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 interaction of cement and water have a significant effect on the compressive strength of concrete?

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

**Q3.** What compressive strength value do we expect for all concrete that ages to 365 days and average values for the other factors?

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. After this, we will construct a model with stepwise regression using AIC criterion and compare it to the Best Subset Regression model to see which one is better. Then, we will check for the strongest interaction terms to add to our model with the General Linear F-test. Then, we will conduct residual analysis to correct ensure our model meets LINE conditions.

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 term, cement and water is zero with the alternative being that they are not both zero.

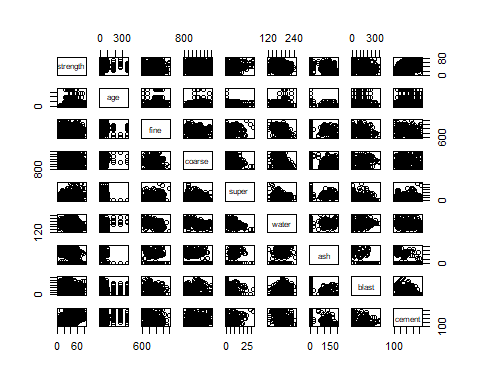
**Q2.** We calculate a 95% prediction interval for a new value of strength given that there is no blast furnace slag nor ash, and average values for cement, water, superplasticizer, and age.

**Q3.** We calculate the 95% confidence interval for the value of strength that concrete has when it ages to 365 days with all other factors fixed.

Regression Analysis, Results, and Interpretation

Determining Multicollinearity from the Data

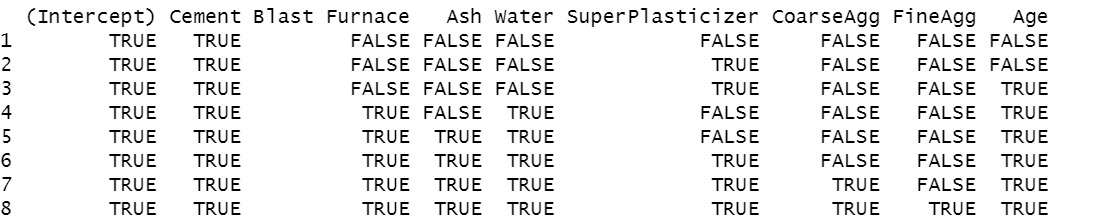
We begin by plotting the scatterplot matrix which includes the response and all predictors.



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 first try to use Best Subset Regression and choose the model with the highest adjusted R-squared value. The results of our Best Subset Regression is the following table which tells us that our full model is the best model. Please refer to the appendix (Best Subset Adjusted R-squared).





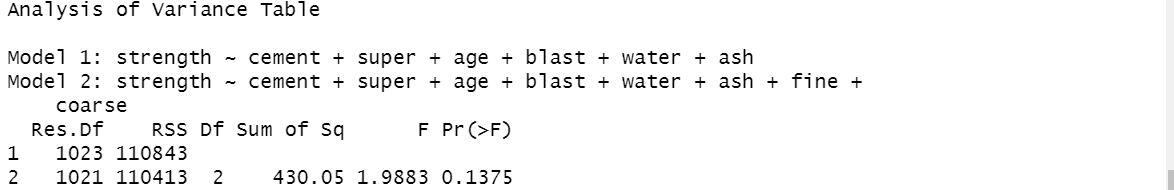


We now use a Stepwise Regression using Akaike’s Information Criterion (AIC) as our criteria to determine which predictors will be present in our final model.

Using this regression methods results in the following model:



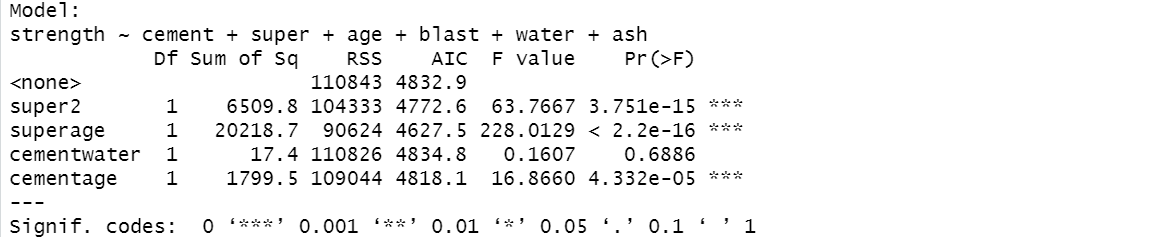
For the R output refer to appendix (Stepwise Regression using AIC). We now apply the General Linear F-Test between the model that we obtained from Stepwise Regression and the Best Subset Regression model with adjusted R-squared.



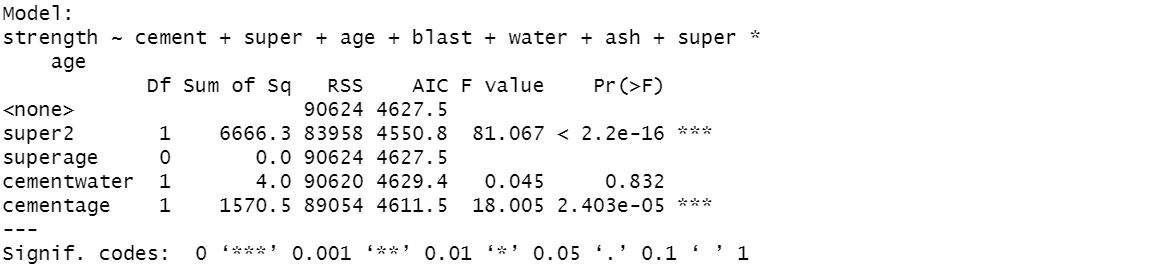
From the table, we realize that we fail to choose the full model and we use our reduced model. As a result, we will choose our Stepwise Regression model.

Addition of Interaction Terms

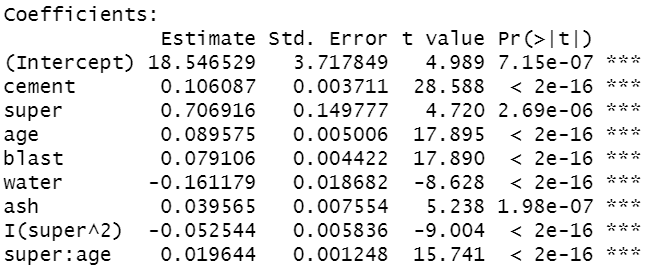
We will now apply the General Linear F-test to find out the best possible interaction terms between our predictors. We would prefer to have more than 6 predictors so we will consider the strengthening effect of superplasticizer with itself, superplasticizer and age, cement and water, cement and age. We will check all these possible interaction terms and see which one is the best to add.



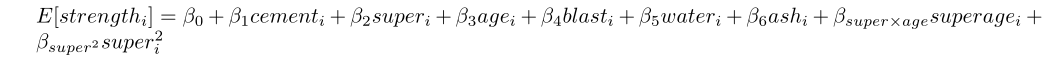
We decide to add the interaction term of superplasticizer and age because this interaction reduces our AIC from 4832.9 to 4627.5. Then, we will rerun the General Linear F-test again.



We decide to add the interaction term of superplasticizer and itself because this interaction reduces our AIC from 4627.5 to 4550.8. We will now check to see if we need to delete any of our variables with the t-test.

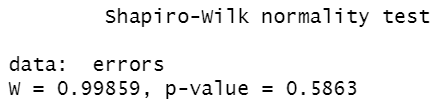
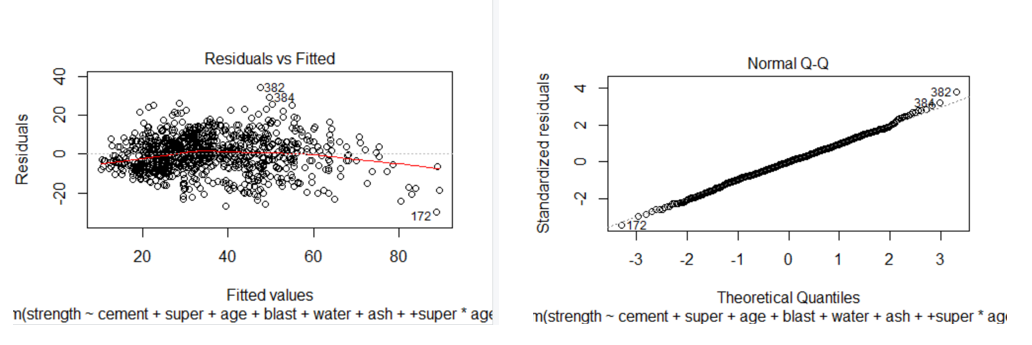


We see that according to the t-test all our variables are significant. Thus, our model after adding the interaction terms is:

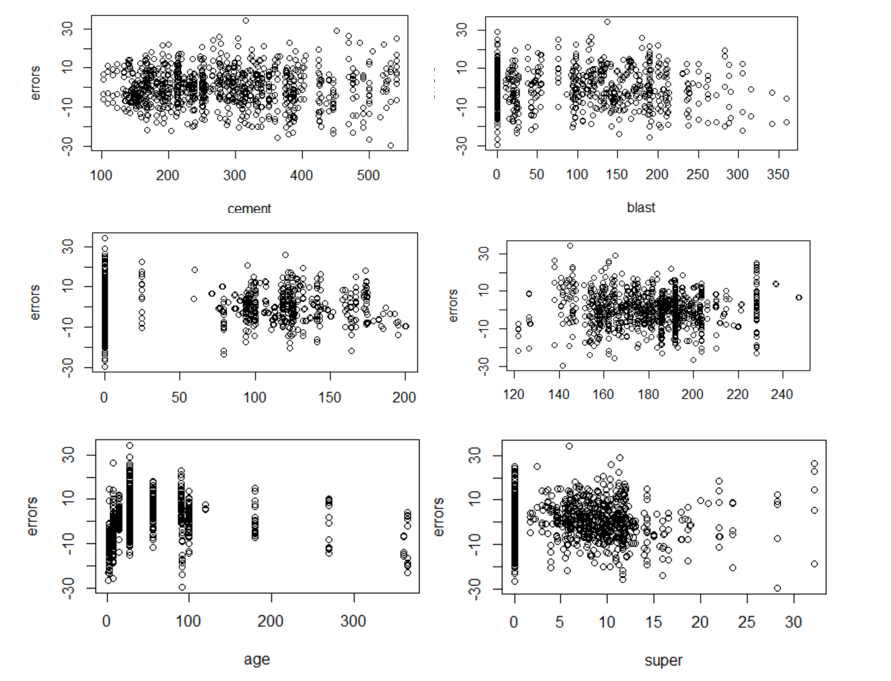


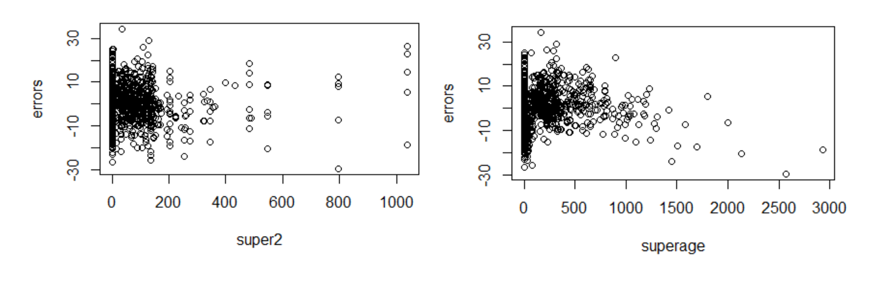
Residual Diagnostics and Transformations

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

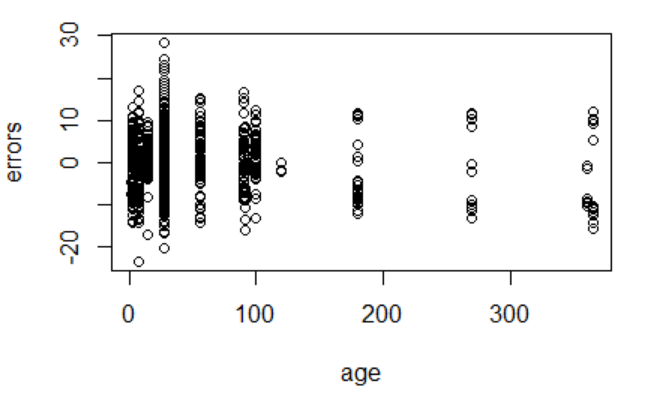


The Residuals vs. Fit plot shows a curved trend of the residuals, which means we fail to meet the linearity condition. Running the Shapiro-Wilk normality test also indicates that our residuals follow a normal distribution. We will now investigate our Residual vs Predictor plots.

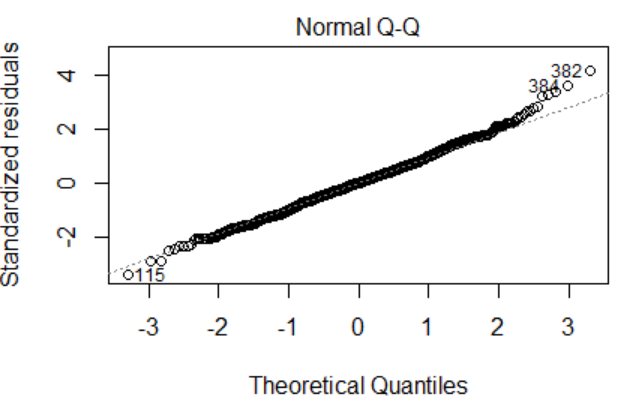
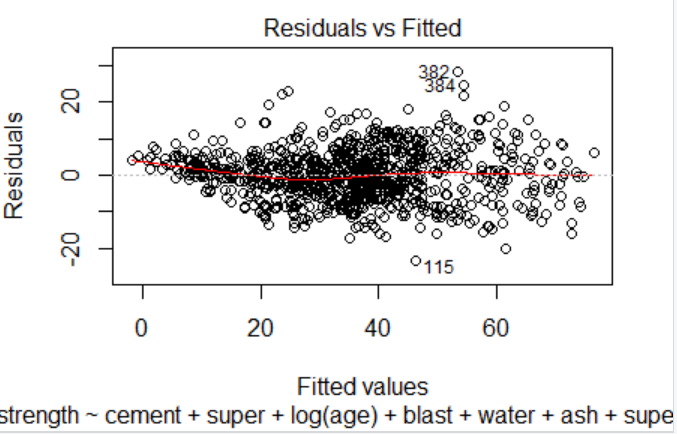


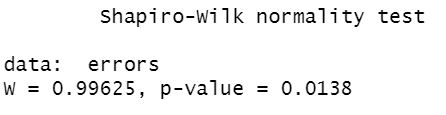


From the above plots we notice that age exhibits a non-linear pattern. We use a log-transformation on age and see if it will reduce how severe the Residual vs. Predictor plot appears.



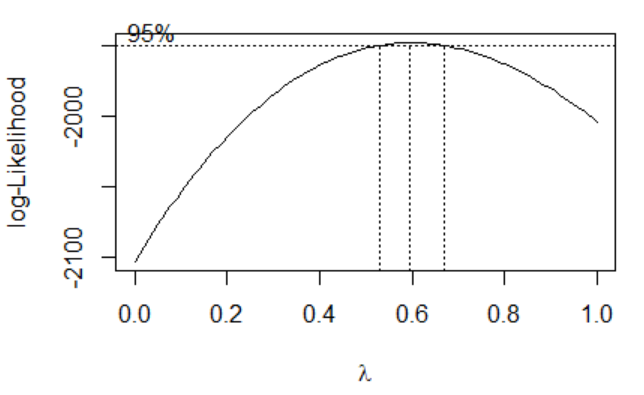
The log-transformation reduces the curvature of the plot. We now perform plot diagnostics with log-transformed age to determine if our linearity condition is met.

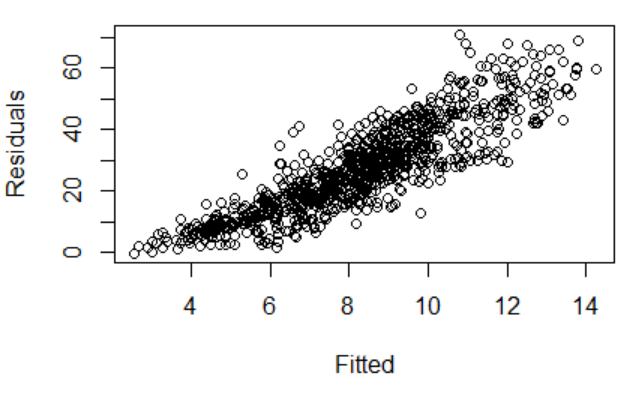




Applying the log transformation results in the linearity condition to be met, however it causes a normality issue as displayed with the Shapiro-Wilk normality test along with a fanning effect at lower fitted values.

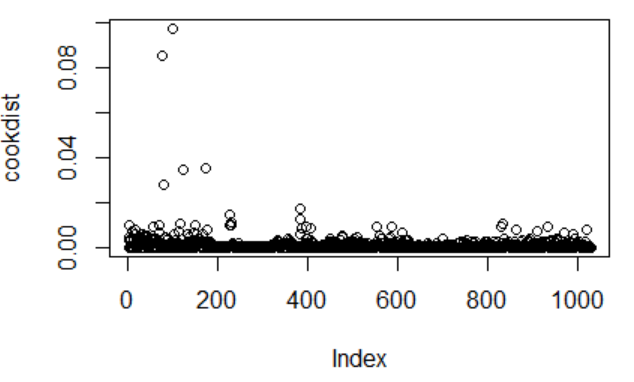
We now attempt to correct our normality problem and fanning effect problem by applying a Box-Cox transformation. As we can see below, our estimated value is power value is 0.6.





As we can see, applying this Box-Cox transformation has caused more problems for our Residual vs Fitted plot. Thus, we will not apply a Box-Cox transformation.

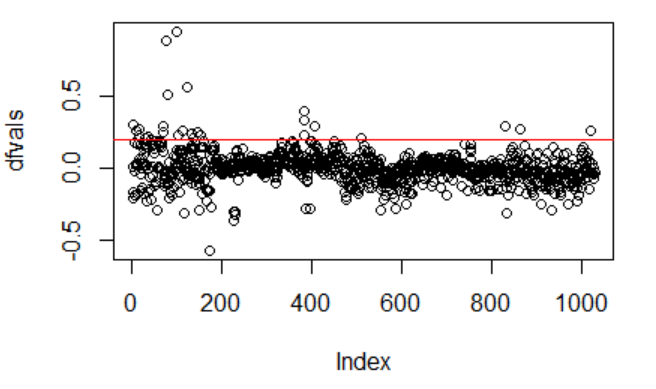
Now, we decide to attempt to fix our normality problem as indicated by our Shapiro-Wilks hypothesis test result which when p< 0.05 indicates that we reject our null hypothesis that our regression is normal. We will use Cook’s Distance to identify our potentially influential points.



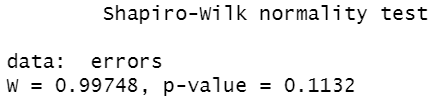
From Cook’s distance, we see that while there are no points greater than 0.5, there are points that are much higher than the other. As a result, we have some influential points. We now check Difference in Fits where our cutoff value for whether a point is influential or not is 0.1981267 from applying the formula below.



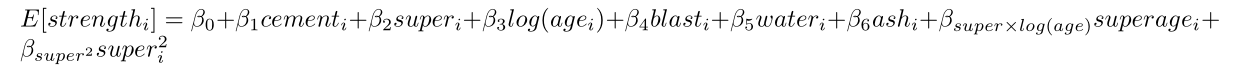
Our plot for our DFFITS values is as below. As we can see, we have some potentially influential points according to our DFFITS criterion.



We will now identify outliers by using our externally studentized residuals. If those outliers are also influential points, we will delete them. We end up deleting the highest externally studentized point with an externally studentized residual values greater than 3. This point is influential according to our DFFITS criterion and above the other points in our Cook’s distance criterion. The point has the index 382.



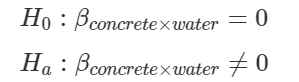
Removing the outlier resulted in our normality condition being met because our p value > 0.05 which means that our regression is normal from the Shapiro-Wilk test. After applying plot diagnostics and transformations our final model now appears as follows:



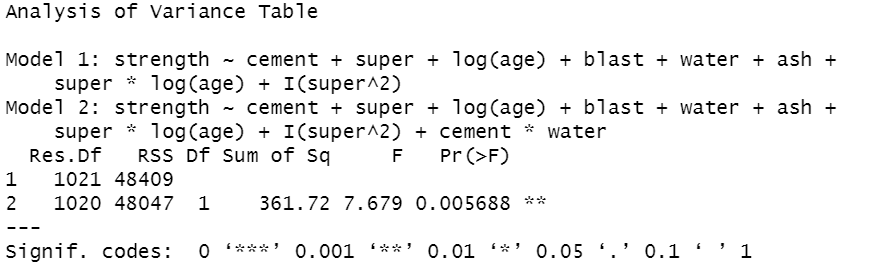
Research Questions

**Does the interaction of cement and water have a significant effect on the compressive strength of concrete?**

We need to determine whether the interaction between cement and water has an effect on the compressive strength of concrete. We check this by using a general linear F-test with the following hypotheses:



After applying the General Linear F-test, we have the following result:



Since our p-value=0.005688 < 0.05, we decide that the interaction between cement and water has an effect on the compressive strength of concrete and that we should add this interaction term to our model.

**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?**

In order to answer this question, we must set blast=0, ash=0, cement=281.1679, water=181.5673, superplasticizer=6.20466, age=45.6621. Now, applying the prediction interval with 95% confidence, we have the following result:



This means that we predict that concrete with no blast furnace slag and ash but with the average values of the other predictors will have a compressive concrete strength of 35.12913 MPa. We are 95% sure that the compressive concrete strength with these values will be between 21.56151 MPa and 48.69675 MPa.

**What compressive strength value do we expect for all concrete that ages to 365 days and average values for the other factors?**

To answer this question, we set the age of the concrete to 365 and set the other predictors to their mean values. We then apply the confidence interval and receive the result below.



From our result, we expect that our compressive concrete strength is 61.31746 MPa. We are 95% confident that our concrete compressive strength will be between 60.19046 and 62.44446 MPa with an age of 365 days and all other values at their average.

Conclusion

To conclude, from applying linear regression and meeting all the LINE conditions, we have a model that demonstrates how the compressive strength of concrete is affected by the amount of cement, superplasticizer, logarithm of age, blast furnace slag, ash, the interaction of superplasticizer and logarithm of age, and the interaction of superplasticizer with itself. With our research questions, we can admit that the interaction between concrete and water influence the strength of concrete. Furthermore, we are 95% sure that when we have no blast furnace slag, no ash, and average values for all other predictors, the strength of a piece of concrete is between 21.56151 MPa and 48.69675 MPa. We also know that the average value of concrete that ages to 365 days with mean values for the other predictors has a strength of 61.31746 MPa.

Since we only have 1030 data points, our lack of data points causes a fanning effect because there aren’t too many concrete samples with low strength. This can be resolved by taking a larger sample of concrete. We could gain more insight in our model if we had predictors such as the type of superplasticizer used or the highest percentage material for the cement mix. Our model is too general and may not work for all types of concrete which is why we need more information.

Appendix

library(leaps)

library(MASS)

data=read.table("Concrete\_Data.txt", sep="\t", header=TRUE)

names(data)<-c("Cement", "Blast Furnace", "Ash", "Water", "SuperPlasticizer", "CoarseAgg",

"FineAgg", "Age", "Strength")

cement=data$Cement

blast=data$`Blast Furnace`

ash=data$Ash

water=data$Water

super=data$SuperPlasticizer

coarse=data$CoarseAgg

fine=data$FineAgg

age=data$Age

strength=data$Strength

# Scatterplot Matrix of predictors

pairs(~strength + age + fine + coarse + super + water + ash + blast + cement)

# Check the correlation matrix for any high correlation values

cor(data)

# Apply stepwise regression

basemod=lm(strength~1)

stepmod=lm(strength ~ age + fine + coarse + super + water + ash + blast + cement)

step(basemod, scope = list(lower=basemod, upper=stepmod))

stepwise=lm(strength ~ cement + super + age + blast + water +

ash)

fittedstep=fitted(stepwise)

errors=strength-fittedstep

summary(stepwise)

# Check for interaction terms

super2=super^2

superage=super\*age

cementwater=cement\*water

cementage=cement\*age

estimate=fitted(stepwise)

errors=strength-estimate

plot(stepwise)

plot(estimate, errors, ylab="Residuals", xlab="Fitted")

plot(cement, errors)

plot(blast, errors)

plot(ash, errors)

plot(water, errors)

plot(age, errors)

plot(super, errors)

plot(super2, errors)

plot(superage, errors)

shapiro.test(errors)

qqnorm(errors)

qqline(errors)

# F-tests to see which terms to add

add1(stepwise, ~.+super2+superage+cementwater+cementage, test="F")

# From the general linear F Test we add super^2 and apply again

stepwise=lm(strength ~ cement + super + age + blast + water +

ash+ super\*age)

# F-tests to see which terms to add

add1(stepwise, ~.+super2+superage+cementwater+cementage, test="F")

stepwise=lm(strength ~ cement + super + age + blast + water +

ash+ super\*age + I(super^2))

estimate=fitted(stepwise)

errors=strength-estimate

plot(stepwise)

plot(estimate, errors, ylab="Residuals", xlab="Fitted")

plot(cement, errors)

plot(blast, errors)

plot(ash, errors)

plot(water, errors)

plot(age, errors)

plot(super, errors)

plot(super2, errors)

plot(superage, errors)

shapiro.test(errors)

qqnorm(errors)

qqline(errors)

# Log transform age because non-linear

stepwise=lm(strength ~ cement + super + log(age) + blast + water +

ash+ super\*log(age) + I(super^2))

estimate=fitted(stepwise)

errors=strength-estimate

plot(stepwise)

plot(estimate, errors, ylab="Residuals", xlab="Fitted")

plot(cement, errors)

plot(blast, errors)

plot(ash, errors)

plot(water, errors)

plot(age, errors)

plot(super, errors)

plot(super2, errors)

plot(superage, errors)

shapiro.test(errors)

qqnorm(errors)

qqline(errors)

# Try Box-Cox transform for fanning effect

trans=boxcox(strength ~ cement + super + log(age) + blast + water +

ash+ super\*log(age)+ I(super^2), data=data, lambda=seq(0, 1, length=10))

str=strength^(0.6)

stepwise2=lm(str ~ cement + super + log(age) + blast + water +

ash+ super\*log(age)+ I(super^2) )

estimate=fitted(stepwise2)

errors=strength-estimate

plot(estimate, errors, xlab="Fitted", ylab="Residuals")

plot(stepwise2)

qqnorm(errors)

qqline(errors)

# Check Cook's Distance for influential points

cookdist=(cooks.distance(stepwise))

plot(cookdist)

# DFFITS Line for influential points

dfvals=(dffits(stepwise))

cutoff=2\*sqrt((10/(n-10-1)))

plot(dfvals)

abline(h=0.1981267, col='red')

# Find externally studentized residual

r\_ext=sort(rstudent(stepwise))

n<-length(strength)

r\_ext[n]

# Delete externally studentized residual

data<-data[-382,]

rownames(data) <- 1:nrow(data)

cement=data$Cement

blast=data$`Blast Furnace`

ash=data$Ash

water=data$Water

super=data$SuperPlasticizer

coarse=data$CoarseAgg

fine=data$FineAgg

age=data$Age

strength=data$Strength

stepwise=lm(strength ~ cement + super + log(age) + blast + water +

ash+ super\*log(age)+ I(super^2) )

# Check whether interaction term cement\*water has effect on concrete strength

stepwise2=lm(strength ~ cement + super + log(age) + blast + water +

ash+ super\*log(age)+ I(super^2) + cement\*water )

anova(stepwise, stepwise2)

fit=fitted(stepwise)

errors=strength-fit

shapiro.test(errors)

# Prediction Interval for strength

p1=data.frame(blast=0, ash=0, cement=mean(cement), water=mean(water), super=mean(super),

age=mean(age))

predictstr=predict(stepwise, p1, se.fit = TRUE, interval = "prediction", level = 0.95)

print(predictstr$fit)

# Confidence interval for strength

p2=data.frame(blast=mean(blast), ash=mean(ash), cement=mean(cement), water=mean(water), super=mean(super), age=365)

confstr=predict(stepwise, p2, se.fit = TRUE, interval = "confidence", level = 0.95)

print(confstr$fit)