analysis of Variance (ANOVA) for WLS model

Source	DF	Sum of Square	Mean Square	F value	Pr(>F)
Relative compactness	1	0.33606	0.33606	4042.453	$< 2.2\mathrm{e}^{-16}$
Surface area	1	0.04954	0.04954	595.887	$< 2.2\mathrm{e}^{-16}$
Overall height	1	0.24296	0.24296	2922.628	$< 2.2\mathrm{e}^{-16}$
Wall area	1	0.00103	0.00103	12.448	0.0004434
Glazing area	1	0.04106	0.04106	493.969	$< 2.2\mathrm{e}^{-16}$
Residuals	762	0.06335	0.00008		
Total	767				

Since each weight is inversely proportional to the error variance, it reflects the information in that observation. So, an observation with small error variance has a large weight since it contains relatively more information than an observation with large error variance (small weight). Set of weights will (legitimately) impact the widths of statistical intervals.

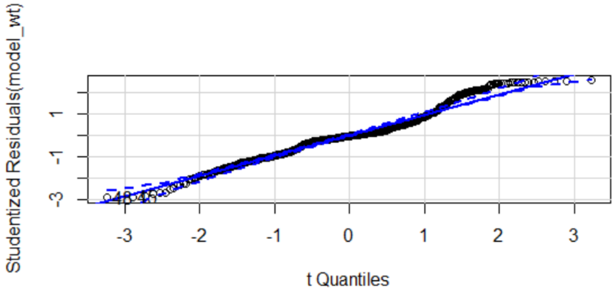


Figure 07: QQ plot of WLS

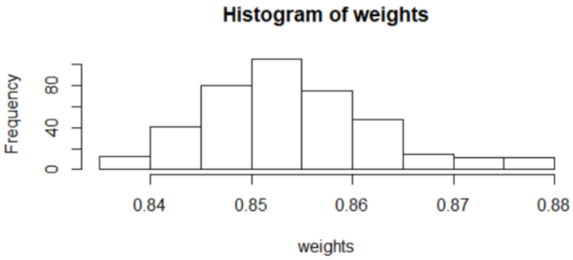


Figure 08: Range of weights and its distribution

From the figure showed below we observe that after fitting WLS model, relative compactness and surface area does not show any significant change in residues. It is showing constant almost. That is meaning that if we change these two predictors by 1 unit there is no significant anomalies to be observed.

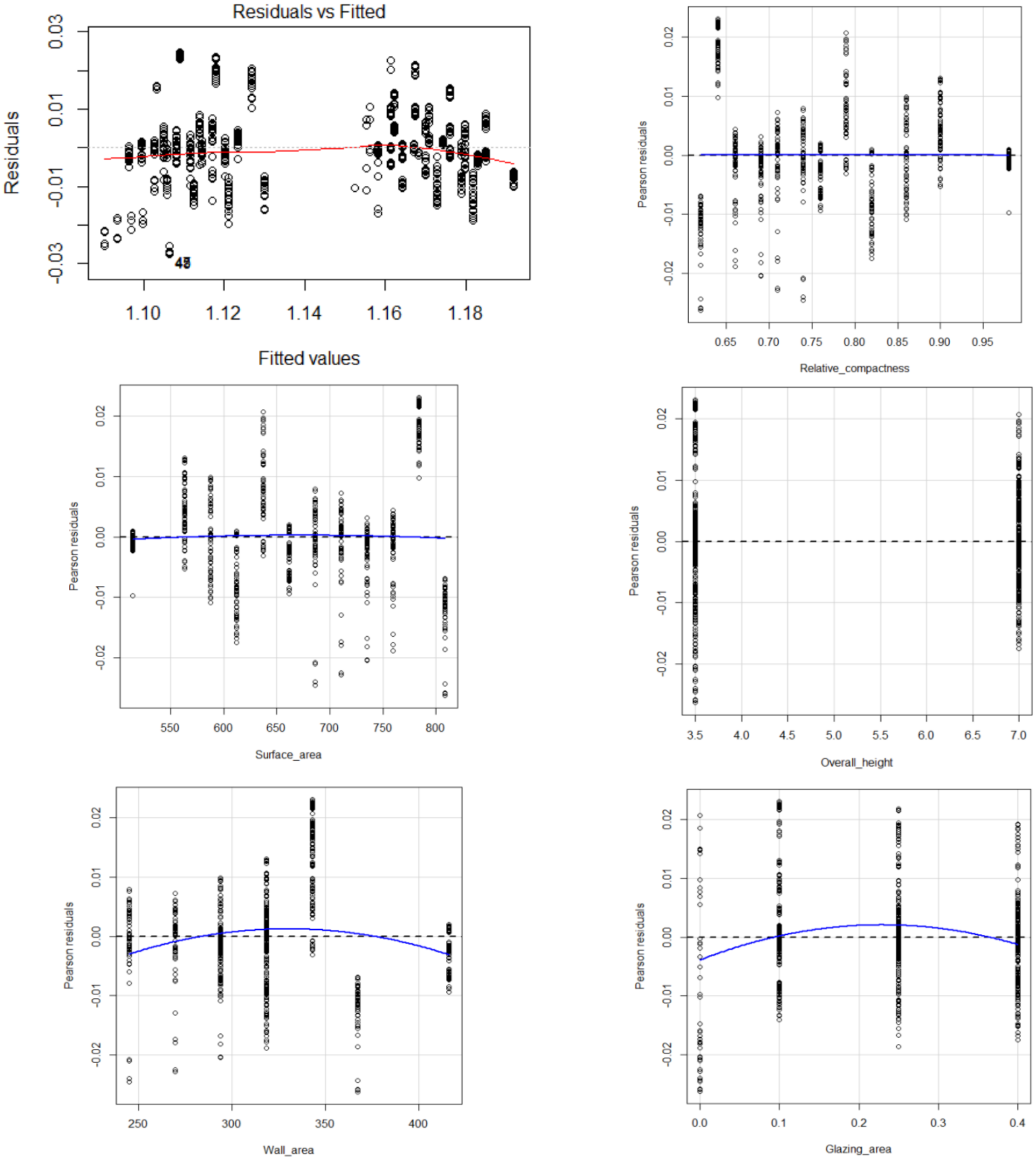


Figure 09: WLS model output

From figure 09 we can see the pattern of each predictors with residuals. Here residual vs. fits plot and what they suggest about the appropriateness of the multiple linear regression model. First of all, the residuals "bounce randomly" around the residual = 0 line. This suggests that the assumption that the relationship is linear is reasonable. We see that our WLS model gives almost linear relationship. Secondly, the residuals roughly form a "horizontal band" around the residual = 0 line. This suggests that the variances of the error terms are equal. However, we see very few non constant variances terms which can be ignorable. Finally, very few residual "stands out" from the basic random pattern of residuals. This suggests that there are few outliers. We can see, 45 and 48.

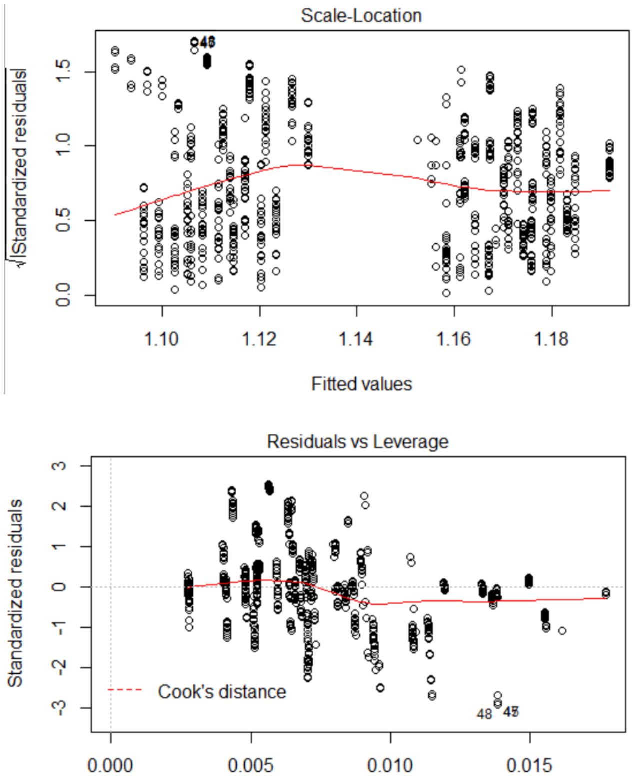


Figure 10: WLS model diagnosis plots

The scale-location plot (square rooted standardized residual vs. predicted value). This is useful for checking the assumption of homoscedasticity. In this particular plot we are checking to see if there is a pattern in the residuals and we found that there is a concave pattern in it. However, because of WLS fit, we found collinearity here. Thus, we need to reduce a predictor and check it again.

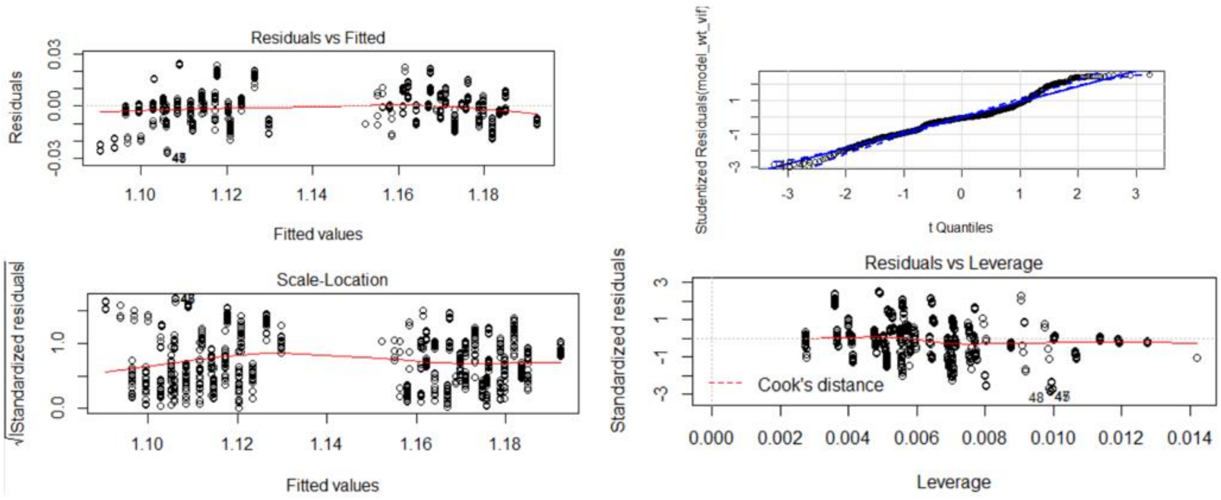


Figure: diagnostic plots for our final model

Cook's D is a commonly used estimate of the influence of a data point when performing a least-squares regression analysis. Here we can see from the figure 15 and 16 we can see there are few values with high cook distance. Figure 16 shows the exact points which has biggest Cook's D.

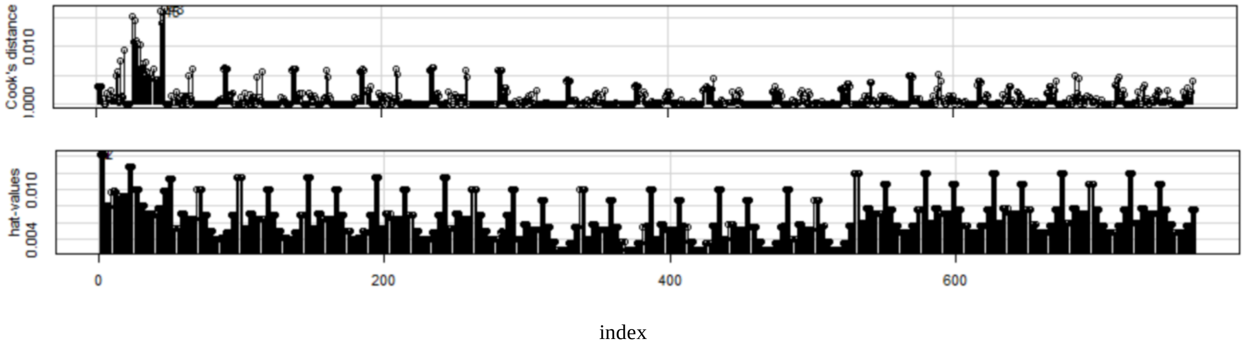


Figure11: Diagnostic plots with Hat values and Cook distance

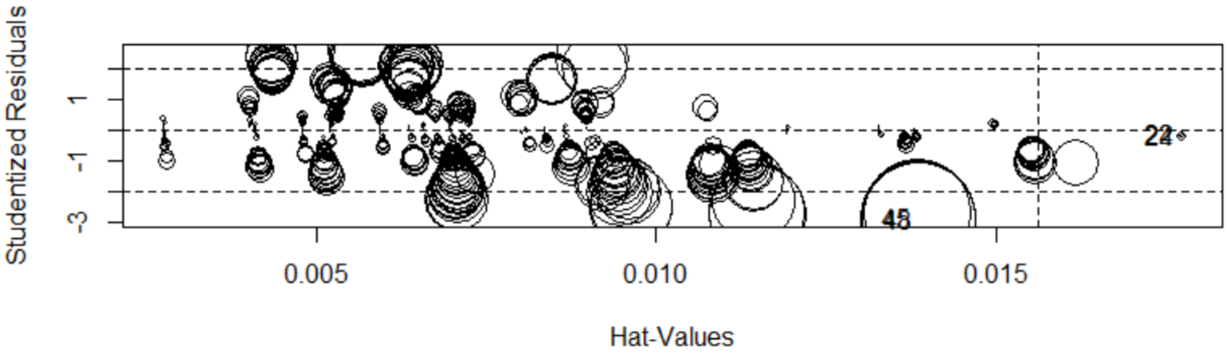


Figure 12: Influence plot after transformation of response variable

From the analysis of influence plot in figure 11, we can see the following output

Table 07: some extreme values from WLS model diagnosis

Data points	Standardized Residual	Hat	Cook-D
22	-0.1372486	0.01772626	$5.672951e^{-05}$
24	-0.2019629	0.01773586	$1.229031e^{-04}$
45	-2.8854279	0.01386842	$1.932885e^{-02}$
48	-2.9225801	0.01387279	$1.983057e^{-02}$

The assumption of a random sample and independent observations cannot be tested with diagnostic plots. It is an assumption that you can test by examining the study design. The 2nd plot in figure 09 is of "Cook's distance", which is a measure of the influence of each observation on the regression coefficients. The Cook's distance statistic is a measure, for each observation in turn, of the extent of change in model estimates when that particular observation is omitted. Any observation for which the Cook's distance is close to 1 or more, or that is substantially larger than other Cook's distances (highly influential data points), requires investigation. Here we have found four data points which may need to account. As we know, outliers may or may not be influential points. In fact, influential outliers are of the greatest concern. They should never be disregarded. Careful scrutiny of the original data may reveal an error in data entry that can be corrected. In our model we did not exclude it. From the QQ plot in figure 07 we can see 45 and 48 becomes influential.

Hypothesis test with final WLS models

$$H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_1: At\ least\ one\ \beta\ is\ not\ zero$$

In this case, because our p-value is very small (almost close to zero) for grazing area, wall area and overall heights, we reject that there is no correlation at all and say that we do have a good model for prediction.

Also, in table 05, we got the test values for VIF to check co-linearity.

Table 08: Confidence interval for the model WLS

	2.5 %	97.5%
(Intercept)	$9.554707e^{-01}$	$9.986528e^{-01}$
Surface area	$7.737760e^{-06}$	$7.440829e^{-05}$
Overall height	$1.666304e^{-02}$	$2.008869e^{-02}$
Wall area	$5.021156e^{-05}$	$1.221353e^{-04}$
Glazing area	$5.345437e^{-02}$	$6.380684e^{-02}$

k-fold cross-validation divides the entire dataset into *5* subsets. In turn, each subset is used as the validation sample, while the other *4-1* subsets are combined to use as the training sample.
There is statistical theory that shows that the appropriate choice of *k* depends on *n* and the type of predictor. Below, we have shown our results for cross validation.

Table 09: Cross validation

Resampling:	Cross-Validated (5-fold)		
Summary of sample sizes:	615, 614, 614, 615, 614		
Resampling results:	RMSE	R-squared	MAE
	0.009677356	0.9138734	0.00707785
RMSE of fitted model	0.009118		

Model Type: Linear Regression with 768 samples and 5 predictors. From the above table we observe that the RMSE of fitted model is less than CV RMSE. Which is approximately same. So, we do not have any over fit for this model.
Note that cross-validation is used to estimate prediction error and sometimes aspects of the predic tion equation such as the number of clusters or number of predictors. The final predictor will be trained on all the training data.
There is a small problem with this method for assessing prediction error. The final predictor will be based on all n, or 100% percent of the sample but the estimated prediction error is based on

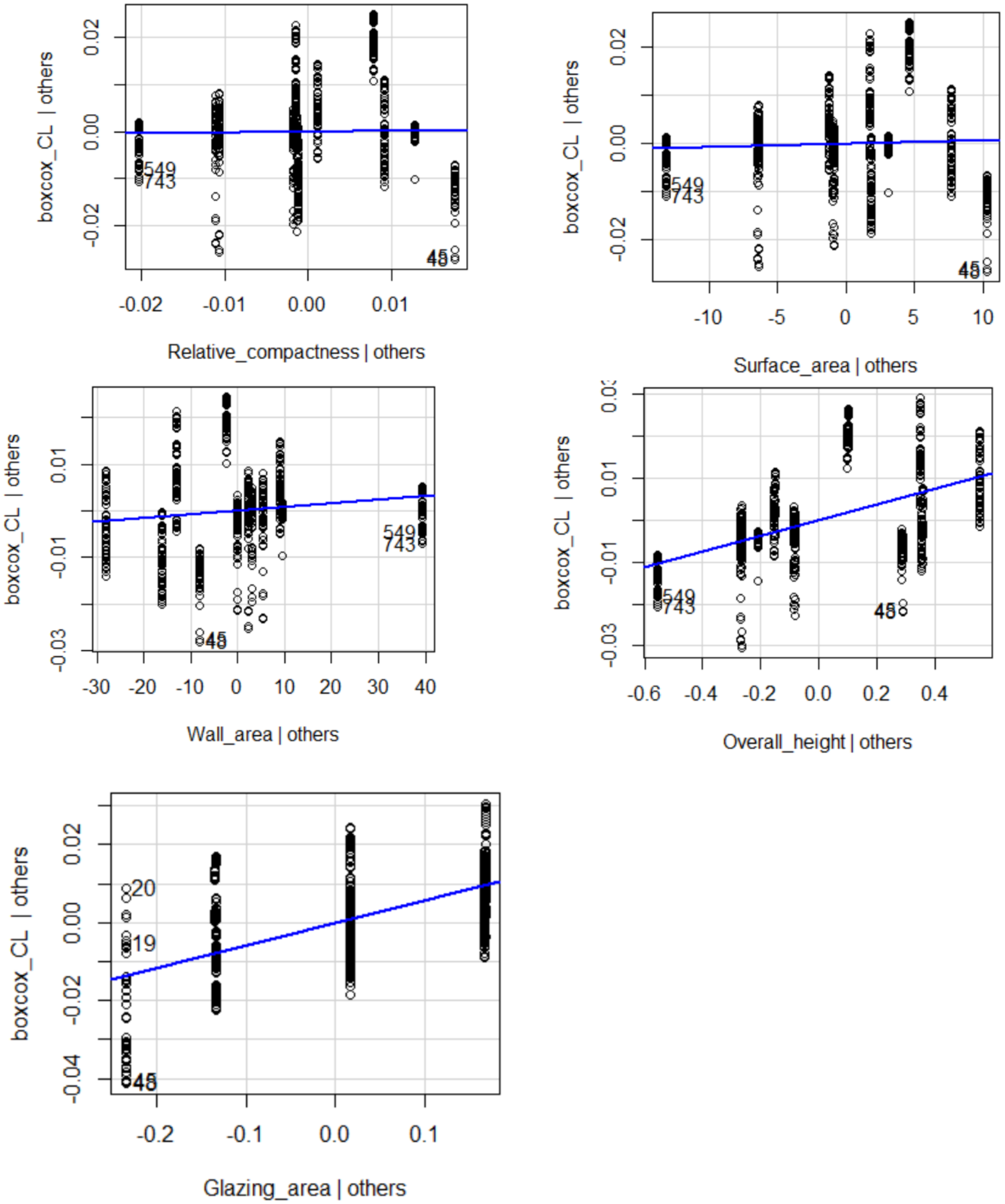
predictor developed on a smaller sample: $n > n - n/k$. So, the cross-validation estimate of prediction error might actually be pessimistic, might have slightly better prediction error than we assumed. However, with 10-fold cross-validation can't be too far off because you are using at least 90% of your samples. For small number of data set like, we choose here, 5-fold CV is a good choice. It also validates our model.

Conclusion

- Rate of change with respect to surface area, overall height, wall area and grazing area has a positive effect on cooling load
- However, wall area as well as surface area is numerically small in case of rate of change.
- MLR is not a good model to predict cooling load because we are losing important predictors.
- Although cross validation verified a good fit.
- Elastic net could be a better model because we can use two different penalties with regularization parameters which avail from CV.

Appendix

Added variables Plots



Worked out Codes in R

```
# How to read file in R
library(readxl)
ENB2012 <- read_excel("C:/Users/gmostafa/Desktop/Regression class/regression
project/ENB2012.xlsx")
data=ENB2012  # Name of my data set, change it according to yours
# LIBRARY should be loaded:
library(lattice)
library(ggplot2)
library(caret)
library(corrplot)
library(car)
library(lmtest)
library(sandwich)
# Name of each predictors/input variables
Relative_compactness= data$X1
Surface_area= data$X2
Wall_area = data$X3
Roof_area = data$X4
Overall_height = data$X5
Orientation = data$X6
Glazing_area = data$X7
Glazing_area_distribution = data$X8
Cooling_load = data$Y2

#Linear model fuction for regression
model2<-lm(Cooling_load~
        Relative_compactness+
        Surface_area+
        Overall_height+
          Wall_area+
        Glazing_area
      , data = data)
summary(model2)
anova(model2)

# testing heteroskedasticity
ncvTest(model2)
# Residuals vs. Fitted only
residualPlots(model2, ~1, fitted=TRUE)
# Outliers - QQ Plots
qqPlot(model2, id.n=3)
### Box cox transformation
y <-boxcox(model2, plotit = TRUE)
boxcox_CL=((Cooling_load^(-0.8)) - 1)/ (-0.8)
model2_tnf<-lm(boxcox_CL~
              Relative_compactness+
              Surface_area+
              Overall_height+
              Wall_area+
              Glazing_area
            , data = data)
summary(model2_tnf)
# Residuals vs. Fitted only
residualPlots(model2_tnf, ~1, fitted=TRUE)
# Outliers - QQ Plots
qqPlot(model2_tnf, id.n=3)
# fitting model with weights/ WLS
model_wt<- update(model2_tnf, weights=1/boxcox_CL)
summary(model_wt)
anova(model_wt)
plot(model_wt)
# testing heteroskedasticity
ncvTest(model_wt)
Grazing_area=Glazing_area
#residual plot
residualPlots(model_wt, fitted=TRUE)
residualPlots(model_wt, ~Relative_compactness, fitted=TRUE)
residualPlots(model_wt, ~Surface_area, fitted=TRUE)
residualPlots(model_wt, ~Overall_height, fitted=TRUE)
residualPlots(model_wt, ~Wall_area, fitted=TRUE)
residualPlots(model_wt, ~Glazing_area, fitted=TRUE)
# to see sqrt(std_res) and leverage
plot(model_wt)
# Outliers - QQ Plots
qqPlot(model_wt, id.n=3)
# High Leverage (hat) points
influenceIndexPlot(model_wt, id.n=3)
#Influence bubble plot
influencePlot(model_wt, id.n=3)

# Confidence Interval
confint(model_wt)
# cross validation for boxcox_CLm2
data5 <-data.frame(boxcox_CL,
                   Relative_compactness,
                   Surface_area,
                   Overall_height,
                  Wall_area,
                   Glazing_area)
train3.control=trainControl(method="cv",number=5)
model3<- train(boxcox_CL~., data =  data5, method = "lm",
            trControl = train3.control)
print(model3)
model3$results
# Influential Variables- Added- variable
avPlots(model_wt, ~Glazing_area,id.n=2, id.cex=0.7)
```

Reference:

1. Notes of Dr. Leif Ellingson, Department of Math & Stat, Texas Tech University.
2. James, Gareth; Witten, Daniela; Hastie, Trevor; Tibshirani, Robert (2017). An Introduction to Statistical Learning (8th ed.). Springer Science+Business Media New York.
3. 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
4. G. Platt, J. Li, R. Li, G. Poulton, G. James, J. Wall, Adaptive HVAC zone modelling for sustainable buildings, Energy and Buildings 42 (2010) 412–421.
5. Sanford Weisberg (2005). Applied Linear Regression, Third Edition. Hoboken NJ: Wiley.
6. An R Companion to Applied Regression, Second Edition / John Fox, Sanford Weisberg, Sage Publications, 2011
7. Data Manipulation with R / Phil Spector, Springer, 2008
8. Applied Econometrics with R / Christian Kleiber, AchimZeileis, Springer, 2008
9. Introductory Statistics with R / Peter Dalgaard, Springer, 2008
10. Complex Surveys. A guide to Analysis Using R / Thomas Lumley, Wiley, 2010
11. L. Perez-Lombard, J. Ortiz, C. Pout, A review on buildings energy consumption, I information, Energy and Buildings 40 (3) (2008) 394–398.
12. B. Dong, C. Cao, S.E. Lee, Applying support vector machines to predict building energy consumption in tropical region, Energy and Buildings 37 (2005) 545–553.
13. T. Catalina, J. Virgone, E. Blanco, Development and validation of regression models to predict monthly heating demand for residential buildings, Energy and Buildings 40 (2008) 1825–1832.
14. Mostafa, GM Fahad Bin, Revisiting the Performance of PCA Versus FDA Versus Simple Projection for Image Recognition (May 02, 2020). http://dx.doi.org/10.2139/ssrn.3606738
15. Q. Li, Q. Meng, J. Cai, H. Yoshino, A. Mochida, Applying support vector machine to predict hourly cooling load in the building, Applied Energy 86 (2009) 2249–2256.
16. J. Zhang, F. Haghighat, Development of artificial neural network based heat convection for thermal simulation of large rectangular cross-sectional area earth-to-earth heat exchanges, Energy and Buildings 42 (4) (2010) 435–440.
17. S.S.K. Kwok, R.K.K. Yuen, E.W.M. Lee, An intelligent approach to assessing the effect of building occupancy on building cooling load prediction, Building and Environment (2011), doi:10.1016/j.buildenv.2011.02.008.