**Introduction**

For this project, we asked, what ingredients and at what concentrations do you put in a concrete mixture to maximize its strength?

Using the dataset “[Concrete Compressive Strength](https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength)” from I-Cheng Yeh first published in Cement and Concrete Research, Vol. 28, No. 12 in 1998, we used multiple linear regression to model the relationship between several variables (Cement (kg/m3), Blast Furnace Slag (kg/m3), Fly Ash (kg/m3), Water (kg/m3), Superplasticizer (kg/m3), Coarse Aggregate (kg/m3), Fine Aggregate (kg/m3), Age (day)) and the compressive strength of the concrete created using the mixture (Pressure (MPa)).

We chose to model the relationship between these variables using multiple linear regression because in every concrete mixture, all of these predictor variables are important and may have significant effects on the outcome. The ingredients cement, water, coarse aggregate, and fine aggregate are the basic components of concrete. In particular, Abrams’ law in civil engineering states that concrete with a lower water-to-cement ratio tends to be stronger and more durable than concrete with a higher water-to-cement ratio. The strength of concrete depends on the ratio of all of these basic components. Next, blast furnace slag, fly ash, and superplasticizer are optional components that are added to concrete to make it more durable or easy to work with. Lastly, the age variable indicates how old the concrete was when the pressure test was done. Because the creation of strong concrete is a process that involves a multitude of ingredients and the age of concrete is also a factor in its strength, we decided that multiple linear regression was the optimal method to model this relationship.

In this paper, we will discuss how the data is structured, how the original model without transformations looked, how we went about deciding which transformations to use, how we used variable selection and anova tests to find the most optimal model, how we decided which model is the most optimal, and how our model can be applied to the real world and the limitations it holds.

**Data Description**

To see the overview of the data set, we report each summary statistics of each variable, such as mean, standard deviation, and correlations among the variables. We will observe the distribution of each variable and relationships among the variables using appropriate graphs. The following table shows the key summary statistics of the data set.

|  | **Min** | **1st Q** | **Median** | **Mean** | **3rd Q** | **Max** | **Std Dev** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Cement** | 102 | 192.4 | 272.9 | 281.2 | 350 | 540 | 104.51 |
| **Blast Furnace Slag** | 0 | 0 | 22 | 73.9 | 142.9 | 359.4 | 86.28 |
| **Fly Ash** | 0 | 0 | 0 | 54.19 | 118.3 | 200.1 | 63.99 |
| **Water** | 121.8 | 164.9 | 185 | 181.6 | 192 | 247 | 21.35 |
| **Superplasticizer** | 0 | 0 | 6.4 | 6.21 | 10.2 | 32.2 | 5.97 |
| **Coarse Aggregate** | 801 | 932 | 968 | 972.9 | 1029.4 | 1145 | 77.75 |
| **Fine Aggregate** | 594 | 731 | 779.5 | 773.6 | 824 | 992.6 | 80.18 |
| **Age** | 1 | 7 | 28 | 45.66 | 56 | 365 | 63.17 |
| **Strength** | 2.33 | 23.71 | 34.45 | 35.82 | 46.13 | 82.6 | 16.71 |

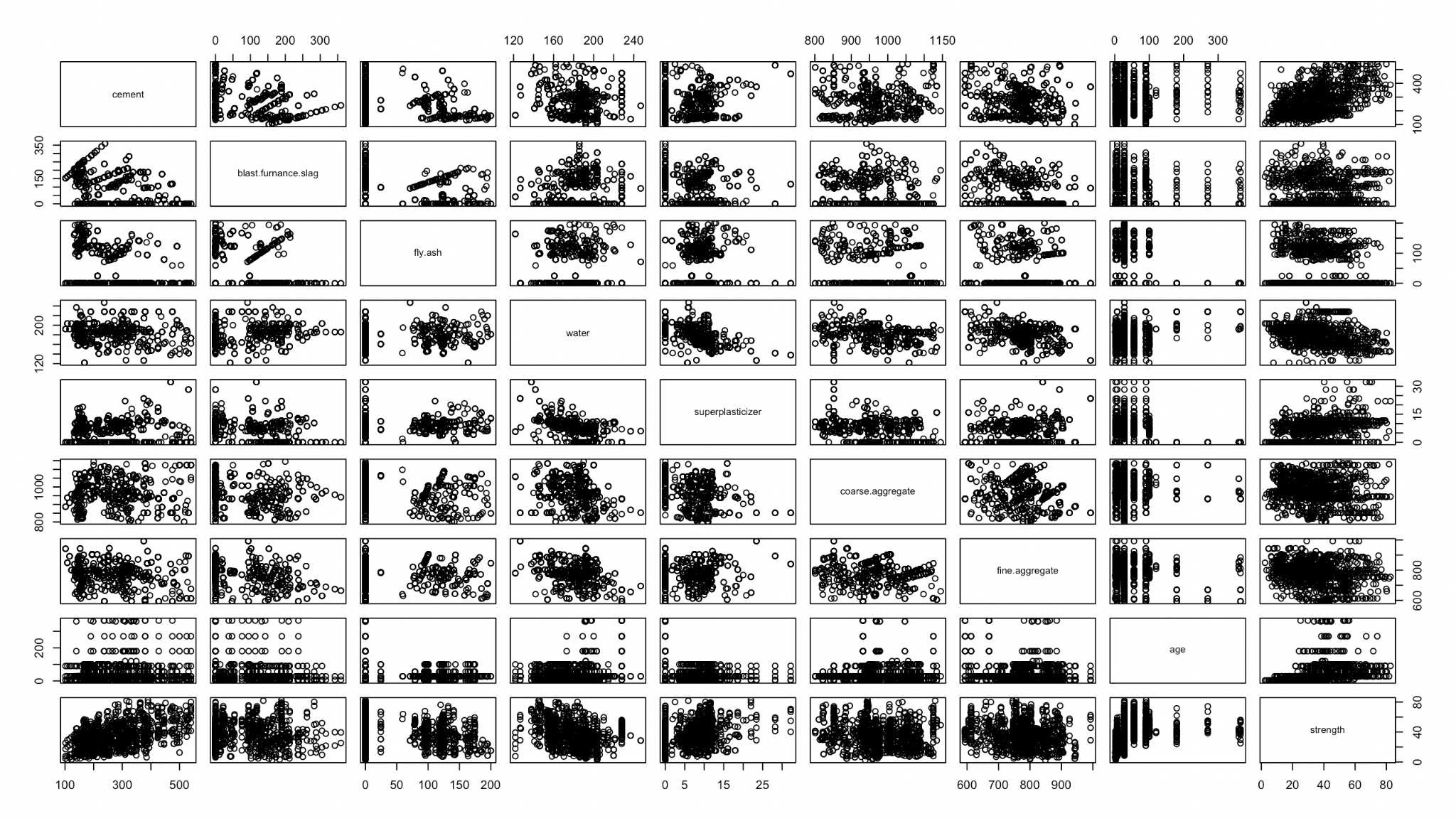
We observe through the distributions of our data that all are either roughly normally distributed or heavily skewed right (see Appendix). Through the following pairs plot, we can examine the correlation between each variable with their distribution.

Figure 1

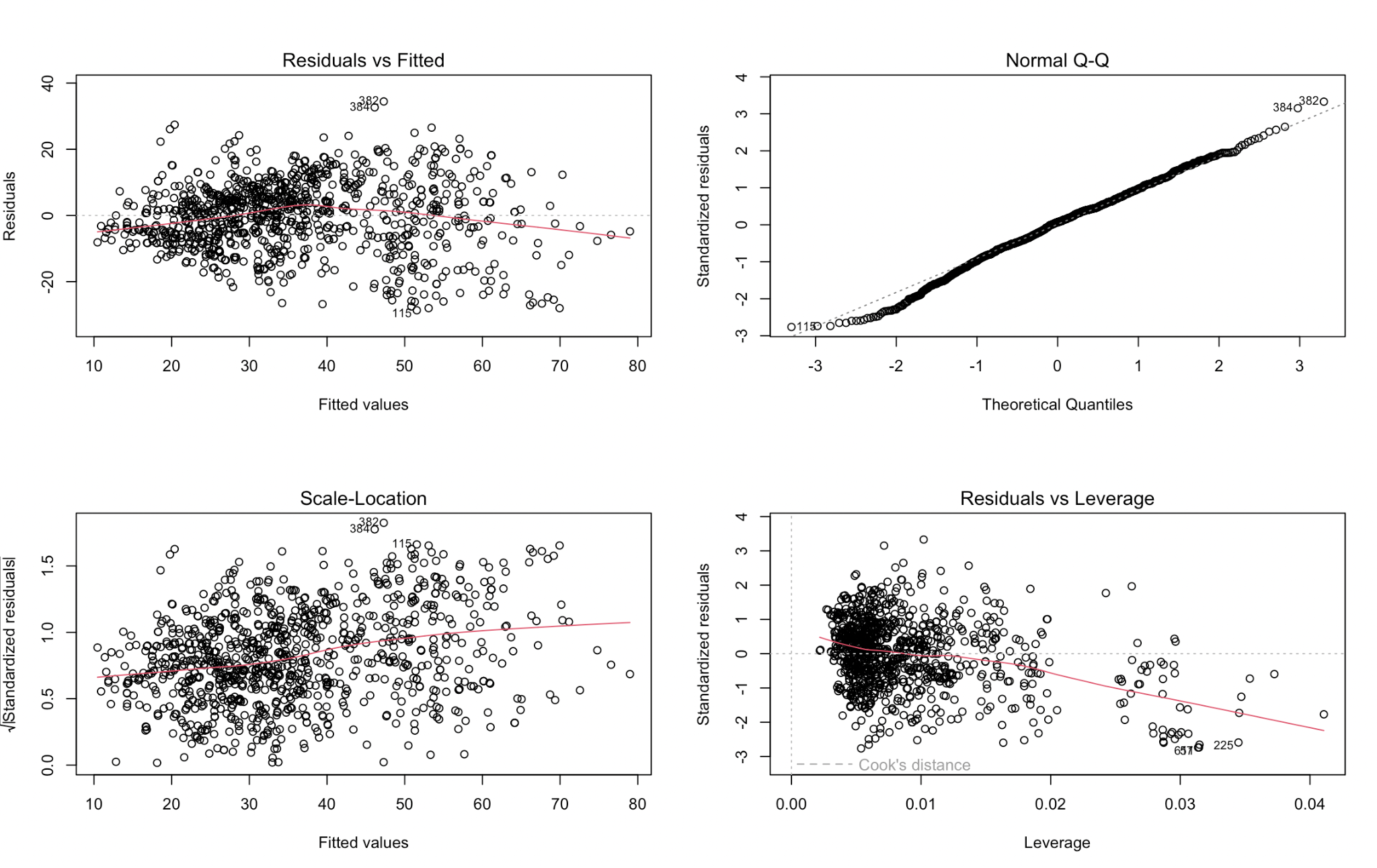
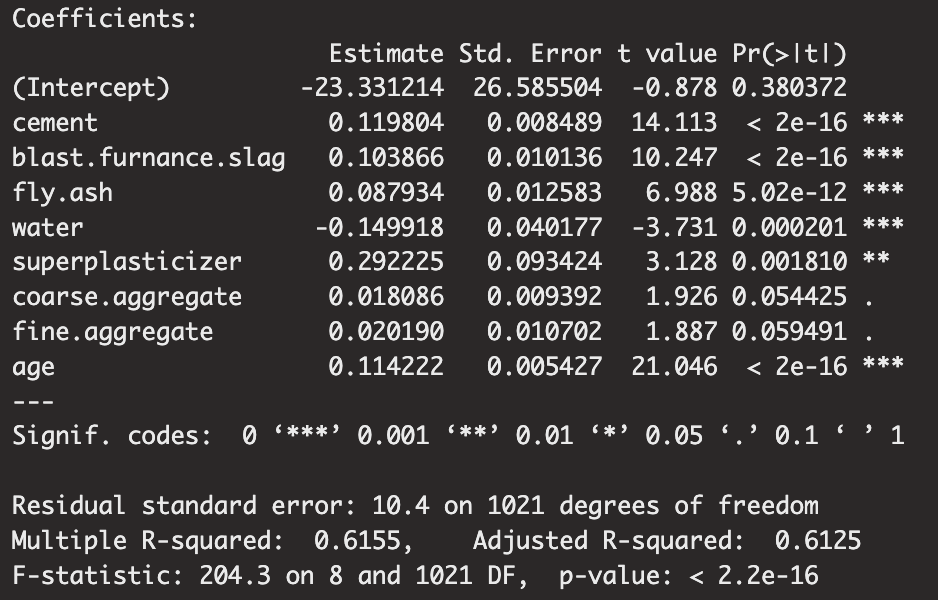


Figure 2 Figure 3

Before transformation, we can note that five out of the eight predictor variables are statistically significant. From the diagnostic plots we can observe that there are bad leverage points like observation 225, 605, and 611; the full list is shown in the appendix. Additionally, the average of error terms is about zero. And, it satisfies the linearity assumption because the residual plot has a random scatter. Additionally, the model supports the normality of error terms because the Q-Q plot looks fairly straight. However, the model violates the constant variance of error terms because the line increases slightly as the number of fitted values increases. Additionally, some standardized residual plots (in Appendix) for individual variables show nonconstant variance (mostly age).

**Results and Interpretation**

**Transformation**

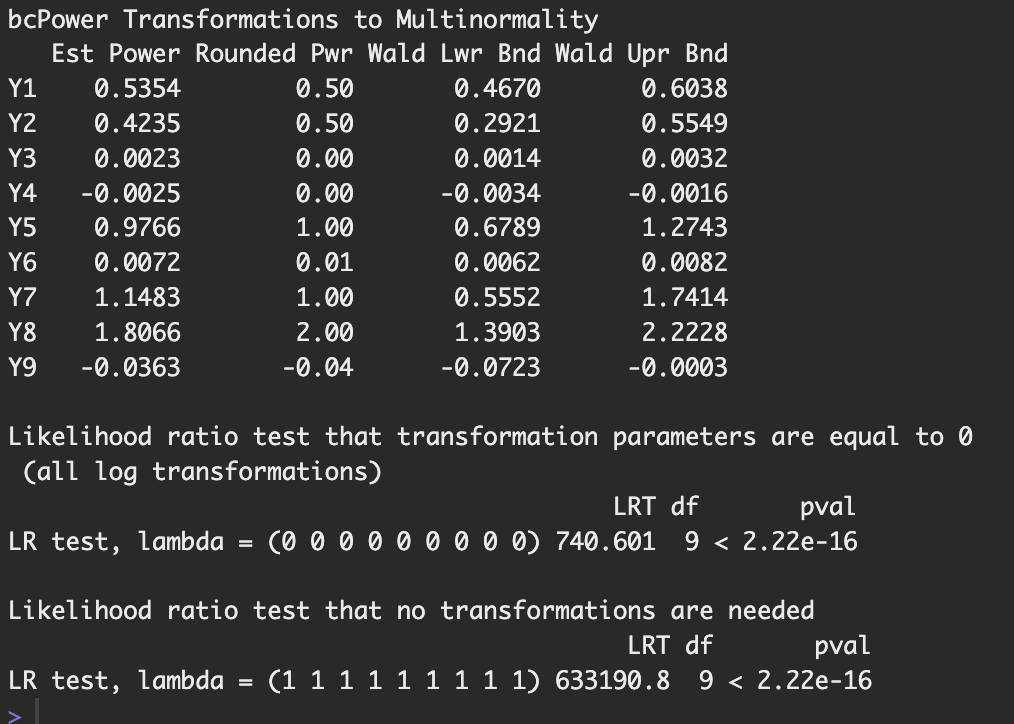
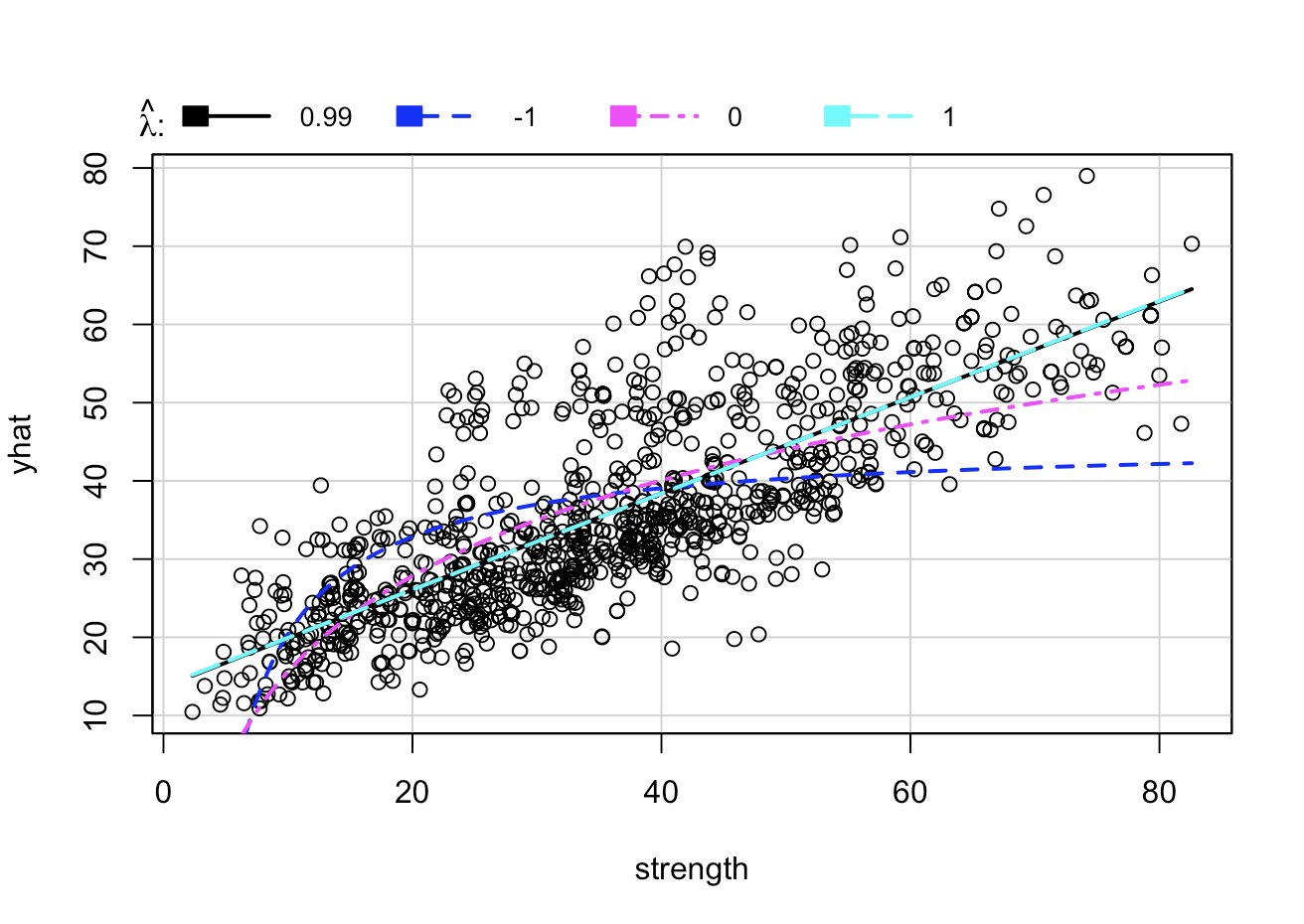


Figure 4 Figure 5

Currently, our value of the model without any transformation is 0.6155, which means that about 61.55% of the variance in concrete strength is accounted for by our model. This is a fairly low value. Because of the low value and the diagnostic plots, we will try to transform the model to find a better fit. We first use the inverse response plot method to see if we should transform the response variable (Figure 4). The inversePlot function suggested a lambda of 0.99, which can be rounded up to 1. Because the recommended transformation is no transformation, we will continue with a different method to find the best fit.

The second method that we utilized was the boxcox method, in which we tested both the responses variable and the predictors. The R output is shown in Figure 5. From the Likelihood Ratio tests we can reject the hypothesis that all lambda values are 0 and the hypothesis that all lambda values are 1.

After transforming all variables according to the rounded power suggestion (Figure 5), the following R output is the summary of our transformed model (Figure 6). The value increased by 0.205 and now all 8 predictors are significant. Similar to the original model, the F-test reveals that the model is significant and now 82% of the variance in concrete strength can be explained by the model. However, 8 predictors may present problems of multicollinearity so we will continue with variable selection.

**Variable Selection**

We first calculate the VIF values of all predictors and find that all of them are less than 5, meaning there is no multicollinearity present. Even though there is no high correlation between predictor variables, to make sure we have the best model, we will complete forward and backward stepwise testing, as well as test all possible subsets. When completing stepwise testing, forward AIC and backward AIC suggested the 8 predictor model while backward BIC suggested 7 predictors. Our final test for variable selection is through testing all possible subsets. After finding the best model for each number of predictors, we calculate the , AIC, AIC corrected, and BIC values (Figure 7).

Observing the and AIC, the best model is the 8 predictor model (highest , and lowest AIC). However, in regard to AIC corrected and BIC, the 7 predictor model fits the best, as it has the lowest AICc and BIC.

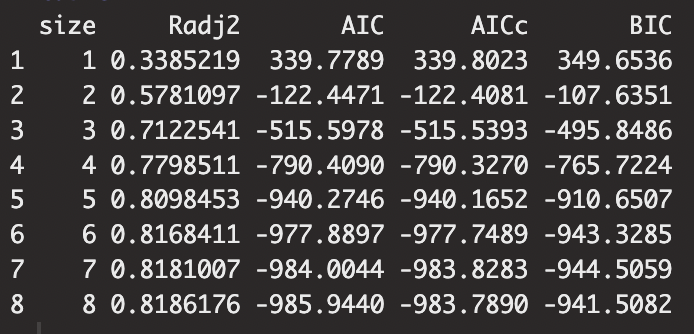
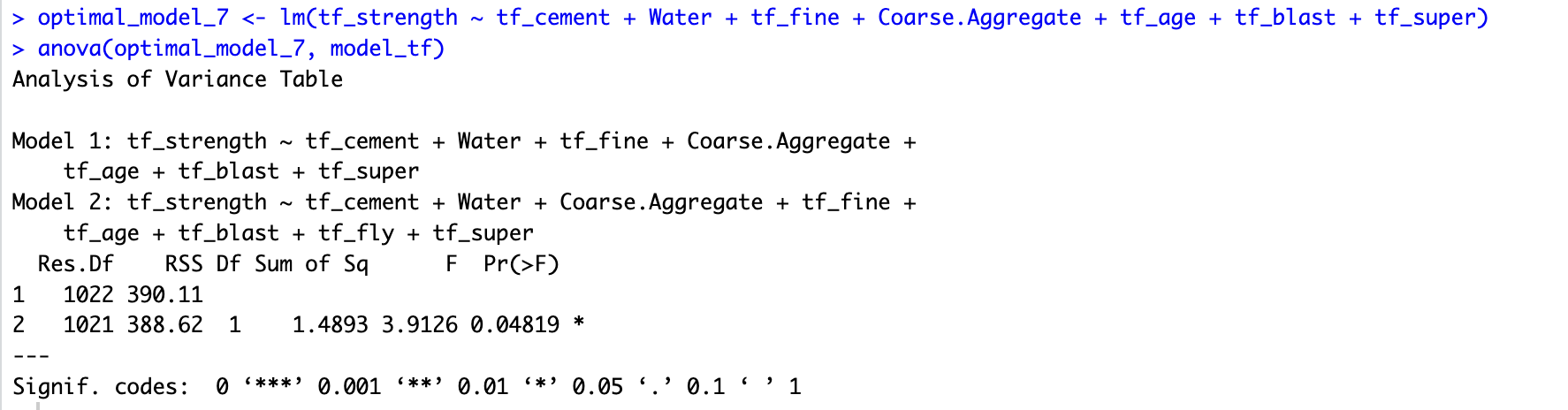


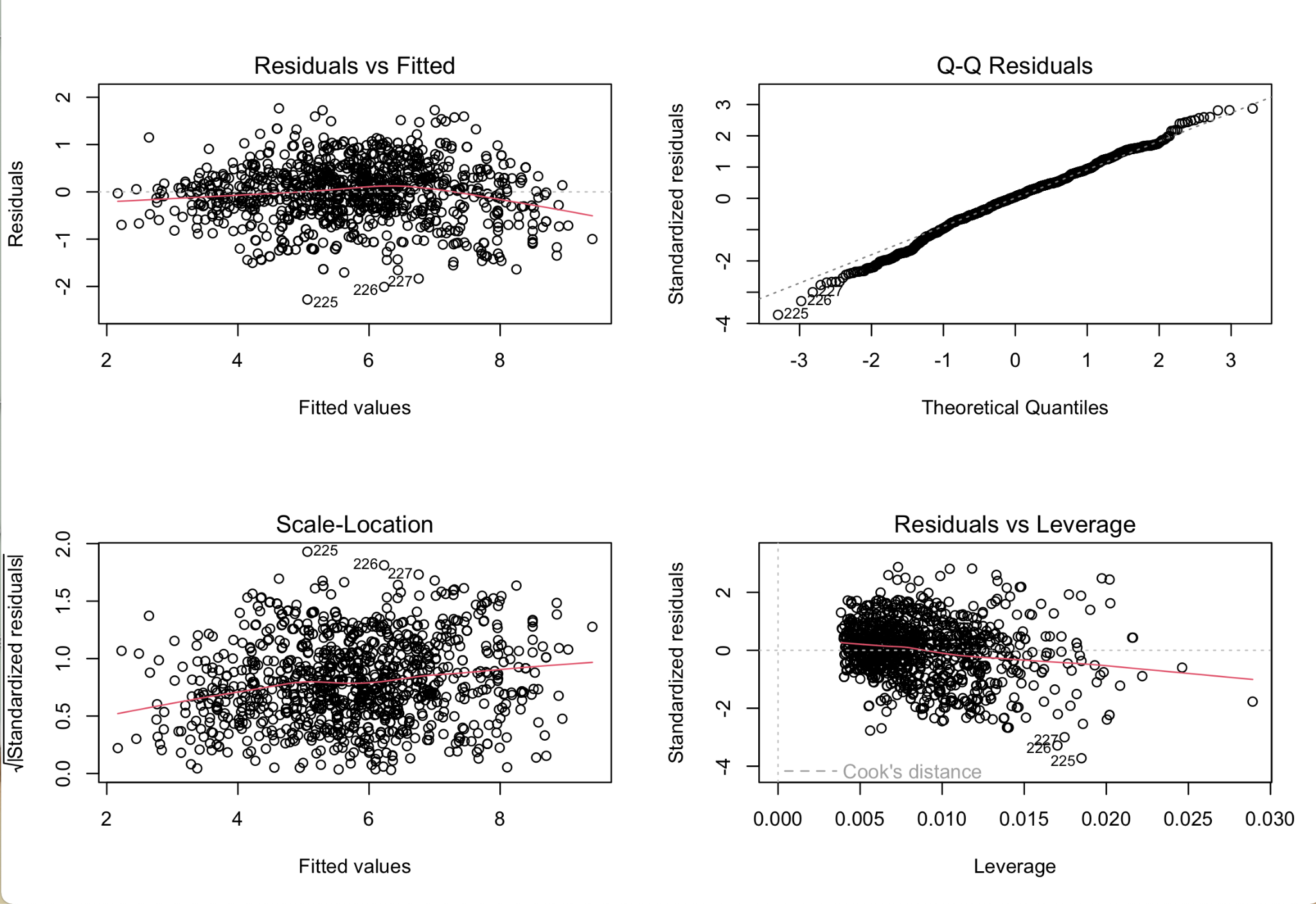
Figure 6 Figure 7

To decide between the models, we check the significance of predictors for both models. However, all predictors are significant in both (see tmodel2 in Appendix), so we move on to conducting a partial F-test between the reduced 7 predictor model and the full 8 predictor model. From our ANOVA results, we observe the p-value is less than 0.05, which means we reject our null hypothesis that the reduced model is a better fit. Thus, our final model remains the full model with all 8 transformed predictors.

Figure 8

**Final Model**

With 8 predictors, the final model is as follows:

Figure 9

The summary output (Figure 6) again reveals that the model and all predictors are significant. Reviewing the final diagnostic plots, we can conclude that the model normality has slightly improved because the normal Q-Q line is a little straighter. The residual and scale-location plots have improved and are a little more scattered, showing constant variance. The individual standardized residuals show overall less variance, especially age (see Appendix), although Blast Furnace Slag, Fly Ash, and Superplasticizier look a bit weird because of how we had to add 1e-100 to log transform them. The residuals vs leverage plot has also improved moderately with less outliers. Additionally, we were also able to remove many bad leverage points. Now, there are only 7 bad leverage points as shown in the appendix. Nevertheless, all of these improvements are very slight, as our original model already showed promising diagnostic plots. Even though the improvements are small, we choose the transformed model because of its greater value and ability to predict our response variable, concrete strength.

**Interpretation**

An increase of 1 kg in square root of concrete leads to a 0.2185 MPa increase in square root of strength.

An increase of 1 kg in log of blast furnace slag leads to 0.002534 MPa increase in square root of strength.

An increase of 1 kg in log of fly ash leads to 0.0005226 MPa decrease in square root of strength.

An increase of 1 kg in log of superplasticizer leads to 0.002608 MPa increase in square root of strength.

An increase in 1 kg of water leads to 0.02128 MPa decrease in square root of strength.

An increase in 1 kg of coarse aggregate leads to 0.001149 MPa decrease in square root of strength.

An increase in 1kg of fine aggregate squared leads to 0.000001764 decrease in square root of strength.

An increase in 1 day of age leads to 0.7598 increase in square root of strength.

**Discussion**

According to our final model, the ideal concrete mixture would maximize the amount of cement added while minimizing the other core components, water, coarse aggregate, and fine aggregate. Additionally, regarding the optional components of a concrete mixture, blast furnace slag and superplasticizer are shown to have a positive effect on concrete strength, whereas fly ash has a negative effect. It is important to note that, however, the increase or decrease in strength brought about by these optional components is only a small effect given how small the coefficients are for these variables in our final model. Lastly, age has a positive effect on strength: an increase in 1 day of age leads to 0.7598 increase in the square root of concrete strength. This result implies that older concrete is stronger.

Examining the coefficients on each of our mixture components, it appears that the amount of cement, water, and coarse aggregate in our mixture has the greatest effect on concrete strength, as they have coefficients of 0.2185, -0.02128, and -0.001149, respectively. These results seem to be compatible with Abrams’ law, which states that concrete strength is inversely related to the concrete mixture’s water-to-cement ratio (Kargari). A ratio of 0.3 to 0.8 is most typical, with a ratio of 0.4 being most desired for high quality concrete (Concrete Countertop Institute). While this ratio is, clearly, not the only factor used to predict concrete strength, it is typically one of the most influential, which is confirmed by our model.

Lastly, something peculiar occurred following the variable transformation that is worth mentioning. In our original model, only water had a negative coefficient, but after transforming, coarse aggregate, fine aggregate, and fly ash all switched signs from positive to negative as well. We did not have any multicollinearity, which is usually the cause of this. We suspect this may be due to the fact that these were the least significant out of all predictors, and since the transformations occurred in a nonlinear fashion, the signs switched because their original relationship wasn’t too strong to begin with.

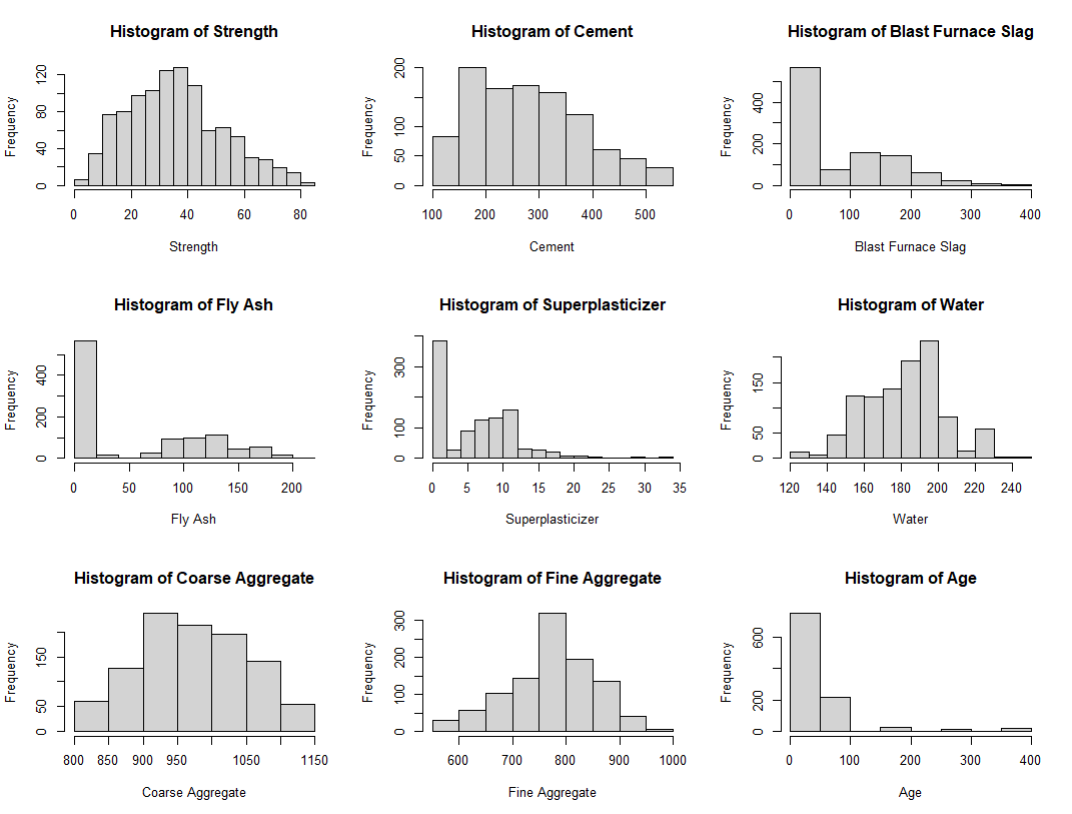
**References**

Concrete Countertop Institute. “The Importance of Water Cement Ratio in Concrete Countertop Mix Design.” *Concrete Countertop Institute*, 6 Dec. 2022, [concretecountertopinstitute.com/free-training/the-importance-of-water-cement-ratio-in-concrete-countertop-mix-design/#:~:text=Typical%20Water%2DCement%20Ratios%20in%20Concrete%20Mixes&text=Normal%20for%20ordinary%20concrete%20](http://concretecountertopinstitute.com/free-training/the-importance-of-water-cement-ratio-in-concrete-countertop-mix-design/#:~:text=Typical%20Water%2DCement%20Ratios%20in%20Concrete%20Mixes&text=Normal%20for%20ordinary%20concrete%20)(sidewalks,quality%20concrete%20is%20desired%3A%200.4.

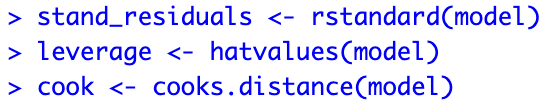
Kargari, A., Eskandari, H. and Kazemi, R. (2018). *Effect of Cement Strength Class on the Generalization of Abrams’ Law*. Structural Concrete (20): 493 – 505.

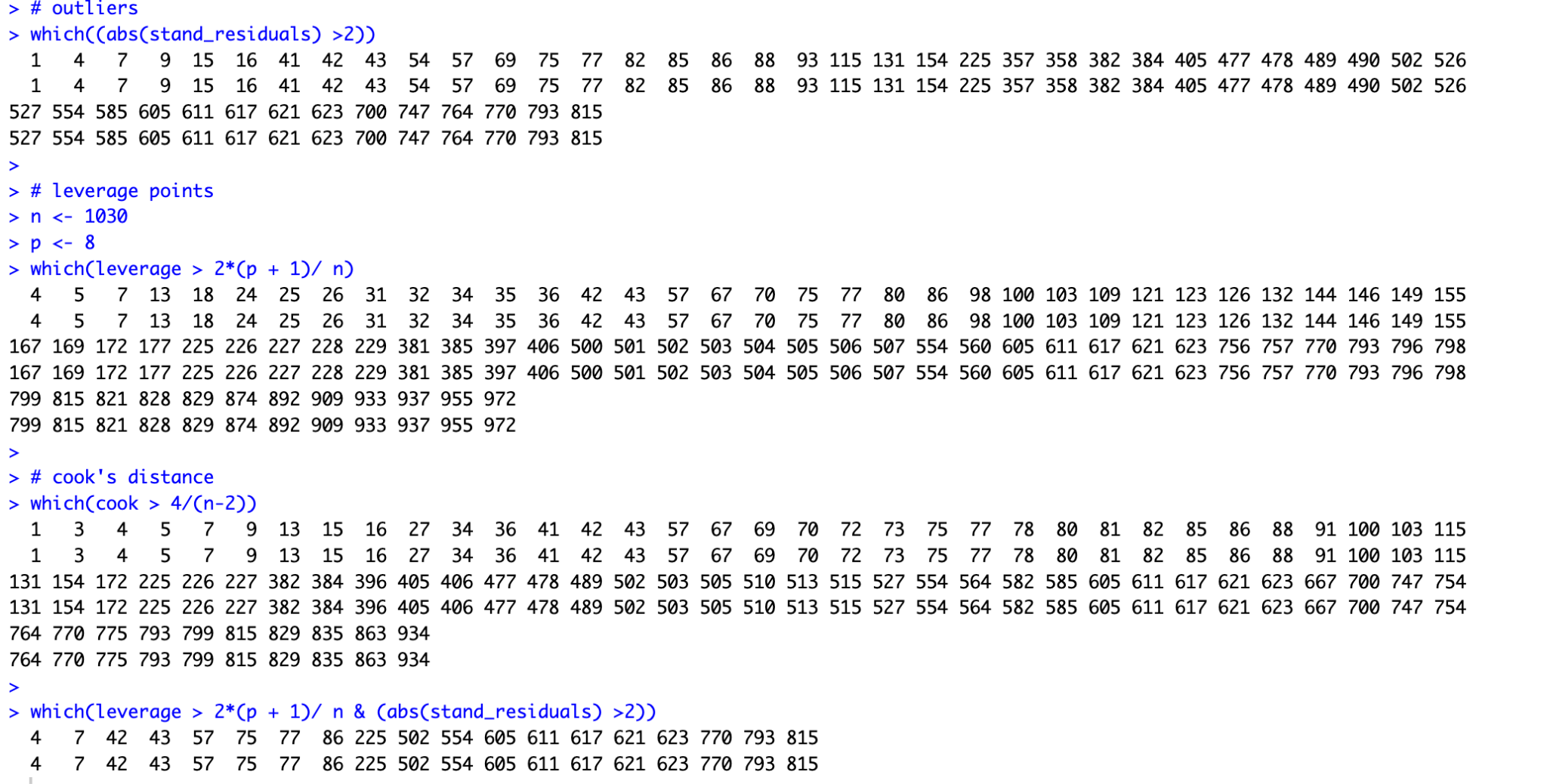
**Appendix**

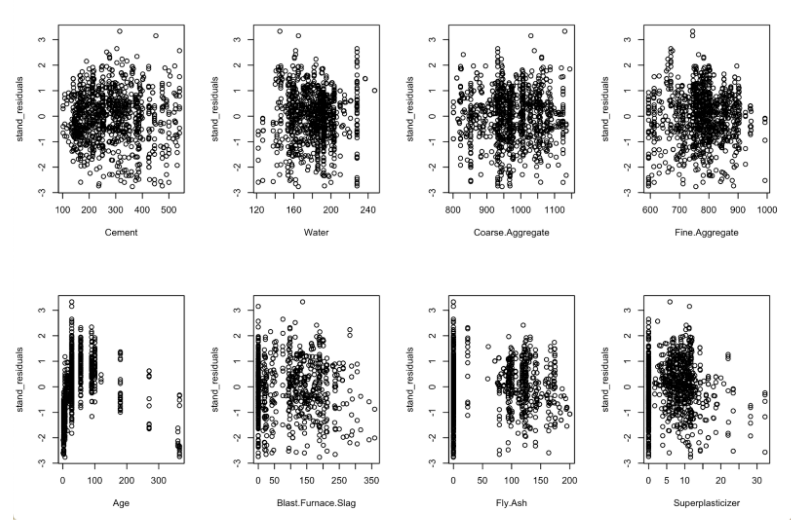
Distribution of data



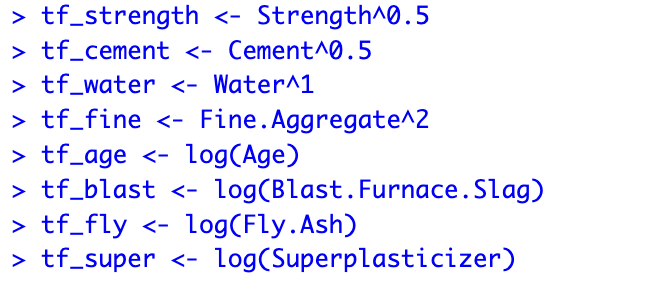
Standardized residuals, leverage, and cook’s distance for full model





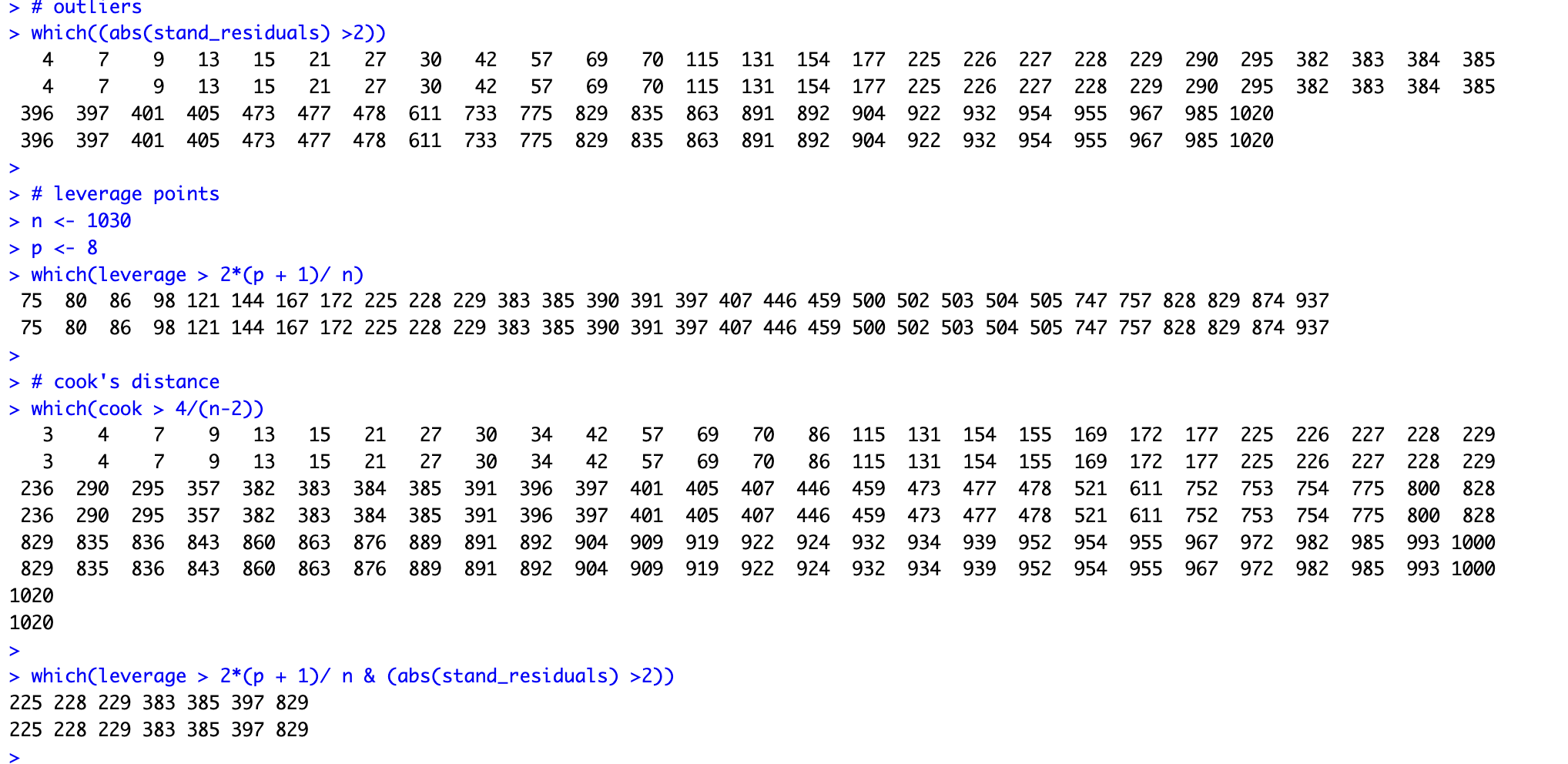
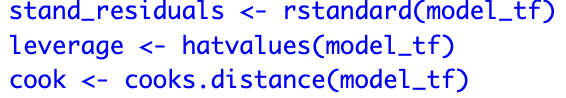


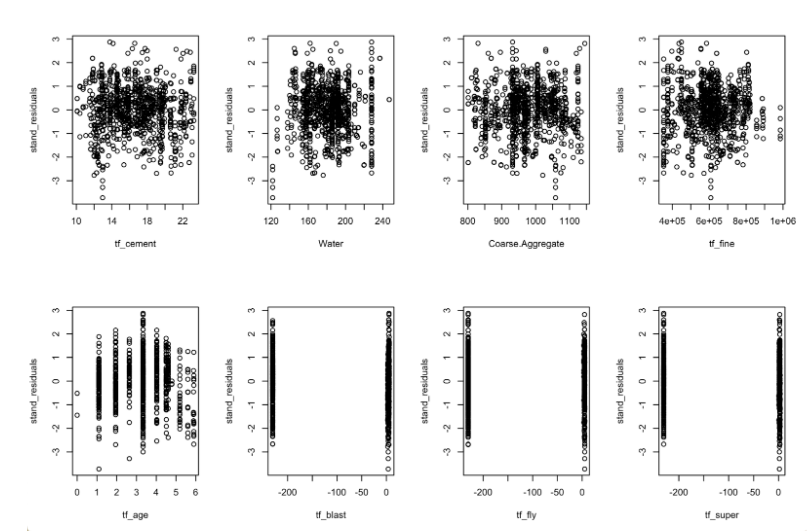
Code for transformed model



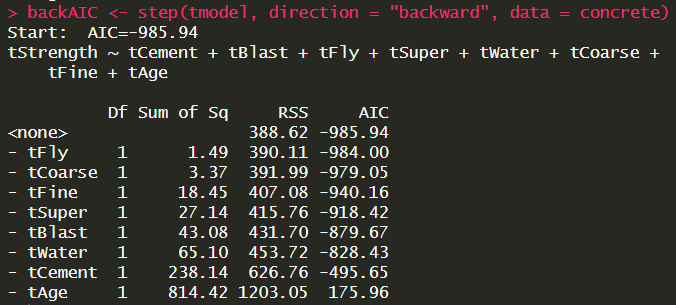


Transformed standardized residuals, leverages, and cook’s distance

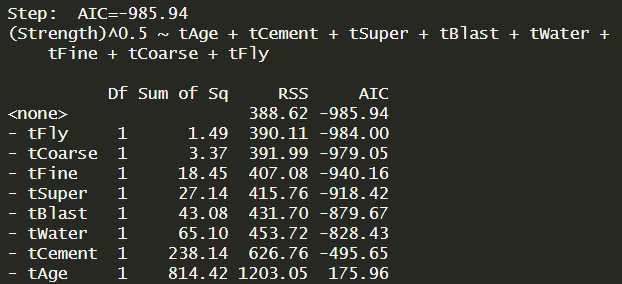




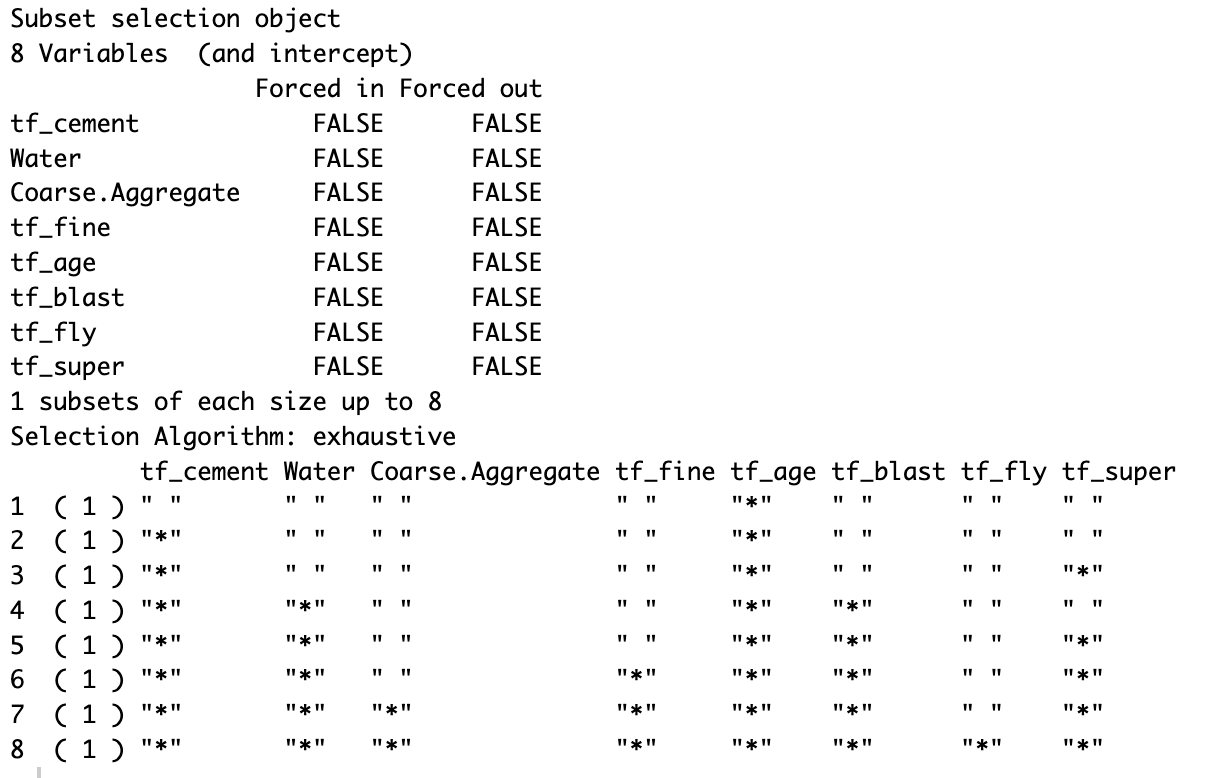
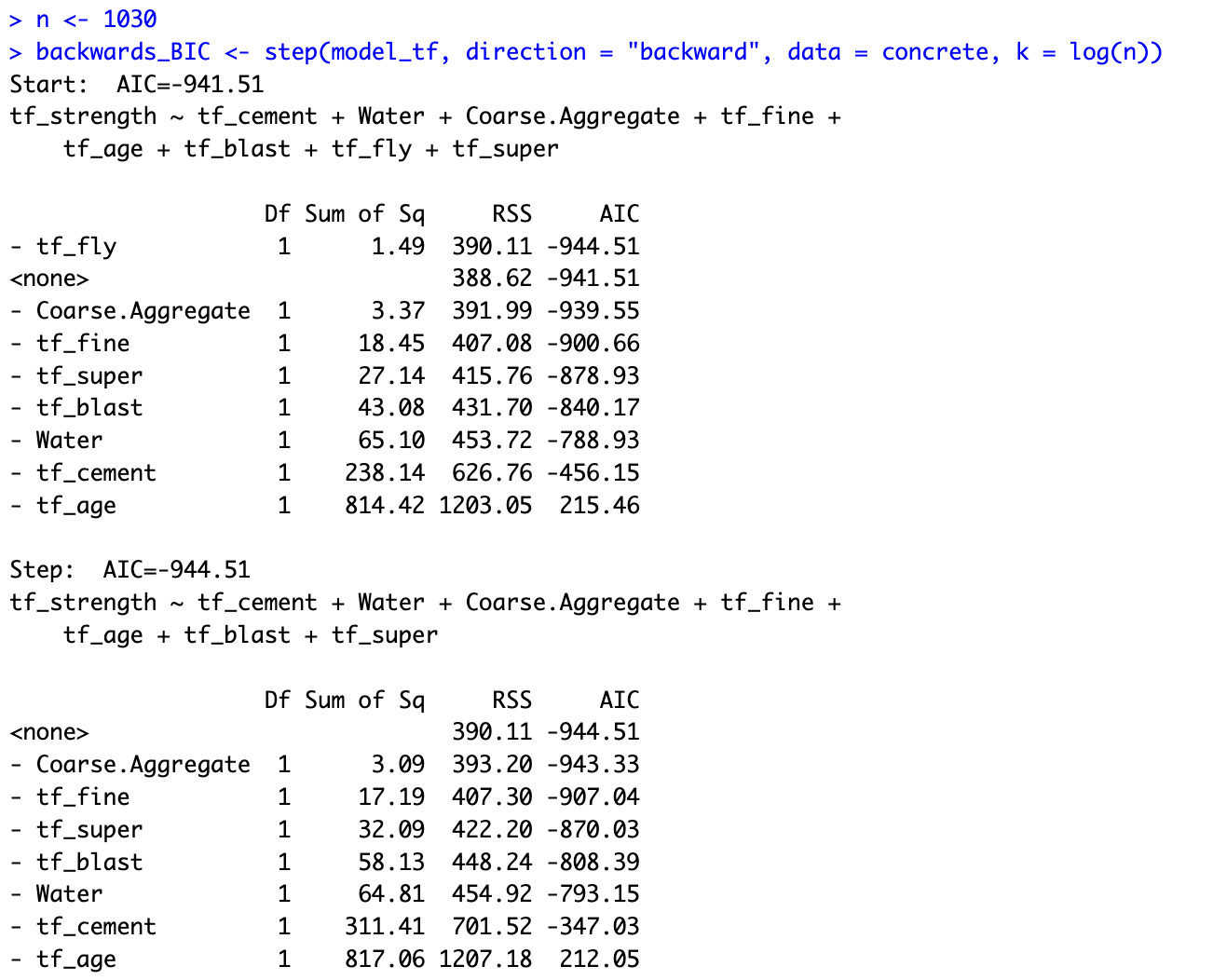
Back AIC



Forward AIC last step



Backward BIC and all possible subsets



Transformed model with only 7 predictors

