Concrete Compressive Strength Analysis

STAT 4355.001 Applied Linear Models

Geodudes (Team 3)

Gabrielle Allin, Ramesh Kanakala, Josh Yoo

Table of Contents

1. Introduction

By observing which variables have the greatest impact in predicting concrete strength, contractors, concrete suppliers, and even scientists among a multitude of other people can efficiently make longer lasting concrete. Concrete is a fundamental aspect of any city used in basic foundations, buildings, roads, pavements, and more. Being used since the Romans, concrete is something that is ever improving. With the help of strength prediction models, models can contribute to many stakeholders involved in concrete production by accelerating production and use it to create high performance material before spending too much capital.

1.1 Data Description

This dataset was taken from kaggle but originally from UCI Machine Learning Repository: https://www.kaggle.com/maajdl/yeh-concret-data. It has 1030 instances available with 9 quantitative data types in the data set. It provides 8 predictor variables that determine the strength of concrete include:

- Cement (measured in kg/m^3): a binder, a substance used for construction that sets, hardens, and adheres to other materials to bind them together.
- Slag (measured in kg/m^3): the glass-like by-product left over after a desired metal has been separated from its raw ore and is considered as a binder along with fly ash.
- Fly Ash (measured in kg/m^3): a byproduct from burning pulverized coal in electric power generating plants and is considered as a binder along with slag.
- Water (measured in kg/m^3): is utilized to mix with cement for preparation.
- Superplasticizer (measured in kg/m^3): ensures better flow properties since this minimizes particle segregation allowing to decrease water-cement ratio which contributes to the increase of compressive strength.
- Coarse Aggregate (measured in kg/m^3): are any particles greater than 0.19 inch, but generally range between 3/8 and 1.5 inches in diameter.
- Fine Aggregate (measured in kg/m^3): generally consist of natural sand or crushed stone with most particles passing through a 3/8-inch sieve.
- Age (measured in days): the age of the cement; concrete hardens with time in which increases strength.

With the Concrete Compressive Strength being the Response Variable:

- csMPa (measured in MPa): Concrete Compressive Strength

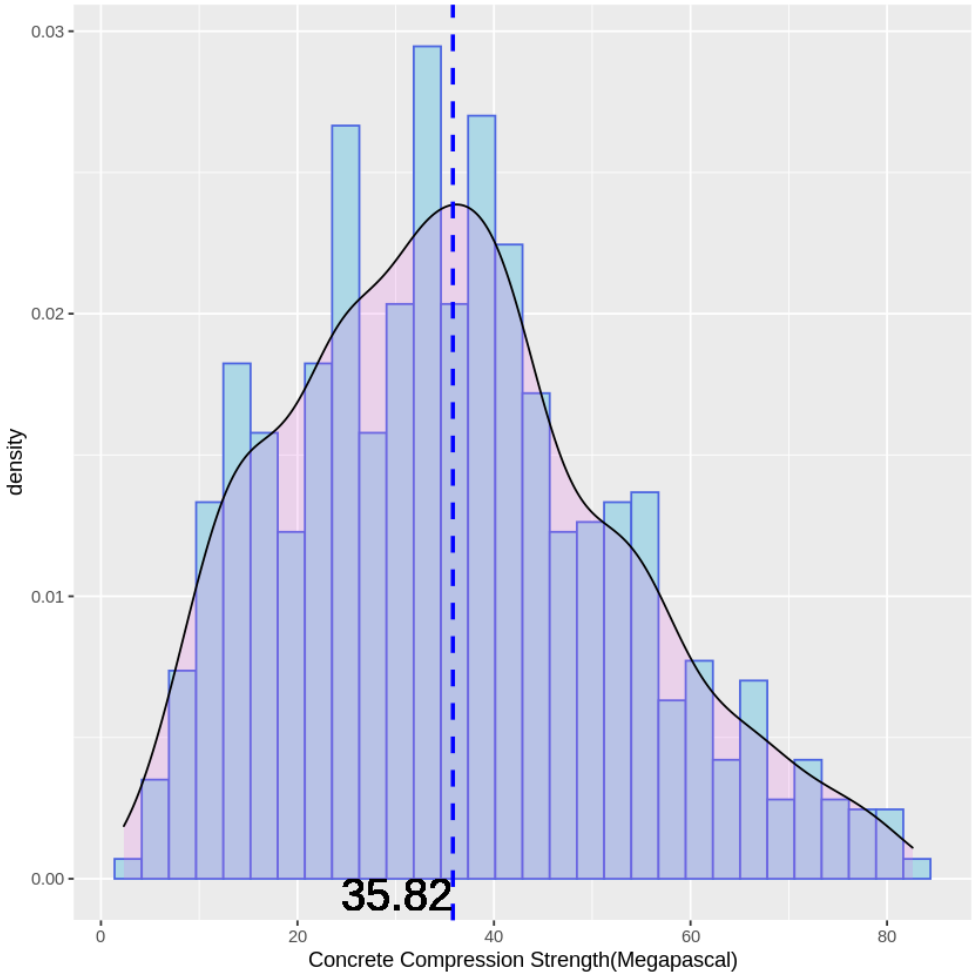
1.2 Analysis Goal

The goal of this analysis is to analyze which parameters influence concrete strength prediction more than others and what transformations are needed to model the data.

1.3 Data Exploration

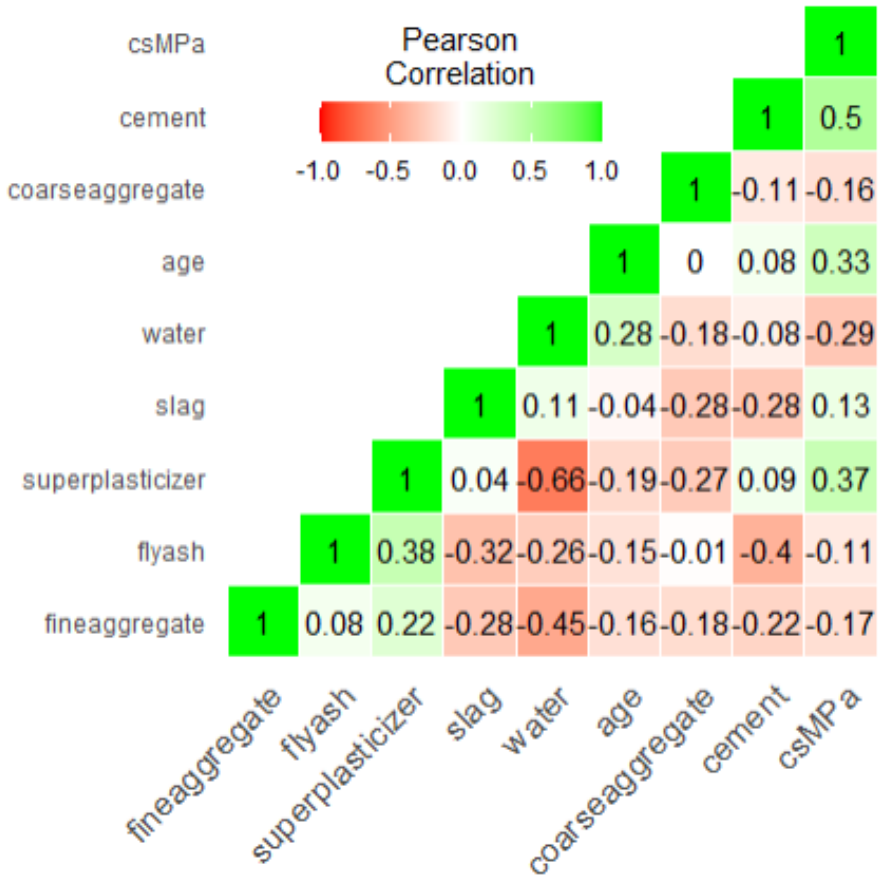
1.3.1 Histogram

With the histogram, there is a mean-response value of 35.82. Likewise, there is a right skew in the histogram but nonetheless the data appears to still be well-distributed.



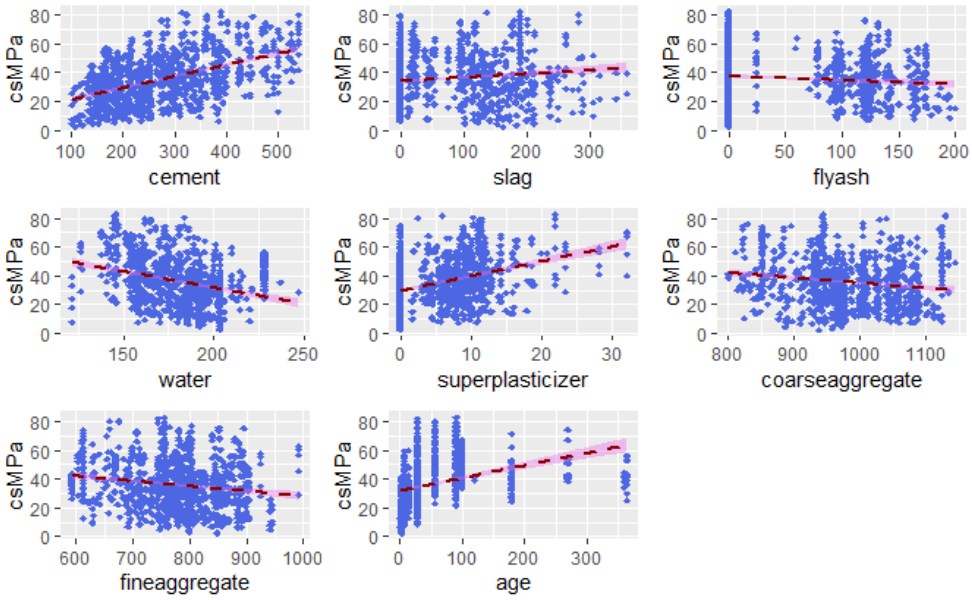
1.3.2 Pearson Correlation Heat Map

With the Pearson Correlation Heat Map, we can observe that there is a positive correlation between Compressive Strength (csMPA) and Cement which is valid because the increasing amount of cement in preparing concrete leads to stronger concrete. Other positive correlations in accordance with the Compressive Strength (csMPA) include: Super Plasticizer, Fly Ash and Age. These correlations aid with understanding how the predictor variables affect the response variable. However, there is a negative correlation presented with water.



1.3.3 Scatterplot

With the scatterplot, we can observe the different linear relations with the 8 predictor variables and the response variable. There are positive linear relations for Cement, Slag Superplasticizer and Age. However, there are negative linear relations for Water, Fly Ash, Fine Aggregate and Coarse Aggregate.



2. Data Analysis

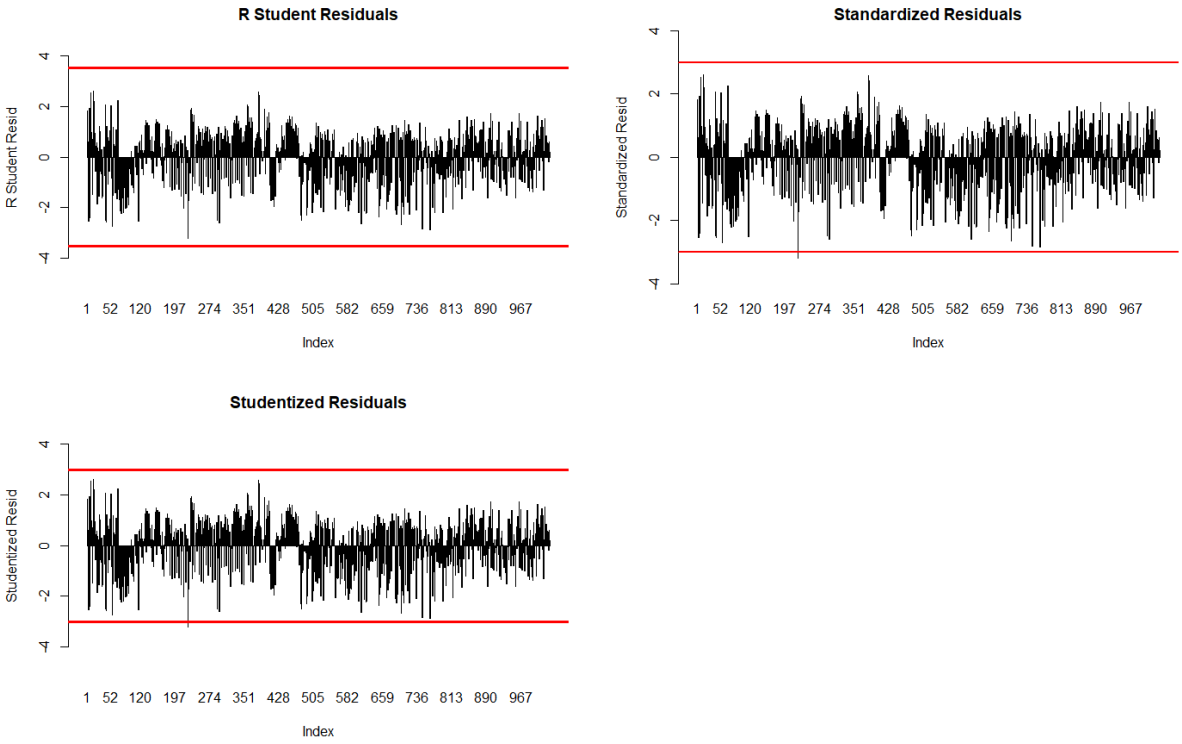
2.1 Full Model Building

Now that the basic understanding of the dataset is known, the next task to tackle was building the full fit linear model. This would allow us to have a greater insight of our variables and see which ones contribute to the response variable using a statistical test.

```
Call:
lm(formula = csMPa ~ ., data = conc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8749 -0.6117  0.1697  0.6539  2.3891

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9744083  2.3441309   0.416 0.677732
cement            0.0102014  0.0007485  13.629  < 2e-16 ***
slag              0.0086351  0.0008937   9.662  < 2e-16 ***
flyash            0.0080147  0.0011095   7.224 9.89e-13 ***
water            -0.0120092  0.0035425  -3.390 0.000726 ***
superplasticizer  0.0256855  0.0082375   3.118 0.001871 **
coarseaggregate   0.0014051  0.0008281   1.697 0.090064 .
fineaggregate     0.0014005  0.0009436   1.484 0.138054
age               0.0101299  0.0004785  21.169  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9169 on 1021 degrees of freedom
Multiple R-squared:  0.6025,    Adjusted R-squared:  0.5994
F-statistic: 193.4 on 8 and 1021 DF,  p-value: < 2.2e-16
```

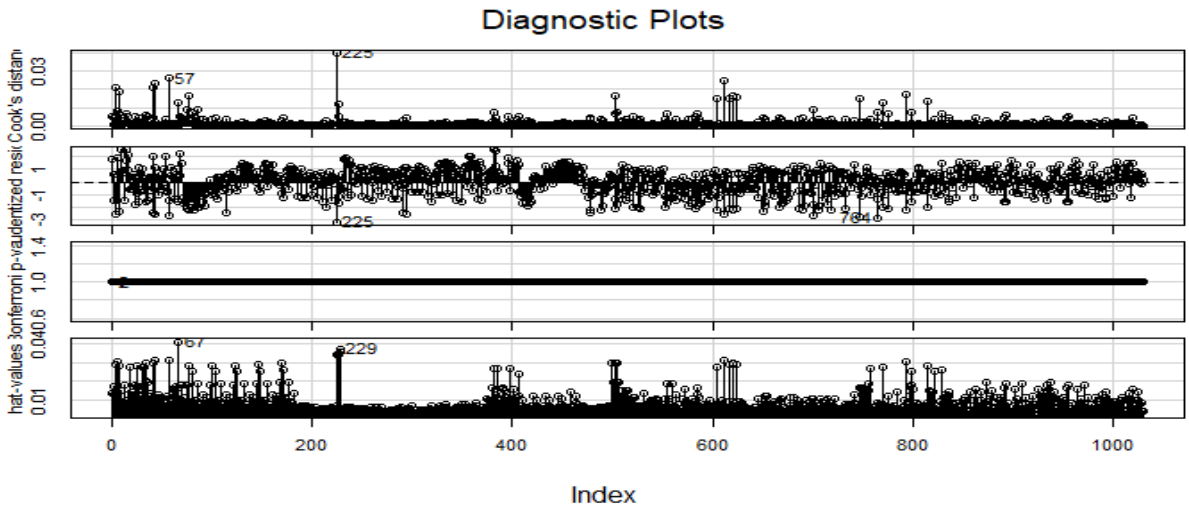
From the table above, it is clear that when considering a significance level of 0.001, superplasticizer, coarseaggregate, and fineaggregate are not seen as significant. This will be important later when the reduced model is explored, but for now, we focus only on the full fit linear model.

Using the model, we take a look at R student, Standardized, and Studentized Residual graphs. Here, we can see that all of the points fall within the boundaries(+-3) except for 225. This point gets close, or sometimes beyond the line on multiple occasions. It is important to note this, as it is clearly an outlier and will likely show up again in future analytical graphs.



2.1.2.1 R student, Standardized, and Studentized Residual graphs

When looking at the Diagnostic Plots, and in cook's Distance in particular, 225 stands out among the rest as the greatest point. This clearly reflects the insight that was given from the previous graphs, and this further provides merit to the idea that 225 is an outlier.



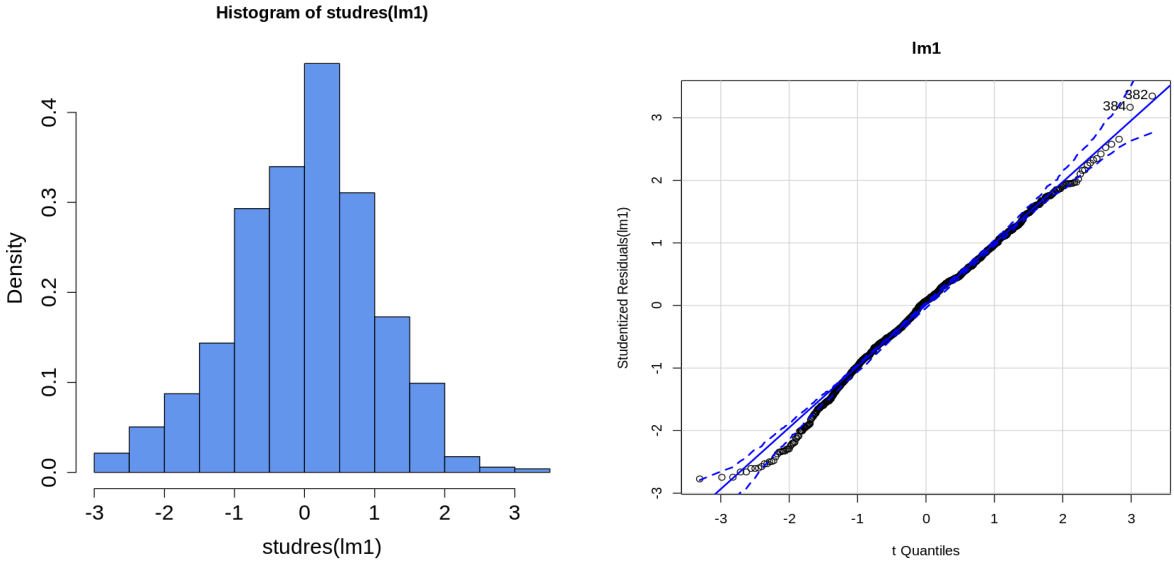
When looking at the variance inflation factors, we can see that the values are lower than 10 for each of the variables. This tells us that none of the x variables are statistically correlated with each other. In other words, our data is good and needs no further changes regarding this test.

cement	slag	flyash
7.488944	7.276963	6.170634
water	superplasticizer	coarseaggregate
7.003957	2.963776	5.074617
fineaggregate	age	
7.005081	1.118367	

2.1.2.2 Diagnostic Plots

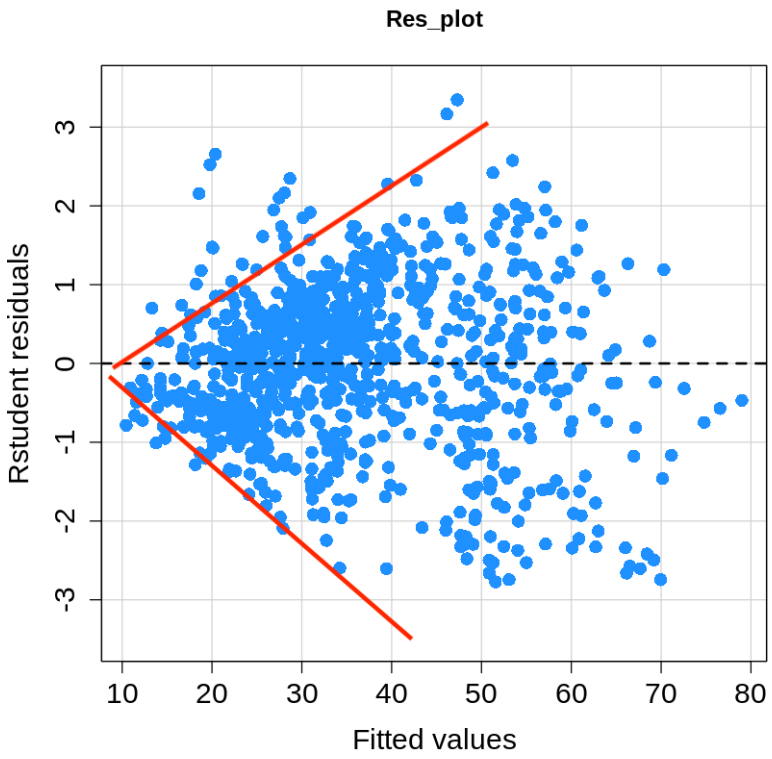
Now we take a look at the residual histogram, qqplot, and residual plot.

From this graph, it is clear that the residuals are nicely spread in a normal distribution, and the qq plot reflects the same idea. There are slight deviations from the linear line along the start and beginning, but overall, it shows a nicely even distribution.



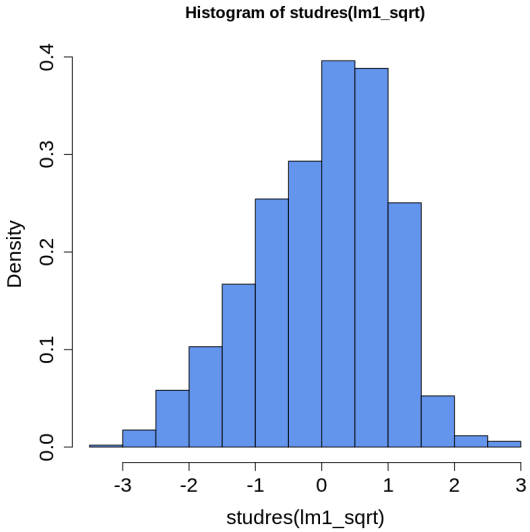
2.1.2.3 Residual Histogram & QQPlot

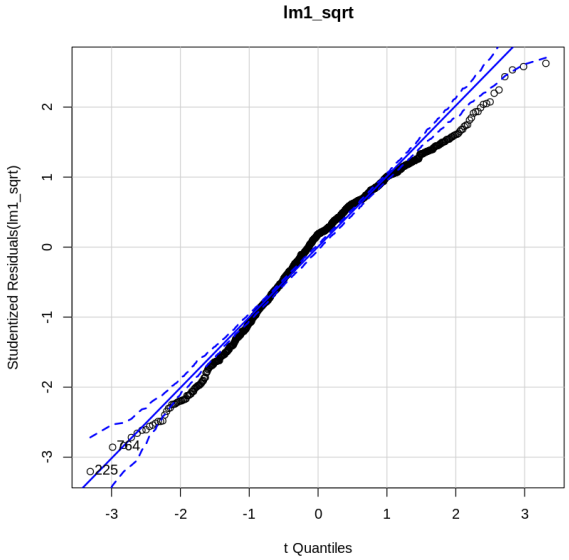
We now come to the residual plot, and it shows something very interesting. There is a gathering of points on the 0 line on the left side, and it starts to gradually spread as the fitted values increase, creating a cone pattern on the plot. We drew a couple lines to make it easier to see. Because this cone shape is so apparent, we can assess that a square root transformation is necessary. So we did the transformation and created the graphs again to see the differences.



2.1.2.3 Residual Plot

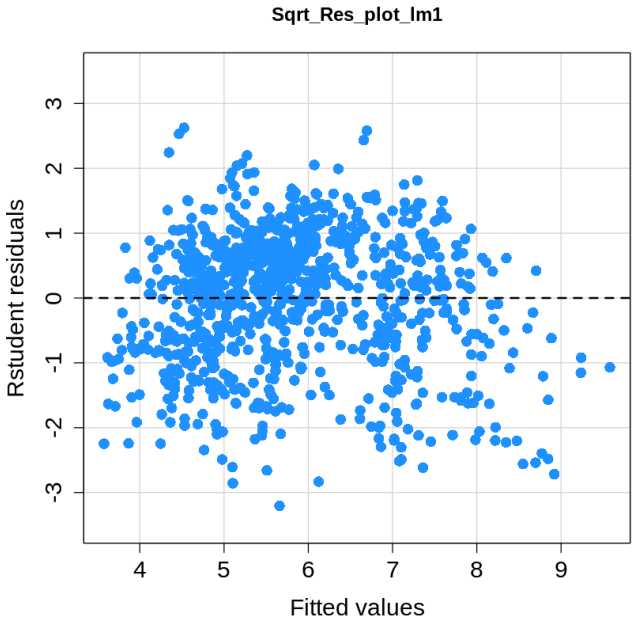
The new residual histogram is showing a strong left skew now, and the qqplot reflects it also as apparent by the bowing in the middle and the two legs of the graph sticking out the opposite direction from it. We can also note that the point 225 sticks out further than any point in this new qqplot, so it confirms the previous graphs that this point is an outlier.





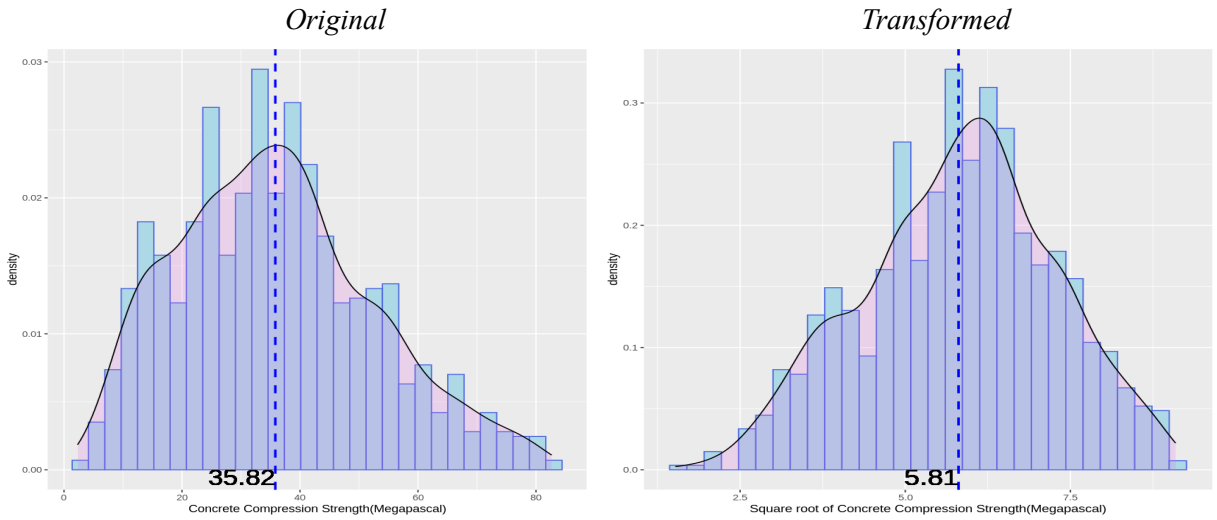
2.1.2.5 Transformed Residual Histogram & QQPlot

The new residual plot below shows great results from the transformation as there is no longer a cone shape in the plot. There is, however, still a slight clutter of points between 4 and 6, but the points are much better distributed than before. Because of this, we can confidently say that the square root transformation has provided us a much more stable and reliable model.



2.1.2.6 Transformed Residual Plot

2.1.3 Density Graph Comparison



Here, we take a look at the density graph of the original dataset compared to the square root transformed dataset. Just from observation, the difference is noticeable. The graph changed from a strong right skew to a slightly less pronounced left skew. This alone projects that the new data has a better distribution, but we will hold our conclusion for now.

The original data set has a mean of 35.82, while the transformed has a mean of 5.81. We compare these means to their respected medians to see if the transformed density graph has a better distribution using numbers. We do this comparison by subtracting the quotient of the mean and median from 1 then multiplying by 100. Doing this calculation, we get 3.986% for the original, and 1.056% for the transformed. Clearly, the mean is more than three times closer to the median in the transformed density graph than the original, meaning the data is much better distributed in the former than the latter. This further proves the usefulness of the transformation.

2.2 Reduced Model Building

Selecting variables for the reduced model was somewhat difficult as most of the time, specific variables are removed as they are not very significant and the model is generally expected to improve; in our case, all of the variables were quite useful in that they were all reasonably correlated and had something to offer in predicting concrete compressive strength. Below, you can observe that forward, backward, and stepwise selection actually ended up recommending the original full model:

```
[1] "Stepwise:"

Call:
lm(formula = csMPa ~ cement + slag + flyash + water + superplasticizer +
    coarseaggregate + fineaggregate + age, data = conc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8749 -0.6117  0.1697  0.6539  2.3891

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.9744083  2.3441309   0.416 0.677732
cement         0.0102014  0.0007485  13.629  < 2e-16 ***
slag           0.0086351  0.0008937   9.662  < 2e-16 ***
flyash         0.0080147  0.0011095   7.224 9.89e-13 ***
water         -0.0120092  0.0035425  -3.390 0.000726 ***
superplasticizer  0.0256855  0.0082375   3.118 0.001871 **
coarseaggregate   0.0014051  0.0008281   1.697 0.090064 .
fineaggregate     0.0014005  0.0009436   1.484 0.138054
age               0.0101299  0.0004785  21.169  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9169 on 1021 degrees of freedom
Multiple R-squared:  0.6025,    Adjusted R-squared:  0.5994
F-statistic: 193.4 on 8 and 1021 DF,  p-value: < 2.2e-16
```

```
[1] "Forward:"

Call:
lm(formula = csMPa ~ cement + slag + flyash + water + superplasticizer +
    coarseaggregate + fineaggregate + age, data = conc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8749 -0.6117  0.1697  0.6539  2.3891

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9744083  2.3441309   0.416 0.677732
cement            0.0102014  0.0007485  13.629  < 2e-16 ***
slag              0.0086351  0.0008937   9.662  < 2e-16 ***
flyash            0.0080147  0.0011095   7.224 9.89e-13 ***
water            -0.0120092  0.0035425  -3.390 0.000726 ***
superplasticizer  0.0256855  0.0082375   3.118 0.001871 **
coarseaggregate   0.0014051  0.0008281   1.697 0.090064 .
fineaggregate     0.0014005  0.0009436   1.484 0.138054
age               0.0101299  0.0004785  21.169  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9169 on 1021 degrees of freedom
Multiple R-squared:  0.6025,    Adjusted R-squared:  0.5994
F-statistic: 193.4 on 8 and 1021 DF,  p-value: < 2.2e-16
```

```
[1] "Backward:"

Call:
lm(formula = csMPa ~ cement + slag + flyash + water + superplasticizer +
    coarseaggregate + fineaggregate + age, data = conc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8749 -0.6117  0.1697  0.6539  2.3891

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.9744083  2.3441309   0.416 0.677732
cement            0.0102014  0.0007485  13.629  < 2e-16 ***
slag              0.0086351  0.0008937   9.662  < 2e-16 ***
flyash            0.0080147  0.0011095   7.224 9.89e-13 ***
water            -0.0120092  0.0035425  -3.390 0.000726 ***
superplasticizer  0.0256855  0.0082375   3.118 0.001871 **
coarseaggregate   0.0014051  0.0008281   1.697 0.090064 .
fineaggregate     0.0014005  0.0009436   1.484 0.138054
age               0.0101299  0.0004785  21.169  < 2e-16 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9169 on 1021 degrees of freedom
Multiple R-squared:  0.6025,    Adjusted R-squared:  0.5994
F-statistic: 193.4 on 8 and 1021 DF,  p-value: < 2.2e-16
```

The ANOVA table confirms that these models are identical:

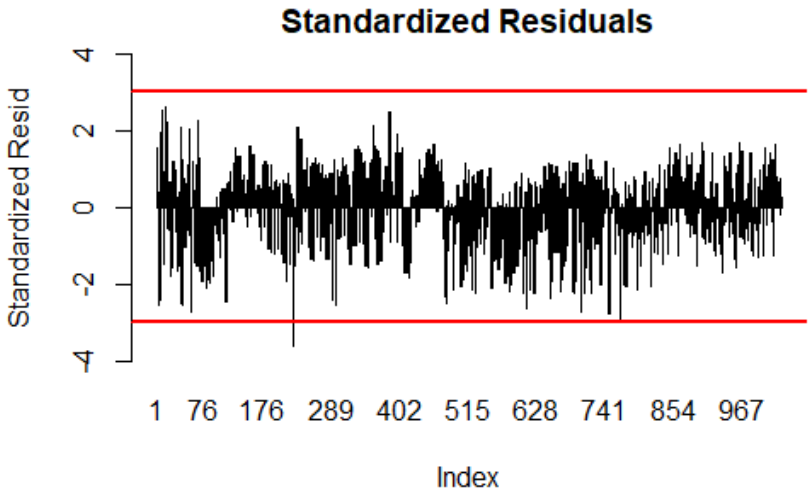
```
Analysis of Variance Table

Model 1: csMPa ~ cement + slag + flyash + water + superplasticizer + coarseaggregate +
    fineaggregate + age
Model 2: csMPa ~ cement + slag + flyash + water + superplasticizer + coarseaggregate +
    fineaggregate + age
Model 3: csMPa ~ cement + slag + flyash + water + superplasticizer + coarseaggregate +
    fineaggregate + age
Model 4: csMPa ~ cement + slag + flyash + water + superplasticizer + coarseaggregate +
    fineaggregate + age
  Res.Df    RSS Df Sum of Sq F Pr(>F)
1   1021 858.41
2   1021 858.41  0         0
3   1021 858.41  0         0
4   1021 858.41  0         0
```

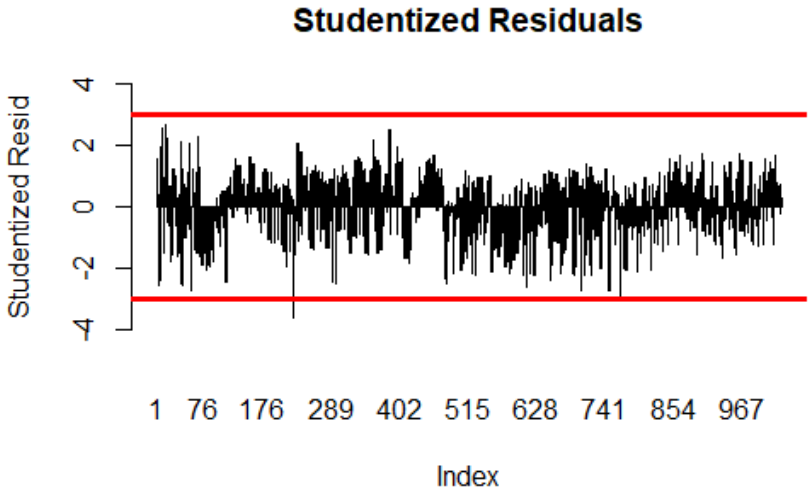
So, in the end, we decided to vary our variables by selecting only those that were significant at the 0.001 level: cement, slag, fly ash, water, and age:

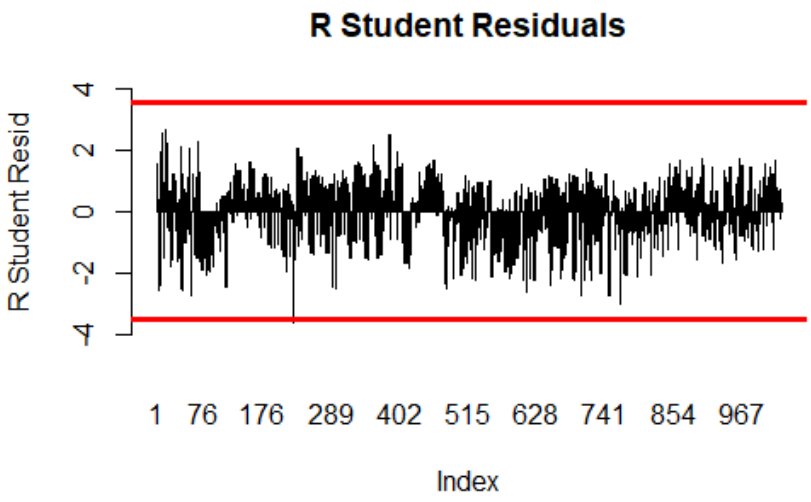
```
Call:
lm(formula = csMPa ~ cement + slag + flyash + water + age, data = conc)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3017 -0.6361  0.1689  0.6652  2.4050

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.3368896  0.3253538   16.40   <2e-16 ***
cement       0.0095799  0.0003467   27.63   <2e-16 ***
slag         0.0079010  0.0003998   19.77   <2e-16 ***
flyash       0.0076090  0.0005928   12.84   <2e-16 ***
water       -0.0202804  0.0014741  -13.76   <2e-16 ***
age          0.0101318  0.0004779   21.20   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9204 on 1024 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5963
F-statistic:   305 on 5 and 1024 DF,  p-value: < 2.2e-16

[1] "Variance:  0.847209052557485"
```

We see that the adjusted R^2 value decreased just a little from 0.5994 to 0.5963. This very minimal difference indicates that the variables superplasticizer, coarse aggregate, and fine aggregate did not offer much to the full model and the R^2 was only barely higher than because of the greater number of variables. We also see the variance increase slightly from 0.841 to 0.847 as well as the RSE from 0.9169 to 0.9204.

Moving on to the residual analysis of the reduced model, we can see a few changes:





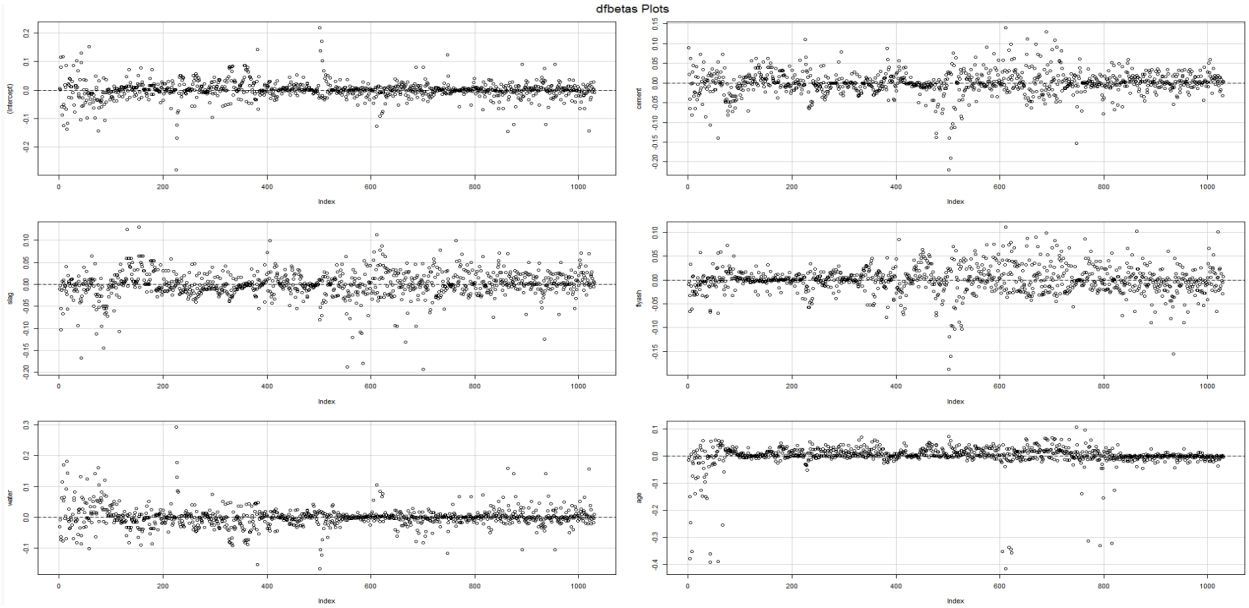


2.2.2.1 RStudent, Standardized and Studentized Residual Graphs

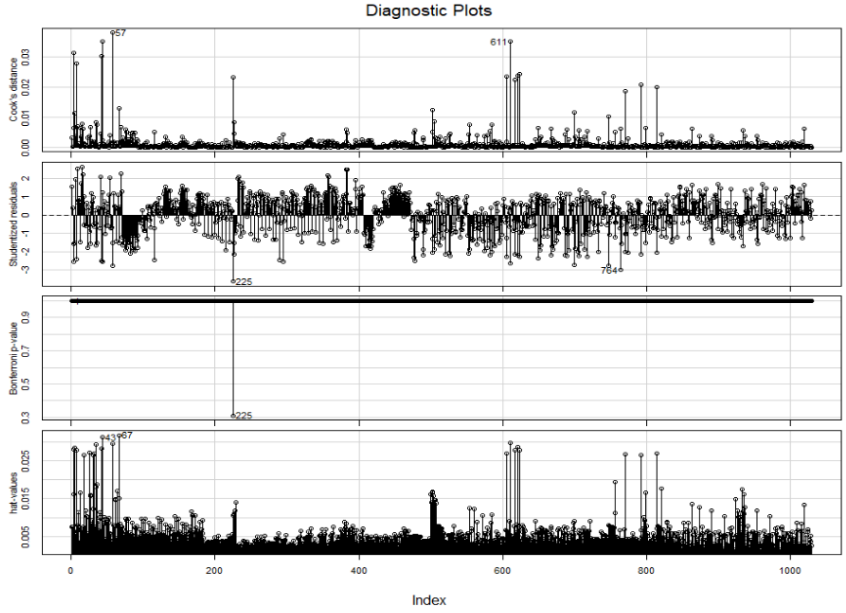
Most variables stay within the cutoff limits but points 225 seems to stick out more than the other variables, passing even the R student residuals plots here. Essentially, the residuals' distribution is the same but they are even larger due to there being less variables. Looking closer at point 225, besides having a somewhat low compressive strength, it is not abnormal in any way so there are no worries. The variance inflation factor values are below 5, actually very close to 1, for all the variables used here which means the variables are not correlated and there is no evidence of multicollinearity:

```
  cement     slag   flyash    water      age
1.594164 1.444858 1.748086 1.203434 1.106834
[1] 225 764
```

We see the point 225 pop up again in the influential analysis as well as point 764. Let's look at how they appear in the dfbetas and diagnostic plots:

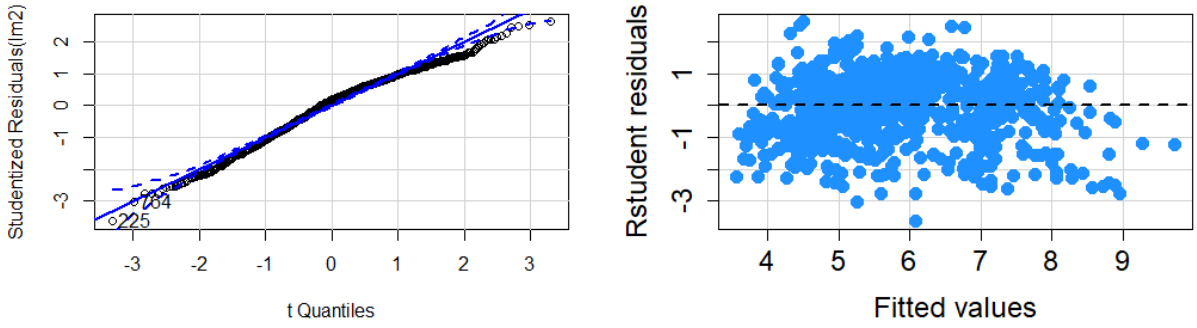


2.2.2.2 DFBETAS Plots



2.2.2.3 Diagnostic Plots

The points in the influential analysis plots mostly hover around the 0 value and all points seem to be in the cutoffs. In the diagnostic plots, Many of the hat-value leverage points at the lower indices have been reduced though point 43 and 67 stick out. Points 229 in the same graph and 225 in Cook's distance have been reduced. Point 611 sticks out for both full and reduced models and point 225 is stuck out greatly in the, again, studentized residuals (as well as 764 here) and the bonferroni p-value plots. Looking directly at these observations, only compressive strength seems low but nothing is really out of the ordinary. Now let's look at the normal probability plot and residuals vs fitted values plot:



2.2.2.4 Normal Probability Plot & Residuals vs. Fitted Values Plot

We can see here with the normal probability plots that, just like with the original, points 382 and 384 stick out above with the untransformed and points 225 and 764 stick out below for the transformed; this means the positively skewed data has been transformed to a more normal plot with perhaps just a little left-skewing. Again, in the residuals vs fitted plot, we see the change from the cone shaped plot of the original residuals, where the variance of errors is not constant to the transformed data, where the residuals are more randomly distributed.

2.3 Model Comparison

Comparing the transformed full and reduced model with the ANOVA table, we see that RSS is higher for model 2:

```
Analysis of Variance Table

Model 1: csMPa ~ cement + slag + flyash + water +
superplasticizer + coarseaggregate +
    fineaggregate + age
Model 2: csMPa ~ cement + slag + flyash + water + age
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1021 858.41
2   1024 867.54 -3   -9.1322 3.6206 0.0128 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This means that the full model outperformed the reduced model. Here is a simple table that allows us to compare all the models:

Model	*RSE*	*Adjusted R^2*	*Variance*
Full	**10.4**	**0.6125**	**108.142**
Reduced	10.45	0.6091	109.1
Full (Transformed)	**0.9169**	**0.5994**	**0.841**
Reduced (Transformed)	0.9204	0.5963	0.847

For all models, we see that the full versions perform better with a lower RSE and Variance and a higher R^2 value; we prefer the variance stabilized full transformed model due to the relatively higher R^2 and lower variance. Moreover, though the other variables don't offer much to the model, the more variables don't hurt by a large margin and are probably useful as they are direct ingredients of concrete mixture. Something interesting we learned is that we can't exactly assess the differences in R^2 between the original and transformed models since R^2 is the proportion of the variance in the dependent variable that is explained by the variation in the independent variables; transforming compressive strength with square root will have also transformed its variance structures too.

3. Conclusion

To summarize our findings: cement, slag, fly ash, water, and age are the most significant predictors in modeling concrete compressive strength and concrete compressive strength is calculated as a nonlinear function of its components. What this means for the future is that concrete mixers/manufacturers should give more weight to or focus on the cement, slag, fly ash, water, and age ingredients in their mixture. When testing various concrete mixtures the compressive strength should be transformed with some function to improve the distribution of the data. This type of research is significant for the foundations (pun intended) of civil engineering, architecture, and infrastructure as a whole. Stronger concrete means structures that endure the test of time as well as natural disasters or any other complication. For future study, perhaps more aspects of concrete mixture such as sub-qualities of air, cement, water, aggregates, and other minerals could be analyzed. Other transformations should also be explored to perhaps create an even more valuable distribution of the data or other types of regression models could be utilized to improve the predictive power.

Reflecting on our project journey, we actually learned much from our experience. First, organizing meetings, assigning roles, and working together as a team has improved all members' communication and efficiency in a group. Outside these logistical aspects, researching the data helped us learn much about concrete such as the many ingredients that it consists of, in what units these are measured, and how extensive the use of concrete is in our modern world. We were successful in many aspects but did have some difficulty with some parts, such as variable selection. Having many useful variables is great but we learned that studying the loss of certain variables allows us to understand the strength of the others in full models; a reduced model should not always be an improvement but is a chance to evaluate specific variables. Also, working with the data over the duration of the project allowed us to catch the skews and shape of the distribution, allowing us to apply our knowledge of transformations and observe how it can make the data more favorable for modeling. However, a non-linear transformation may not have to be our square-root function; we learned about other transformations that we could use in the future with more knowledge.

This project was a great experience with linear modeling but even more with the analysis of models. We hope to use this experience in the future to evaluate and create better, more *concrete* models.

4. Appendix

4.1 Team Responsibility

Gabrielle Allin: Documentation and oversight of deliverables such as proposal, presentation, and the final report while working alongside team members with data exploration, model building, and analysis. Also responsible for making sure the team is consistent with meeting times.

Ramesh Kanakala: Data exploration, primarily model building, primarily analysis, and contributing equally alongside team members for the proposal, presentation, and report.

Hyun Guk Yoo: Primarily data exploration, primarily model building, analysis, and contributing equally alongside team members for the proposal, presentation, and report.

4.2 References

I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

Maajdl. (2018, June 15). Concrete Strength Regression. Retrieved May 10, 2021, from https://www.kaggle.com/maajdl/yeh-concret-data

4.3 R Notebook attached below: