Predicting Compressive Strength of Concrete

Preston Weaver

2025-02-24

Introduction to the Dataset

This dataset provides an analysis of concrete mixtures. Each entry includes the composition and curing age of concrete samples along with their measured compressive strength in megapascals (MPa). The goal of this report is to predict compressive strength (`strength`) using the these eight predictor variables:

- `cem`: Amount of Cement in the mix (kg)
- `slag`: Amount of Blast Furnace Slag in the mix (kg)
- `FA`: Amount of Fly Ash in the mix (kg)
- `h2o`: Amount of Water in the mix (kg)
- `plast`: Amount of Superplasticizer in the mix (kg)
- `cAgg`: Amount of Coarse Aggregate in the mix (kg)
- `fAgg`: Amount of Fine Aggregate in the mix (kg)
- `age`: Curing Time (1-365 days)

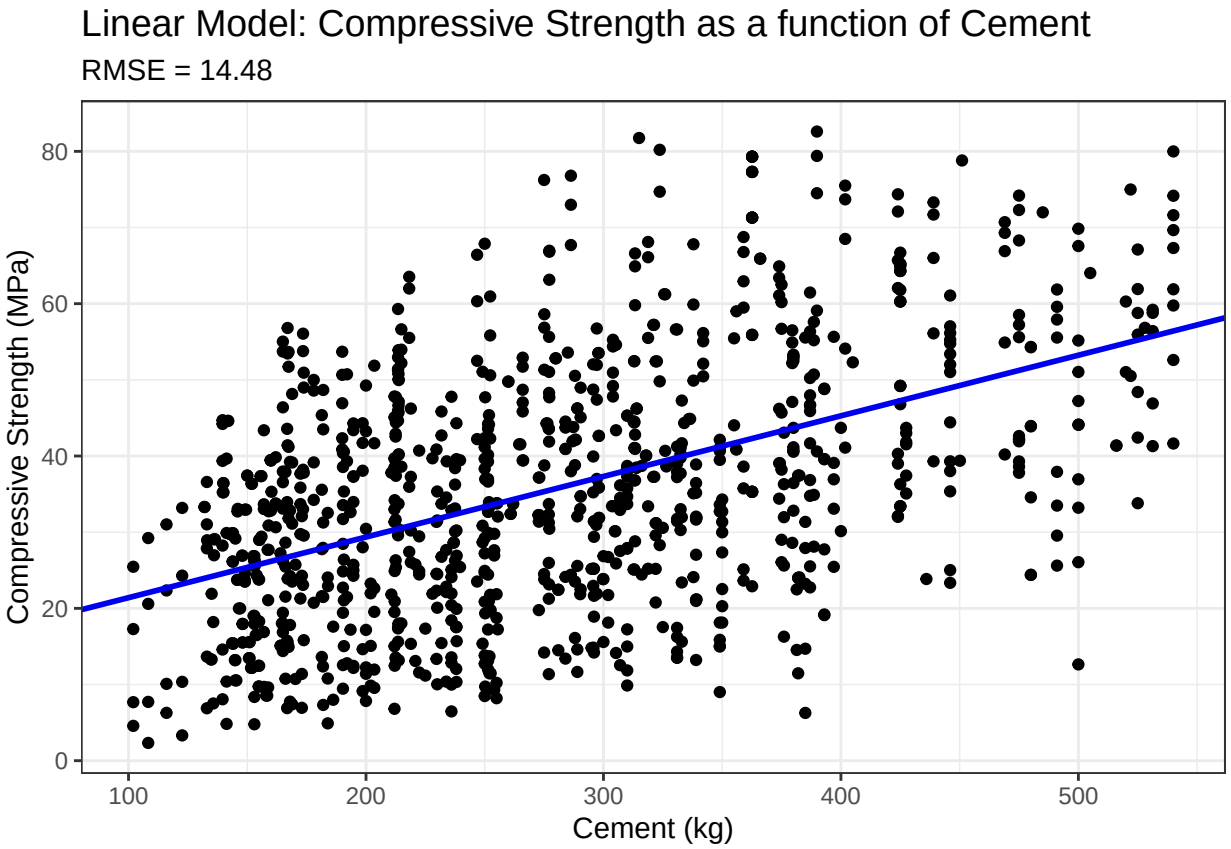
SLR with Best Predictor

To create a baseline, we will create a simple linear model using the predictor that best correlates with compressive strength. This variables happens to be the amount of cement in the concrete mix. Below is the summary output for the linear model that predicts compressive strength by cement content.

```
##
## Call:
## lm(formula = strength ~ cem, data = concrete_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.593 -10.952  -0.569   9.990  43.240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.442528   1.296948   10.37   <2e-16 ***
## cem          0.079580   0.004324   18.40   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.5 on 1028 degrees of freedom
## Multiple R-squared:  0.2478, Adjusted R-squared:  0.2471
## F-statistic: 338.7 on 1 and 1028 DF,  p-value: < 2.2e-16
```

From the output, it is observed that both the intercept term and the cement variable are significant predictors in the model. Looking at the R^2 value, this model explains about 25% of the variance in the compressive strength variable. This model also produces a RMSE of about 14.5, meaning the average prediction is off by ± 14.5.

Below is a visualization of the amount of cement in the mixture vs the compressive strength of the sample. The line of best fit calculated by the linear model has been laid over top.



Best MLR Model

The SLR above gave a good baseline, but a better model is definitely attainable by including the other predictors. To find the best model, the technique of backward elimination was used. Below is the summary output of the first iteration, which predicted compressive strength by all the other variables.

```
##
## Call:
## lm(formula = strength ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.4073  -6.1274   0.6145   6.7041  27.6941
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.291703  29.971572  -1.411  0.15863
## cem           0.126211   0.009657  13.069  < 2e-16 ***
```

```
## slag          0.110904   0.011479   9.662  < 2e-16 ***
## FA            0.098783   0.014145   6.984 6.26e-12 ***
## h2o          -0.132353   0.044960  -2.944  0.00334 **
## plast         0.213882   0.101483   2.108  0.03539 *
## cAgg          0.024742   0.010557   2.344  0.01935 *
## fAgg          0.029172   0.012093   2.412  0.01608 *
## age           0.113641   0.006082  18.683  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.2 on 763 degrees of freedom
## Multiple R-squared:  0.6179, Adjusted R-squared:  0.6139
## F-statistic: 154.2 on 8 and 763 DF,  p-value: < 2.2e-16
```

From the summary output, it is observed that all predictors are significant. The R^2 value of 0.618 indicates that the model can explain about 62% of the variance in the strength variable. A RMSE value of 10.1 is also achieved.

For the second iteration, the three variables with the highest p values (plast, cAgg, fAgg) were taken out. The variables used for the rest of the iterations are cem, slag, FA, h2o, and age. For iterations 3 through 6, a combination of log transformations were done to the predictor variables. A summary table of the model iterations is included below.

Table 1: Model Iterations and Their Respective Metrics

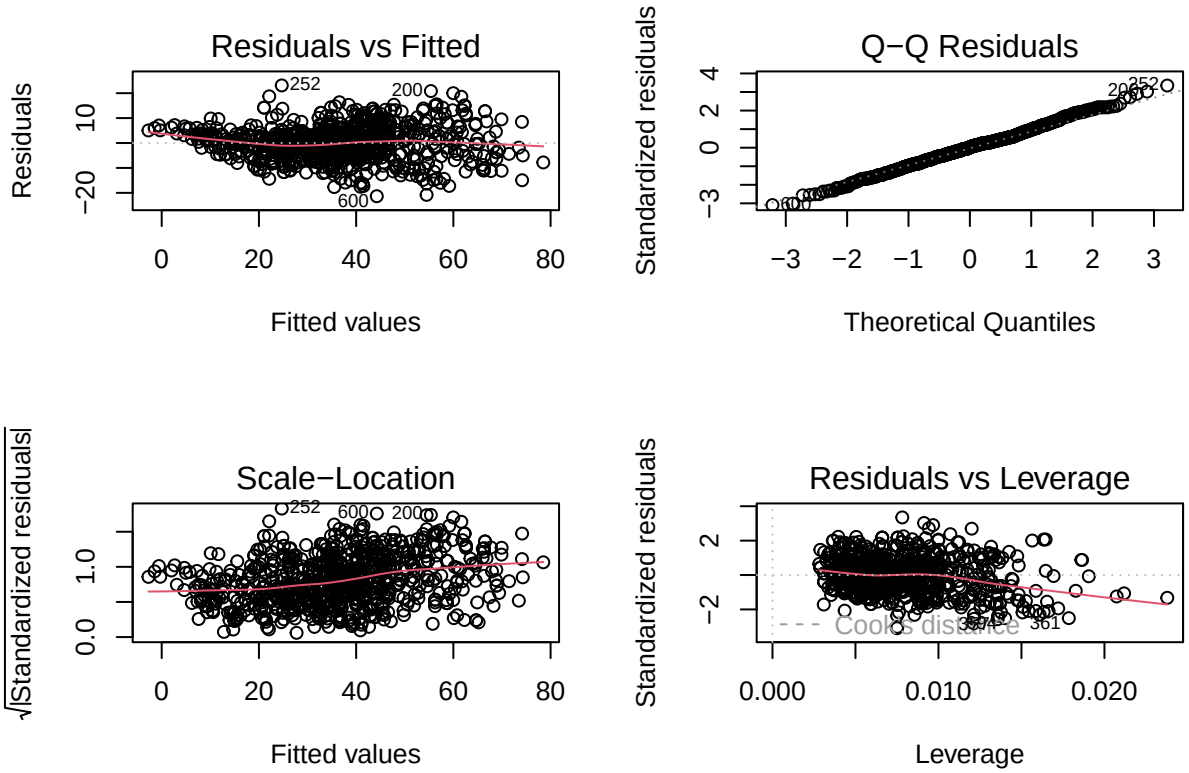
Iteration	R Squared	RMSE
Model 1	0.614	10.14
Model 2	0.611	10.20
Model 3	0.778	7.70
Model 4	0.782	7.64
Model 5	0.785	7.59
Model 6	0.822	6.90

The final model iteration uses the five variables previously mentioned: cem, slag, FA, h2o, and age. It also applies a log transformation to both the FA and age predictors. Taking a look at the model summary of iteration 6, a R^2 value of 0.82 is achieved, meaning 82% of the variance in compressive strength can be explained by the model. In addition to the high R^2, Model 6 also achieved the lowest RMSE with a value of 6.90.

```
##
## Call:
## lm(formula = strength ~ cem + slag + log(FA) + h2o + log(age),
##     data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -21.3458  -4.4717   0.1572   3.9701  23.1101
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.709315   2.510770   6.655 5.39e-11 ***
## cem          0.109905   0.002869  38.313  < 2e-16 ***
## slag         0.087320   0.003489  25.025  < 2e-16 ***
```

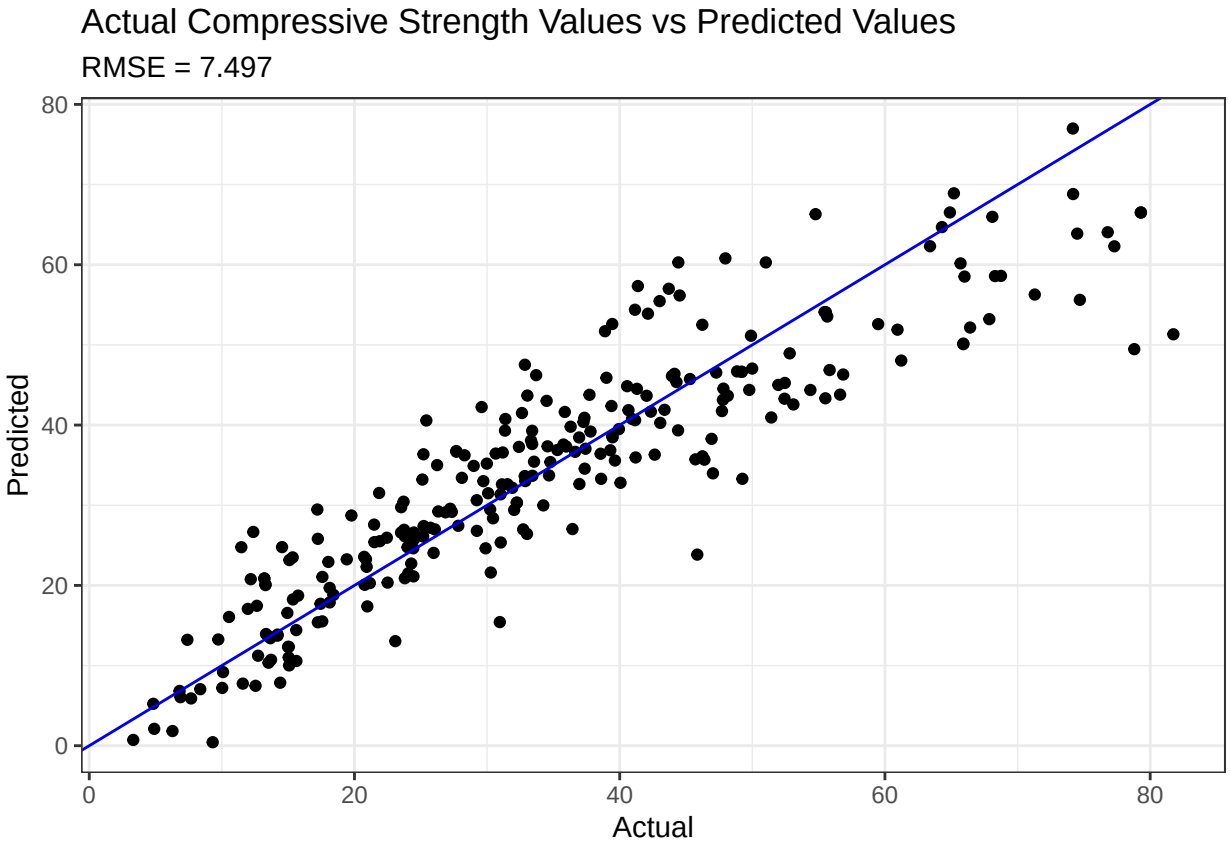
```
## log(FA)      0.629536   0.045503  13.835  < 2e-16 ***
## h2o         -0.242047   0.012195 -19.847  < 2e-16 ***
## log(age)     8.681322   0.214017  40.564  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.929 on 766 degrees of freedom
## Multiple R-squared:  0.8231, Adjusted R-squared:  0.8219
## F-statistic: 712.6 on 5 and 766 DF,  p-value: < 2.2e-16
```

Now that a final model has been derived, it is time to check out the diagnostic plots.



Nothing concerning jumps out from the diagnostic plots.

After running the model on the test set, a plot of values predicted by the model vs actual values can be created. It is observed that most predictions hover around the line which indicates the model is fairly accurate. Something to note is that the model tends to underestimate the larger values of strength. Also, the RMSE of the predicted values vs the actual values of the test set is around 7.5, meaning the model has an error of ± 7.5 when preforming on unseen data.



Prediction

To test the validity of the model, a T Test is prefomed to see if the mean of the residuals is zero.

```
##
##  One Sample t-test
##
## data:  test$strength - test$predictions
## t = 0.55932, df = 257, p-value = 0.5764
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.6589875  1.1818298
## sample estimates:
## mean of x
## 0.2614211
```

A p-value of 0.576 is obtained, meaning we fail to reject the null hypothesis. This suggests that there is no significant evidence to conclude that the true mean difference is different from zero. Additionally, since 0 is contained within the confidence interval (-0.659 to 1.182), we can infer that the model is not systematically bias in its predictions.

Conclusion

A reasonably reliable model was developed. To confirm that the seed did not influence its performance, seven additional random seeds were tested. The table below presents the R^2 and RMSE values for the models

generated with these different seeds.

Table 2: Testing Model with other Seeds

Seed	R Squared	RMSE
1568	0.8248184	6.958788
1176	0.8217026	7.024565
1911	0.8222733	7.023508
562	0.8164890	7.041637
574	0.8100913	7.251013
1803	0.8158317	7.076510
1500	0.8229715	7.122970
Average	0.8191682	7.071284

The average R^2 of 0.819 is very close to the 0.821 achieved by the final model. Similarly, the average RMSE of 7.071 is comparable to the final model's 6.90. Since the results show minimal variation, we can confidently conclude that the model is reasonably reliable and performs well.