concrete data

Yogesh Kumar

Introduction

Importance of Concrete

Concrete is of utmost importance in most types of construction as it provides **strength** to the underlying structure over a long period of time thereby showcasing its **durability** aspect. Now, in concrete, compressive strength is the property of great use because as a construction material, concrete is usually employed to resist compressive stresses. Even in situations when tensile stresses are predominant, compressive strength can be used to estimate those properties.

Objective

We want to identify the dependency of concrete's compressive strength on the concrete's mixture, explore the degree of dependency through known parameters and formulate a suitable linear regression model

Data Preparation

Importing the libraries

```
library(readxl)
library(ggfortify)
library(ggplot2)
library(psych)
library(tidyverse)
library(GGally)
library(ggthemes)
library(ggpubr)
library(patchwork)
```

Importing the data

Let's have a glimpse of the data

```
a = read_excel("C:\\Users\\patel\\Downloads\\Concrete_Data.xls")
glimpse(a)
```

```
## Rows: 1,030
## Columns: 9
## $ `Cement (component 1)(kg in a m^3 mixture)`             <dbl> 540.0, 540.0, ~
## $ `Blast Furnace Slag (component 2)(kg in a m^3 mixture)` <dbl> 0.0, 0.0, 142.~
## $ `Fly Ash (component 3)(kg in a m^3 mixture)`            <dbl> 0, 0, 0, 0, 0,~
## $ `Water  (component 4)(kg in a m^3 mixture)`             <dbl> 162, 162, 228,~
## $ `Superplasticizer (component 5)(kg in a m^3 mixture)`   <dbl> 2.5, 2.5, 0.0,~
## $ `Coarse Aggregate  (component 6)(kg in a m^3 mixture)`  <dbl> 1040.0, 1055.0~
```

```
## $ `Fine Aggregate (component 7)(kg in a m^3 mixture)`     <dbl> 676.0, 676.0, ~
## $ `Age (day)`                                             <dbl> 28, 28, 270, 3~
## $ `Concrete compressive strength(MPa, megapascals)`       <dbl> 79.986111, 61.~
```

Removing Duplicates

First we will check if our data contain duplicate rows of data and if it contains some then we will remove them as these data points do not conveying any new information to the model.

```
dim(a[duplicated(a),])
```

```
## [1] 25  9
```

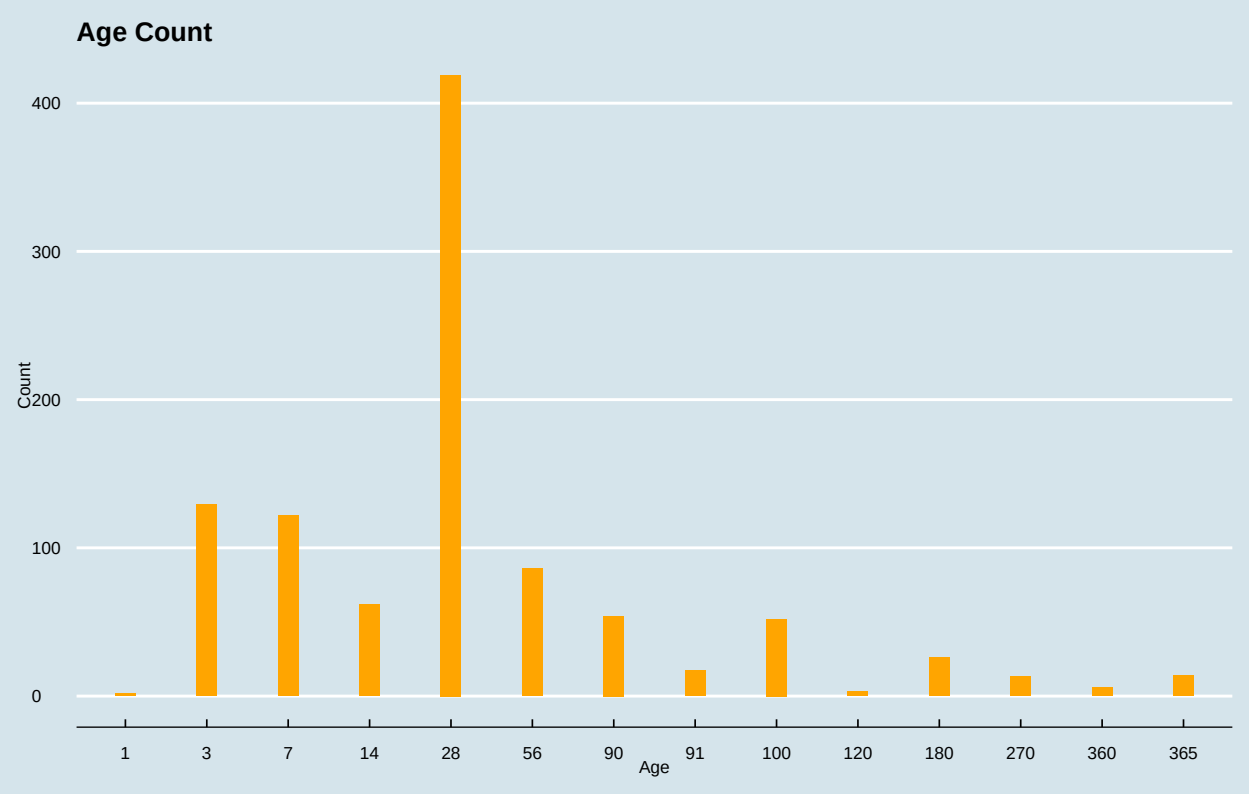
```
a<-a%>%
  unique(fromLast=TRUE)
colnames(a) = c("cement","slag","ash","water","superplastic",
                "coarseAgg","fineAgg","age","strength")

a = as.data.frame(a)
```

Data Exploration

```
p<-ggplot(data=a,aes(x=strength),color='blue')+geom_density(
  aes(y=..density..),fill='orange')+theme_economist()
```

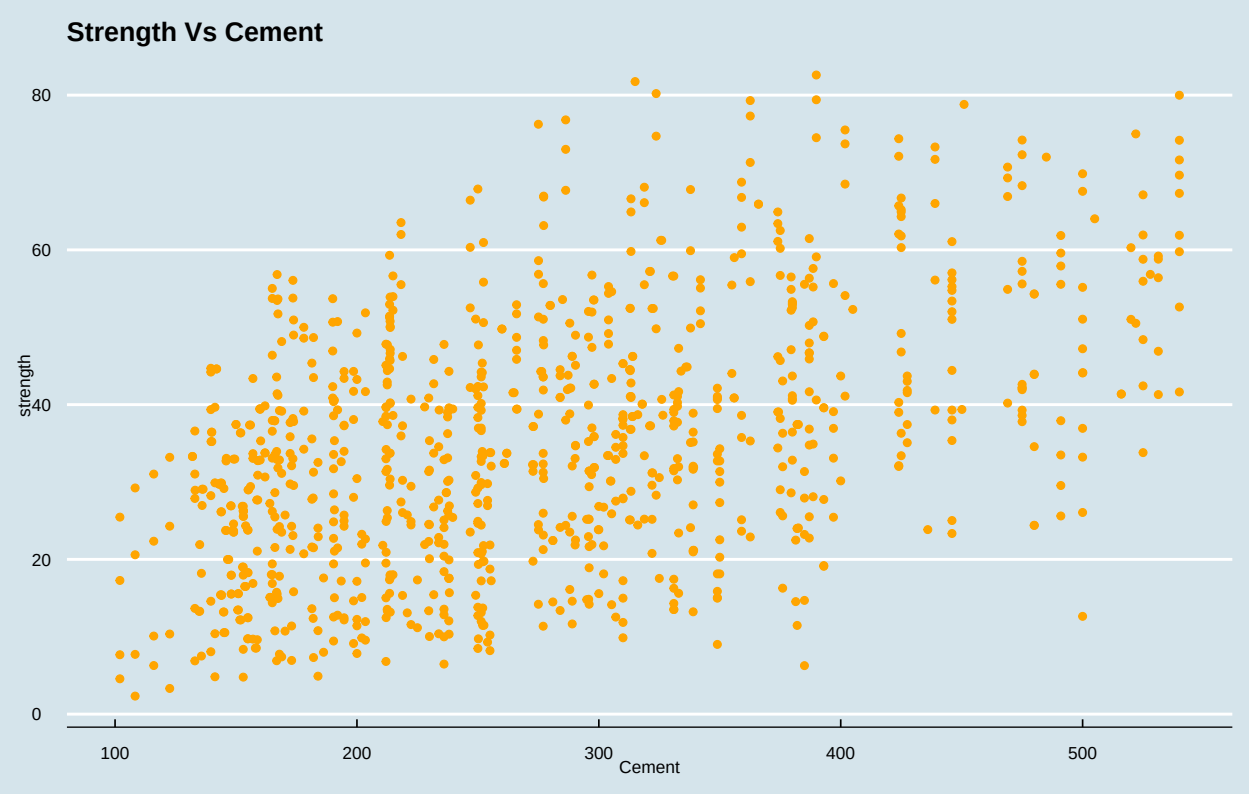
So, here we can see that our target variable is approximately normal distributed and it is a little skewed towards right.

```
p<-ggplot(data=a)
p+geom_bar(aes(x=as.factor(age)),width=0.25,fill='orange')+
  labs(x='Age',y='Count')+
  ggtitle("Age Count ")+
  theme_economist()
```



So, at first glance, age seems like a continuous variable but it is only taking certain number of values and the value "28" have ocurred more number of times.This may be due to the fact that the strength gained by concrete is maximum after 28 days and then it starts declining.

```
p+geom_point(aes(x=cement,y=strength),size=2,color='orange')+
  labs(x="Cement",y="strength")+
  ggtitle("Strength Vs Cement ")+
  theme_economist()
```



From the above plot it is clear that cement is linearly related to concrete's strength. So, it may be an important factor.

Analysis

Regression on all variables

We start with a linear regression model involving all the variables

```
lr = lm(strength~.,a)
summary(lr)
```

```
##
## Call:
## lm(formula = strength ~ ., data = a)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.139  -6.301   0.602   6.646  34.991
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.748084  26.419314  -0.672 0.501877
## cement         0.117221   0.008495  13.799  < 2e-16 ***
## slag           0.099445   0.010158   9.790  < 2e-16 ***
## ash            0.085632   0.012478   6.862 1.19e-11 ***
## water         -0.152630   0.039801  -3.835 0.000134 ***
## superplastic   0.283380   0.092929   3.049 0.002353 **
## coarseAgg      0.015621   0.009321   1.676 0.094092 .
```

```
## fineAgg        0.018291   0.010675   1.713 0.086940 .
## age            0.112181   0.005393  20.800  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 996 degrees of freedom
## Multiple R-squared:  0.6038, Adjusted R-squared:  0.6006
## F-statistic: 189.8 on 8 and 996 DF,  p-value: < 2.2e-16
```

From the summary of the regression model, we see that the variables corresponding to columns `coarseAgg` and `fineAgg` are not statistically significant as their corresponding p value is more than 0.05, so they fail to reject the null hypothesis in the t test

Removing the non-significant variables

```
a2 = a[,-which(summary(lr)$coefficients[-1,4]>0.05)]
lr2 = lm(strength~.,a2)
summary(lr2)
```

```
##
## Call:
## lm(formula = strength ~ ., data = a2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.829  -6.519   0.555   6.603  35.207
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  28.504229   4.170758   6.834 1.43e-11 ***
## cement        0.104254   0.004227  24.662  < 2e-16 ***
## slag          0.083801   0.004980  16.827  < 2e-16 ***
## ash           0.068410   0.007697   8.887  < 2e-16 ***
## water        -0.212540   0.020953 -10.144  < 2e-16 ***
## superplastic  0.240087   0.084439   2.843  0.00456 **
## age           0.111464   0.005369  20.759  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.3 on 998 degrees of freedom
## Multiple R-squared:  0.6026, Adjusted R-squared:  0.6002
## F-statistic: 252.2 on 6 and 998 DF,  p-value: < 2.2e-16
```

Even after removing the non-significant columns `fineAgg` and `coarseAgg`, we are unaware about any interaction between two independent variables or a possibility of concrete's compressive strength being dependent on higher orders of a particular variable.

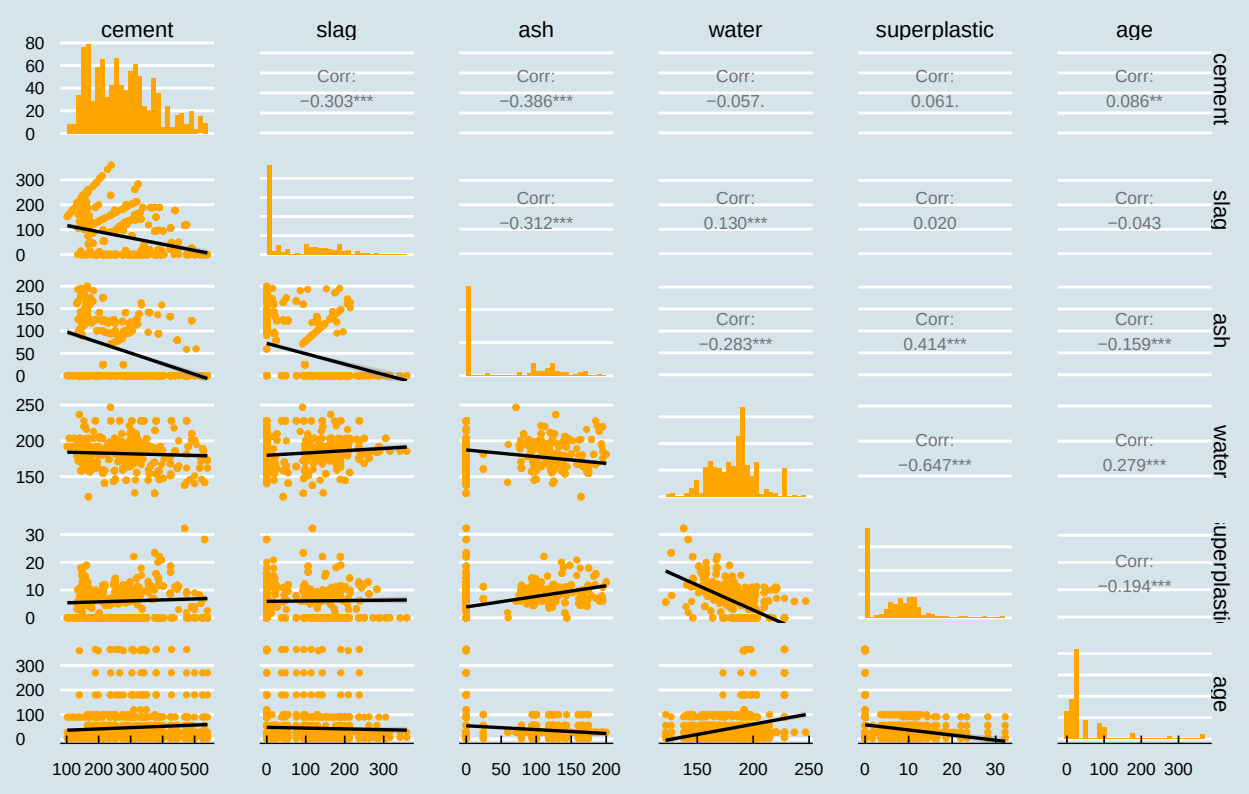
Descriptive statistics for identifying interaction and higher order terms

```
a2 %>%
  gather(-strength, key = "var", value = "value") %>%
  ggplot(aes(x = value, y = strength)) +
  ggtitle("Plot of strength against each significant regressor \n") +
    geom_point(color = "orange") +
```

```
  facet_wrap(~ var, scales = "free") +
theme(panel.spacing.y=unit(1,"lines"),plot.title=element_text(vjust=7))+
  theme_economist()+
theme(panel.spacing=unit(4,"lines"))
```

We see from the above plot that strength is almost `linearly varying` with cement, water, superplastic, slag and ash but from the strength vs age plot, we can immediately identify that there is a `quadratic component` to their bivariate relationship.

```
ggpairs(a2[,-7],
    lower=list(continuous=wrap("smooth",colour="orange")),
    diag=list(continuous=wrap("barDiag", fill="orange",bins=30))) +
    theme_economist() +
  theme(panel.spacing.y=unit(1,"lines")) +
  theme(panel.spacing=unit(2,"lines"))
```



We find that two pairs, one being, `water` and `superplastic` and second being, `ash` and `superplastic` are moderately high negative correlated so an interaction term between the variables being represented by the two columns is definitely deserving in the regression model. Also, none of the other independent variables seem to have moderate to high correlation with any of the other independent variables thereby extinguishing the possibility of more interaction term.

Adding interaction terms in the model

```
lr4 = lm(strength~.+I(age^(2))+
           superplastic:water+ash:superplastic,a2)
summary(lr4)
```

```
##
## Call:
## lm(formula = strength ~ . + I(age^(2)) + superplastic:water +
##     ash:superplastic, data = a2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.322  -4.708  -0.309   4.645  33.679
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          3.517e+01  4.158e+00   8.457  < 2e-16 ***
## cement               1.041e-01  3.412e-03  30.496  < 2e-16 ***
## slag                 7.982e-02  4.074e-03  19.592  < 2e-16 ***
## ash                  6.747e-02  9.775e-03   6.902 9.11e-12 ***
## water               -2.770e-01  2.122e-02 -13.055  < 2e-16 ***
```

```
## superplastic       -2.051e+00  3.871e-01  -5.298 1.44e-07 ***
## age                 3.528e-01  1.133e-02  31.141  < 2e-16 ***
## I(age^(2))         -7.932e-04  3.539e-05 -22.413  < 2e-16 ***
## water:superplastic  1.425e-02  2.350e-03   6.064 1.88e-09 ***
## ash:superplastic   -2.828e-03  9.602e-04  -2.946   0.0033 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.254 on 995 degrees of freedom
## Multiple R-squared:  0.7454, Adjusted R-squared:  0.7431
## F-statistic: 323.7 on 9 and 995 DF,  p-value: < 2.2e-16
```

The increase in `Adjusted R-Suared` is **14.29%** which is a huge increment given that the terms added are age^2 , interaction terms where one interaction term is between columns `water` and `superplastic`and other interaction term is between columns `ash` and `superplastic`

Diagnostics

Residuals vs Fitted values

```
p1<-autoplot(lr4,1) + geom_point(color = "orange")+
  theme_economist()
```

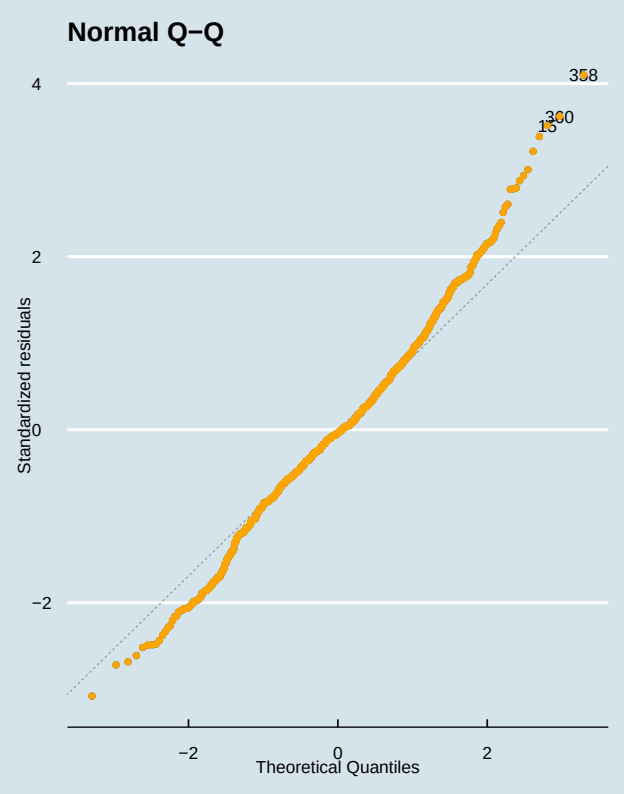
From the `Residuals` vs `Fitted values` plot, we notice a couple of things:

- The residuals are **scattered randomly around the 0 line** which can be inferred from the fact that the red line is approximately equal to the dashed line in the graph. This implies that the linear regression model is reasonable.

- The spread of residuals is approximately same across x-axis implying that the variance of the error terms are equal or in other words, the model is **homoscedastic**

Normal Q-Q

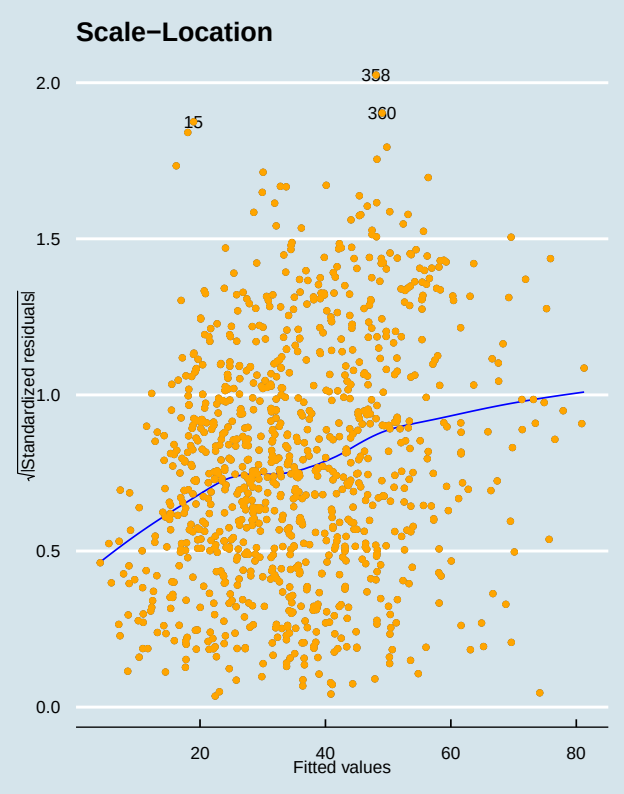
```
autoplot(lr4,2) + geom_point(color = "orange")+ theme_economist()
```



From the Normal Q-Q plot, we notice that the `Standardized residuals` when plotted against the `Theoretical Quantiles` lie on the 45° line , at least in the middle region, implying that the residuals are approximately normal distributed

Scale Location

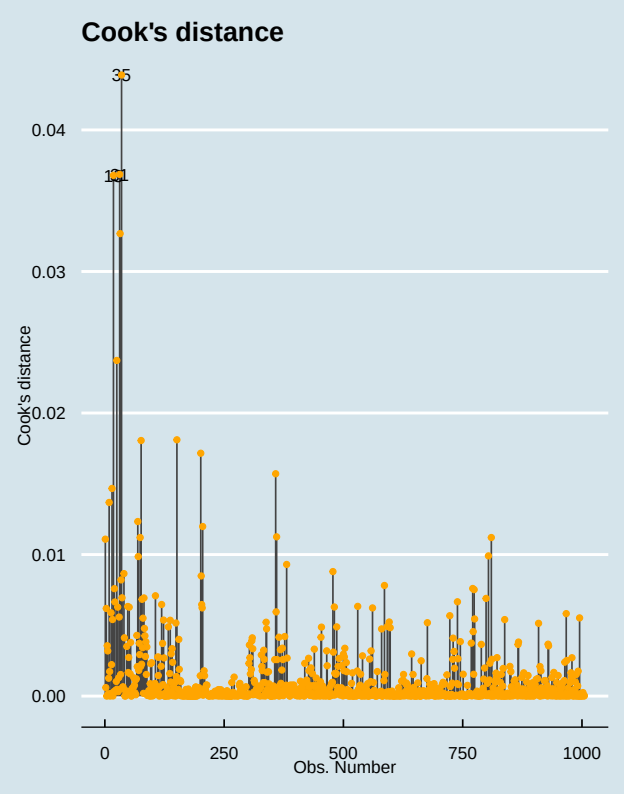
```
autoplot(lr4,3) + geom_point(color = "orange")+ theme_economist()
```



This plot is plotting $\sqrt{\text{Standardized Residuals}}$ against Fitted values and is used to find spread in the residuals as a function of the predictor's range. Here, we can further verify our assumption on top of `Residuals vs Fitted values` plot that the model is `homoscedastic` as the red line remains nearly constant and the spread of the points around the red line doesn't vary with the fitted values.

Residuals vs Leverage

```
autoplot(lr4,4) + geom_point(color = "orange")+ theme_economist()
```



`Cook's distance`is a measurement which is formed by combining an observation's leverage with its residual to find an influential set of outliers which can significantly hamper the regression model.

Recreating model

Now, from the four diagnostics plots, we find that the observations 15,18,31,35,358 and 360 are influential outliers and thus, we are removing them to again generate the linear regression model

```
a3 = a2[-c(15,18,31,35,358,360),]
lr5 = lm(strength~.+I(age^(2))+
           superplastic:water+ash:superplastic,a3)
summary(lr5)
```

```
##
## Call:
## lm(formula = strength ~ . + I(age^(2)) + superplastic:water +
##     ash:superplastic, data = a3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -25.040  -4.720  -0.262   4.694  28.033
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.567e+01  4.058e+00   8.790  < 2e-16 ***
## cement              1.042e-01  3.304e-03  31.531  < 2e-16 ***
## slag                7.985e-02  3.950e-03  20.216  < 2e-16 ***
## ash                 6.795e-02  9.466e-03   7.178 1.39e-12 ***
## water              -2.812e-01  2.074e-02 -13.558  < 2e-16 ***
```

```
## superplastic       -2.110e+00  3.764e-01  -5.605 2.70e-08 ***
## age                 3.607e-01  1.106e-02  32.611  < 2e-16 ***
## I(age^(2))         -8.381e-04  3.532e-05 -23.730  < 2e-16 ***
## water:superplastic  1.451e-02  2.285e-03   6.350 3.27e-10 ***
## ash:superplastic   -2.751e-03  9.297e-04  -2.959  0.00316 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.989 on 989 degrees of freedom
## Multiple R-squared:  0.758,  Adjusted R-squared:  0.7558
## F-statistic: 344.2 on 9 and 989 DF,  p-value: < 2.2e-16
```

Our `Adjusted R-squared` has increased by **1.27%** after removing the influential outliers