Gulshankumar Bakle **Project Report**

Description of data and Research Objective:

Concrete forms to be the most essential material in building blocks of any civil engineering project. The compressive strength of the concrete material depends upon several components. Given a dataset [1] collected by Department of Information Management, Chung Hua University, Taiwan, details the information about the compressive strength of concrete measured in Megapascal(MPa), and the factors on which this strength depends upon. The dataset consists of 7 such factors (measured in kilogram meter^3-kgm^3) and 1 age factor (measured in days) which govern the response concrete compressive strength variable. Dataset consists of 1030 distinct rows. The factor variables are mentioned below:

Component 1: Cement	Component 2: Blast furnace Slag
Component 3: Fly Ash	Component 4: Water
Component 5: Superplasticizer	Component 6: Coarse aggregate
Component 7: Age	Component 8: Fine aggregate

Given this dataset, the following are the research objectives:

- To find which of the above components are most significant in determining the strength of the concrete.
- What is estimated strength of concrete in MPa with distribution of components given as cement-320 kgm^3, blast furnace slag- 5 kgm^3, fly ash- 55 kgm^3, water-200 kgm^3, Superplasticizer -30 kgm^3, Age-25 days, coarse aggregate 980 kgm^3, fine aggregate 800 kgm^3.
- What variables are also associated for the strength of concrete, after accounting for cement?

Summary of Statistical findings:

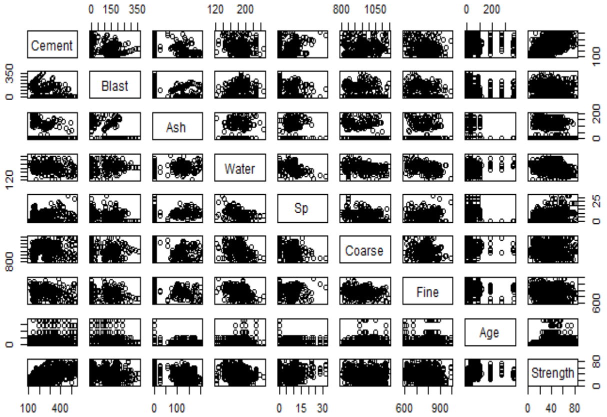
On performing statistical analysis on the given dataset, it was inferred that the relative compressive strength of the concrete is better governed by selective components which are, cement, fly ash, blast furnace slag, water, superplasticizer. Components like coarse aggregate and fine aggregate hold less significance in determining the strength of concrete. Significantly higher p-value of 0.354 was observed without coarse aggregate and fine aggregate as compared with these components. With the given research objective parameters, the predicted strength was 35.57 MPa and with 95% confidence interval the predicted strength was between 29.64 MPa and 41.5 MPa. After accounting for cement, strength of concrete depends upon blast furnace slag, fly ash and water.

Scope of Inference:

The collection of data for compressive strength of concrete details about significant components determining the strength. From the study performed, it seemed significantly true that it was a completely observational study. The values for factor variables cement, fly ash, water, coarse aggregate, superplasticizer, age and blast furnace slag were recorded based on observations done in the laboratories. The strength of concrete was also recorded from the observations. Thus, response variable does not give any evidence for any treatment effect. Hence no causal effect can be established. The materials were also sampled randomly, hence inference can be made to general population. Thus, no causal effects can be established since no treatment effect was noticed.

Details:

After performing statistical analysis on the data, it was observed that there were 8 explanatory variables (components) responsible for the determining the compressive strength of the concrete material. The distribution of observations of explanatory variables was not completely normally distributed. Observations for blast furnace slag and ash were slightly left skewed. Applying logarithmic transformation did not removed the skewness in the observations, as observations were right skewed. Hence any kind of transformation was not considered. There were no incomplete or null values in the dataset. Following plots were plotted to visualize the data better:



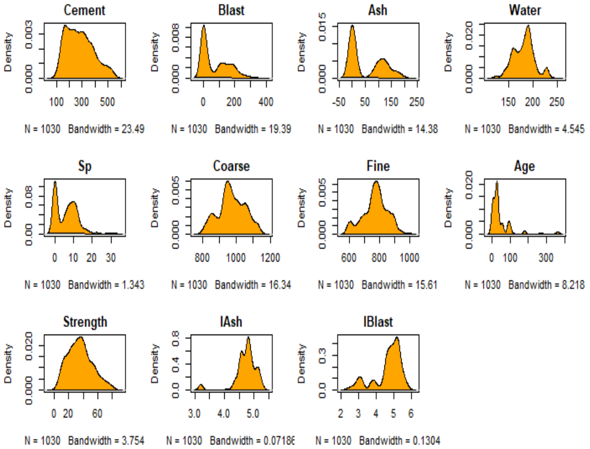


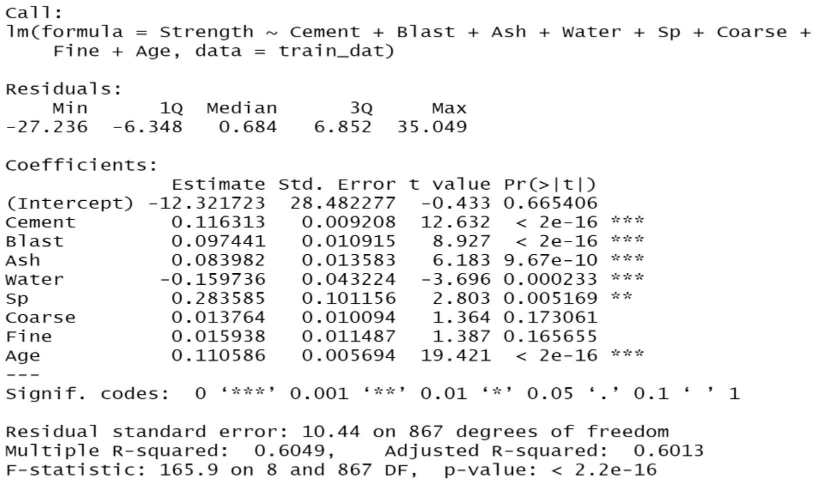
Figure 1: matrix plot of all variables *Figure 2: Histogram plot of distribution*

Since there were 1030 observations in the dataset, the data was randomly split into training and testing data in the ratio of 85:15. 85% of data (876) was randomly considered into training sample, while 15% (154) observations as testing data. Initially model fitting was considered only for the training data, predictions were analyzed with testing data.

To answer the first research question following linear model was considered:

Model 1:μ(Strength|,)=β0+ β1*Cement + β2*Blast + β3*FlyAsh + β4*Water + β5*Superplasticizer + β6*CoarseAggregate + β7*FineAggregate + β8*Ash

It was observed that coefficients of coarse aggregate and fine aggregate had a higher p-value greater than 0.05. Following table summarizes the p-values for all parameters.



```
Call:
lm(formula = Strength ~ Cement + Blast + Ash + Water + Sp + Coarse +
    Fine + Age, data = train_dat)

Residuals:
    Min      1Q  Median      3Q     Max
-27.236  -6.348   0.684   6.852  35.049

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.321723  28.482277  -0.433 0.665406
Cement        0.116313   0.009208  12.632  < 2e-16 ***
Blast         0.097441   0.010915   8.927  < 2e-16 ***
Ash           0.083982   0.013583   6.183 9.67e-10 ***
Water        -0.159736   0.043224  -3.696 0.000233 ***
Sp            0.283585   0.101156   2.803 0.005169 **
Coarse        0.013764   0.010094   1.364 0.173061
Fine          0.015938   0.011487   1.387 0.165655
Age           0.110586   0.005694  19.421  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.44 on 867 degrees of freedom
Multiple R-squared:  0.6049,    Adjusted R-squared:  0.6013
F-statistic: 165.9 on 8 and 867 DF,  p-value: < 2.2e-16
```

Figure 3: Model1 summary

As it can be observed coarse and fine aggregate have p values 0.173 and 0.1656 which are significantly higher as compared to others. This could be an indicative that these two predictors are not much significant in determining strength of concrete. So, another reduced model was considered which did not include these variables. Following equation describes the reduced model:

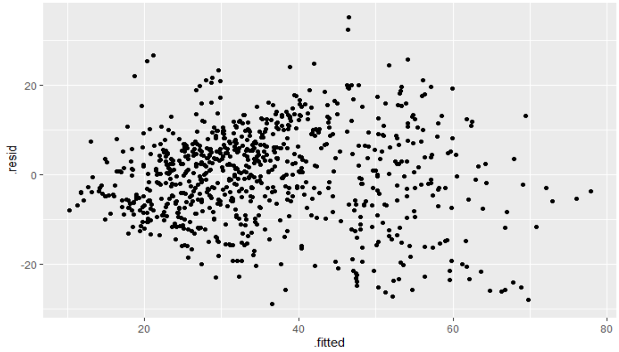
Model 2: $\mu(Strength|,)=\beta_0+ \beta_1*Cement + \beta_2*Blast + \beta_3*FlyAsh + \beta_4*Water + \beta_5*Superplasticizer+ \beta_6*Age$

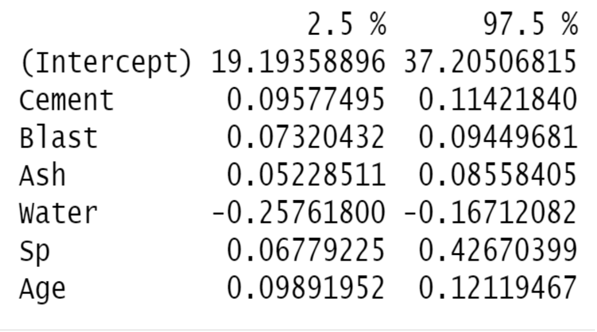
Analysis of variance (ANOVA) test was performed in order to compare model1 and model2. Following null and alternative hypothesis was considered:

Ho: The coefficients of coarse aggregate or fine aggregate or both are zero.

Ha: The coefficients are not zero.

Null hypothesis is an indicative for significance of reduced model, while alternative hypothesis considers full model. It was observed that significantly higher p-value of 0.354 was observed. This concludes that we fail to reject the null hypothesis, since we do not have enough evidence. Thus, reduced model without coarse aggregate and fine aggregate is better model. Hence, factors cement, fly ash, blast furnace slag, water, age and superplasticizer are much significant in determining compressive strength of material. Following figure 3 describes residual vs fitted values based on reduced model. The plot seems to be approximately symmetrical below and above line of zero, hence justifying equal variance assumption.





```
                   2.5 %      97.5 %
(Intercept) 19.19358896 37.20506815
Cement       0.09577495  0.11421840
Blast        0.07320432  0.09449681
Ash          0.05228511  0.08558405
Water       -0.25761800 -0.16712082
Sp           0.06779225  0.42670399
Age          0.09891952  0.12119467
```

Figure 3: Residual vs fitted plot

Figure 4: 95%Confidence interval of betas

Figure 4 lists the confidence interval of coefficients of reduced model. To answer our second research question, we consider model 1, since values of coarse aggregate and fine aggregate are also given. On predicting the values by plugging the given values for parameters, strength of 35.57 MPa was predicted. With 95% confidence interval, the predicted strength was between 29.64MPa and 41.5 MPa.

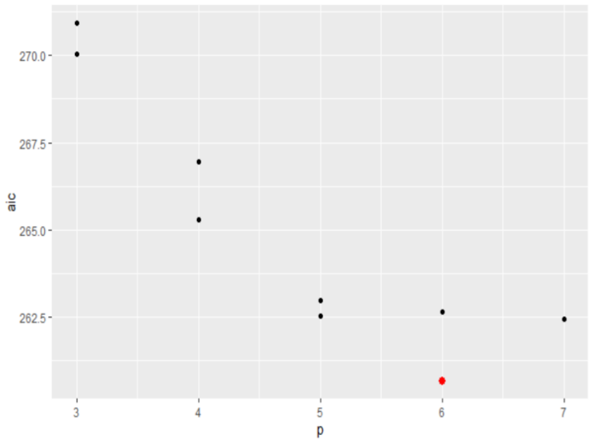
Adjusted R squared values for both models were compared. Model 1 had 0.601, while model2 had 0.60 values for adjusted R-squared. There was not much significant difference in the adjusted r squared values. Thus model 2 was consider as best model to answer third research question. Following observations were noted after considering different criteria for model selection:

Minimum Akaike Information Criterion (Aic): 262.4402

Minimum Bayesian Information Criterion (Bic): -931.9739

Minimum Mallow Cp: 7.00000

Maximum adjusted R sq: 0.611



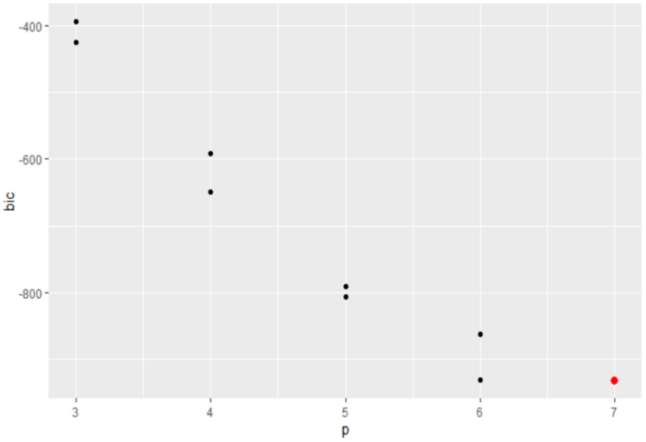
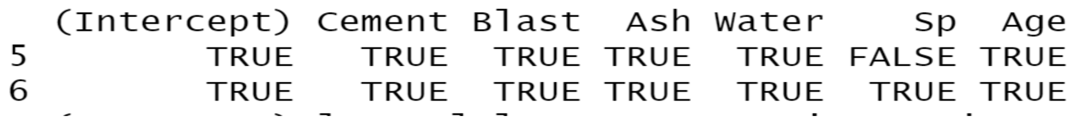


Figure 6: Aic vs # of predictors

Figure 7: Bic vs # of predictors

Above two plots describe the aic and bic plots. As it can be observed that maximum number of variables that the model should have after accounting for cement, could be 6 or 7 (indicated by red dot). Below table summarizes variable selected results:



```
  (Intercept) Cement Blast  Ash Water    Sp  Age
5        TRUE   TRUE  TRUE TRUE  TRUE FALSE TRUE
6        TRUE   TRUE  TRUE TRUE  TRUE  TRUE TRUE
```

Thus, after accounting for cement, the strength of concrete depends upon blast furnace slag, fly ash, water and age. The minimum number of parameters the model could have is 6, by excluding superplasticizer.

References:

[1]. (n.d.). Retrieved from https://archive.ics.uci.edu/ml/datasets/Concrete Compressive Strength