Modeling Compressive Strength of Concrete

Khang Tong & Xinyi Huang

6 December 2021

1. Introduction

Concrete is the second most consumed material in the world after water [1]. In 2020, 4.1 billion metric tons of cement was produced worldwide and is expected to continue growing each year, especially from emerging countries, as more nations industrialize [2]. Furthermore, America's aging infrastructure means that additional concrete will be needed to maintain them. Concrete has a wide range of uses, from building material to hydroelectric dams.

Concrete is, at its most basic, a mixture of three components: cement, water, and aggregates [3]. There may also be additional ingredients which help to strengthen the final product. Concerns regarding concrete production stem from economic as well as environmental impacts. Cement accounts for 8% of the world's global greenhouse gas emissions and concrete production accounts for 1.7% of global water use each year [1][4]. Additionally, roughly 50 billion metric tons of sand are used to create concrete each year. This has become an issue in places such as Florida where the state's beaches are receding due to the overharvesting of sand allowing tides to reach cities and cause damage [5].

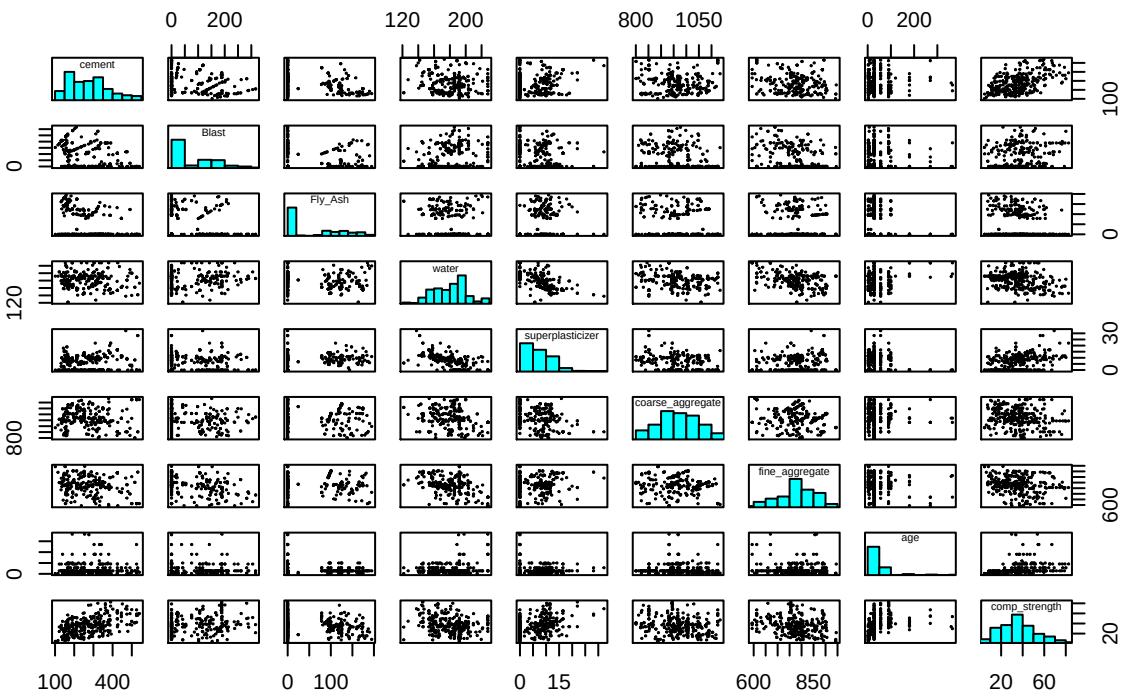
Our primary concern for this project was to figure out which of the components of concrete contributes the most to its final compressive strength so that we can better understand how to make concrete more efficiently to minimize economic and environmental costs as well as maximize strength for safety and longevity.

2. Data

Our data comes from the UCI Machine Learning Repository which was donated to them by Dr. I-Cheng Yeh from the Department of Information Management at the Chung-Hua University in Taiwan [6]. Previous analysis on this data used machine learning methods and is different from our present analysis. The data consists of one response variable, compressive strength, and 8 regressor variables, 7 of which are ingredients which might be found in concrete plus a time variable. Our sample size is 1030. Below is a description of each variable on how it relates to concrete production as well as a brief numerical summary:

Variable	Unit	Range	Mean	Description
Compressive Strength	mPA	2.3-82.6	35.82	Our response variable and a measure of how much pressure can be applied to the concrete before it cracks.
Cement	kg/m^3	102.0-540.0	281.2	A fine powder of limestone, clay, and gypsum.
Water	kg/m^3	121.8-247.0	181.6	Activates the cement and other chemical compounds in the concrete mixture so that it can glue the aggregates together while filling in the empty space between them.

Variable	Unit	Range	Mean	Description
Fine Aggregate	kg/m^3	594.0-992.6	773.6	Fine and coarse aggregate account for roughly 3/4ths of concrete's volume. It is typically used as a filler since it is cheap and provides concrete with its structure. Typically, aggregates are made from eroded rocks such as basalts, granite, or limestone.
Coarse Aggregate	kg/m^3	801.0-1,145.0	972.9	An aggregate is coarse if it is over 5mm in diameter. A mixture of fine and coarse aggregate is desired as the coarse aggregate can fill a lot of volume while the fine aggregate can provide support by filling in the space between the coarse aggregates.
Fly Ash	kg/m^3	0.0-200.1	54.19	A by-product of coal plants. Fly ash can replace up to 60% of cement by volume which is good since fly ash is cheaper to produce than cement. Fly ash also has properties which can help keep cement more robust when it is wet.
Blast Furnace Slag	kg/m^3	0.0-359.4	73.9	A by-product of steel production. It can replace up to 80% of cement by volume and has properties which increases the compressive strength and durability of the final concrete mixture.
Superplasticizer	kg/m^3	0.0-32.2	6.203	Helps to keep wet concrete more moldable and workable, thereby decreasing the amount of water needed by as much as 15% - 30%.
Age	Days	1.0-365.0	45.66	Concrete tends to increase in strength over time. Concrete gets roughly half of its strength in 3 days, almost 3/4ths by the first week, and 90% by the second week. Barring damage, concrete will continue to strengthen over years [3].



From the pairs plot, we notice that there is a high amount of collinearity between fly ash and blast furnace slag. This observation is in concurrence with their similar role in concrete.

Initially, we wanted to see if we were able to remove one of these variables. An ANOVA test was carried out using the following models:

Model 0.1: Compressive Strength ~ 1

Model 0.2: Compressive Strength $\sim$ Blast Furnace Slag

Model 0.3: Compressive Strength $\sim$ Blast Furnace Slag + Fly Ash

```
## Anova Table (Type II tests)
##
## Response: comp_strength
##           Sum Sq   Df F value    Pr(>F)
## Fly_Ash     1238    1  4.5296 0.0335515 *
## Blast       3247    1 11.8775 0.0005912 ***
## Residuals 280715 1027
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since both variables explain a significant amount of variation in compressive strength, we should not remove one of these variables. To deal with our collinearity, we combined them into one variable called fly_blast by taking their average.

$$\text{Let fly\_blast} = \frac{\text{Fly\_Ash} + \text{Blast}}{2}.$$

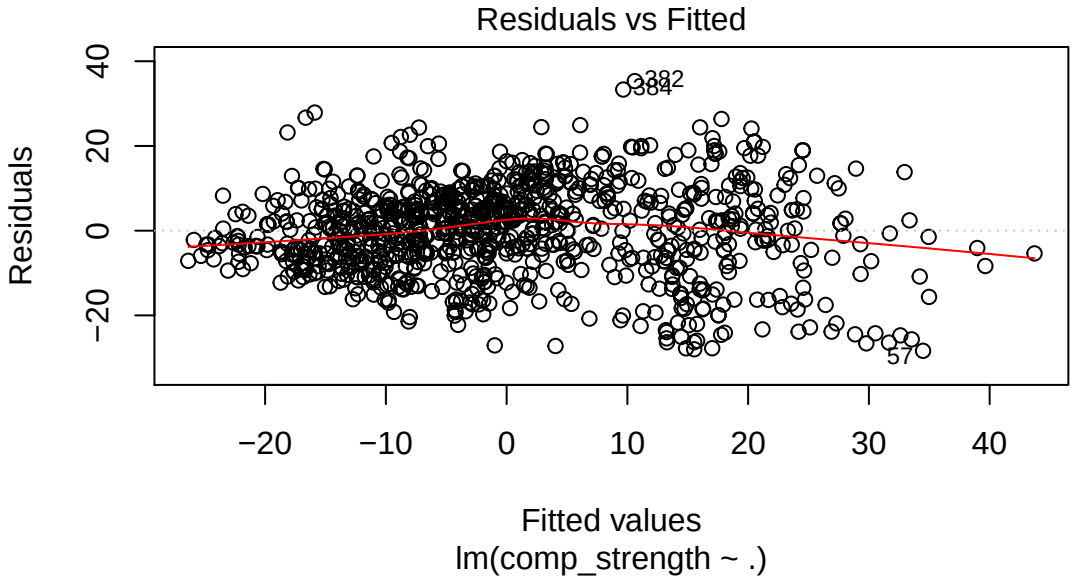
3. Initial Model

In order to attain a more interpretable intercept, we subtracted the mean value from each variable.

Our initial regression model was just a regression on all the given variables where the only change was introducing fly_blast as discussed in the previous section and subtracting the mean value from each variable.

Model 1: Compressive Strength ~ All Variables

```
## Summary of Model 1

##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.000      0.325   0.000    1.000
## cement               0.127      0.008  15.727    0.000
## water               -0.136      0.040  -3.408    0.001
## superplasticizer     0.230      0.091   2.544    0.011
## coarse_aggregate     0.021      0.009   2.196    0.028
## fine_aggregate       0.025      0.011   2.384    0.017
## age                  0.115      0.005  21.164    0.000
## fly_blast            0.214      0.020  10.594    0.000

## Adjusted R-squared: 0.61
```



We notice that some of the assumptions for a linear model are not met, namely:

1. The conditional expectation of the residuals are not 0 towards both ends of our fitted values.
2. The variance appears to increase as the fitted values increase.

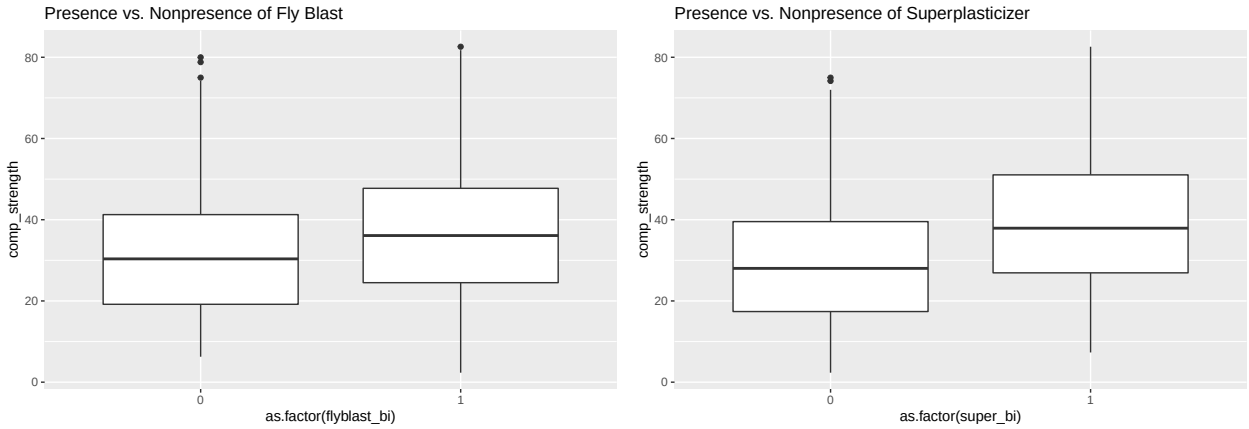
We then proceeded by addressing two concerns we noticed in the pairs plot:

1. The large number of zeros in the fly blast and superplasticizer variables.
2. The apparent quadratic trend in age.

4. Analysis

4.1 The Large Number of Zeros in the Fly Blast and Superplasticizer Variables

We were concerned that the large number of zeros in both the fly blast and superplasticizer variables would influence their estimated coefficients. So we created subsets of our data, one where fly blast was 0 and another where fly blast was not 0. We did the same thing for superplasticizer. We did this so that we could answer a sub-question: does the presence versus non-presence of either fly blast or superplasticizer result in greater average compressive strengths? We hypothesized that it would given their known roles in concrete production.




```
##
##  Welch Two Sample t-test
##
## data:  fly_blast0 and fly_blast1
## t = -4.0897, df = 389.29, p-value = 2.626e-05
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -2.948822
## sample estimates:
## mean of x mean of y
##  31.98998  36.93070
```

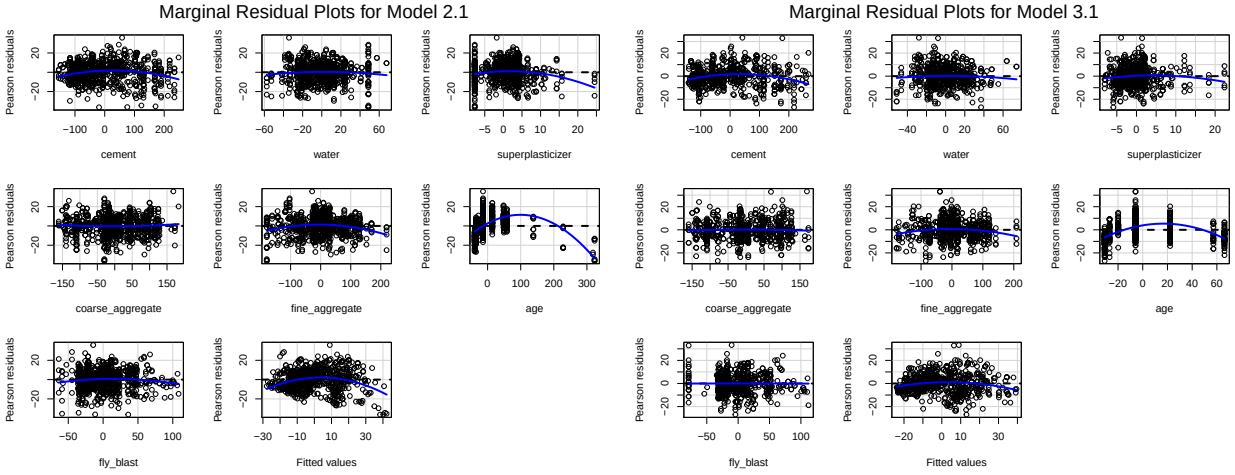
```
##
##  Welch Two Sample t-test
##
## data:  super0 and super1
## t = -9.9656, df = 872.39, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -8.321075
## sample estimates:
## mean of x mean of y
##  29.51761  39.48571
```

The result of our t-test provides overwhelming evidence that concrete mixtures with the presence of either fly ash, blast furnace slag or superplasticizer have a higher average compressive strength than mixtures that do not. Now that we have answered that question, we focus our attention to analyzing the subset of our data that includes either fly blast or superplasticizer. Doing so will remove 232 data entries (22.5%) for the subset concerning fly blast and 379 data entries (36.8%) for superplasticizer. We will also subtract the means from each variable again to ensure a more interpretable intercept.

Model 2.1: Compressive Strength $\sim$ All variables where fly blast is nonzero.

Model 3.1: Compressive Strength $\sim$ All variables where superplasticizer is nonzero.

```
## Summary of Model 2.1

##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.000      0.372   0.000    1.000
## cement               0.136      0.010  14.293    0.000
## water               -0.071      0.048  -1.478    0.140
## superplasticizer     0.316      0.109   2.895    0.004
## coarse_aggregate     0.035      0.012   2.938    0.003
## fine_aggregate       0.047      0.013   3.702    0.000
## age                  0.148      0.007  19.932    0.000
## fly_blast            0.233      0.026   8.864    0.000

## Adjusted R-squared: 0.606

## Summary of Model 3.1

##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.000      0.342   0.000    1.000
## cement               0.142      0.008  16.739    0.000
## water               -0.126      0.039  -3.218    0.001
## superplasticizer    -0.264      0.102  -2.582    0.010
## coarse_aggregate     0.022      0.010   2.110    0.035
## fine_aggregate       0.023      0.011   2.059    0.040
## age                  0.325      0.013  25.264    0.000
## fly_blast            0.230      0.024   9.567    0.000

## Adjusted R-squared: 0.727
```



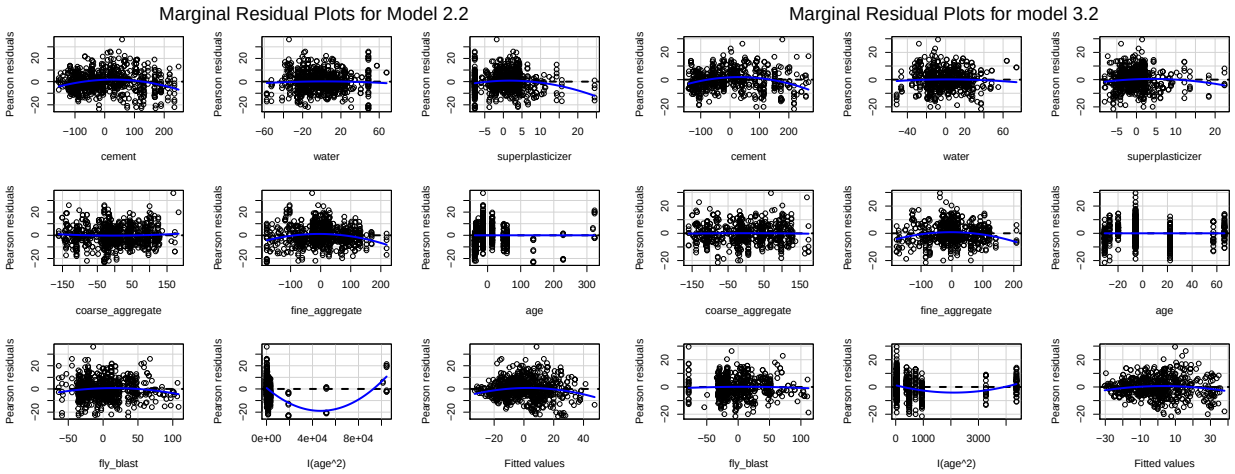
Models 2.1 and 3.1 have an adjusted R^2 of 0.61 and 0.73 respectively. We notice that there still appears to be a large quadratic trend in the residuals of age.

4.2 The Quadratic Trend in Age

Next we add a quadratic term for age to see if that improves our model and reduces the quadratic residuals we see in the age variable.

Model 2.2: Compressive Strength $\sim$ All variables where fly blast is nonzero + Age^2.

Model 3.2: Compressive Strength $\sim$ All variables where superplasticizer is nonzero + Age^2.



The conditional expectation for the marginal residuals of the age terms now appear to be much closer to zero across their ranges, especially in model 3.2. Model 3.2 also has a better residual plot for the fitted values over model 2.2.

Additionally, the adjusted R^2 value in both models 2.2 and 3.2 increase from their 2.1 and 3.1 counterparts, from 0.61 to 0.74 and from 0.73 to 0.81 respectively. Based off of the improved residual plots and R^2 values, we conclude that adding the Age^2 term improves our model.

4.3 Model Selection

In this section we perform various tests on models 2.2 and 3.2 to determine which one fits our data better.

As mentioned in the previous section, model 3.3 has a higher adjusted R^2 value at 0.81 compared to 0.74 for model 2.2.

We calculated the Akaike information criterion for both of our models and got:

```
## [1] "Model 2.2 AIC: 5692.348"
```

```
## [1] "Model 3.2 AIC: 4451.255"
```

Since model 3.2 has a lower AIC, model 3.2 is a better fit for our data.

We also conducted a 10-fold cross validation test on our data and got:

```
## [1] "Model 2.2 RMSE: 16.638"
```

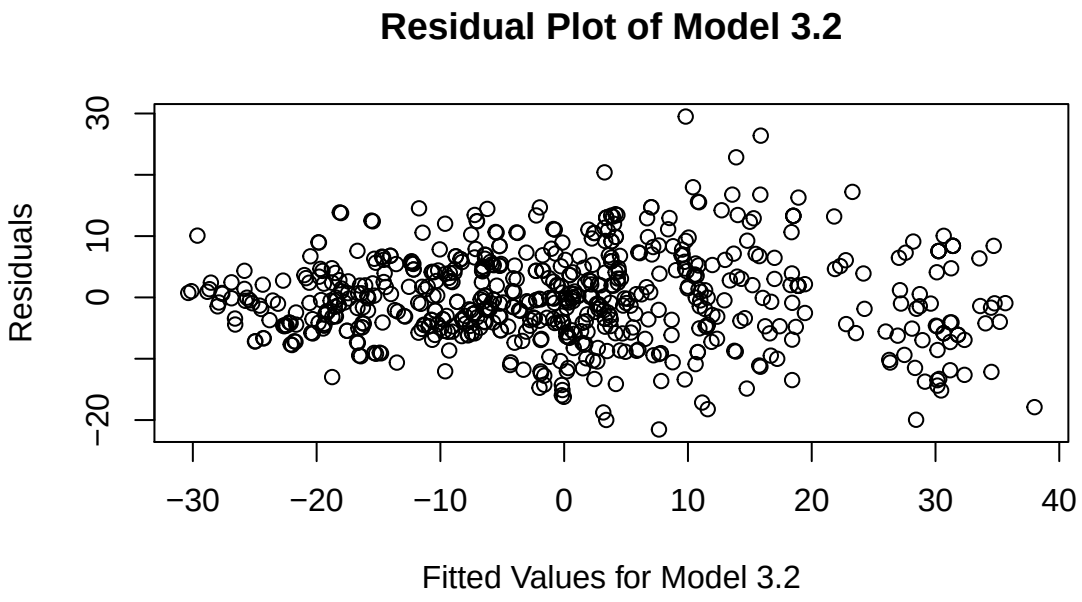
```
## [1] "Model 3.2 RMSE: 16.247"
```

Since the RMSE of model 3.2 is lower than the RMSE of model 2.2, then the results of 10-fold cross validation suggests that model 3.2 is a better fit for our data.

Given evidence from analyzing adjusted R^2 values, AIC, and RMSE, we will use model 3.2 since it fits our data better.

4.4 Heteroscedasticity

Next, we address the heteroscedasticity in our data.

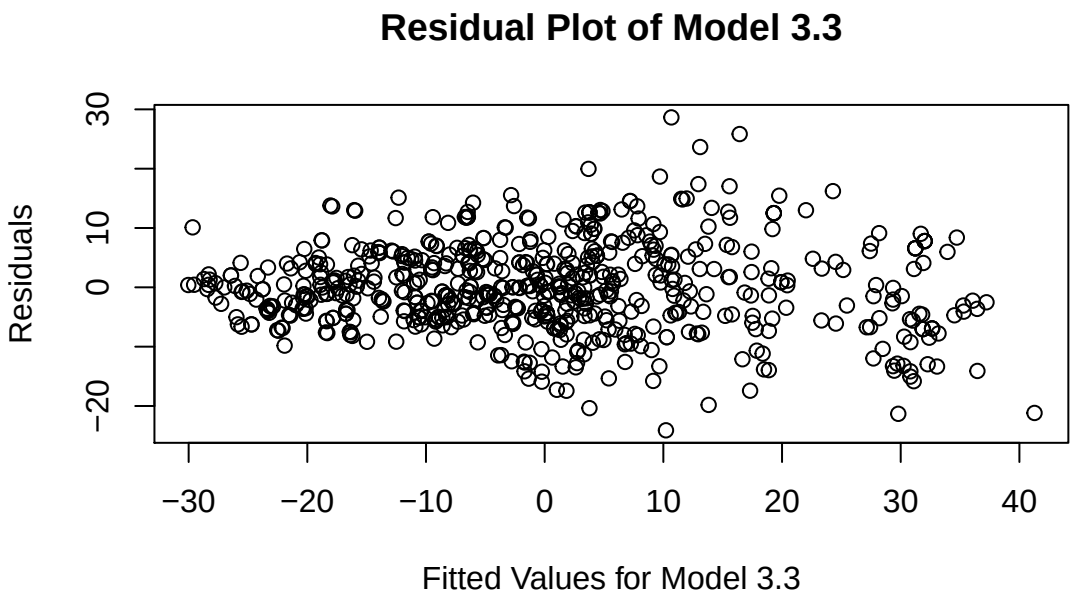



```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 46.20777, Df = 1, p = 1.0635e-11
```

As we can see in the residual plot, variance appears to increase as the fitted values increase. A non-constant variance score test also provides evidence that the heteroscedasticity is significant.

We fit a weighted least squares model using the predicted values of the absolute residuals of model 3.2 which we will call model 3.3:

```
w1 = lm(abs(residuals(m3.2)) ~ predict(m3.2))
m3.3 = lm(comp_strength~.+I(age^2)-super_bi, data = df2_super, weights=1/predict(w1)^2)
```



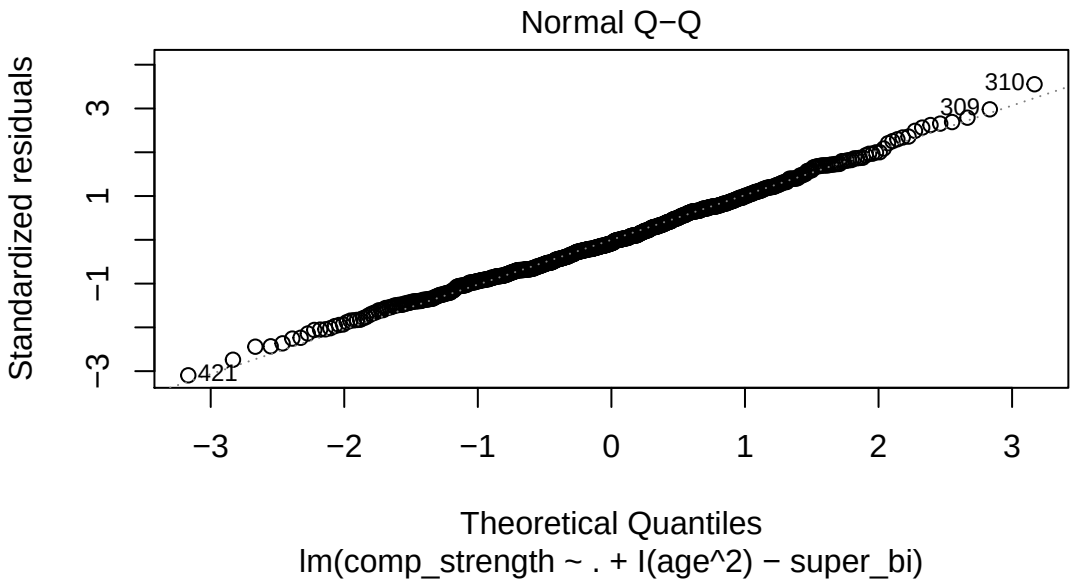

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 2.04237, Df = 1, p = 0.15297
```

While there appears to be little change in the residual plot, the non-constant variance score test no longer results in significant heteroscedasticity. While this does not prove homoscedasticity, we are less concerned about heteroscedasticty in our model.

4.5 Miscellaneous Diagnostics

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 310 3.586285         0.00036096      0.23498
```

Additionally, an outlier test does not find any significant outliers.



The Q-Q plot also shows reasonable normality in the residuals of our model.

5. Conclusion

Below is a summary of our final model, model 3.3, using weighted least squares:

```
## Summary of Model 3.3

##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)         4.694      0.412  11.380    0.000
## cement              0.188      0.007  26.373    0.000
## water               0.011      0.031   0.352    0.725
## superplasticizer   -0.213      0.085  -2.496    0.013
## coarse_aggregate    0.079      0.009   9.028    0.000
## fine_aggregate      0.083      0.009   8.798    0.000
## age                 0.530      0.015  36.130    0.000
## fly_blast           0.330      0.020  16.444    0.000
## I(age^2)           -0.006      0.000 -17.531    0.000

## Adjusted R-squared: 0.822
```

We notice that age has the highest t value in our model. This supports the known knowledge that concrete strength increases over time. The other 2 additional additive ingredients that had a significant estimated coefficient were fly blast and superplasticizer. Of all the ingredients with comparable units, fly blast has the highest coefficient in terms of magnitude. Superplasticizer appears to have a significant negative association with concrete compressive strength. While the results of one of our earlier t tests showed that the average compressive strength of the subset of the sample which contained superplasticizer was significantly greater than the subset of the sample without superplasticizer, we did not conclude how varying quantities of

superplasticizer is associated with compressive strength. Based off of our knowledge that superplasticizer helps to keep concrete more moldable and easy to work with (thereby decreasing the amount of water needed), the negative coefficient of superplasticizer could be explained by a trade-off in benefits. On average, samples with superplasticizer are stronger because they use less water, but superplasticizer itself weakens concrete by making it more moldable which can help save money on site by making it easier to work with.

The only insignificant coefficient in our model was water. This could be explained by the nonlinear role that water has in concrete production. While a certain amount of water is needed to activate the other ingredients, after that threshold, additional water only serves to dilute the mixture. We attempted to add a quadratic term for water into our model but it made the fit worse. This could be a point for further investigation.

Based off of our weighted model, we recommend that fly ash and blast furnace slag be maximized in concrete mixtures. We also recommend conducting a cost-benefit analysis on superplasticizer use on site so that savings in working with concrete and final compressive strength can be maximized while minimizing the reductive association of strength with superplasticizer.

References

[1] C. R. Gagg, "Cement and concrete as an engineering material: An historic appraisal and case study analysis," *Engineering Failure Analysis*, vol. 40, pp. 114–140, 2014, doi: https://doi.org/10.1016/j.engfailanal.2014.02.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1350630714000387

[2] "Cement production in the united states from 2010 to 2020," 2021 [Online]. Available: https://www.statista.com/statistics/219343/cement-production-worldwide/

[3] A. M. Neville and J. J. Brooks, *Concrete technology*. Longman Scientific & Technical England, 1987.

[4] "Study: Water use in concrete production higher than expected." Fluence, 2018 [Online]. Available: https://www.fluencecorp.com/concrete-industry-water-use/

[5] A. Niranjan, "Sand crisis: Mafias thrive as shortages loom." DW, 2021 [Online]. Available: https://p.dw.com/p/3pxxa

[6] I.-C. Yeh, "Concrete compressive strength data set." 1998 [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength