Compressive Strength of Concrete Prediction using R Software

TTTR6124 – Assignment 3

Name : Chairani Tiara Sayyu
Student ID : P104718

1. Introduction

Concrete is a composite material composed of fine and coarse aggregate bonded together with a fluid cement (cement paste) that hardens over time. Concrete frequently used in building construction due its high compressive strength, high durability, and superior fire resistance.

Compressive strength is important in concrete as a construction material since it is used to resist compressive stresses. Compressive strength of concrete is affected by the quantity of each components. The main components of concrete are cement, water, aggregates, admixture, and concrete additions and supplementary cementitious material (SCM). Different mixture of components give different compressive strength. In this research, the dataset used was the "predicting compressive strength of concrete" dataset taken from UCI machine learning repository. There are 8 input variables used in this research, which are cement, slag, fly ash, water, super plasticizer, fine aggregate, coarse aggregate, and age. By using R software, the prediction of compressive strength of concrete with different composition can be done.

2. Research Objective

There are 3 objective in this research, which are:

1. To evaluate the correlation between input variables and between input and output variables
2. To evaluate variables that are most significant in predicting compressive strength
3. To make a linear equation predicting the compressive strength of concrete

3. Methodology

3.1 Data

Data used in this research are data titled "Predicting Compressive Strength of Concrete" dataset taken from UCI machine learning repository. The data consist of 1030 instances and 9 attributes. There are 8 input variables, which are cement (kg/m^3), slag (kg/m^3), fly ash (kg/m^3), water (kg/m^3), super plasticizer (kg/m^3), coarse aggregate (kg/m^3), fine aggregate (kg/m^3), and age (day). The output variable is the compressive strength of concrete (MPa).

3.2 Linear Regression using R Software

Simple linear regression is used to predict quantitative output y on the basis of one single input variables x. Multiple linear regression is an extension of simple linear regression with multiple input variables x. With three input variables x, the prediction of y can be expressed as equation below:

$$y = b_0 + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3$$

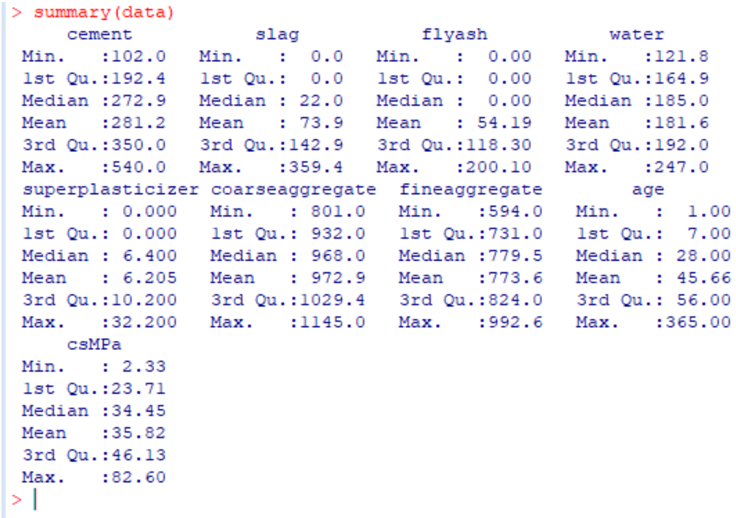
The b values in the equation above is called regression weight (or beta coefficients). B values measure the association between input variable and output variable.

In R software, the linear regression model can be made and the equation will obtained through the summary of linear regression model. Also, the accuracy of the model can be assessed by examining the R-squared and residual standard error (RSE).

4. Data Analysis

4.1 Descriptive Data Analysis

By using R software, the summary of each variables can be obtained, which can be seen in figure 1.



```
> summary(data)
     cement           slag            flyash           water      
 Min.   :102.0   Min.   :  0.0   Min.   :  0.00   Min.   :121.8  
 1st Qu.:192.4   1st Qu.:  0.0   1st Qu.:  0.00   1st Qu.:164.9  
 Median :272.9   Median : 22.0   Median :  0.00   Median :185.0  
 Mean   :281.2   Mean   : 73.9   Mean   : 54.19   Mean   :181.6  
 3rd Qu.:350.0   3rd Qu.:142.9   3rd Qu.:118.30   3rd Qu.:192.0  
 Max.   :540.0   Max.   :359.4   Max.   :200.10   Max.   :247.0  
 superplasticizer coarseaggregate  fineaggregate        age        
 Min.   : 0.000   Min.   : 801.0   Min.   :594.0   Min.   :  1.00  
 1st Qu.: 0.000   1st Qu.: 932.0   1st Qu.:731.0   1st Qu.:  7.00  
 Median : 6.400   Median : 968.0   Median :779.5   Median : 28.00  
 Mean   : 6.205   Mean   : 972.9   Mean   :773.6   Mean   : 45.66  
 3rd Qu.:10.200   3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00  
 Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00  
     csMPa      
 Min.   : 2.33  
 1st Qu.:23.71  
 Median :34.45  
 Mean   :35.82  
 3rd Qu.:46.13  
 Max.   :82.60  
> 
```

Figure 1 Summary of Each Variables in the Dataset

By using R software also, the boxplot of each variables can be made, to check whether there are any outliers or not. By knowing the interquartile range of each variables, the outliers can be treated. The boxplot of all variables can be seen in figure 2. From figure 2, it can be seen that the variables which have outliers are slag, water, super plasticizer, fine aggregate, age, and compressive strength of concrete.

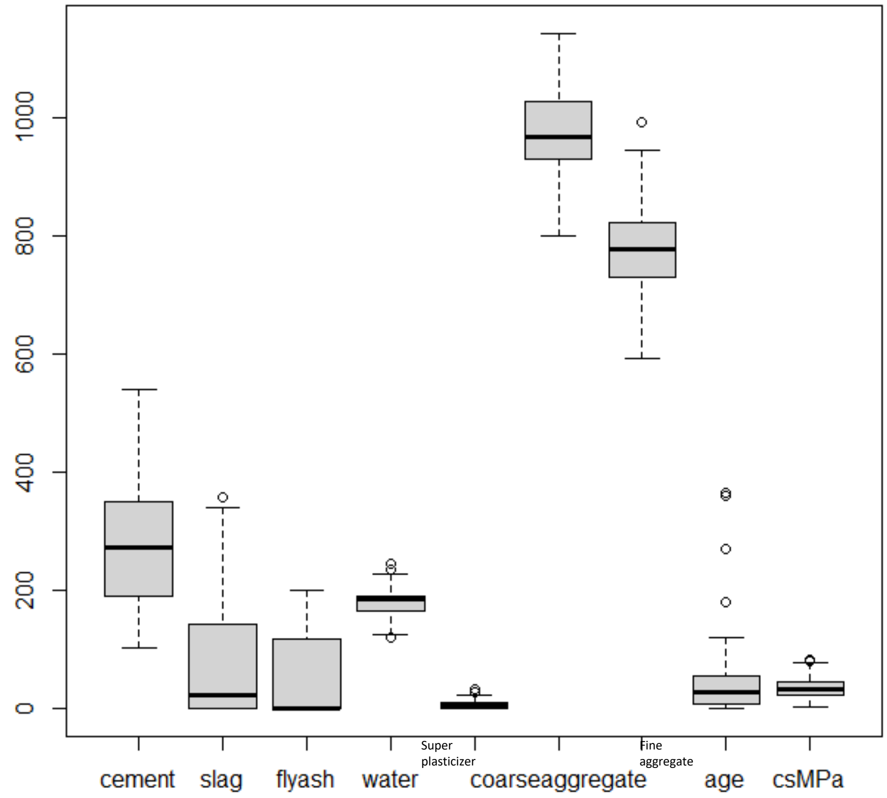


Figure 2 Boxplot of Each Variables

4.2 Inference Data Analysis

4.2.1 The Correlation between All Variables

In the first inference data analysis, the correlation between all variables are assessed using R software. The scatter plot between variables can be seen in figure 3.

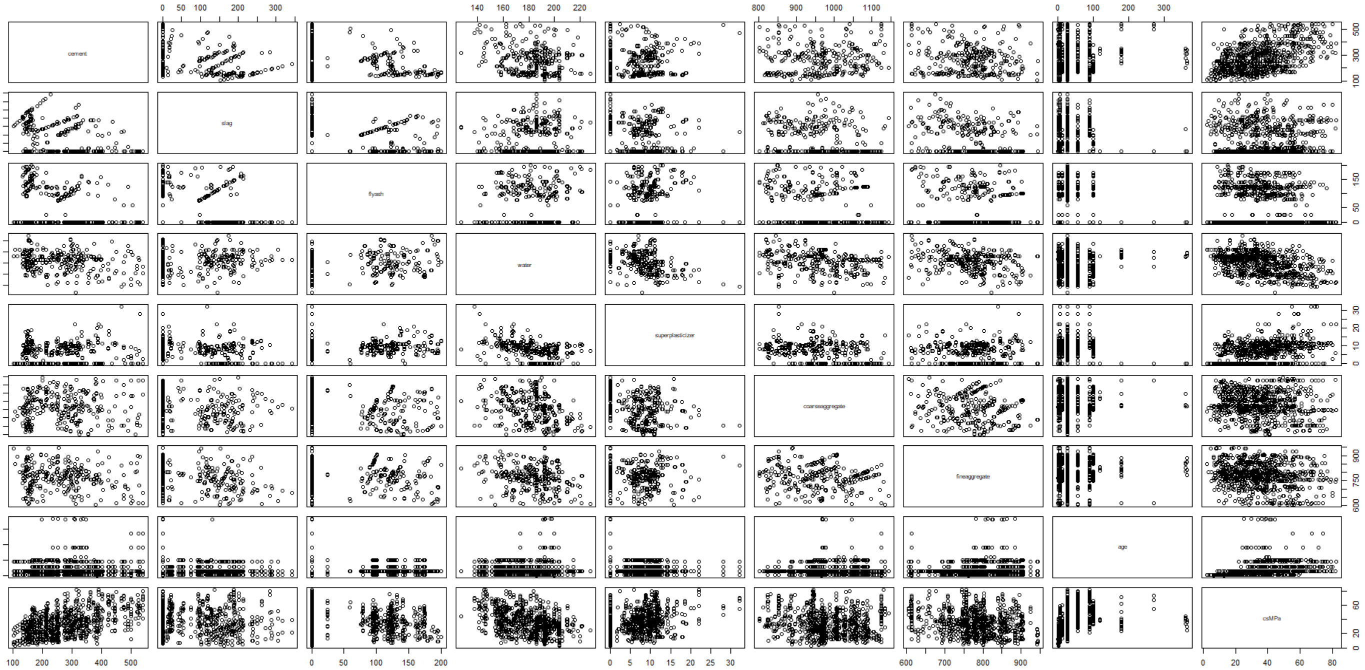
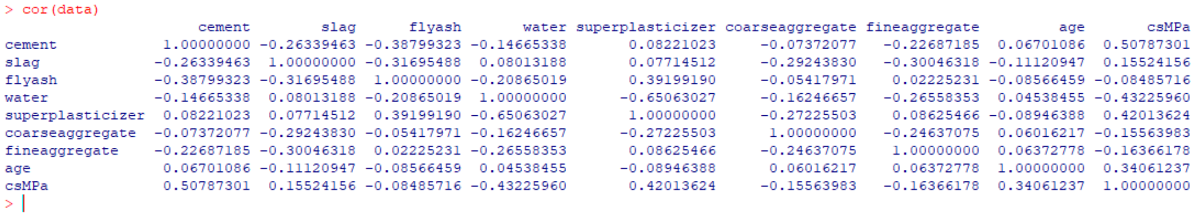


Figure 3 Scatter Plots between All Variables

By just seeing the scatter plots between all variables, the correlation between variables cannot be seen since it does not show a linear trend. By using R software, the correlation between variables can be seen in figure 4.



```
> cor(data)
                     cement        slag      flyash       water superplasticizer coarseaggregate fineaggregate         age       csMPa
cement           1.00000000 -0.26339463 -0.38799323 -0.14665338       0.08221023     -0.07372077   -0.22687185  0.06701086  0.50787301
slag            -0.26339463  1.00000000 -0.31695488  0.08013188       0.07714512     -0.29243830   -0.30046318 -0.11120947  0.15524156
flyash          -0.38799323 -0.31695488  1.00000000 -0.20865019       0.39199190     -0.05417971    0.02225231 -0.08566459 -0.08485716
water           -0.14665338  0.08013188 -0.20865019  1.00000000      -0.65063027     -0.16246657   -0.26558353  0.04538455 -0.43225960
superplasticizer 0.08221023  0.07714512  0.39199190 -0.65063027       1.00000000     -0.27225503    0.08625466 -0.08946388  0.42013624
coarseaggregate -0.07372077 -0.29243830 -0.05417971 -0.16246657      -0.27225503      1.00000000   -0.24637075  0.06016217 -0.15563983
fineaggregate   -0.22687185 -0.30046318  0.02225231 -0.26558353       0.08625466     -0.24637075    1.00000000  0.06372778 -0.16366178
age              0.06701086 -0.11120947 -0.08566459  0.04538455      -0.08946388      0.06016217    0.06372778  1.00000000  0.34061237
csMPa            0.50787301  0.15524156 -0.08485716 -0.43225960       0.42013624     -0.15563983   -0.16366178  0.34061237  1.00000000
> |
```

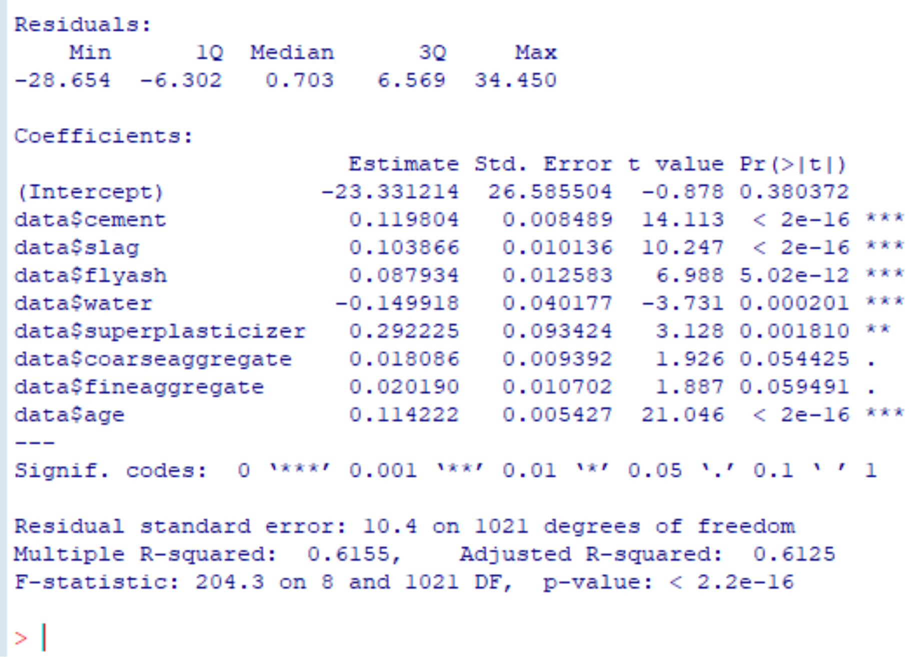
Figure 4 Correlation Coefficient between All Variables

From figure 4, it can be seen that cement and strength has the highest correlation coefficient in positive direction. It can be concluded that the higher quantity of cement added in concrete mix, the higher compressive strength of concrete can be obtained. From figure 4, it also can be seen that water and strength has the second highest correlation coefficient in negative direction, which means the higher the water content in concrete mix, the lower compressive strength of concrete can be obtained.

From figure 4, the correlation between input variables can also be seen. For example, the correlation coefficient between water and super plasticizer is -0.65, which means the quantity of water and quantity of super plasticizer have a strong negative linear relationship. Another example is between fly ash and super plasticizer, where the correlation coefficient is 0.39, which means the increase of usage of fly ash makes an increase of usage of super plasticizer.

4.2.2 Multiple Linear Regression

By using R software, multiple linear regression can be done. The results summary of multiple linear regression between 8 input variables and 1 output variables can be seen in figure 5.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-28.654  -6.302   0.703   6.569  34.450 

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)    
(Intercept)            -23.331214  26.585504  -0.878 0.380372    
data$cement              0.119804   0.008489  14.113  < 2e-16 ***
data$slag                0.103866   0.010136  10.247  < 2e-16 ***
data$flyash              0.087934   0.012583   6.988 5.02e-12 ***
data$water              -0.149918   0.040177  -3.731 0.000201 ***
data$superplasticizer    0.292225   0.093424   3.128 0.001810 ** 
data$coarseaggregate     0.018086   0.009392   1.926 0.054425 .  
data$fineaggregate       0.020190   0.010702   1.887 0.059491 .  
data$age                 0.114222   0.005427  21.046  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.4 on 1021 degrees of freedom
Multiple R-squared:  0.6155,	Adjusted R-squared:  0.6125 
F-statistic: 204.3 on 8 and 1021 DF,  p-value: < 2.2e-16

> |
```

Figure 5 Summary of Multiple Linear Regression

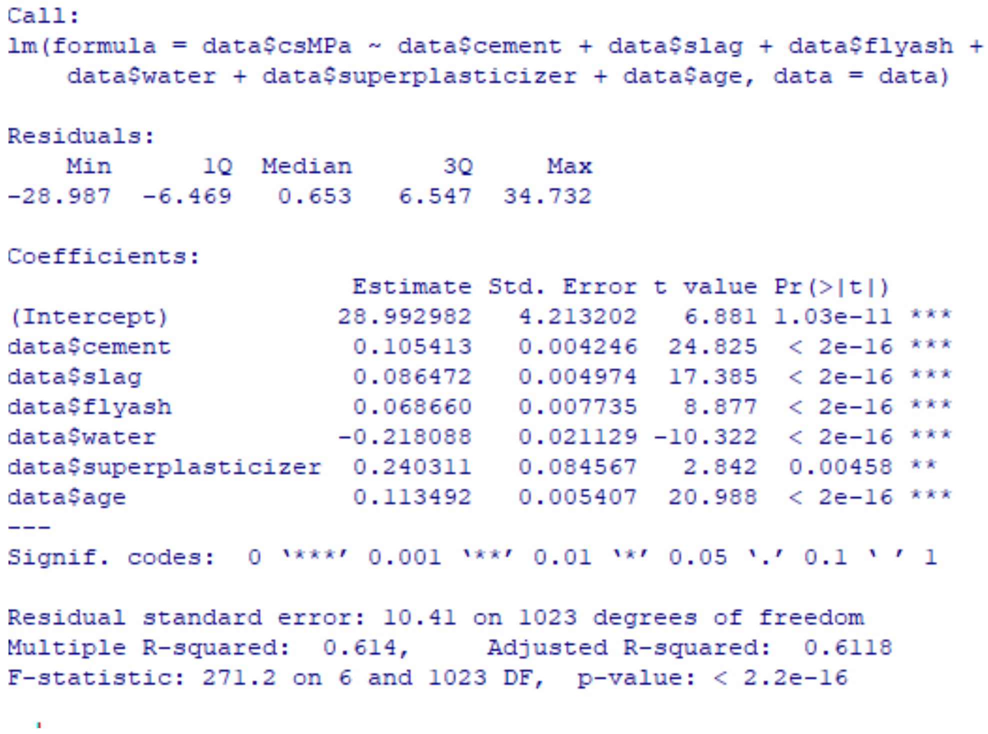
From the results of multiple linear regression, the p-value of each input variables can be seen. From figure 5, cement, slag, fly ash, and water are highly significance input variables, followed by water. Meanwhile, the least significant variables are coarse aggregate and fine aggregate. From the result, the linear model of predicting concrete’s compressive strength is as follow:

$$\begin{aligned}
&\mathit{Concrete\ Compressive\ Strength} \\
&\qquad = -23.33 + 0.12 \times (cement) + 0.10 \times slag \\
&\qquad + \ 0.088 \times (flyash) - 0.15 \times (water) \\
&\qquad + 0.29 \times (superplasticizer) \\
&\qquad + 0.018 \times (coarse\ aggregate) \\
&\qquad + 0.02 \times (fine\ aggregate) + 0.11 \times (age)
\end{aligned}$$

From the result, the adjusted R-squared is 0.6125, which means 61.25% of the variation in concrete’s compressive strength can be explained by the model containing cement, slag, fly ash, water, super plasticizer, coarse aggregate, fine aggregate, and age.

From the results, the multiple R-squared is 0.6155 which means that 8 input variables explain 61.55% of the variability in concrete’s compressive strength.

Since variables coarse aggregate and fine aggregate are least significance, another linear regression model can be made. The results of the new linear regression model can be seen in figure 6.



```
Call:
lm(formula = data$csMPa ~ data$cement + data$slag + data$flyash + 
    data$water + data$superplasticizer + data$age, data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-28.987  -6.469   0.653   6.547  34.732 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           28.992982   4.213202   6.881 1.03e-11 ***
data$cement            0.105413   0.004246  24.825  < 2e-16 ***
data$slag              0.086472   0.004974  17.385  < 2e-16 ***
data$flyash            0.068660   0.007735   8.877  < 2e-16 ***
data$water            -0.218088   0.021129 -10.322  < 2e-16 ***
data$superplasticizer  0.240311   0.084567   2.842  0.00458 ** 
data$age               0.113492   0.005407  20.988  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.41 on 1023 degrees of freedom
Multiple R-squared:  0.614,	Adjusted R-squared:  0.6118 
F-statistic: 271.2 on 6 and 1023 DF,  p-value: < 2.2e-16
```

Figure 6 Summary of the New Multiple Linear Regression Model

From figure 6, it can be seen that the multiple R-squared for the new model is 0.614, which is lower than the original linear regression model. Since the multiple R-squared for the new model is lower, the original linear regression model will be used.

5. Conclusion

The correlation between all variables can be evaluated using R software, where cement and concrete compressive strength has a high correlation coefficient compared to other variables. The usage of fly ash and super plasticizer has a moderate linear relationship in positive direction, while the usage of water and super plasticizer has a strong linear relationship in negative direction.

From the results, variables with high significance are cement, slag, fly ash, and water, while coarse aggregate and fine aggregate are the least significant variable. By doing linear regression, a model was obtained to predict the compressive strength of concrete. The model is as below.

$$\begin{aligned}
&Concrete\ Compressive\ Strength \\
&\quad = -23.33 + 0.12 \times (cement) + 0.10 \times slag + 0.088 \times (flyash) \\
&\quad - 0.15 \times (water) + 0.29 \times (superplasticizer) \\
&\quad + 0.018 \times (coarse\ aggregate) + 0.02 \times (fine\ aggregate) \\
&\quad + 0.11 \times (age)
\end{aligned}$$

The model has accuracy of 61.55%, which is still low. It can be increased by applying different model, such as random forest.