Abstract

Ensuring that the correct type and grade of concrete is chosen is the one of the most critical steps in any project. Compressive strength of concrete and concrete workability are the two most significant determining factors when selecting a concrete type. A mathematical model for the prediction of the compressive strength of concrete was determined using statistical analysis on the data set obtained online. The proposed model proved to be accurate in predicting the compressive strength. Statistical modeling of this data set has a significant advantage over alternative techniques since it is mathematically rigorous, can be used to find a confidence interval for the predictions and can also provide insight into the major factors that influence the compressive strength of concrete through correlation and interaction analysis.

For this study, we aim to investigate the following questions:

1) Which component(s) has the greatest/least effect on the compressive strength?
2) Are there any relationships between the different components that affect compressive strength?
3) Is there any predictor variable(s) that is unusual in the chosen model?

Introduction

Concrete is a very important mixture of materials in civil engineering that is commonly used when constructing buildings in modern times. The Compressive Strength of Concrete is a crucial factor of the performance and durability of the concrete and can be determined by altering the proportions of cement, coarse and fine aggregates, water, and various other components. Estimating compressive strength also assists in meeting all construction quality control parameters such as avoiding excessive loading. This study aims to investigate the impact of the variables on the compressive strength. The dataset that was used in this study contains one thousand and thirty observations of the X variables and the resulting response variable. The data was sourced online from the University of California Irvine Machine Learning Repository and was originally donated by Professor I-Cheng Yeh (Department of Information Management, Chung-Hua University, Taiwan)

The Variables from our dataset that we will consider in this study are:

Cement (X Variable 1) measured in KG per M3
Blast Furnace Slag (X Variable 2) measured in KG per M3
Fly Ash (X Variable 3) measured in KG per M3
Water (X Variable 4) measured in KG per M3

Superplasticizer (X Variable 5) measured in KG per M3
Coarse Aggregate (X Variable 6) measured in KG per M3
Fine Aggregate (X Variable 7) measured in KG per M3
Age (X Variable 8) measured in Days (1 - 365)
Concrete Compressive Strength (Response Variable) measured in MPa

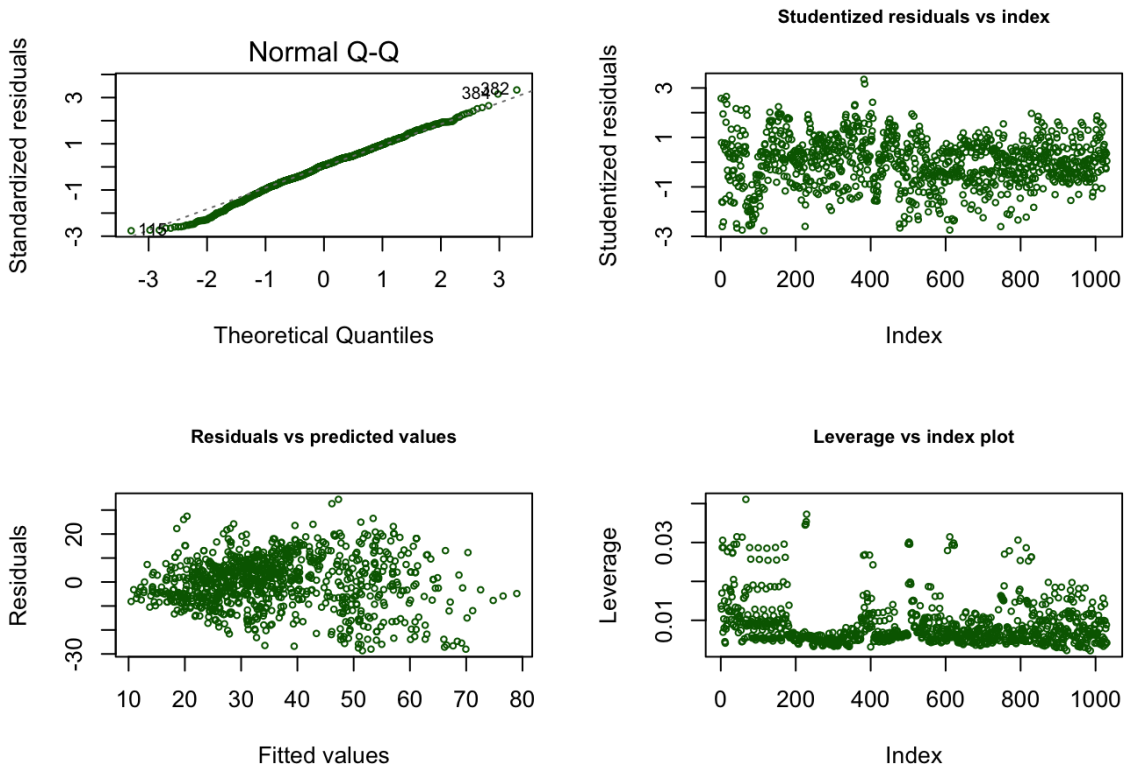
Analysis

Correlation:
Variance Inflation Factors on the regular model with all variables (no interactions).

Variable	VIF
Cement	7.488657
Blast Furnace Slag	7.276529
Fly Ash	6.171455
Water	7.004663
Superplasticizer	2.965297
Coarse Aggregate	5.076044
Fine Aggregate	7.005346
Age	1.118357

The VIFs for each variable are not too large relative to each other. Also, scatterplots were plotted and no high correlations were seen. None of the VIFs exceed 10, which indicates that multicollinearity does not exist.

Diagnostics:



1) The data looks to follow a straight line on the qq-plot, so the assumption of normality is appropriate.
2) Based on the studentized residuals vs index plot, there are 2 points greater than 3 which indicates those two are outliers. However those are only 2 out of 1030 points, so it can be assumed that there are no outliers
3) The residuals vs predicted values data looks to have a funnel-shape, so the data has non-constant variance. Non-constant variance can be assumed
4) Most of the leverage points are small as they are less than 0.03. But there is one point that is greater than 0.03, which suggests the possibility of an outlier in x-space

Since non-constant variance and normality is assumed, robust regression must be used for the model creation.

Interaction variables:

Consider the two-way interaction variable of Age and Super Plasticizer (note both are continuous variables), let it be called asuper where asuper = Age * Super Plasticizer. This variable improves the model fit since it increases R-Squared by a good amount:

R-squared value of the full model is 0.6142, while the Rsquared value of the full model with the interaction variable added is 0.6846. As such, asuper will be added to the data for model selection testing.

To interpret asuper, consider the contour plot of the function (coefficients are calculated from robust regression of the full model with interaction variable):

$$y = 2.2658697 + 0.0876651Age - 0.3834723SuperPlasticizer + 0.0194690asuper$$

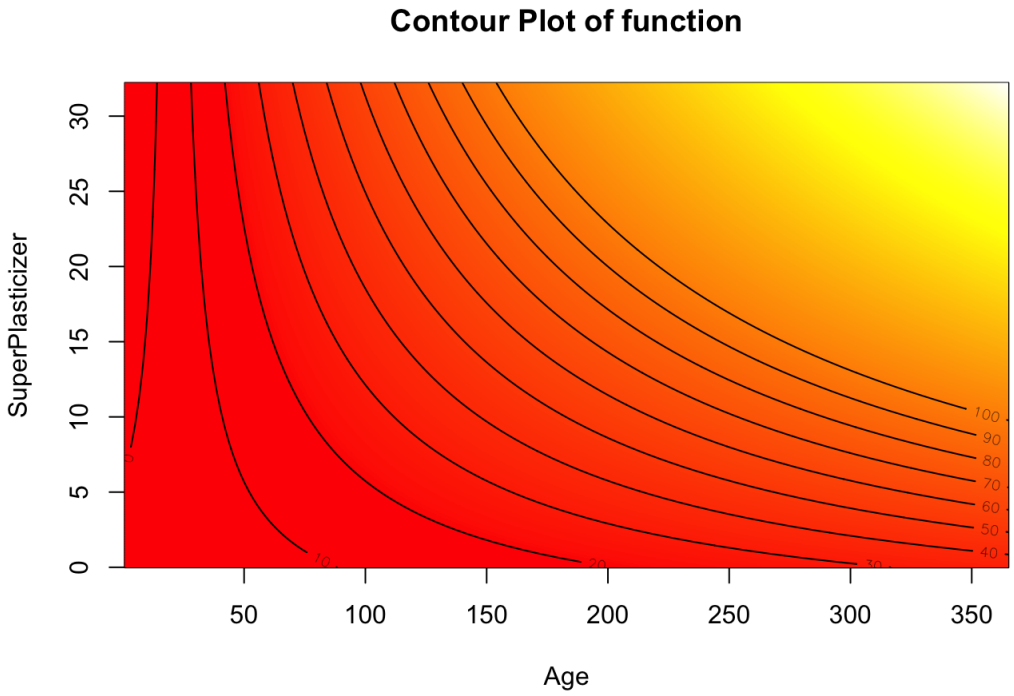


Figure 1: Contour plot

As the segments of the contour plot become darker, they indicate the y value is increasing. It looks that the y value increases the lower SuperPlasticizer is, and the higher Age is.

Model Selection:

Forward Selection with AIC: The model with the lowest AIC adds the following x-variables: asuper, Cement, Blast Furnace Slag, Age, Fly Ash, Water, Superplasticizer

Backward Selection with AIC: The model with the lowest AIC removes the following x-variables: Fine Aggregate, Coarse Aggregate

Both methods suggest that the full model without the x-variables Fine Aggregate and Coarse Aggregate should be used. Using robust regression to gather the coefficients for each x-variable, the model is:

$$y = 22.146796 + 0.019568asuper + 0.108191Cement + 0.087040BlastFurnaceSlag + 0.070291FlyAsh$$
$$- 0.179393Water - 0.406861SuperPlasticizer + 0.087259Age$$

Diagnostics of chosen model (please refer to Appendix for the plots)

1) The data follows a more straight line on the qq-plot
2) Most of the studentized residuals are less than 3, while there are 2 points greater than 3. Again, since only 2 out of 1030 points are greater than 3, there are no outliers.
3) The residuals vs predicted values look more clustered together and look to have a funnel-shape, so non-constant variance can be assumed.
4) Most of the leverage points are small since less than 0.01, while there are a few points greater than 0.01.

Discussion

In this section, the questions that are to be investigated will be discussed.

Which component(s) has the greatest/least effect on the compressive strength?

Since the variables CoarseAggregate and FineAggregate are not included in the chosen model, it suggests that these 2 variables have very little effect on compressive strength.

Consider the null hypothesis for each individual case when Coarse Aggregate = 0, Fine Aggregate = 0, and Coarse Aggregate = Fine Aggregate = 0. The original full model (does not contain the x-variable asuper) and a significance level of 0.05 will be used for the test.

Hypothesis	Result
Coarse Aggregate = 0	p-value = 0.0763448 Since p-value > 0.05, we accept the hypothesis that CoarseAggregate = 0. Hence CoarseAggregate is not

	statistically significant
Fine Aggregate = 0	p-value = 0.0813378 Since p-value > 0.05, we accept the hypothesis that FineAggregate = 0. Hence FineAggregate is not statistically significant
Coarse Aggregate = Fine Aggregate = 0	p-value = 0.1375 Since p-value > 0.05, we accept the hypothesis that FineAggregate = 0. Hence FineAggregate is not statistically significant

Based on the results, both Coarse Aggregate and Fine Aggregate are not statistically significant which suggests that they produce little to no effect on the compressive strength.

On the other hand, based on the chosen model, it looks that Superplasticizer has a great effect on the compressive strength. It has a regression coefficient of -0.406861 which is rather big compared to the other x-variables. As such, 0.406861 is the estimated decrease in compressive strength on average, per kg/m^3 of Superplasticizer.

So a suggestive answer is:

- Greatest effect: Superplasticizer
- Least effect: Coarse Aggregate, Fine Aggregate

Are there any relationships between the different components that affect compressive strength?

As the interaction variable asuper is used to improve the model fit, there may be an interaction between Age and Superplasticizer. Consider the null hypothesis when asuper = 0. The chosen model with the interaction variable asuper, and a significance level of 0.05 will be used for the test. Results are as follows:

The p-value < 2.2e^(-16). Since p-value < 0.05, we reject the hypothesis that asuper = 0. Hence asuper is statistically significant, so there is an interaction between Age and Superplasticizer.
R^2 improved by roughly 7%.

Is there any predictor variable(s) that is unusual in the chosen model?

The Superplasticizer material is generally used as an additive to strengthen the compressive strength of concrete. Based on the chosen model, the regression coefficient for Superplasticizer is negative which indicates that it reduces the compressive strength when Superplasticizer > 0. It is also suggested previously that Superplasticizer has a great effect on compressive strength as it decreases it by a relatively big amount. Hence, the coefficient estimate of Superplasticizer possibly contradicts the purpose of the material itself.

The 99% confidence interval for the coefficient estimate of Superplasticizer on 1022 degrees of freedom is: [-0.6331599, -0.1805627]. Based on the interval, it strongly suggests that the coefficient is negative in the chosen model, which makes it an unusual observation given the purpose of Superplasticizer.

So a suggestive answer is:

Superplasticizer since it can reduce the compressive strength by a big amount given that Superplasticizer > 0

Conclusion

Based on the analysis, it is highly suggested that the following model should be used for prediction of compressive strength of concrete:

$$y = 22.146796 + 0.019568asuper + 0.108191Cement + 0.087040BlastFurnaceSlag + 0.070291FlyAsh$$

$$- 0.179393Water - 0.406861SuperPlasticizer + 0.087259Age$$

Appendix

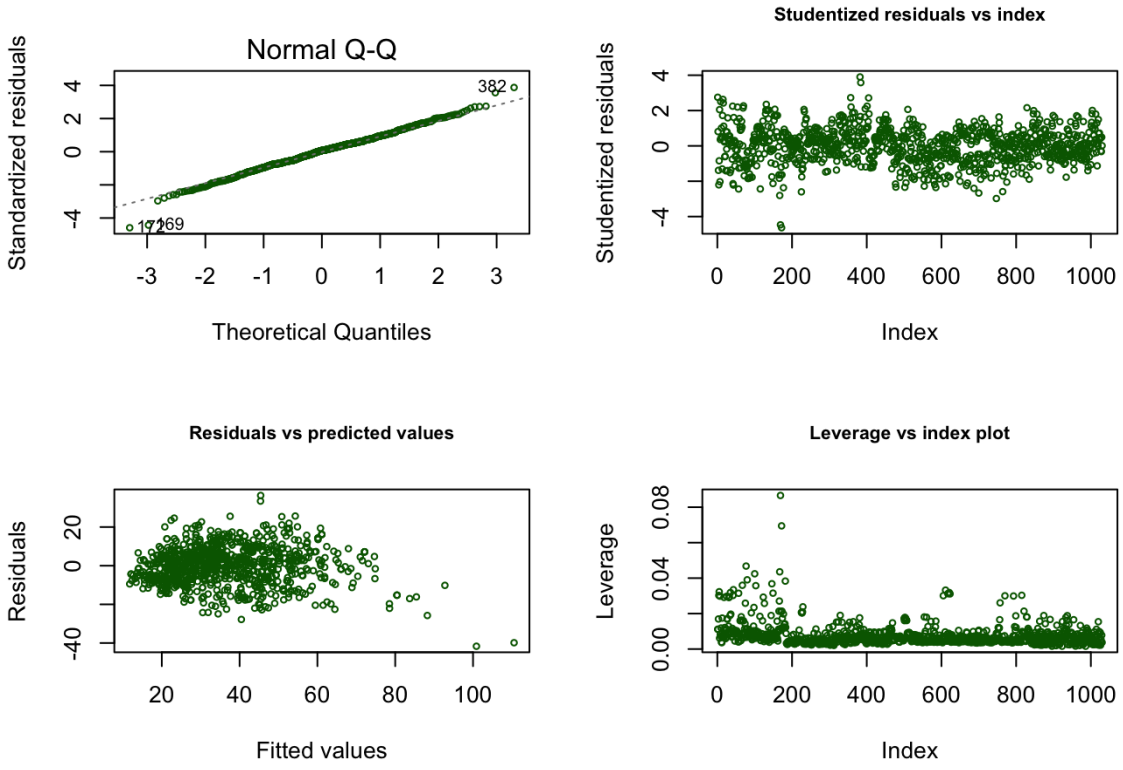
```
concrete <- read.csv("Concrete-Data.csv")

# rho value for contour plot
rho <- function(age, super) {
  2.2658697 + (0.0876651*age) + (-0.3834723*super) + (0.0194690*age*super)
}

xage <- seq(min(concrete$Age), max(concrete$Age), length = 1000)
xsup <- seq(0, max(concrete$Superplasticizer), length = 1000)
Rho <- outer(xage, xsup, "rho")

image(xage, xsup, Rho, col = heat.colors(1000), main = "Contour Plot of function",
      xlab = "Age", ylab = "SuperPlasticizer")
contour(xage, xsup, Rho, add = T, levels = c(0,10,20,30,40,50,60,70,80,90,100))
```

Code to produce contour plot (Figure 1)



Diagnostics of chosen model