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.

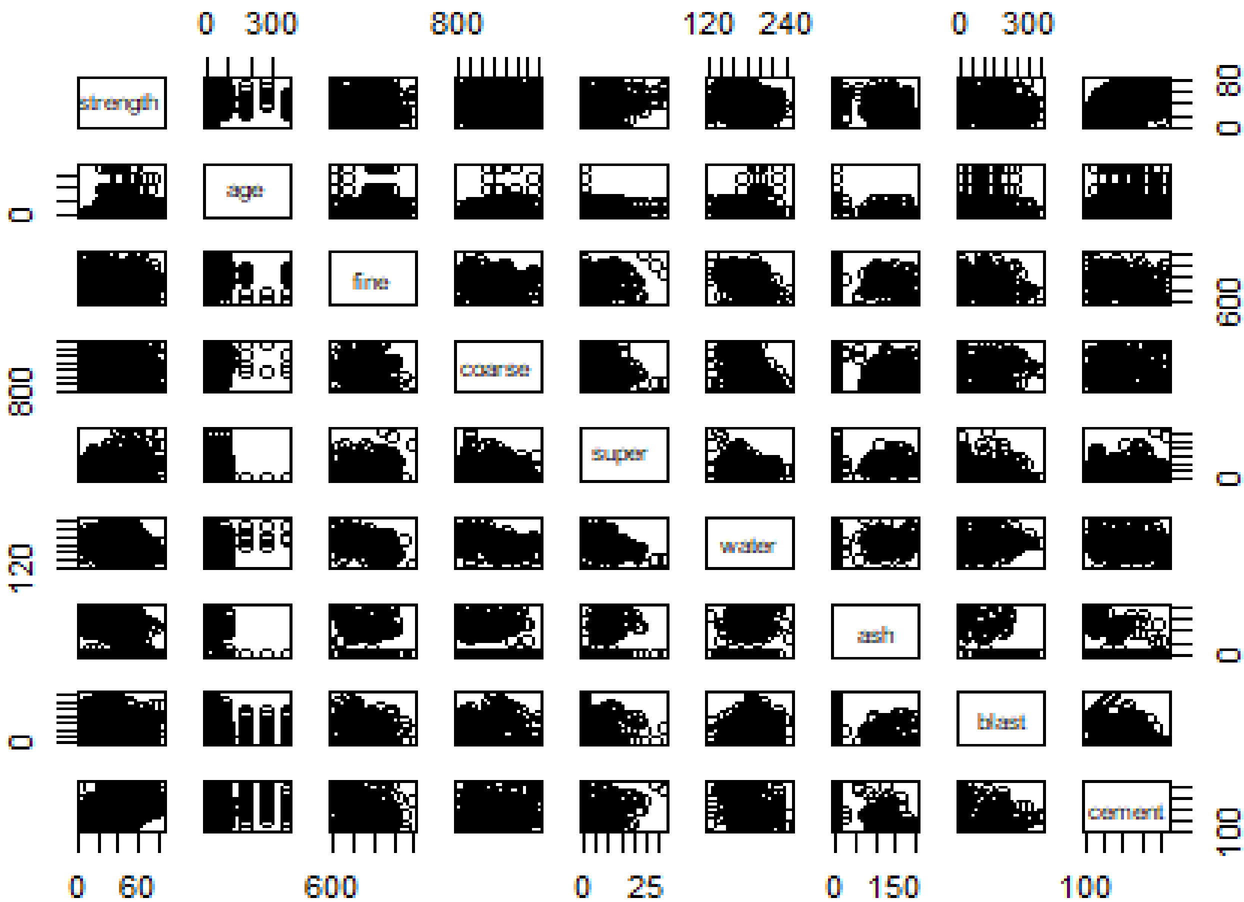
Q2. We calculate a 95% prediction interval for a new value of strength given that there is no blast furnace slag nor ash, and average values for cement, water, superplasticizer, and age.

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.

```
  (Intercept) Cement Blast Furnace   Ash Water SuperPlasticizer CoarseAgg FineAgg   Age
1        TRUE   TRUE         FALSE FALSE FALSE            FALSE     FALSE   FALSE FALSE
2        TRUE   TRUE         FALSE FALSE FALSE             TRUE     FALSE   FALSE FALSE
3        TRUE   TRUE         FALSE FALSE FALSE             TRUE     FALSE   FALSE  TRUE
4        TRUE   TRUE          TRUE FALSE  TRUE            FALSE     FALSE   FALSE  TRUE
5        TRUE   TRUE          TRUE  TRUE  TRUE            FALSE     FALSE   FALSE  TRUE
6        TRUE   TRUE          TRUE  TRUE  TRUE             TRUE     FALSE   FALSE  TRUE
7        TRUE   TRUE          TRUE  TRUE  TRUE             TRUE      TRUE   FALSE  TRUE
8        TRUE   TRUE          TRUE  TRUE  TRUE             TRUE      TRUE    TRUE  TRUE

[1] 0.2471049 0.3499102 0.4802385 0.5560241 0.6090761 0.6117586 0.6115369 0.6125073
```

$$E[strength_i] = \beta_0 + \beta_1 cement_i + \beta_2 super_i + \beta_3 age_i + \beta_4 blast_i + \beta_5 water_i + \beta_6 ash_i + \beta_7 coarse_i + \beta_8 fine_i$$

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:

$$\boldsymbol{E[strength_i] = \beta_0 + \beta_1 cement_i + \beta_2 super_i + \beta_3 age_i + \beta_4 blast_i + \beta_5 water_i + \beta_6 ash_i}$$

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.

```
Analysis of Variance Table

Model 1: strength ~ cement + super + age + blast + water + ash
Model 2: strength ~ cement + super + age + blast + water + ash + fine +
    coarse
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1023 110843
2   1021 110413  2    430.05 1.9883 0.1375
```

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.

```
Model:
strength ~ cement + super + age + blast + water + ash
           Df Sum of Sq    RSS    AIC  F value    Pr(>F)    
<none>                  110843 4832.9                       
super2      1    6509.8 104333 4772.6  63.7667 3.751e-15 ***
superage    1   20218.7  90624 4627.5 228.0129 < 2.2e-16 ***
cementwater 1      17.4 110826 4834.8   0.1607    0.6886    
cementage   1    1799.5 109044 4818.1  16.8660 4.332e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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.

```
Model:
strength ~ cement + super + age + blast + water + ash + super * 
    age
           Df Sum of Sq   RSS    AIC F value    Pr(>F)    
<none>                  90624 4627.5                      
super2      1    6666.3 83958 4550.8  81.067 < 2.2e-16 ***
superage    0       0.0 90624 4627.5                      
cementwater 1       4.0 90620 4629.4   0.045     0.832    
cementage   1    1570.5 89054 4611.5  18.005 2.403e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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.

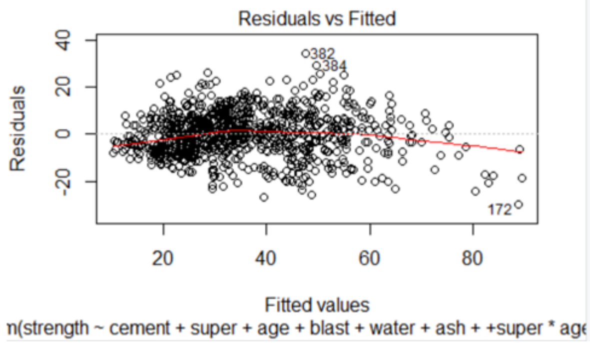
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.546529   3.717849   4.989 7.15e-07 ***
cement       0.106087   0.003711  28.588  < 2e-16 ***
super        0.706916   0.149777   4.720 2.69e-06 ***
age          0.089575   0.005006  17.895  < 2e-16 ***
blast        0.079106   0.004422  17.890  < 2e-16 ***
water       -0.161179   0.018682  -8.628  < 2e-16 ***
ash          0.039565   0.007554   5.238 1.98e-07 ***
I(super^2)  -0.052544   0.005836  -9.004  < 2e-16 ***
super:age    0.019644   0.001248  15.741  < 2e-16 ***
```

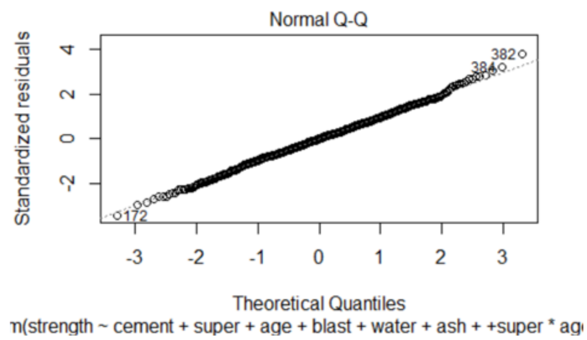
We see that according to the t-test all our variables are significant. Thus, our model after adding the interaction terms is:

$$E[strength_i] = \beta_0 + \beta_1 cement_i + \beta_2 super_i + \beta_3 age_i + \beta_4 blast_i + \beta_5 water_i + \beta_6 ash_i + \beta_{super \times age} superage_i + \beta_{super^2} super_i^2$$

Residual Diagnostics and Transformations

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



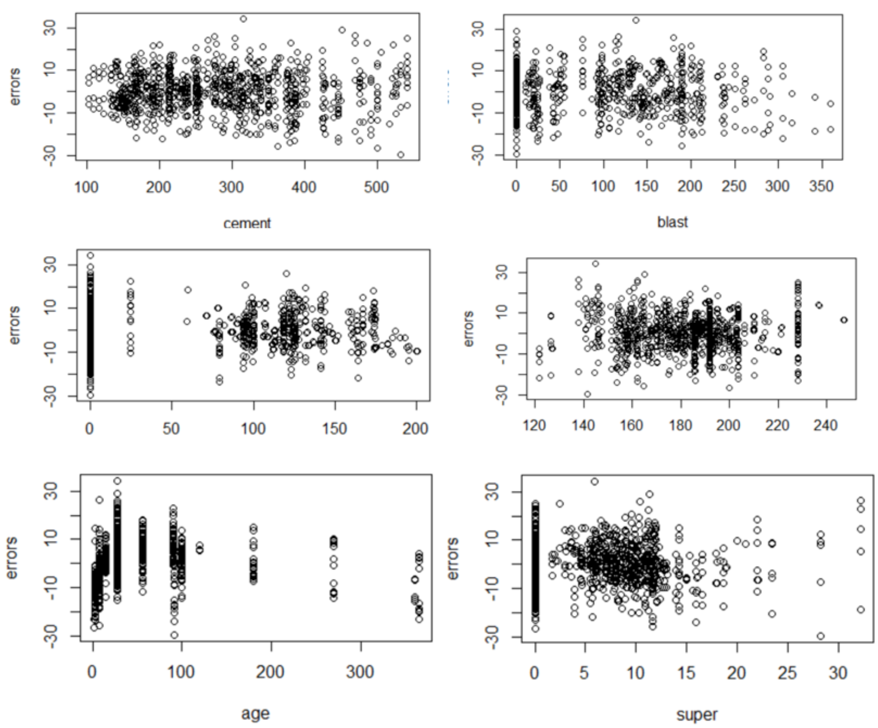


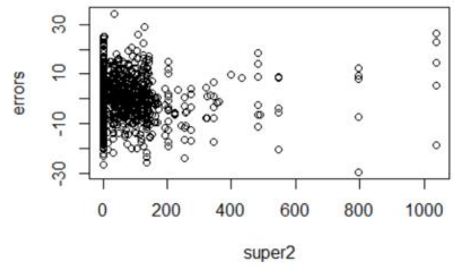
The Residuals vs. Fit plot shows a curved trend of the residuals, which means we fail to meet the

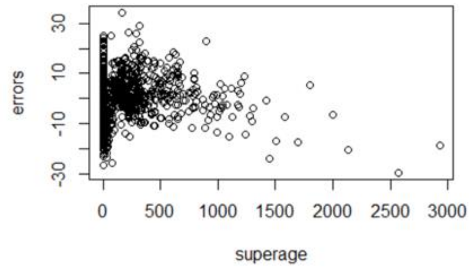
```
        Shapiro-Wilk normality test

data:  errors
W = 0.99859, p-value = 0.5863
```

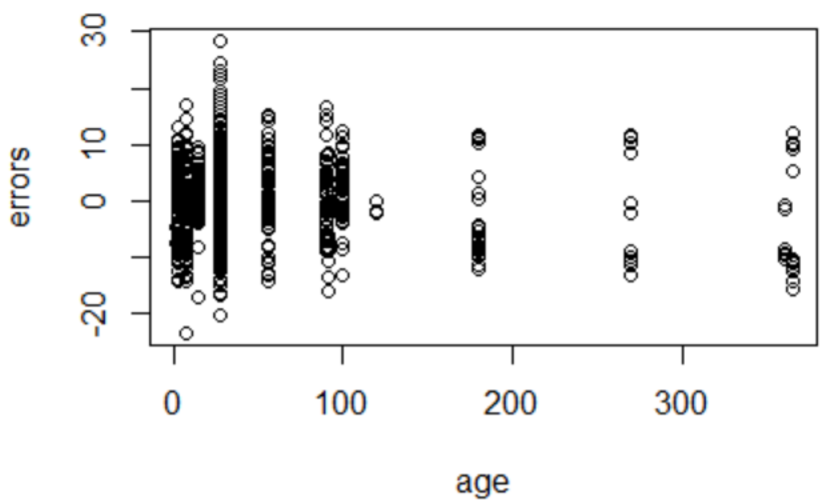
linearity condition. Running the Shapiro-Wilk normality test also indicates that our residuals follow a normal distribution. We will now investigate our Residual vs Predictor plots.



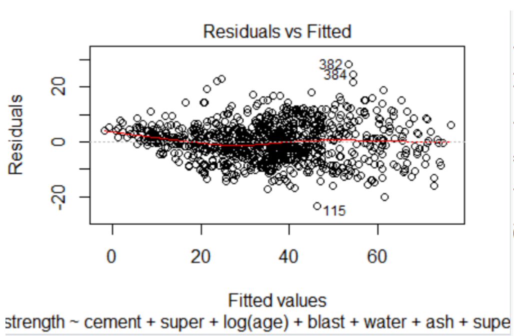


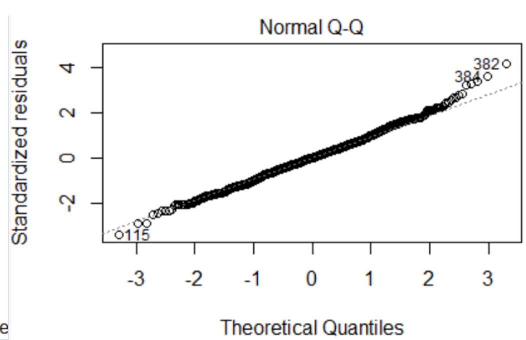


From the above plots we notice that age exhibits a non-linear pattern. We use a log-transformation on age and see if it will reduce how severe the Residual vs. Predictor plot appears.



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.



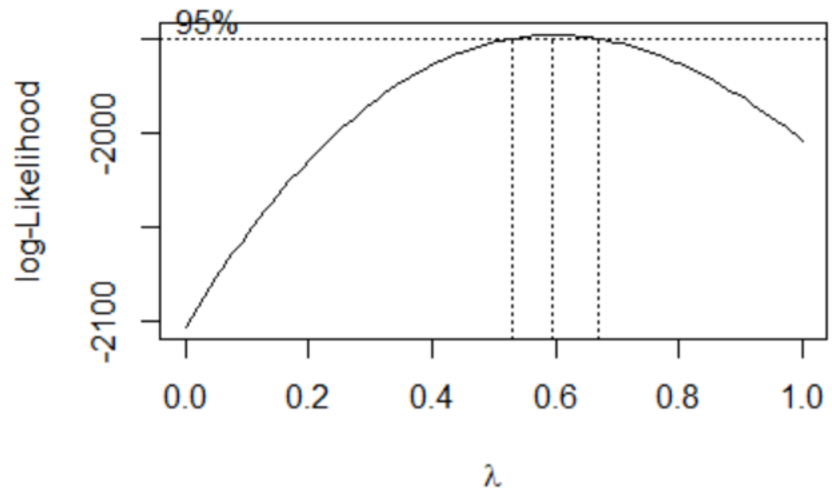


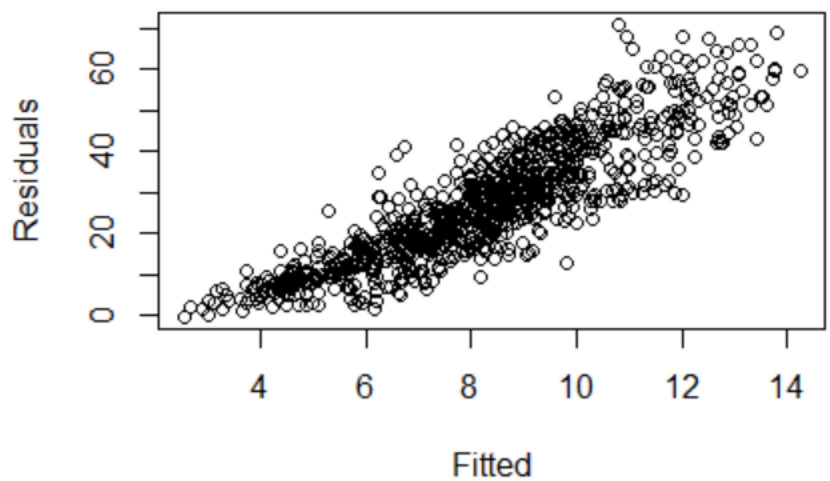

```
        Shapiro-Wilk normality test

data:  errors
W = 0.99625, p-value = 0.0138
```

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.

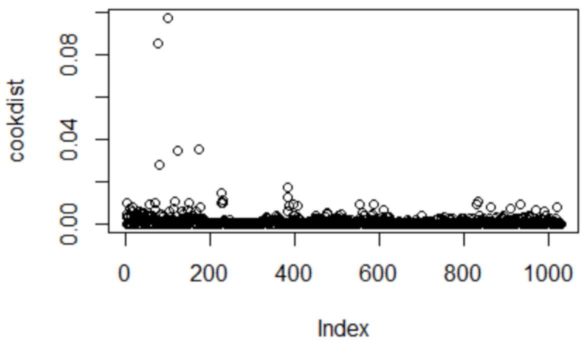
We now attempt to correct our normality problem and fanning effect problem by applying a Box-Cox transformation. As we can see below, our estimated value is power value is 0.6.





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.

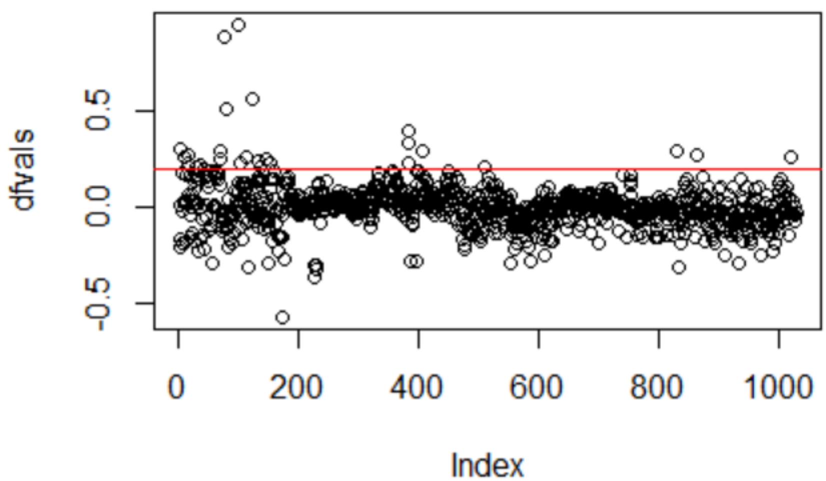
Now, we decide to attempt to fix our normality problem as indicated by our Shapiro-Wilks hypothesis test result which when $p< 0.05$ indicates that we reject our null hypothesis that our regression is normal. We will use Cook's Distance to identify our potentially influential points.



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.

$$2\sqrt{\frac{p+1}{n-p-1}}$$

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.

```
        Shapiro-Wilk normality test

data:  errors
W = 0.99748, p-value = 0.1132
```

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:

$$E[strength_i] = \beta_0 + \beta_1 cement_i + \beta_2 super_i + \beta_3 log(age_i) + \beta_4 blast_i + \beta_5 water_i + \beta_6 ash_i + \beta_{super \times log(age)} superage_i + \beta_{super^2} super_i^2$$

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:

$$H_0 : \beta_{concrete \times water} = 0$$

$$H_a : \beta_{concrete \times water} \neq 0$$

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

```
Analysis of Variance Table

Model 1: strength ~ cement + super + log(age) + blast + water + ash +
    super * log(age) + I(super^2)
Model 2: strength ~ cement + super + log(age) + blast + water + ash +
    super * log(age) + I(super^2) + cement * water
  Res.Df   RSS Df Sum of Sq     F   Pr(>F)
1   1021 48409
2   1020 48047  1    361.72 7.679 0.005688 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

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:

```
       fit      lwr      upr
1 35.12913 21.56151 48.69675
```

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.

```
       fit      lwr      upr
1 61.31746 60.19046 62.44446
```

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)

Correlation Matrix:

	Cement	Blast Furnace	Ash	Water	SuperPlasticizer	CoarseAgg	FineAgg	Age	Strength
Cement	1.00000000	-0.27521591	-0.397467341	-0.08158675	0.09238617	-0.109348994	-0.22271785	0.08194602	0.4978319
Blast Furnace	-0.27521591	1.00000000	-0.323579901	0.10725203	0.04327042	-0.283998612	-0.28160267	-0.04424602	0.1348293
Ash	-0.39746734	-0.32357990	1.000000000	-0.25698402	0.37750315	-0.009960828	0.07910849	-0.15437052	-0.1057549
Water	-0.08158675	0.10725203	-0.256984023	1.00000000	-0.65753291	-0.182293602	-0.45066117	0.27761822	-0.2896334
SuperPlasticizer	0.09238617	0.04327042	0.377503146	-0.65753291	1.00000000	-0.265999148	0.22269123	-0.19270003	0.3660788
CoarseAgg	-0.10934899	-0.28399861	-0.009960828	-0.18229360	-0.26599915	1.000000000	-0.17848096	-0.00301588	-0.1649346
FineAgg	-0.22271785	-0.28160267	0.079108491	-0.45066117	0.22269123	-0.178480957	1.00000000	-0.15609470	-0.1672412
Age	0.08194602	-0.04424602	-0.154370516	0.27761822	-0.19270003	-0.003015880	-0.15609470	1.00000000	0.3288730
Strength	0.49783192	0.13482926	-0.105754916	-0.28963338	0.36607883	-0.164934614	-0.16724125	0.32887300	1.0000000

Stepwise Regression using AIC

```
Start:  AIC=5801.45
strength ~ 1

         Df Sum of Sq    RSS    AIC
+ cement  1     71173 216003 5510.1
+ super   1     38485 248690 5655.2
+ age     1     31060 256115 5685.5
+ water   1     24090 263085 5713.2
+ fine    1      8032 279143 5774.2
+ coarse  1      7812 279363 5775.0
+ blast   1      5221 281955 5784.6
+ ash     1      3212 283963 5791.9
<none>                287175 5801.4

Step:  AIC=5510.1
strength ~ cement

         Df Sum of Sq    RSS    AIC
+ super   1     29676 186327 5359.9
+ age     1     23993 192009 5390.8
+ blast   1     22961 193042 5396.3
+ water   1     17927 198076 5422.9
+ coarse  1      3549 212454 5495.0
+ ash     1      2894 213109 5498.2
+ fine    1       960 215043 5507.5
<none>                216003 5510.1
- cement  1     71173 287175 5801.4

Step:  AIC=5359.88
strength ~ cement + super

         Df Sum of Sq    RSS    AIC
+ age     1     37499 148827 5130.4
+ blast   1     19467 166860 5248.2
+ fine    1      5867 180459 5328.9
+ water   1       777 185550 5357.6
+ ash     1       743 185583 5357.8
<none>                186327 5359.9
+ coarse  1       242 186085 5360.5
- super   1     29676 216003 5510.1
- cement  1     62363 248690 5655.2
```

```
Step:  AIC=5130.43
strength ~ cement + super + age

         Df Sum of Sq    RSS    AIC
+ blast   1     19922 128905 4984.4
+ water   1      4858 143969 5098.2
+ fine    1      3390 145437 5108.7
+ ash     1       325 148502 5130.2
<none>                148827 5130.4
+ coarse  1        37 148790 5132.2
- age     1     37499 186327 5359.9
- super   1     43182 192009 5390.8
- cement  1     52281 201109 5438.5

Step:  AIC=4984.4
strength ~ cement + super + age + blast

         Df Sum of Sq    RSS    AIC
+ water   1      9524 119381 4907.3
+ ash     1      6519 122386 4933.0
+ coarse  1      1735 127170 4972.4
<none>                128905 4984.4
+ fine    1         4 128901 4986.4
- blast   1     19922 148827 5130.4
- age     1     37955 166860 5248.2
- super   1     39056 167961 5255.0
- cement  1     67040 195945 5413.7

Step:  AIC=4907.34
strength ~ cement + super + age + blast + water

         Df Sum of Sq    RSS    AIC
+ ash     1      8537 110843 4832.9
+ fine    1      1895 117486 4892.9
<none>                119381 4907.3
+ coarse  1        24 119357 4909.1
- super   1      7623 127003 4969.1
- water   1      9524 128905 4984.4
- blast   1     24588 143969 5098.2
- age     1     44581 163962 5232.2
- cement  1     67229 186609 5365.4
```

```
Step:  AIC=4832.91
strength ~ cement + super + age + blast + water + ash

         Df Sum of Sq    RSS    AIC
<none>                110843 4832.9
+ coarse  1        45 110798 4834.5
+ fine    1        29 110814 4834.6
- super   1       875 111718 4839.0
- ash     1      8537 119381 4907.3
- water   1     11543 122386 4933.0
- blast   1     32750 143593 5097.5
- age     1     47728 158571 5199.7
- cement  1     66774 177617 5316.6

Call:
lm(formula = strength ~ cement + super + age + blast + water +
    ash)

Coefficients:
(Intercept)       cement        super          age        blast        water          ash
   28.99298      0.10541      0.24031      0.11349      0.08647     -0.21809      0.06866
```