Stann-Omar Jones
Advanced Statistics for Data Science
Professor Chong Liu

Regression Theory

Formally, the model for multiple linear regression, given n observations, is:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i \text{ for } i = 1,2, \ldots n.$$

Building Regression #1

```
model1 = lm(mydata2$heating_load~
                    mydata2$roof_area +
                    mydata2$surf_area +
                    mydata2$glaz_area,
                    data = mydata2)
```

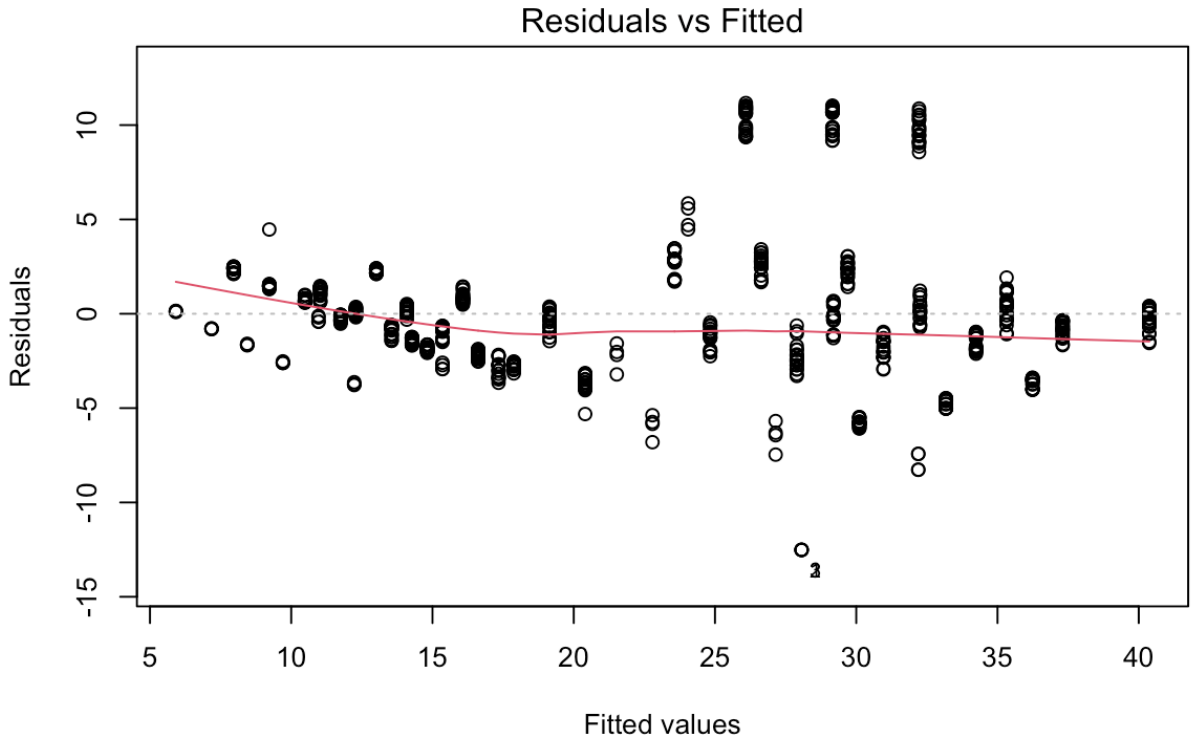
```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       32.539063   1.343003   24.23   <2e-16 ***
mydata2$roof_area -0.281039   0.006363  -44.17   <2e-16 ***
mydata2$surf_area  0.051526   0.003263   15.79   <2e-16 ***
mydata2$glaz_area 20.437968   1.021794   20.00   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.77 on 764 degrees of freedom
Multiple R-squared:  0.861,     Adjusted R-squared:  0.8604
F-statistic:  1577 on 3 and 764 DF,  p-value: < 2.2e-16
```

In our first regression, we are saying that heating load can be predicted by roof area, surface area, and glazing area. All of these parameters have a significance level very close to zero i.e. **<0.0001.**

Regression Assumptions

The four assumptions of linear regressions are the following:

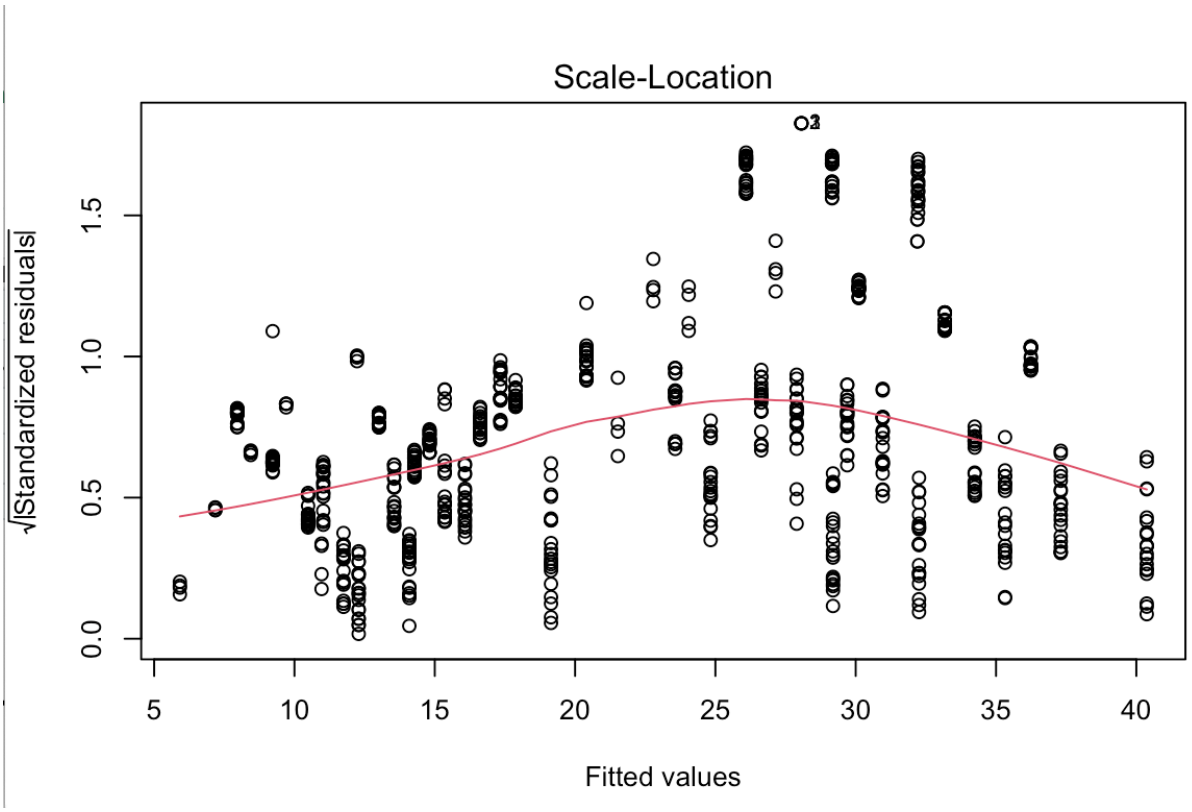
1. **Linearity**: The relationship between X and the mean of Y is linear.



The data points demonstrate a pattern on the residual plot but the line roughly fits the central line of 0. This suggests that there is a linear relationship between the predictors and the outcome variables but our sample may be too small, leading to predictability across the scatter plots.

2. **Homoscedasticity**: The variance of residual is the same for any value of X.

Based on the scale-location plot, the line is roughly horizontally plotted. However, there may be slight heteroskedasticity i.e. residuals are not equally spread across predictors.

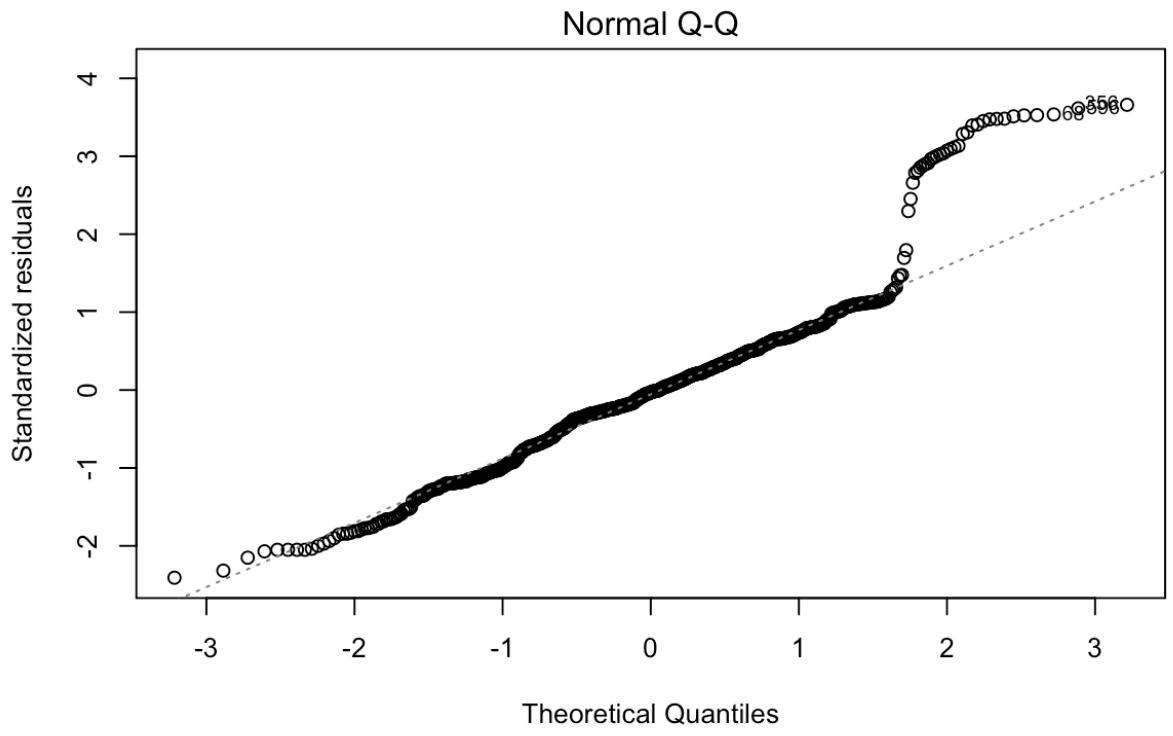


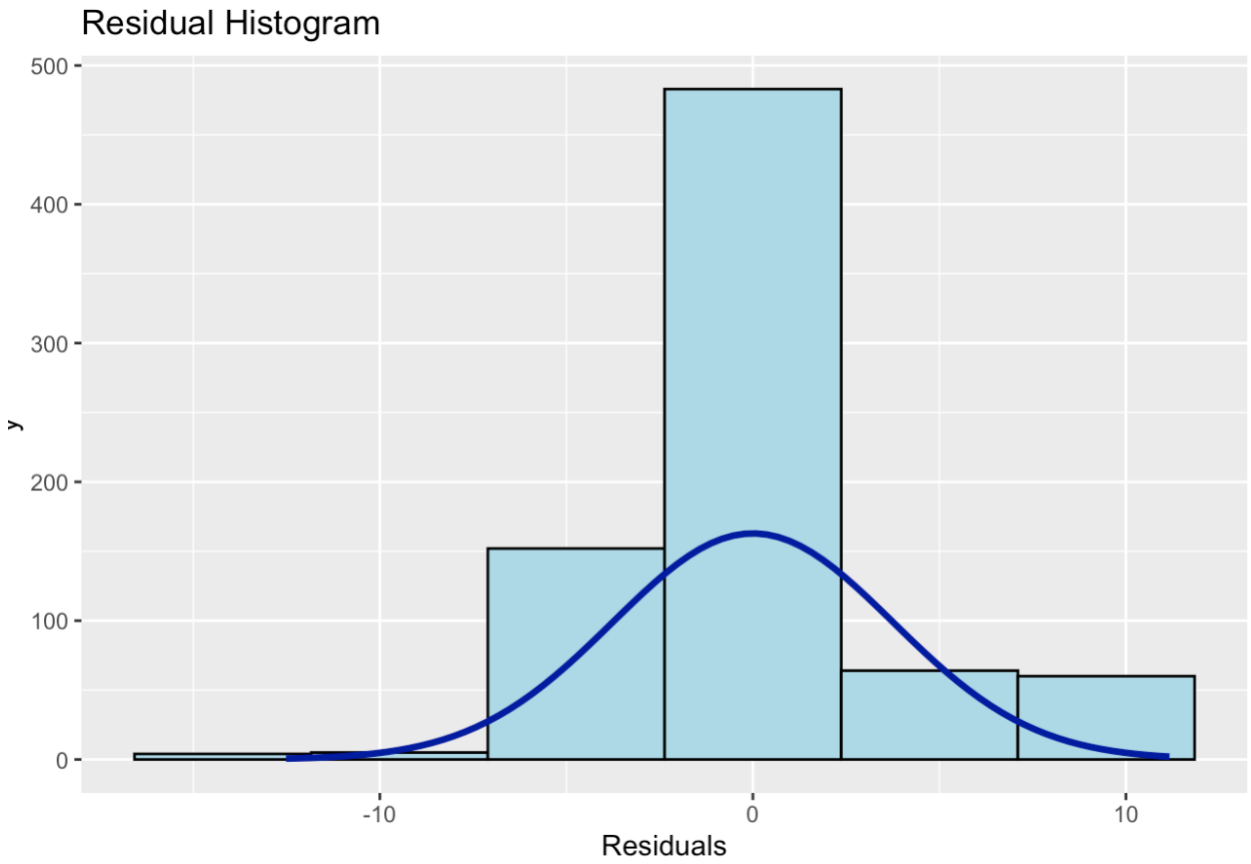
3. **Independence**: Observations are independent of each other.

This criterion is satisfied because each observation is of a different building structure i.e. each observation is unique.

4. **Normality**: For any fixed value of X, Y is normally distributed.

This residuals have a normal distribution: the residuals follow the diagonal line in the quantile-quantile graph but there is strong deviation at the end of the line suggesting a slight rightward skew. This is confirmed in the residual histogram as we can see a heavy right tail.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
3
-2
-1
0
1
2
3
4


Residual Histogram
y
Residuals
500
400
300
200
100
0
-10
0
10

	rel_compact	surf_area	wall_area	roof_area	height
median	7.500000e-01	6.737500e+02	3.185000e+02	1.837500e+02	5.250000e+00
mean	7.641667e-01	6.717083e+02	3.185000e+02	1.766042e+02	5.250000e+00
SE.mean	3.816916e-03	3.178534e+00	1.574235e+00	1.629786e+00	6.318884e-02
CI.mean.0.95	7.492841e-03	6.239658e+00	3.090321e+00	3.199370e+00	1.240436e-01
var	1.118887e-02	7.759164e+03	1.903270e+03	2.039963e+03	3.066493e+00
std.dev	1.057775e-01	8.808612e+01	4.362648e+01	4.516595e+01	1.751140e+00
coef.var	1.384220e-01	1.311374e-01	1.369748e-01	2.557468e-01	3.335506e-01
skewness	4.935786e-01	-1.246425e-01	5.313356e-01	-1.621288e-01	0.000000e+00
skew.2SE	2.797545e+00	-7.064591e-01	3.011548e+00	-9.189267e-01	0.000000e+00
kurtosis	-7.157384e-01	-1.065419e+00	9.994467e-02	-1.776395e+00	-2.002602e+00
kurt.2SE	-2.030976e+00	-3.023229e+00	2.836025e-01	-5.040689e+00	-5.682575e+00
normtest.W	9.334065e-01	9.496748e-01	9.268989e-01	7.459520e-01	6.365761e-01
normtest.p	5.323092e-18	1.628230e-15	7.145804e-19	1.125483e-32	2.196983e-37

5. **Little to no multi-collinearity**: Linear regression analysis assumes that there is no perfect exact relationship among exploratory variables

Based on our VIF output, we can see that all our inputs have VIF < 5, so there is some correlation within the expected range as the default VIF cutoff value is 5.

```
mydata2$roof_area mydata2$surf_area mydata2$glaz_area
         4.457656          4.457656          1.000000
```

Regression Equation

Therefore, the regression prediction equation is:

```
Heating load = 32.539 + (-0.281) * roof area + 0.515 * surface
area + 20.438 * glazing area
```

The coefficients or multipliers describe the size of the effect the independent variables have on your dependent variable *Y* (heating load), and the constant, 32.539 is the value *Y* is predicted to have when all the independent variables are equal to zero.

This tells us that for every single unit increase in roof area, our heating load decreases by -0.281 units. As surface area increases by 1 unit, heating increases by 0.515 units. And finally, as glazing area increases by 1 unit, the heating load increases by 20.438 units.

The prediction equation makes intuitive sense. Glass is extremely inefficient construction material and leaks heat on cold winter nights which will likely have a strong impact on the

heating load of a building. Increased building surface area given that there will be more space to heat. Since the roof as a building surface has the most exposed area to the sun, it contributes most of heat gains in the building.

Model Specification and Significance

Adjusted R-squared: The adjusted R2 is 0.8604. This means that 86% of the variance in the heating load is explained by our inputs.

F-test: The F-test statistic of 1577 is larger than the critical value of 3.807333 so we #then reject H0. and determine that at least one x variable is a significant determinant of y.

P-values: The p-values are highly significant and close to 0 so we can be certain that the model is well-specified.

Regression #2

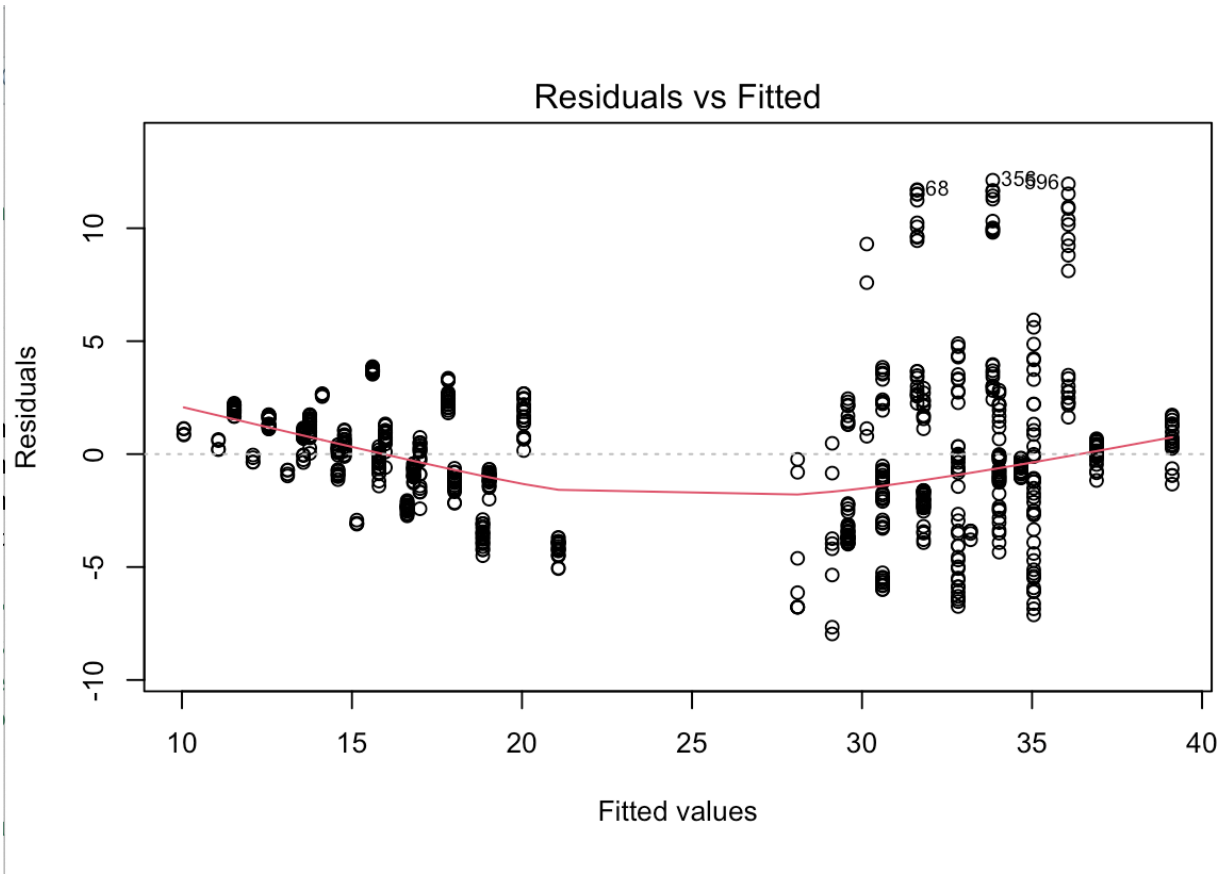
In our second regression model, we are saying that cooling load can be predicted by height, wall area, and glazing area. All of these parameters have a significance level very close to zero i.e. **<0.0001.**

```
model2 = lm(mydata2$cooling_load~
                        mydata2$glaz_area +
                        mydata2$height +
                        mydata2$wall_area,
                      data = mydata2)
```

Regression Assumptions

The four assumptions of linear regressions are the following:

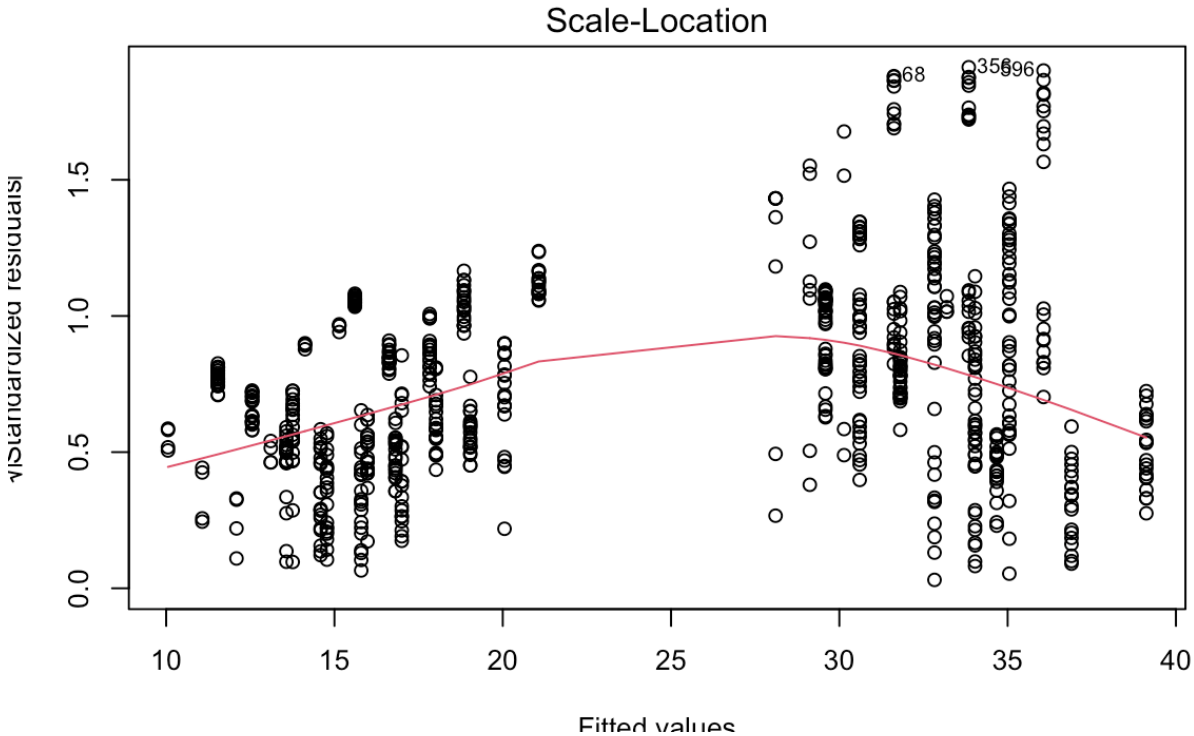
1. **Linearity**: The relationship between X and the mean of Y is linear.



The data points demonstrate a pattern on the residual plot. The line sags under the central line of 0. This suggests that there may not be a linear relationship between the predictors and the outcome variables or that our sample may be too small.

2. **Homoscedasticity**: The variance of residual is the same for any value of X.

Based on the scale-location plot, the line is curved so there may be heteroskedasticity present i.e. residuals are not equally spread across predictors.

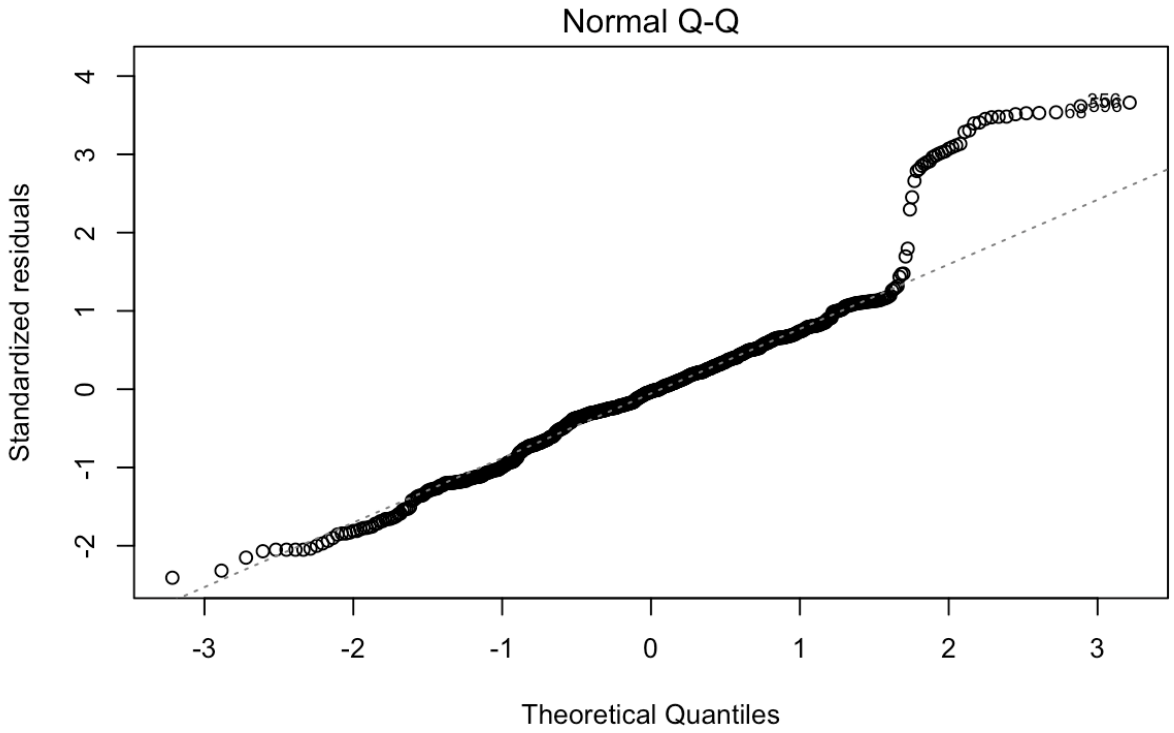


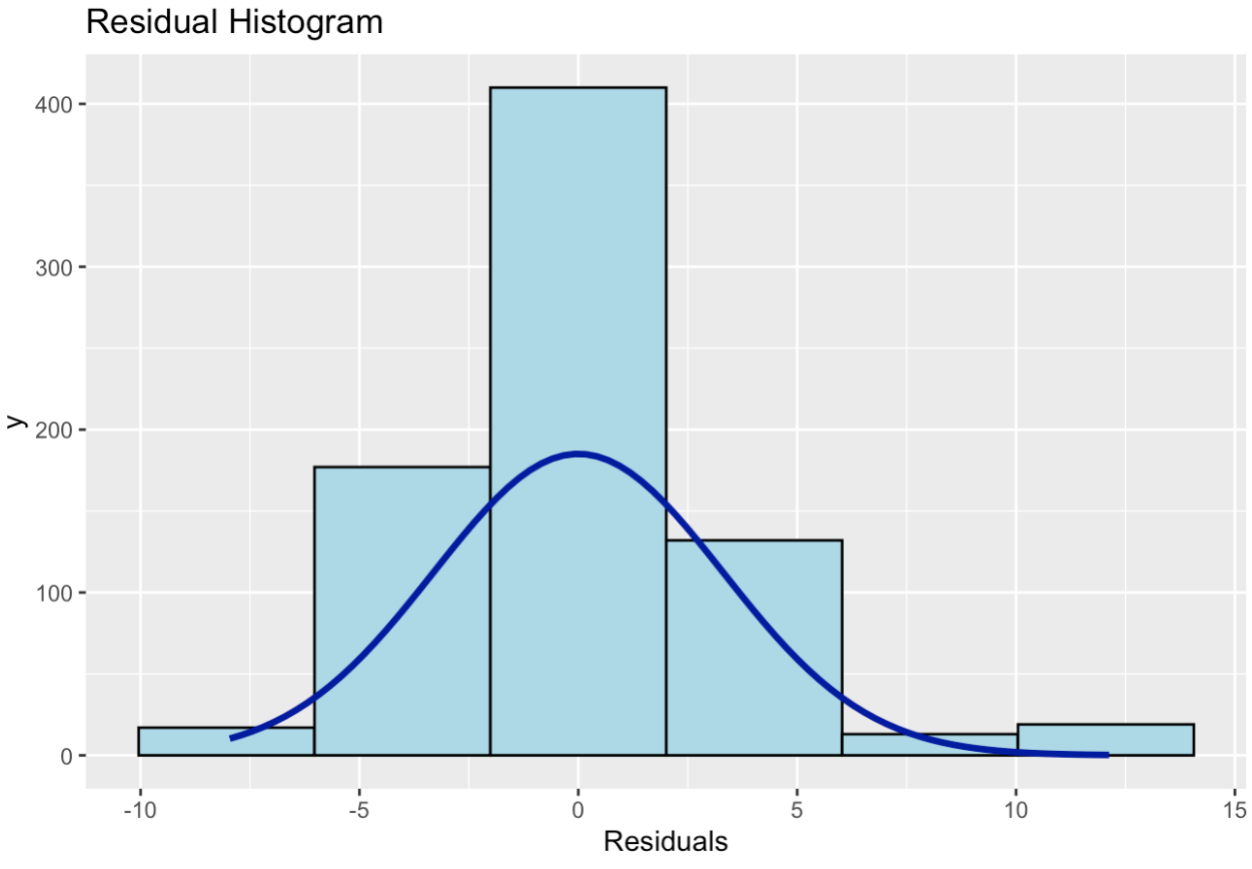
3. **Independence**: Observations are independent of each other.

This criterion is satisfied because each observation is of a different building structure i.e. each observation is unique.

4. **Normality**: For any fixed value of X, Y is normally distributed.

The residuals have a normal distribution: the residuals follow the diagonal line in the quantile-quantile graph but there is strong deviation at the end of the line suggesting a rightward skew. However, in the residual histogram, we can see that the residuals are roughly normal with a very slight left skew.


Normal Q-Q
Standardized residuals
Theoretical Quantiles
-3
-2
-1
0
1
2
3
-2
-1
0
1
2
3
4
356


Residual Histogram
y
Residuals
0
100
200
300
400
-10
-5
0
5
10
15

5. **Little to no multi-collinearity**: Linear regression analysis assumes that there is no perfect exact relationship among exploratory variables

Based on our VIF output, we can see that all our inputs have VIF ~ 1, so there is not strong correlation between the inputs as the default VIF cutoff value is 5.

```
mydata2$glaz_area    mydata2$height mydata2$wall_area
         1.000000          1.085714          1.085714
```

Regression Equation

The regression prediction equation is:

```
Cooling load = -16.136 + 14.818 * glazing area + 4.576 * height
+ 0.0415 * wall area
```

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -16.135847   0.915031  -17.63   <2e-16 ***
mydata2$glaz_area  14.817971   0.899052   16.48   <2e-16 ***
mydata2$height      4.575749   0.071268   64.20   <2e-16 ***
mydata2$wall_area   0.041532   0.002861   14.52   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.317 on 764 degrees of freedom
Multiple R-squared:  0.8789,    Adjusted R-squared:  0.8784
F-statistic:  1848 on 3 and 764 DF,  p-value: < 2.2e-16
```

The coefficients or multipliers describe the size of the effect the independent variables have on your dependent variable *Y* (cooling load), and the constant, -16.136 is the value *Y* is predicted to have when all the independent variables are equal to zero.

This tells us that for every single unit increase in glazing area, our cooling load increases by 14.818 units. If height increases by 1 unit, cooling load increases by 4.576 units. And finally, as wall area increases by 1 unit, the cooling load increases by 0.0415 units.

The prediction equation makes intuitive sense. Glass (glazing) is an extremely inefficient construction parameter and traps heat during periods of warm weather. This will naturally

increase the cooling load of a building. Furthermore, increased height and wall area will increase the surface area for cooling thus increasing the cooling load.

Model Specification and Significance

Adjusted R-squared: The adjusted R2 is 0.8784 .This means that 88% of the variance in the dependent variable, cooling load, is explained by our inputs.

F-test: The F-test statistic of 1848 is larger than the critical value of 3.807333 so we then reject H0 and determine that at least one x variable is a significant determinant of y.

P-values: The p-values are highly significant and close to 0 so we can be certain that the model is well-specified.

Takeaway

When one or more of the model assumptions underlying the linear model is violated, we can no longer believe our inferential procedures e.g. our confidence intervals and p-values may no longer be reliable.

A larger sample could probably introduce more variation across the residual data points, eliminating some of the predictability we see in the scatterplots. However, apart from this, I would suggest the following to eliminate some of the assumption violations evident in this sample:

1. In order to correct possible non-linearity and dispersion issues across the models, I would transform the y and/or x variables using sqrt() and log() functions to see if this helped the model conform to the assumption of homoskedasticity.
2. I would also convert height, orientation, glazing area and glazing area distribution variables into categorical variables (factors) as they seem to have a finite number of values. This is likely the variables that are affecting the randomness of the plotted residuals, creating a visible pattern on the residual scatterplots. By creating different subgroups within our sample, this may help us to satisfy more of our regression assumptions and understand if there is a relationship between different building subtypes within our sample. However, this approach would require a much larger sample than our current 768 observations.