Project Report

Predicting the Concrete Compressive Strength by Regression Models

Shunya Sunami

1 Introduction

1.1 Background and Motivation

Concrete is a fundamental building material in modern society, and its quality control is extremely important. Concrete is used in many of our familiar buildings such as bridges, houses, and dams. The compressive strength of concrete is one of the most important indicators to guarantee the safety of such buildings, and various compressive strength standards are established for different types of buildings. Compressive strength is a quantitative measure of how much weight a concrete can withstand and is a basic indicator to measure its reliability as a building material.[1]

This report tries to answer the research question: how various concrete formulation components and concrete age effect on compressive strength, by developing several predictive models.

We hope that the insights gained from this research will benefit building materials researchers, engineers, and related industries.

1.2 Data set

We use the Concrete Compressive Strength data set created by I-Cheng-Yeh.[2] The variables in the data set and the units are as shown below:

name	**description**	**units**
Cement	Joining agent used in water	kg in m^3 of mixture
Blast Furnace Slag	By-products generated in the steelmaking process	kg in m^3 of mixture
Fly Ash	By-product of burning coal	kg in m^3 of mixture
Water		kg in m^3 of mixture
Superplasticizer	Additives to reduce water requirements in concrete mixtures	kg in m^3 of mixture
Coarse Aggregate	It is commonly refer to as gravel	kg in m^3 of mixture
Fine Aggregate	Sand used in concrete	kg in m^3 of mixture
Age		Day(1 to 365)
Concrete compressive strength	Indicates how much force (weight) the concrete can withstand	MPa

Table 1: Variables of the data set with units. Obtained from [2]

This data set has a 1030 instances in total. Since there is no missing values nor corrupted data, no data cleaning is required prior to the analysis. Seven different ingredients of concrete, Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate, and Fine Aggregate, and age of the concrete will be used as the explanatory variables to try and predict the response variable, Concrete compressive strength.

1.3 Pre-Analysis

First of all, we visualize the scatter plots between the response variable, concrete compressive strength and each explanatory variables.

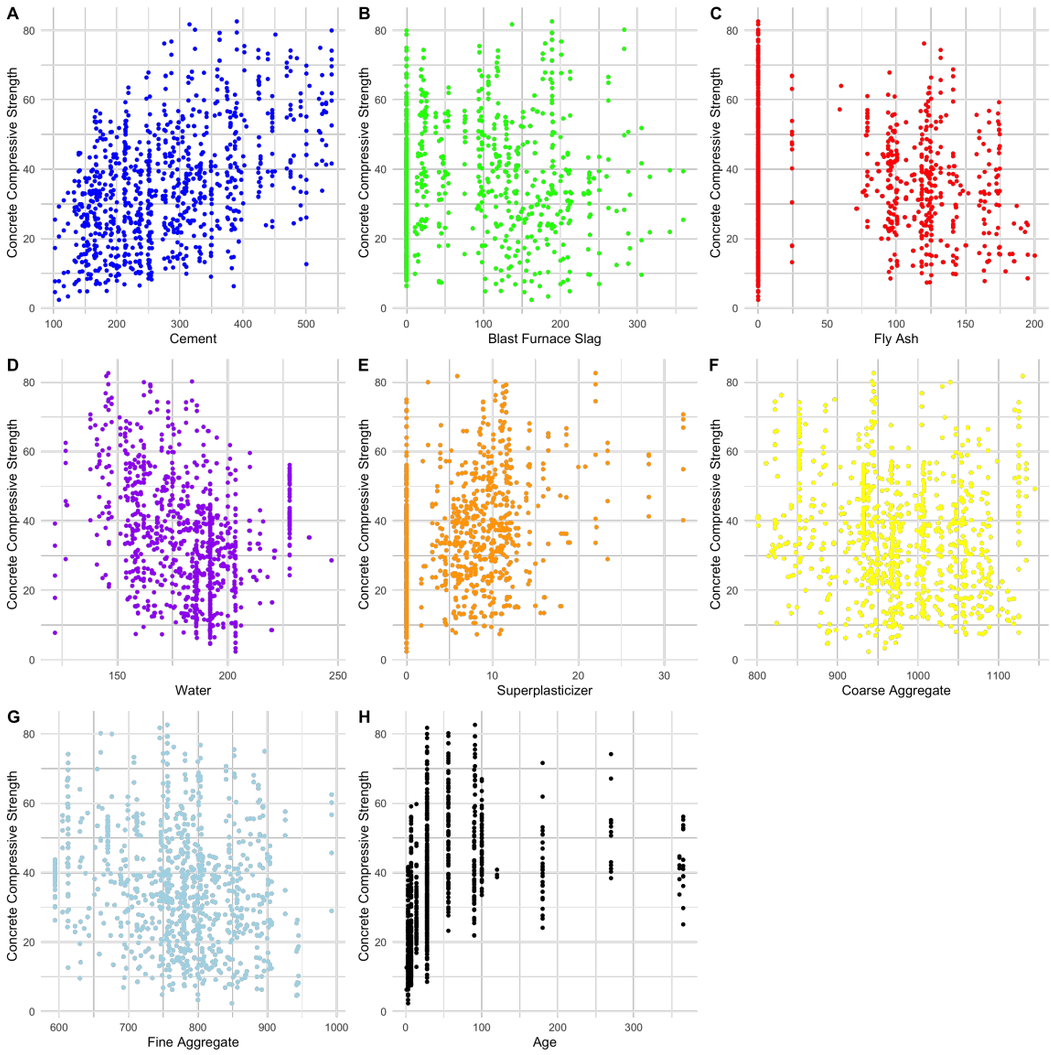

Figure 1: Concrete Compressive Strength vs Explanatory Variables

1.3.1 Cement

Cement seems to have the strongest positive linear trend against against concrete compressive strength according to Fig 1.A. The correlation coefficient of 0.5 is also indicating that the linear positive relationship between two variables are likely to be significant.

1.3.2 Blast Furnance Slag, Fly Ash, Coarse Aggregate, Fine Aggregate

From blast furnance slag(Fig 1.B), fly ash(Fig 1.C), coarse aggregate(Fig 1.F), and fine aggregate(Fig 1.G), we are not able to visually find the obvious patterns. This is also backed up by the very low correlation coefficient (0.13, -0.16, -0.16, and -0.17 respectively). However, it is very interesting to see that the correlation coefficient of fly ash, coarse aggregate, and fine aggregate are very close.

1.3.3 Water

Water (Fig 1.D) shows a negative trend against concrete compressive strength(corr coefficient = -0.29). We can also see the concave pattern, indicating that there maybe a quadratic relationship between two variables.

1.3.4 Superplasticizer

Superplasticizer(Fig 1.E) seems to have a positive trend against concrete compressive strength(corr coefficient = 0.37). It also shows a slight concave pattern, which might indicate the quadratic relationship between those variables.

1.3.5 Age

We can see that age(Fig 1.H) behaves like a categorical variable, though it is continuous. Though the range of the age is large, the data is biased towards some particular age, which could influence the reliability of the model. It is showing positive trend (corr coefficient = 0.33) with slight concave pattern.

1.4 Flow of Analysis

We will try to find the best predictive model as follows. First, full linear regression model, including all of the explanatory variables, is fitted and analyzed. We will then find the best combination of variables in linear regression. Subsequently, the model with polynomial terms and the model with interaction terms are fitted and analyzed respectively. Finally, those models are assessed using cross validation and the best predictive model is chosen.

2 Analysis

In this section, we will fit the different models to predict concrete compressive strength using the ingredients of concrete and the age, and analyze how each model fits. We define:

x_1 = Cement x_2 = Blast Furnace Slag x_3 = Fly Ash x_4 = Water x_5 = Superplasticizer
x_6 = Coarse Aggregate x_7 = Fine Aggregate x_8 = Age

2.1 Full Linear Model

First of all, full linear model, includes all of the explanatory variables, is fitted. After the model is fitted, we obtained the model as follows:

$$y = -23.163756 + 0.119785x_1 + 0.103847x_2 + 0.087943x_3 - 0.150298x_4 + 0.290687x_5 + 0.018030x_6 + 0.020154x_7 + 0.114226x_8$$

In this model, we obtained adjusted R2 value of 0.6125, BIC value of 7807.437 and Cp mallow's value of 9. The global f-test shows that the model is significant overall, and the p-values of coefficients also shows that all of the coefficients are significant in explaining the concrete compressive strength at $\alpha = 0.01$ except for coarse aggregate and fine aggregate. At this point, we keep all the variables though there are some insignificant variables for the sake of having this model be our baseline full linear model for the comparison. However, we will further investigate if we can improve the model by excluding the insignificant variables, fitting higher order model, or adding interaction terms.

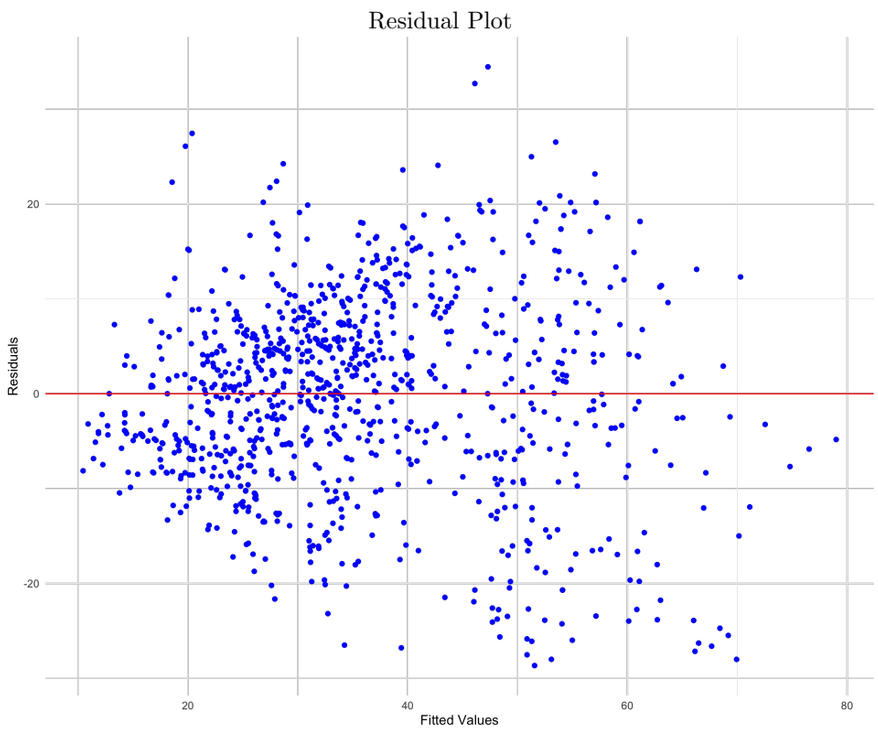

Figure 2: Residual vs Fitted Value for Model 2.1

Observing the residual plot, it seems like there is no major pattern, though the variance with small fitted value is slightly smaller than that with large fitted value. This would not be enough evidence to invalidate the model, but we may want to try different models to see how they behave.

2.2 Linear Model: Best Subset Selection

In order to obtain the linear model which fits better than the full linear model, we try to find the combination of variables which can predict the concrete compressive strength accurately. One of the approaches is exhaustive search, which tries all possible combinations of predictors to find the best model.

We use regsubsets function from leaps package. This function carries out exhaustive search, and returns the best model for each number of variables. In this case, the maximum predictors are 8, so we obtain the best combination for each 1 to 8 variables. The results are summarized in Table 2 and Table 3.

Best Subsets Regression

Model Index	**Predictors**
1	Cement
2	Cement, Superplasticizer
3	Cement, Superplasticizer, Age
4	Cement, Blast Furnance Slag, Age, Water
5	Cement, Blast Furnance Slag, Age, Water, Fly Ash
6	Cement, Blast Furnance Slag, Age, Water, Fly Ash, Superplasticizer
7	Cement, Blast Furnance Slag, Age, Water, Fly Ash, Superplasticizer. Coarse Agg.
8	Cement, Blast Furnance Slag, Age, Water, Fly Ash, Superplasticizer. Coarse Agg., Fine Agg.

Table 2: Variables included in best model of each size selected by exhaustive search

Subsets Regression Summary

Model	Adj R-Square	C_p	BIC
1	0.2471	971.1068	8449.920
2	0.3498	698.9999	8304.796
3	0.4801	354.3021	8080.327
4	0.5560	154.2440	7923.725
5	0.6091	14.9545	7798.615
6	0.6117	8.9555	7797.545
7	0.6115	10.5461	7804.209
8	0.6125	9.0000	7807.437

Table 3: Model selection criteria summary

Looking at the summary, we choose the model with 8 predictors as the best linear model. Though the model with 6 predictors have the lowest BIC value (lower the BIC, the better model), C_p value of the model with 8 predictors is very close to the number of predictors, and it also has the highest adjusted R-squared value, making it the best linear model.

Therefore, the best linear model found by the exhaustive search is:

$$\begin{aligned} y = -23.163756 + 0.119785x_1 + 0.103847x_2 + 0.087943x_3 - 0.150298x_4 \\ +0.290687x_5 + 0.018030x_6 + 0.020154x_7 + 0.114226x_8 \end{aligned}$$

This is the full linear model found in the previous section.

2.3 Polynomial Regression

We also considered adding some polynomial terms to our model in order to obtain better fit. Back in section 1.3, we found that water, superplasticizer, and age have a concave pattern against the response variable, so we started by adding 2nd degree polynomial terms of those variables. After fitting the function, we found that the polynomial term of water and linear terms of coarse aggregate and fine aggregate are insignificant. Here, we do not remove coarse aggregate and fine aggregate, since we found in the previous section that removing those variables lowered the adjusted R-squared value (comparing model 6 and model 8). We can assume that they are contributing to explain the response variable, though they are insignificant. After deleting the polynomial term of water, we obtained adjusted R-squared value of 0.76 with BIC and Cp values of 7327.477 and -379.8752, respectively. This gave us the model of:

$$\begin{aligned} y = 19.71 + 0.1085x_1 - 0.08467x_2 + 0.04238x_3 - 0.173x_4 \\ + 66.78x_5 - 86.49x_5^2 + 0.004171x_6 + 0.005667x_7 + 228.9x_8 - 190.8x_8^2 \end{aligned}$$

Figure 3: Residual vs Fitted Value for Model 2.3

Observing the residual plot above, it seems like it does not have any major patterns. Therefore, this does not help on deciding if this model is valid or not.

2.4 Interaction Model

In addition, we tried to fit a model including some interaction terms to the linear model. We started by adding all possible two-way interaction terms and chose the terms that are significant at $\alpha = 0.001$ to be included in the model. The 8 interaction terms are included in the model, and fitted again. Then, we further removed the insignificant interaction terms. The interaction terms included in the final model are: blast furnace slag*fine aggregate, blast furnace slag*fly ash, water*coarse aggregate, blast furnace slag*age, fly ash*fine aggregate, and fly ash*age.

The model is:

$$y = -271.8 + 0.1283x_1 - 0.06029x_2 - 0.1485x_3 - 1.261x_4 + 0.4147x_5 + 0.2749x_6 + 0.009424x_7 + 0.06835x_8 + 0.0002x_2x_7 + 0.0002248x_2x_3 - 0.001417x_4x_6 + 0.0003216x_2x_8 + 0.0001972x_3x_7 + 0.001896x_3x_8$$

We obtained adjusted R squared value of 0.6914, Cp value of -191.6932, and BIC value of 7608.457. The residual plot of our interaction model is shown below:

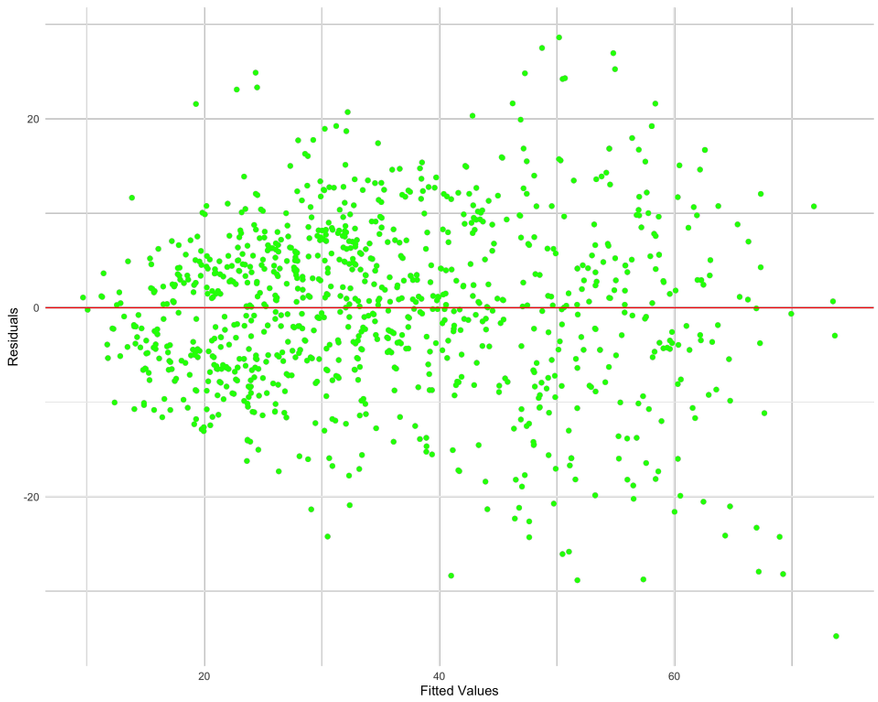

Figure 4: Residual vs Fitted Value for Model 2.4

We can see that this graph also does not have any pattern except that there is a outlier in the right bottom corner. The points are approximately evenly plotted to each side of the horizontal line.

2.5 Cross Validation

5-fold cross validation was carried out in order to compare the performance of 3 models fitted in the above sections, and to see which model performs the best in predicting the concrete compressive strength.

Procedure of the test is the following:
1.Randomly and equally split the samples into 5 groups
2.Take one group as test set and the other 4 sets as training set
3.Fit the model with the training set and find RMSE values for the test set
4.Repeat steps 2 and 3 for 5 times and take the average of the RMSE values

Model	RMSE	adj R^2	C_p	BIC
Full linear	10.46027	0.6125	9.000	7807.437
Polynomial	8.237803	0.7596	-379.8752	7327.477
Interaction	9.396861	0.6914	-191.6932	7608.457

Table 4: RMSE values from 5-fold cross validation + model summary

The result above shows that the polynomial model performs the best among those 3 models. Though it has the negative C_p value, it has the smallest RMSE value, the largest adjusted R-squared value, and the smallest BIC value, making it the best model to predict the concrete compressive strength.

3 Conclusion

In conclusion, we fitted three different models to predict the concrete compressive strength; full linear model, model with some polynomial terms, and the model with some interaction terms. After applying cross validation to each model, we found that the model with polynomial terms we developed is the best model we can find so far. The main limitation of this study is that only 3 regression models are examined in this study, though there are many other possible regression models that can be examined such as including interaction terms with more than 2 variables, including higher degree polynomial terms and including both polynomial and interaction terms. The other limitation is that the data only contains the concrete with age of 365 days or younger, thus it is questionable that if the result can be generalized. We hope that further research on the relationship between the concrete compressive strength and concrete formulation components and concrete age could be done to obtain the better model.

References

[1] Ramadhansyah Putra Jaya. (2020). *New Materials in Civil Engineering.* https://www.sciencedirect.com/science/article/pii/B9780128189610000144

[2] Yeh,I-Cheng. (2007). *Concrete Compressive Strength. UCI Machine Learning Repository.* https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength