Assessing Concrete Compressive Strength in Relation to Age and Composition

STAT 448: Advanced Data Analysis
Fall 2022

Benny Zhao

Instructor: Professor Darren Glosemeyer

Introduction

As one of the most widely used substances in the world, concrete is a fundamental building material that has a history dating to a time even before Ancient Egypt. As such, having a coherent understanding of the properties of concrete is critical in the construction industry, especially with regards to the composition of concrete, compressive strength, and age. In order to gain insight on these properties of concrete, the various analyses outlined below have been performed using a modified version of Professor I-Cheng Yeh's concrete compressive strength data set; while the unmodified version was obtained from the UCI Machine Learning Repository, the data set is owned and was donated by Professor Yeh, who is in Chung-Hua University's Department of Information Management.

To give a broad overview of the data set that will be used, the data set has 1,080 samples of concrete and the variables included in the data set are as follows:

- Six quantitative variables that describe ratio of a component of concrete to water, measured as kilograms in a cubic meter; the components are cement, blast furnace slag, fly ash, superplasticizer, coarse aggregate, and fine aggregate
- Age, measured in days; all concrete in the data set are within 365 days old
- Concrete compressive strength, measured in megapascals (MPa)
- A categorical variable for age group, consisting of the six age groups which were stated by the construction manager as follows:
    - Age group 1: Less than 1 week
    - Age group 2: At least 1 week to within 4 weeks
    - Age group 3: At least 4 weeks to within 8 weeks
    - Age group 4: At least 8 weeks to within 90 days
    - Age group 5: At least 90 days to within 180 days
    - Age group 6: At least 180 days

This report aims to address the topics, which have been grouped into five analyses as follows:

- Analysis 1: General descriptive overview of the concrete data set, with emphasis on concrete strength by age and concrete composition by age
- Analysis 2: Grouping the concrete samples based on concrete composition and age
- Analysis 3: Modeling the strength of concrete based on concrete composition, assuming the concrete is at least 90 days old
- Analysis 4: Modeling whether the compressive strength of a concrete sample is at least 50 MPa, assuming the concrete is between 90 and 100 days old
- Analysis 5: Determining the age group, as defined above, based on concrete composition and strength

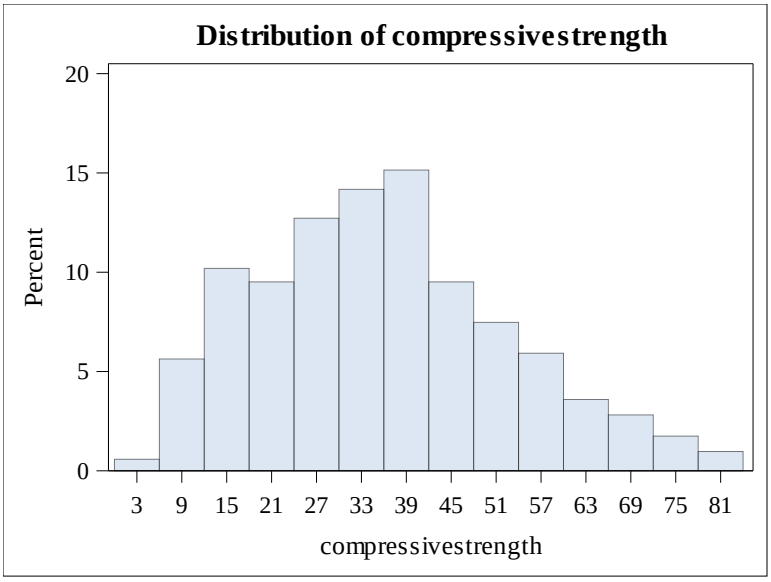
Note that any additional technical details will be listed in the Appendix section, and will be stated when referred to during the analyses.

Analysis 1: Exploratory analysis

Beginning with a general descriptive overview of the features of the concrete in the data set, an examination of the distribution of the concrete compressive strength and ages is performed in order to determine the limits of what can be predicted before extrapolation.

Variable: compressivestrength

Basic Statistical Measures			
Location		**Variability**	
Mean	35.81784	**Std Deviation**	16.70568
Median	34.44277	**Variance**	279.07972
Mode	33.39822	**Range**	80.26742
		Interquartile Range	22.50519



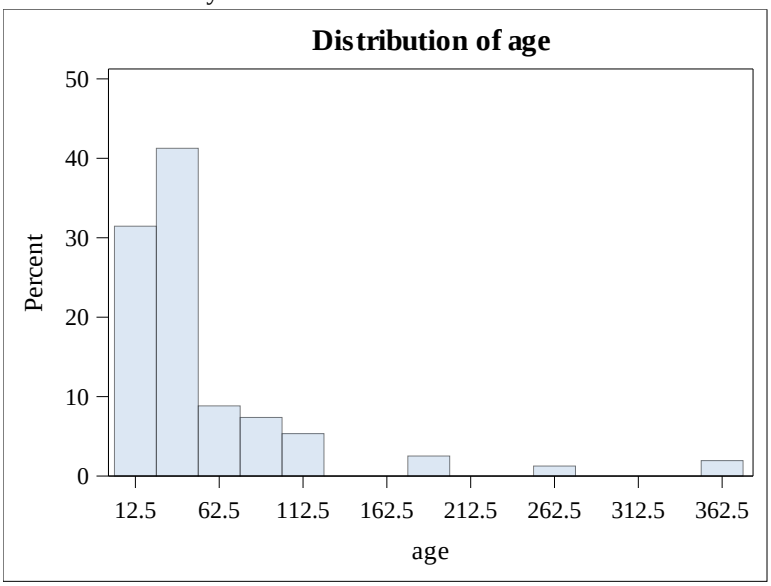
When it comes to the distribution of compressive strength, it appears to be somewhat symmetrical, while being unimodular. Hence, mean and standard deviation can be used to describe the distribution's center and spread. In this case, the average compressive strength appears to be 35.82 MPa, and on average, the compressive strength of a concrete sample is 16.7 MPa different from the mean compressive strength. The highest compressive strength in the data set is around 81 MPa, so attempting to model for compressive strengths that are substantially greater than 81 MPa is infeasible as the results are likely to be unreliable.

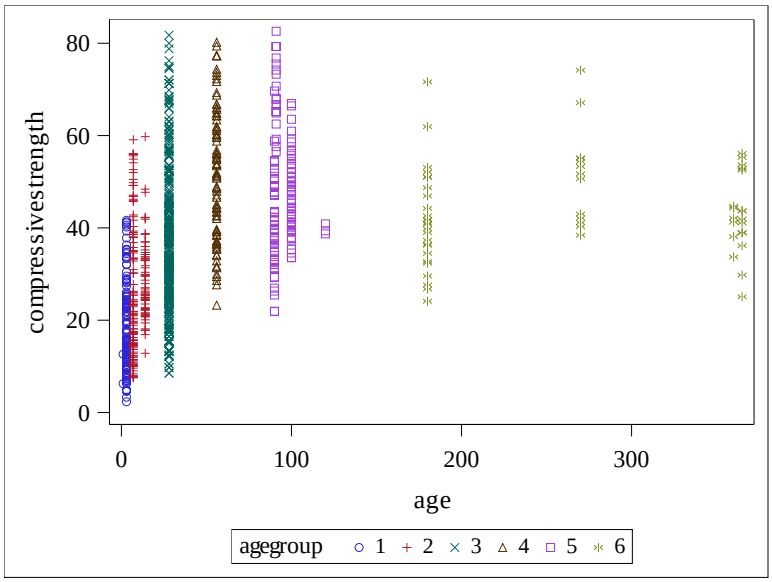
Variable: age

Basic Statistical Measures			
Location		**Variability**	
Mean	45.66214	**Std Deviation**	63.16991
Median	28.00000	**Variance**	3990
Mode	28.00000	**Range**	364.00000
		Interquartile Range	49.00000

When it comes to the distribution of age, it appears to be skewed to the right with a few extreme points that have greater ages relative to the other concrete samples. Hence, median and interquartile range should be used to describe the distribution's center and mass. In this case, the middle age of concrete in the data set is about 28 days while the middle 50%

of the data spans 28 days. As the maximum age of the concrete samples in the data set is 365 days, attempting to model beyond this age is infeasible as the results are likely to be unreliable.





Now moving on to the relationship between concrete compressive strength and age, the scatter plot does not exhibit an immediately apparent linear relationship.

agegroup=1

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
136	18.8409587	15.6128393	41.6374556	2.3318078	39.3056478

agegroup=2

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
188	26.9411856	24.3557397	59.7637797	7.5070147	52.2567650

agegroup=3

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
425	36.7484803	33.7622608	81.7511693	8.5357129	73.2154564

agegroup=4

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
91	51.8900612	51.7244895	80.1998483	23.2451909	56.9546575

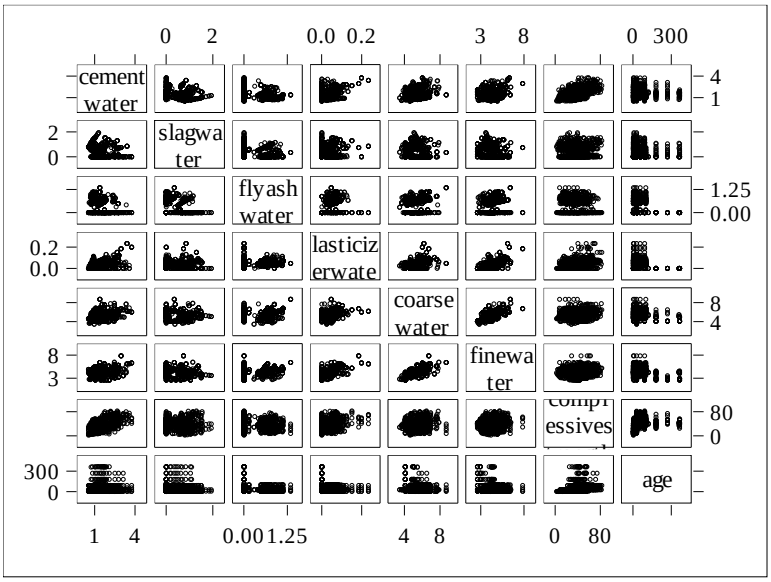
agegroup=5

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
131	48.2399568	46.2293658	82.5992248	21.8591471	60.7400777

agegroup=6

Analysis Variable : compressivestrength					
N	Mean	Median	Maximum	Minimum	Range
59	44.1614169	42.1269836	74.1669333	24.1040810	50.0628524

The mean compressive strength is the highest for the concrete in age group 4 at 51.9 MPa. Hence, compressive strength does not necessarily track with the age of the concrete, but the mean compressive strength for the younger age groups of 1, 2, and 3 are smaller than the other age groups.



As for the relationship between the composition of concrete samples and compressive strength, along with age, a pairwise scatter plot matrix was generated to get an overview.

For compressive strength, the most apparent relationship was with the cement content, which follows Abram's law, which claims the compressive strength of concrete is expected to increase as the content of cement increases. This is demonstrated in the scatterplot as a line could be reasonable fitted in order to exhibit a linear relationship. The relationship between compressive strength and the other components of concrete was inconclusive. As for age, it could appear that a line could be fitted over the scatter plots for the age versus the various concrete component, but this is due to the small spread of the content of each concrete component for the older compared to the larger spread of the content of each component for the younger samples.

In order to better understand the relationship between each concrete component and the compressive strength and age, Spearman correlations were computed. Since it is unreasonable to assume a linear relationship between the content of a particular component of concrete and the strength and age of the sample, Spearman correlations were used instead of Pearson correlations; for example, the benefit in strength provided by an increase in the content of a particular component may be diminishing.

Spearman Correlation Coefficients, N = 1030 Prob > |r| under H0: Rho=0				
	cementwater	**slagwater**	**flyashwater**	**superplasticizerwater**
cementwater	1.00000	-0.20064 <.0001	-0.28986 <.0001	0.24141 <.0001
slagwater	-0.20064 <.0001	1.00000	-0.25956 <.0001	0.13814 <.0001
flyashwater	-0.28986 <.0001	-0.25956 <.0001	1.00000	0.49200 <.0001
superplasticizerwater	0.24141 <.0001	0.13814 <.0001	0.49200 <.0001	1.00000
coarsewater	0.25225 <.0001	-0.17852 <.0001	0.29368 <.0001	0.47318 <.0001
finewater	0.15065 <.0001	-0.14236 <.0001	0.30711 <.0001	0.58176 <.0001
compressivestrength	0.52190 <.0001	0.19235 <.0001	-0.05821 0.0618	0.36174 <.0001
age	-0.02823 0.3654	-0.02438 0.4345	0.00232 0.9408	-0.01123 0.7189

Spearman Correlation Coefficients, N = 1030 Prob > |r| under H0: Rho=0				
	coarsewater	**finewater**	**compressivestrength**	**age**
cementwater	0.25225 <.0001	0.15065 <.0001	0.52190 <.0001	-0.02823 0.3654
slagwater	-0.17852 <.0001	-0.14236 <.0001	0.19235 <.0001	-0.02438 0.4345
flyashwater	0.29368 <.0001	0.30711 <.0001	-0.05821 0.0618	0.00232 0.9408
superplasticizerwater	0.47318 <.0001	0.58176 <.0001	0.36174 <.0001	-0.01123 0.7189
coarsewater	1.00000	0.61825 <.0001	0.14697 <.0001	-0.08813 0.0046
finewater	0.61825 <.0001	1.00000	0.08894 0.0043	-0.07034 0.0240
compressivestrength	0.14697 <.0001	0.08894 0.0043	1.00000	0.59602 <.0001
age	-0.08813 0.0046	-0.07034 0.0240	0.59602 <.0001	1.00000

For compressive strength, there appears to be statistically significant correlations between compressive strength and each component. Compressive strength tends to increase as the content of each component except for fly ash content increases individually, whereas compressive strength tends to decrease as fly ash content increases; while this correlation is moderately strong for cement content, the correlation is weak for the other components.

As for age, the only statistically significant correlation is between age and coarse aggregate and between age and fine aggregate. There is a weak negative correlation between age and coarse aggregate and between age and fine aggregate, meaning the age of the concrete tends to decrease as each of coarse aggregate and fine aggregate increase individually.

Analysis 2: Cluster analysis

To identify any groupings of the concrete samples using the component to water ratios and the age of the sample, cluster analysis is used. The variables were standardized due to the large disparity in the variability for the concrete component ratios compared to the age, where age would have a very large impact on the distance between points compared to the component ratios.

Variable	N	Mean	Std Dev	Minimum	Maximum	Range
cementwater	1030	1.5782748	0.6481051	0.5312500	3.7468265	3.2155765
slagwater	1030	0.4068528	0.4719605	0	1.9353796	1.9353796
flyashwater	1030	0.3134171	0.3756394	0	1.3456263	1.3456263
superplasticizerwater	1030	0.0374018	0.0391309	0	0.2336720	0.2336720
coarsewater	1030	5.4431809	0.8429658	3.4534413	8.6956879	5.2422466
finewater	1030	4.3447628	0.8249082	2.6052632	7.8404423	5.2351792
age	1030	45.6621359	63.1699116	1.0000000	365.0000000	364.0000000

The type of linkage used was average linkage so that the pseudo F and pseudo t-squared statistics could be taken as additional consideration when it came to the diagnostics for determining the number of clusters to choose. After taking into account the pseudo F, pseudo t-squared, and the cubic clustering criterion (CCC), along with the distances illustrated on the dendrogram, 6 clusters were chosen (see Appendix A).

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	53	2.5361637	0.2907642	1.9785764	2.8071334
slagwater	53	0.9162982	0.3028846	0.6149462	1.3883898
flyashwater	53	0	0	0	0
superplasticizerwater	53	0.0996181	0.0243569	0.0406897	0.1507882
coarsewater	53	6.1905898	0.7657617	5.3964535	7.8740157
finewater	53	5.4927621	0.5776158	3.9311241	6.4724409
age	53	36.4905660	32.3351575	3.0000000	91.0000000
compressivestrength	53	54.8588214	15.5996182	24.4005556	82.5992248

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	904	1.4820112	0.5634684	0.5312500	3.3333333
slagwater	904	0.3814977	0.4709107	0	1.9353796
flyashwater	904	0.3491991	0.3748964	0	1.1288843
superplasticizerwater	904	0.0330535	0.0301054	0	0.1317044
coarsewater	904	5.4041054	0.7670866	3.4534413	7.6426288
finewater	904	4.2787551	0.6647207	2.6052632	6.1568359
age	904	34.0077434	32.4824332	1.0000000	180.0000000
compressivestrength	904	33.7144857	15.8915883	2.3318078	80.1998483

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	15	3.3707984	0.3324755	2.9620853	3.7468265
slagwater	15	0.5304748	0.3910199	0	0.8505080
flyashwater	15	0	0	0	0
superplasticizerwater	15	0.2057926	0.0212504	0.1848341	0.2336720
coarsewater	15	6.3078050	0.3181459	6.0091678	6.7306477
finewater	15	6.7474669	0.8045677	6.0994194	7.8404423
age	15	37.0000000	34.0608699	3.0000000	91.0000000
compressivestrength	15	54.5784605	11.7870557	28.9993606	70.6988690

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	5	1.3794661	0	1.3794661	1.3794661
slagwater	5	0.3456263	0	0.3456263	0.3456263
flyashwater	5	1.3456263	0	1.3456263	1.3456263
superplasticizerwater	5	0.0469815	0	0.0469815	0.0469815
coarsewater	5	8.6956879	0	8.6956879	8.6956879
finewater	5	6.4074743	0	6.4074743	6.4074743
age	5	40.2000000	38.8741559	3.0000000	100.0000000
compressivestrength	5	24.3798714	12.3686988	7.7497102	39.2311844

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	48	1.5962473	0.5843612	0.7270833	3.1213873
slagwater	48	0.3319878	0.3680812	0	1.0906250
flyashwater	48	0	0	0	0
superplasticizerwater	48	0	0	0	0
coarsewater	48	4.5914738	0.7255255	4.0877193	6.5028902
finewater	48	3.2163518	0.6720916	2.6052632	4.5854922
age	48	280.8333333	78.6612271	180.0000000	365.0000000
compressivestrength	48	46.7520892	9.7889821	25.0831369	74.1669333

CLUSTER=6

Variable	N	Mean	Std Dev	Minimum	Maximum
cementwater	5	3.4778203	0.1320149	3.3219178	3.5753425
slagwater	5	0	0	0	0
flyashwater	5	0.0830986	0.1858141	0	0.4154930
superplasticizerwater	5	0.0083903	0.0126824	0	0.0285714
coarsewater	5	6.9155200	0.7827605	6.1369863	7.7323944
finewater	5	5.6720733	0.7048738	4.5140845	6.1369863
age	5	23.8000000	9.3914855	7.0000000	28.0000000
compressivestrength	5	64.4563533	10.2963743	50.5110118	74.9874098

In terms the average compressive strength across clusters, the order of the clusters descending magnitude is cluster 6, 1, 3, 5, 2, 4, with mean compressive strengths of 64.5 MPa, 54.9 MPa, 54.6 MPa, 46.8 MPa, 33.7 MPa, and 24.4 MPa, respectively. The spread of the compressive strength across clusters is roughly equal across the clusters with a minimum of 9.8 MPa for cluster 5 and a maximum of 15.9 MPa for cluster 2.

Cluster 6 had the highest mean cement ratio of 3.5 kg and the lowest mean blast furnace slag ratio of 0 kg.

When comparing clusters 1 and 3, which had similar mean compressive strengths, cluster 3 had a higher mean cement ratio of 3.4 kg whereas cluster 1 had a mean cement ratio of 2.5 kg. Cluster 3's mean blast furnace slag ratio was a little higher than half of cluster 1's mean blast furnace slag ratio, whereas the mean superplasticizer ratio for cluster 3 was twice as much as that of cluster 1's mean ratio of 100 g. Both had mean blast furnace slag ratios of 0 kg, like cluster 6.

Across all clusters, the ratio of coarse aggregate and fine aggregate had a mean value of at least 5, so they are not distinctive enough for determining which cluster a particular sample would belong to.

Using analysis of variance, along with Welch’s adjustment due to the homogeneity of the variance not holding, we see that the model for compressive strength is overall significant.

Levene's Test for Homogeneity of compressivestrength Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
CLUSTER	5	1532834	306567	2.71	0.0193
Error	1024	1.1584E8	113123		

Dependent Variable: compressivestrength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	38988.1571	7797.6314	32.17	<.0001
Error	1024	248184.8714	242.3680		
Corrected Total	1029	287173.0285			

R-Square	Coeff Var	Root MSE	compressivestrength Mean
0.135765	43.46486	15.56817	35.81784

However, because the R-squared value implies that only 13.6% of the variation in compressive strength can be explained by the cluster of a concrete sample, using the grouping of a concrete sample would be unviable for predicting the concrete sample’s compressive strength.

Welch's ANOVA for compressivestrength			
Source	**DF**	**F Value**	**Pr > F**
CLUSTER	5.0000	40.48	<.0001
Error	19.6601		

Analysis 3: Multiple linear regression

First, a subset of the data set was generated, which only included the concrete samples that had an age of at least 90 days. Then, using stepwise selection to select from the variables of the concrete components and age to model compressive strength, a multiple linear regression model was generated, which included the following variables that were considered significant: ratio of cement to water, ratio of blast furnace slag to water, ratio of fly ash to water, and the ratio of fine aggregate to water.

Model: MODEL1
Dependent Variable: compressivestrength

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	cementwater		1	0.3512	0.3512	248.588	101.76	<.0001
2	slagwater		2	0.2031	0.5543	114.524	85.23	<.0001
3	flyashwater		3	0.1506	0.7049	15.6556	94.92	<.0001
4	finewater		4	0.0155	0.7204	7.2939	10.23	0.0016

Model: MODEL1
Dependent Variable: compressivestrength

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	4	22245	5561.28909	119.15	<.0001
Error	185	8634.55161	46.67325		
Corrected Total	189	30880			

The overall model is statistically significant and we can conclude that there is evidence that indicates some relationship between compressive strength and at least one of the ratios of concrete to water, blast furnace slag to water, fly ash to water, or fine aggregate to water.

In terms of regression diagnostics, the model does not appear to violate any of the assumptions for linear regression or have collinearity issues with its variables (see Appendix C).

Root MSE	6.83178	**R-Square**	0.7204
Dependent Mean	46.97346	**Adj R-Sq**	0.7143
Coeff Var	14.54392		

As for how well the model fits the compressive strength for concrete that is at least 90 days old, the model explains around 72.04% of the variation in compressive strength, which is not spectacular, but not exactly abysmal either.

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	15.98000	2.49284	6.41	<.0001	0
cementwater	1	18.28954	0.95387	19.17	<.0001	1.30680
slagwater	1	18.45323	1.25495	14.70	<.0001	1.20637
flyashwater	1	19.40942	1.91617	10.13	<.0001	1.74084
finewater	1	-1.94854	0.60908	-3.20	0.0016	1.44135

In terms of the significance of each individual variable when the other variables are already in the model, each one is significant. The regression model would

$$y_{\text{comp. str.}} = 15.98 + 18.28954x_{\text{cement ratio}} + 18.45323x_{\text{blast f. slag ratio}} + 19.40942x_{\text{fly ash ratio}} - 1.94854x_{\text{fine agg. ratio}}$$

What the model implies is that the compressive strength of concrete that is at least 90 days old is expected to increase by 18.3 MPa for each kilogram increment of the ratio of cement to water, when the ratio of blast furnace slag to water, ratio of fly ash to water, and ratio of fine aggregate to water is kept the same; additionally, the compressive strength of concrete that is at least 90 days old is expected to increase by 18.5 MPa for each kilogram increment of the ratio of blast furnace slag to water, when the ratio of cement to water, fly ash to water, and fine aggregate to water is kept the same. The compressive strength for concrete that is at least 90 days old is expected to increase by 19.4 for each kilogram increment of the ratio of fly ash to water when the ratio of cement to water, ratio of blast furnace slag to water, ratio of fine aggregate to water is the same. Likewise, the compressive strength for concrete that is at least 90 days old is expected to decrease by 1.95 MPa for each kilogram increment of the ratio of fine aggregate to water when the ratio of cement to water, ratio of blast furnace slag to water, and the ratio of fly ash to water is kept the same.

As for the intercept, the interpretation lacks meaning as it assumes the concrete mixture has zero cement (along with zero blast furnace slag, fly ash, and fine aggregate), but the definition of concrete implies the use of cement as an ingredient.

Ultimately, this model suggests increasing the ratio of a component of concrete to water in the order of fly ash, blast furnace slag, followed by cement if the goal is to increase the compressive strength of concrete that is to last for at least 90 days. Moreover, the ratio of fine aggregate to water can also be increased if the goal is to gradually decrease the compressive strength of concrete that is to least for at least 90 days.

Analysis 4: Logistic model

After restricting the data set to the concrete samples that have an age between 90 and 100 days, a binary variable was designated for whether the compressive strength of each concrete sample was at least 50 MPa.

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	superplasticizerwate		1	1	39.4759		<.0001
2	cementwater		1	2	22.0435		<.0001
3	slagwater		1	3	26.6605		<.0001
4	finewater		1	4	4.2477		0.0393
5		finewater	1	3		3.6925	0.0547

Then using stepwise selection to select the variables considered most significant in a logistic model for predicting whether a concrete sample had a compressive strength of at least 50 MPa out of all concrete component ratio variables and age, the variables to include were the variables for following the concrete components: superplasticizer, cement, and blast furnace slag.

Due to unduly influential points identified in the Cbar plot, the concrete sample that had the highest Cbar value was removed from the data set, and the logistic model was refit. This was performed 5 times, resulting in 5 duly influential points being removed in total (see Appendix D).

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	136.7169	3	<.0001
Score	65.9427	3	<.0001
Wald	11.4586	3	0.0095

Even after removing the 5 unduly influential points, the global tests still indicate that the logistic model is significant in predicting the log odds of whether a concrete sample which had an age between 90 and 100 days had a compressive strength of at least 50 MPa.

The Hosmer-Lemeshow test indicates there is not enough evidence that suggests this logistic model does not fit the data well, and as for the diagnostics for the assumption of the variance in the residuals changing, there is some departure shown in the residual plots; model fit statistics also indicate this model is better than an intercept only model (see Appendix E).

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-32.1535	10.0071	10.3238	0.0013
superplasticizerwate	1	208.7	63.7765	10.7080	0.0011
cementwater	1	12.9832	4.0442	10.3064	0.0013
slagwater	1	15.9966	5.3557	8.9211	0.0028

In terms of the individual variables, the content of superplasticizer, cement, and blast furnace slag are considered significant in explaining the variation in whether a concrete sample which had an age between 90 and 100 days had a compressive strength of at least 50 MPa.

The log-odds model for whether a concrete sample which had an age between 90 and 100 days had a compressive strength of at least 50 MPa is

$$\log\left(\frac{\pi_{\geq 50\text{ MPa}}}{1-\pi_{\geq 50\text{ MPa}}}\right) = -32.1535 + 208.7x_{\text{superplast. ratio}} + 12.9832x_{\text{cement ratio}} + 15.9966x_{\text{bl. f. slag ratio}}$$

Therefore, the expected probability model for predicting if concrete sample which had an age between 90 and 100 days had a compressive strength of at least 50 MPa is

$$\pi_{\geq 50\text{ MPa}} = \frac{\exp\left(-32.1535 + 208.7x_{\text{superplast. ratio}} + 12.9832x_{\text{cement ratio}} + 15.9966x_{\text{bl. f. slag ratio}}\right)}{1 + \exp\left(-32.1535 + 208.7x_{\text{superplast. ratio}} + 12.9832x_{\text{cement ratio}} + 15.9966x_{\text{bl. f. slag ratio}}\right)}$$

Since the expected probability model is complex to interpret, the expected odds model should be used as an interpretation in terms of how the odds of a concrete sample having a compressive strength of at least 50 MPa changes as the content of superplasticizer, cement, and blast furnace slag changes. Since the range of the content of the superplasticizer, cement, and blast furnace slag measured in kilograms in the data set is small (e.g. the ranges are 0.23 kg, 3.22, and 1.94 kg for the content of the superplasticizer, cement, and blast furnace slag, respectively), the interpretation of the change in odds will be given as increments of 100 grams of each component, which is 0.1 of a kilogram, assuming the age of the concrete is between 90 and 100 days:

- The odds of the compressive strength being at least 50 MPa is expected to be 1×10^9 times more for every 100 gram increase in the superplasticizer ratio.
- The odds of the compressive strength being at least 50 MPa is expected to be 3.66 times more for every 100 gram increase in the cement ratio.
- The odds of the compressive strength being at least 50 MPa is expected to be 4.95 times for every 100 gram increase in the superplasticizer ratio.

Analysis 5: Discriminant Analysis

Using stepwise selection, the best predictors for determining the age group out of a concrete sample out of the concrete component ratios and the compressive strength of the sample were all of the variables:

Stepwise Selection Summary								
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	compressivestrength		0.3559	113.17	<.0001	0.64409164	<.0001
2	2	cementwater		0.2988	87.19	<.0001	0.45162892	<.0001
3	3	slagwater		0.1699	41.85	<.0001	0.37487606	<.0001
4	4	flyashwater		0.2271	59.99	<.0001	0.28974765	<.0001
5	5	finewater		0.0734	16.17	<.0001	0.26847098	<.0001
6	6	superplasticizerwater		0.0324	6.82	<.0001	0.25977622	<.0001
7	7	coarsewater		0.0200	4.16	0.0009	0.25456903	<.0001

Based on the equal covariance test, there is evidence that suggests a significant differences in covariance; hence, quadratic discriminant analysis should be used. Proceed with proportional-prior quadratic discriminant analysis since the distribution of the age groups within the data set are not equal as revealed in the exploratory analysis.

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	compressivestrength		0.07118167	<.0001
2	2	cementwater		0.11199218	<.0001
3	3	slagwater		0.12909978	<.0001
4	4	flyashwater		0.16227206	<.0001
5	5	finewater		0.17602881	<.0001
6	6	superplasticizerwater		0.18077211	<.0001
7	7	coarsewater		0.18407300	<.0001

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2855.013554	140	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Multivariate Statistics and F Approximations					
S=5 M=0.5 N=508					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.25456903	47.06	35	4284.8	<.0001
Pillai's Trace	0.92036501	32.94	35	5110	<.0001
Hotelling-Lawley Trace	2.28333577	66.33	35	2884.8	<.0001
Roy's Greatest Root	1.99063872	290.63	7	1022	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

Then, evaluating the p-values of all of the MANOVA tests, we conclude that it is possible to obtain some level of discrimination between some of the age groups using some of the predictors we identified as being viable for predicting age group.

Classification Summary for Calibration Data: WORK.CONCRETERATS
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into agegroup							
From agegroup	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1	74 54.41	16 11.76	3 2.21	0 0.00	0 0.00	43 31.62	136 100.00
2	11 5.85	55 29.26	49 26.06	4 2.13	0 0.00	69 36.70	188 100.00
3	0 0.00	33 7.76	266 62.59	32 7.53	18 4.24	76 17.88	425 100.00
4	0 0.00	2 2.20	23 25.27	32 35.16	32 35.16	2 2.20	91 100.00
5	0 0.00	0 0.00	11 8.40	36 27.48	31 23.66	53 40.46	131 100.00
6	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	59 100.00	59 100.00
Total	85 8.25	106 10.29	352 34.17	104 10.10	81 7.86	302 29.32	1030 100.00
Priors	0.13204	0.18252	0.41262	0.08835	0.12718	0.05728	

Error Count Estimates for agegroup							
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Rate	0.4559	0.7074	0.3741	0.6484	0.7634	0.0000	0.4981
Priors	0.1320	0.1825	0.4126	0.0883	0.1272	0.0573	

Using cross-validation to estimate error rates, the classification summary indicates that only concrete samples in age group 6 are reliably classified into the respective age group every time, whereas the classification of the other age groups has rather poor reliability.

In terms of age groups that are difficult to distinguish, it appears that concrete samples in age group 1 will occasionally be classified into age group 6, or more uncommonly be classified into age group 2. If a concrete sample is not correctly classified into age group 2, then it will likely be classified into age group 6 or age group 3. Age group 3 concrete

samples will be incorrectly classified into age group 6 most of the time that it is not properly classified into age group 3. When age group 4 samples are not properly classified into age group 4, there is a roughly equal likelihood that it will be incorrectly classified into age group 3 or age group 5. For age group 5 samples, they will more than likely be incorrectly classified into age group 6 than into age group 5; otherwise, age group 5 concrete samples will be incorrectly classified as age group 4.

The overall error rate is nearly 50%, which is marginally better than flipping a coin, but it is better than random assignment, which has an error rate of five-sixths. There appears to be a relatively high false classification for all age groups except for age group 4 to be misclassified as age group 6.

Due to the poor reliability of this classification model, a discriminant analysis using a training and testing partition of the data set was performed, but the results were equally as unfruitful as the one described here (see Appendix F).

Conclusion

In summary, the first analysis revealed weak correlations between compressive strength and most of the concrete components; the exception was the ratio of cement to water, which had a moderately strong positive correlation, but this is to be expected due to Abram's law. As the ratio of each component increases, compressive strength tended to increase, except for fly ash. Additionally, age was significantly negatively correlated with coarse aggregate and fine aggregate; however, the correlation was rather weak, so that the age of a concrete sample tended to decrease as the age increased.

From the cluster analysis, while it was possible to form clusters that were considered statistically significant for an ANOVA model for modelling compressive strength, less than 15% of the variation in compressive strength was explained by the clustering, meaning clustering is not recommended for determining a concrete sample's strength.

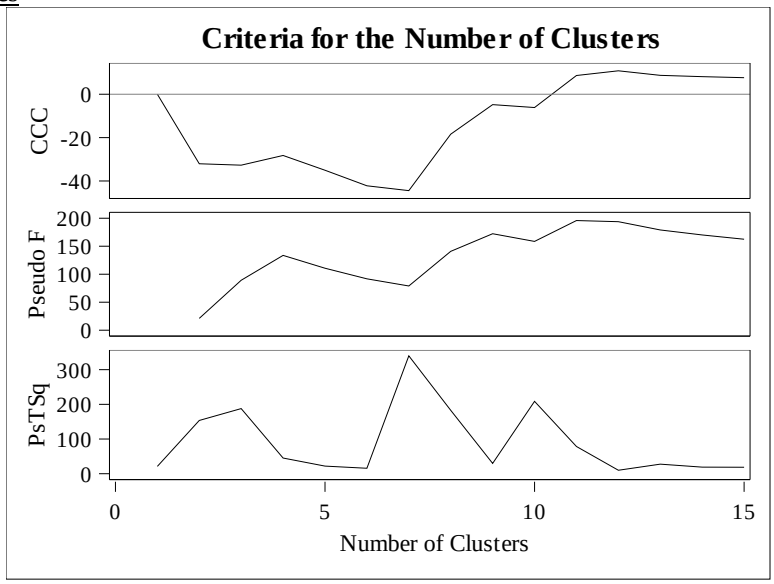
As for the multiple regression model, the model indicates increasing the ratio of cement, blast furnace slag, fly ash in a mixture to be significant in increasing compressive strength, while increasing the ratio of fine aggregate to be effective in decreasing strength, assuming the concrete sample was to least for at least 90 days old.

When it came modeling whether a concrete sample's compressive strength was at least 50 MPa, given that the age was between 90 and 100 days, the odds of the compressive strength being at least 50 is expected to have a multiplicative increase as the ratio of at least one of the superplasticizer, cement, and blast furnace slag increased.

The classification of a concrete sample's age group proved to be unfruitful for most of the age groups, other than age group 6, which was the concrete samples that were between 180 days to 365 days old. This also occurred when partitioning the data set into 80:20 training and testing sets, which would appear to suggest that the age of a concrete sample in general cannot be accurately modeled by the composition nor compressive strength.

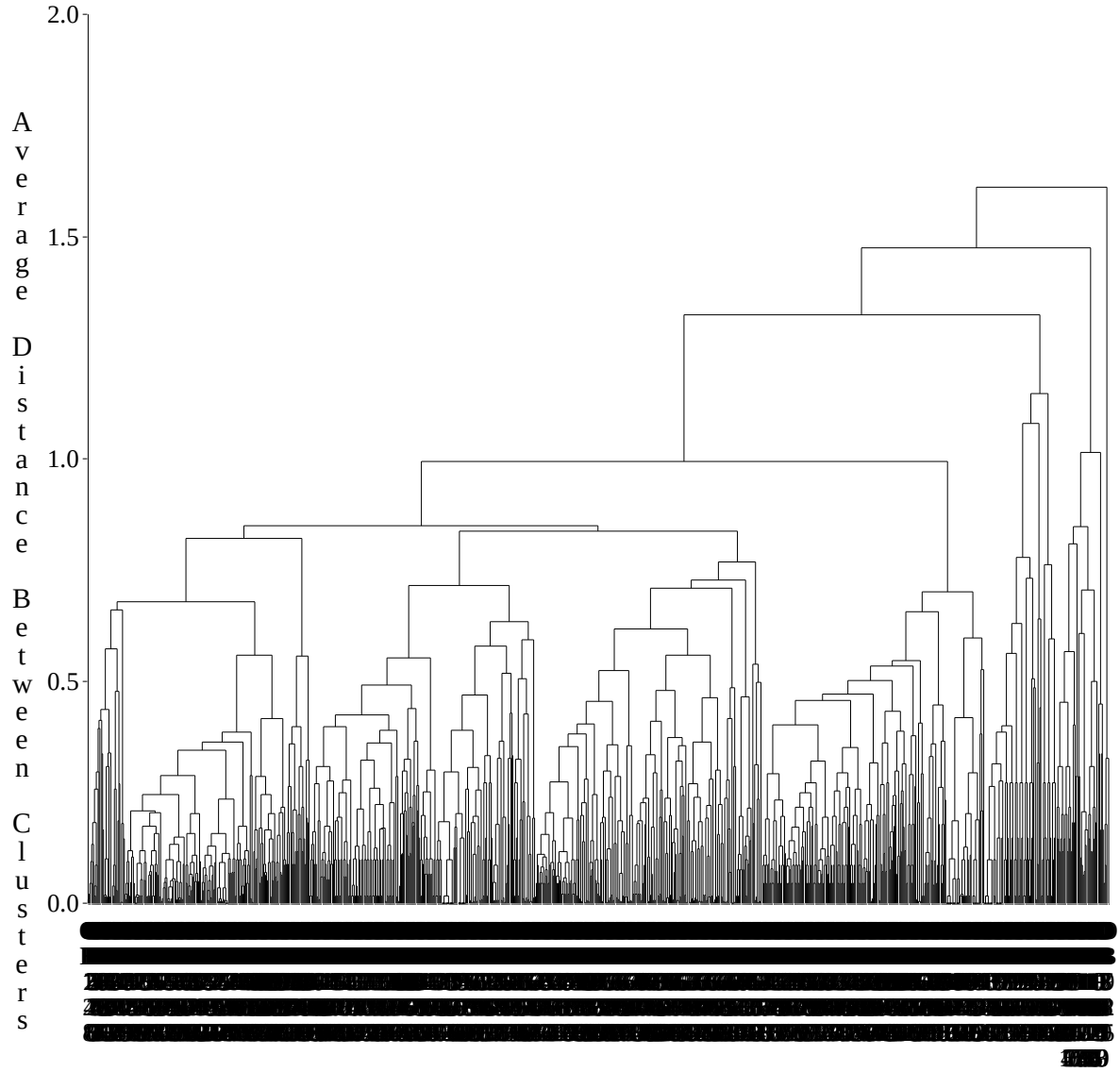
Appendix

Appendix A: Cluster diagnostics



Average Linkage Cluster Analysis

Since average linkage was used, the pseudo F and pseudo t-squared statistics are appropriate to take into consideration when choosing the number of clusters. In this case, the pseudo F plot has peaks at 4, 9, and 11 clusters. The pseudo t-squared plot has troughs at 4, 6, 9, and 12 clusters, but the most notable jumps are from 7 to 6 clusters and from 10 to 9 clusters. Lastly, the CCC plot has peaks at 4, 9, and 12 clusters.



Looking at the dendrogram, at appears that 6 clusters would be a reasonable number of clusters to be well-separated, with the first 4 clusters having an average distance below 1 and the last 2 clusters having an average distance of around 1.

I made the decision to proceed with 6 clusters due to the jump in the pseudo t-squared statistic from 7 to 6 clusters and the distance between clusters exhibited in the dendrogram, while also due to the desire to keep the grouping of concrete samples at a practical size.

Appendix B: Analysis of variance using 9 clusters

Levene's Test for Homogeneity of compressivestrength Variance ANOVA of Squared Deviations from Group Means					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
CLUSTER	8	3632721	454090	4.03	<.0001
Error	1021	1.1503E8	112666		

Dependent Variable: compressivestrength

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	45358.5911	5669.8239	23.94	<.0001
Error	1021	241814.4374	236.8408		
Corrected Total	1029	287173.0285			

R-Square	Coeff Var	Root MSE	compressivestrength Mean
0.157949	42.96639	15.38963	35.81784

Welch's ANOVA for compressivestrength			
Source	**DF**	**F Value**	**Pr > F**
CLUSTER	8.0000	32.87	<.0001
Error	33.4910		

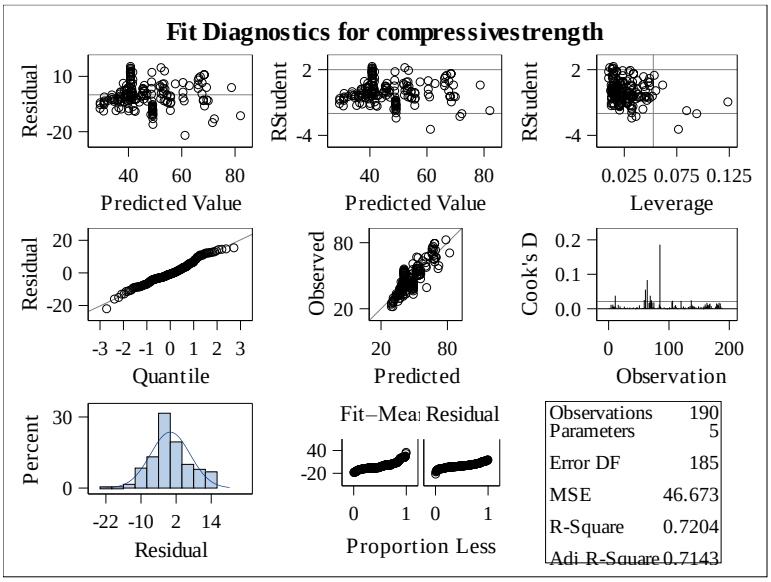
Appendix C: Multiple linear regression diagnostics

Model: MODEL1
Dependent Variable: compressivestrength

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	1	15.98000	2.49284	6.41	<.0001	0
cementwater	1	18.28954	0.95387	19.17	<.0001	1.30680
slagwater	1	18.45323	1.25495	14.70	<.0001	1.20637
flyashwater	1	19.40942	1.91617	10.13	<.0001	1.74084
finewater	1	-1.94854	0.60908	-3.20	0.0016	1.44135

Model: MODEL1
Dependent Variable: compressivestrength

In terms of collinearity, it does not appear that any of the variables have issues with collinearity given that the variance inflation factors are much smaller than 10, so neither must be removed from the model.

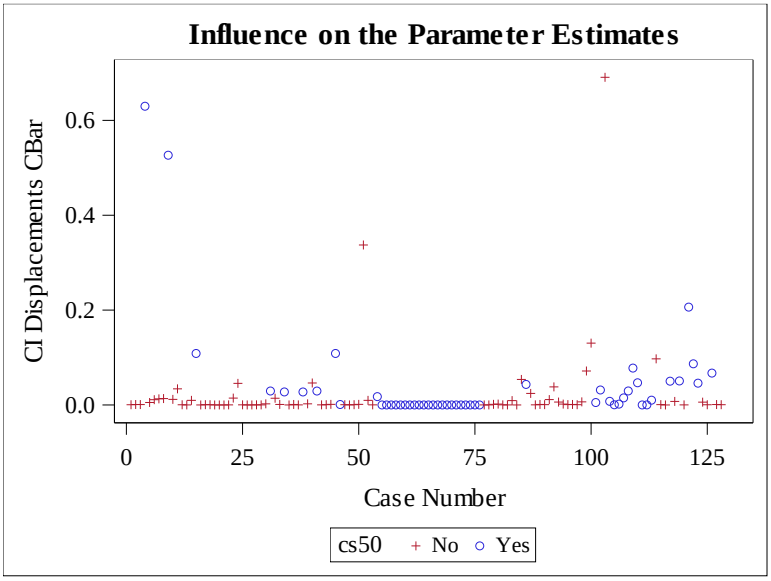


As for the other diagnostics, there does not appear to be a trend between the residuals and the predicted values, and there does not appear to be any readily apparent indicators of homoscedasticity within the same plot. In the Cook's distances plot, there are some concrete samples with high Cook's distances relative to the other samples, but overall, they are all much less than 1; hence, there are no unduly influential observations that require their removal from the model.

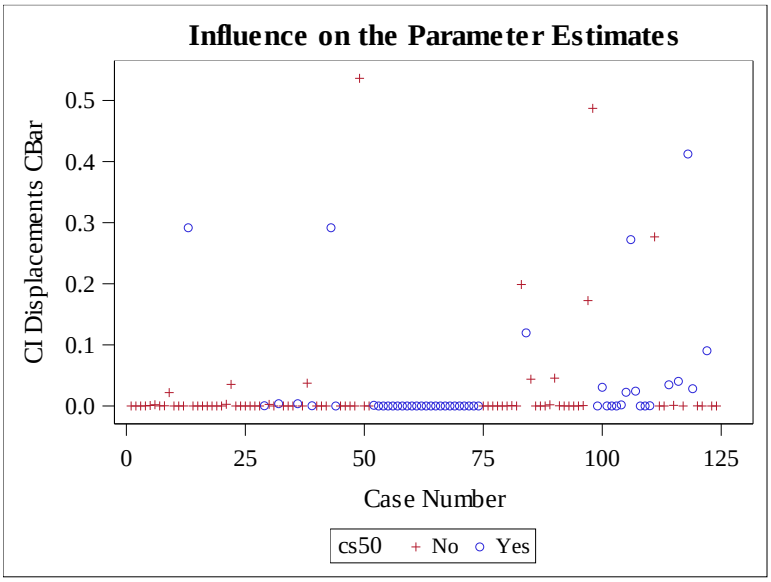
As for the normality assumption, there is a very minimal departure from the normal distribution in the histogram of the residuals, but the quantile plot shows a very minimal departure from the straight line. Therefore, it would not be unreasonable to consider the residuals to be approximately normally distributed.

Appendix D: Removal of unduly influential points

After model selection, the first Cbar plot was generated, which indicated that the model included a few unduly influential concrete samples:



Therefore, the model was refitted after removing the concrete sample with the highest Cbar value, but this had to be performed an addition 4 times, resulting in 5 concrete samples being removed from the model after refitting, which resulted in the following Cbar plot



At this point, the observations with the highest Cbar values were relatively close to the other observations, and the issue of unduly influential observations was resolved.

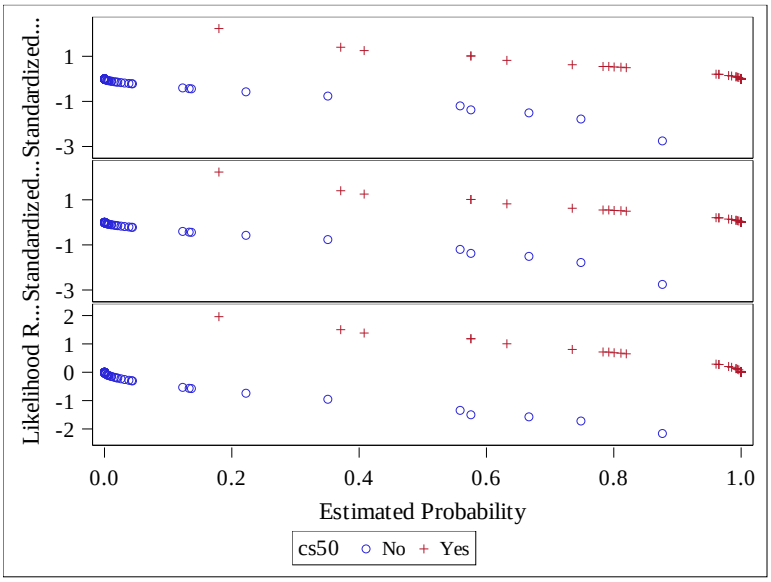
Appendix E: Logistic regression residual diagnostics plot

Model Fit Statistics		
Criterion	**Intercept Only**	**Intercept and Covariates**
AIC	167.523	36.806
SC	170.343	48.087
-2 Log L	165.523	28.806

Since the AIC statistic of the logistic model is much smaller that the AIC for the intercept only model, it can be concluded that the logistic model is significant in predicting whether a concrete sample with an age between 90 days to 100 days had a compressive strength of at least 50 MPa.

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	**DF**	**Pr > ChiSq**
1.2677	8	0.9959

The null hypothesis of the Hosmer-Lemeshow test is not rejected, and therefore, it cannot be concluded that there is enough evidence that suggests the model does not fit the data well.



As for the plots for the predicted probability versus the standard Pearson residuals, the standardized deviance residuals, and the likelihood residuals, the shape of the various measures of residuals remains fairly flat, parallel with residuals equal to zero, up to the estimated probability of 0.7, in which the residuals begin to curve away from the residuals equal to 0; however, since this curvature is not very steep and a more of a gradual bend, this does not entirely invalidate logistic model.

Appendix F: Discriminant analysis using training & testing partition

After randomly partitioning the data set into a training set made up of 80% of the data set and a testing set made up of the remaining 20%, stepwise selection was performed on the training set to determine which variables were the most effective for classifying the age group of a concrete sample.

Stepwise Selection Summary								
Step	**Number In**	**Entered**	**Removed**	**Partial R-Square**	**F Value**	**Pr > F**	**Wilks' Lambda**	**Pr < Lambda**
1	1	compressivestrength		0.3624	94.26	<.0001	0.63755315	<.0001
2	2	cementwater		0.2947	69.20	<.0001	0.44964773	<.0001
3	3	slagwater		0.1617	31.92	<.0001	0.37691798	<.0001
4	4	flyashwater		0.2219	47.12	<.0001	0.29326491	<.0001
5	5	finewater		0.0719	12.78	<.0001	0.27218096	<.0001
6	6	superplasticizerwater		0.0358	6.12	<.0001	0.26243603	<.0001
7	7	coarsewater		0.0154	2.57	0.0255	0.25840000	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	compressivestrength		0.07248937	<.0001
2	2	cementwater		0.11274651	<.0001
3	3	slagwater		0.12893516	<.0001
4	4	flyashwater		0.16258488	<.0001
5	5	finewater		0.17587921	<.0001
6	6	superplasticizerwater		0.18141079	<.0001
7	7	coarsewater		0.18401992	<.0001

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
2289.847755	140	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Like before, the analysis uses quadratic discriminant analysis, and the MANOVA statistics indicate that some discrimination is reasonable between some age groups using some of the variables.

Multivariate Statistics and F Approximations					
S=5 M=0.5 N=410.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.25840000	37.56	35	3464.5	<.0001
Pillai's Trace	0.92009962	26.64	35	4135	<.0001
Hotelling-Lawley Trace	2.22281213	52.19	35	2329.8	<.0001
Roy's Greatest Root	1.91937875	226.76	7	827	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					

Classification Summary for Calibration Data: WORK.CRTRAIN
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into agegroup							
From agegroup	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1	63 56.76	11 9.91	2 1.80	0 0.00	0 0.00	35 31.53	111 100.00
2	9 5.77	41 26.28	43 27.56	4 2.56	0 0.00	59 37.82	156 100.00
3	0 0.00	27 8.06	204 60.90	30 8.96	13 3.88	61 18.21	335 100.00
4	0 0.00	3 3.53	18 21.18	38 44.71	24 28.24	2 2.35	85 100.00
5	0 0.00	0 0.00	7 7.00	35 35.00	20 20.00	38 38.00	100 100.00
6	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	48 100.00	48 100.00
Total	72 8.62	82 9.82	274 32.81	107 12.81	57 6.83	243 29.10	835 100.00
Priors	0.13293	0.18683	0.4012	0.1018	0.11976	0.05749	

Error Count Estimates for agegroup							
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Rate	0.4324	0.7372	0.3910	0.5529	0.8000	0.0000	0.5042
Priors	0.1329	0.1868	0.4012	0.1018	0.1198	0.0575	

Classification Summary for Test Data: WORK.CRTEST

Classification Summary using Quadratic Discriminant Function

In the training data set, we see similar results as before, in that age group 6 has an estimated error rate of 0, whereas the other age groups have very unreliable classification rates.

Number of Observations and Percent Classified into agegroup							
From agegroup	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
1	16 64.00	1 4.00	0 0.00	0 0.00	0 0.00	8 32.00	25 100.00
2	1 3.13	12 37.50	6 18.75	1 3.13	0 0.00	12 37.50	32 100.00
3	0 0.00	6 6.67	57 63.33	7 7.78	4 4.44	16 17.78	90 100.00
4	0 0.00	0 0.00	2 33.33	3 50.00	1 16.67	0 0.00	6 100.00
5	0 0.00	0 0.00	4 12.90	9 29.03	3 9.68	15 48.39	31 100.00
6	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 100.00	11 100.00
Total	17 8.72	19 9.74	69 35.38	20 10.26	8 4.10	62 31.79	195 100.00
Priors	0.13293	0.18683	0.4012	0.1018	0.11976	0.05749	

Error Count Estimates for agegroup							
	1	**2**	**3**	**4**	**5**	**6**	**Total**
Rate	0.3600	0.6250	0.3667	0.5000	0.9032	0.0000	0.4708
Priors	0.1329	0.1868	0.4012	0.1018	0.1198	0.0575	

Even when using a training data set, the error rate estimated using the test data set is concerning, meaning that classification of concrete that is not in age group 6 is unrealiable.