PSTAT 126 – REGRESSION ANALYSIS

FALL 2019

Analysis of Compressive Strength of Concrete

Andrew Kung

Brian Ngo

Section: Wang, W: 11-11:50 am

Introduction

Our project is focused around studying the compressive strength of concrete (strength), using 8 attributes within the "Concrete Compressive Strength Data Set" provided by UCI Machine Learning Repository. We want to figure out which attributes can be used in our model to predict concrete compressive strength measured in megapascal pressure units (MPa). We would also like to determine which predictors are best used in estimating compressive strength.

Questions of Interest

We consider the following questions:

**Q1.** Does the effect of superplasticizer on water depend on age?

**Q2.** What compressive strength can we expect from a concrete with no blast furnace slag, fly ash, and average values for cement, water, superplasticizer, and age?

**Q3.** Does the effect of superplasticizer on itsel

Regression Method

To answer our research questions, we must create an appropriate linear regression model. We first determine if our data has multicollinearity issues, and if any of our predictors must be discarded. Using the remaining predictors, we will construct a model using a stepwise regression using AIC criterion, and then conduct residual analysis to correct ensure our model meets line conditions. Lastly, we add any interaction terms necessary and check and remove studentized points to improve our model.

We can then answer our research questions by using the following methods:

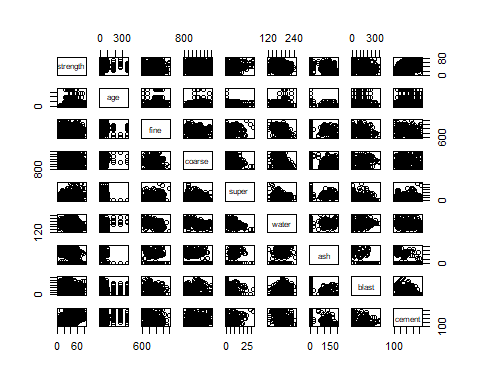
**Q1.** We conduct a general linear F-test with the null hypothesis being that the regression coefficient of the interaction terms, superplasticizer on water and superplasticizer on age are zero, and the alternative being that they are not both zero.

**Q2.** We calculate a 95% prediction interval for a new value of strength given that there is no blast furnace slag nor ash, and average values for cement, water, superplasticizer, and age.

Regression Analysis, Results, and Interpretation

Determining Multicollinearity from the Data

We begin by plotting the scatterplot matrix which includes the response and all predictors.



From the scatterplot matrix there seems to be a positive relationship of strength with our predictors: age, cement, and super. We can also discern a negative relationship of strength with our predictor, water. By observing the correlation matrix, we determine that there is no severe correlation issues between our predictors. Multicollinearity will not be problematic in our regression. As such, there is no need to remove any of the predictors. For the correlation matrix, please refer to the appendix (Correlation Matrix).

Selecting our Predictors

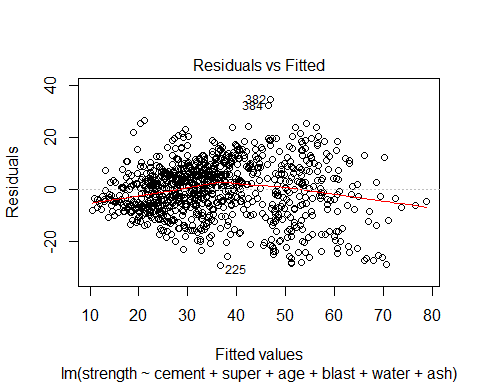
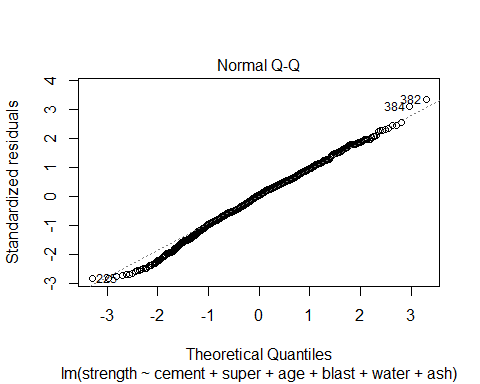
We use a Stepwise Regression using Akaike’s Information Criterion (AIC) as our criteria to determine which predictors will be present in our final model.

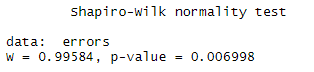
Using this regression methods results in the following model:



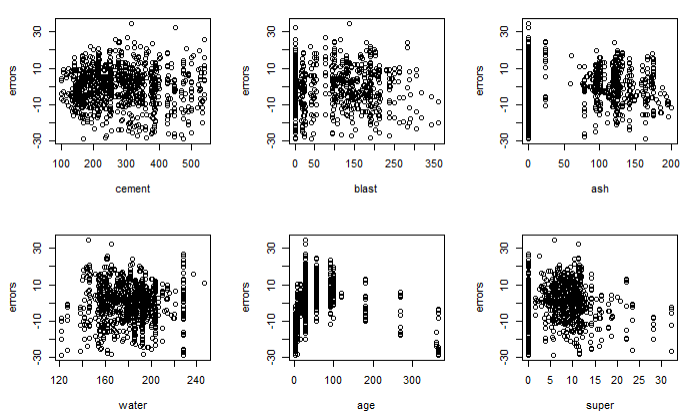
For the R output refer to appendix (Stepwise Regression using AIC).

Residual Diagnostics and Transformations

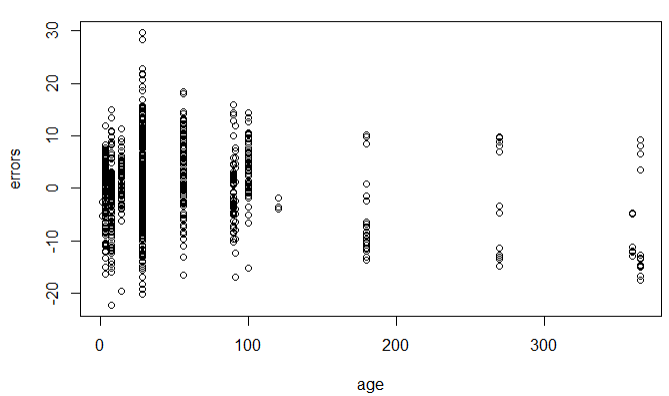
Now we test to see if our model meets our LINE conditions for a linear model:



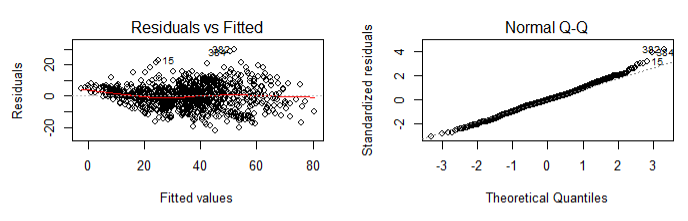
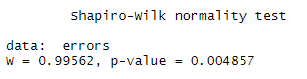
The Residuals vs. Fit plot shows a curved trend of the residuals, which means we fail to meet the linearity condition. Running the Shapiro-Wilk normality test also indicates that our residuals does not follow normal. We perform transformations to correct these issues.



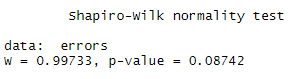
From the above plots we notice that age exhibits a curved pattern. We use a log-transformation on age and see if it will reduce how severe the Residual vs. Predictor plot appears.



The log-transformation reduces the curvature of the plot. We now perform plot diagnostics with log-transformed age to determine if our linearity condition is met.



Applying the log transformation results in the linearity condition to be met, however it does not fix our normality issue as displayed with the Shapiro-Wilk normality test. We can correct the normality condition by removing the externally studentized residuals until our model meets the condition. We delete 2 outliers as defined by externally studentized points and check normality with the Shapiro-Wilk test.



Removing the 2 outliers resulted in our normality condition being met. After applying plot diagnostics and transformations our model now appears as follows:



Addition of Interaction Terms

We will now check for interaction terms with the general linear F-test. In particular, we are interested in the interaction between super and water, and super and age as they will be needed to answer our research question. Adding the interaction terms results in the model:

We can determine to add additional interaction terms between the predictor, super, and another predictor. We choose to check whether adding the interaction of super on itself should be added to the model. We check this by using a general linear F-test with the following hypotheses:





The p-value comes out as p =