Statistical Modeling Project

Prateek Sethi

2022-12-01

STATISTICAL MODELLING PROJECT

Regression Analysis on the Concrete Compressive Strength Dataset

Load required packages

```
library("readxl")

## Warning: package 'readxl' was built under R version 4.2.2

library(faraway)

## Warning: package 'faraway' was built under R version 4.2.2

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 4.2.2

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.2.2

library(lmtest)

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

library(dplyr)

## Warning: package 'dplyr' was built under R version 4.2.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.2.2

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':
##
##     combine

## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

Load dataset

```
data <- read_excel("Concrete_Data.xls")
```

Change column names

```
#changing column names to make them shorter
colnames(data)[colnames(data) == "Cement (component 1)(kg in a m^3 mixture)"]
="cement"
colnames(data)[colnames(data) == "Blast Furnace Slag (component 2)(kg in a
m^3 mixture)"] ="blast_furnace_slag"
colnames(data)[colnames(data) == "Fly Ash (component 3)(kg in a m^3
mixture)"] ="fly_ash"
colnames(data)[colnames(data) == "Superplasticizer (component 5)(kg in a m^3
mixture)"] ="superplasticizer"
colnames(data)[colnames(data) == "Water  (component 4)(kg in a m^3 mixture)"]
="water"
colnames(data)[colnames(data) == "Coarse Aggregate  (component 6)(kg in a m^3
mixture)"] ="coarse_agg"
colnames(data)[colnames(data) == "Fine Aggregate (component 7)(kg in a m^3
mixture)"] ="fine_agg"
```

```r
colnames(data)[colnames(data) == "Age (day)"] ="age"
colnames(data)[colnames(data) == "Concrete compressive strength(MPa,
megapascals)"] ="concrete_strength"
```

Univariate Analysis

Data Description

```r
#column names
names(data)

## [1] "cement"             "blast_furnace_slag" "fly_ash"
## [4] "water"              "superplasticizer"   "coarse_agg"
## [7] "fine_agg"           "age"                "concrete_strength"

#data size
dim(data)

## [1] 1030    9

#variable description
str(data)

## tibble [1,030 × 9] (S3: tbl_df/tbl/data.frame)
##  $ cement            : num [1:1030] 540 540 332 332 199 ...
##  $ blast_furnace_slag: num [1:1030] 0 0 142 142 132 ...
##  $ fly_ash           : num [1:1030] 0 0 0 0 0 0 0 0 0 0 ...
##  $ water             : num [1:1030] 162 162 228 228 192 228 228 228 228
228 ...
##  $ superplasticizer  : num [1:1030] 2.5 2.5 0 0 0 0 0 0 0 0 ...
##  $ coarse_agg        : num [1:1030] 1040 1055 932 932 978 ...
##  $ fine_agg          : num [1:1030] 676 676 594 594 826 ...
##  $ age               : num [1:1030] 28 28 270 365 360 90 365 28 28 28 ...
##  $ concrete_strength : num [1:1030] 80 61.9 40.3 41.1 44.3 ...

#top 5 rows
head(data)

## # A tibble: 6 × 9
##   cement blast_furnace_slag fly_ash water superp…¹ coars…² fine_…³   age
concr…⁴
##    <dbl>              <dbl>   <dbl> <dbl>    <dbl>   <dbl>   <dbl> <dbl>
<dbl>
## 1   540                  0        0   162      2.5   1040     676     28
80.0
## 2   540                  0        0   162      2.5   1055     676     28
61.9
## 3   332.               142.       0   228      0      932     594    270
40.3
## 4   332.               142.       0   228      0      932     594    365
41.1
## 5   199.               132.       0   192      0      978.    826.   360
44.3
```

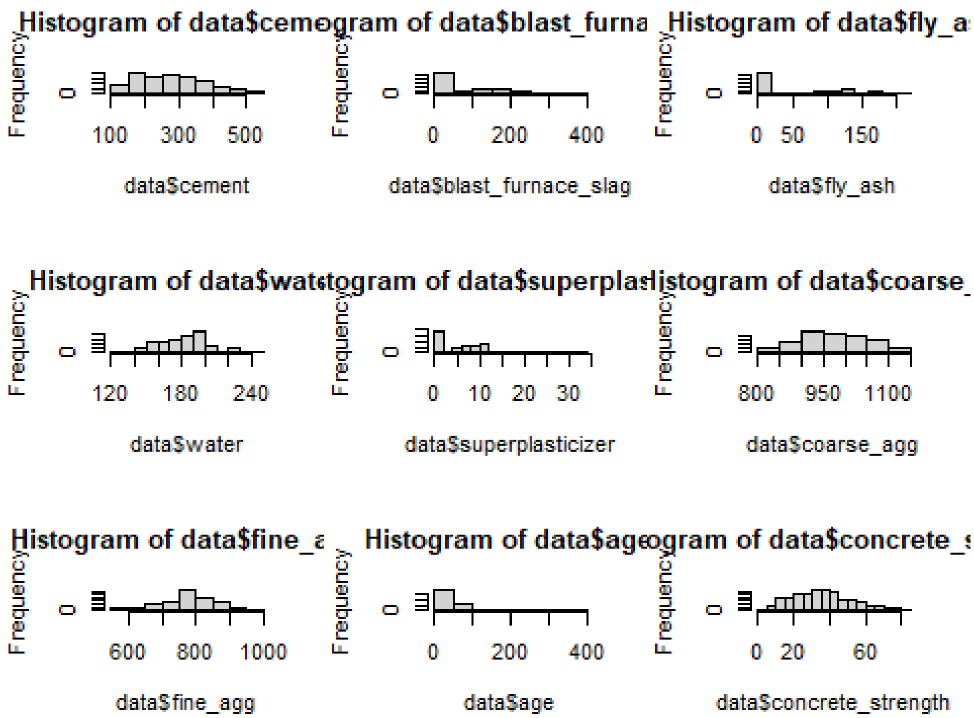
```
## 6   266                  114         0    228       0       932      670      90
47.0
## # … with abbreviated variable names ¹superplasticizer, ²coarse_agg, ³
fine_agg,
## #   ⁴concrete_strength
```

```
#data summary
summary(data)
```

```
##      cement       blast_furnace_slag    fly_ash            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.:118.27   3rd Qu.:192.0  
##  Max.   :540.0   Max.   :359.4      Max.   :200.10   Max.   :247.0  
##  superplasticizer   coarse_agg        fine_agg          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 : 6.350   Median : 968.0   Median :779.5   Median : 28.00  
##  Mean   : 6.203   Mean   : 972.9   Mean   :773.6   Mean   : 45.66  
##  3rd Qu.:10.160   3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00  
##  Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00  
##  concrete_strength
##  Min.   : 2.332   
##  1st Qu.:23.707   
##  Median :34.443   
##  Mean   :35.818   
##  3rd Qu.:46.136   
##  Max.   :82.599   
```

Data Distribution

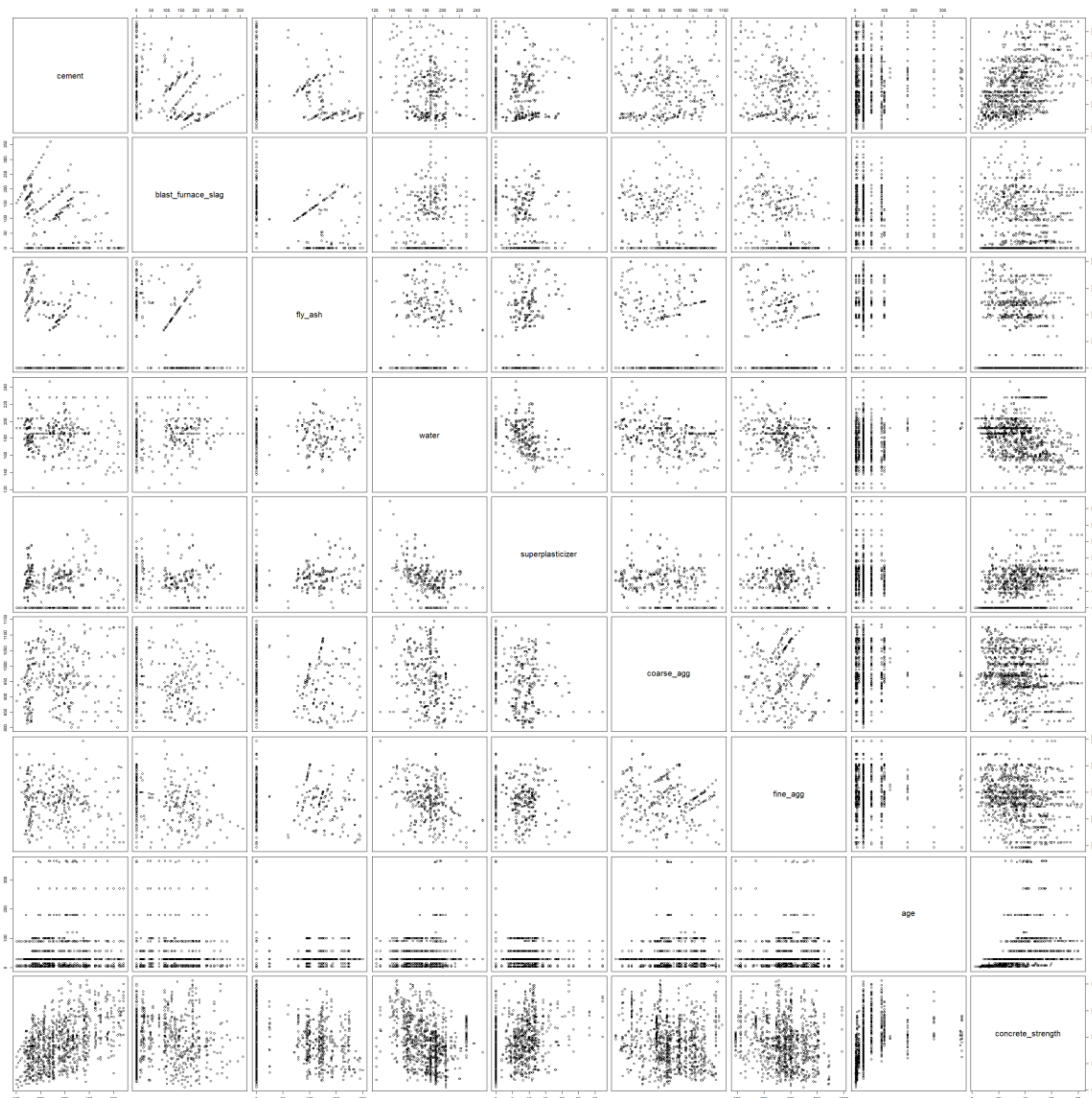
```
par(mfrow=c(3,3))
hist(data$cement)
hist(data$blast_furnace_slag)
hist(data$fly_ash)
hist(data$water)
hist(data$superplasticizer)
hist(data$coarse_agg)
hist(data$fine_agg)
hist(data$age)
hist(data$concrete_strength)
```



Multivariate Analysis

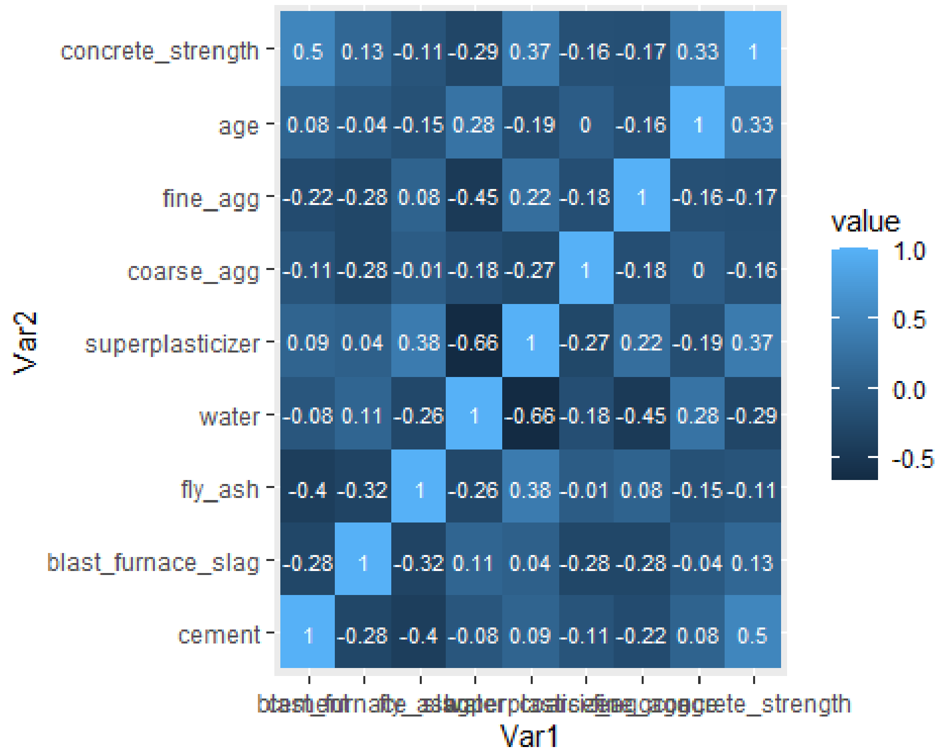
Correlation Matrix

```
pairs(data)
```



Correlation Heatmap

```
corr_mat <- round(cor(data),2)
melted_corr_mat <- melt(corr_mat)
ggplot(data = melted_corr_mat, aes(x=Var1, y=Var2,
                                   fill=value), size = 10) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value),
            color = "white", size = 3)
```



Split train test set

```
seed = 21

##train test split
set.seed(seed)
dt = sort(sample(nrow(data), nrow(data)*.80))
train<-data[dt,]
test<-data[-dt,]
```

Baseline Model

```
baseline = lm(concrete_strength~.,data=train)
```

Variance Inflation Factor

```
vif(baseline)

##             cement blast_furnace_slag            fly_ash
water
##           7.278260           7.043331           6.042083
6.628763
##   superplasticizer         coarse_agg           fine_agg
age
##           2.844048           4.954347           6.824790
1.110492
```

Outliers

```
nrow(data[which(abs(rstandard(baseline)) > 2) , ])
```

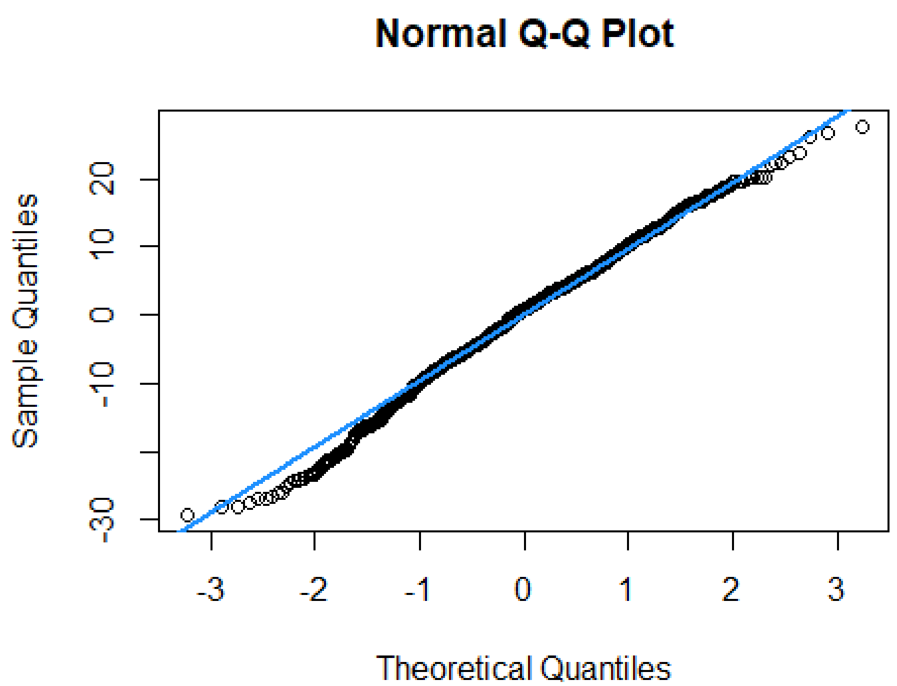
```
## [1] 39
```

Influential Points

```
indices = cooks.distance(baseline) > 4 / length(cooks.distance(baseline))
nrow(data[which(indices) , ])

## [1] 62
```

Test Model Assumptions

```
plot_residuals <- function(model) {
  plot(fitted(model), resid(model), col = "grey", pch = 20,
       xlab = "Fitted", ylab = "Residuals", main = "Resid plot")
  abline(h = 0, col = "darkorange", lwd = 2)
}

plot_qq <- function(model) {
  qqnorm(resid(model))
  qqline(resid(model), col = "dodgerblue", lwd = 2)
}

check_model_assumptions <- function(model) {
  #check by graphs
  plot_residuals(model)
  #invisible(readline(prompt="Press [enter] to continue"))
  plot_qq(model)

  #bptest for equal variance
  print(bptest(model))

  #shapiro wilk test for normality
  print(shapiro.test(resid(model)))
}

check_model_assumptions(baseline)
```



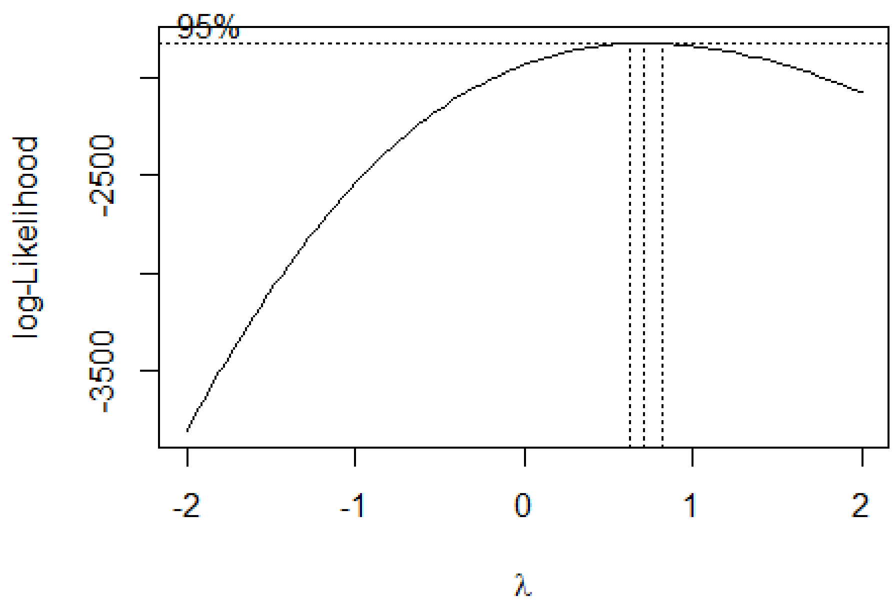

```
##
##  studentized Breusch-Pagan test
##
## data:  model
```

```
## BP = 109.42, df = 8, p-value < 2.2e-16
##
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.99357, p-value = 0.001314
```

Applying Response Transformation to satisfy model assumptions

Using BoxCox graph

```
par(mfrow=c(1,1))
boxcox(concrete_strength~.,data=train)
```



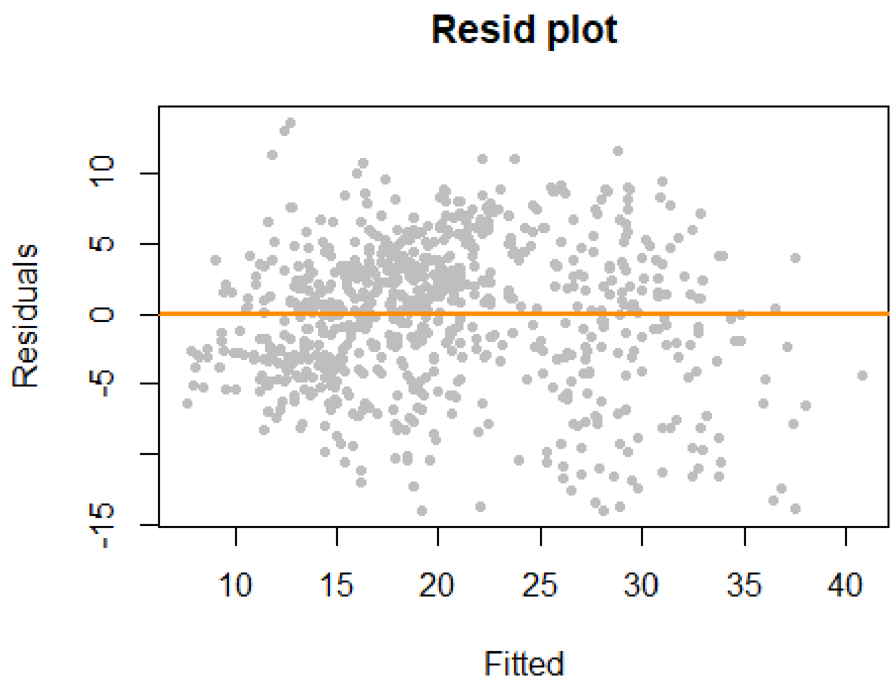
lambda = 0.8

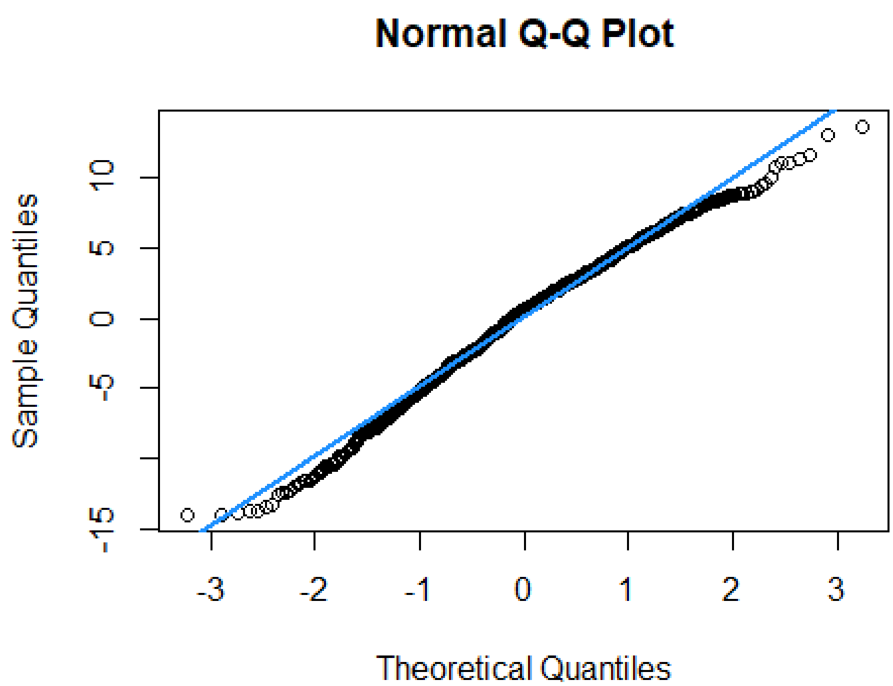
Transform response variable and check model assumptions

```
lambda  =  0.8
transformed_model = lm(((concrete_strength^lambda)-1)/lambda ~., data=train)
summary(transformed_model)

##
## Call:
## lm(formula = ((concrete_strength^lambda) - 1)/lambda ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0292  -3.2096   0.5557   3.4607  13.6373
##
```

```
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)    
## (Intercept)        -10.837406  14.199158  -0.763  0.44554    
## cement               0.059654   0.004559  13.084  < 2e-16 ***
## blast_furnace_slag   0.051994   0.005428   9.579  < 2e-16 ***
## fly_ash              0.044438   0.006791   6.543 1.06e-10 ***
## water               -0.068413   0.021315  -3.210  0.00138 ** 
## superplasticizer     0.165895   0.050768   3.268  0.00113 ** 
## coarse_agg           0.009295   0.005038   1.845  0.06539 .  
## fine_agg             0.010102   0.005748   1.758  0.07920 .  
## age                  0.057435   0.003014  19.057  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 5.081 on 815 degrees of freedom
## Multiple R-squared:  0.6201, Adjusted R-squared:  0.6164 
## F-statistic: 166.3 on 8 and 815 DF,  p-value: < 2.2e-16
```

```
check_model_assumptions(transformed_model)
```






```
##
##  studentized Breusch-Pagan test
##
## data:  model
```

```
## BP = 72.581, df = 8, p-value = 1.503e-12
##
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.98927, p-value = 1.006e-05
```

The model assumptions still fail for linearity, normality and equal variance.

Prediction Performance

Variable Addition [Increasing model complexity]

We will now start adding non linear predictor variables in order to capture any nonlinearity that's present in the data.

Fit a simple model with all predictors and calculate MSE and PRESS score

```
mlr <- lm(concrete_strength~.,data=train)
summary(mlr)

##
## Call:
## lm(formula = concrete_strength ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.2826  -6.4525   0.7969   6.6006  27.6096
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -28.982063  28.748522  -1.008 0.313694
## cement              0.122050   0.009231  13.221  < 2e-16 ***
## blast_furnace_slag  0.107661   0.010990   9.797  < 2e-16 ***
## fly_ash             0.088705   0.013750   6.451  1.9e-10 ***
## water              -0.143393   0.043155  -3.323 0.000931 ***
## superplasticizer    0.332813   0.102787   3.238 0.001253 **
## coarse_agg          0.019683   0.010200   1.930 0.053991 .
## fine_agg            0.022415   0.011637   1.926 0.054423 .
## age                 0.115418   0.006102  18.914  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 815 degrees of freedom
## Multiple R-squared:  0.6235, Adjusted R-squared:  0.6198
## F-statistic: 168.7 on 8 and 815 DF,  p-value: < 2.2e-16

n = nrow(train)
press_mlr = sqrt(sum((resid(mlr)/(1-hatvalues(mlr)))^2)/n)
press_mlr
```

```
## [1] 10.37549

calculate_mse_test <- function(model, test) {
  ypred = predict(model, newdata = test)
  resid = ypred - test$concrete_strength
  mse = mean(resid^2)
  return(mse)
}

mse_mlr = calculate_mse_test(mlr, test)
mse_mlr

## [1] 117.934
```

Add squared polynomial terms and calculate MSE and PRESS score

```
mlr_squared <- lm(concrete_strength ~ cement + blast_furnace_slag +
                    fly_ash + water + superplasticizer + coarse_agg +
                    fine_agg + age + I(cement^2) + I(blast_furnace_slag^2) +
                    I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                    I(coarse_agg^2) +
                    I(fine_agg^2) + I(age^2), data = train)
summary(mlr_squared)

##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(cement^2) + I(blast_furnace_slag^2) + I(fly_ash^2) +
##     I(water^2) + I(superplasticizer^2) + I(coarse_agg^2) + I(fine_agg^2) +
##     I(age^2), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.6484  -4.5178   0.2729   5.0186  27.1957
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             -5.039e+01  5.046e+01  -0.999  0.31823
## cement                   1.343e-01  1.816e-02   7.396 3.51e-13 ***
## blast_furnace_slag       1.214e-01  1.348e-02   9.001  < 2e-16 ***
## fly_ash                  9.231e-02  2.007e-02   4.599 4.94e-06 ***
## water                   -5.919e-01  1.834e-01  -3.228  0.00130 **
## superplasticizer         9.715e-01  1.606e-01   6.050 2.22e-09 ***
## coarse_agg              -2.092e-02  8.258e-02  -0.253  0.80004
## fine_agg                 2.529e-01  5.727e-02   4.417 1.14e-05 ***
## age                      3.568e-01  1.243e-02  28.695  < 2e-16 ***
## I(cement^2)             -3.377e-05  2.633e-05  -1.283  0.19997
## I(blast_furnace_slag^2) -1.201e-04  4.326e-05  -2.775  0.00564 **
## I(fly_ash^2)            -2.794e-04  1.214e-04  -2.302  0.02161 *
## I(water^2)               1.173e-03  5.098e-04   2.301  0.02164 *
```

```
## I(superplasticizer^2)   -4.124e-02  6.618e-03  -6.231 7.44e-10 ***
## I(coarse_agg^2)          1.503e-05  4.195e-05   0.358  0.72025
## I(fine_agg^2)           -1.563e-04  3.616e-05  -4.322 1.74e-05 ***
## I(age^2)                -8.103e-04  3.844e-05 -21.080  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.916 on 807 degrees of freedom
## Multiple R-squared:  0.7793, Adjusted R-squared:  0.7749
## F-statistic: 178.1 on 16 and 807 DF,  p-value: < 2.2e-16
```

```
press_sq = sqrt(sum((resid(mlr_squared)/(1-hatvalues(mlr_squared)))^2)/n)
press_sq
```

```
## [1] 8.045019
```

```
mse_sq = calculate_mse_test(mlr_squared, test)
mse_sq
```

```
## [1] 70.66705
```

Add cubic polynomial terms and calculate MSE and PRESS score

```
mlr_cubed <- lm(concrete_strength ~ cement + blast_furnace_slag +
                  fly_ash + water + superplasticizer + coarse_agg +
                  fine_agg + age + I(cement^2) + I(blast_furnace_slag^2) +
                  I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                  I(coarse_agg^2) +
                  I(fine_agg^2) + I(age^2) + I(cement^3) +
                  I(blast_furnace_slag^3) +
                  I(fly_ash^3) + I(water^3) + I(superplasticizer^3) +
                  I(coarse_agg^3) +
                  I(fine_agg^3) + I(age^3), data = train)
summary(mlr_cubed)
```

```
##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(cement^2) + I(blast_furnace_slag^2) + I(fly_ash^2) +
##     I(water^2) + I(superplasticizer^2) + I(coarse_agg^2) + I(fine_agg^2) +
##     I(age^2) + I(cement^3) + I(blast_furnace_slag^3) + I(fly_ash^3) +
##     I(water^3) + I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
##     I(age^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.7998  -4.0615  -0.0275   4.2222  21.1396
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -1.544e+03  4.601e+02  -3.357 0.000825 ***
```

```
## cement                   4.053e-01  6.138e-02   6.603 7.35e-11 ***
## blast_furnace_slag       7.397e-02  2.098e-02   3.526 0.000447 ***
## fly_ash                 -5.691e-02  5.987e-02  -0.951 0.342080
## water                    8.236e+00  1.185e+00   6.950 7.58e-12 ***
## superplasticizer         1.390e+00  2.636e-01   5.272 1.74e-07 ***
## coarse_agg               1.685e+00  1.288e+00   1.309 0.191069
## fine_agg                 1.889e+00  5.034e-01   3.752 0.000188 ***
## age                      6.315e-01  2.133e-02  29.608  < 2e-16 ***
## I(cement^2)             -9.224e-04  2.041e-04  -4.518 7.17e-06 ***
## I(blast_furnace_slag^2)  3.826e-04  1.873e-04   2.043 0.041371 *
## I(fly_ash^2)             1.590e-03  8.278e-04   1.920 0.055188 .
## I(water^2)              -4.880e-02  6.578e-03  -7.419 3.02e-13 ***
## I(superplasticizer^2)   -1.295e-01  2.587e-02  -5.007 6.79e-07 ***
## I(coarse_agg^2)         -1.730e-03  1.336e-03  -1.294 0.195892
## I(fine_agg^2)           -2.339e-03  6.559e-04  -3.566 0.000384 ***
## I(age^2)                -3.725e-03  1.977e-04 -18.843  < 2e-16 ***
## I(cement^3)              8.983e-07  2.122e-07   4.234 2.56e-05 ***
## I(blast_furnace_slag^3) -1.232e-06  4.313e-07  -2.856 0.004399 **
## I(fly_ash^3)            -5.887e-06  2.899e-06  -2.031 0.042585 *
## I(water^3)               9.274e-05  1.200e-05   7.731 3.21e-14 ***
## I(superplasticizer^3)    2.554e-03  6.366e-04   4.012 6.58e-05 ***
## I(coarse_agg^3)          5.908e-07  4.601e-07   1.284 0.199545
## I(fine_agg^3)            9.610e-07  2.828e-07   3.398 0.000712 ***
## I(age^3)                 6.129e-06  4.091e-07  14.983  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.669 on 799 degrees of freedom
## Multiple R-squared:  0.8449, Adjusted R-squared:  0.8402
## F-statistic: 181.4 on 24 and 799 DF,  p-value: < 2.2e-16
```

```
press_cub = sqrt(sum((resid(mlr_cubed)/(1-hatvalues(mlr_cubed)))^2)/n)
press_cub
```

```
## [1] 6.841195
```

```
mse_cub = calculate_mse_test(mlr_cubed, test)
mse_cub
```

```
## [1] 52.50221
```

Add square root terms and calculate MSE and PRESS score

```
mlr_sqrt <- lm(concrete_strength ~ cement + blast_furnace_slag +
                 fly_ash + water + superplasticizer + coarse_agg +
                 fine_agg + age + I(cement^2) + I(blast_furnace_slag^2) +
                 I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                 I(coarse_agg^2) +
                 I(fine_agg^2) + I(age^2) + I(cement^3) +
                 I(blast_furnace_slag^3) +
                 I(fly_ash^3) + I(water^3) + I(superplasticizer^3) +
                 I(coarse_agg^3) +
```

```
                 I(fine_agg^3) + I(age^3) + I(sqrt(cement)) +
                 I(sqrt(blast_furnace_slag)) +
                 I(sqrt(fly_ash)) + I(sqrt(water)) +
                 I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                 I(sqrt(fine_agg)) + I(sqrt(age)), data = train)
summary(mlr_sqrt)

##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(cement^2) + I(blast_furnace_slag^2) + I(fly_ash^2) +
##     I(water^2) + I(superplasticizer^2) + I(coarse_agg^2) + I(fine_agg^2) +
##     I(age^2) + I(cement^3) + I(blast_furnace_slag^3) + I(fly_ash^3) +
##     I(water^3) + I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
##     I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
##     I(sqrt(fly_ash)) + I(sqrt(water)) + I(sqrt(superplasticizer)) +
##     I(sqrt(coarse_agg)) + I(sqrt(fine_agg)) + I(sqrt(age)), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.6318  -3.6333   0.1838   3.8447  18.2204
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.255e+04  3.004e+04   1.084   0.2788
## cement                      4.081e+00  8.589e-01   4.751 2.40e-06 ***
## blast_furnace_slag          1.778e-01  7.929e-02   2.242   0.0252 *
## fly_ash                    -1.472e+00  2.816e-01  -5.228 2.19e-07 ***
## water                      -2.311e+02  4.437e+01  -5.210 2.42e-07 ***
## superplasticizer            6.783e-01  1.154e+00   0.588   0.5569
## coarse_agg                  1.779e+02  9.044e+01   1.967   0.0496 *
## fine_agg                   -3.027e+01  2.597e+01  -1.166   0.2440
## age                        -5.110e-01  8.941e-02  -5.716 1.55e-08 ***
## I(cement^2)                -5.073e-03  1.021e-03  -4.969 8.27e-07 ***
## I(blast_furnace_slag^2)    -2.653e-05  3.401e-04  -0.078   0.9378
## I(fly_ash^2)                1.056e-02  1.827e-03   5.779 1.08e-08 ***
## I(water^2)                  4.029e-01  8.347e-02   4.826 1.67e-06 ***
## I(superplasticizer^2)      -1.097e-01  5.472e-02  -2.004   0.0454 *
## I(coarse_agg^2)            -6.282e-02  3.130e-02  -2.007   0.0451 *
## I(fine_agg^2)               1.113e-02  1.128e-02   0.986   0.3244
## I(age^2)                    5.431e-04  3.704e-04   1.466   0.1430
## I(cement^3)                 3.528e-06  6.869e-07   5.136 3.54e-07 ***
## I(blast_furnace_slag^3)    -4.927e-07  6.069e-07  -0.812   0.4171
## I(fly_ash^3)               -2.875e-05  4.936e-06  -5.825 8.29e-09 ***
## I(water^3)                 -4.108e-04  9.312e-05  -4.412 1.17e-05 ***
## I(superplasticizer^3)       2.374e-03  1.027e-03   2.313   0.0210 *
## I(coarse_agg^3)             1.324e-05  6.479e-06   2.043   0.0414 *
## I(fine_agg^3)              -2.385e-06  2.920e-06  -0.817   0.4144
## I(age^3)                   -2.522e-07  6.074e-07  -0.415   0.6781
```

```
## I(sqrt(cement))             -6.498e+01  1.498e+01  -4.337 1.63e-05 ***
## I(sqrt(blast_furnace_slag)) -5.365e-01  5.150e-01  -1.042   0.2979
## I(sqrt(fly_ash))             7.574e+00  1.508e+00   5.023 6.29e-07 ***
## I(sqrt(water))               3.374e+03  6.270e+02   5.382 9.73e-08 ***
## I(sqrt(superplasticizer))    1.247e+00  2.253e+00   0.553   0.5802
## I(sqrt(coarse_agg))         -5.815e+03  2.990e+03  -1.944   0.0522 .
## I(sqrt(fine_agg))            9.652e+02  7.651e+02   1.261   0.2075
## I(sqrt(age))                 9.063e+00  6.942e-01  13.055  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.827 on 791 degrees of freedom
## Multiple R-squared:  0.8828, Adjusted R-squared:  0.878
## F-statistic: 186.1 on 32 and 791 DF,  p-value: < 2.2e-16
```

```
press_sqrt = sqrt(sum((resid(mlr_sqrt)/(1-hatvalues(mlr_sqrt)))^2)/n)
press_sqrt
```

```
## [1] 5.987787
```

```
mse_sqrt = calculate_mse_test(mlr_sqrt, test)
mse_sqrt
```

```
## [1] 39.8393
```

Add logarithmic terms and calculate MSE and PRESS score

```
mlr_log <- lm(concrete_strength ~ cement + blast_furnace_slag +
                fly_ash + water + superplasticizer + coarse_agg +
                fine_agg + age + I(cement^2) + I(blast_furnace_slag^2) +
                I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                I(coarse_agg^2) +
                I(fine_agg^2) + I(age^2) + I(cement^3) +
                I(blast_furnace_slag^3) +
                I(fly_ash^3) + I(water^3) + I(superplasticizer^3) +
                I(coarse_agg^3) +
                I(fine_agg^3) + I(age^3) + I(sqrt(cement)) +
                I(sqrt(blast_furnace_slag)) +
                I(sqrt(fly_ash)) + I(sqrt(water)) +
                I(sqrt(superplasticizer))+ I(sqrt(coarse_agg)) +
                I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                I(log(blast_furnace_slag+1)) +
                I(log(fly_ash+1)) + I(log(water+1)) +
                I(log(superplasticizer+1)) + I(log(coarse_agg + 1)) +
                I(log(fine_agg+1)) + I(log(age + 1)) , data = train)
summary(mlr_log)
```

```
##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(cement^2) + I(blast_furnace_slag^2) + I(fly_ash^2) +
```

```
##     I(water^2) + I(superplasticizer^2) + I(coarse_agg^2) + I(fine_agg^2) +
##     I(age^2) + I(cement^3) + I(blast_furnace_slag^3) + I(fly_ash^3) +
##     I(water^3) + I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
##     I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
##     I(sqrt(fly_ash)) + I(sqrt(water)) + I(sqrt(superplasticizer)) +
##     I(sqrt(coarse_agg)) + I(sqrt(fine_agg)) + I(sqrt(age)) +
##     I(log(cement + 1)) + I(log(blast_furnace_slag + 1)) + I(log(fly_ash +
##     1)) + I(log(water + 1)) + I(log(superplasticizer + 1)) +
##     I(log(coarse_agg + 1)) + I(log(fine_agg + 1)) + I(log(age +
##     1)), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.7365  -3.7549   0.2675   3.7292  18.2079
##
## Coefficients: (2 not defined because of singularities)
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.695e+04  1.074e+05  -0.158  0.87458
## cement                       -2.910e+00  9.177e+00  -0.317  0.75127
## blast_furnace_slag           -2.728e+00  6.881e-01  -3.965 8.02e-05 ***
## fly_ash                       5.865e+00  5.993e+00   0.979  0.32807
## water                         5.919e+02  1.221e+03   0.485  0.62805
## superplasticizer             -1.890e-02  1.169e+00  -0.016  0.98710
## coarse_agg                    2.345e+02  9.126e+01   2.570  0.01035 *
## fine_agg                     -3.125e+01  2.588e+01  -1.208  0.22758
## age                          -2.841e-01  4.484e-01  -0.634  0.52651
## I(cement^2)                  -1.074e-03  5.669e-03  -0.189  0.84984
## I(blast_furnace_slag^2)       5.554e-03  1.346e-03   4.125 4.10e-05 ***
## I(fly_ash^2)                 -8.299e-03  1.531e-02  -0.542  0.58787
## I(water^2)                   -3.715e-01  1.140e+00  -0.326  0.74452
## I(superplasticizer^2)        -9.551e-02  9.272e-02  -1.030  0.30332
## I(coarse_agg^2)              -8.302e-02  3.160e-02  -2.627  0.00878 **
## I(fine_agg^2)                 1.157e-02  1.124e-02   1.029  0.30386
## I(age^2)                      1.053e-04  9.680e-04   0.109  0.91344
## I(cement^3)                   1.775e-06  2.654e-06   0.669  0.50384
## I(blast_furnace_slag^3)      -6.943e-06  1.615e-06  -4.298 1.94e-05 ***
## I(fly_ash^3)                  3.053e-06  2.584e-05   0.118  0.90598
## I(water^3)                    1.684e-04  8.470e-04   0.199  0.84245
## I(superplasticizer^3)         2.111e-03  1.633e-03   1.292  0.19661
## I(coarse_agg^3)               1.754e-05  6.545e-06   2.679  0.00753 **
## I(fine_agg^3)                -2.511e-06  2.910e-06  -0.863  0.38841
## I(age^3)                      2.504e-07  1.197e-06   0.209  0.83443
## I(sqrt(cement))               1.854e+02  3.155e+02   0.588  0.55687
## I(sqrt(blast_furnace_slag))   4.899e+01  1.172e+01   4.182 3.22e-05 ***
## I(sqrt(fly_ash))             -1.064e+02  9.366e+01  -1.136  0.25647
## I(sqrt(water))               -2.003e+04  3.488e+04  -0.574  0.56601
## I(sqrt(superplasticizer))     8.674e+00  2.454e+01   0.353  0.72382
## I(sqrt(coarse_agg))          -7.660e+03  3.016e+03  -2.539  0.01130 *
## I(sqrt(fine_agg))             9.939e+02  7.625e+02   1.303  0.19280
## I(sqrt(age))                  5.173e+00  7.217e+00   0.717  0.47376
```

```
## I(log(cement + 1))             -6.519e+02  7.883e+02  -0.827  0.40850
## I(log(blast_furnace_slag + 1)) -5.487e+01  1.300e+01  -4.220 2.73e-05 ***
## I(log(fly_ash + 1))             1.219e+02  1.007e+02   1.211  0.22619
## I(log(water + 1))               4.916e+04  7.360e+04   0.668  0.50442
## I(log(superplasticizer + 1))   -7.363e+00  2.979e+01  -0.247  0.80485
## I(log(coarse_agg + 1))                 NA         NA      NA       NA
## I(log(fine_agg + 1))                   NA         NA      NA       NA
## I(log(age + 1))                 4.376e+00  7.762e+00   0.564  0.57303
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.778 on 785 degrees of freedom
## Multiple R-squared:  0.8856, Adjusted R-squared:  0.8801
## F-statistic: 159.9 on 38 and 785 DF,  p-value: < 2.2e-16
```

```
press_log = sqrt(sum((resid(mlr_log)/(1-hatvalues(mlr_log)))^2)/n)
press_log
```

```
## [1] 5.972964
```

```
mse_log = calculate_mse_test(mlr_log, test)
```

```
## Warning in predict.lm(model, newdata = test): prediction from a rank-
deficient
## fit may be misleading
```

```
mse_log
```

```
## [1] 39.77537
```

Adding 2nd order interaction terms

```
mlr_int <- lm(concrete_strength ~ cement + blast_furnace_slag + fly_ash +
                water + superplasticizer + coarse_agg + fine_agg + age +
                cement*blast_furnace_slag + cement*fly_ash + cement*water +
                cement*superplasticizer + cement*coarse_agg
                +cement*fine_agg+
                cement*age + blast_furnace_slag*fly_ash +
                blast_furnace_slag*fly_ash+
                blast_furnace_slag*water +
                blast_furnace_slag*superplasticizer +
                blast_furnace_slag*coarse_agg + blast_furnace_slag*fine_agg+
                blast_furnace_slag*age + fly_ash*water +
                fly_ash*superplasticizer +
                fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                water*superplasticizer + water*coarse_agg + water*fine_agg+
                water*age + superplasticizer*coarse_agg +
                superplasticizer*fine_agg+
                superplasticizer*age + coarse_agg*fine_agg + coarse_agg*age+
                fine_agg*age , data = train)
summary(mlr_int)
```

```
##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + cement * blast_furnace_slag + cement * fly_ash + cement *
##     water + cement * superplasticizer + cement * coarse_agg +
##     cement * fine_agg + cement * age + blast_furnace_slag * fly_ash +
##     blast_furnace_slag * fly_ash + blast_furnace_slag * water +
##     blast_furnace_slag * superplasticizer + blast_furnace_slag *
##     coarse_agg + blast_furnace_slag * fine_agg + blast_furnace_slag *
##     age + fly_ash * water + fly_ash * superplasticizer + fly_ash *
##     coarse_agg + fly_ash * fine_agg + fly_ash * age + water *
##     superplasticizer + water * coarse_agg + water * fine_agg +
##     water * age + superplasticizer * coarse_agg + superplasticizer *
##     fine_agg + superplasticizer * age + coarse_agg * fine_agg +
##     coarse_agg * age + fine_agg * age, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.9441  -5.4361   0.1303   5.7533  30.4008
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -8.268e+01  1.546e+02  -0.535 0.592973
## cement                               2.853e-01  1.680e-01   1.698 0.089822
.
## blast_furnace_slag                  -1.573e-02  2.349e-01  -0.067 0.946635
## fly_ash                             -4.970e-01  3.354e-01  -1.482 0.138778
## water                                1.540e+00  5.053e-01   3.049 0.002376
**
## superplasticizer                    -1.670e+00  5.472e+00  -0.305 0.760364
## coarse_agg                           1.718e-02  1.066e-01   0.161 0.872010
## fine_agg                            -1.680e-01  1.059e-01  -1.587 0.112983
## age                                 -7.643e-01  5.539e-01  -1.380 0.168000
## cement:blast_furnace_slag            1.621e-04  6.356e-05   2.550 0.010964
*
## cement:fly_ash                       3.396e-04  9.064e-05   3.746 0.000192
***
## cement:water                        -1.748e-03  4.202e-04  -4.159 3.54e-05
***
## cement:superplasticizer             -3.489e-03  2.123e-03  -1.643 0.100797
## cement:coarse_agg                    3.547e-05  7.397e-05   0.480 0.631711
## cement:fine_agg                      1.128e-04  6.442e-05   1.751 0.080279
.
## cement:age                           6.896e-04  1.986e-04   3.472 0.000544
***
## blast_furnace_slag:fly_ash           4.476e-04  1.299e-04   3.444 0.000603
***
## blast_furnace_slag:water            -8.750e-04  6.027e-04  -1.452 0.146936
## blast_furnace_slag:superplasticizer  2.810e-05  2.550e-03   0.011 0.991212
```

```
## blast_furnace_slag:coarse_agg       -3.314e-05  9.975e-05  -0.332 0.739774
## blast_furnace_slag:fine_agg          2.777e-04  8.009e-05   3.467 0.000555
***
## blast_furnace_slag:age               9.611e-04  1.970e-04   4.878 1.30e-06
***
## fly_ash:water                       -1.406e-03  7.183e-04  -1.958 0.050599
.
## fly_ash:superplasticizer            -4.799e-03  3.280e-03  -1.463 0.143833
## fly_ash:coarse_agg                   2.181e-04  1.415e-04   1.542 0.123582
## fly_ash:fine_agg                     5.906e-04  1.478e-04   3.996 7.06e-05
***
## fly_ash:age                          2.117e-03  3.344e-04   6.330 4.11e-10
***
## water:superplasticizer               9.074e-03  6.876e-03   1.320 0.187334
## water:coarse_agg                    -8.435e-04  2.937e-04  -2.872 0.004193
**
## water:fine_agg                      -3.249e-04  2.968e-04  -1.094 0.274094
## water:age                            6.539e-05  9.568e-04   0.068 0.945533
## superplasticizer:coarse_agg          1.576e-03  2.057e-03   0.766 0.443696
## superplasticizer:fine_agg            5.520e-05  2.380e-03   0.023 0.981503
## superplasticizer:age                 4.756e-03  2.584e-03   1.841 0.066042
.
## coarse_agg:fine_agg                  1.433e-04  7.385e-05   1.941 0.052670
.
## coarse_agg:age                       1.300e-04  1.694e-04   0.767 0.443174
## fine_agg:age                         5.879e-04  2.367e-04   2.483 0.013224
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.237 on 787 degrees of freedom
## Multiple R-squared:  0.7669, Adjusted R-squared:  0.7563
## F-statistic: 71.93 on 36 and 787 DF,  p-value: < 2.2e-16
```

```
press_int = sqrt(sum((resid(mlr_int)/(1-hatvalues(mlr_int)))^2)/n)
press_int
```

```
## [1] 8.607692
```

```
mse_int = calculate_mse_test(mlr_int, test)
mse_int
```

```
## [1] 85.61925
```

Add both polynomial and 2nd order interaction terms

```
mlr_allt_int <- lm(concrete_strength ~ cement + blast_furnace_slag + fly_ash
                  + water + superplasticizer + coarse_agg + fine_agg + age +
                  cement*blast_furnace_slag + cement*fly_ash + cement*water+
                  cement*superplasticizer + cement*coarse_agg
                  +cement*fine_agg +
                  cement*age + blast_furnace_slag*fly_ash +
```

```
                   blast_furnace_slag*fly_ash+
                   blast_furnace_slag*water +
                   blast_furnace_slag*superplasticizer +
                   blast_furnace_slag*coarse_agg +
                   blast_furnace_slag*fine_agg
                   +blast_furnace_slag*age + fly_ash*water +
                   fly_ash*superplasticizer +
                   fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                   water*superplasticizer + water*coarse_agg +
                   water*fine_agg+
                   water*age + superplasticizer*coarse_agg +
                   superplasticizer*fine_agg+
                   superplasticizer*age + coarse_agg*fine_agg +
                   coarse_agg*age
                   +fine_agg*age + I(cement^2) + I(blast_furnace_slag^2) +
                   I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                   I(coarse_agg^2) + I(fine_agg^2) + I(age^2) + I(cement^3) +
                   I(blast_furnace_slag^3) + I(fly_ash^3) + I(water^3) +
                   I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
                   I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
                   I(sqrt(fly_ash)) + I(sqrt(water)) +
                   I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                   I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                   I(log(blast_furnace_slag+1)) +
                   I(log(fly_ash+1)) + I(log(water+1)) +
                   I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                   I(log(fine_agg+1)) + I(log(age)), data = train)

summary(mlr_allt_int)

##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + cement * blast_furnace_slag + cement * fly_ash + cement *
##     water + cement * superplasticizer + cement * coarse_agg +
##     cement * fine_agg + cement * age + blast_furnace_slag * fly_ash +
##     blast_furnace_slag * fly_ash + blast_furnace_slag * water +
##     blast_furnace_slag * superplasticizer + blast_furnace_slag *
##     coarse_agg + blast_furnace_slag * fine_agg + blast_furnace_slag *
##     age + fly_ash * water + fly_ash * superplasticizer + fly_ash *
##     coarse_agg + fly_ash * fine_agg + fly_ash * age + water *
##     superplasticizer + water * coarse_agg + water * fine_agg +
##     water * age + superplasticizer * coarse_agg + superplasticizer *
##     fine_agg + superplasticizer * age + coarse_agg * fine_agg +
##     coarse_agg * age + fine_agg * age + I(cement^2) +
I(blast_furnace_slag^2) +
##     I(fly_ash^2) + I(water^2) + I(superplasticizer^2) + I(coarse_agg^2) +
##     I(fine_agg^2) + I(age^2) + I(cement^3) + I(blast_furnace_slag^3) +
##     I(fly_ash^3) + I(water^3) + I(superplasticizer^3) + I(coarse_agg^3) +
```

```
##     I(fine_agg^3) + I(age^3) + I(sqrt(cement)) +
I(sqrt(blast_furnace_slag)) +
##     I(sqrt(fly_ash)) + I(sqrt(water)) + I(sqrt(superplasticizer)) +
##     I(sqrt(coarse_agg)) + I(sqrt(fine_agg)) + I(sqrt(age)) +
##     I(log(cement + 1)) + I(log(blast_furnace_slag + 1)) + I(log(fly_ash +
##     1)) + I(log(water + 1)) + I(log(superplasticizer + 1)) +
##     I(log(coarse_agg)) + I(log(fine_agg + 1)) + I(log(age)),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.3310  -3.1586  -0.2125   3.2715  19.8164
##
## Coefficients: (2 not defined because of singularities)
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         1.083e+05  1.157e+05   0.936 0.349417
## cement                             -4.543e-01  1.022e+01  -0.044 0.964558
## blast_furnace_slag                 -1.395e+00  1.012e+00  -1.379 0.168274
## fly_ash                             1.852e+01  6.210e+00   2.982 0.002955
**
## water                              -9.205e+02  1.346e+03  -0.684 0.494219
## superplasticizer                    1.142e+01  7.738e+00   1.475 0.140553
## coarse_agg                          2.230e+02  9.456e+01   2.358 0.018609
*
## fine_agg                           -1.167e+02  3.003e+01  -3.888 0.000110
***
## age                                -6.768e-01  4.612e-01  -1.467 0.142697
## I(cement^2)                        -1.629e-03  6.295e-03  -0.259 0.795873
## I(blast_furnace_slag^2)             4.894e-03  1.379e-03   3.548 0.000412
***
## I(fly_ash^2)                       -3.607e-02  1.559e-02  -2.313 0.020989
*
## I(water^2)                          8.842e-01  1.246e+00   0.710 0.478077
## I(superplasticizer^2)              -6.896e-02  9.296e-02  -0.742 0.458421
## I(coarse_agg^2)                    -7.772e-02  3.278e-02  -2.371 0.017973
*
## I(fine_agg^2)                       4.945e-02  1.312e-02   3.769 0.000177
***
## I(age^2)                            6.617e-04  7.383e-04   0.896 0.370405
## I(cement^3)                         1.963e-06  2.942e-06   0.667 0.504863
## I(blast_furnace_slag^3)            -5.732e-06  1.648e-06  -3.478 0.000534
***
## I(fly_ash^3)                        4.946e-05  2.630e-05   1.880 0.060432
.
## I(water^3)                         -6.683e-04  9.192e-04  -0.727 0.467406
## I(superplasticizer^3)               1.115e-03  1.574e-03   0.709 0.478722
## I(coarse_agg^3)                     1.618e-05  6.792e-06   2.382 0.017469
*
## I(fine_agg^3)                      -1.245e-05  3.418e-06  -3.642 0.000289
***
```

```
## I(age^3)                          -5.165e-07  9.842e-07  -0.525 0.599914
## I(sqrt(cement))                    1.568e+02  3.494e+02   0.449 0.653664
## I(sqrt(blast_furnace_slag))        4.895e+01  1.192e+01   4.106 4.46e-05
***
## I(sqrt(fly_ash))                  -2.757e+02  9.588e+01  -2.875 0.004151
**
## I(sqrt(water))                     2.611e+04  3.864e+04   0.676 0.499402
## I(sqrt(superplasticizer))         -2.926e+00  2.494e+01  -0.117 0.906633
## I(sqrt(coarse_agg))               -7.245e+03  3.124e+03  -2.319 0.020638
*
## I(sqrt(fine_agg))                  3.577e+03  8.888e+02   4.025 6.27e-05
***
## I(sqrt(age))                       5.565e+00  4.281e+00   1.300 0.194030
## I(log(cement + 1))                -5.562e+02  8.720e+02  -0.638 0.523765
## I(log(blast_furnace_slag + 1))    -5.554e+01  1.329e+01  -4.180 3.25e-05
***
## I(log(fly_ash + 1))                3.045e+02  1.032e+02   2.951 0.003261
**
## I(log(water + 1))                 -5.387e+04  8.190e+04  -0.658 0.510904
## I(log(superplasticizer + 1))       9.638e+00  3.016e+01   0.320 0.749355
## I(log(coarse_agg))                        NA         NA      NA       NA
## I(log(fine_agg + 1))                      NA         NA      NA       NA
## I(log(age))                        4.204e+00  3.819e+00   1.101 0.271303
## cement:blast_furnace_slag         -4.578e-04  2.443e-04  -1.874 0.061379
.
## cement:fly_ash                    -5.506e-04  2.889e-04  -1.906 0.057038
.
## cement:water                      -3.214e-03  7.514e-04  -4.277 2.14e-05
***
## cement:superplasticizer           -4.988e-04  2.186e-03  -0.228 0.819580
## cement:coarse_agg                 -5.465e-04  2.343e-04  -2.332 0.019951
*
## cement:fine_agg                   -6.290e-04  2.665e-04  -2.360 0.018530
*
## cement:age                         1.434e-04  1.335e-04   1.074 0.283257
## blast_furnace_slag:fly_ash        -4.908e-04  3.513e-04  -1.397 0.162837
## blast_furnace_slag:water          -2.145e-03  9.028e-04  -2.376 0.017755
*
## blast_furnace_slag:superplasticizer  3.748e-03  2.642e-03   1.419 0.156429
## blast_furnace_slag:coarse_agg     -4.656e-04  2.569e-04  -1.812 0.070315
.
## blast_furnace_slag:fine_agg       -3.375e-04  3.054e-04  -1.105 0.269424
## blast_furnace_slag:age             3.724e-04  1.341e-04   2.778 0.005611
**
## fly_ash:water                     -3.643e-03  9.935e-04  -3.667 0.000262
***
## fly_ash:superplasticizer           2.549e-03  3.286e-03   0.776 0.438037
## fly_ash:coarse_agg                -6.138e-04  3.077e-04  -1.995 0.046444
*
## fly_ash:fine_agg                  -4.995e-04  3.530e-04  -1.415 0.157553
```

```
## fly_ash:age                             6.028e-04  2.310e-04   2.609 0.009251
**
## water:superplasticizer                 -3.471e-02  1.117e-02  -3.109 0.001949
**
## water:coarse_agg                       -4.044e-03  9.227e-04  -4.383 1.34e-05
***
## water:fine_agg                         -3.810e-03  1.074e-03  -3.547 0.000413
***
## water:age                               1.836e-05  6.344e-04   0.029 0.976919
## superplasticizer:coarse_agg            -3.135e-03  2.760e-03  -1.136 0.256390
## superplasticizer:fine_agg              -3.638e-03  2.858e-03  -1.273 0.203554
## superplasticizer:age                    8.391e-04  1.729e-03   0.485 0.627688
## coarse_agg:fine_agg                    -7.843e-04  3.088e-04  -2.540 0.011293
*
## coarse_agg:age                          1.170e-04  1.129e-04   1.037 0.300277
## fine_agg:age                            1.092e-04  1.555e-04   0.702 0.482785
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.292 on 757 degrees of freedom
## Multiple R-squared:  0.9075, Adjusted R-squared:  0.8994
## F-statistic: 112.5 on 66 and 757 DF,  p-value: < 2.2e-16
```

```
press_allt_int = sqrt(sum((resid(mlr_allt_int)/(1-
hatvalues(mlr_allt_int)))^2)/n)
press_allt_int
```

```
## [1] 5.685851
```

```
mse_all = calculate_mse_test(mlr_allt_int, test)
```

```
## Warning in predict.lm(model, newdata = test): prediction from a rank-
deficient
## fit may be misleading
```

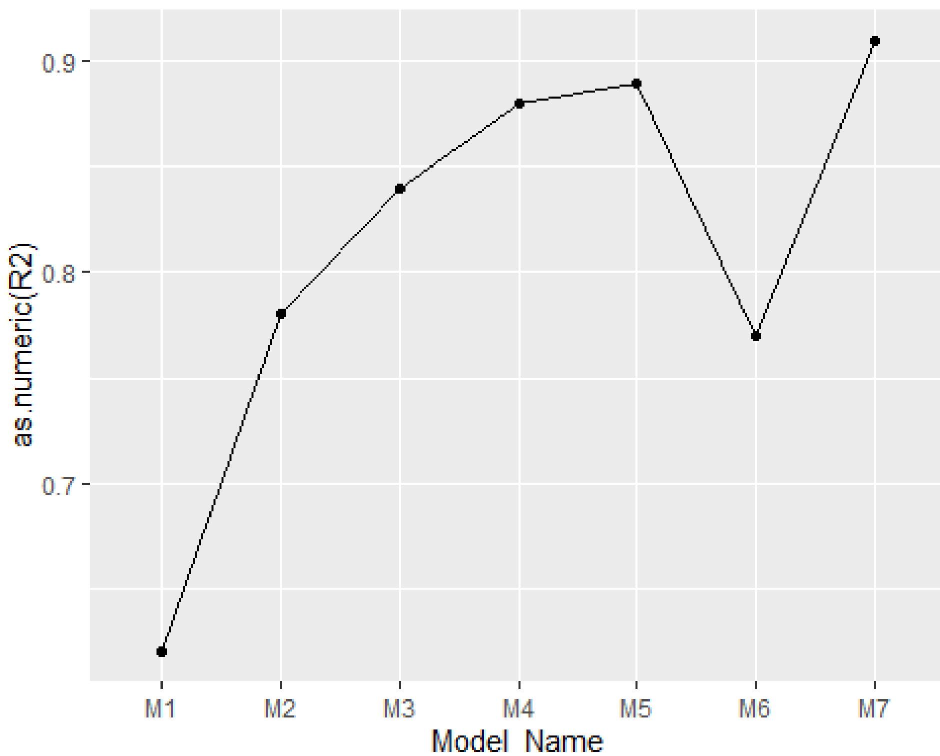
```
mse_all
```

```
## [1] 35.10078
```

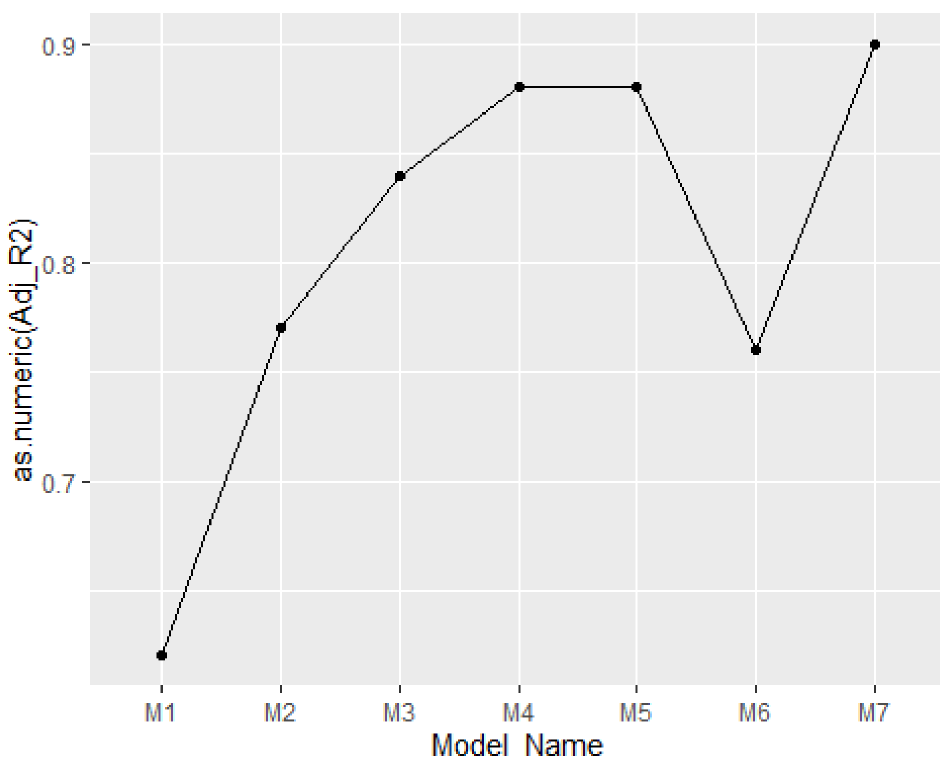
Plotting model metrics

```
tab <- matrix(c('M1', 'M2', 'M3', 'M4', 'M5', 'M6', 'M7',
                'Baseline', 'Baseline+Squared', 'Baseline+Squared+Cubic',
                'Baseline+Squared+Cubic+SquareRoot',
                'Baseline+Squared+Cubic+SquareRoot+Log' ,
                'Baseline+Interaction',
                'Baseline+Squared+Cubic+SquareRoot+Log+Interaction',
                mlr$rank-1, mlr_squared$rank-1, mlr_cubed$rank-1,
                mlr_sqrt$rank-1, mlr_log$rank-1,
                mlr_int$rank-1, mlr_allt_int$rank-1,
                summary(mlr)$r.squared, summary(mlr_squared)$r.squared,
                summary(mlr_cubed)$r.squared, summary(mlr_sqrt)$r.squared,
```

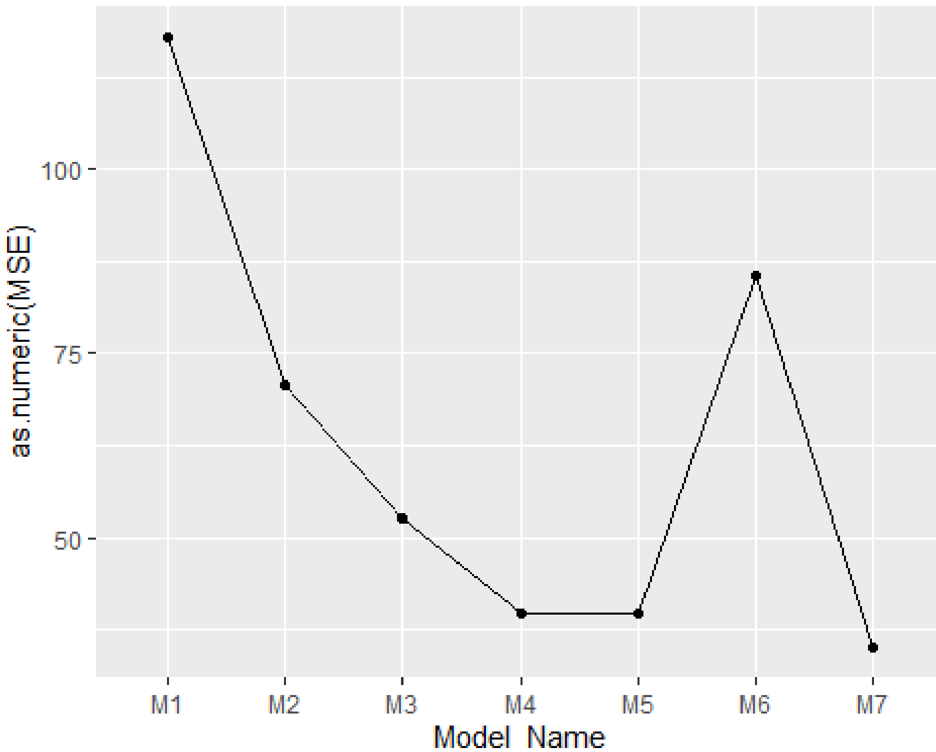
```
                summary(mlr_log)$r.squared,
                summary(mlr_int)$r.squared, summary(mlr_allt_int)$r.squared,
                summary(mlr)$adj.r.squared,
                summary(mlr_squared)$adj.r.squared,
                summary(mlr_cubed)$adj.r.squared,
                summary(mlr_sqrt)$adj.r.squared,
                summary(mlr_log)$adj.r.squared,
                summary(mlr_int)$adj.r.squared,
                summary(mlr_allt_int)$adj.r.squared,
                mse_mlr, mse_sq, mse_cub, mse_sqrt, mse_log, mse_int,
                mse_all,
                press_mlr, press_sq, press_cub, press_sqrt, press_log,
                press_int, press_allt_int
                ), ncol=7)
colnames(tab) <- c('Model_Name', 'Model_Description','No_of_predictors','R2',
'Adj_R2', 'MSE', 'PRESS')
tab <- as.table(tab)

metrics_df = as.data.frame.matrix(tab)

metrics_df$R2 = as.numeric(as.character(metrics_df$R2))
metrics_df$Adj_R2 = as.numeric(as.character(metrics_df$Adj_R2))
metrics_df$MSE = as.numeric(as.character(metrics_df$MSE))
metrics_df$PRESS = as.numeric(as.character(metrics_df$PRESS))

metrics_df = metrics_df %>% mutate(across(is.numeric, round, digits=2))

## Warning: Use of bare predicate functions was deprecated in tidyselect
1.1.0.
## ℹ Please use wrap predicates in `where()` instead.
##   # Was:
##   data %>% select(is.numeric)
##
##   # Now:
##   data %>% select(where(is.numeric))

ggplot(metrics_df, aes(x=Model_Name, y=as.numeric(R2))) +
geom_line(aes(group=1)) + geom_point()
```



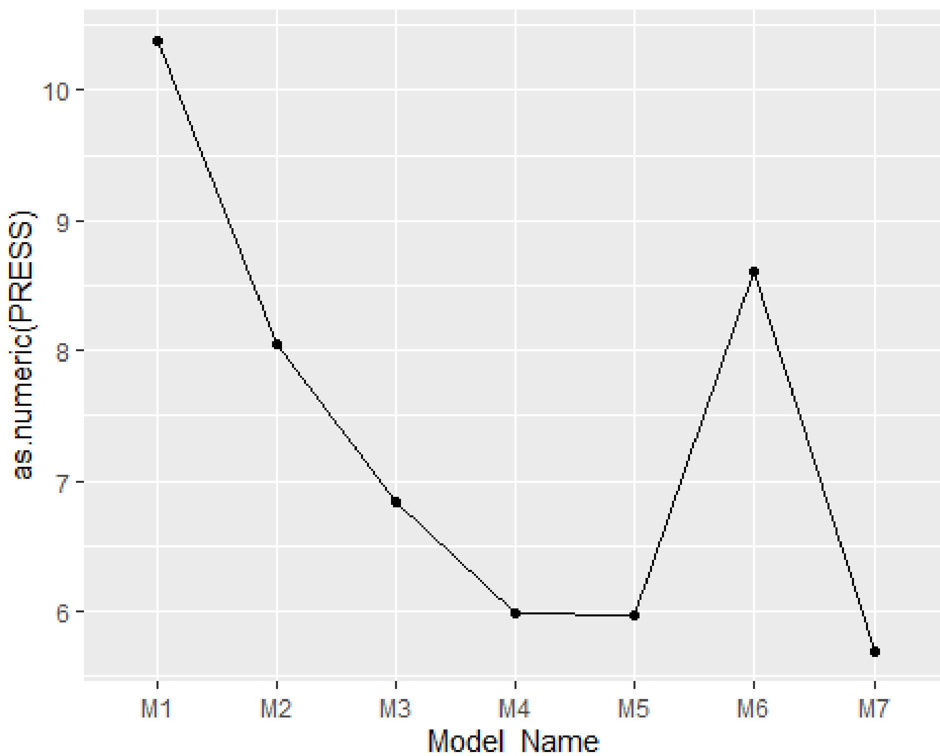

```
ggplot(metrics_df, aes(x=Model_Name, y=as.numeric(Adj_R2))) +
geom_line(aes(group=1)) + geom_point()
```



```
ggplot(metrics_df, aes(x=Model_Name, y=as.numeric(MSE))) +
geom_line(aes(group=1)) + geom_point()
```




```
ggplot(metrics_df, aes(x=Model_Name, y=as.numeric(PRESS))) +
geom_line(aes(group=1)) + geom_point()
```



Variable Reduction

We will now attempt to reduce the number of variables (model complexity) by penalizing predictors in the full model(both polynomial and interaction) in order to reduce overfitting.

Backward AIC Regression

```
fit_back_aic = step(mlr_allt_int, direction = "backward", trace = 0)
summary(fit_back_aic)

##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(blast_furnace_slag^2) + I(fly_ash^2) + I(coarse_agg^2) +
##     I(fine_agg^2) + I(age^2) + I(cement^3) + I(blast_furnace_slag^3) +
##     I(fly_ash^3) + I(coarse_agg^3) + I(fine_agg^3) + I(sqrt(cement)) +
##     I(sqrt(blast_furnace_slag)) + I(sqrt(fly_ash)) + I(sqrt(water)) +
##     I(sqrt(coarse_agg)) + I(sqrt(fine_agg)) + I(sqrt(age)) +
##     I(log(cement + 1)) + I(log(blast_furnace_slag + 1)) + I(log(fly_ash +
##     1)) + I(log(superplasticizer + 1)) + I(log(age)) +
cement:blast_furnace_slag +
##     cement:fly_ash + cement:water + cement:coarse_agg + cement:fine_agg +
##     cement:age + blast_furnace_slag:water +
blast_furnace_slag:superplasticizer +
##     blast_furnace_slag:coarse_agg + blast_furnace_slag:age +
##     fly_ash:water + fly_ash:superplasticizer + fly_ash:coarse_agg +
```

```
##     fly_ash:age + water:superplasticizer + water:coarse_agg +
##     water:fine_agg + superplasticizer:coarse_agg +
superplasticizer:fine_agg +
##     coarse_agg:fine_agg + coarse_agg:age + fine_agg:age, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.0618  -3.2247  -0.1762   3.1585  20.7852
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          3.062e+04  2.981e+04   1.027 0.304776
## cement                              -3.546e+00  8.097e-01  -4.379 1.35e-05
***
## blast_furnace_slag                  -2.346e+00  6.755e-01  -3.472 0.000545
***
## fly_ash                              1.759e+01  5.819e+00   3.023 0.002582
**
## water                                2.410e+00  4.651e-01   5.181 2.82e-07
***
## superplasticizer                     7.106e+00  2.616e+00   2.717 0.006745
**
## coarse_agg                           2.087e+02  9.049e+01   2.306 0.021364
*
## fine_agg                            -1.253e+02  2.610e+01  -4.800 1.91e-06
***
## age                                 -5.493e-01  1.494e-01  -3.676 0.000254
***
## I(blast_furnace_slag^2)              5.431e-03  1.307e-03   4.155 3.61e-05
***
## I(fly_ash^2)                        -3.640e-02  1.478e-02  -2.462 0.014022
*
## I(coarse_agg^2)                     -7.268e-02  3.134e-02  -2.319 0.020641
*
## I(fine_agg^2)                        5.304e-02  1.131e-02   4.691 3.21e-06
***
## I(age^2)                             2.718e-04  1.236e-04   2.200 0.028111
*
## I(cement^3)                          1.251e-06  2.180e-07   5.736 1.39e-08
***
## I(blast_furnace_slag^3)             -6.176e-06  1.574e-06  -3.924 9.48e-05
***
## I(fly_ash^3)                         5.089e-05  2.492e-05   2.042 0.041510
*
## I(coarse_agg^3)                      1.513e-05  6.490e-06   2.331 0.020011
*
## I(fine_agg^3)                       -1.335e-05  2.921e-06  -4.570 5.67e-06
***
## I(sqrt(cement))                      2.395e+02  4.271e+01   5.607 2.88e-08
***
```

```
## I(sqrt(blast_furnace_slag))          5.238e+01  1.140e+01   4.596 5.02e-06
***
## I(sqrt(fly_ash))                    -2.779e+02  9.063e+01  -3.067 0.002238
**
## I(sqrt(water))                       1.248e+02  1.859e+01   6.712 3.71e-11
***
## I(sqrt(coarse_agg))                 -6.808e+03  2.989e+03  -2.277 0.023035
*
## I(sqrt(fine_agg))                    3.801e+03  7.698e+02   4.938 9.68e-07
***
## I(sqrt(age))                         3.568e+00  2.308e+00   1.546 0.122611
## I(log(cement + 1))                  -7.621e+02  1.433e+02  -5.318 1.38e-07
***
## I(log(blast_furnace_slag + 1))      -5.940e+01  1.267e+01  -4.687 3.27e-06
***
## I(log(fly_ash + 1))                  3.066e+02  9.746e+01   3.146 0.001720
**
## I(log(superplasticizer + 1))         9.198e+00  1.393e+00   6.601 7.59e-11
***
## I(log(age))                          5.846e+00  2.425e+00   2.411 0.016140
*
## cement:blast_furnace_slag           -1.589e-04  7.047e-05  -2.255 0.024398
*
## cement:fly_ash                      -1.422e-04  8.876e-05  -1.602 0.109567
## cement:water                        -2.845e-03  2.689e-04 -10.580  < 2e-16
***
## cement:coarse_agg                   -2.730e-04  1.111e-04  -2.457 0.014243
*
## cement:fine_agg                     -3.334e-04  7.295e-05  -4.571 5.66e-06
***
## cement:age                           1.785e-04  7.482e-05   2.385 0.017323
*
## blast_furnace_slag:water            -1.738e-03  4.060e-04  -4.281 2.10e-05
***
## blast_furnace_slag:superplasticizer  3.947e-03  8.374e-04   4.713 2.90e-06
***
## blast_furnace_slag:coarse_agg       -1.749e-04  1.006e-04  -1.738 0.082673
.
## blast_furnace_slag:age               4.136e-04  8.192e-05   5.049 5.54e-07
***
## fly_ash:water                       -2.998e-03  4.863e-04  -6.165 1.13e-09
***
## fly_ash:superplasticizer             3.473e-03  1.240e-03   2.800 0.005239
**
## fly_ash:coarse_agg                  -1.954e-04  1.311e-04  -1.491 0.136492
## fly_ash:age                          6.551e-04  1.407e-04   4.656 3.80e-06
***
## water:superplasticizer              -2.926e-02  5.500e-03  -5.320 1.36e-07
***
## water:coarse_agg                    -3.466e-03  4.784e-04  -7.244 1.06e-12
```

```
***
## water:fine_agg                        -3.294e-03  4.297e-04  -7.665 5.39e-14
***
## superplasticizer:coarse_agg           -1.940e-03  1.293e-03  -1.501 0.133821
## superplasticizer:fine_agg             -2.824e-03  1.261e-03  -2.240 0.025373
*
## coarse_agg:fine_agg                   -4.490e-04  1.336e-04  -3.361 0.000814
***
## coarse_agg:age                         1.138e-04  6.621e-05   1.718 0.086164
.
## fine_agg:age                           1.296e-04  6.087e-05   2.129 0.033596
*
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.266 on 771 degrees of freedom
## Multiple R-squared:  0.9067, Adjusted R-squared:  0.9004
## F-statistic: 144.1 on 52 and 771 DF,  p-value: < 2.2e-16
```

```
press_fit_back_aic = sqrt(sum((resid(fit_back_aic)/(1-
hatvalues(fit_back_aic)))^2)/n)
press_fit_back_aic
```

```
## [1] 5.541012
```

```
mse_fbaic = calculate_mse_test(fit_back_aic, test)
mse_fbaic
```

```
## [1] 35.28844
```

Backward BIC Regression

```
n = nrow(train)
fit_back_bic = step(mlr_allt_int, direction = "backward", k=log(n), trace =
0)
summary(fit_back_bic)
```

```
##
## Call:
## lm(formula = concrete_strength ~ cement + blast_furnace_slag +
##     fly_ash + water + superplasticizer + coarse_agg + fine_agg +
##     age + I(blast_furnace_slag^2) + I(fly_ash^2) + I(fine_agg^2) +
##     I(cement^3) + I(blast_furnace_slag^3) + I(fine_agg^3) +
I(sqrt(cement)) +
##     I(sqrt(blast_furnace_slag)) + I(sqrt(fly_ash)) + I(sqrt(water)) +
##     I(sqrt(fine_agg)) + I(log(cement + 1)) + I(log(blast_furnace_slag +
##     1)) + I(log(fly_ash + 1)) + I(log(superplasticizer + 1)) +
##     I(log(age)) + cement:water + cement:fine_agg +
blast_furnace_slag:water +
##     blast_furnace_slag:superplasticizer + blast_furnace_slag:age +
##     fly_ash:water + fly_ash:superplasticizer + fly_ash:age +
##     water:superplasticizer + water:coarse_agg + water:fine_agg,
```

```
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.1432  -3.3149  -0.0804   3.3689  20.6600
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -3.323e+04  6.323e+03  -5.254 1.91e-07
***
## cement                              -3.679e+00  7.969e-01  -4.617 4.55e-06
***
## blast_furnace_slag                  -2.461e+00  6.601e-01  -3.728 0.000207
***
## fly_ash                              5.218e+00  8.176e-01   6.383 2.97e-10
***
## water                                1.036e+00  3.017e-01   3.434 0.000626
***
## superplasticizer                     2.162e+00  4.786e-01   4.517 7.23e-06
***
## coarse_agg                           4.572e-01  4.790e-02   9.545  < 2e-16
***
## fine_agg                            -1.248e+02  2.480e+01  -5.030 6.07e-07
***
## age                                 -5.412e-02  5.810e-03  -9.315  < 2e-16
***
## I(blast_furnace_slag^2)              5.476e-03  1.279e-03   4.281 2.09e-05
***
## I(fly_ash^2)                        -5.827e-03  1.063e-03  -5.483 5.63e-08
***
## I(fine_agg^2)                        5.276e-02  1.075e-02   4.908 1.12e-06
***
## I(cement^3)                          1.234e-06  2.136e-07   5.780 1.07e-08
***
## I(blast_furnace_slag^3)             -6.398e-06  1.531e-06  -4.178 3.27e-05
***
## I(fine_agg^3)                       -1.329e-05  2.777e-06  -4.785 2.05e-06
***
## I(sqrt(cement))                      2.244e+02  4.122e+01   5.443 7.02e-08
***
## I(sqrt(blast_furnace_slag))          5.152e+01  1.115e+01   4.619 4.50e-06
***
## I(sqrt(fly_ash))                    -8.628e+01  1.515e+01  -5.697 1.73e-08
***
## I(sqrt(water))                       1.114e+02  1.445e+01   7.711 3.78e-14
***
## I(sqrt(fine_agg))                    3.754e+03  7.320e+02   5.129 3.68e-07
***
## I(log(cement + 1))                  -7.212e+02  1.388e+02  -5.197 2.58e-07
***
```

```
## I(log(blast_furnace_slag + 1))      -5.826e+01  1.238e+01  -4.708 2.96e-06
***
## I(log(fly_ash + 1))                  9.963e+01  1.748e+01   5.701 1.68e-08
***
## I(log(superplasticizer + 1))         8.108e+00  9.552e-01   8.488  < 2e-16
***
## I(log(age))                          9.255e+00  2.832e-01  32.685  < 2e-16
***
## cement:water                        -2.640e-03  2.284e-04 -11.558  < 2e-16
***
## cement:fine_agg                     -1.751e-04  3.543e-05  -4.944 9.37e-07
***
## blast_furnace_slag:water            -2.081e-03  3.247e-04  -6.410 2.50e-10
***
## blast_furnace_slag:superplasticizer  3.243e-03  6.002e-04   5.403 8.66e-08
***
## blast_furnace_slag:age               2.338e-04  4.271e-05   5.473 5.95e-08
***
## fly_ash:water                       -3.090e-03  3.906e-04  -7.910 8.70e-15
***
## fly_ash:superplasticizer             2.725e-03  1.047e-03   2.604 0.009396
**
## fly_ash:age                          5.235e-04  1.062e-04   4.928 1.01e-06
***
## water:superplasticizer              -2.260e-02  3.567e-03  -6.335 3.98e-10
***
## water:coarse_agg                    -2.385e-03  2.588e-04  -9.213  < 2e-16
***
## water:fine_agg                      -2.336e-03  2.502e-04  -9.338  < 2e-16
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.357 on 788 degrees of freedom
## Multiple R-squared:  0.9013, Adjusted R-squared:  0.8969
## F-statistic: 205.6 on 35 and 788 DF,  p-value: < 2.2e-16
```

```
press_fit_back_bic = sqrt(sum((resid(fit_back_bic)/(1-
hatvalues(fit_back_bic)))^2)/n)
press_fit_back_bic
```

```
## [1] 5.505326
```

```
mse_fbbic = calculate_mse_test(fit_back_bic, test)
mse_fbbic
```

```
## [1] 36.41173
```

Forward AIC Regression

```
fit_null = lm(concrete_strength~1,data=train)
fit_forw_aic = step(fit_null,
```

```
                  scope = concrete_strength ~ cement + blast_furnace_slag +
                  fly_ash
                  + water + superplasticizer + coarse_agg + fine_agg + age +
                  cement*blast_furnace_slag + cement*fly_ash + cement*water
                  +
                  cement*superplasticizer + cement*coarse_agg
                  +cement*fine_agg +
                  cement*age + blast_furnace_slag*fly_ash +
                  blast_furnace_slag*fly_ash+
                  blast_furnace_slag*water +
                  blast_furnace_slag*superplasticizer +
                  blast_furnace_slag*coarse_agg +
                  blast_furnace_slag*fine_agg
                  +blast_furnace_slag*age + fly_ash*water +
                  fly_ash*superplasticizer +
                  fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                  water*superplasticizer + water*coarse_agg +
                  water*fine_agg+
                  water*age + superplasticizer*coarse_agg +
                  superplasticizer*fine_agg+
                  superplasticizer*age + coarse_agg*fine_agg +
                  coarse_agg*age
                  +fine_agg*age + I(cement^2) + I(blast_furnace_slag^2) +
                  I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                  I(coarse_agg^2) + I(fine_agg^2) + I(age^2) + I(cement^3) +
                  I(blast_furnace_slag^3) + I(fly_ash^3) + I(water^3) +
                  I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
                  I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
                  I(sqrt(fly_ash)) + I(sqrt(water)) +
                  I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                  I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                  I(log(blast_furnace_slag+1)) +
                  I(log(fly_ash+1)) + I(log(water+1)) +
                  I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                  I(log(fine_agg+1)) + I(log(age)),
                  direction = "forward", trace = 0)

summary(fit_forw_aic)

##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement +
I(log(superplasticizer +
##     1)) + blast_furnace_slag + water + superplasticizer + I(log(fly_ash +
##     1)) + I(age^2) + I(log(cement + 1)) + I(superplasticizer^2) +
##     I(superplasticizer^3) + I(cement^3) + I(sqrt(cement)) +
I(blast_furnace_slag^3) +
##     fly_ash + I(log(fine_agg + 1)) + I(sqrt(fine_agg)) +
I(log(coarse_agg)) +
##     I(age^3) + I(water^3) + I(log(water + 1)) + I(sqrt(water)) +
```

```
##     fine_agg + I(sqrt(age)) + I(cement^2) + I(sqrt(fly_ash)) +
##     I(fly_ash^2) + I(log(blast_furnace_slag + 1)) +
I(sqrt(blast_furnace_slag)) +
##     I(blast_furnace_slag^2) + I(fine_agg^2) + I(fine_agg^3) +
##     I(sqrt(coarse_agg)) + I(fly_ash^3) + water:superplasticizer +
##     blast_furnace_slag:superplasticizer + cement:water +
blast_furnace_slag:fly_ash +
##     water:fly_ash + cement:fine_agg + cement:blast_furnace_slag,
##     data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.778  -3.200   0.004   3.473  20.330
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -3.692e+06  7.596e+05  -4.861 1.41e-06
***
## I(log(age))                          7.446e+00  1.115e+00   6.675 4.68e-11
***
## cement                              -4.615e+00  9.331e+00  -0.495 0.621075
## I(log(superplasticizer + 1))         1.432e+00  2.769e+00   0.517 0.605048
## blast_furnace_slag                  -2.679e+00  6.603e-01  -4.057 5.47e-05
***
## water                                1.821e+02  4.896e+01   3.720 0.000213
***
## superplasticizer                     1.891e+00  1.565e+00   1.208 0.227386
## I(log(fly_ash + 1))                  2.893e+02  9.898e+01   2.923 0.003566
**
## I(age^2)                            -6.462e-04  1.762e-04  -3.667 0.000262
***
## I(log(cement + 1))                  -7.070e+02  7.991e+02  -0.885 0.376618
## I(superplasticizer^2)               -1.045e-01  6.167e-02  -1.694 0.090678
.
## I(superplasticizer^3)                1.966e-03  1.096e-03   1.794 0.073157
.
## I(cement^3)                          6.882e-07  2.713e-06   0.254 0.799830
## I(sqrt(cement))                      2.326e+02  3.203e+02   0.726 0.468057
## I(blast_furnace_slag^3)             -6.037e-06  1.576e-06  -3.830 0.000139
***
## fly_ash                              1.621e+01  5.911e+00   2.742 0.006253
**
## I(log(fine_agg + 1))                 1.288e+06  2.690e+05   4.789 2.01e-06
***
## I(sqrt(fine_agg))                   -2.951e+05  6.200e+04  -4.759 2.31e-06
***
## I(log(coarse_agg))                   3.340e+02  1.218e+02   2.743 0.006230
**
## I(age^3)                             1.271e-06  3.786e-07   3.356 0.000829
***
```

```
## I(water^3)                         -9.982e-05  3.346e-05  -2.984 0.002937
**
## I(log(water + 1))                   2.374e+04  5.874e+03   4.042 5.83e-05
***
## I(sqrt(water))                     -8.145e+03  2.098e+03  -3.882 0.000112
***
## fine_agg                            4.958e+03  1.048e+03   4.731 2.65e-06
***
## I(sqrt(age))                        1.185e+00  6.222e-01   1.904 0.057225
.
## I(cement^2)                         7.724e-04  5.779e-03   0.134 0.893715
## I(sqrt(fly_ash))                   -2.628e+02  9.214e+01  -2.852 0.004453
**
## I(fly_ash^2)                       -3.451e-02  1.506e-02  -2.291 0.022237
*
## I(log(blast_furnace_slag + 1))     -5.728e+01  1.241e+01  -4.614 4.62e-06
***
## I(sqrt(blast_furnace_slag))         5.011e+01  1.120e+01   4.474 8.82e-06
***
## I(blast_furnace_slag^2)             5.094e-03  1.305e-03   3.904 0.000103
***
## I(fine_agg^2)                      -1.060e+00  2.266e-01  -4.678 3.41e-06
***
## I(fine_agg^3)                       1.806e-04  3.902e-05   4.629 4.29e-06
***
## I(sqrt(coarse_agg))                -2.067e+01  7.849e+00  -2.634 0.008609
**
## I(fly_ash^3)                        4.771e-05  2.541e-05   1.878 0.060785
.
## water:superplasticizer             -7.676e-03  4.369e-03  -1.757 0.079287
.
## blast_furnace_slag:superplasticizer  3.083e-03  6.548e-04   4.708 2.96e-06
***
## cement:water                       -8.128e-04  1.624e-04  -5.007 6.85e-07
***
## blast_furnace_slag:fly_ash         -8.529e-05  7.336e-05  -1.163 0.245371
## water:fly_ash                      -1.215e-03  2.752e-04  -4.413 1.16e-05
***
## cement:fine_agg                    -1.723e-04  3.613e-05  -4.768 2.22e-06
***
## cement:blast_furnace_slag          -6.744e-05  4.668e-05  -1.445 0.148916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.459 on 782 degrees of freedom
## Multiple R-squared:  0.8983, Adjusted R-squared:  0.8929
## F-statistic: 168.4 on 41 and 782 DF,  p-value: < 2.2e-16
```

```
press_fit_forw_aic = sqrt(sum((resid(fit_forw_aic)/(1-
hatvalues(fit_forw_aic)))^2)/n)
press_fit_forw_aic

## [1] 5.64857

mse_ffaic = calculate_mse_test(fit_forw_aic, test)
mse_ffaic

## [1] 36.34937
```

Forward BIC Regression

```
fit_null = lm(concrete_strength~1,data=train)
fit_forw_bic = step(fit_null,
                  scope = concrete_strength ~ cement + blast_furnace_slag +
                  fly_ash
                  + water + superplasticizer + coarse_agg + fine_agg + age +
                  cement*blast_furnace_slag + cement*fly_ash + cement*water +
                  cement*superplasticizer + cement*coarse_agg
                  +cement*fine_agg +
                  cement*age + blast_furnace_slag*fly_ash +
                  blast_furnace_slag*fly_ash+
                  blast_furnace_slag*water +
                  blast_furnace_slag*superplasticizer +
                  blast_furnace_slag*coarse_agg + blast_furnace_slag*fine_agg
                  +blast_furnace_slag*age + fly_ash*water +
                  fly_ash*superplasticizer +
                  fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                  water*superplasticizer + water*coarse_agg + water*fine_agg+
                  water*age + superplasticizer*coarse_agg +
                  superplasticizer*fine_agg+
                  superplasticizer*age + coarse_agg*fine_agg + coarse_agg*age
                  +fine_agg*age + I(cement^2) + I(blast_furnace_slag^2) +
                  I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                  I(coarse_agg^2) + I(fine_agg^2) + I(age^2) + I(cement^3) +
                  I(blast_furnace_slag^3) + I(fly_ash^3) + I(water^3) +
                  I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
                  I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
                  I(sqrt(fly_ash)) + I(sqrt(water)) +
                  I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                  I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                  I(log(blast_furnace_slag+1)) +
                  I(log(fly_ash+1)) + I(log(water+1)) +
                  I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                  I(log(fine_agg+1)) + I(log(age)),
                  direction = "forward", k = log(n), trace = 0)

summary(fit_forw_bic)

##
## Call:
```

```
## lm(formula = concrete_strength ~ I(log(age)) + cement +
I(log(superplasticizer +
##     1)) + blast_furnace_slag + water + superplasticizer + I(log(fly_ash +
##     1)) + I(age^2) + I(log(cement + 1)) + I(superplasticizer^2) +
##     I(superplasticizer^3) + I(cement^3) + I(sqrt(cement)) +
I(blast_furnace_slag^3) +
##     fly_ash + I(log(fine_agg + 1)) + I(sqrt(fine_agg)) +
I(log(coarse_agg)) +
##     water:superplasticizer + blast_furnace_slag:superplasticizer +
##     cement:water + blast_furnace_slag:fly_ash + water:fly_ash,
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -26.9007  -4.0294  -0.3753   3.8406  17.7254
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           1.083e+02  4.053e+02   0.267 0.789310
## I(log(age))                           9.146e+00  2.146e-01  42.611  < 2e-16
***
## cement                               -4.952e+00  8.263e-01  -5.993 3.12e-09
***
## I(log(superplasticizer + 1))         -6.236e+00  2.825e+00  -2.207 0.027570
*
## blast_furnace_slag                    1.161e-01  9.438e-03  12.297  < 2e-16
***
## water                                 3.458e-01  6.438e-02   5.371 1.03e-07
***
## superplasticizer                      7.486e+00  1.480e+00   5.059 5.22e-07
***
## I(log(fly_ash + 1))                   6.640e-01  4.188e-01   1.585 0.113300
## I(age^2)                             -8.857e-05  1.354e-05  -6.541 1.09e-10
***
## I(log(cement + 1))                   -8.314e+02  1.441e+02  -5.769 1.14e-08
***
## I(superplasticizer^2)                -2.415e-01  6.419e-02  -3.762 0.000181
***
## I(superplasticizer^3)                 3.445e-03  1.162e-03   2.965 0.003121
**
## I(cement^3)                           1.573e-06  2.249e-07   6.992 5.71e-12
***
## I(sqrt(cement))                       2.614e+02  4.295e+01   6.087 1.78e-09
***
## I(blast_furnace_slag^3)              -3.163e-07  8.519e-08  -3.713 0.000219
***
## fly_ash                               2.831e-01  5.256e-02   5.387 9.45e-08
***
## I(log(fine_agg + 1))                  2.832e+02  6.643e+01   4.263 2.26e-05
***
```

```
## I(sqrt(fine_agg))                    -1.860e+01  4.862e+00  -3.826 0.000140
***
## I(log(coarse_agg))                    2.381e+01  6.447e+00   3.693 0.000236
***
## water:superplasticizer               -2.137e-02  3.507e-03  -6.095 1.70e-09
***
## blast_furnace_slag:superplasticizer   4.591e-03  6.316e-04   7.269 8.65e-13
***
## cement:water                         -8.508e-04  1.412e-04  -6.027 2.55e-09
***
## blast_furnace_slag:fly_ash           -1.725e-04  6.108e-05  -2.824 0.004866
**
## water:fly_ash                        -1.086e-03  2.693e-04  -4.034 6.00e-05
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.044 on 800 degrees of freedom
## Multiple R-squared:  0.8724, Adjusted R-squared:  0.8687
## F-statistic: 237.8 on 23 and 800 DF,  p-value: < 2.2e-16
```

```
press_fit_forw_bic = sqrt(sum((resid(fit_forw_bic)/(1-
hatvalues(fit_forw_bic)))^2)/n)
press_fit_forw_bic
```

```
## [1] 6.164206
```

```
mse_ffbic = calculate_mse_test(fit_forw_bic, test)
mse_ffbic
```

```
## [1] 47.26784
```

```
fit_null = lm(concrete_strength~1,data=train)
fit_step_aic = step(fit_null,
                  scope = concrete_strength ~ cement + blast_furnace_slag +
                  fly_ash
                  + water + superplasticizer + coarse_agg + fine_agg + age +
                  cement*blast_furnace_slag + cement*fly_ash + cement*water +
                  cement*superplasticizer + cement*coarse_agg
                  +cement*fine_agg +
                  cement*age + blast_furnace_slag*fly_ash +
                  blast_furnace_slag*fly_ash+
                  blast_furnace_slag*water +
                  blast_furnace_slag*superplasticizer +
                  blast_furnace_slag*coarse_agg + blast_furnace_slag*fine_agg
                  +blast_furnace_slag*age + fly_ash*water +
                  fly_ash*superplasticizer +
                  fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                  water*superplasticizer + water*coarse_agg + water*fine_agg+
                  water*age + superplasticizer*coarse_agg +
                  superplasticizer*fine_agg+
```

```
                     superplasticizer*age + coarse_agg*fine_agg + coarse_agg*age
                     +fine_agg*age + I(cement^2) + I(blast_furnace_slag^2) +
                     I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                     I(coarse_agg^2) + I(fine_agg^2) + I(age^2) + I(cement^3) +
                     I(blast_furnace_slag^3) + I(fly_ash^3) + I(water^3) +
                     I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
                     I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
                     I(sqrt(fly_ash)) + I(sqrt(water)) +
                     I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                     I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                     I(log(blast_furnace_slag+1)) +
                     I(log(fly_ash+1)) + I(log(water+1)) +
                     I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                     I(log(fine_agg+1)) + I(log(age)),
                     direction = "both", trace = 0)

summary(fit_step_aic)

##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement + blast_furnace_slag
+
##     water + superplasticizer + I(age^2) + I(superplasticizer^2) +
##     I(superplasticizer^3) + I(cement^3) + I(sqrt(cement)) +
I(blast_furnace_slag^3) +
##     fly_ash + I(log(fine_agg + 1)) + I(sqrt(fine_agg)) +
I(log(coarse_agg)) +
##     I(age^3) + I(water^3) + I(log(water + 1)) + I(sqrt(water)) +
##     fine_agg + I(sqrt(age)) + I(cement^2) + water:superplasticizer +
##     blast_furnace_slag:superplasticizer + cement:water + water:fly_ash +
##     cement:fine_agg + water:fine_agg, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.3288  -3.7358   0.0099   3.7636  19.4723
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -7.612e+04  1.249e+04  -6.094 1.71e-09
***
## I(log(age))                          6.412e+00  1.148e+00   5.585 3.22e-08
***
## cement                               4.673e+00  8.475e-01   5.514 4.75e-08
***
## blast_furnace_slag                   1.059e-01  9.226e-03  11.476  < 2e-16
***
## water                                2.066e+02  4.605e+01   4.486 8.35e-06
***
## superplasticizer                     2.790e+00  7.793e-01   3.580 0.000364
***
```

```
## I(age^2)                              -7.104e-04  1.832e-04  -3.879 0.000114
***
## I(superplasticizer^2)                 -1.374e-01  2.279e-02  -6.028 2.53e-09
***
## I(superplasticizer^3)                  2.467e-03  5.488e-04   4.496 7.95e-06
***
## I(cement^3)                            4.058e-06  6.767e-07   5.997 3.05e-09
***
## I(sqrt(cement))                       -6.931e+01  1.496e+01  -4.632 4.22e-06
***
## I(blast_furnace_slag^3)               -2.061e-07  7.811e-08  -2.638 0.008498
**
## fly_ash                                2.730e-01  4.955e-02   5.510 4.86e-08
***
## I(log(fine_agg + 1))                   6.871e+03  1.402e+03   4.901 1.16e-06
***
## I(sqrt(fine_agg))                     -9.614e+02  2.037e+02  -4.719 2.80e-06
***
## I(log(coarse_agg))                     1.991e+01  6.246e+00   3.188 0.001488
**
## I(age^3)                               1.369e-06  3.939e-07   3.475 0.000538
***
## I(water^3)                            -1.212e-04  3.184e-05  -3.807 0.000151
***
## I(log(water + 1))                      2.626e+04  5.491e+03   4.781 2.08e-06
***
## I(sqrt(water))                        -9.110e+03  1.967e+03  -4.631 4.26e-06
***
## fine_agg                               8.523e+00  1.848e+00   4.611 4.67e-06
***
## I(sqrt(age))                           1.710e+00  6.428e-01   2.660 0.007960
**
## I(cement^2)                           -5.694e-03  1.009e-03  -5.641 2.35e-08
***
## water:superplasticizer                -9.069e-03  3.857e-03  -2.351 0.018954
*
## blast_furnace_slag:superplasticizer  3.680e-03  5.526e-04   6.661 5.09e-11
***
## cement:water                          -7.271e-04  1.653e-04  -4.400 1.23e-05
***
## water:fly_ash                         -1.075e-03  2.611e-04  -4.119 4.21e-05
***
## cement:fine_agg                       -1.509e-04  3.581e-05  -4.213 2.81e-05
***
## water:fine_agg                        -3.246e-04  1.966e-04  -1.652 0.099016
.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.707 on 795 degrees of freedom
```

```
## Multiple R-squared:  0.887,  Adjusted R-squared:  0.883
## F-statistic: 222.8 on 28 and 795 DF,  p-value: < 2.2e-16
```

```
press_fit_step_aic = sqrt(sum((resid(fit_step_aic)/(1-
hatvalues(fit_step_aic)))^2)/n)
press_fit_step_aic
```

```
## [1] 5.841791
```

```
mse_fsaic = calculate_mse_test(fit_step_aic, test)
mse_fsaic
```

```
## [1] 39.06844
```

```
fit_null = lm(concrete_strength~1,data=train)
fit_step_bic = step(fit_null,
                  scope = concrete_strength ~ cement + blast_furnace_slag +
                  fly_ash
                  + water + superplasticizer + coarse_agg + fine_agg + age +
                  cement*blast_furnace_slag + cement*fly_ash + cement*water +
                  cement*superplasticizer + cement*coarse_agg
                  +cement*fine_agg +
                  cement*age + blast_furnace_slag*fly_ash +
                  blast_furnace_slag*fly_ash+
                  blast_furnace_slag*water +
                  blast_furnace_slag*superplasticizer +
                  blast_furnace_slag*coarse_agg + blast_furnace_slag*fine_agg
                  +blast_furnace_slag*age + fly_ash*water +
                  fly_ash*superplasticizer +
                  fly_ash*coarse_agg + fly_ash*fine_agg + fly_ash*age +
                  water*superplasticizer + water*coarse_agg + water*fine_agg+
                  water*age + superplasticizer*coarse_agg +
                  superplasticizer*fine_agg+
                  superplasticizer*age + coarse_agg*fine_agg + coarse_agg*age
                  +fine_agg*age + I(cement^2) + I(blast_furnace_slag^2) +
                  I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                  I(coarse_agg^2) + I(fine_agg^2) + I(age^2) + I(cement^3) +
                  I(blast_furnace_slag^3) + I(fly_ash^3) + I(water^3) +
                  I(superplasticizer^3) + I(coarse_agg^3) + I(fine_agg^3) +
                  I(age^3) + I(sqrt(cement)) + I(sqrt(blast_furnace_slag)) +
                  I(sqrt(fly_ash)) + I(sqrt(water)) +
                  I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                  I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                  I(log(blast_furnace_slag+1)) +
                  I(log(fly_ash+1)) + I(log(water+1)) +
                  I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                  I(log(fine_agg+1)) + I(log(age)),
                  direction = "both", trace = 0)

summary(fit_step_bic)
```

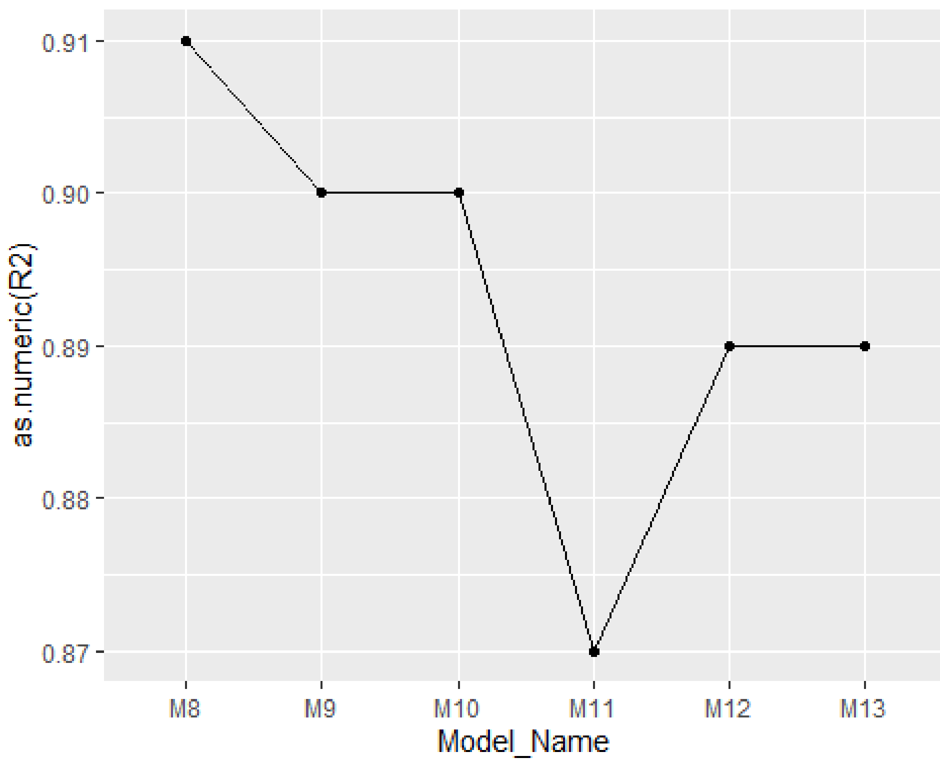
```
##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement + blast_furnace_slag
+
##     water + superplasticizer + I(age^2) + I(superplasticizer^2) +
##     I(superplasticizer^3) + I(cement^3) + I(sqrt(cement)) +
I(blast_furnace_slag^3) +
##     fly_ash + I(log(fine_agg + 1)) + I(sqrt(fine_agg)) +
I(log(coarse_agg)) +
##     I(age^3) + I(water^3) + I(log(water + 1)) + I(sqrt(water)) +
##     fine_agg + I(sqrt(age)) + I(cement^2) + water:superplasticizer +
##     blast_furnace_slag:superplasticizer + cement:water + water:fly_ash +
##     cement:fine_agg + water:fine_agg, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.3288  -3.7358   0.0099   3.7636  19.4723
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         -7.612e+04  1.249e+04  -6.094 1.71e-09
***
## I(log(age))                          6.412e+00  1.148e+00   5.585 3.22e-08
***
## cement                               4.673e+00  8.475e-01   5.514 4.75e-08
***
## blast_furnace_slag                   1.059e-01  9.226e-03  11.476  < 2e-16
***
## water                                2.066e+02  4.605e+01   4.486 8.35e-06
***
## superplasticizer                     2.790e+00  7.793e-01   3.580 0.000364
***
## I(age^2)                            -7.104e-04  1.832e-04  -3.879 0.000114
***
## I(superplasticizer^2)               -1.374e-01  2.279e-02  -6.028 2.53e-09
***
## I(superplasticizer^3)                2.467e-03  5.488e-04   4.496 7.95e-06
***
## I(cement^3)                          4.058e-06  6.767e-07   5.997 3.05e-09
***
## I(sqrt(cement))                     -6.931e+01  1.496e+01  -4.632 4.22e-06
***
## I(blast_furnace_slag^3)             -2.061e-07  7.811e-08  -2.638 0.008498
**
## fly_ash                              2.730e-01  4.955e-02   5.510 4.86e-08
***
## I(log(fine_agg + 1))                 6.871e+03  1.402e+03   4.901 1.16e-06
***
## I(sqrt(fine_agg))                   -9.614e+02  2.037e+02  -4.719 2.80e-06
***
```

```
## I(log(coarse_agg))                 1.991e+01  6.246e+00   3.188 0.001488
**
## I(age^3)                           1.369e-06  3.939e-07   3.475 0.000538
***
## I(water^3)                        -1.212e-04  3.184e-05  -3.807 0.000151
***
## I(log(water + 1))                  2.626e+04  5.491e+03   4.781 2.08e-06
***
## I(sqrt(water))                    -9.110e+03  1.967e+03  -4.631 4.26e-06
***
## fine_agg                           8.523e+00  1.848e+00   4.611 4.67e-06
***
## I(sqrt(age))                       1.710e+00  6.428e-01   2.660 0.007960
**
## I(cement^2)                       -5.694e-03  1.009e-03  -5.641 2.35e-08
***
## water:superplasticizer            -9.069e-03  3.857e-03  -2.351 0.018954
*
## blast_furnace_slag:superplasticizer  3.680e-03  5.526e-04   6.661 5.09e-11
***
## cement:water                      -7.271e-04  1.653e-04  -4.400 1.23e-05
***
## water:fly_ash                     -1.075e-03  2.611e-04  -4.119 4.21e-05
***
## cement:fine_agg                   -1.509e-04  3.581e-05  -4.213 2.81e-05
***
## water:fine_agg                    -3.246e-04  1.966e-04  -1.652 0.099016
.
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.707 on 795 degrees of freedom
## Multiple R-squared:  0.887,  Adjusted R-squared:  0.883
## F-statistic: 222.8 on 28 and 795 DF,  p-value: < 2.2e-16

press_fit_step_bic = sqrt(sum((resid(fit_step_bic)/(1-
hatvalues(fit_step_bic)))^2)/n)
press_fit_step_bic

## [1] 5.841791

mse_fsbic = calculate_mse_test(fit_step_bic, test)
mse_fsbic

## [1] 39.06844
```

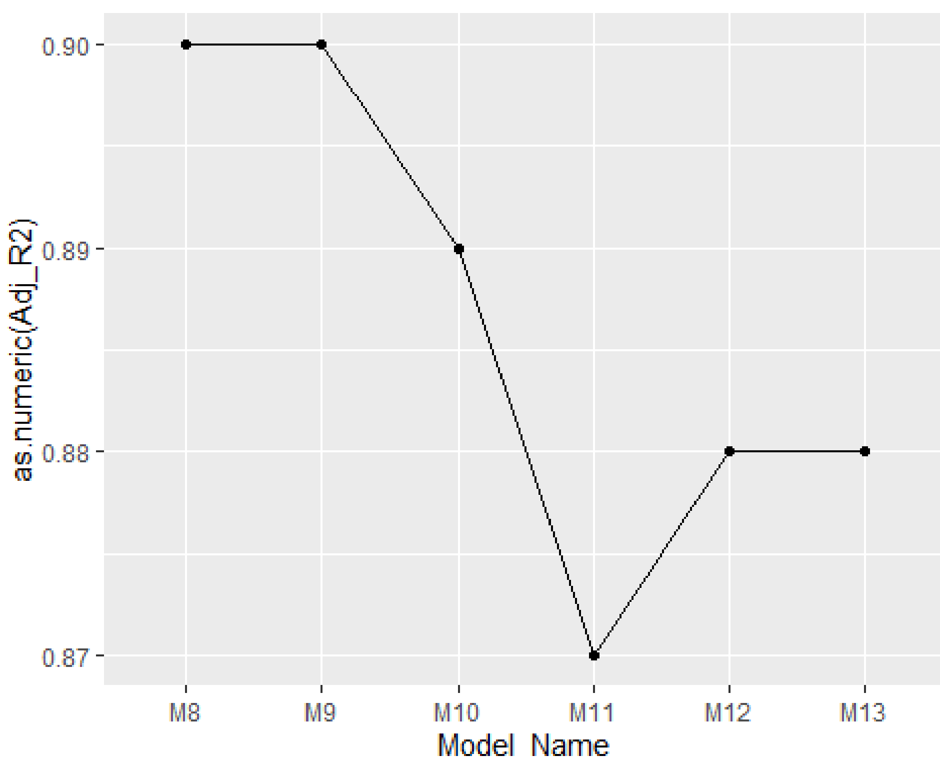
Plotting metrics for models after variable reduction

```
tab2 <- matrix(c('M8', 'M9', 'M10', 'M11', 'M12', 'M13',
                'FitBack_AIC', 'FitBack_BIC', 'FitForward_AIC',
                'FitForward_BIC',
                'FitStep_AIC' , 'FitStep_BIC',
```

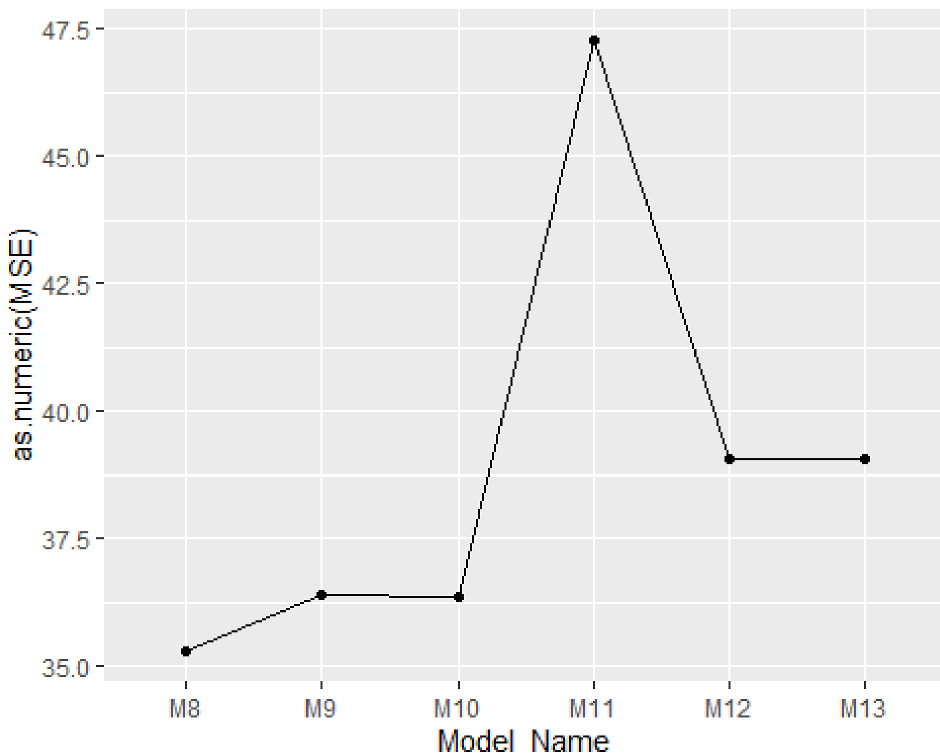
```
                fit_back_aic$rank-1, fit_back_bic$rank-1,
                fit_forw_aic$rank-1, fit_forw_bic$rank-1,
                fit_step_aic$rank-1,
                fit_step_bic$rank-1,
                summary(fit_back_aic)$r.squared,
                summary(fit_back_bic)$r.squared,
                summary(fit_forw_aic)$r.squared,
                summary(fit_forw_bic)$r.squared,
                summary(fit_step_aic)$r.squared,
                summary(fit_step_bic)$r.squared,
                summary(fit_back_aic)$adj.r.squared,
                summary(fit_back_bic)$adj.r.squared,
                summary(fit_forw_aic)$adj.r.squared,
                summary(fit_forw_bic)$adj.r.squared,
                summary(fit_step_aic)$adj.r.squared,
                summary(fit_step_bic)$adj.r.squared,
                mse_fbaic, mse_fbbic, mse_ffaic, mse_ffbic, mse_fsaic,
                mse_fsbic,
                press_fit_back_aic, press_fit_back_bic, press_fit_forw_aic,
                press_fit_forw_bic, press_fit_step_aic, press_fit_step_bic
                ), ncol=7)
colnames(tab2) <- c('Model_Name',
'Model_Description','No_of_predictors','R2', 'Adj_R2', 'MSE', 'PRESS')
tab2 <- as.table(tab2)

metrics_df2 = as.data.frame.matrix(tab2)

metrics_df2$R2 = as.numeric(as.character(metrics_df2$R2))
metrics_df2$Adj_R2 = as.numeric(as.character(metrics_df2$Adj_R2))
metrics_df2$MSE = as.numeric(as.character(metrics_df2$MSE))
metrics_df2$PRESS = as.numeric(as.character(metrics_df2$PRESS))

metrics_df2 = metrics_df2 %>% mutate(across(is.numeric, round, digits=2))

metrics_df2$Model_Name <- factor(metrics_df2$Model_Name, levels =
metrics_df2$Model_Name)

ggplot(metrics_df2, aes(x=Model_Name, y=as.numeric(R2))) +
geom_line(aes(group=1)) + geom_point()
```



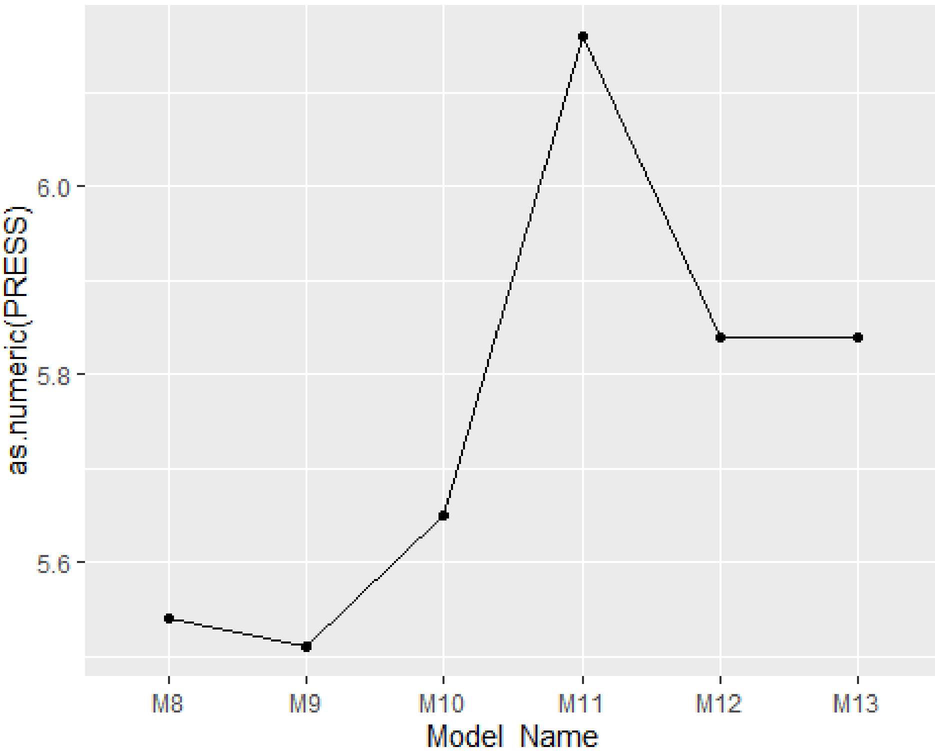

```
ggplot(metrics_df2, aes(x=Model_Name, y=as.numeric(Adj_R2))) +
geom_line(aes(group=1)) + geom_point()
```



```
ggplot(metrics_df2, aes(x=Model_Name, y=as.numeric(MSE))) +
geom_line(aes(group=1)) + geom_point()
```




```
ggplot(metrics_df2, aes(x=Model_Name, y=as.numeric(PRESS))) +
geom_line(aes(group=1)) + geom_point()
```



K Fold Cross Validation for the selected models

K-Fold cross validation is useful for understanding how well the model generalises on the data.

For this experiment, we will select some of the most promising models from the previous experiments and compare their performance over 5 fold cross validation.

Selected Models:
1. M1: Baseline Model
2. M7: Baseline+Squared+Cubic+SquareRoot+Log+Interaction Model
3. M10: Fit Forward AIC
4. M9: Fit Backward BIC
5. M12: Fit Step AIC
6. M13: Fit Step BIC

```
k=8

#m1 baseline M1
#m2 all M7
#m3 ff aic M10
#m4 fb bic M9
#m5 fs aic M12
```

```r
#m6 fs bic M13

RMSE_m1 = RMSE_m2 = RMSE_m3 = RMSE_m4 = RMSE_m5 = RMSE_m6 = numeric(k)

#Create k equally size folds
folds <- cut(1:n,breaks=k,labels=FALSE)

#Perform a k-fold cross validation
for(i in 1:k)
{
  # Find the indices for test data
  test_index = which(folds==i)

  # Obtain training/test data
  test_data = data[test_index, ]
  training_data = data[-test_index, ]


  model_1 = lm(concrete_strength ~ cement + blast_furnace_slag +
                 fly_ash + water + superplasticizer + coarse_agg +
                 fine_agg + age, data=training_data)

  model_2 = lm(concrete_strength ~ cement + blast_furnace_slag + fly_ash +
                 water + superplasticizer + coarse_agg + fine_agg + age +
                 cement:blast_furnace_slag + cement:fly_ash + cement:water +
                 cement:superplasticizer + cement:coarse_agg +
                 cement:fine_agg +
                 cement:age + blast_furnace_slag:fly_ash +
                 blast_furnace_slag:fly_ash+
                 blast_furnace_slag:water +
                 blast_furnace_slag:superplasticizer +
                 blast_furnace_slag:coarse_agg + blast_furnace_slag:fine_agg
                 +blast_furnace_slag:age + fly_ash:water +
                 fly_ash:superplasticizer +
                 fly_ash:coarse_agg + fly_ash:fine_agg + fly_ash:age +
                 water:superplasticizer + water:coarse_agg + water:fine_agg+
                 water:age + superplasticizer:coarse_agg +
                 superplasticizer:fine_agg+
                 superplasticizer:age + coarse_agg:fine_agg + coarse_agg:age
                 +fine_agg:age + I(cement^2) + I(blast_furnace_slag^2) +
                 I(fly_ash^2) + I(water^2) + I(superplasticizer^2) +
                 I(coarse_agg^2) +
                 I(fine_agg^2) + I(age^2) + I(cement^3) +
                 I(blast_furnace_slag^3) +
                 I(fly_ash^3) + I(water^3) + I(superplasticizer^3) +
                 I(coarse_agg^3) +
                 I(fine_agg^3) + I(age^3) + I(sqrt(cement)) +
                 I(sqrt(blast_furnace_slag)) +
                 I(sqrt(fly_ash)) + I(sqrt(water)) +
```

```r
                 I(sqrt(superplasticizer)) + I(sqrt(coarse_agg)) +
                 I(sqrt(fine_agg)) + I(sqrt(age)) + I(log(cement+1)) +
                 I(log(blast_furnace_slag+1)) +
                 I(log(fly_ash+1)) + I(log(water+1)) +
                 I(log(superplasticizer+1)) + I(log(coarse_agg)) +
                 I(log(fine_agg+1)) + I(log(age)), data=training_data)

  model_3 = lm(concrete_strength~ I(log(age))+ cement+ I(log(superplasticizer
+ 1))+ blast_furnace_slag+ water+ superplasticizer+
                 I(log(fly_ash + 1))+ I(age^2)+ I(log(cement + 1))+
                 I(superplasticizer^2)+
                 I(superplasticizer^3)+ I(cement^3)+ I(sqrt(cement))+
                 I(blast_furnace_slag^3)+
                 fly_ash+ I(log(fine_agg + 1))+ I(sqrt(fine_agg))+
                 I(log(coarse_agg))+
                 I(age^3)+ I(water^3)+ I(log(water + 1))+ I(sqrt(water))+
                 fine_agg+ I(sqrt(age))+ I(cement^2)+ I(sqrt(fly_ash))+
                 I(fly_ash^2)+
                 I(log(blast_furnace_slag + 1))+ I(sqrt(blast_furnace_slag))+
                 I(blast_furnace_slag^2)+ I(fine_agg^2)+ I(fine_agg^3)+
                 I(sqrt(coarse_agg))+
                 I(fly_ash^3) + water:superplasticizer +
                 blast_furnace_slag:superplasticizer +
                 cement:water + blast_furnace_slag:fly_ash + water:fly_ash +
                 cement:fine_agg +
                 cement:blast_furnace_slag, data=training_data)

  model_4 = lm(concrete_strength~ cement+ blast_furnace_slag+ fly_ash+
               water+ superplasticizer+ coarse_agg+ fine_agg+ age+
               I(blast_furnace_slag^2)+
               I(fly_ash^2)+ I(fine_agg^2)+ I(cement^3)+
               I(blast_furnace_slag^3)+
               I(fine_agg^3)+ I(sqrt(cement))+ I(sqrt(blast_furnace_slag))+
               I(sqrt(fly_ash))+ I(sqrt(water))+ I(sqrt(fine_agg))+
               I(log(cement + 1))+ I(log(blast_furnace_slag + 1))+
               I(log(fly_ash + 1))+ I(log(superplasticizer + 1))+ I(log(age))
+cement:water + cement:fine_agg + blast_furnace_slag:water +
               blast_furnace_slag:superplasticizer + blast_furnace_slag:age +
fly_ash:water + fly_ash:superplasticizer + fly_ash:age +
               water:superplasticizer + water:coarse_agg + water:fine_agg,
data=training_data)

  model_5 = lm(concrete_strength~ I(log(age))+ cement+ blast_furnace_slag+
               water+ superplasticizer+ I(age^2)+ I(superplasticizer^2)+
               I(superplasticizer^3)+ I(cement^3)+ I(sqrt(cement))+
               I(blast_furnace_slag^3)+
               fly_ash+ I(log(fine_agg + 1))+ I(sqrt(fine_agg))+
               I(log(coarse_agg))+
               I(age^3)+ I(water^3)+ I(log(water + 1))+ I(sqrt(water))+
               fine_agg+ I(sqrt(age))+ I(cement^2) + water:superplasticizer +
```

```
               blast_furnace_slag:superplasticizer + cement:water +
               water:fly_ash +
               cement:fine_agg + water:fine_agg, data=training_data)

  model_6 = lm(concrete_strength~ I(log(age))+ cement+ blast_furnace_slag+
               water+ superplasticizer+ I(log(fly_ash + 1))+ I(age^2)+
               I(log(cement + 1))+ I(superplasticizer^2)+
               I(superplasticizer^3) +
               water:superplasticizer + blast_furnace_slag:superplasticizer +
               cement:water, data=training_data)

  # Obtain RMSE on the 'test' data
  resid_m1 = test_data["concrete_strength"] - predict(model_1,
newdata=test_data)
  RMSE_m1[i] = sqrt(sum(resid_m1^2)/nrow(test_data))

  resid_m2 = test_data[,"concrete_strength"] - predict(model_2,
newdata=test_data)
  RMSE_m2[i] = sqrt(sum(resid_m2^2)/nrow(test_data))

  resid_m3 = test_data[,"concrete_strength"] - predict(model_3,
newdata=test_data)
  RMSE_m3[i] = sqrt(sum(resid_m3^2)/nrow(test_data))

  resid_m4 = test_data[,"concrete_strength"] - predict(model_4,
newdata=test_data)
  RMSE_m4[i] = sqrt(sum(resid_m4^2)/nrow(test_data))

  resid_m5 = test_data[,"concrete_strength"] - predict(model_5,
newdata=test_data)
  RMSE_m5[i] = sqrt(sum(resid_m5^2)/nrow(test_data))

  resid_m6 = test_data[,"concrete_strength"] - predict(model_6,
newdata=test_data)
  RMSE_m6[i] = sqrt(sum(resid_m6^2)/nrow(test_data))
}

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading
```

```
## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

## Warning in predict.lm(model_2, newdata = test_data): prediction from a
rank-
## deficient fit may be misleading

cat("Baseline model: ", mean(RMSE_m1))

## Baseline model:  11.8895

cat("Complete  model with both all interaction and polynomial terms: ",
mean(RMSE_m2))

## Complete  model with both all interaction and polynomial terms:  7.549809

cat("Fit Forward AIC model: ",mean(RMSE_m3))

## Fit Forward AIC model:  6.717115

cat("Fit Backward BIC model: ",mean(RMSE_m4))

## Fit Backward BIC model:  6.774654

cat("Fit Stepwise AIC model: ",mean(RMSE_m5))

## Fit Stepwise AIC model:  6.760802

cat("Fit Stepwise BIC model: ",mean(RMSE_m6))

## Fit Stepwise BIC model:  7.38123
```

From all these experiments, we can conclude that the model we obtain after applying Stepwise AIC on the complete model(polynomial + interaction) is the best model for predicting the strength of concrete.

Model Interpretability

Variable Importance

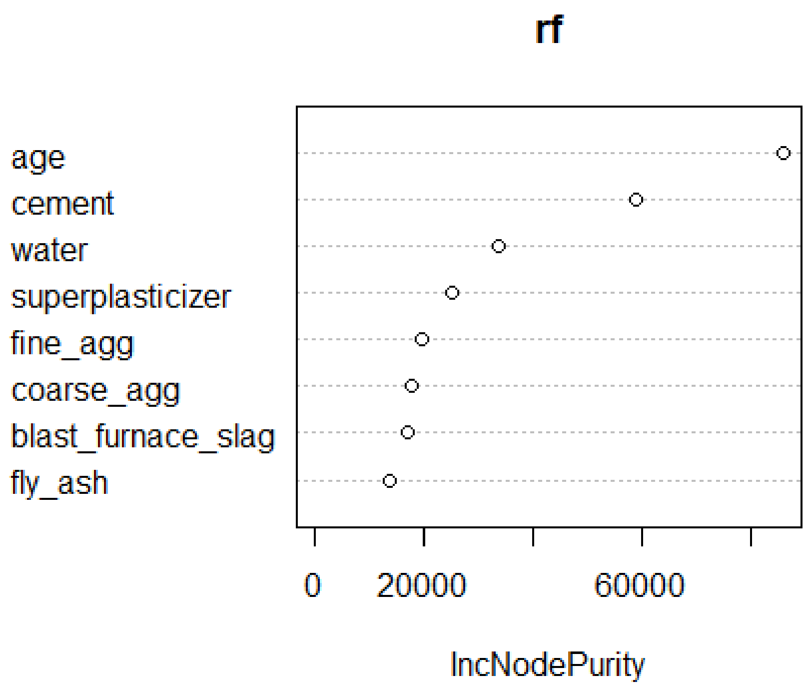
Using Standardized Model Coefficients

```
data_std = as.data.frame(scale(data, center=TRUE, scale=TRUE))

mlr_std <- lm(concrete_strength~.,data=data_std)
summary(mlr_std)

##
## Call:
## lm(formula = concrete_strength ~ ., data = data_std)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71518 -0.37728  0.04213  0.39280  2.06194
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.883e-16  1.940e-02   0.000 1.000000
## cement               7.494e-01  5.311e-02  14.110  < 2e-16 ***
## blast_furnace_slag   5.363e-01  5.235e-02  10.245  < 2e-16 ***
## fly_ash              3.369e-01  4.821e-02   6.988 5.03e-12 ***
## water               -1.921e-01  5.136e-02  -3.741 0.000194 ***
## superplasticizer     1.039e-01  3.342e-02   3.110 0.001921 **
## coarse_agg           8.392e-02  4.372e-02   1.919 0.055227 .
## fine_agg             9.673e-02  5.137e-02   1.883 0.059968 .
## age                  4.319e-01  2.052e-02  21.046  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6225 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

sort(abs(mlr_std$coefficients), decreasing =TRUE)

##             cement blast_furnace_slag                age
fly_ash
##       7.493508e-01       5.363354e-01       4.319263e-01       3.368942e-
01
##              water   superplasticizer           fine_agg
coarse_agg
##       1.921321e-01       1.039417e-01       9.672711e-02       8.391850e-
02
##        (Intercept)
##       1.882770e-16
```

Top influential features : cement, blast_furnace, age, fly_ash

Random Forest Variable Importance

```
rf <- randomForest(concrete_strength~., data=data, proximity=TRUE)
summary(rf)

##                 Length  Class  Mode
## call                  4 -none- call
## type                  1 -none- character
## predicted          1030 -none- numeric
## mse                 500 -none- numeric
## rsq                 500 -none- numeric
## oob.times          1030 -none- numeric
## importance            8 -none- numeric
## importanceSD          0 -none- NULL
## localImportance       0 -none- NULL
## proximity       1060900 -none- numeric
## ntree                 1 -none- numeric
## mtry                  1 -none- numeric
## forest               11 -none- list
## coefs                 0 -none- NULL
## y                  1030 -none- numeric
## test                  0 -none- NULL
## inbag                 0 -none- NULL
## terms                 3 terms  call

importance(rf)

##                    IncNodePurity
## cement                  58876.62
## blast_furnace_slag      16942.41
## fly_ash                 13491.83
## water                   33778.05
## superplasticizer        25254.10
## coarse_agg              17913.97
## fine_agg                19550.34
## age                     85823.92

varImpPlot(rf)
```



Top influential features : age, cement, water, superplasticizer

R2 from Single Predictor Model

```
calculate_r2 <- function(var_name, data) {
  print(var_name)
  fm <- as.formula(paste("concrete_strength", "~", var_name))
  model = lm(fm, data = data)
  r2 = summary(model)$r.squared
  return(r2)
}


for(el in names(data)) {
  if (el != "concrete_strength")
  {
  r2 = calculate_r2(el, data)
  cat("R2 is:  ", r2, "\n")
  }
}

## [1] "cement"
## R2 is:   0.2478374
## [1] "blast_furnace_slag"
## R2 is:   0.01817763
## [1] "fly_ash"
## R2 is:   0.01118377
## [1] "water"
```

```
## R2 is:   0.08387597
## [1] "superplasticizer"
## R2 is:   0.1340309
## [1] "coarse_agg"
## R2 is:   0.02720119
## [1] "fine_agg"
## R2 is:   0.02797222
## [1] "age"
## R2 is:   0.1081601
```

Top influential features : cement, superplasticizer, age, water

Based on all these experiments, we can conclude that the following are highly likely to be features of significant importance to the model:

1. Cement
2. Superplasticizer
3. Water
4. Age

Final Balanced Model

We choose the Model we got by using Stepwise BIC as a balanced model. It has 13 predictors as well as considerably high performance metrics.

```
best_model = lm(concrete_strength~ I(log(age))+ cement+ blast_furnace_slag+
               water+ superplasticizer+ I(log(fly_ash + 1))+ I(age^2)+
               I(log(cement + 1))+ I(superplasticizer^2)+
               I(superplasticizer^3) +
               water:superplasticizer + blast_furnace_slag:superplasticizer +
               cement:water, data=train)
summary(best_model)

##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement + blast_furnace_slag
+
##     water + superplasticizer + I(log(fly_ash + 1)) + I(age^2) +
##     I(log(cement + 1)) + I(superplasticizer^2) + I(superplasticizer^3) +
##     water:superplasticizer + blast_furnace_slag:superplasticizer +
##     cement:water, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8502  -4.5383  -0.1061   4.0948  18.9215
##
## Coefficients:
##                                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          -8.666e+01  1.674e+01  -5.177 2.85e-07
***
```

```
## I(log(age))                          9.106e+00  2.274e-01  40.049  < 2e-16
***
## cement                               1.775e-01  2.385e-02   7.442 2.54e-13
***
## blast_furnace_slag                   6.938e-02  4.535e-03  15.297  < 2e-16
***
## water                                5.437e-02  4.507e-02   1.206  0.22807
## superplasticizer                     5.183e+00  6.896e-01   7.515 1.51e-13
***
## I(log(fly_ash + 1))                  1.460e+00  1.957e-01   7.457 2.28e-13
***
## I(age^2)                            -8.111e-05  1.432e-05  -5.664 2.05e-08
***
## I(log(cement + 1))                   9.417e+00  3.224e+00   2.921  0.00358
**
## I(superplasticizer^2)               -1.648e-01  2.490e-02  -6.618 6.60e-11
***
## I(superplasticizer^3)                2.539e-03  5.756e-04   4.411 1.17e-05
***
## water:superplasticizer              -2.012e-02  3.159e-03  -6.369 3.20e-10
***
## blast_furnace_slag:superplasticizer  2.593e-03  5.318e-04   4.876 1.30e-06
***
## cement:water                        -5.749e-04  1.191e-04  -4.827 1.66e-06
***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.452 on 810 degrees of freedom
## Multiple R-squared:  0.8528, Adjusted R-squared:  0.8505
## F-statistic: 361.1 on 13 and 810 DF,  p-value: < 2.2e-16
```

```
vif(best_model)
```

```
##                        I(log(age))                              cement
##                           1.419282                          123.518185
##                 blast_furnace_slag                               water
##                           3.049695                           18.381712
##                   superplasticizer                 I(log(fly_ash + 1))
##                         325.454468                            4.281335
##                           I(age^2)                  I(log(cement + 1))
##                           1.616513                           29.739565
##              I(superplasticizer^2)               I(superplasticizer^3)
##                         163.431815                           56.562710
##             water:superplasticizer blast_furnace_slag:superplasticizer
##                         172.479383                            3.965155
##                       cement:water
##                         108.058971
```

Dropping interaction and polynomial terms with high VIF

Features Dropped:
- cement:water
- blast_furnace_slag:superplasticizer
- water:superplasticizer
- superplasticizer^2
- log(cement)

```
best_model = lm(concrete_strength~ I(log(age))+ cement+ blast_furnace_slag+
               water+ superplasticizer+ I(log(fly_ash + 1))+ I(age^2)+
               I(superplasticizer^3), data=train)
summary(best_model)

##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement + blast_furnace_slag
+
##     water + superplasticizer + I(log(fly_ash + 1)) + I(age^2) +
##     I(superplasticizer^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1593  -4.4585  -0.0811   4.2591  21.8556
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             6.050e+00  3.101e+00   1.951 0.051382 .
## I(log(age))             9.221e+00  2.399e-01  38.443  < 2e-16 ***
## cement                  1.100e-01  3.062e-03  35.922  < 2e-16 ***
## blast_furnace_slag      8.535e-02  3.785e-03  22.549  < 2e-16 ***
## water                  -2.223e-01  1.555e-02 -14.294  < 2e-16 ***
## superplasticizer        3.226e-01  9.031e-02   3.572 0.000375 ***
## I(log(fly_ash + 1))     1.381e+00  1.713e-01   8.064 2.63e-15 ***
## I(age^2)               -7.320e-05  1.471e-05  -4.976 7.91e-07 ***
## I(superplasticizer^3) -6.601e-04  1.320e-04  -5.002 6.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.828 on 815 degrees of freedom
## Multiple R-squared:  0.8341, Adjusted R-squared:  0.8325
## F-statistic: 512.4 on 8 and 815 DF,  p-value: < 2.2e-16

vif(best_model)

##          I(log(age))                cement    blast_furnace_slag
##             1.410247              1.817292              1.896671
##                water      superplasticizer   I(log(fly_ash + 1))
##             1.953770              4.983428              2.926611
```

```
##               I(age^2) I(superplasticizer^3)
##               1.523283              2.654322
```

```
press_best = sqrt(sum((resid(best_model)/(1-hatvalues(best_model)))^2)/n)
press_best
```

```
## [1] 6.886499
```

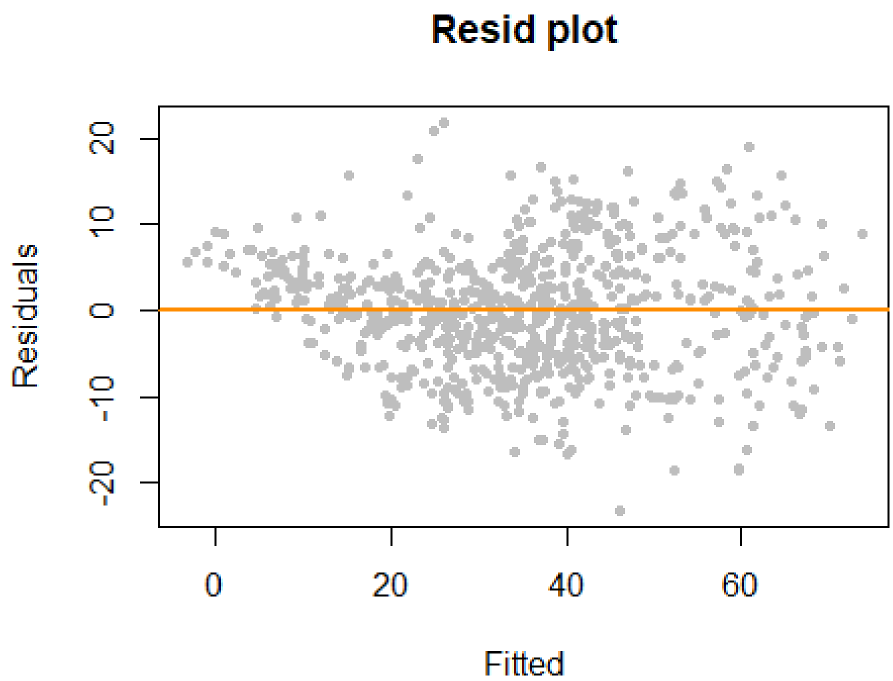
```
mse_best = calculate_mse_test(best_model, test)
mse_best
```

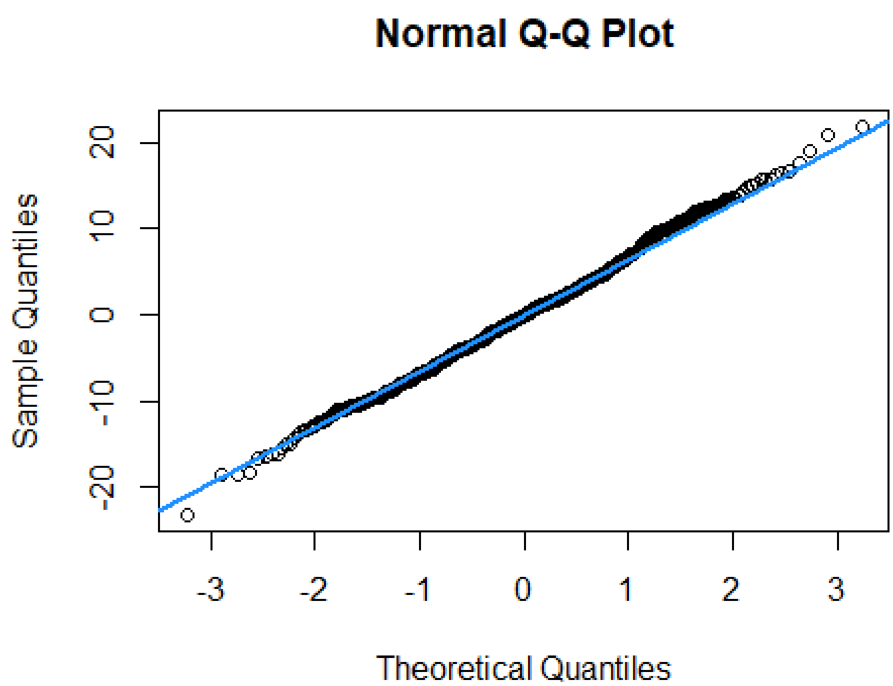
```
## [1] 48.58791
```

This model gives us a good balance of predictability as well as interpretability.

Checking Model Assumptions for the Final Selected model

```
check_model_assumptions(best_model)
```






```
## 
##  studentized Breusch-Pagan test
## 
## data:  model
```

```
## BP = 62.263, df = 8, p-value = 1.674e-10
##
##
##  Shapiro-Wilk normality test
##
## data:  resid(model)
## W = 0.9976, p-value = 0.2806
```

The normality assumption holds true, as the p-value for the Shapiro Wilks test is greater than the significance level (0.05).

The equal variance assumption fails as the p value for the BP test is still less than 0.05.

The linearity assumption seems to get violated as the data points are not equally distributed on either side of the axis.

Significance of Regression (Testing Hypothesis)

Verifing initial assumptions

We can use the final model to verify whether some of the initial hypothesis we formulated hold true.

1. Water is a significant predictor for the model
2. Cement is a significant variable for predicting concrete strength.
3. Fine aggregate and coarse aggregate do not contribute significantly towards the prediction of concrete strength.

```
summary(best_model)

##
## Call:
## lm(formula = concrete_strength ~ I(log(age)) + cement + blast_furnace_slag
+
##     water + superplasticizer + I(log(fly_ash + 1)) + I(age^2) +
##     I(superplasticizer^3), data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.1593  -4.4585  -0.0811   4.2591  21.8556
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.050e+00  3.101e+00   1.951 0.051382 .
## I(log(age))           9.221e+00  2.399e-01  38.443  < 2e-16 ***
## cement                1.100e-01  3.062e-03  35.922  < 2e-16 ***
## blast_furnace_slag    8.535e-02  3.785e-03  22.549  < 2e-16 ***
## water                -2.223e-01  1.555e-02 -14.294  < 2e-16 ***
## superplasticizer      3.226e-01  9.031e-02   3.572 0.000375 ***
## I(log(fly_ash + 1))   1.381e+00  1.713e-01   8.064 2.63e-15 ***
## I(age^2)             -7.320e-05  1.471e-05  -4.976 7.91e-07 ***
```

```
## I(superplasticizer^3) -6.601e-04  1.320e-04  -5.002 6.97e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.828 on 815 degrees of freedom
## Multiple R-squared:  0.8341, Adjusted R-squared:  0.8325
## F-statistic: 512.4 on 8 and 815 DF,  p-value: < 2.2e-16
```

```
summary(baseline)
```

```
##
## Call:
## lm(formula = concrete_strength ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.2826  -6.4525   0.7969   6.6006  27.6096
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -28.982063  28.748522  -1.008 0.313694
## cement               0.122050   0.009231  13.221  < 2e-16 ***
## blast_furnace_slag   0.107661   0.010990   9.797  < 2e-16 ***
## fly_ash              0.088705   0.013750   6.451  1.9e-10 ***
## water               -0.143393   0.043155  -3.323 0.000931 ***
## superplasticizer     0.332813   0.102787   3.238 0.001253 **
## coarse_agg           0.019683   0.010200   1.930 0.053991 .
## fine_agg             0.022415   0.011637   1.926 0.054423 .
## age                  0.115418   0.006102  18.914  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.29 on 815 degrees of freedom
## Multiple R-squared:  0.6235, Adjusted R-squared:  0.6198
## F-statistic: 168.7 on 8 and 815 DF,  p-value: < 2.2e-16
```

By looking at the model summary for both the best model and the baseline model, we can conclude that:

1. Water is an important predictor. [p-value < 0.05]
2. Cement is an important predictor. [p-value < 0.05]
3. Fine aggregate and coarse aggregate are not significant predictors. [p-value > 0.05 and not in best model]