Statistical Computing Final Project

Concrete Compressive Strength Data Set

Spring 2018

Author - Anurag Jain

Part 1: Data Description

The data set for our report is Concrete Compressive Strength Data Set. Using this data set we are predicting the compressive strength of concrete from its given ingredients. We got this data set from UCI machine learning repository but the original data source is Department of Information Management, Chung-Hua University, Taiwan. This data set contains 1030 records, eight input variables and one output variable. The response variable is Concrete Compressive Strength and predictor variables are Cement, Blast Furnace Slag, Fly ash, Water, Superplasticizer, Coarse aggregate, Fine Aggregate and Age. All these predictor variables are of interest in this case as they play a significant role in strength of concrete. We checked for any NULL or unintuitive values using R but did not find anything wrong with the data set. We checked the structure and all the variables are numeric in nature.

Part 2: Summary Statistics

We plot the histograms of all the variables using the hist() function in R and see different distributions for different variables.
Shape of the distribution: We see that variables 'Blast_Furnace_slag', 'Fly_Ash', 'Superplasticizer' and 'Age', are positively skewed. Variables 'Water', 'Fine_Aggregate', 'Coarse_aggregate' and 'cement' have a normal distribution. From this we can infer that the ingredients with positive skew are added in low quantity in the concrete mixture.

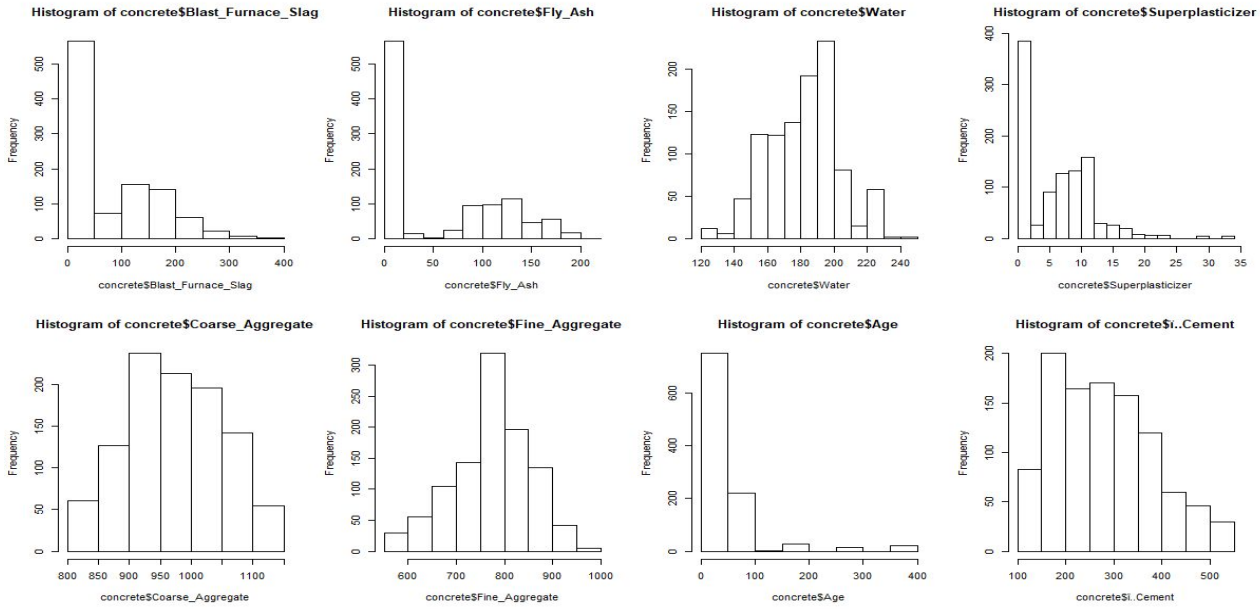

Fig 1. Histograms for predictor variables

The majority data distribution can be inferred from the histogram in Fig. 1.

The histogram below shows distribution of response variable which is a normal distribution curve suggesting that the strength can vary from less to more depending on the ingredients.

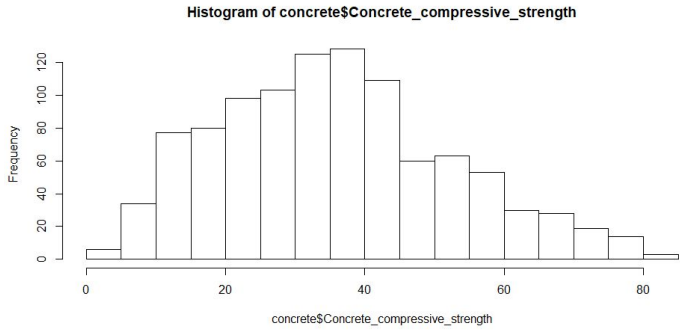

Fig 2. Histogram for response variable (Concrete compressive strength)

The range and center of data is show in table below:

Variable	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
Min.	102	0	0	121.8	0	801	594	1	2.33
Max.	540	359.4	200.1	247	32.2	1145	992.6	365	82.6
Mean	281.2	73.9	54.19	181.6	6.205	972.9	773.6	45.7	35.82

Table 1. Distribution of variables

Part 3: Analysis

We find the correlation between variables cement, water, superplasticizer, Fine_aggregate which are ingredients and are related to each other and also the response variable.

Pearson Correlation Coefficients, N = 1030 Prob > |r| under H0: Rho=0					
	Cement	Water	Superplasticizer	Fine_Aggregate	Concrete_compressive_strength
Cement	1.00000	-0.08159 0.0088	0.09239 0.0030	-0.22272 <.0001	0.49783 <.0001
Water	-0.08159 0.0088	1.00000	-0.65753 <.0001	-0.45066 <.0001	-0.28963 <.0001
Superplasticizer	0.09239 0.0030	-0.65753 <.0001	1.00000	0.22269 <.0001	0.36608 <.0001
Fine_Aggregate	-0.22272 <.0001	-0.45066 <.0001	0.22269 <.0001	1.00000	-0.16724 <.0001
Concrete_compressive_strength	0.49783 <.0001	-0.28963 <.0001	0.36608 <.0001	-0.16724 <.0001	1.00000

Table 2. Correlation matrix

From the table above we see positive correlation between concrete_compressive_strength and cement which tells us that the strength of concrete increases as more cement is added in concrete mixture. We see a inverse relationship between superplasticizer and water which makes sense as superplasticizer is a water reducer. There is a direct relationship between concrete strength and superplasticizer too.

We performed linear regression on our dataset with Compressive Strength as the response variable whereas Cement, Blast Furnace Slag, Fly ash, Water, Superplasticizer, Coarse aggregate, Fine Aggregate and Age as the predictor variables. We chose these predictor variables because all of them play a significant role in compressive strength of concrete. On running linear regression on this model we get the below parameter estimates.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.33121	26.58550	-0.88	0.3804
Cement	1	0.11980	0.00849	14.11	<.0001
Blast_Furnace_Slag	1	0.10387	0.01014	10.25	<.0001
Fly_Ash	1	0.08793	0.01258	6.99	<.0001
Water	1	-0.14992	0.04018	-3.73	0.0002
Superplasticizer	1	0.29222	0.09342	3.13	0.0018
Coarse_Aggregate	1	0.01809	0.00939	1.93	0.0544
Fine_Aggregate	1	0.02019	0.01070	1.89	0.0595
Age	1	0.11422	0.00543	21.05	<.0001

Table 3. Parameter estimates

The first row gives the estimated value of the intercept as -23.33 which is the value of response when all ingredients are zero. The following rows give estimated values for other variables. We can parameter estimate for cement as 0.11 which tell us that for every addition of 1 kg cement in concrete mixture there is an increase of compressive strength by 0.11 units. The same goes for other estimate values too. The column 'Standard error' measures the accuracy of the estimate. A smaller value of standard error is considered better. The next column ‘t Value’ is calculated by dividing Parameter Estimate by Standard Error. This gives the difference in original and predicted value represented in units of standard error. From P value we can infer if that variable is a significant participant of the model or not. If the P value is more than 0.05, that variable is not significant in the model. In this case, ‘Coarse_Aggregate’ and ‘Fine_Aggregate’ have P values more than 0.05. Thus, we can infer that they are not playing significant role in prediction.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	176762	22095	204.32	<.0001
Error	1021	110413	108.14217		
Corrected Total	1029	287175			

Table 4. ANOVA Table

Root MSE	10.39914	R-Square	0.6155
Dependent Mean	35.81796	Adj R-Sq	0.6125
Coeff Var	29.03332		

Table 5. Rsquare values

From ANOVA table above we can see that there are 8 independent variables so the model has 8 degrees of freedom. Total observations (1030) less one gives the corrected total degrees of freedom (as 1029). F value is the mean square due to model divided by mean square error and can be compared to F critical value whose result is shown in the P value. Here, since P value is <0.05, there is a linear relationship between the response and predictor variables.

Table 5 shows the values for R-Square and Adjusted R square approximately 0.61 which shows that the prediction accuracy is around 61%. We concluded earlier that ‘Coarse_Aggregate’ and ‘Fine_Aggregate’ are insignificant of this prediction model. Thus, the final prediction equation can be written as:

Concrete_Compressive_Strength = **-23.33 + 0.12***Cement **+ 0.1***Blast_Furnace_Slag **+ 0.09***Fly_ash **- 0.15***Water **+ 0.29***Superplasticizer **+ 0.02***Coarse_aggregate Fine_Aggregate **+0.11***Age

Appendix:

Structure of Dataset:

Variable	Classification	Units	Nature
Cement (component 1)	quantitative	kg in a m3 mixture	Input Variable
Blast Furnace Slag (component 2)	quantitative	kg in a m3 mixture	Input Variable
Fly Ash (component 3)	quantitative	kg in a m3 mixture	Input Variable
Water (component 4)	quantitative	kg in a m3 mixture	Input Variable
Superplasticizer (component 5)	quantitative	kg in a m3 mixture	Input Variable
Coarse Aggregate (component 6)	quantitative	kg in a m3 mixture	Input Variable
Fine Aggregate (component 7)	quantitative	kg in a m3 mixture	Input Variable
Age	quantitative	Day (1~365)	Input Variable
Concrete compressive strength	quantitative	MPa	Output Variable

Table 6 : Structure of Concrete conpressive strength dataset

Summary Statistics:

Variable	Cement	Blast Furnace Slag	Fly Ash	Water	Superplasticizer	Coarse Aggregate	Fine Aggregate	Age	Concrete compressive strength
Min.	102	0	0	121.8	0	801	594	1	2.33
1st Qu.	192.4	0	0	164.9	0	932	731	7	23.71
Median	272.9	22	0	185	6.4	968	779.5	28	34.45
Mean	281.2	73.9	54.19	181.6	6.205	972.9	773.6	45.7	35.82
3rd Qu.	350	142.9	118.3	192	10.2	1029.4	824	56	46.13
Max.	540	359.4	200.1	247	32.2	1145	992.6	365	82.6

Table 7 : Summary statistics for dataset

Data source:

https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength