**CONCRETE COMPRESSIVE STRENGTH**

MATH 644 FALL 2019 FINAL PROJECT

Submitted by: Chhavi Tyagi

**Description of the Dataset**

The data set was found on the UCI Machine Learning Repository. It was chosen for three reasons: a large number of observations (~1000) for ensure there were many more observations than features, a moderate number of attributes (n=8) for quick analysis, and a single dependent variable for multivariable Linear Regression which is the Compressive Strength of the concrete being made. The large number of observations also enabled 10-fold cross validation which became important at the end of the project to examine potential over-fitting.

NOTE: Reuse of this database is unlimited with retention of copyright notice for Prof. I-Cheng Yeh and the following published paper:

Dataset Citation **:** 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).

**Introduction to the study:**

Per the paper authors:

*“Concrete is the most important material in civil engineering. The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate…Several studies independently have shown that concrete strength development is determined not only by the water-to-cement ratio, but that it also is inﬂuenced by the content of other concrete ingredients.*

The author’s comment modeling concrete compressive strength is challenging - keep in mind that both physical interactions and time-dependent chemical reactions are occurring to create cement so it is not surprising that the project ultimately found many important pair-wise interactions and multiple transformations were required to afford the best models.

*“...High-performance concrete is a highly complex material, which makes modeling its behavior a very difﬁcult task. This paper is aimed at demonstrating the possibilities of adapting artiﬁcial neural networks (ANN) to predict the compressive strength of high-performance concrete…..This study led to the following conclusions: 1) A strength model based on ANN is more accurate than a model based on regression analysis; and 2) It is convenient and easy to use ANN models for numerical experiments to review the effects of the proportions of each variable on the concrete mix…”*

**Objective of the study:**

The Math644 project objective was to build more accurate models on Concrete Compressive Strength with linear regression.This is in contrast to the paper’s authors who abandoned linear regression and moved to other algorithms. The work in this Math 644 project is unique and does not leverage the analysis in the article. In particular, this study fills in the blank between the author’s basic attempts at simple linear regression and their move to ANN.

**Data Set**

UCI ML Data Set Main Page: <https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength>

Links: <https://archive.ics.uci.edu/ml/machine-learning-databases/concrete/compressive/Concrete_Readme.txt>

**Attribute Information:**

Given are the variable name, variable type, the measurement unit and a brief description. The concrete compressive strength is the regression problem. The order of this listing corresponds to the order of numerals along the rows of the database.

*“...Data Characteristics:*

*The actual concrete compressive strength (MPa) for a given mixture under a specific age (days) was determined from laboratory. Data is in raw form (not scaled).*

*Data Set Information:*

*Number of instances 1030  
Number of Attributes 9  
Attribute breakdown 8 quantitative input variables, and 1 quantitative output variable  
Missing Attribute Values None..”*

*Name -- Data Type -- Measurement -- Description  
Cement (component 1) -- quantitative -- kg in a m3 mixture -- Input Variable  
Blast Furnace Slag (component 2) -- quantitative -- kg in a m3 mixture -- Input Variable  
Fly Ash (component 3) -- quantitative -- kg in a m3 mixture -- Input Variable  
Water (component 4) -- quantitative -- kg in a m3 mixture -- Input Variable  
Superplasticizer (component 5) -- quantitative -- kg in a m3 mixture -- Input Variable  
Coarse Aggregate (component 6) -- quantitative -- kg in a m3 mixture -- Input Variable  
Fine Aggregate (component 7) -- quantitative -- kg in a m3 mixture -- Input Variable  
Age -- quantitative -- Day (1~365) -- Input Variable  
Concrete compressive strength -- quantitative -- MPa -- Output Variable…”*

**Water-to-Binder Ratio**

There are fundamentally three ingredients in concrete: aggregate which is fine and coarse rock, water, and various binders like Portland cement to react and bind the rock together. Additives like slag, ash, and plasticizer act in concert with cement as supplementary binders. Within the main article above, the authors describe that the water-to-cement ratio (w/c), is historically valuable in understanding concrete strength. We referred to another article (<http://www.ajer.org/papers/Vol-8-issue-1/V0801172183.pdf>) that further emphasizes the role of the more complex variable, water-to-binder ratio (w/b), and how it plays an important role towards modeling the complexities of concrete compressive strength. The given formula was used to calculate w/b ratio and we added this new feature, “WaterBind”, to the data set.

In the original article, the w/c and w/b ratios were used to build regression formulas of a single type:

where t = age at test; X = w/c or w/b ratio; and a, b, c, and d are regression coefﬁcients. These clearly included a limited number of non-linear transformations. Our project included many other transformed variables and interactions.

**Approach to modeling**

We took a three step approach to modeling the dataset.

1. We created simple models with limited features in a trial-and-error fashion to understand sensitivities in the data, build in scientific principles, study if pair-wise interactions were valuable, and determine if we should develop new features which the original articles authors chose not to do. This semi-manual effort led to a much-improved “hand-model” we denoted ‘Reg4’.
2. Expanding on the learning from the Reg4 hand-model, we then created a full model of many types of potential transformed variables and all pair-wide interactions and used AIC/BIC criteria to perform feature reduction and produced a reduced “machine-model”, that we denoted ‘Reg3’. We removed some low-value features from this model to produce a simpler version of Reg3 we denoted ‘Reg5’. The machine models better allowed us to understand the main effects and interactions by looking at a comprehensive set of features.
3. In order to understand if over-fitting was occurring and to compare models, we studied Reg3, Reg4, and Reg5 via 10-fold cross-validation and compared RSME versus the number of features to choose a ‘best’ model which struck a balance between the best prediction of the real measurement and an economical number of features for efficiency in practice.

**Statistical Analysis**

Using only the original raw variables, the following summary and correlations were generated in R to describe the initial dataset. Because the data came from a designed experiment, several of the variables have many points where the value is zero, meaning they were left out of the ingredient mix for that observation. In some cases, transformations for regression demanded that a value of 1 was added to all observations of an independent variable value to enable functions such as 1/x or log(x) which do not exist at zero.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | Cement | BlastF | FlyAsh | Water | SuperP | Coarse | Fine | Age | Strength |
| Minimum | 102.0 | 0.0 | 0.00 | 121.8 | 0.000 | 801.0 | 594.0 | 1.00 | 2.332 |
| 1st Quartile | 192.4 | 0.0 | 0.00 | 164.9 | 0.000 | 932.0 | 731.0 | 7.00 | 23.707 |
| Median | 272.9 | 22.0 | 0.00 | 185.0 | 6.350 | 968.0 | 779.5 | 28.00 | 34.443 |
| Mean | 281.2 | 73.9 | 54.19 | 181.6 | 6.203 | 972.9 | 773.6 | 45.66 | 35.818 |
| 3rd Quartile | 350.0 | 142.9 | 118.27 | 192.0 | 10.160 | 1029.4 | 824.0 | 56.00 | 46.136 |
| Maximum | 540.0 | 359.4 | 200.10 | 247.0 | 32.200 | 1145.0 | 992.6 | 365.00 | 82.599 |

Correlation Coefficients:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Cement | BlastF | FlyAsh | Water | SuperP | Coarse | Fine | Age | Strength | WaterBind |
| Cement | 1.000 |  |  |  |  |  |  |  |  |  |
| BlastF | -0.275 | 1.000 |  |  |  |  |  |  |  |  |
| FlyAsh | -0.397 | -0.324 | 1.000 |  |  |  |  |  |  |  |
| Water | -0.082 | 0.107 | -0.257 | 1.000 |  |  |  |  |  |  |
| SuperP | 0.093 | 0.043 | 0.377 | -0.657 | 1.000 |  |  |  |  |  |
| Coarse | -0.109 | -0.284 | -0.010 | -0.182 | -0.266 | 1.000 |  |  |  |  |
| Fine | -0.223 | -0.282 | 0.079 | -0.451 | 0.223 | -0.179 | 1.000 |  |  |  |
| Age | 0.082 | 0.044 | -0.154 | 0.278 | -0.193 | -0.003 | -0.156 | 1.000 |  |  |
| Strength | 0.498 | 0.135 | -0.106 | -0.290 | 0.366 | -0.165 | -0.167 | 0.329 | 1.000 |  |
| WaterBind | -0.458 | -0.269 | -0.161 | 0.547 | -0.628 | 0.225 | 0.231 | 0.157 | -0.618 | 1.000 |

Then we take a new model with interaction terms with water to bind ratio and obtain the following correlation coefficient.

newxy<-data.frame(FineSuperP, FineCement, FineFly, FSW, FFW, LogOneOverAge, Fine, SuperP, Strength, Predicted)

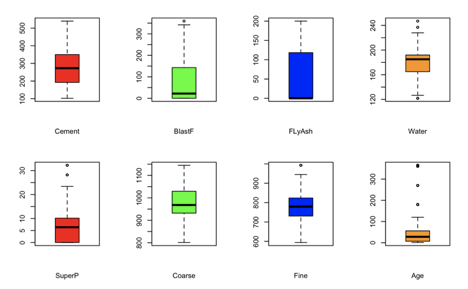
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FineSuperP | FineCement | FineFly | FSW | FFW | LogOneOverAge | Fine | SuperP | Strength | Predicted |
| FineSuperP | 1.000 |  |  |  |  |  |  |  |  |  |
| FineCement | 0.2383 | 1.000 |  |  |  |  |  |  |  |  |
| FineFly | 0.3546 | -0.3693 | 1.000 |  |  |  |  |  |  |  |
| FSW | 0.9866 | 0.1811 | 0.4133 | 1.000 |  |  |  |  |  |  |
| FFW | 0.3325 | -0.3777 | 0.9894 | 0.4040 | 1.000 |  |  |  |  |  |
| LogOneOverAge | 0.0441 | 0.0417 | 0.0160 | 0.0386 | 0.0123 | 1.000 |  |  |  |  |
| Fine | 0.3071 | 0.0897 | 0.1369 | 0.2724 | 0.1066 | 0.1139 | 1.000 |  |  |  |
| SuperP | 0.9896 | 0.2017 | 0.3764 | 0.9882 | 0.3578 | 0.0430 | 0.2225 | 1.000 |  |  |
| Strength | 0.3547 | 0.4650 | -0.1105 | 0.3329 | -0.1299 | -0.5522 | -0.1672 | 0.3661 | 1.000 |  |
| Predicted | 0.3917 | 0.5135 | -0.1221 | 0.3676 | -0.1435 | -0.6097 | -0.1847 | 0.4043 | 0.9056 | 1.000 |

**Model Scoring**

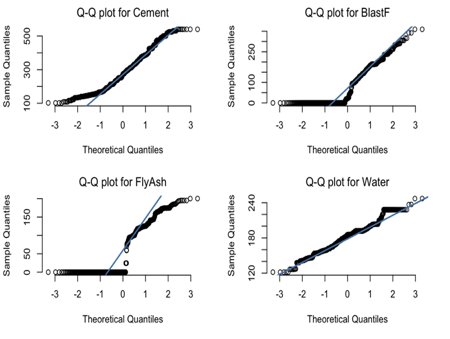
In all cases, regression models were evaluated using RMSE, R2, and Adj-R2, where RMSE = Root Mean Squared Error = , R2 = R2 score, and Adj-R2 = adjusted R2. Except for the pair-wise term generated by the machine regressions, all the added polynomial and non-linear features were pre-calculated and entered into the R lm functions as apparently-linear terms so the reported R2 values seem appropriate. However, it should be noted that these models are in-fact polynomial and other wise complex, and R2  would not be expected to be the correct measure. So, RMSE was the primary gauge of the model, though the R2 values move directionally with the RMSE.

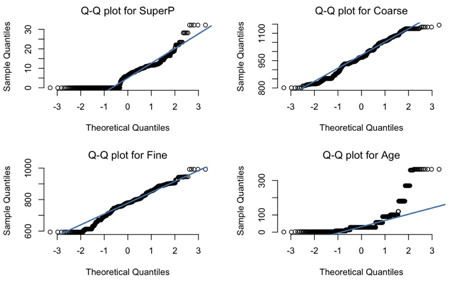
**Outliers**

We plotted box plots to look for outliers. Given that the recipes were predetermined in the paper at each point, we did not expect outliers. There are few points outside the 1.5\*IQR criteria so no outliers were identified or removed. The ‘outliers’ in AgeTime again are not outliers - we know scientifically reactions like concrete hardening can be aged indefinitely.

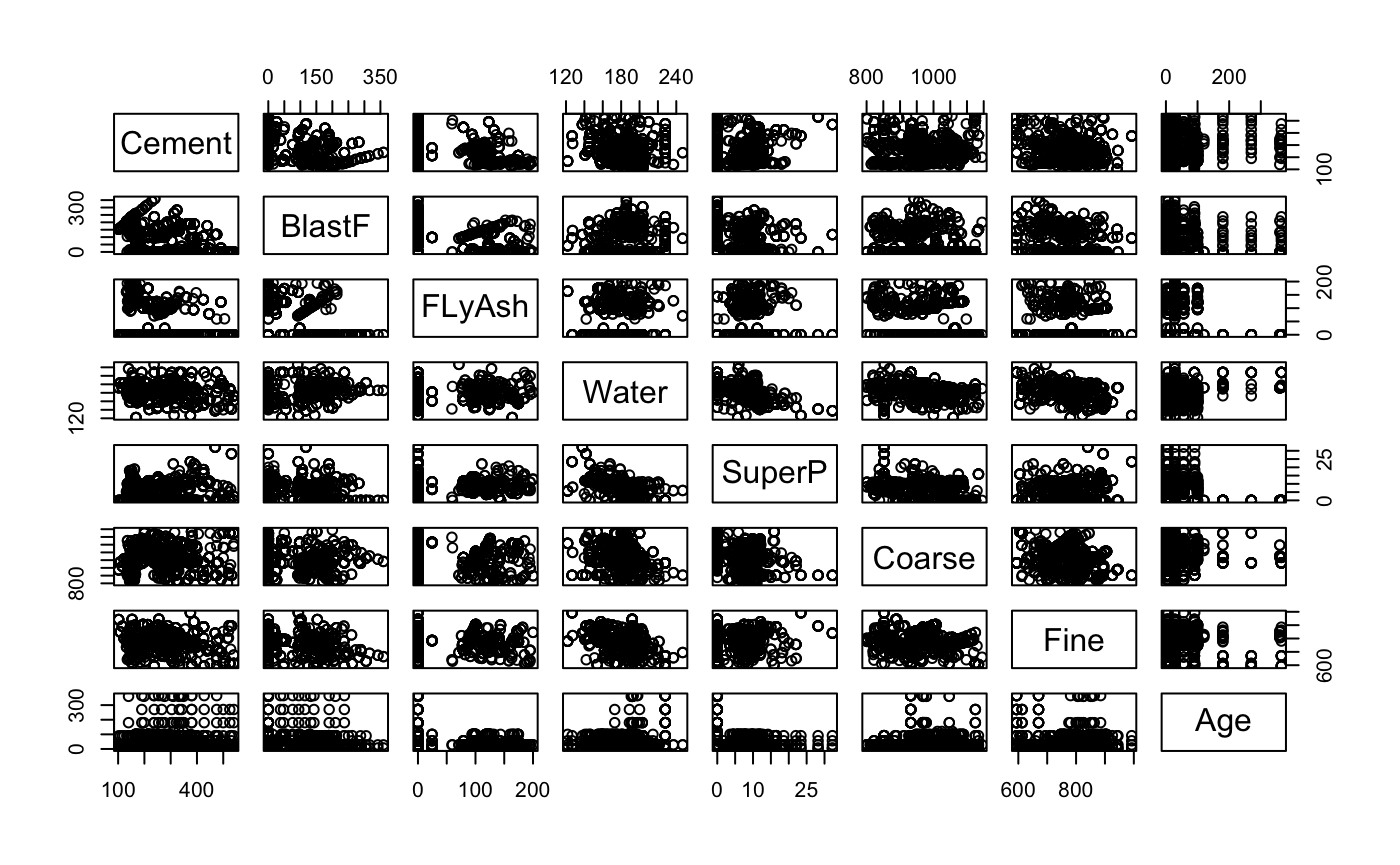


We also plotted sample quantiles with theoretical quantiles for all the features in order to look for linearity of the underlying independent variable and for outliers. Five of the Q-Q plots were clearly linear. The Q-Q plot for Age time was non-linear but that is to be expected - the rate of a chemical reaction may not be linear and the author’s own regression model used the log of Age. The plots for Fly Ash and Super Plasticizer are not perfectly linear but are linear for portions of the plot. Again, given that recipes were prepared intentionally in a designed experiment, it did not seem appropriate to remove any points from the study.





We also studied the pairwise relationships between the independent variables. Given that the experiments were planned with preset mixes of the concrete, we did not expect to see relationships. From these plots one can see the planned nature of the experiments, for example, in the even spacing between points particularly in the Age column and row. The plots were helpful to point out that there were many zero values which had to be carefully corrected when using 1/x and log(x) functions.



**Model Building:**

We use several methods to generate the regression models.

**1.** **Reg - Simple Linear Model with 8 predictors.**

lm(formula = Strength ~ Cement + BlastF + FlyAsh + Water + SuperP +

Coarse + Fine + Age)

Residuals:

Min 1Q Median 3Q Max

-28.653 -6.303 0.704 6.562 34.446

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -23.163756 26.588421 -0.871 0.383851

Cement 0.119785 0.008489 14.110 < 2e-16 \*\*\*

BlastF 0.103847 0.010136 10.245 < 2e-16 \*\*\*

FlyAsh 0.087943 0.012585 6.988 5.03e-12 \*\*\*

Water -0.150298 0.040179 -3.741 0.000194 \*\*\*

SuperP 0.290687 0.093460 3.110 0.001921 \*\*

Coarse 0.018030 0.009394 1.919 0.055227 .

Fine 0.020154 0.010703 1.883 0.059968 .

Age 0.114226 0.005427 21.046 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 1021 degrees of freedom

Multiple R-squared: 0.6155, Adjusted R-squared: 0.6125

F-statistic: 204.3 on 8 and 1021 DF, p-value: < 2.2e-16

Discussion: This model does not include any interactions, transformations, polynomials, nor the scientifically-important parameter water-to-binder ratio (w/b)

Interpretation: R-squared is 0.6155 which is the starting point for the work. FIne and Coarse features were found insignificant (Significance code 0.1).

**2.** **Reg1 - Model after removing Fine and Coarse**

Call:

lm(formula = Strength ~ Cement + BlastF + FlyAsh + Water + SuperP +

Age)

Residuals:

Min 1Q Median 3Q Max

-29.014 -6.474 0.650 6.546 34.726

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 29.030224 4.212476 6.891 9.64e-12 \*\*\*

Cement 0.105427 0.004248 24.821 < 2e-16 \*\*\*

BlastF 0.086494 0.004975 17.386 < 2e-16 \*\*\*

FlyAsh 0.068708 0.007736 8.881 < 2e-16 \*\*\*

Water -0.218292 0.021128 -10.332 < 2e-16 \*\*\*

SuperP 0.239003 0.084586 2.826 0.00481 \*\*

Age 0.113495 0.005408 20.987 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.41 on 1023 degrees of freedom

Multiple R-squared: 0.614, Adjusted R-squared: 0.6117

F-statistic: 271.2 on 6 and 1023 DF, p-value: < 2.2e-16

Discussion: The fine and coarse features were removed from the initial Reg model.

Interpretation: The R-squared value of 0.6140 is essentially unchanged from the initial model. This change was not beneficial.

**3.** **Regwb Model which includes Waterbind ratio as an additional feature.**

Call:

lm(formula = Strength ~ Coarse + Fine + Age + WaterBind)

Residuals:

Min 1Q Median 3Q Max

-36.312 -7.283 0.819 7.487 35.627

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 60.463138 6.126821 9.869 < 2e-16 \*\*\*

Coarse 0.001605 0.004664 0.344 0.73080

Fine 0.014084 0.004620 3.048 0.00236 \*\*

Age 0.118999 0.005607 21.223 < 2e-16 \*\*\*

WaterBind -91.800716 2.911627 -31.529 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

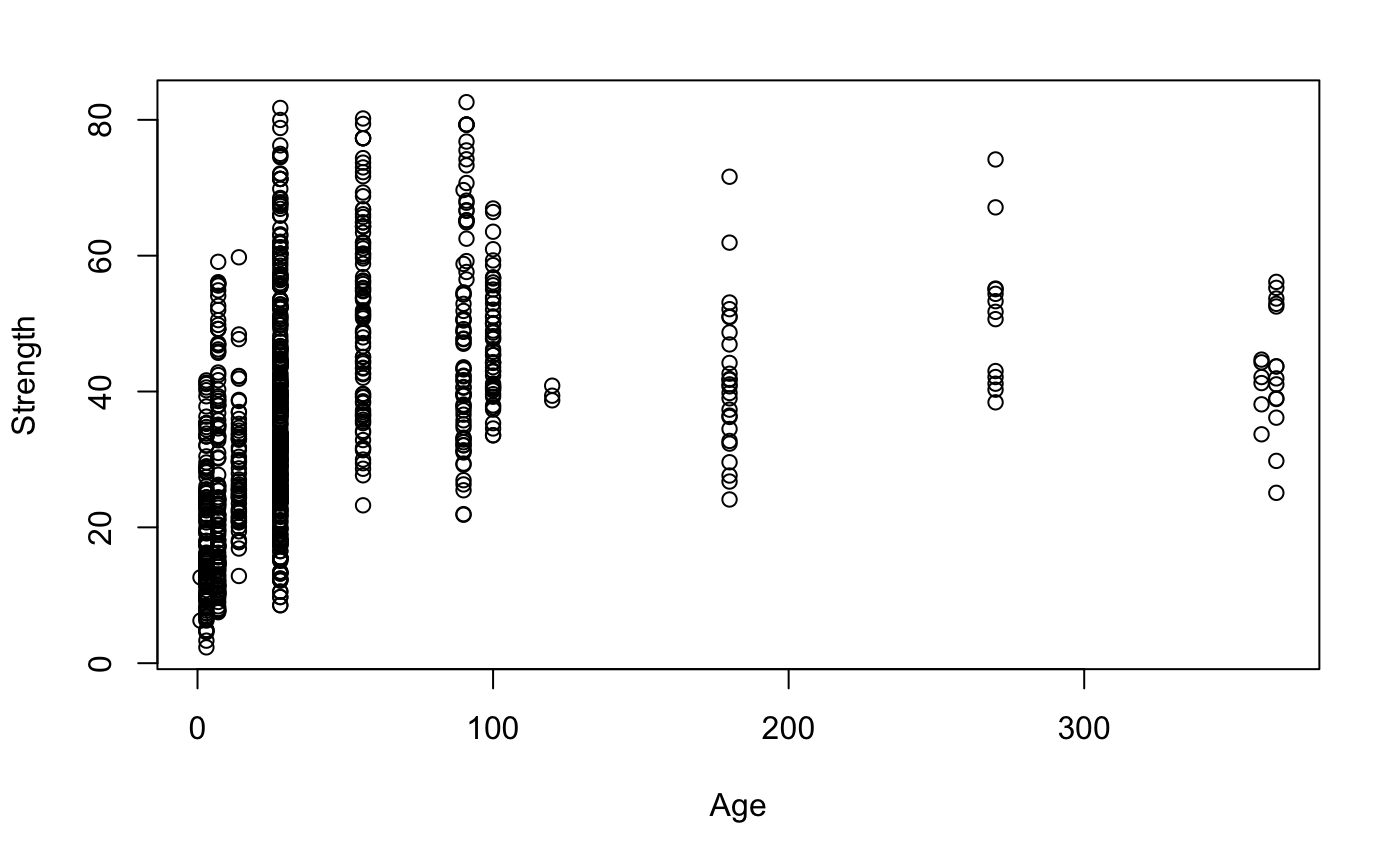
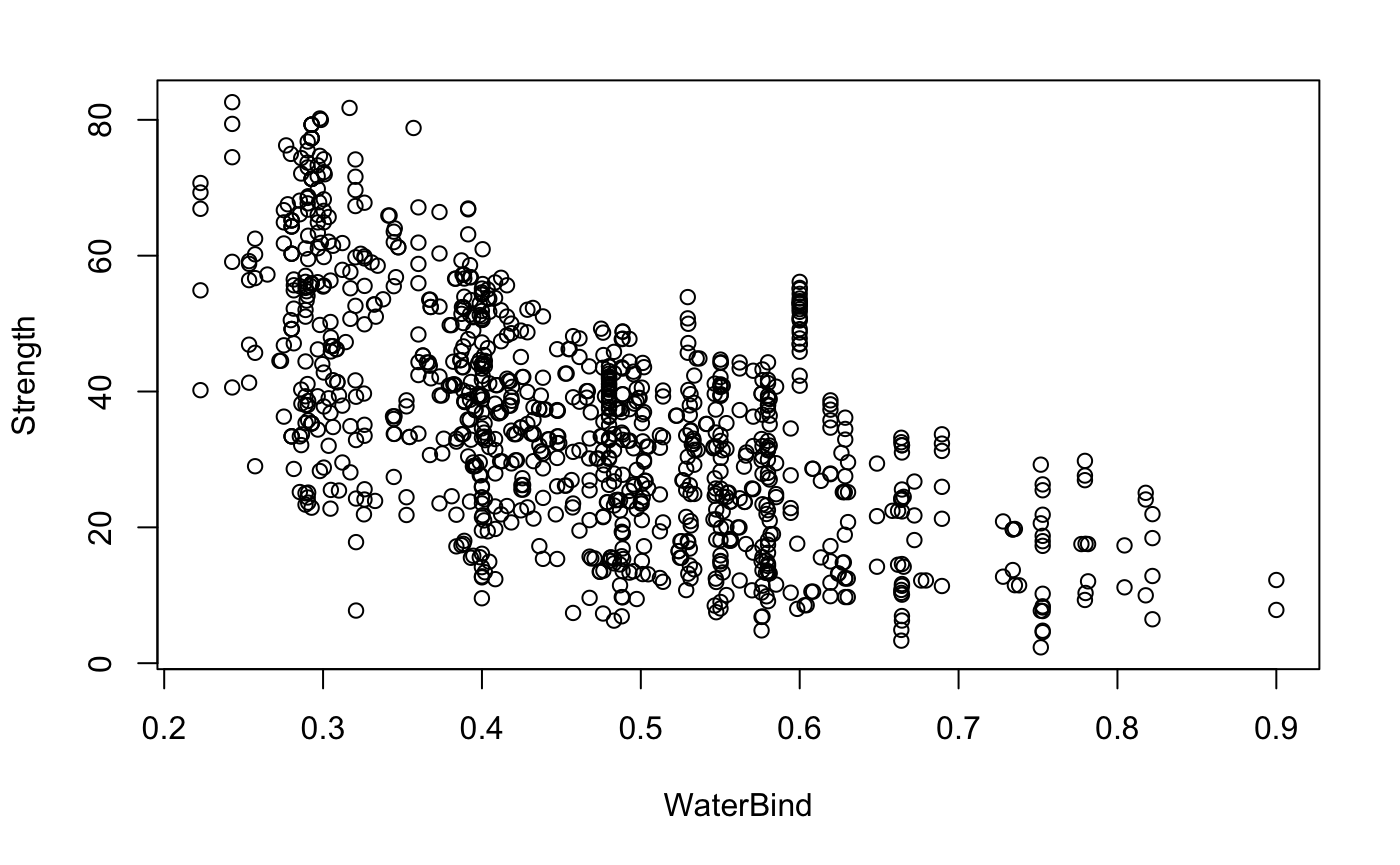
Residual standard error: 10.95 on 1025 degrees of freedom

Multiple R-squared: 0.5722, Adjusted R-squared: 0.5706

F-statistic: 342.8 on 4 and 1025 DF, p-value: < 2.2e-16

Discussion: The removal of Fine and Coarse was not helpful so they were put back into the models. The scientifically-important feature water-to-binder ratio (w/b, WaterBind) was added to the initial model Reg to include a feature that is potentially more influential based on journal articles.

Interpretation: The R-squared value of 0.5722 was not improved versus the prior models. However, several features contained within the calculated WaterBind feature were found to be insignificant (Significance code 0.1) suggesting WaterBind could replace those terms, so it was retained. Also, the journal articles on concrete also indicated known relationships to the water-to-binder ratio, suggesting an inverse relations, which is visible in the Concrete dataset using the calculated w/b variable (below, left). Also visible is the non-linear relationship to Age time (below,right).



**4.** **RegA2 Model which includes Age squared term in addition to waterbind ratio**

Call:

lm(formula = Strength ~ Fine + WaterBind + Age + AgeSquared)

Residuals:

Min 1Q Median 3Q Max

-30.572 -6.311 -0.566 5.750 35.916

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.847e+01 2.921e+00 20.018 <2e-16 \*\*\*

Fine 9.024e-03 3.818e-03 2.364 0.0183 \*

WaterBind -8.797e+01 2.384e+00 -36.896 <2e-16 \*\*\*

Age 3.465e-01 1.257e-02 27.555 <2e-16 \*\*\*

AgeSquared -7.723e-04 3.947e-05 -19.564 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.342 on 1025 degrees of freedom

Multiple R-squared: 0.6885, Adjusted R-squared: 0.6873

F-statistic: 566.4 on 4 and 1025 DF, p-value: < 2.2e-16

Discussion: Based on the unusual Q-Q plot, and the time-based chemical reactions occurring, the age term was thought to be important, so an AgeSquared term was added to the model.

Interpretation: The R-squared value of 0.6885 was improved versus the prior models using WaterBind to provide the influence of other features. An additional model (not shown) with Age-cubed was not better (R-squared = 0.664) so further work was done with polynomials up to age^2 only.

**5.** **RegBW Model where we have log(1/waterbind) as a feature instead of waterbind**

Call:

lm(formula = Strength ~ SuperP + Fine + LogBW + LogAge)

Residuals:

Min 1Q Median 3Q Max

-28.928 -4.836 0.165 4.307 33.900

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -43.705285 3.347364 -13.057 < 2e-16 \*\*\*

SuperP -0.413537 0.061430 -6.732 2.78e-11 \*\*\*

Fine 0.017682 0.003546 4.987 7.20e-07 \*\*\*

LogBW 47.379106 1.297408 36.518 < 2e-16 \*\*\*

LogAge 9.288272 0.223580 41.543 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.843 on 1025 degrees of freedom

Multiple R-squared: 0.7804, Adjusted R-squared: 0.7796

F-statistic: 910.8 on 4 and 1025 DF, p-value: < 2.2e-16

Discussion: Based on their importance in the literature, the next model examined the role of WaterBind and Age using log transformations. In order to facilitate calculations containing zero values, log transformations were of the type log (x+1).

Interpretation: The R-squared value of 0.7804 was again improved versus the prior models using log transformations of key variables.

**6.** **RegInterect Model which contains several interaction terms.**

Call:

lm(formula = Strength ~ FineSuperP + FineBW + FSW + FFW + FineCement +

FineB + FineFly + FineW + SuperP + Fine + LogBW + LogAge)

Residuals:

Min 1Q Median 3Q Max

-22.608 -4.252 -0.067 4.452 29.579

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.674e+01 9.204e+00 -3.991 7.04e-05 \*\*\*

FineSuperP -2.979e-03 6.117e-04 -4.870 1.29e-06 \*\*\*

FineBW -3.925e-03 5.583e-03 -0.703 0.482224

FSW 8.428e-06 4.006e-06 2.104 0.035658 \*

FFW -4.497e-07 2.867e-07 -1.569 0.117036

FineCement 7.746e-05 4.169e-05 1.858 0.063454 .

FineB 4.191e-05 4.201e-05 0.998 0.318720

FineFly 7.674e-05 6.411e-05 1.197 0.231591

FineW -1.368e-04 9.648e-05 -1.418 0.156538

SuperP 1.244e+00 4.174e-01 2.980 0.002955 \*\*

Fine 3.116e-02 1.383e-02 2.253 0.024467 \*

LogBW 3.165e+01 9.153e+00 3.458 0.000568 \*\*\*

LogAge 9.280e+00 2.085e-01 44.512 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.163 on 1017 degrees of freedom

Multiple R-squared: 0.8183, Adjusted R-squared: 0.8162

F-statistic: 381.7 on 12 and 1017 DF, p-value: < 2.2e-16

Discussion: Based on the nature of chemical reactions, the project team decided to investigate interactions between components. Fine was chosen as a test case because the reactions to form cement require binding to the aggregate particles.

Interpretation: The R-squared value of 0.8181 was again improved versus the prior models with one interactions of significance: FSW = Fine \* Water \* SuperP (plasticizer). This three component interaction suggested other interactions may be worth studying.

**7.** **RegLog Model which has log features.**

Call:

lm(formula = Strength ~ FineSuperP + FSW + FFW + FineCement +

FineFly + SuperP + Fine + LogBW + LogOneOverAge)

Residuals:

Min 1Q Median 3Q Max

-22.646 -4.200 -0.126 4.433 29.618

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.210e+01 3.949e+00 -10.659 < 2e-16 \*\*\*

FineSuperP -2.950e-03 4.765e-04 -6.191 8.66e-10 \*\*\*

FSW 7.902e-06 2.845e-06 2.778 0.005576 \*\*

FFW -6.580e-07 2.345e-07 -2.806 0.005109 \*\*

FineCement 3.632e-05 4.031e-06 9.012 < 2e-16 \*\*\*

FineFly 7.334e-05 3.996e-05 1.835 0.066748 .

SuperP 1.262e+00 3.875e-01 3.257 0.001162 \*\*

Fine 1.727e-02 4.430e-03 3.899 0.000103 \*\*\*

LogBW 3.878e+01 1.494e+00 25.957 < 2e-16 \*\*\*

LogOneOverAge -8.621e+00 1.905e-01 -45.254 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.116 on 1020 degrees of freedom

Multiple R-squared: 0.8201, Adjusted R-squared: 0.8186

F-statistic: 516.8 on 9 and 1020 DF, p-value: < 2.2e-16

Discussion: Non-significant interaction were removed and the next model studied 1/log(age) versus log(age).

Interpretation: The R-squared value of 0.8201 was roughly the same suggesting no specific difference.

**8.** **Reg4 is our hand made model with all the good features from the above models included.**

Call:

lm(formula = Strength ~ WB + FSW + FFW + FineCement + FineB +

FineW + Fine + LogWB + LogAge + OneOverFlyAsh + OneOverWater +

OneOverBlastF + OneOverFine + OneOverAge + SquaredFlyAsh +

SquaredWater + SquaredBlastF + SquaredSuperP + SquaredCoarse +

SquaredFine + SquaredAge + SquaredBW + LogCement + LogFlyAsh +

LogWater + LogSuperP + LogCoarse + LogFine)

Residuals:

Min 1Q Median 3Q Max

-26.9493 -3.7268 0.0589 3.8745 23.9077

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.852e+05 5.611e+04 3.302 0.000995 \*\*\*

WB 6.242e+03 9.916e+02 6.295 4.58e-10 \*\*\*

FSW -1.757e-05 2.098e-06 -8.373 < 2e-16 \*\*\*

FFW -6.739e-07 2.627e-07 -2.566 0.010444 \*

FineCement -3.546e-04 4.557e-05 -7.782 1.77e-14 \*\*\*

FineB -2.225e-04 4.827e-05 -4.609 4.57e-06 \*\*\*

FineW -1.307e-03 2.454e-04 -5.324 1.25e-07 \*\*\*

Fine 3.595e+01 1.207e+01 2.979 0.002966 \*\*

LogWB -7.964e+03 1.190e+03 -6.694 3.62e-11 \*\*\*

LogAge 8.889e+00 5.089e-01 17.468 < 2e-16 \*\*\*

OneOverFlyAsh -4.464e+01 8.429e+00 -5.296 1.45e-07 \*\*\*

OneOverWater -1.600e+05 2.035e+04 -7.862 9.77e-15 \*\*\*

OneOverBlastF 5.953e+00 1.037e+00 5.742 1.24e-08 \*\*\*

OneOverFine -7.738e+06 2.320e+06 -3.336 0.000882 \*\*\*

OneOverAge -1.425e+01 6.112e+00 -2.331 0.019962 \*

SquaredFlyAsh -3.593e-04 1.097e-04 -3.276 0.001090 \*\*

SquaredWater 7.812e-03 9.139e-04 8.548 < 2e-16 \*\*\*

SquaredBlastF -2.534e-04 3.418e-05 -7.415 2.59e-13 \*\*\*

SquaredSuperP 2.345e-02 6.040e-03 3.883 0.000110 \*\*\*

SquaredCoarse 7.220e-05 1.651e-05 4.373 1.36e-05 \*\*\*

SquaredFine -7.358e-03 2.622e-03 -2.806 0.005113 \*\*

SquaredAge -9.275e-05 1.431e-05 -6.481 1.42e-10 \*\*\*

SquaredBW -1.217e+03 2.190e+02 -5.557 3.52e-08 \*\*\*

LogCement 1.151e+01 3.226e+00 3.568 0.000376 \*\*\*

LogFlyAsh -1.311e+01 2.334e+00 -5.616 2.54e-08 \*\*\*

LogWater -1.083e+03 1.969e+02 -5.500 4.81e-08 \*\*\*

LogSuperP 8.809e+00 1.037e+00 8.492 < 2e-16 \*\*\*

LogCoarse -1.145e+02 3.146e+01 -3.639 0.000288 \*\*\*

LogFine -2.871e+04 9.189e+03 -3.125 0.001831 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.893 on 1001 degrees of freedom

Multiple R-squared: 0.8789, Adjusted R-squared: 0.8756

F-statistic: 259.6 on 28 and 1001 DF, p-value: < 2.2e-16

Discussion: Based on the potential improvements from a variety of transformations, the project team manually constructed a “hand model” including many different types of transformations : 1/x, x^2, and log(x) for the significant features found in prior models.

Interpretation: The R-squared value of 0.8789 was again a step improvement versus the prior models with every single term found to be significant. This was at the cost of a more complex model with many features. However, the adjusted R-squared and R-squared terms were both increasing and the RMSE was also decreasing through this entire effort suggesting we were not over-fitting the model.

**9.** **RegFull Model where all possible interactions are included.**

Call:

lm(formula = Strength ~ (WB + BW + Cement + FlyAsh + Water +

BlastF + SuperP + Fine + Coarse + Age)^2 + LogBW + LogWB +

LogAge + OneOverCement + OneOverFlyAsh + OneOverWater + OneOverBlastF +

OneOverSuperP + OneOverCoarse + OneOverFine + OneOverAge +

OneOverCement2 + OneOverFlyAsh2 + OneOverWater2 + OneOverBlastF2 +

OneOverSuperP2 + OneOverCoarse2 + OneOverFine2 + OneOverAge2 +

SquaredCement + SquaredFlyAsh + SquaredWater + SquaredBlastF +

SquaredSuperP + SquaredCoarse + SquaredFine + SquaredAge +

SquaredBW + LogCement + LogFlyAsh + LogWater + LogBlastF +

LogSuperP + LogCoarse + LogFine)

Residuals:

Min 1Q Median 3Q Max

-26.0617 -3.0723 -0.2553 3.1652 23.8616

Coefficients: (4 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.036e+07 1.842e+06 -5.623 2.47e-08 \*\*\*

WB 2.311e+04 2.691e+04 0.859 0.390795

BW -7.341e+02 1.260e+03 -0.583 0.560357

Cement -5.561e+00 5.298e+00 -1.050 0.294201

FlyAsh -2.497e+00 5.445e+00 -0.459 0.646603

Water -1.618e+02 2.534e+02 -0.638 0.523406

BlastF -4.422e+00 5.062e+00 -0.874 0.382547

SuperP 5.363e+00 1.029e+01 0.521 0.602546

Fine -1.322e+03 2.280e+02 -5.801 9.02e-09 \*\*\*

Coarse -3.857e+01 1.500e+01 -2.571 0.010297 \*

Age -7.747e-01 4.078e-01 -1.900 0.057775 .

LogBW 3.556e+03 8.305e+03 0.428 0.668578

LogWB -2.680e+04 4.224e+04 -0.635 0.525882

LogAge 1.482e+01 2.568e+00 5.769 1.08e-08 \*\*\*

OneOverCement 3.016e+04 9.352e+04 0.322 0.747157

OneOverFlyAsh -1.142e+03 3.975e+03 -0.287 0.773916

OneOverWater 5.450e+06 7.826e+06 0.696 0.486345

OneOverBlastF -2.090e+02 2.198e+02 -0.951 0.341943

OneOverSuperP 4.833e+01 1.703e+02 0.284 0.776633

OneOverCoarse -3.784e+07 1.378e+07 -2.745 0.006161 \*\*

OneOverFine 7.818e+08 1.318e+08 5.930 4.24e-09 \*\*\*

OneOverAge 4.536e+01 3.298e+01 1.375 0.169380

OneOverCement2 -3.743e+05 2.544e+06 -0.147 0.883072

OneOverFlyAsh2 8.121e+02 3.358e+03 0.242 0.808923

OneOverWater2 -1.202e+08 1.682e+08 -0.714 0.475124

OneOverBlastF2 1.929e+02 1.826e+02 1.056 0.291143

OneOverSuperP2 -4.760e+00 8.660e+01 -0.055 0.956174

OneOverCoarse2 9.002e+09 3.275e+09 2.749 0.006095 \*\*

OneOverFine2 -7.445e+10 1.243e+10 -5.990 2.98e-09 \*\*\*

OneOverAge2 -7.508e+01 5.125e+01 -1.465 0.143292

SquaredCement 4.993e-04 2.769e-03 0.180 0.856924

SquaredFlyAsh -3.907e-03 3.550e-03 -1.101 0.271299

SquaredWater 1.158e-01 1.736e-01 0.667 0.504902

SquaredBlastF -1.645e-04 2.669e-03 -0.062 0.950876

SquaredSuperP 1.262e-01 5.729e-02 2.203 0.027848 \*

SquaredCoarse 1.027e-02 3.924e-03 2.617 0.009017 \*\*

SquaredFine 2.136e-01 3.722e-02 5.739 1.29e-08 \*\*\*

SquaredAge 2.581e-04 7.264e-05 3.553 0.000399 \*\*\*

SquaredBW -3.795e+03 4.050e+03 -0.937 0.348999

LogCement 3.261e+02 6.062e+02 0.538 0.590755

LogFlyAsh -9.845e+01 1.570e+02 -0.627 0.530766

LogWater 4.651e+04 6.754e+04 0.689 0.491217

LogBlastF -3.657e+00 9.852e+00 -0.371 0.710546

LogSuperP 3.766e+01 4.121e+01 0.914 0.361147

LogCoarse NA NA NA NA

LogFine 1.531e+06 2.608e+05 5.870 6.04e-09 \*\*\*

WB:BW NA NA NA NA

WB:Cement 9.901e+00 4.595e+00 2.155 0.031430 \*

WB:FlyAsh 9.929e+00 4.779e+00 2.078 0.038015 \*

WB:Water -8.700e+00 3.978e+00 -2.187 0.029007 \*

WB:BlastF 9.881e+00 4.619e+00 2.139 0.032674 \*

WB:SuperP NA NA NA NA

WB:Fine -9.558e-02 4.339e-01 -0.220 0.825683

WB:Coarse -3.925e-01 3.691e-01 -1.063 0.287842

WB:Age -4.190e-02 1.079e-01 -0.388 0.697891

BW:Cement 7.576e-01 7.960e-01 0.952 0.341444

BW:FlyAsh 7.321e-01 7.622e-01 0.960 0.337076

BW:Water NA NA NA NA

BW:BlastF 7.291e-01 7.970e-01 0.915 0.360504

BW:SuperP -2.868e+00 1.345e+00 -2.132 0.033292 \*

BW:Fine -6.471e-02 1.438e-01 -0.450 0.652842

BW:Coarse -6.648e-02 1.439e-01 -0.462 0.644235

BW:Age 1.145e-01 7.246e-02 1.581 0.114265

Cement:FlyAsh -3.509e-04 5.062e-03 -0.069 0.944741

Cement:Water -1.570e-03 1.409e-02 -0.111 0.911305

Cement:BlastF -3.253e-04 5.269e-03 -0.062 0.950784

Cement:SuperP 1.204e-02 1.001e-02 1.203 0.229364

Cement:Fine -2.824e-04 9.075e-04 -0.311 0.755686

Cement:Coarse -5.615e-04 8.808e-04 -0.638 0.523954

Cement:Age -3.991e-04 4.114e-04 -0.970 0.332259

FlyAsh:Water -2.508e-03 1.363e-02 -0.184 0.854103

FlyAsh:BlastF -2.509e-04 5.141e-03 -0.049 0.961092

FlyAsh:SuperP 9.564e-03 9.591e-03 0.997 0.318953

FlyAsh:Fine 2.294e-06 8.881e-04 0.003 0.997939

FlyAsh:Coarse -7.686e-04 8.579e-04 -0.896 0.370511

FlyAsh:Age 6.795e-05 4.182e-04 0.162 0.870950

Water:BlastF -1.725e-03 1.402e-02 -0.123 0.902075

Water:SuperP -4.256e-02 2.629e-02 -1.619 0.105832

Water:Fine -4.692e-03 2.685e-03 -1.748 0.080851 .

Water:Coarse -3.922e-03 2.437e-03 -1.609 0.107861

Water:Age 1.258e-03 1.116e-03 1.127 0.259860

BlastF:SuperP 1.380e-02 9.948e-03 1.387 0.165646

BlastF:Fine -3.452e-05 9.048e-04 -0.038 0.969572

BlastF:Coarse -6.202e-04 8.751e-04 -0.709 0.478679

BlastF:Age -1.584e-04 4.172e-04 -0.380 0.704219

SuperP:Fine -1.419e-03 3.019e-03 -0.470 0.638469

SuperP:Coarse -4.303e-04 2.978e-03 -0.144 0.885154

SuperP:Age -2.265e-03 2.086e-03 -1.086 0.277878

Fine:Coarse -6.531e-04 3.007e-04 -2.172 0.030122 \*

Fine:Age 2.201e-04 1.373e-04 1.603 0.109218

Coarse:Age 1.297e-04 9.971e-05 1.300 0.193788

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.17 on 943 degrees of freedom

Multiple R-squared: 0.9122, Adjusted R-squared: 0.9042

F-statistic: 114 on 86 and 943 DF, p-value: < 2.2e-16C

Discussion: Given the success of the prior hand-model with more features, the team decided to take an automated approach and move to very complete models including all pair-wise interactions and all conceivable transformations, followed by automated evaluation in R using both the standard regression and the AIC criteria and algorithms.

Interpretation: The R-squared value of 0.9122 was further improved versus the best hand model. Of the 86 potential features, the standard regression identified 19 features of significance including 6 pair-wise interactions and many transformation that had not yet been studied. From the same starting point, the AIC algorithm (not shown) produced a model with 65 features that was far less-optimal and was discarded in favor of using BIC.

**10.** **Reg3 where we used BIC to perform feature reduction from the full model.**

Call:

lm(formula = Strength ~ WB + BW + Cement + FlyAsh + Water + BlastF +

SuperP + Fine + Coarse + Age + LogBW + LogAge + OneOverCement +

OneOverFlyAsh + OneOverBlastF + OneOverCoarse + OneOverFine +

OneOverBlastF2 + OneOverFine2 + SquaredCement + SquaredFlyAsh +

SquaredFine + SquaredAge + LogCement + LogFlyAsh + LogWater +

LogSuperP + LogFine + WB:Cement + WB:FlyAsh + WB:BlastF +

BW:Cement + BW:FlyAsh + BW:BlastF + BW:Age + Cement:Fine +

Cement:Age + FlyAsh:Water + Water:Fine + Water:Coarse + BlastF:SuperP)

Residuals:

Min 1Q Median 3Q Max

-25.1368 -3.1221 -0.0441 3.2470 26.5815

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -6.715e+06 1.062e+06 -6.322 3.91e-10 \*\*\*

WB 2.703e+02 4.669e+01 5.789 9.50e-09 \*\*\*

BW -3.866e+02 6.487e+01 -5.960 3.51e-09 \*\*\*

Cement -2.912e+00 4.163e-01 -6.994 4.92e-12 \*\*\*

FlyAsh -5.841e-02 3.425e-01 -0.171 0.86462

Water -1.845e+00 6.898e-01 -2.675 0.00761 \*\*

BlastF -1.809e+00 2.059e-01 -8.787 < 2e-16 \*\*\*

SuperP -4.837e-01 2.559e-01 -1.890 0.05901 .

Fine -8.631e+02 1.393e+02 -6.198 8.40e-10 \*\*\*

Coarse 3.913e-01 5.216e-02 7.500 1.41e-13 \*\*\*

Age -1.608e-01 2.191e-02 -7.341 4.42e-13 \*\*\*

LogBW 1.783e+03 2.443e+02 7.300 5.91e-13 \*\*\*

LogAge 1.179e+01 4.155e-01 28.387 < 2e-16 \*\*\*

OneOverCement 2.122e+04 7.377e+03 2.876 0.00412 \*\*

OneOverFlyAsh -2.046e+02 2.538e+01 -8.062 2.15e-15 \*\*\*

OneOverBlastF -1.278e+02 2.921e+01 -4.373 1.36e-05 \*\*\*

OneOverCoarse -6.923e+04 2.580e+04 -2.683 0.00741 \*\*

OneOverFine 5.198e+08 8.118e+07 6.403 2.35e-10 \*\*\*

OneOverBlastF2 1.264e+02 2.835e+01 4.458 9.22e-06 \*\*\*

OneOverFine2 -4.997e+10 7.676e+09 -6.509 1.20e-10 \*\*\*

SquaredCement 6.898e-04 2.193e-04 3.145 0.00171 \*\*

SquaredFlyAsh -3.717e-03 5.102e-04 -7.285 6.57e-13 \*\*\*

SquaredFine 1.380e-01 2.261e-02 6.106 1.47e-09 \*\*\*

SquaredAge 1.728e-04 3.829e-05 4.513 7.18e-06 \*\*\*

LogCement 2.944e+02 9.133e+01 3.223 0.00131 \*\*

LogFlyAsh -6.924e+01 8.710e+00 -7.949 5.10e-15 \*\*\*

LogWater 7.822e+02 7.398e+01 10.573 < 2e-16 \*\*\*

LogSuperP 6.259e+00 7.899e-01 7.924 6.19e-15 \*\*\*

LogFine 1.008e+06 1.600e+05 6.301 4.46e-10 \*\*\*

WB:Cement 5.078e+00 7.521e-01 6.752 2.49e-11 \*\*\*

WB:FlyAsh 4.852e+00 8.326e-01 5.827 7.63e-09 \*\*\*

WB:BlastF 4.937e+00 7.480e-01 6.600 6.68e-11 \*\*\*

BW:Cement 3.199e-01 4.476e-02 7.148 1.71e-12 \*\*\*

BW:FlyAsh 2.633e-01 5.081e-02 5.182 2.66e-07 \*\*\*

BW:BlastF 2.770e-01 4.137e-02 6.695 3.61e-11 \*\*\*

BW:Age 6.145e-02 7.455e-03 8.243 5.30e-16 \*\*\*

Cement:Fine -2.443e-04 3.296e-05 -7.412 2.66e-13 \*\*\*

Cement:Age -2.491e-04 3.429e-05 -7.264 7.63e-13 \*\*\*

FlyAsh:Water -9.008e-04 2.444e-04 -3.686 0.00024 \*\*\*

Water:Fine -2.935e-03 2.334e-04 -12.579 < 2e-16 \*\*\*

Water:Coarse -2.497e-03 2.468e-04 -10.117 < 2e-16 \*\*\*

BlastF:SuperP 1.912e-03 5.841e-04 3.273 0.00110 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.214 on 988 degrees of freedom

Multiple R-squared: 0.9065, Adjusted R-squared: 0.9026

F-statistic: 233.5 on 41 and 988 DF, p-value: < 2.2e-16

Discussion: The BIC algorithm was used to perform feature reduction from the RegFull model with all potential transformations and pair-wise interactions.

Interpretation: The R-squared value of 0.9065 and RMSE of 5.214 were close to the values for the RegFull hand-model (R-squared = 0.9121, RMSE = 5.170) but with far fewer terms (Full = 86, AIC = 65, BIC = 41). Of the 86 potential features, the BIC algorithm identified 41 features of significance including 13 pair-wise interactions and many transformations. The ANOVA table for Reg3 was checked (not shown) and a reduced model was built (Reg5) with the intent to eliminate features without significance.

**11.**  **Reg5 Model where we eliminated insignificant features from ANOVA table for Reg3 - OPTIMAL MODEL**

Call:

lm(formula = Strength ~ WB + BW + Cement + FlyAsh + Water + BlastF +

SuperP + Coarse + Age + LogBW + LogAge + OneOverCement +

OneOverFlyAsh + OneOverBlastF + OneOverFine + OneOverBlastF2 +

SquaredFine + LogCement + LogFlyAsh + LogSuperP + WB:Cement +

WB:FlyAsh + WB:BlastF + BW:FlyAsh + BW:BlastF + BW:Age +

Cement:Fine + Cement:Age + Water:Fine + Water:Coarse + BlastF:SuperP)

Residuals:

Min 1Q Median 3Q Max

-24.9153 -3.6587 -0.4049 3.5458 25.3866

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.882e+02 1.425e+02 -1.321 0.186744

WB 8.906e+01 3.684e+01 2.417 0.015820 \*

BW -2.647e+01 1.761e+01 -1.503 0.133145

Cement 1.498e-01 4.895e-02 3.060 0.002275 \*\*

FlyAsh 3.982e-01 1.521e-01 2.617 0.009007 \*\*

Water -1.190e+00 7.838e-01 -1.518 0.129409

BlastF -1.793e-01 9.828e-02 -1.825 0.068342 .

SuperP -1.826e-01 2.847e-01 -0.641 0.521401

Coarse 1.290e-01 3.115e-02 4.141 3.75e-05 \*\*\*

Age -4.591e-02 1.353e-02 -3.394 0.000716 \*\*\*

LogBW 1.836e+02 8.554e+01 2.146 0.032125 \*

LogAge 1.050e+01 3.212e-01 32.693 < 2e-16 \*\*\*

OneOverCement -2.978e+03 2.981e+03 -0.999 0.317989

OneOverFlyAsh -3.102e+01 1.001e+01 -3.099 0.001995 \*\*

OneOverBlastF -2.216e+01 3.261e+01 -0.680 0.496884

OneOverFine -9.773e+04 1.457e+04 -6.706 3.34e-11 \*\*\*

OneOverBlastF2 2.443e+01 3.163e+01 0.772 0.440037

SquaredFine 1.416e-05 1.071e-05 1.322 0.186436

LogCement 2.743e+00 2.550e+01 0.108 0.914338

LogFlyAsh -7.875e+00 2.736e+00 -2.878 0.004092 \*\*

LogSuperP 5.287e+00 8.856e-01 5.970 3.29e-09 \*\*\*

WB:Cement 2.283e+00 7.858e-01 2.905 0.003750 \*\*

WB:FlyAsh 1.963e+00 8.865e-01 2.214 0.027043 \*

WB:BlastF 2.661e+00 7.982e-01 3.334 0.000888 \*\*\*

BW:FlyAsh -7.459e-02 3.110e-02 -2.399 0.016635 \*

BW:BlastF 2.680e-02 2.042e-02 1.313 0.189643

BW:Age 3.356e-02 8.115e-03 4.135 3.85e-05 \*\*\*

Cement:Fine -1.536e-04 3.430e-05 -4.478 8.40e-06 \*\*\*

Cement:Age -1.842e-04 3.886e-05 -4.740 2.45e-06 \*\*\*

Water:Fine -6.613e-04 1.474e-04 -4.486 8.10e-06 \*\*\*

Water:Coarse -6.105e-04 1.636e-04 -3.731 0.000201 \*\*\*

BlastF:SuperP 2.452e-03 6.231e-04 3.936 8.88e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.01 on 998 degrees of freedom

Multiple R-squared: 0.8745, Adjusted R-squared: 0.8706

F-statistic: 224.3 on 31 and 998 DF, p-value: < 2.2e-16

Discussion: After manually removing terms without significance from the Reg3 ANOVA table, the Reg5 model with 31 features was evaluated via standard linear regression in R.

Interpretation: The R-squared value of 0.8745 and RMSE of 6.01 for the Reg5 model were similar to the Reg3 or marginally worse (R-squared = 0.9065, RMSE = 5.124) but again the Reg5 model afforded the use of fewer terms (Reg3 BIC = 41, Reg5 = 31). The ANOVA table for Reg5 was checked (below) and almost all the terms were significant.

**Reg5 Model: ANOVA - OPTIMAL MODEL**

Analysis of Variance Table

Response: Strength

Df Sum Sq Mean Sq F value Pr(>F)

WB 1 109854 109854 3041.7472 < 2.2e-16 \*\*\*

BW 1 7673 7673 212.4566 < 2.2e-16 \*\*\*

Cement 1 11216 11216 310.5539 < 2.2e-16 \*\*\*

FlyAsh 1 1700 1700 47.0691 1.201e-11 \*\*\*

Water 1 247 247 6.8314 0.0090919 \*\*

BlastF 1 796 796 22.0284 3.061e-06 \*\*\*

SuperP 1 117 117 3.2525 0.0716178 .

Coarse 1 142 142 3.9362 0.0475313 \*

Age 1 45900 45900 1270.9387 < 2.2e-16 \*\*\*

LogBW 1 813 813 22.5079 2.398e-06 \*\*\*

LogAge 1 58693 58693 1625.1691 < 2.2e-16 \*\*\*

OneOverCement 1 1052 1052 29.1169 8.512e-08 \*\*\*

OneOverFlyAsh 1 1358 1358 37.6030 1.248e-09 \*\*\*

OneOverBlastF 1 135 135 3.7485 0.0531385 .

OneOverFine 1 1743 1743 48.2711 6.689e-12 \*\*\*

OneOverBlastF2 1 695 695 19.2504 1.268e-05 \*\*\*

SquaredFine 1 236 236 6.5462 0.0106575 \*

LogCement 1 64 64 1.7801 0.1824470

LogFlyAsh 1 91 91 2.5127 0.1132497

LogSuperP 1 1634 1634 45.2515 2.914e-11 \*\*\*

WB:Cement 1 442 442 12.2392 0.0004886 \*\*\*

WB:FlyAsh 1 638 638 17.6728 2.859e-05 \*\*\*

WB:BlastF 1 798 798 22.0964 2.957e-06 \*\*\*

BW:FlyAsh 1 674 674 18.6492 1.728e-05 \*\*\*

BW:BlastF 1 1241 1241 34.3501 6.250e-09 \*\*\*

BW:Age 1 192 192 5.3207 0.0212780 \*

Cement:Fine 1 603 603 16.7087 4.708e-05 \*\*\*

Cement:Age 1 889 889 24.6038 8.273e-07 \*\*\*

Water:Fine 1 468 468 12.9510 0.0003355 \*\*\*

Water:Coarse 1 466 466 12.9109 0.0003426 \*\*\*

BlastF:SuperP 1 559 559 15.4883 8.877e-05 \*\*\*

Residuals 998 36043 36

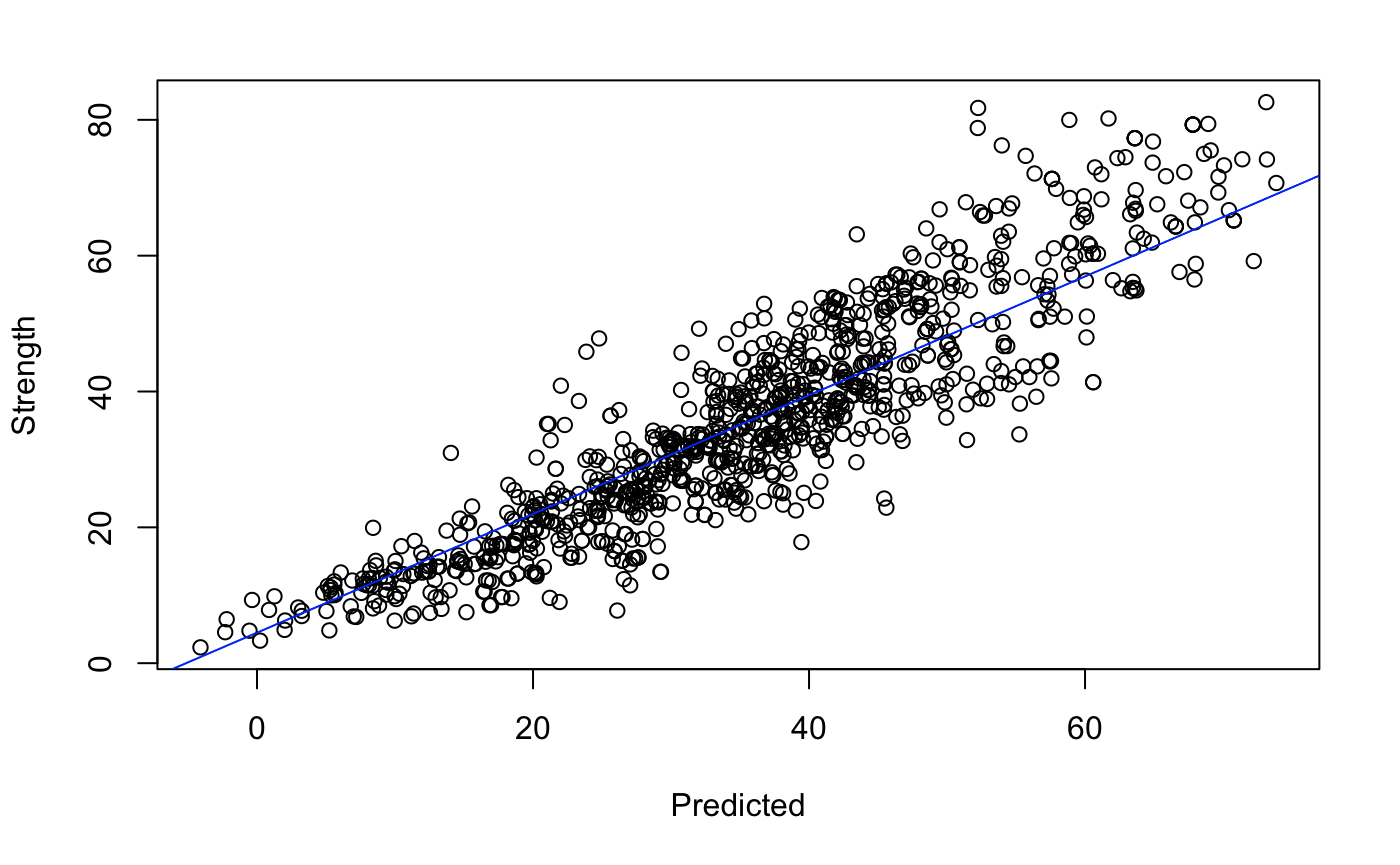
---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

**LINE Assumption checks:**

The team examined the final model (Reg5) and the initial model (Reg) against the LINE assumptions.

LINEARITY OF REGRESSION FUNCTION - Verified using the Predicted vs. Actual x-y scatter plot for model Reg5. The relationship shown is linear. The Q-Q plot below also shows a linear relationship indicating the data is normally distributed. While linear, the Predicted vs Actual plot does show more variation within the model at higher values of strength, the cause is not known. It is possible that the variance is due to the models, but it also may be from the experimental design which produced higher strength concrete not only with different recipes but also longer Age times with tend unidirectionally toward higher strengths.



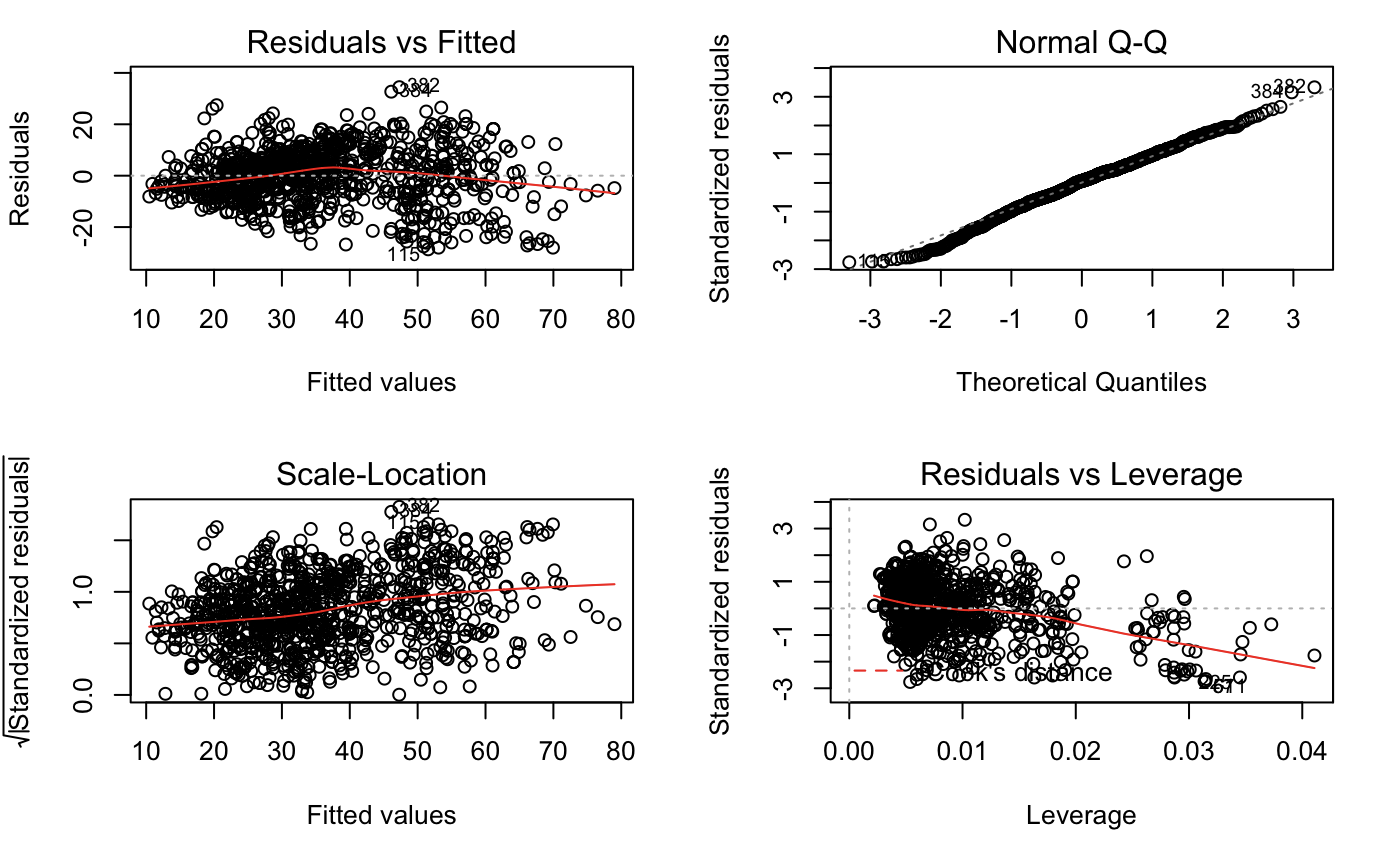
INDEPENDENT / UNCORRELATED ERROR TERMS - The data points are independent. The data set is the result of large planned experiment with different recipes for each point, producing a tangible, physical sample - a separate slug of cement which hardened over time and was tested for compressive strength.

NORMALITY OF ERROR TERMS - The ‘Residuals vs Fitted’ plot below shows that for both the original Reg model and the optimal Reg5 model, the residuals are scattered randomly.

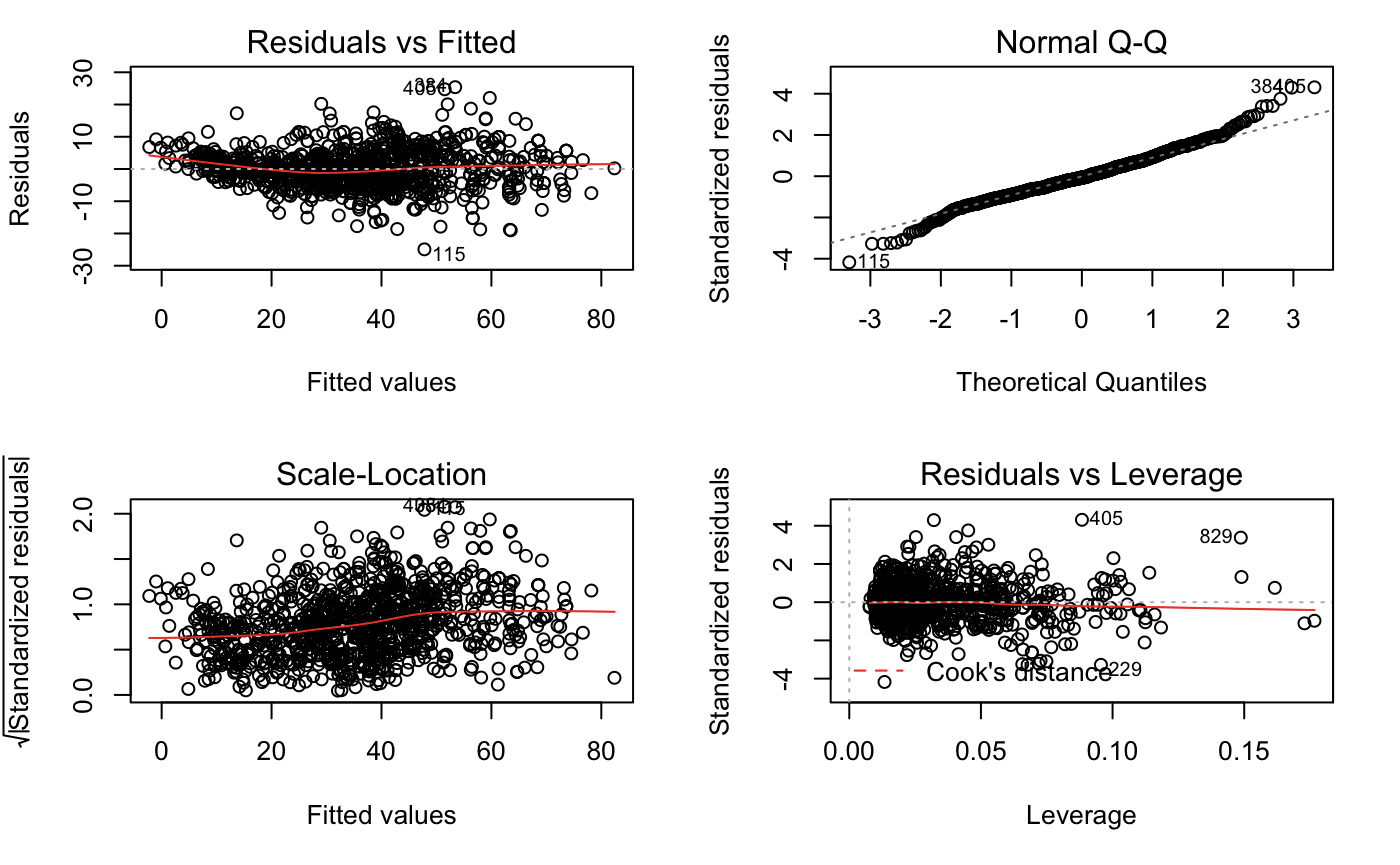
EQUAL ERROR VARIANCE - The ‘Residuals vs Fitted’ plot below shows that for both the original Reg model and the optimal Reg5 model, the residuals are, at every value of X, following a roughly normal distribution with the same variance above and below the axis across all X’s.

The models meet the LINE criteria, thus, we conclude the models are valid.

**Initial Reg Model**

****

**Reg5 Model**

****

**Model Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **No. of Predictors** | **RMSE** | **R squared** | **Adjusted R squared** |
| **Initial model and ‘Hand-models’ from trial and error** | | | | |
| **Reg** | **8** | **10.4** | **0.6155** | **0.6125** |
| **Reg1** | **6** | **10.41** | **0.614** | **0.6118** |
| **Regwb** | **4** | **10.95** | **0.5723** | 0.5706 |
| **RegA2** | **4** | **9.342** | **0.6885** | **0.6873** |
| **RegBW** | **4** | **7.843** | **0.7804** | **0.7796** |
| **RegInteract** | **12** | **7.162** | **0.8183** | **0.8162** |
| **RegLog** | **9** | **7.116** | **0.8202** | **0.8186** |
| **Reg4** | **28** | **5.894** | **0.8789** | **0.8755** |
| **Full Model and Reduced Models created using Akaike criteria** | | | | |
| **RegFull** | **86** | **5.168** | **0.9123** | **0.9043** |
| **Reg1 (AIC)** | **65** | **OVERLY COMPLEX MODEL** | | |
| **Reg3 (BIC)** | **41** | **5.218** | **0.9063** | **0.9024** |
| **Reg5 (BIC, red.)** | **31** | **6.011** | **0.8744** | **0.8705** |
| **Cross-Validation for Candidate Models** | | | | |
| **Reg3 with cross validation** | **41** | **5.38** | **0.8967** | **MAE=4.1** |
| **Reg 5 with cross validation** | **31** | **6.190** | **0.8636** | **MAE=4.7** |
| **Reg 4 with cross validation** | **28** | **6.035** | **0.8694** | **MAE=4.68** |

The data for all models is presented below in tabular form. Models from Reg to Reg 4 are ‘hand-models” created manually. RegFull and its reduced variants Reg1 AIC, Reg3 BIC, Reg5 BIC (reduced) were machine-generated within the linear regression algorithm. The transition from Reg3 BIC to Reg5 BIC (reduced) did include manual remove of insignificant terms.

Overall, the model with the most terms, RegFull, afforded the lowest RSME and highest R-squared but with the penalty of high-complexity (86 terms). The BIC algorithm was able to provide substantial feature reduction (45 terms, 52% of all terms) with acceptable erosion of R-squared and RMSE (~1%). Further removal, by hand, of 10 more terms (to 31 terms) afforded only an additional 4% erosion of R-squared.

In order to assure the large number of added features were not due to over-fitting with 100% of the data in the model, the three best models (Reg3, Reg4, and Reg5) were compared again using 10-fold cross validation. This analysis uses 90% of the data to train and 10% to test so over fitting should become apparent. No over-fitting was observed. R-squared values were high, RMSE values were low, and trends were similar between models.

**Conclusions:**

1. Since the P-value << 0.05, we reject the null hypothesis; we have a model with significant terms, the model is linear, and the model is valid.
2. Our optimal model is Reg5 BIC (reduced) with 31 predictors (see below). This model is a trade-off of slightly lower R-square for a substantially lower number of features.
3. It was critical to include both calculated and transformed variables plus interactions.
4. While the best hand-built model (Reg4) was very close to the optimal model (Reg5) in most measures, it was not selected as the optimal model. Reg4 was built manually with a trial and error approach and was biased toward certain features chosen to be included by the team. The models derived from the full set of all potential features (RegFull) like the optimal model Reg5, were machine-generated and do not contain this bias.
5. Compared to the journal article, the project team, greatly exceeded the R-squared value of the author’ linear models (0.76-0.79) and equalled the performance of the artificial neural networks (0.90-0.92). Interestingly, the largest terms in our regression by the ANOVA sum of squares are the water-to-binder ratio and the log(Age) which are the only two terms in the authors regression. However, our model adds several additional terms with large sum of squares (Cement, Age, 1/WaterBind) in the ANOVA analysis (below) and many smaller - but significant - terms which account for the higher performance versus the author’s model.

**Final Model Reg5 with 31 predictors:**

**Strength = -1.882e+02 + 8.906e+01\*WB - 2.647e+01\*BW + 1.498e-01\*Cement + 3.982e-01\*FlyAsh - 1.190e+00\*Water - 1.793e-01\*BlastF - 1.826e-01\*SuperP + 1.290e-01\*Coarse - 4.591e-02\*Age + 1.836e+02\*LogBW + 1.050e+01\*LogAge - 2.978e+03\*OneOverCement - 3.102e+01\*OneOverFlyAsh - 2.216e+01\*OneOverBlastF - 9.773e+04\*OneOverFine + 2.443e+01 \*OneOverBlastF2 + 1.416e-05\*SquaredFine + 2.743e+00\*LogCement - 7.875e+00\*LogFlyAsh + 5.287e+00\*LogSuperP + 2.283e+00\*WB\*Cement + 1.963e+00\*WB\*FlyAsh + 2.661e+00\*WB\*BlastF - 7.459e-02\*BW\*FlyAsh + 2.680e-02\*BW\*BlastF + 3.356e-02\*BW\*Age - 1.536e-04\*Cement\*Fine - 1.842e-04\*Cement\*Age - 6.613e-04\*Water\*Fine - 6.105e-04\*Water\*Coarse + 2.452e-03\*BlastF\*SuperP**

Analysis of Variance Table

Response: Strength

Df Sum Sq Mean Sq F value Pr(>F)

WB 1 109854 109854 3041.7472 < 2.2e-16 \*\*\*

BW 1 7673 7673 212.4566 < 2.2e-16 \*\*\*

Cement 1 11216 11216 310.5539 < 2.2e-16 \*\*\*

FlyAsh 1 1700 1700 47.0691 1.201e-11 \*\*\*

Water 1 247 247 6.8314 0.0090919 \*\*

BlastF 1 796 796 22.0284 3.061e-06 \*\*\*

SuperP 1 117 117 3.2525 0.0716178 .

Coarse 1 142 142 3.9362 0.0475313 \*

Age 1 45900 45900 1270.9387 < 2.2e-16 \*\*\*

LogBW 1 813 813 22.5079 2.398e-06 \*\*\*

LogAge 1 58693 58693 1625.1691 < 2.2e-16 \*\*\*

OneOverCement 1 1052 1052 29.1169 8.512e-08 \*\*\*

OneOverFlyAsh 1 1358 1358 37.6030 1.248e-09 \*\*\*

OneOverBlastF 1 135 135 3.7485 0.0531385 .

OneOverFine 1 1743 1743 48.2711 6.689e-12 \*\*\*

OneOverBlastF2 1 695 695 19.2504 1.268e-05 \*\*\*

SquaredFine 1 236 236 6.5462 0.0106575 \*

LogCement 1 64 64 1.7801 0.1824470

LogFlyAsh 1 91 91 2.5127 0.1132497

LogSuperP 1 1634 1634 45.2515 2.914e-11 \*\*\*

WB:Cement 1 442 442 12.2392 0.0004886 \*\*\*

WB:FlyAsh 1 638 638 17.6728 2.859e-05 \*\*\*

WB:BlastF 1 798 798 22.0964 2.957e-06 \*\*\*

BW:FlyAsh 1 674 674 18.6492 1.728e-05 \*\*\*

BW:BlastF 1 1241 1241 34.3501 6.250e-09 \*\*\*

BW:Age 1 192 192 5.3207 0.0212780 \*

Cement:Fine 1 603 603 16.7087 4.708e-05 \*\*\*

Cement:Age 1 889 889 24.6038 8.273e-07 \*\*\*

Water:Fine 1 468 468 12.9510 0.0003355 \*\*\*

Water:Coarse 1 466 466 12.9109 0.0003426 \*\*\*

BlastF:SuperP 1 559 559 15.4883 8.877e-05 \*\*\*

Residuals 998 36043 36

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1