PREDICTION OF COMPRESSIVE CONCRETE STRENGHTH

A PROJECT REPORT
Presented to the Department of Mathematics and Statistics
California State University, Long Beach

In Partial Fulfillment of the Requirements for the Degree
Master of Science in Mathematics
Option in Statistics

Faculty Reviewer:

Kagba Suaray, Ph.D.

May 2013

Table of Contents

I. Introduction

Concrete is one of the most essential elements in many cities infrastructure in terms of buildings, homes, bridges and roads. Although we may think of concrete as just a water/cement ratio, many other elements can be added to the mixture to give it strength, more specifically, compressive strength. Often when concrete is ordered on a job site, a specific compressive strength is specified. Compressive strength is measured by the maximum uniaxial loan divided by the cross sectional area of a sample concrete cylinder when tested using hydraulically operated compression machines. Understanding how much of each element should be used for a given compressive strength can greatly impact the process of concrete production and reduce the time limit to deliver concrete to job sites. The aim of this paper is determine how different elements impact compressive concrete strength, and predict compressive concrete strength given a set of input values.

II. Data

The data was collected by Prof. I-Cheng Yeh from Chung-Hua University and donated to the UCI Machine Learning Depository in 2009. The data consists of 1,030 observations and 9 attributes. Figure 1 displays the table of the attributes and a brief description of each variable. The independent variable used in this analysis will be concrete compressive strength that is measured in MPa. MPa's are defined is megapascals and describes units of pressure. We split the data into a model building set and a validation set using PROC SURVEYSELECT using 70% of the data into the model building set and 30% into the validation set.

Figure 1- Overview of Variables

Variable	Measurement	Scale	Type	Description
Cement	Continous	kg in a m3 mixture	Input	Substance used in production of concrete that binds and hardens the mixture
Blast Furnace Slag	Continous	kg in a m3 mixture	Input	Product formed when iron pellets, coke and flux are melted together in a blast furnace
Fly Ash	Continous	kg in a m3 mixture	Input	Residue generated by combustion of ground and powedered coal
Water	Continous	kg in a m3 mixture	Input	Liquid used by combining concrete mixture
Superplasticizer	Continous	kg in a m3 mixture	Input	chemicals used as admixtures where well-dispersed particle suspension are required
Coarse Aggregate	Continous	kg in a m3 mixture	Input	Crushed stone or gravel used in concrete mixture
Fine Aggregate	Continous	kg in a m3 mixture	Input	Finely crushed stone or gravel used in concrete mixture
Age	Continous	Days (1-365)	Input	Number of days concrete was allowed to cure before testing
Concrete	Continous	Mpa	Output	Compressive strength measured in Mpa (Megapascals- Units of Pressure)

Figure 2- Summary Statistics for Independent Variables

Variable	N	Mean	Std Dev	Minimum	Maximum
Cement	721	281.44	104.98	102	540
Blast Furnace Slag	721	71.82	86.75	0	359.4
Fly Ash	721	53.84	64.30	0	200.1
Water	721	182.01	21.64	121.8	247
Superplasticizer	721	6.25	6.13	0	32.2
Coarse Aggregate	721	971.60	78.00	801	1145
Fine Aggregate	721	775.14	80.93	594	992.6
Age	721	48.58	67.13	1	365

III. Methodology

In our analysis of the data, we will be using multiple linear regression to determine which subset of variables are important in determining the compressive strength of concrete. For interpretability and comparison of model parameters, we will center the input variables by subtracting the mean for each variable.

We will use several different methods to help guide us in finding the "best" model for predicting concrete compressive strength. Once have determined a "best" model, we will then ensure the model is free of multicollinearity and conduct a residual analysis to check the model for any violations of the underlying assumptions of the multiple linear regression model.

Finally, we will apply the model to the validation set to check for any bias that may be in the model. Assuming the model has little bias, we will then conclude with interpretations of parameters, confidence intervals, and conclusions that can be made from our analysis.

IV. Diagnostics of Independent variables

Figure 3 shows the correlation matrix for the variables. The last column on the right side displays the correlations between the predictor variables and the dependent variable concrete strength. Looking at the correlations, we see that blast furnace slag, superplastlclzer, water, and age have the highest correlations to concrete which helps us identify which variables may be significant in the model. Looking at the signs on the correlations, it gives us an idea of the direction of the association for each predictor variable and the dependent variable. Black furnace slag, fly ash, superplastlclzer, and age have positive associations while water, coarse aggregate, and fine aggregate have negative associations.

Figure 3 also can give us an idea of whether multicollinearity may be present in our predictor variables. Although multicollinearty does not have impact on the model as a whole, when we try to interpret individual parameters in the model, multicollinearty can lead to inaccurate or inflated confidence intervals, t statistics and predictions. In our date, highlighted in red, there are some variables that are somewhat highly correlated with each other. This may be a concern with us and something we will diagnosis as we build the model and check the validity of the model.

Figure 3- Correlation Matri

	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Strength
Cement	1.00	-0.28	-0.37	-0.08	0.10	-0.12	-0.25	0.08	0.51
Blast Furnace Slag	-0.28	1.00	-0.33	0.12	0.04	-0.27	-0.29	-0.05	0.37
Fly Ash	-0.37	-0.33	1.00	-0.28	0.38	-0.01	0.10	-0.17	-0.08
Water	-0.08	0.12	-0.28	1.00	-0.66	-0.19	-0.46	0.29	-0.30
Superplasticizer	0.10	0.04	0.38	-0.66	1.00	-0.27	0.23	-0.20	0.10
Coarse Aggregate	-0.12	-0.27	-0.01	-0.19	-0.27	1.00	-0.14	-0.01	-0.16
Fine Aggregate	-0.25	-0.29	0.10	-0.46	0.23	-0.14	1.00	-0.15	-0.17
Age	0.08	-0.05	-0.17	0.29	-0.20	-0.01	-0.15	1.00	0.32
Strength	0.51	0.37	-0.08	-0.30	0.10	-0.16	-0.17	0.32	1.00

V. Model Building

We will be utilizing three selection criteria including R^2, adjusted R^2, AIC, and BIC to help determine the "best" model, along with automatic selection methods such as forwards selection, backwards selection and stepwise selection to help validate our choice. Ideally, we would like to have justification for interaction terms that should be added in our model, but because I am not an expert in the concrete field, we created all possible interactions with our predictor variables and will try to determine whether any of the terms should be included in the model.

Figure 5 shows the selection criteria for eight different models which only include the centered original variables. I selected the model with 5 predictor variables because the R^2 doesn't increase much by adding in more predictors, and both the AIC and BIC are both fairly minimized as desired.

Figure 5- Preliminary Regression

Number in Model	R-Square	AIC	BIC	Variables in Model
1	0.24	3855.69	3855.29	Cement
2	0.36	3733.73	3733.12	Cement, superplasticizer
3	0.49	3580.91	3580.88	Cement, superplasticizer, age
4	0.55	3480.77	3481.51	Cement, blast furnace slag, water, age
5	0.61	3388.86	3390.85	Cement, blast furnace slag, fly ash, water, age
6	0.61	3383.46	3385.59	Cement, blast furnace slag, fly ash, water, superplasticizer, age
7	0.61	3384.64	3386.80	Cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, age
8	0.61	3385.10	3387.33	Cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, age

Although our selection criteria suggests cement, blast furnace slag, fly ash, water and age should be in the model, we will use added variable plots to determine if the variables should be added linearly or in a curved way. We begin with cement as the first variable included in the model and add in each variable one at a time. Looking each added variable plot, blast furnace slag, fly ash and water should be added into the model linearly, while age should be added in a curved way.

Figure 6- Added Variable Plots for Original Parameters

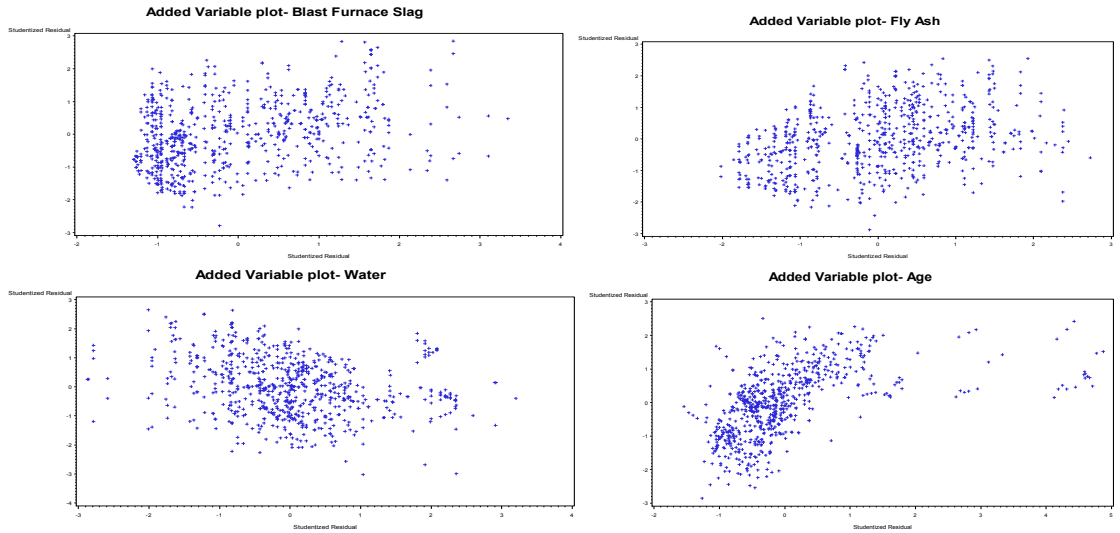


Figure 7 shows the regression output from SAS along with the VIF for each coefficient. The VIF gives us a good indication whether multicollinearity is present in the model. Looking at the ANOVA output, we have a very large F-value of 619.58 and significant p-value of <.0001 suggesting the model is significant. Next, the regression output suggests that all of the estimated parameters are significant at the 1% level. All of the parameters are positively associated with cement strength except for water, which we noted earlier when looking at the correlation matrix. The VIF's are all very low suggesting that multicollinearity may not be an issue in this model (we would be concerned if we had VIF's greater than 10).

Figure 7- Preliminary Regression Output

ANOVA Table						
Source	**DF**	**SS**	**MS**	**F-Value**	**P-Value**	**R-Square (adj)**
Model	5	161651.00	32330.00	619.58	<.0001	0.8112
Error	715	37309.00	52.18			
Corrected Total	720	198961.00				

Regression Output						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**P-Value**	**VIF**
Intercept	1	8.91772	0.77318	11.53	<.0001	0
Cement	1	0.10861	0.00327	33.24	<.0001	1.62346
Blast Furnace Slag	1	0.08525	0.00369	23.11	<.0001	1.41347
Fly Ash	1	0.0645	0.00548	11.76	<.0001	1.71504
Water	1	-0.2514	0.0134	-18.76	<.0001	1.16041
Log Age	1	8.35713	0.22573	37.02	<.0001	1.03007

The model selection criteria had an R^2 = 61% when we chose the model with 5 parameters. Once we added age in a nonlinearly, the R^2 increased to 81%, therefore 81% of the variation in strength was explained by the model. We will now introduce interaction variables for the given variables in the previous model. Using the same model selection criteria (R^2, AIC, BIC), we find that in figure 8, the model with the original 5 variables looks to be sufficient. By sufficient, it looks like adding in any interaction terms does not significantly increase the R^2 and the AIC and BIC criteria does not appear to be lowered significantly as well. It does not appear that any interaction terms are necessary in our model.

Figure8- Model Selection with Interaction Terms

Number in Model	R-Square	AIC	BIC	Variables in Model
1	0.31	3789.99	3787.80	Cement
2	0.54	3498.12	3495.47	Cement, blast furnace slag
3	0.67	3251.30	3248.90	Cement, water, *age
4	0.79	2934.87	2934.71	Cement, blast furnace slag, water, *age
5	0.82	2807.76	2809.14	Cement, blast furnace slag, fly ash, water, *age
6	0.83	2795.39	2796.94	Cemenet, blast furnace slag, fly ash, water, *age, interaction(9)
7	0.83	2786.80	2788.51	Cemenet, blast furnace slag, fly ash, water, *age, interaction(9,10)
8	0.83	2781.55	2783.41	Cemenet, blast furnace slag, fly ash, water, *age, interaction(7,9,10)
9	0.83	2776.06	2778.11	Cemenet, blast furnace slag, fly ash, water, *age, interaction(4,7,9,10)
10	0.84	2772.91	2775.12	Cemenet, blast furnace slag, fly ash, water, *age, interaction(1,4,7,9,10)
11	0.84	2773.12	2775.41	Cemenet, blast furnace slag, fly ash, water, *age, interaction(1,4,5,7,9,10)
12	0.84	2772.73	2775.14	Cemenet, blast furnace slag, fly ash, water, *age, interaction(1,2,4,5,7,9,10)
13	0.84	2771.48	2774.04	Cemenet, blast furnace slag, fly ash, water, *age, interaction(1,3,4,6,7,8,9,10)
14	0.84	2772.36	2775.01	Cemenet, blast furnace slag, fly ash, water, *age, interaction(1,3,4,5,6,7,8,9,10)
15	0.84	2773.66	2776.39	Cement, blast furnace slag, fly ash, water, *age, interaction(1-10)

*Age- Log(Age)

VI. Residual Diagnostics

In the previous section, we built our model using various methods and concluded that no interaction terms were necessary in the model. The Final model included predictors cement, blast furnace slag, fly ash, water and log of age. Now that we have our final model we need to make sure assumptions of the multiple linear regression model ae not violated. The assumptions of the model are the following:

1. $E\{ \varepsilon_i \}=0$
2. ε_i Independent of ε_j
3. $\varepsilon_i \sim N(0,\sigma 2)$
4. $Var(\varepsilon_i) = \sigma 2$

We assume that the error terms are independent since we are not dealing with time series data but the other 3 assumtpions can be shown graphically by plotting the residuals verses the fitted values and a normal probability plot. Figure 9 shows the residual plot and we can actually see a pattern in the residuals. The residuals tend to disperse more and more as predicted strength increases suggesting a non constant variance and a serious violation of the MLR model. Using the Box Cox transformation procedure in SAS, the optimal lambda is 0.5 which indicates a square root on our Y variable strength is necessary. After computing the transformation, the residual plot is much more random around 0 and ranges from -3 to 3 as desired (empirical rule). Looking at the normal probability plot, except for a few observations at the tails, the data looks to fall in a straight line giving us evidence of normality in the error terms. Once we made the transformation on our dependent variable strength, the assumptions of constant variance and normality of the error terms looked to be satisfied.

Figure 9- Studentized Residuals Vs. Fitted Values- Before Transformation

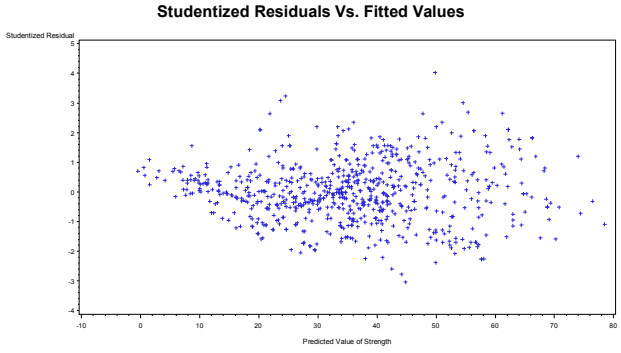


Figure 10- Studentized Residuals Vs. Fitted Values- After Transformation

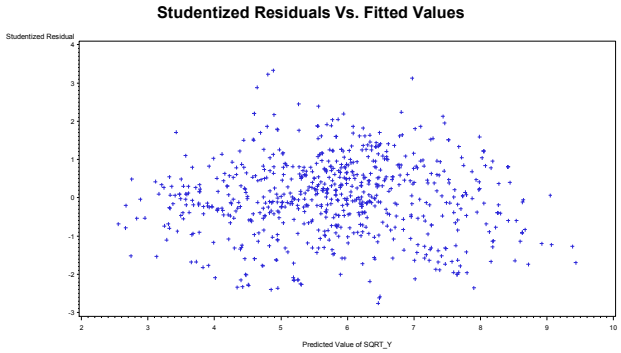
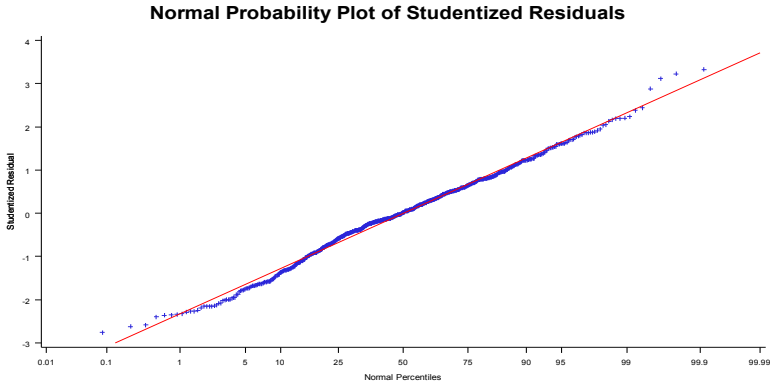


Figure 11- Normal Probability Plot



Lastly, we can actually test whether a model specific type of regression coefficient fits the data. The test is called the F-test lack of fit. Assumptions of the test are the following:

1.Observations of Y for a given X are independent
2. Observations of Y for a given X are normally distributed
3.Distribution of Y have the same variance
4. Requires repeat observations at one or more X levels

The idea of the test is that it breaks down the residual error (SSE) into two components which include lack of fit sum of squares (SSLF) and pure error sum of squares (SSPE). SSLF is due to lack of model fit while SSPE is due to pure random error. When the SSLF is large, then we may have evidence that the linear model is not appropriate. SSE can be broken down into the following:

SSE=SSLF+SSPE

$$\sum_i \sum_j (y_{ij}-\hat{y}_{ij})^2 = \sum_i \sum_j (\bar{y}_i-\hat{y}_{ij})^2 + \sum_i \sum_j (y_{ij}-\bar{y}_i)^2$$

Full Model: $\mu_j=\beta_0+\beta_{1X_j}$

Reduced Model: $Yij=\beta_0 +\beta_{1X_i}+\varepsilon_{ij}$

The null hypothesis consists of:

Ho: $\mu_i =\beta_0 +\beta_{1X_i+\ldots+}\beta_{p-1}x_{p-1}$
H1: $\mu_i \neq\beta_0 +\beta_{1X_i+\ldots+}\beta_{p-1}x_{p-1}$

F Statistic

$$\frac{SSLF/c-p}{SSPE/n-c}$$

Figure 12 shows the SAS output for the Lack of Fit Test. We obtain a p-value for the test of 0.2656 which we fail to reject the null and conclude a linear model is appropriate. Although the lack of fit is extremely high, I believe that the power of this test is very low because o the fact we have 5 predictors and very little replications. I would be cautious when using this statistics in determining a linear relationship.

Figure 12- Lack of Fit Statistic

Source	DF	Sum of Squares	Mean Square	F-Value	Pr>F
Model	5	1235.59979	247.11996	661.05	<.0001
Error	715	267.28861	0.37383		
Lack of Fit	673	253.78861	0.37710	1.17	0.2656
Pure Error	42	13.5	0.32143		
Corrected Total	720	1502.8884			

VII. Automatic Selection Methods

Much of this paper has focused on manually building a MLR model to determine important variables in predicting compressive concrete strength. Although a very exhaustive process, there does exist automatic selection methods to help determine a model. Automatic selection methods include forward selection, backwards selection and stepwise selection.

Forward selection pre-selects a significance level, alpha to entry, and continues to add in variables meeting the alpha criteria. The backwards selection method includes all variables in the initial model and drops the covariates one by one that meet the alpha to drop criteria. Lastly, the stepwise selection method uses two significance levels to add and drop variables at each step. One drawback of automatic selection methods is that even though we may think a variable, significant or not, needs to be in the model, if it does not meet the criteria to be selected, it will not be selected in the model. We will utilize these selection methods to compare to the final model we chose earlier using manual selection criteria. We will compare all three methods to the final model we selected earlier using significance level 0.15. Figure 13 summarizes the results for each

method. Although the automatic selection methods chose to have more variables in the model, notice the R^2 does not change very much. In this case I chose to have the more parsimonious model which is easier to interpret.

Figure 13- Summary of Automatic Selection Methods

Model Selection Method	Variables Included	R- Square (Adj)
Manual	Cement, Blast Furnace Slag, Fly Ash, Water, *Age	0.8237
Forward Selection	Cement, Blast Furnace Slag, Fly Ash, Water, *Age, Superplasicizer, Coarse Aggregate, Fine Aggregate, Interaction (1,2,4,5,7,8,9,10)	0.8411
Backwards Selection	Cement, Blast Furnace Slag, Fly Ash, Coarse Aggregate, Fine Aggregate, *Age, Interaction (1,2,3,4,6,7,8,9,10)	0.8394
Stepwise Selection	Cement, Blast Furnace Slag, Fly Ash, Coarse Aggregate, Fine Aggregate, *Age, Interaction (1,2,3,4,6,7,8,9,10)	0.8394

VIII. Final Model

After a very exhaustive process of building our model and checking the underlying assumptions of the model, we have finally arrived at a final model. The ANOVA suggests the model is valid with a p-value <.0001 and each parameter estimate is significant at the 1% level. The R^2(adj)= 0.8237 which is fairly high telling us that 82.3% of the variation in square root of strength is explained by the model. Below are also the 95% confidence intervals for the parameter estimates. An interpretation for the 95% confidence interval for cement would include that we are 95% confident that the true value of the parameter estimate for cement lies between 0.00916 and 0.01022. An interpretation of the parameter itself suggests that on average (because we centered the variable), a one unit increase in cement increases the square root of strength by a factor of .00969.

Figure 14- ANOVA and Parameter Estimates

ANOVA					
Source	df	Sum of Squares	Mean Square	F Value	Pr >F
Model	5	1249.28611	249.85722	673.86	<.0001
Error	715	265.11021	0.37078		
Corrected Total	720	1514.39632			

Variable	df	Parameter Estimate	Standard Error	t Value	P-Value	95% Confidence Interval	
Intercept	1	3.3355	0.06455	51.67	<.0001	3.20877	3.46223
Cement	1	0.00969	0.00026933	35.96	<.0001	0.00916	0.01022
Blast Furnace Slag	1	0.00737	0.00031602	23.31	<.0001	0.00675	0.00799
Fly Ash	1	0.00632	0.00046375	13.63	<.0001	0.00541	0.00723
Water	1	-0.01984	0.00113	-17.59	<.0001	-0.02205	-0.01762
*Age	1	0.76879	0.01899	40.49	<.0001	0.73151	0.80607

We may also want to create joint confidence intervals for our parameters. Using the Bonferroni joint confidence interval method we use the following procedure:

$b_k = Bs\{b_o\}$ where $B = t(1-\alpha/2g; n-p)$ where g are the number of parameters to be estimated jointly. Figure 15 summarizes the results. In this example, g=6, n=721 and α=.10.

Figure 15- 90% Joint Confidence Interval

Variable	90% Joint Confidence Interval	
Intercept	3.490396419	3.180603581
Cement	0.010336294	0.009043706
Blast Furnace Slag	0.008128333	0.006611667
Fly Ash	0.007432831	0.005207169
Water	-0.017128413	-0.022551587
*Age	0.814359063	0.723220937

Lastly, we would like to create both a 95% confidence interval for the mean response and for a prediction of a given subset of variables. Below are the two formulas used to give the 95% confidence intervals for the mean response and prediction. As an example we will estimate both the mean response and prediction intervals for cement=258.56, blast furnace slag = -71.82, fly ash = -53.84, water = -20.01 and *age = 3.3322.

95% CI for mean given X_h's
$Y(hat) = t(n-p, 1-\alpha/2)s\{Y_h\}$

95% CI for predicition given X_h's
$Y(hat) = t(n-p, 1-\alpha/2)s\{pred\}$

Figure 16- Confidence Intervals for Mean Response and New Prediction

	95% Confidence Interval	
Mean Response	7.8062	8.0524
Prediction	6.7275	9.1311

IX. Validation of Model

Now that we have a final model, we need to know the predictive power of the model. Earlier, we stated that the data was split into a model building set and a validation set. We will employ two methods of validation of our model. First we transformed the data in the same exact way as we built the original model and ran the same regression model. Figure 17 displays the parameter estimates and standard errors for the model building model and the model using the validation set. Notice that the parameter estimates are fairly similar to each other and all of the signs on the parameters are the same. Secondly, the standard errors are also fairly similar, thus giving us evidence the model has pretty high predictive power.

Figure 17- Comparison of Model Build Vs. Validation Sets

		Model Building		Validation	
Variable	df	Parameter Estimate	Standard Error	Parameter Estimate	Standard Error
Intercept	1	3.3355	0.06455	3.24362	0.10221
Cement	1	0.00969	0.00026933	0.00945	0.00043579
Blast Furnace Slag	1	0.00737	0.00031602	0.00741	0.00047754
Fly Ash	1	0.00632	0.00046375	0.00551	0.00072605
Water	1	-0.01984	0.00113	-0.02069	0.00181
*Age	1	0.76879	0.01899	0.8232	0.03085

The second method of validation is to calculate the mean squared prediction error MSPR. We want the MSPR to be fairly close to MSE in the original model MSPR is defined as the following:

$$\left(\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right)/n^*$$

where
1. Y_i is the value of the response variable in the ith validation case
2. $Y_{(hat)i}$ is the predicted value of the ith validation case based on the model building data set
3. n* is the number of cases in the validation set

For our data, we get MSPR 0.4232 to be and MSE in the final model was 0.37078 indicating that the model does not seem to be biased and has a good applicability for predicting strength in concrete.

X. Conclusion

The purpose of this analysis was to determine the important factors that can predict concrete compressive strength. Through an exhaustive analysis using various model selection criteria, transformations and analysis, we concluded that cement, blast furnace slag, fly ash, water and the log of age were important variables in predicting strength with 82% of the variance in strength explained by the model. Once we developed the model, we then checked to ensure the validity of the model by checking the assumptions of the MLR model. Lastly, we checked the predictive power of our model by use of a validation set checking to ensure parameter estimates and standard errors using the validation set were consistent with those from the model building model and then using MSPR as another gauge to determine the predictability of or model. Although much analysis was covered in this paper, many other methods such as principal components analysis, factor analysis, and other types of regressions may also be applicable to the data and may have a more powerful predictive model. A lot of research has been done for the use of engineering, and many other methods have been utilized. This paper only proposed one method, multiple linear regression, to determine a predictive model.

Appendix I-Bibliography

Kutner, Michael H. Applied linear statistical models. Boston: McGraw-Hill Irwin, 2005.

Noorzaei, J, Hakim, SJS, Jaafar, M.S. and Thanoon,W.A.M.. "Development of Artificial Neural Networks for Predicting Concrete Compressive Strength." International Journal of Engineering and Technology, Vol 4, Nov 2,2007,141-153.

Appendix II- SAS Code

```
LIBNAME MICHAEL "j:\STATS 510 PROJECT ";

PROC SURVEYSELECT DATA=MICHAEL.CEMENT OUT=MICHAEL.SPLIT SAMPRATE=.7
OUTALL;
RUN;

PROC SQL;
CREATE TABLE MICHAEL.MODEL_BUILD AS
SELECT *
FROM MICHAEL.SPLIT
WHERE SELECTED = 1
;QUIT;

PROC SQL;
CREATE TABLE MICHAEL.VALIDATION AS
SELECT *
FROM MICHAEL.SPLIT
WHERE SELECTED = 0
;QUIT;



/*SUMMARY STATISTICS FOR EACH VARIABLE*/
PROC MEANS DATA = MICHAEL.MODEL_BUILD;
VAR CEMENT BLAST_FURNACE_SLAG FLY_ASH WATER SUPERPLASTICIZER
COARSE_AGGREGATE FINE_AGGREGATE AGE;
RUN;

/*CENTERING THE DATA*/
DATA MICHAEL.MODEL_BUILD_CENTERED;
SET MICHAEL.MODEL_BUILD;
CEMENT_CTR= CEMENT-281.44;
BLAST_FURNACE_SLAG_CTR=BLAST_FURNACE_SLAG-71.82;
FLY_ASH_CTR=FLY_ASH-53.84;
WATER_CTR=WATER-182.01;
SUPERPLASTICIZER_CTR=SUPERPLASTICIZER-6.25;
COARSE_AGGREGATE_CTR=COARSE_AGGREGATE-971.60;
FINE_AGGREGATE_CTR=FINE_AGGREGATE-775.14;
AGE_CTR=AGE-45.58;
RUN;


/*CORRELATION STRUCTURE OF DATA*/
PROC CORR DATA = MICHAEL.MODEL_BUILD_CENTERED;
VAR CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
SUPERPLASTICIZER_CTR COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR AGE_CTR
STRENGTH;
RUN;


/*INITIAL PROC REG WITH MODEL SELECTION*/
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED;
```

```
MODEL STRENGTH = CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR
WATER_CTR
SUPERPLASTICIZER_CTR COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR AGE_CTR
/SELECTION=RSQUARE AIC BIC BEST=1;
RUN;

/****************************************/
/*                                      */
/*         ADDED VARIABLES PLOTS        */
/*                                      */
/****************************************/
/*ADDING IN BLAST FURNACE SLAG*/
PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL STRENGTH = CEMENT_CTR/;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
PLOT STRENGTH*CEMENT_CTR;
RUN;

PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL BLAST_FURNACE_SLAG_CTR = CEMENT_CTR;
OUTPUT OUT=BLAST_FURNACE STUDENT=BLAST_FURNACE_SLAG_RESIDS;
RUN;

DATA BLAST_FURNACE_SLAG;
MERGE RESIDS BLAST_FURNACE;
RUN;

PROC GPLOT DATA = BLAST_FURNACE_SLAG;
PLOT RESIDS*BLAST_FURNACE_SLAG_RESIDS;
RUN;
/****************************************/
/*ADDING IN FLY ASH*/
PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL STRENGTH = CEMENT_CTR BLAST_FURNACE_SLAG_CTR;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
RUN;

PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL  FLY_ASH_CTR= CEMENT_CTR BLAST_FURNACE_SLAG_CTR;
OUTPUT OUT=FLY_ASH STUDENT=FLY_ASH_RESIDS;
RUN;

DATA FLY_ASH_NEW;
MERGE RESIDS FLY_ASH;
RUN;

PROC GPLOT DATA = FLY_ASH_NEW;
PLOT RESIDS*FLY_ASH_RESIDS;
RUN;

/****************************************/
/*ADDING IN WATER*/
PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL STRENGTH = CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
```

```
RUN;

PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL  WATER_CTR= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR;
OUTPUT OUT=WATER STUDENT=WATER_RESIDS;
RUN;

DATA WATER_RESIDS;
MERGE RESIDS WATER;
RUN;

PROC GPLOT DATA = WATER_RESIDS;
PLOT RESIDS*WATER_RESIDS;
RUN;

/*****************************************/
/*ADDING IN AGE*/
PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL STRENGTH = CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR
WATER_CTR ;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
RUN;

PROC REG DATA =  MICHAEL.MODEL_BUILD_CENTERED;
MODEL  AGE_CTR= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR
WATER_CTR;
OUTPUT OUT=AGE STUDENT=AGE_RESIDS;
RUN;

DATA AGE_RESIDS;
MERGE RESIDS AGE;
RUN;

PROC GPLOT DATA = AGE_RESIDS;
PLOT RESIDS*AGE_RESIDS;
RUN;

/*ADD IN AGE AS LOG(AGE)*/
DATA MICHAEL.MODEL_BUILD_CENTERED_1;
SET MICHAEL.MODEL_BUILD_CENTERED;
LOG_AGE=LOG(AGE);
RUN;


PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_1;
MODEL STRENGTH= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
LOG_AGE/VIF;
RUN;

/*************************************/
/*                                   */
/*         Interaction Var.          */
/*                                   */
/*************************************/
```

```
/*CREATION OF INTERACTION VARIABLES*/
DATA MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
SET MICHAEL.MODEL_BUILD_CENTERED_1;
SQRT_Y=SQRT(STRENGTH);
INTERACTION1=CEMENT_CTR*BLAST_FURNACE_SLAG_CTR;
INTERACTION2=CEMENT_CTR*FLY_ASH_CTR;
INTERACTION3=CEMENT_CTR*WATER_CTR;
INTERACTION4=CEMENT_CTR*LOG_AGE;
INTERACTION5=BLAST_FURNACE_SLAG_CTR*FLY_ASH_CTR;
INTERACTION6=BLAST_FURNACE_SLAG_CTR*WATER_CTR;
INTERACTION7=BLAST_FURNACE_SLAG_CTR*LOG_AGE;
INTERACTION8=FLY_ASH_CTR*WATER_CTR;
INTERACTION9=FLY_ASH_CTR*LOG_AGE;
INTERACTION10=WATER_CTR*LOG_AGE;
RUN;


/*MODEL SELECTION WITH INERACTION TERMS */
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL STRENGTH = CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR
WATER_CTR LOG_AGE INTERACTION1 INTERACTION2 INTERACTION3
INTERACTION4 INTERACTION5 INTERACTION6 INTERACTION7 INTERACTION8
INTERACTION9 INTERACTION10 /SELECTION=RSQUARE AIC BIC BEST=1;
RUN;
/****************************************/
/*                                      */
/*         RESIDUAL DIAGNOSTICS         */
/*                                      */
/****************************************/
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL STRENGTH= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
LOG_AGE/VIF;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
RUN;

PROC GPLOT DATA = RESIDS;
PLOT RESIDS*FITTED;
RUN;

PROC TRANSREG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL BOXCOX(strength)=identity(CEMENT_CTR BLAST_FURNACE_SLAG_CTR
FLY_ASH_CTR WATER_CTR LOG_AGE);
run;

PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL SQRT_Y= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
LOG_AGE/p clm clb cli ;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
RUN;

PROC GPLOT DATA = RESIDS;
PLOT RESIDS*FITTED;
RUN;

/*normal probability plot*/
```

```
PROC UNIVARIATE DATA = RESIDS;
VAR RESIDS;
PROBPLOT / NORMAL(MU=0 SIGMA=1 COLOR=RED) NOFRAME;
RUN;

/*****************************************/
/*                                       */
/*         Automated Selection           */
/*                                       */
/*****************************************/
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL SQRT_Y=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
SUPERPLASTICIZER_CTR COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR  lOG_AGE
INTERACTION1 INTERACTION2 INTERACTION3 INTERACTION4 INTERACTION5
INTERACTION6 INTERACTION7 INTERACTION8 INTERACTION9
INTERACTION10/SELECTION=F SLE=.15;
RUN;
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL STRENGTH=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
SUPERPLASTICIZER_CTR COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR  lOG_AGE
INTERACTION1 INTERACTION2 INTERACTION3 INTERACTION4 INTERACTION5
INTERACTION6 INTERACTION7 INTERACTION8 INTERACTION9
INTERACTION10/SELECTION=B SLE=.15;
RUN;
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL STRENGTH=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
SUPERPLASTICIZER_CTR COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR  lOG_AGE
INTERACTION1 INTERACTION2 INTERACTION3 INTERACTION4 INTERACTION5
INTERACTION6 INTERACTION7 INTERACTION8 INTERACTION9
INTERACTION10/SELECTION=STEPWISE SLS=.1;
RUN;

PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL SQRT_Y=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR  lOG_AGE SUPERPLASTICIZER_CTR
INTERACTION1 INTERACTION2 INTERACTION4  INTERACTION5 INTERACTION7
INTERACTION8 INTERACTION9 INTERACTION10;
RUN;
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL SQRT_Y=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR
COARSE_AGGREGATE_CTR FINE_AGGREGATE_CTR  lOG_AGE
INTERACTION1 INTERACTION2 INTERACTION3 INTERACTION4  INTERACTION6
INTERACTION7 INTERACTION8 INTERACTION9 INTERACTION10;
RUN;


/*****************************************/
/*                                       */
/*               Final Model             */
/*                                       */
/*****************************************/
PROC REG DATA = MICHAEL.MODEL_BUILD_CENTERED_INTERATION;
MODEL SQRT_Y= CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
LOG_AGE/p clm clb cli ;
OUTPUT OUT=RESIDS STUDENT=RESIDS P=FITTED;
```

```
RUN;



/*****************************************/
/*                                       */
/*              Model Validation         */
/*                                       */
/*****************************************/
PROC MEANS DATA = MICHAEL.VALIDATION;
VAR CEMENT BLAST_FURNACE_SLAG FLY_ASH WATER AGE;
RUN;

DATA MICHAEL.VALIDATION_NEW;
SET MICHAEL.VALIDATION;
CEMENT_CTR= CEMENT-273.88;
BLAST_FURNACE_SLAG_CTR=BLAST_FURNACE_SLAG-78.87;
FLY_ASH_CTR=FLY_ASH-55.71;
WATER_CTR=WATER-180.03;
AGE_CTR=AGE-39.7;
SQRT_Y=SQRT(STRENGTH);
lOG_AGE=LOG(AGE);
RUN;

PROC REG DATA = MICHAEL.VALIDATION_NEW;
MODEL SQRT_Y=CEMENT_CTR BLAST_FURNACE_SLAG_CTR FLY_ASH_CTR WATER_CTR
LOG_AGE/p;
OUTPUT OUT=VALID P=FITTED;
RUN;

PROC SQL;
CREATE TABLE MSPR AS
SELECT *,
       3.3355+ .00969*CEMENT_CTR+
.00737*BLAST_FURNACE_SLAG_CTR+.00632*FLY_ASH_CTR-
.01984*WATER_CTR+.76879*LOG_AGE AS MODEL_DATA
FROM VALID
;QUIT;
```