Predicting Compressive Strength of Concrete Mixtures

Katie Corcoran, Samia Rubaiat, Justin Loertscher, Jonathan Scoresby

April 17, 2020

Introduction

Used in almost every manner of construction project, concrete is vital to modern infrastructure. A concrete's ability to withstand stress can vary depending on the mixture of its components such as water, cement, slag, etc. The desired strength of concrete for any given structure could vary depending on cost, availability of materials, or required strength.

In this paper, the authors hope to provide a model that can accurately predict the strength of a particular concrete based on the given quantity of elements. This idea is not novel and previously explored models include multiple regression analysis and artificial neural networks predicting the compressive strength of concrete containing different ratios of nanoparticles and fine aggregate along with different curing lengths (Chithra & Ashmita, 2016),(Sobhani, 2010), and (Sadrmomtazi, 2013). These models are complex and need a lot of calculations to predict compressive strength.

The aim of this project is to propose a simple but effective model to predict compressive strength of concrete using concrete ingredients and curing days as variables. This could be very useful in a wide variety of construction planning to ensure stress requirements are met and to potentially reduce costs.

Data

The data comes from the study done by Yeh (1998) and contains the following variables and information:

Variable	Description
Compressive strength	Measured in Mega Pascals (MPa)
Cement	Measured in kg/m^3 of mixture
Blast Furnace Slag	Measured in kg/m^3 of mixture
Fly Ash	Measured in kg/m^3 of mixture
Water	Measured in kg/m^3 of mixture
Superplasticizer	Measured in kg/m^3 of mixture
Coarse Aggregate	Measured in kg/m^3 of mixture
Fine Aggregate	Measured in kg/m^3 of mixture
Age	Measured in days. This is from 1 - 365

Table 1: Variables

There are some clear relationships between the response variable and the explanatory variables. In Figure 1, it is fairly plain that cement has a linear effect on the compressive strength. In Figure 2, though not linear, it is clear there is a relationship between age and compressive strength.

Many of the variables are left skewed as seen in Figures 3, 4, 5.

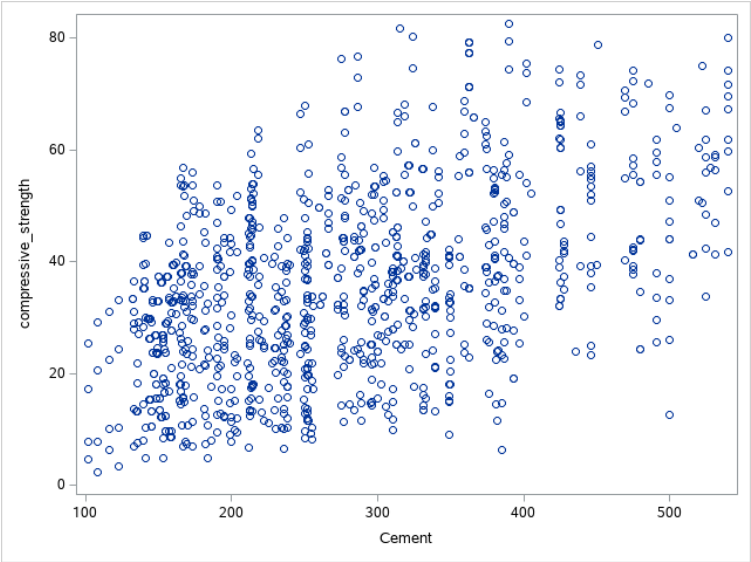


Figure 1: Cement Scatter Plot

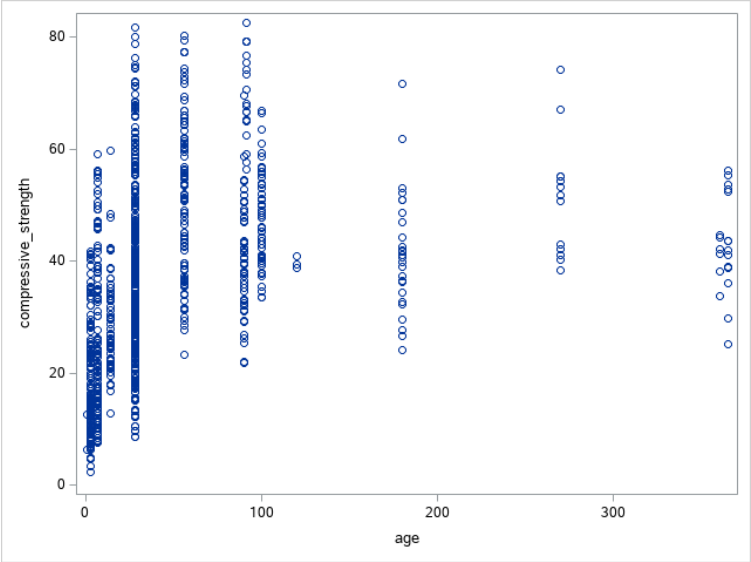


Figure 2: Age Scatter Plot

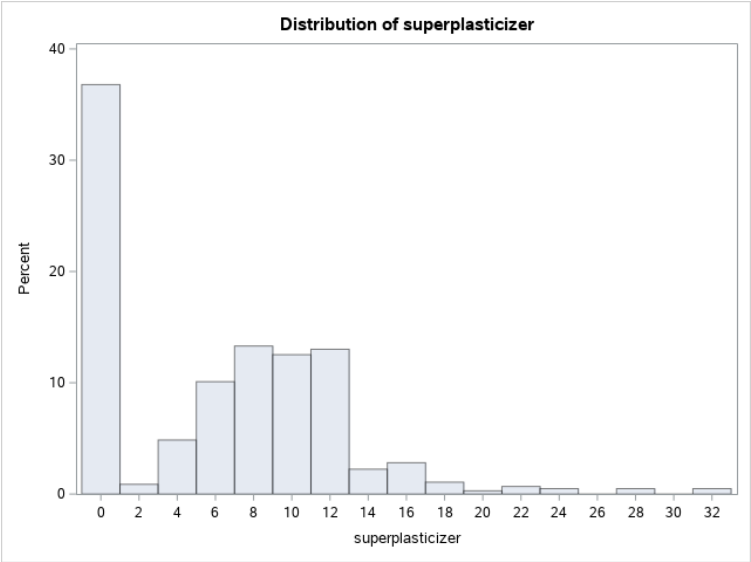


Figure 3: Superplasticizer Histogram

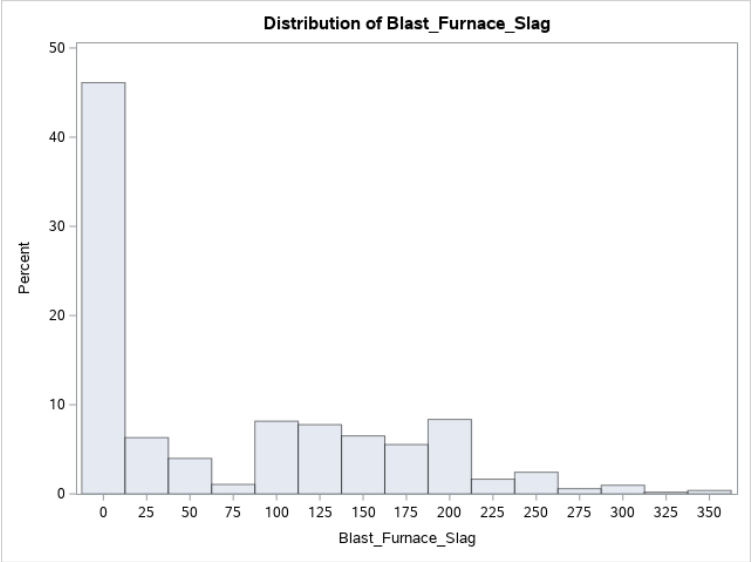


Figure 4: Blast Furnace Slag Histogram

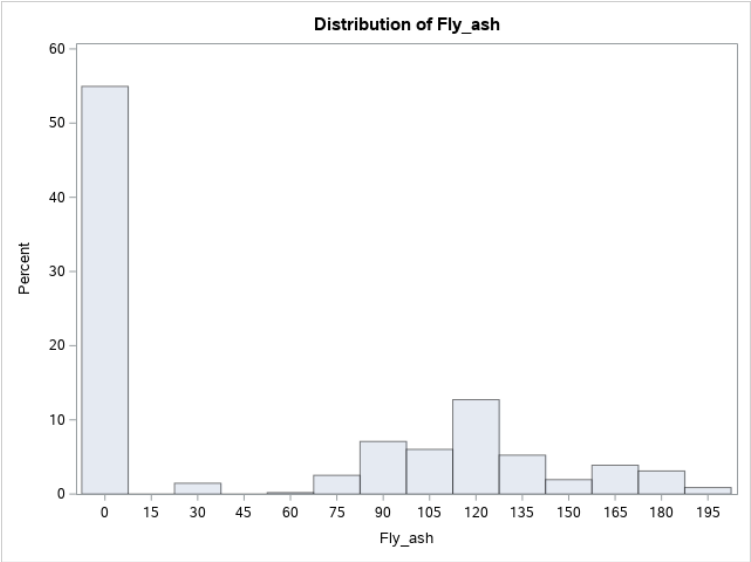


Figure 5: Fly Ash Histogram

Modeling Assumptions

The initial model presents a few problems. Firstly, the residuals are not random nor do they possess constant variance. It is clear from Figure 6 that there is a slight linear trend in the residuals and a Brown-Forsythe Test confirms non-constant variance in the data with a p-value ≈ 0.

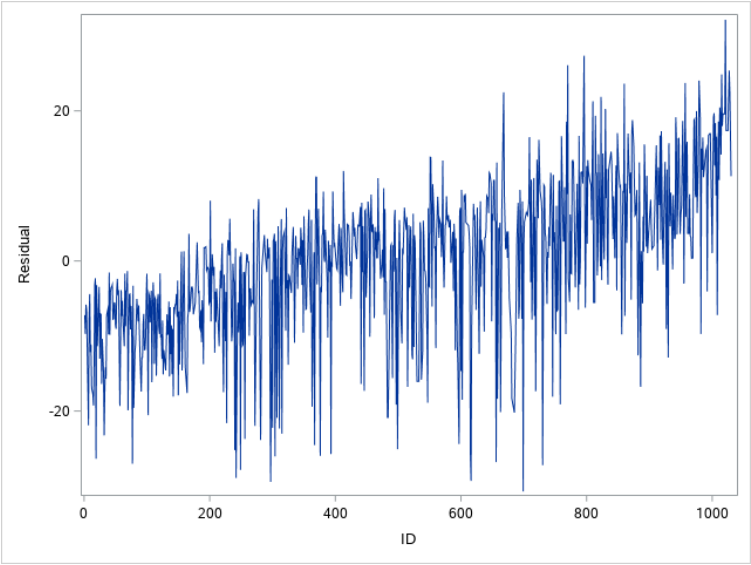


Figure 6: Residual Sequence

The residuals are, however, normally distributed as seen in Figure 7 and the error is not significantly correlated with any of the variables.

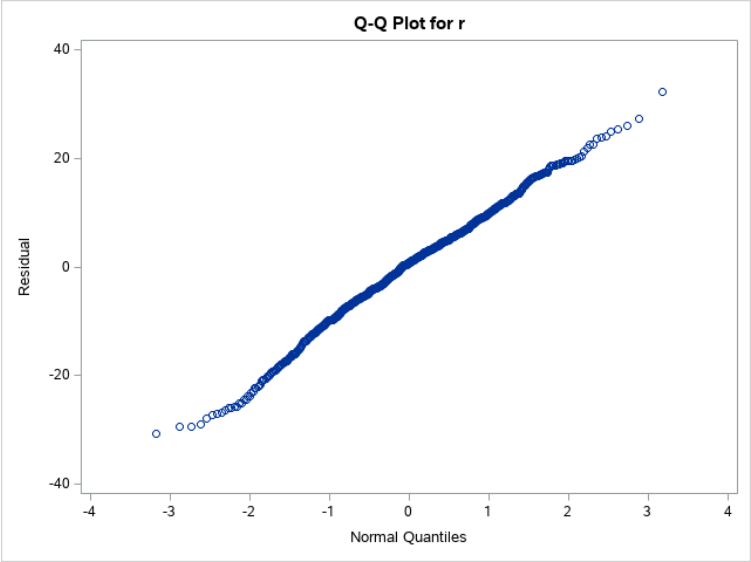


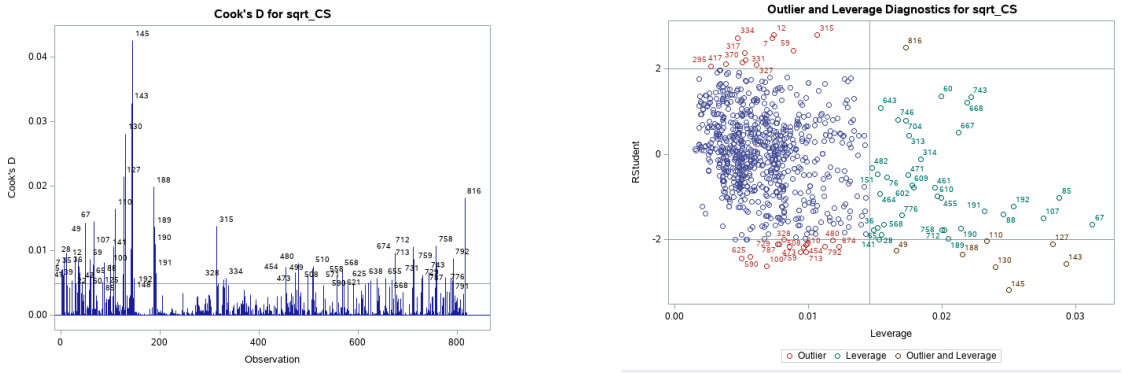
Figure 7: Residual Sequence

Finally, there is some correlation between the variables. Specifically, water and superplasticizer are very correlated. Also, fly ash and fine aggregate are somewhat correlated with almost all the other variables.

Variable Selection and Interaction Terms

To select the variables for our final model, we used a combination of forward selection, backwards elimination, and stepwise selection, all with entry and exit alpha values of .05. All methods agreed that fine aggregate and coarse aggregate should be removed from the model. Backward elimination and stepwise selection suggested that we also remove superplasticizer, however, because all methods did not agree in this case, we chose to tentatively keep it in the model. As we continued the analysis, we found that superplasticizer is significantly correlated with concrete strength and kept it in our final model. Although fly_ash was not taken out during variable selection, we found throughout our analysis that it was not as significantly correlated with the square root of compressive strength as the other predictor variables, and the mean squared error is reduced when it is removed. Because of this, we also chose to remove fly_ash from the model.

We then looked at diagnostics for our final model to identify any influential points or outliers. These diagnostics are shown in figure 8. There are many points that fall outside of the thresholds in these plots, however, no individual observations stand out in both plots. Additionally, the maximum Cook's Distance and leverage values are relatively low; because of this, we determined that we did not have any highly influential points in our data and continued with our analysis.



(a) Cook's Distance

(b) Leverage

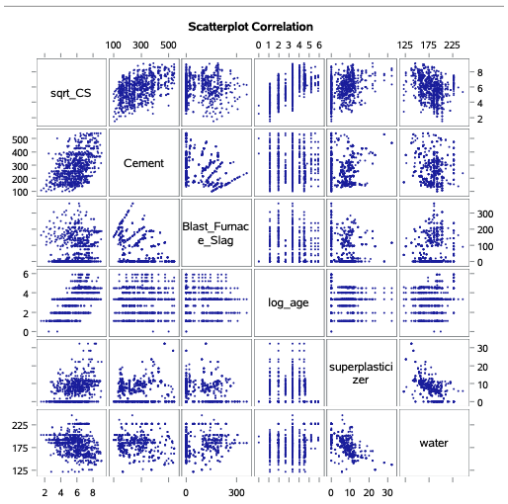
Figure 8

Next, we looked at correlations between variables to diagnose any multicollinearity problems in our model. The correlation matrix and scatterplots are shown in figure 9. There appear to be highly significant correlations between log_age and water, water and superplasticizer, and blast_furnace_slag and cement. We tested the significance of interaction terms for each of these pairings, and found that the interaction log_age and water has a significant relationship with concrete strength. No other interaction terms had p-values small enough to suggest that we add them to the model.

Because we did not include the interactions of water and superplasticizer and blast_furnace_slag and cement, we looked at the variance inflation factors (VIF's) and condition indexes with the proportions of variation to further test for multicollinearity. The VIF's averaged to be slightly over one, and there was one condition index greater than 10, however, 50% of the variance was not explained in two or more variables. Because the variance inflation factors suggested that there could be multicollinearity, we further tested the correlated terms water and superplasticizer and blast_furnace_slag and cement by looking at the significance of each term without the associated

Pearson Correlation Coefficients, N = 824 Prob > |r| under H0: Rho=0						
	sqrt_CS	Cement	Blast_Furnace_Slag	log_age	superplasticizer	water
sqrt_CS	1.00000	0.48374 <.0001	0.11154 0.0013	0.57366 <.0001	0.34797 <.0001	-0.25866 <.0001
Cement	0.48374 <.0001	1.00000	-0.27862 <.0001	-0.01813 0.6034	0.09001 0.0097	-0.07557 0.0301
Blast_Furnace_Slag	0.11154 0.0013	-0.27862 <.0001	1.00000	-0.01145 0.7428	0.04685 0.1791	0.10864 0.0018
log_age	0.57366 <.0001	-0.01813 0.6034	-0.01145 0.7428	1.00000	-0.05227 0.1339	0.18825 <.0001
superplasticizer	0.34797 <.0001	0.09001 0.0097	0.04685 0.1791	-0.05227 0.1339	1.00000	-0.67033 <.0001
water	-0.25866 <.0001	-0.07557 0.0301	0.10864 0.0018	0.18825 <.0001	-0.67033 <.0001	1.00000

(a)



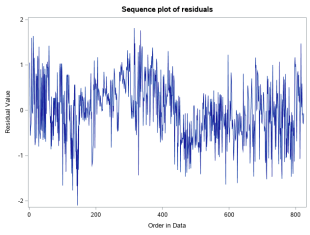
(b)

Figure 9

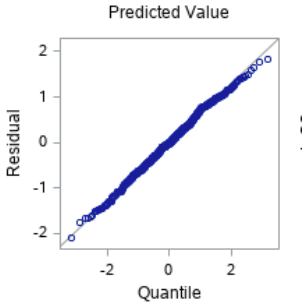
correlated term. All terms were significant individually; due to this and the proportions of variance being inconclusive, we concluded that there was not a multicollinearity problem in our model.

Lastly, we checked the model assumptions for our final model. The visual diagnostics are shown in figure 10. The normal probability plot is fairly linear, and the histogram is relatively normally distributed. Additionally, the correlation coefficient from the correlation test of normality is 0.999, which is large enough to suggest that the model has normally distributed residuals. The sequence plot does not display any highly influential points, so we can also conclude that the residuals are independent. We did not include the residual plots for all variables in the model, however, none displayed high levels of non-constant variance. To confirm this, we performed the Brown-Forsythe test of constant variance. The p-value given by the test was 0.131, which is large enough to suggest that there is not evidence of non-constant variance. We concluded from these metrics that the model assumptions are met and no further transformations are needed.

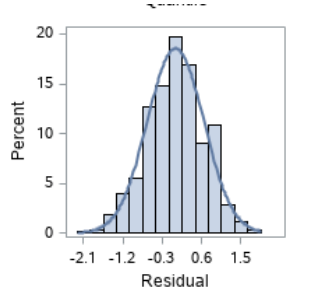
We tested the predictive power of our model by calculating the MSPR with the test portion of the data. First, we calculated the MSPR of a model including only the intercept; the value of this metric was 2.224. The mean squared error (MSE) of the original model is 0.419. We expect the error of the intercept only model to be much greater than the MSE; this holds true with our model. Next, we calculated the MSPR using our model; it was 0.352. Because the MSPR of the test set is lower than the MSE, we determined that the predictive power of our model is good.



(a) Sequence Plot



(b) Normal Probability Plot



(c) Histogram

Figure 10

The regression equation after we fit our model is:

$$\sqrt{\hat{Y}} = 3.32377 + .00497 \cdot X_1 + .00768 \cdot X_2 - .01638 \cdot X_3$$
$$+.92826 \cdot \log X_4 + .02723 \cdot X_5 - .00001856 \cdot X_3 \cdot \log X_4$$

Y	Compressive Strength
X_1	Blast Furnace Slag
X_2	Cement
X_3	Water
X_4	age
X_5	superplasticizer

Table 2: OLS model after variable selection

The beta coefficients tell us how each predictor variable will effect the square root of compressive strength of concrete. For example, $\sqrt{\hat{Y}}$ will increase by 0.00768 for each point increase in the amount of cement in the concrete, holding all other predictors constant. This will hold true for each single predictor variable term and associated coefficient. The interaction term tells us that the effect of water on the square root of compressive strength depends on the value of log(age), and vice versa. Specifically, for each point increase in the amount of water in the concrete, $\sqrt{\hat{Y}}$ will change by $0.92826 * -0.00001856 * \log x_4$. For example, if there is a high amount of water in the concrete, the effect of age on the square root of compressive strength will not be as significant. Conversely, if there is a low amount of water, a higher age could have a greater effect on the square root of compressive strength.

Alternative Models

One of the Alternatives that we used to OLS was Ridge Regression. We chose to consider ridge regression as a way to balance any potential variance in the beta coefficients of the OLS model. We applied ridge regression was applied to the data using the selected variables mentioned above. Using SAS we computed an optimal ridge parameter of .325. Not that at this point in Figure 11 the lines intersect, and the VIF is very close to 1 and the standardized coefficients are very close to 0.After using ridge regression the following model was found,

$$\sqrt{\hat{Y}} = 4.32 + .00299 \cdot X_1 + .00525 \cdot X_2 - 0.01191 \cdot X_3 + 0.54552 \cdot \log(X_4) + .03859 \cdot X_5$$

Here the variables are given by Table 2.

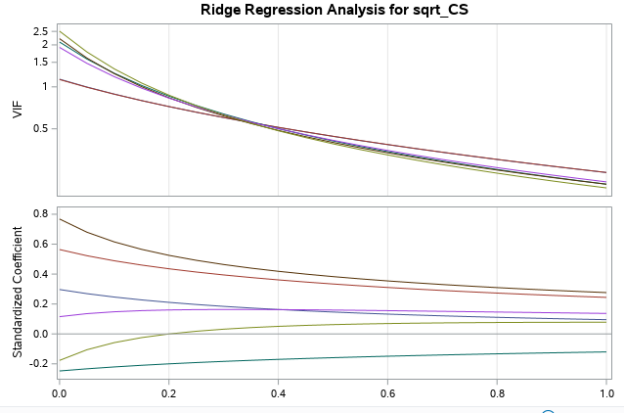


Figure 11: Ridge Regression

The disadvantage of using ridge regression is that it creates a biased model. For some situations, the ability to perform model influence would be worth this sacrifice, however there does not seem to be a huge problem with multicollinearity in our OLS model. Moreover, the predictive power of this model was not as good as the OLS model. Hence, ridge regression is not preferable for modeling this data-set.

As another alternative to ordinary least squares regression we used a regression tree to model our data. We chose to use a regression tree as it does not require many assumptions about the data. Our regression tree split the data into 94 different leaf nodes. The following (Figure: 12) is a scatter plot that shows the predictive power of our regression tree. This scatter plot suggests that the regression tree models the data really well. One possible concern with a regression tree is that it is easy to over-fit the data. This does not seem to be too much of a problem in this case. It has an MSPR of .377 which is really close to the same as the MSPR's of our other models.

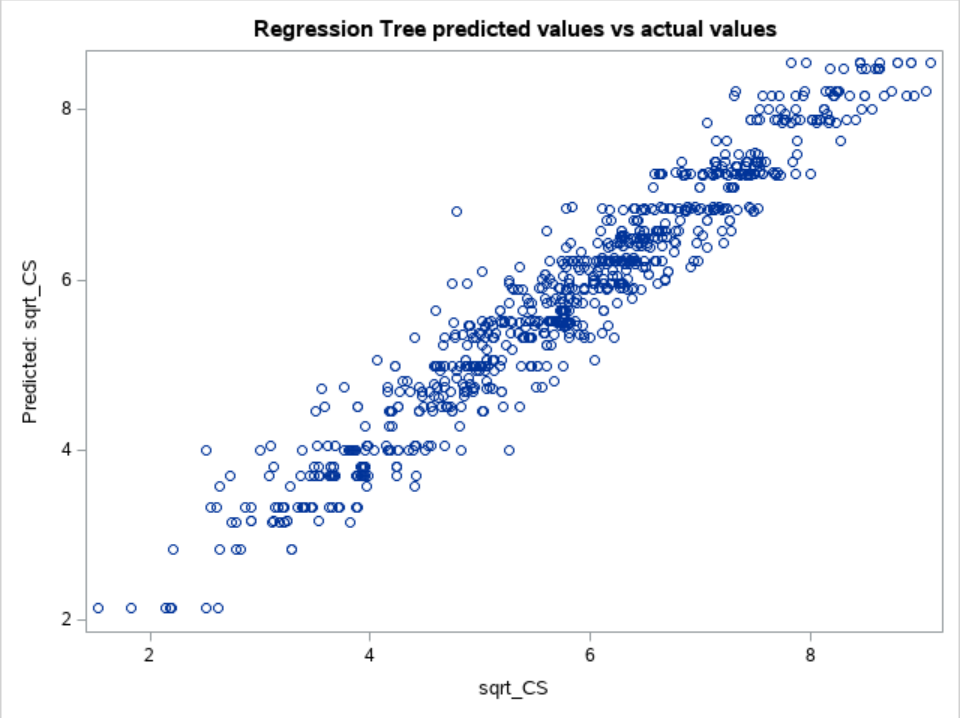


Figure 12: Regression Tree Predictions

Variable Importance			
	Training		
Variable	Relative	Importance	Count
Cement	1.0000	290.2	35
age	0.9126	264.8	28
water	0.5581	162.0	24
Blast_Furnace_Slag	0.4535	131.6	17
superplasticizer	0.3411	99.0063	21
coarse_aggregate	0.2528	73.3595	25
fine_aggregate	0.2292	66.5135	20
Fly_ash	0.1325	38.4639	9

One of our hopes with our regression tree is that there would be some consensus between the regression tree and our OLS Model. This certainly is the case. The results from our regression tree suggest that the least important variable (at least as far as creating a predictive model is concerned) is the Fly ash. This is what are variable selection techniques suggested. There is also consensus between the predictive power of the regression tree and the OLS model. This gives credence to our linear model.

Model Accuracy

We decided to use our OLS model as our final model because as demonstrated above, the remedial measures that we took are sufficient to meet the assumptions for Ordinary Least Squares regression. Moreover, this model preformed the best on test data. This can be clearly shown from Table 3.

Model	**MSPR**
Linear OLS model	0.352
Regression Tree	0.377
Ridge Regression	.5
Null Model	2.22

Table 3: MSPRs of Different Models

Our final model has an R-square value of .8221. This means that it accounts for 82% of the variation in the data. This along with the MSPR sugesst that our model is valuable for predicting the compressive strength of concrete. Moreover, a subset f-test of all the variables sugessts that our model is highly significant.

Conclusions

Our model accounts for the majority of the variance in the data. This is not too surprising, it seems logical to conclude that the strength of concrete can be accurately predicted from the ingredients and the age of the concrete. However, our data is lacking in information about the manufacturing processes used to create the concrete. We would like to see data on the way that the concrete was created to see if there are other variables besides ingredients that affect the compressive strength.

In general concrete tends to have a high compressive strength, but it does not have high tensile strength. Because of this, in almost all practical construction settings, reinforcing steel is used to give the concrete more tensile strength. It would be interesting to know the affect that the given ingredients might increase tensile strength. This could be useful because it has the potential to reduce the cost of construction by reducing the amount of steel needed to reinforce concrete.

Another area of research that we would like to peruse in the future is the long term affect that age has on concrete strength. Our data only includes concrete that is less that a year old. However concrete structures are expected to last for several decades. You only need walk down an old sidewalk to see that concrete can degrade substantially over a couple of decades. It is probable that different mixtures of concrete are less susceptible to the wear and tear of time. We would like to see how different ingredients hold up in the long run.

Currently the world is facing a shortage of sand. This is a major problem because of its use in concrete as a fine aggregate. In the next couple of decades we could be forced to change our mixtures of concrete because of scarcity of materials. For this reason research into alternative mixtures of concrete is extremely important.

References

Chithra, K. S. C. K., S., & Ashmita, F. (2016). A comparative study on the compressive strength prediction models for high performance concrete containing nano silica and copper slag using regression analysis and artificial neural networks. *Construction and Building Materials*, *114*, 528-535.

Sadrmomtazi, S. J. . M. M., A. (2013). Modeling compressive strength of eps lightweight concrete using regression, neural network and anfis. *Construction and Building Materials*, *42*, 205-216.

Sobhani, N. M. P. A. . P. T., J. (2010). Prediction of the compressive strength of no-slump concrete: A comparative study of regression, neural network and anfis models. *Construction and Building Materials*, *24*(5), 719-718.

Yeh, I.-C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. *Cement and Concrete Research*, *28*(12), 1797-1808.

Appendix

```
/* Import Data- Might have to change*/
FILENAME REFFILE '/home/u45031667/Homework/Project/Concrete_Data.csv';
PROC IMPORT DATAFILE=REFFILE replace
DBMS=CSV
OUT=WORK.concrete;
GETNAMES=YES;
RUN;

/* Histograms*/
proc univariate data=concrete;
var compressive_strength cement blast_furnace_slag
fly_ash water superplasticizer coarse_aggregate
fine_aggregate age;
histogram compressive_strength cement
blast_furnace_slag fly_ash water superplasticizer
coarse_aggregate fine_aggregate age;
run;

/* Sequence Plot of Compressive Strength*/
proc sort data=concrete; by compressive_strength;
data concrete; set concrete; ID = _n_;

/*Scatter Plots*/
proc sgplot data=concrete;
scatter x=cement y=compressive_strength;
run;

proc sgplot data=concrete;
scatter x=age y=compressive_strength;
run;


/* Split data for training and testing */
/* Sort into training and test data */
proc surveyselect data=concrete seed=12345 out=concrete2
rate=0.2 outall; /* Withold 20% for validation */
run;
data train; set concrete2;
if Selected = 0;
run;
data test; set concrete2;
if Selected = 1;
run;
```

```
/* Initial Model Fit */
proc reg data=train plots=(CooksD RStudentByLeverage DFFITS DFBETAS);
model compressive_strength = Cement Blast_Furnace_Slag Fly_ash water
superplasticizer coarse_aggregate fine_aggregate age / vif collin;
output out=posttrain residual=r predicted=p;
store regModel;
run;


/* Residual Plots */
proc sgplot data=posttrain;
series x=ID y=r;
run;

proc univariate data=posttrain;
qqplot r;
run;


/* Correlation Matrix */
proc corr data=posttrain;
var r Cement Blast_Furnace_Slag Fly_ash water
superplasticizer coarse_aggregate fine_aggregate age;
title1 'Correlation matrix';
run;

/* Brown Forsythe Test */
%resid_num_diag(dataset=posttrain, datavar=r,label='Residual', predvar=p,
predlabel='Predicted Value');

/*Ridge Regression*/
PROC IMPORT DATAFILE="/home/u45015337/Concrete-F.csv"
DBMS=CSV
OUT=WORK.Concrete_Data;
GETNAMES=YES;
RUN;
PROC CONTENTS DATA=WORK.Concrete_Data; RUN;
%web_open_table(WORK.Concrete_Data);
data con ;
set WORK.Concrete_Data;
run;

/* Separate Into Training and Test Sets.Only Fit Models to the Training Set.
The variable"Selected" separates training (0) from test (1) */
proc surveyselect data=con seed=12345 out=con1 rate=0.2 outall;
/* Withold 20% for validation */
run;
```

```
data train;
set con1;
if Selected = 0;
run;
data test; set con1;
if Selected = 1;
run;
data con2; set train;
log_CS= log(compressive_strength);
sqrt_CS=sqrt(compressive_strength);
   interaction = age*water;
   log_age=log(age);
run;

proc reg data=train;
model sqrt_CS = Cement Blast_Furnace_Slag age Coarse_aggregate fine_aggregate superplasticizer
/ selection= stepwise slentry=.01 slstay=.01;
title1 'Stepwise Selection';
run;

proc reg data=train;
model sqrt_CS = Cement Blast_Furnace_Slag age Coarse_aggregate fine_aggregate superplasticizer
/ selection=backward slstay=.01;
title1 'Backwards Elimination';
run;

proc reg data=train;
model sqrt_CS = Cement Blast_Furnace_Slag age Coarse_aggregate fine_aggregate superplasticizer
/ selection= forward slentry=.01;
title1 'Forward Selection';
run;

data train; set train;
log_age = log(age);
id = _n_;
/*if id ne 145;*/
proc reg data=train plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
model sqrt_CS = Cement Blast_Furnace_Slag water log_age superplasticizer / vif collin;
title1 'Tentative Model';
run;

proc corr data=train;
var sqrt_CS Cement Blast_Furnace_Slag log_age superplasticizer water;
title1 'Correlation matrix';
run;

proc sgscatter data=train;
matrix  sqrt_CS Cement Blast_Furnace_Slag log_age superplasticizer water/
```

```
markerattrs=(symbol=CIRCLEFILLED size=5pt);
title1 'Scatterplot Correlation';
run;

data temp; set train;
   interaction = Blast_Furnace_Slag*Cement;
proc reg data=temp;
   model sqrt_CS = Blast_Furnace_Slag Cement interaction;
   title1 'Interaction model';
run;

data temp; set train;
   interaction = age*water;
proc reg data=temp;
model sqrt_CS = Blast_Furnace_Slag cement water log_age superplasticizer interaction / vif coll
output out=trainout r=resid p=pred;
title1 'Tentative Model w/ Interaction';
run;

data temp; set trainout;
order = _n_;
proc sgplot data=temp;
series x=order y=resid / lineattrs=(pattern=solid);
xaxis label='Order in Data';
yaxis label='Residual Value';
title1 'Sequence plot of residuals';
run;

%resid_num_diag(dataset=trainout, datavar=resid, label='residual',predvar=pred, predlabel='Pred

data test; set test;
log_age = log(age);
   CS_hat = 3.32377+
0.00497*Blast_Furnace_Slag+
0.00768*Cement
-0.01638*water+
0.92826*log_age+
0.02723*superplasticizer
-0.00001856*water*age;
   SqPredError = (sqrt(compressive_strength) - CS_hat)**2;
proc means data=test mean;
   var SqPredError;
   title1 'MSPR for test set';
run;

proc reg data=train;
model sqrt_CS = ;
title1 'Intercept Model';
```

```
run;

data test; set test;
   SqPredError = (sqrt(compressive_strength) - 5.83043)**2;
proc means data=test mean;
   var SqPredError;
   title1 'MSPR for test set (null model)';
run;

/* Try ridge regression as a remedial measure */
proc reg data=con2 ridge=0 to 1 by .05
 outvif outest=ridgests
 plots(only)=ridge(VIFaxis=log);
model sqrt_CS = Blast_Furnace_Slag cement water log_age superplasticizer
interaction/ vif;
 title1 'Concrete Data Ridge Regression';

/* Now look at variable coeffs with ridge parameter 0.325 */
proc reg data=con2 outest=ridgenew outseb ridge=0.325
outvif noprint;
model sqrt_CS = Blast_Furnace_Slag cement water log_age superplasticizer
interaction;
title1 'Concrete Data Ridge Regression (c=.325)';
run;
proc print data=ridgenew;
var _type_ _rmse_ Blast_Furnace_Slag
cement water log_age superplasticizer interaction;
title1 'Ridge Estimates for Variable Coefficients,';
title2 'with ridge parameter c = 0.3';
run;

/* Get intercept term in ridge regression */
proc means data=con2 mean;
var sqrt_CS  Blast_Furnace_Slag
cement water log_age superplasticizer
interaction;
title1 'Summary Statistics';
run;
data temp;
b0 = 5.830 - .00299*72.824  -.00525*
283.84 +.01191* 181.547 -.54552*
3.157 -.03859*6.259 ;
proc print data=temp;
var b0;
title1'Ridge Intercept'
run;

/** Check validity of the model i.e. any overfitting using test data **/
```

```
data test_2 ; set test;

sqrt_CS=sqrt(compressive_strength);
log_age=log(age);
sqrt_CShat = 4.32+.00525*cement+.00299* Blast_Furnace_Slag+.54552* log_age
+.03859* superplasticizer-0.01191* water  ;
SqPredError = (sqrt_CS - sqrt_CShat )**2;
run;
proc means data=test_2 ;
var SqPredError;
title1'MSPR for test set';
run;
/*reduced model*/

proc reg data= con2;
model sqrt_CS=;

title1 'Reduced Model';

run;


data test_3; set test_2;
g=sqrt_CS -5.83043;
SqPrError =(g)**2;
run;
proc means data=test_3 ;
var SqPrError;
title1'MSPR for test set-reduced model';
run;


    /* Regression tree */
proc hpsplit data=train seed=16661 maxdepth=15 maxbranch=2;
model sqrt_CS= Cement Blast_Furnace_Slag age Coarse_aggregate
fine_aggregate superplasticizer water Fly_ash;

output out=out1;
code file='/home/u45031682/Stats5100/output_concrete_tree.sas';
run;



proc sgplot data=out1;
title 'Regression Tree predicted values vs actual values';
scatter x=sqrt_cs y=p_sqrt_cs;
run;
```

```
data scored; set test;

%include '/home/u45031682/Stats5100/output_concrete_tree.sas'; /*change
this path */
run;

data testTree; set scored;
ASE = (sqrt_cs - P_sqrt_cs)**2;
run;

proc means data=testTree;
 var ASE;
run;
```