Predicting Concrete Compressive Strength using Multiple Linear Regression

Jonathan Then, Lihai Geng, Yan Liu, and Yujia Zhai

University of Sydney, Camperdown, New South Wales, Sydney, Australia, 2006

This version was compiled on November 20, 2020

This report demonstrates the process of using Multiple Linear Regression on predicting concrete compressive strength of 28 days using the Concrete dataset. The data exploration is performed to visualise initial observations between all predictors and the target. Detailed analysis on the model selection process are presented to give the final chosen model. Further inference, interpretion and performance analysis on final model led to the conclusion of this process.

Introduction. The main aim of this report is to demonstrate the possibilities of adapting Multiple Linear Regression (MLR) to predict the concrete compressive strength of 28 days using the Concrete dataset.The main question to answer is: Are all variables in the dataset significantly in predicting the concrete compressive strength of 28 days?

Dataset. The dataset can be found on the UCI machine learning repository. Here is the link : https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength. It is donated by its original owner — Professor I-Cheng Yeh, Department of Information Management Chung-Hua University, Hsin Chu, Taiwan 30067, R.O.C..In the original paper, experimental data came from 17 different sources. So during the data collection, there is a determination made to ensure that all these mixtures are fairly representative group governing all of the major parameters that influence the strength of high-performance concrete, for example, they deleted some observations which contain larger aggregates. In this dataset, there are 1030 observations with 9 quantitative variables There is one dependent variable Compressive Concrete Strength, which is the index used to show how much pressure the concrete can withstand. Opposite with that one, the other 8 variables are all independent variables. They are

- Cement, a kind of binder, is used for combining materials together.
- Blast Furnace Slag, is a non-metallic coproduct in the production of iron.
- Fly Ash, which can improve the strength and segregation of the concrete and make concrete easier to pump.
- Water, used to form a paste that binds the aggregates together.
- Superplasticizers, they are chemical compounds using for reducing water in the production of concrete.
- Coarse Aggregate & Fine Aggregate, they are particles free of absorbed chemicals or coatings of clay and other fine materials that could cause the deterioration of concrete.
- Age, the number of days that how long has the concrete been made for And there are no missing values in this dataset.

Analysis. The dataset was cleaned and filtered to only include data points where the concrete was tested at 28 days in order to answer our main question. The resulting dataset contains 425 observations.

Data Exploration. Initial data exploration was conducted to see if there was any linear relationship between our dependent variable, strength, when compared to all the other variables. By looking at the correlation coefficient and the ggpairs plot, we can infer that cement has a strong positive linear relationship, water has a negative linear relationship, while the rest has a weak relationship with strength. From there, we decided to use a full model containing all predictors in the dataset so as to find out the p-values when fitted on a multiple linear regression.

```
#             Estimates p-value
#  Intercept  -95.67    0.17
#  Cement     0.14      0.11
#  Slag       -0.06     0.11
#  FlyAsh     0.04      0.05
#  Water      0.001     <0.001
#  Plasticizer <0.001   <0.001
#  Coarse Agg 0.126     0.260
#  Fine Agg   <0.001    <0.001
```

- Both the p-values of water and plasticizer are greater than 0.05. Therefore, they are insignificant at the 5% level of significance.
- We cannot immediately drop both of them from the model as the p-values are only testing individual coefficients.
- The AIC for the full model is 2881.297.

Backward & Forward Selection Model. Stepwise model selection methods of forward and backward selection were used to select the final model.

```
#            Estimates p-value Estimates p-value
#  Intercept  -81.93    0.003   -81.93    0.003
#  Cement     0.17      <0.001  0.17      <0.001
#  Slag       0.14      <0.001  0.14      <0.001
#  FlyAsh     0.11      <0.001  0.11      <0.001
#  Water      -0.09     0.012   -0.09     0.012
#  Coarse Agg 0.04      <0.001  0.04      <0.001
#  Fine Agg   0.05      <0.001  0.05      <0.001
```

```
#                    Backward Model Forward Model
#  Observations      425            425
#  R^2/R^2 Adjusted  0.771/0.767    0.771/0.767
#  AIC               2880.591       2880.591
```

Both forward and backward models arrived at the same model and has slightly lower AIC when compared to the full model. Therefore, this will be our final model.

Assumptions Checking (Refer to Figure 1).

- Linearity: There is a slight curved pattern in the residual vs fitted values plot. We are overestimating concrete strength for low and high fitted values.

- Independence: It is assumed that the observations are not related to one another as this was dealt with in the experimental design phase, before data collection.
- Homoscedasticity: There does seem to be some fanning out of the residuals in the residual vs fitted value plot, indicating that there may be some heteroscedasticity in our data.
- Normality: Based on the QQPlot, the bottom and top points are not on the line. However, we could appeal to the Central Limit Theorem because we have a reasonable amount of observations meaning our inferences are approximately valid.

Results. The final model is

Strength = -81.93 + 0.17(Cement) + 0.14(Blast Furnace Slag) + 0.11(Fly Ash) - 0.09(Water) + 0.05(Fine Aggregate) + 0.04(Coarse Aggregate) + ϵ

Interpretation. Holding all other variables constant:

- Increase in 1 kg/m^3 of Cement results in 0.17 MPa increase in strength on average.
- Increase in 1 kg/m^3 of Blast Furnace Slag results in 0.14 MPa increase in strength on average.
- Increase in 1 kg/m^3 of Fly Ash results in 0.11 MPa increase in strength on average.
- Increase in 1 kg/m^3 of Water results in 0.09 MPa decrease in strength on average.
- Increase in 1 kg/m^3 of Fine Aggregate results in 0.05 MPa increase in strength of on average.
- Increase in 1 kg/m^3 of Coarse Aggregate results in 0.11 MPa increase in strength on average.

Statistical Inference. Individual t – test was conducted on each slope parameter to check if it is significantly different from zero, i.e significantly in predicting strength. The test procedures are as follows:

Hypothesis: $H_0 : \beta_i = 0$ vs $H_1 : \beta_i \neq 0$

Assumptions: The residuals are iid normal distribution with mean of zero and constant variance of σ^2 and there is a linear relationship between strength and each predictor.

P-value: Each of the p-value is less than 0.05.

Decision: We reject the null hypothesis and conclude that there is very strong evidence in the data to indicate a linear relationship between each predictor and the strength.

However, since there is a slight heteroscedasticity in our data, the inference might be invalid.

Performance.

```
#                       Full Model Final Model
#  Adjusted R Squared      0.7733      0.7699
#  RMSE                    7.1688      7.0898
#  MAE                     5.3890      5.3254
```

10-folds cross validation method was used to fit the full model and the final model to compare performance results.

- In Sample Performance: The full model has a slightly higher adjusted R^2, indicating a stronger fit of the model as compared to the final model. The adjusted R^2 indicates how much percentage of total variability in strength is explained by the model.
- Out of Sample Performance: The Root Mean Squared Error (RMSE) is the square root of the variance of the residuals and the Mean Absolute Error (MAE) is the average over the absolute differences between prediction and actual observation. A smaller RMSE and MAE is desired. The final model has a lower RMSE and MAE, indicating that it is better at predicting observations that are not used to build the model. This is another reason why the final model was chosen.

Discussion and Conclusion.

Conclusion. Based on the final model, Super Plasticizer is excluded in the multiple linear regression model because it is not significant for predicting concrete compressive strength of 28 days.

Limitation. According to the assumption checking of the final model, there is slight heteroscedasticity in the data, so the standard errors computed for the least squares estimators are incorrect. This can affect confidence intervals and hypothesis testing that use those standard errors, which could lead to misleading conclusions.

Improvement. This limitation can be improved by using robust standard errors to correct the issue of incorrect standard errors so that the interval estimates and hypothesis tests are valid.

References.

- Gopal Mishra. (2014, October). Factors Affecting Strength of Concrete. The Constructor. https://theconstructor.org/concrete/factors-affecting-strength-of-concrete/6220/Humad, A. M. (2019).
- https://www.frontiersin.org/articles/10.3389/fmats.2019.00009/fullPatel, H. (2019, November 9). 11 Factors that can Affect the Strength of Concrete. GharPedia.
- https://gharpedia.com/blog/factors-that-affect-strength-of-concrete/Portland Cement Association. (n.d.). Aggregates. Retrieved November 10, 2020, from
- Association. (n.d.). Slag Cement Questions. Retrieved November 10, 2020, from https://www.slagcement.org/resources/faqs.aspxStelsel, K. (2015, March 7). Using Fly Ash in Concrete. National
- Yeh, I.C. (1998). Modeling of strength of high-performance concrete using artificial neural networks. Cement and Concrete Research, Vol. 28, No. 12, pp. 1797–1808, DOI: 10.1016/S0008-8846(98)00165-3
- ‘Cement’. (n.d.). In Wikipedia. Retrieved November 13, 2020, from https://en.wikipedia.org/wiki/Cement
- U.S. Department of transportation Federal Highway Administration. (2016). User Guidelines for Waste and Byproduct Materials in Pavement Construction. Retrieved from https://www.fhwa.dot.gov/publications/research/infrastructure/structures/97148/bfs1.cfm
- Rodriguez, J. (2019). Uses, Benefits, and Drawbacks of Fly Ash in Construction. Retrieved from https://www.thebalancesmb.com/fly-ash-applications-844761
- ‘Superplasticizer’. (n.d.). In Wikipedia. Retrieved November 13, 2020, from https://en.wikipedia.org/wiki/Superplasticizer
- Aggregates. (2019). Portland Cement Association. Retrieved from https://www.cement.org/cement-concrete-applications/concrete-materials/aggregates

Link to GitHub.

- https://github.sydney.edu.au/YZHA2588/T11A_Ontime_6

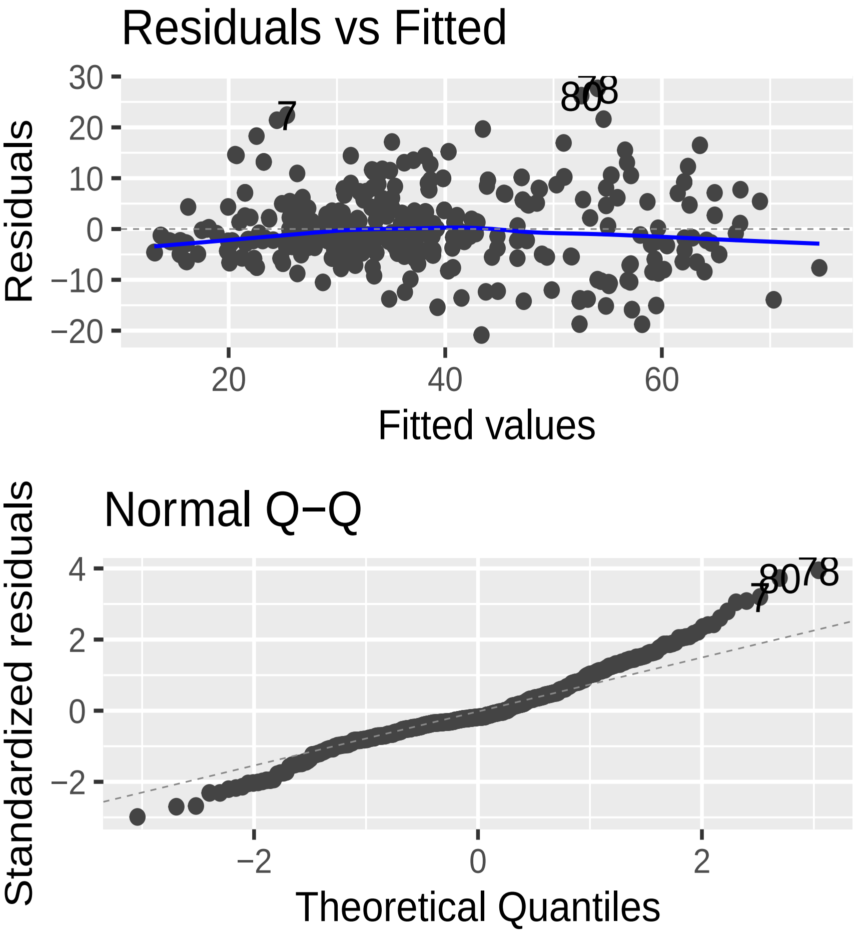

Fig. 1. Residuals vs Fitted Plots and QQPlot