THE ANALYSIS OF CEMENT MIXTURES TO IMPROVE COMPRESSIVE STRENGTH

BY ALEX SHEHDULA

1. Introduction. Concrete is one of the most commonplace construction materials in the world. Effective use of concrete has been a staple of large-scale construction for thousands of years. To minimize costs and improve duration of constructions, variations in component mix have been made to increase the compressive strength of concrete. Current consensus says that the relationship between concrete components and compressive strength can be modeled by non-linear means, specifically using an exponential relationship between water and cement as described by Abram's law (Abrams, 1918).

This report aims to analyze the compressive strength of concrete based on different manufacturing parameters to identify trends. These trends are intended to inform future analysis and composition methods for concrete. A particular point of interest is to examine whether the existing beliefs on the concrete mixture and its impact on strength can be challenged at all. This primarily involves the belief that the mixture of components have a non-linear relationship to compressive strength. This research can be used to guide future manufacturing decisions and improve the effectiveness of concrete.

2. Data.

2.1. *Data Source.* The data collected was based on 1030 different concrete samples explicitly created for this assessment (UCI Machine Learning Repository, 2007). Each sample had a different mixture of cement, sand, and additives. The pieces were made and tested following American Society for Testing and Materials (ASTM) standards, which cover the procedures for creating test specimens of concrete for laboratory analysis. The specimens were tested using a Toni Technick machine, which specializes in recording the behavior of building materials. Samples were also submerged in water during the testing process (International, 2015).

Compressive strength measures a material's resistance to breaking under compression. It is one of the most used measures of quality control for high-grade concrete and will, as such, be used as the response variable for the data at hand. The unit of measurement for compressive strength is megapascals (MPa), indicating the pressure the concrete can withstand before breaking.

2.2. *Structure.* The structure of the data includes the composition of each test specimen and its corresponding compressive strength results. The composition includes the density (kg/m^3) of water, fly ash, cement powder, coarse aggregate, fine aggregate, superplasticizer, and blast furnace slag. It also includes the age of the sample (in days). The relationship between the compressive strength of each sample and its covariates was analyzed (Neville, 2011).

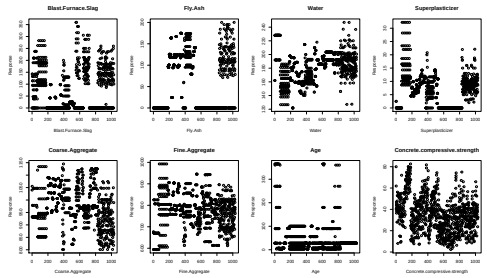

FIG 1. *Scatter plots showing the relationship between response and different covariates*

2.3. *Exploratory Analysis.* None of the relationships appear to be linear. This suggests that a linear model with no interactions would likely not show any substantial trends or accurately model the relationship. This does not mean that linear models are not useful. Interactions might still capture linear relationships, and approaches like LASSO or ridge regression might offer better coefficient insights. However, a non-linear relationship is expected.

The dataset's covariates mostly share the same units, except for the age covariate. This impacts the effectiveness of LASSO and ridge regression. As such, since part of our research question includes exploring the effectiveness of linear models on concrete composition, the data will be scaled. The scaling may impact the ability to interpret the coefficient values of the covariates. However, we will still be able to understand a relationship between the covariates based on the magnitude of their coefficients.

3. Methods.

3.1. *Linear Models: Ridge and LASSO.* As discussed, linear models will be fit to assess whether a linear relationship can be captured. By using Ridge regression and LASSO along with our scaled data, we can develop a form of feature selection. Coefficients that are penalized heavily (or reduced to zero by LASSO) will help determine whether some features are important and influential on the compressive strength of concrete. We will use every covariate to try and capture any relevant information. The models are as follows:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8$$

where β_0 represents the model's intercept coefficient, and β_1 through β_8 are the coefficients for the amount of cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, fine aggregate, and age, respectively. x_1 through x_8 are the values of the covariates. Additionally, the objective functions for the two functions are as follows:

3.1.1. *Ridge Regression Objective Function.*

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

3.1.2. *LASSO Regression Objective Function.*

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

The models will be constructed and an optimal penalization parameter will be selected using cross-validation. Cross-validation aims to find an optimal value for the penalization parameter λ by partitioning the dataset into k folds. Each fold is then used as a validation set for the other $k-1$ folds. A sequence of different λ values are then generated and tested across these k validation sets. The mean squared error for each attempted λ is calculated, and the λ that minimizes the mean squared error is chosen as an optimal parameter.

For the dataset D, being partitioned into the k folds (by default 10 using the built-in R `glmnet` package) $D_1, \ldots, D_k$, we find the MSE as follows. We note that this model is first trained on $D \setminus D_i$, which is to say the D not involved in the validation folds. We proceed to average the validation errors across all folds, and find the cross-validated MSE:

$$\text{MSE}_i(\lambda) = \frac{1}{|D_i|} \sum_{j \in D_i} \left(y_j - \mathbf{x}_j^T \hat{\beta}(\lambda) \right)^2$$

$$\text{CV}(\lambda) = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i(\lambda)$$

3.2. *Additive Models.* We also fit additive models, which can capture non-linear relationshpis that may be helpful based on our experimental analysis of the dataset. Additive models use smoothing functions (splines) on each predictor. Therefore, instead of each predictor having a single coefficient, it has a function. We plan on using all covariates and as such the additive model will look like this:

$$g(\mathbb{E}[Y]) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

where each $f_j(X_j)$ is a smooth function for the j-th predictor.

3.2.1. *Additive Model Objective Function.* The corresponding objective function is:

$$\text{RSS} + \lambda \sum_{j=1}^{p} \int \left(f_j''(X_j) \right)^2 dX_j$$

Parameter selection is done automatically by the `mgcv` package in R, which optimizes itself based on the second derivative of the smooth functions to help prevent overfitting.

4. Results.

4.1. *Linear Models.*

4.1.1. *Linear Models: Performance.* The LASSO and Ridge regression models perform poorly. Using a test train split of 80%, we are able to find that the R^2 is about 0.63. This could be a good performance, but we will compare it to other models to see how it holds up to non-linear approaches. Similar findings are found for LASSO with an R^2 of 0.64. It is worth noting that given a strong penalization parameter, the LASSO model eventually reduces the fine aggregate coefficient to zero.

4.1.2. *Linear Models: Assumptions.* Additionally, we check the assumptions. A ridge regression requires the residuals to be normally distributed around 0. We see from the residual plot of the ridge regression that this assumption is not met. The LASSO regression also does not meet the assumptions set with clear non-randomness for the residuals.

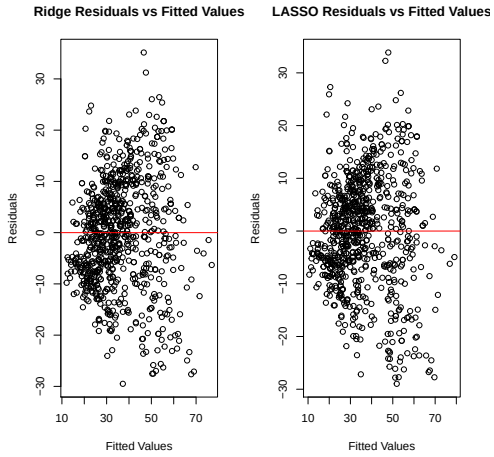

FIG 2. *Residuals vs Fitted Values for Ridge and LASSO*

4.2. *Additive Model.*

4.2.1. *Additive Models: Performance.* We test the performance of our model using cross-validation. By setting the number of folds to 10, we are able to determine what the R^2 value is, as well as the mean absolute error. This allows us to determine how effective the model is at explaining the variance, and the error between predictions and the final value.

We now fit an additive model with all covariates. With cross-validation testing, $k = 10$, we are able to see that the expected MAE (4.87), and R^2 (0.87) significantly outperform the linear models. One concern we have is possible collinearity. The covariance matrix suggests no high correlation between covariates, but multicollinearity may exist.

We check the concurvity of the additive model's results, which checks whether any of the smoothing terms can be estimated by a combination of other selected smoothing terms. After selecting our ideal number of knots, and creating an additive model using all the covariates, we now check the concurvity.

	s(Cement)	s(Blast.Furnace.Slag)	s(Fly.Ash)	s(Water)	s(Superplasticizer)	s(Coarse.Aggregate)	s(Fine.Aggregate)	s(Age)
worst	0.00	0.91	0.90	0.90	0.90	0.80	0.86	0.31
observed	0.00	0.90	0.90	0.85	0.74	0.52	0.79	0.07
estimate	0.00	0.70	0.74	0.76	0.77	0.73	0.67	0.11

TABLE 1
Table showing the values for different parameters

In the worst-case scenario, the worst-valued covariate is superplasticizer. With the removal of that covariate we then find that the concurvity is still high, so we remove the next highest-valued covariate. After doing this twice, we end up removing superplasticizer, the fine aggregate, and the fly ash.

With the three covariates being removed, we note low levels of concurvity. We proceed with this set of covariates. Next, we select the number of knots. The goal is to select a number of knots that avoid overfitting the smoothing parameters while maintaining appropriate error measures. We find that 5 knots seem to strike an effective balance between model performance and having the smoothing terms avoid overfitting.

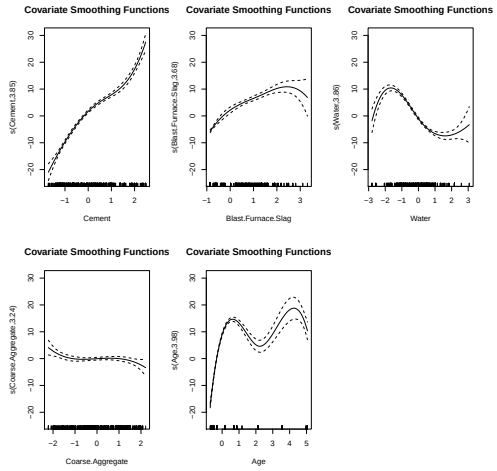

FIG 3. *Covariate Smoothing Functions*

The expected mean average error for this model is 5.418 and the R^2 is 0.827. This fit suggests improved performance versus linear models. The use of cross-validation when testing the fit rather than just a test train split reduces the bias and variance. Additionally, looking at the smoothing functions, we notice that all covariates except for potentially the coarse aggregate appear to be non-linear. We further confirm this by viewing the model summary and seeing the edf values all being above 4.

4.2.2. *Additive Models: Assumptions.* Testing the assumptions of the additive model is straightforward. The residuals are relatively normally distributed, with some small deviations that need to be noted. Additionally, we have a plot that closely resembles a normal QQ plot with slight deviations at the tails. This suggests that the model may have some more complex relationships that the additive model can't capture. The histogram of residuals also appears normal.

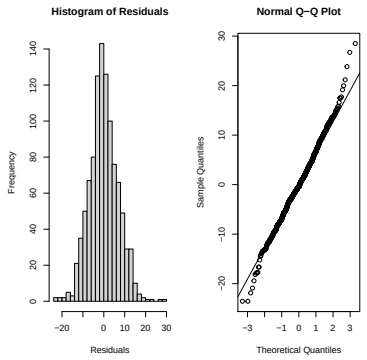

FIG 4. *Histogram of Residuals and Normal QQ Plot*

4.2.3. *Additive Models: Implications.* The smoothing charts (section 4.21) suggest that cement powder has the most consistently important effect. Blast furnace slag and water also appear significant, with clear non-linear trends. Age has the largest impact, especially in the first 50 days before tapering off.

The removed covariates, superplasticizer, fly ash, and fine aggregate, make sense given domain knowledge. Superplasticizer is a water reducer, with its amount typically depending on the water-to-cement ratio. Fly ash reduces the need for water, cement powder, and aggregates. Fine aggregate affects workability rather than strength, heavily influenced by water and coarse aggregate amounts. Removing these covariates and reducing the concurvity is sensible.

5. Conclusions. We find that the impact of compressive strength is determined through a non-linear relationship between its components. Age has two inflection points, indicating peak periods of strength for the concrete: early after mixing, then again after additional settling. Cement powder is always positively impactful. Water has two inflections, suggesting too little or too much water can be problematic. Blast furnace seems to offer some gains on compressive strength with minimal amounts, with diminishing returns eventually.

Interpreting the covariates directly is difficult due to scaling, so the magnitudes are the main source of interpretation. The insights may improve with more complex models. Models such as neural networks and random trees should be researched further, as they may better capture the extremely non-linear relationships suggested by this paper's results.

The results support existing beliefs about concrete compressive strength. Water and cement seem to have an exponential relationship with compressive strength, which linear models cannot accurately capture. Age has a significant impact, indicating concrete must be set without disturbance to improve strength. The results confirm that a linear model does not sufficiently explain the relationship between the components of a cement mix and its compressive strength.

REFERENCES

ABRAMS, D. A. (1918). Design of concrete mixtures. *Structural Materials Research Laboratory* **1**.

INTERNATIONAL, A. (2015). ASTM Standards. Retrieved from `https://www.astm.org/Standards/concrete-and-aggregates.html`.

NEVILLE, A. M. (2011). *Properties of concrete*. Pearson Education.

UCI MACHINE LEARNING REPOSITORY (2007). Concrete Compressive Strength Data Set. Accessed: 2024-04-02.