Predicting concrete compressive strength with regularized linear models

Saulo Mendes de Melo
Applied Computational Intelligence
PPGETI
Fortaleza, Brazil
saulomelo96@hotmail.com

***Abstract*—The prediction of concrete strength has valueable applications in civil engineering. This article has done and analysis of performance of linear models in modeling and predicting the concrete strength of samples with different ingredients and ages. The results found have shown that linear regression and its main variants can model such data with limited results, obtaining approximately 7.84 RMSE and a R^2 of 0.656 and more complex methods ca be used more successfully.**

***Index Terms*—Linear Regression, Concrete strength, Ridge**

I. Introduction

Concrete is a material present in almost every construction project. It's a basic material that depending on its composition and proportion of ingredients might have particular properties and specific applications. Conventional concrete is mainly made of three basic ingredients, cement, fine and coarse aggregate, and water. High-performance concrete (HPC) is a more complex material that requires supplements such as fly ash, blast furnace slag and superplasticizer, and its behavior is more difficult to model.

In civil engineering, a very traditional method for modeling the strength of concrete is through the Abrams' Law, developed by Duff Abrams in 1919, which states a linear relationship between the logarithm of the strength and the water-to-cement ratio [1]. Studies have shown that other components present in the concrete mixture also have a considerable influence in concrete strength [2].

A study done in 1998 [3] has compared the modeling of HPC strength with artificial neural networks (ANN) and regression based analysis and found that ANN models perform better, due to the complexity of the behavior of HPC. This work presents an evaluation of regularized regression models to predict the compressive strength of concrete based on its ingredient proportions and age.

II. Methodology

A. Data

The data set used in this work was the Concrete Compressive Strength Data Set and it is publically available at the UCI Machine Learning Repository [4], and it is made up of 1030 samples, each with 9 attributes. The age variable is quantitative and it corresponds to the amount of time in days passed since the mixing of the concrete, but occurs in fairly regular time intervals instead of a more random distribution, such as 1, 2, 4 and 8 weeks in the majority of the data, but a few samples of 3, 4, 6 and 12 months are present. Apart from MPa, which is the compressive strength of the concrete and the output of the model, all other variables are ingredients and are measured in kilograms per cubic meter(kg/m^3).

The skewness is a parameter which indicates how much the values of a distribution are concentrated in one side in relation to the center. The distributions for slag, fly ash and superplasticizer are heavily right-skewed, this is due to the frequent presence of zero values, and it is also the reason that their Max-Min Ratio is infinite. A general rule of thumb that allows us to determine whether a distribution is skewed or not is checking if the Max-min Ratio is greater than 20 [5], and by such metric we confirm that slag, fly ash and superplasticizer are skewed and require transformation. Cement, water, fine and coarse aggregate have fairly normal distributions and do not require heavy transformations for modeling.

B. Linear Regression

We can define the linear regression as model that relates the predictors and the outputs linearly through a weighted sum of the variables. So for a given dataset with N observations and P predictors, where $X = x_1, x_2, \ldots, x_P$ are the predictor variables and y is the output, the objective is to find coefficients $\beta = \{\beta_0, \beta_1, \ldots, \beta_n\}$ that satisfy equation 1, where β_0 is a bias term.

$$y = \beta_0 + \beta_1 x_1 + \beta_1 x_1 + \cdots + \beta_P x_P \tag{1}$$

From 1 we can see that such an equation presumes that the relationship between the variables and the response falls along a certain flat hyperplane with intercept β_0, which means that data with more complex relationships might not be well modeled. On the upside this provides a very interpretable model, when adequately fit, where we can analyse the influence of each predictor from weights β [6].

There are two main algebraic solutions to linear regression that may be pointed out. Both lead to the same conclusion, but each important considerations to the problem.

1) Ordinary Least Squares (OLS): Consider the linear model $y = f(X) + \varepsilon$ where the response y is a function the predictors X and ε is an amount of error. In OLS we minimize the residual sum of squares of the model, that is, the squared

distances between the hyperplane predicted by the coefficients and the actual response, as seen in 2.

TABLE I
DESCRIPTIVE STATISTICS OF THE DATASET

	Mean	Std. Dev.	Skewness	Max-Min Ratio
Cement	281.17	104.51	0.51	5.29
Slag	73.9	86.28	0.8	inf
Fly Ash	54.19	64.0	0.54	inf
Water	181.57	21.35	0.07	2.03
Superplasticizer	6.2	5.97	0.91	inf
Coarse Aggregate	972.92	77.75	-0.04	1.43
Fine Aggregate	773.58	80.18	-0.25	1.67
Age	45.66	63.17	3.27	365.0
MPa	35.82	16.71	0.42	35.45

$$\begin{aligned}\text{RSS}(\beta) &= \sum_{i=1}^{N}(y_i - f(x_i))^2 \\ &= \sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \\ &= (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\end{aligned} \tag{2}$$

Considering a matrix $\mathbf{X} \in \mathbb{R}^{N\times P}$ and a N-vector of outputs we get the matrix form residual sum of squares in 2. In 3 we compute the first derivative of RSS.

$$\frac{\partial \text{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} = \mathbf{X}\beta) \tag{3}$$

The second derivative of RSS is equals $2\mathbf{X}^T\mathbf{X}$. If we assume that $\mathbf{X}$ is full-rank, then $\mathbf{X}^T\mathbf{X}$ is positive definite, and we can set 3 to zero, arriving in 4 with an unique solution to $\hat{\beta}$

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \rightarrow \hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4}$$

This solution makes the assumption that the columns of $\mathbf{X}$ are linearly independent. If it is the case that $\mathbf{X}$ is not full-rank, then $\mathbf{X}^T\mathbf{X}$ is singular and coefficients β are not unique, that is, there is no unique solution to OLS [7]. In practice, this means that collinearity in data is generally undesired and confers instability to the coefficient estimations.

2) Maximum Likelihood Estimation (MLE): MLE takes a statistical approach of modeling the regression problem as a likelihood function given by REF, and we want to find the most likely parameters β for the data.

$$p(y \mid x) = \mathcal{N}(y \mid \mathbf{x}^T\beta, \sigma^2) \tag{5}$$

where

$$y = \mathbf{x}^T\beta + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

To find the optimal parameters, the approach taken is often to minimize the negative log-likelihood, since that simplifies the problem algebraic and prevents the problem of numerical underflow [8].

$$\begin{aligned}-\log p(\mathbf{y} \mid \mathbf{X}, \beta) &= -\log \prod_{n=1}^{N} p(y_n \mid \mathbf{x}_n, \beta) \\ &= -\sum_{n=1}^{N} \log p(y_n \mid \mathbf{x}_n, \beta)\end{aligned}$$

Since this model has Gaussian noise term added, we can consider this likelihood Gaussian, and we have an error function as described in 6 which is equivalent to the RSS function 2.

$$\begin{aligned}\mathcal{L}(\beta) &= \frac{1}{2\sigma^2}\sum_{n=1}^{N}(y_n - x_n^T\beta)^2 \\ &= \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)\end{aligned} \tag{6}$$

From 6 we take $\frac{\partial \mathcal{L}(\beta)}{\partial \beta} = 0$ and arrive in the same $\hat{\beta}$ OLS estimation defined in 4.

C. Measures of Performance

The most common performance metric for regression is the Mean Squared Error (MSE) and it is the average squared distance between the predicted values and the true values.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{N}\big(y_i - \hat{f}(x)\big)^2$$

It gives straight forward notion of amount of error and it is related to RSS in the sense that provides a measurement squared distances. Another aspect of the MSE is that it penalizes more heavily the larger the difference between prediction and label. A variation of this metric is the root mean squared error (RMSE), which is equals to $\sqrt{\text{MSE}}$. Its main advantage is that the error measured by it is on the same scale as the predicted variable, making it more interpretable.

The coefficient of determination, also known as R^2, is another goodness of fit metric that instead of measuring accuracy or error rate it provides the amount of variance of the output explained by the variance of the input. There are different formulas, but in general it is a value in the 0 to 1 range representing the amount of variance explained by the model, e.g., a model with an R^2 of 0.75 means that 75% of the variance of the output can be explained by the model [6].

D. The bias-variance trade off

The bias-variance trade off is a dynamic that occurs in linear models that relates the fitness of the model to the flexibility and capacity of generalization. A very simple model that fits loosely to that data is considered to have low variance and a high bias, while a model that models said data very precisely would be said to have high variance and low bias.

We can see that to be on either end of this dynamic would be a problem. A model with a high bias would describe the data very poorly and not have a satisfiable precision. On the other hand a model with very high variance would describe the data so well that the it would be disturbed by the slightest change. In general, the process of fitting involves finding a reasonable compromise between bias and variance.

E. Penalized models

A possible issue with OLS estimations is that it with large correlations between predictors, the general variance of the model becomes large. Penalized models deal with this issue by introducing bias terms in the least squares, which may reduce the overall MSE and make the model more flexible. Other terms that refer to such a practice are shrinkage and regularization, referring the effect produced which is to control the size of the least square coefficients.

One of such models is the Ridge Regression, which minimizes a version of the residual sum of squares with the bias term seen in 7, also called L_2 regularization.

$$\text{RSS}_{L_2}(\beta) = \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \quad (7)$$

This quadratic term is small if β is near zero, makes the β estimations become smaller if their scale is too large, essentially shrinking them towards zero, but never to exactly zero. The parameter λ controls the amount of shrinkage provided.

The regularization may also be done by a $|\beta_j|$ instead of β_j^2. This is called a L_1 regularization and the resulting model is called the Lasso. The main effect of this is that this term may set some variables to exactly zero, performing a variable selection. It is also worth mentioning the Elastic Net model, which uses both L_1 and L_2 penalizations and is a compromise between Ridge and Lasso.

F. The Principal Components Regression (PCR)

The Principal Components Regression (PCR) consists of using the Principal Components (PC) of a data set for dimensional reduction. The PCs correspond to the directions of highest variation in the column space of the data set, and they might hold more predictive value.

For a data matrix $X \in \mathbb{R}^{I \times P}$, we 1. Compute the PC_X 2. Select a subset of them with reasonably high explained variance 3. Apply regression on $m < P$ set of PC_X. The premise is that a subset of the total PCs may hold enough information to fit a linear model with less over-fitting (higher bias).

III. Results

The predictors in the concrete data in general do not have large correlations among themselves. Its visible in Figure 2 that the most prominent are between superplasticizer and flyash and water. Collinearity is not so likely to be an issue in the linear models used. In Figure 1 we can see that there are variables with predictive value. Cement, superplaticizer and age have shown higher correlations with the output but some info seems to be present in other predictors as well.

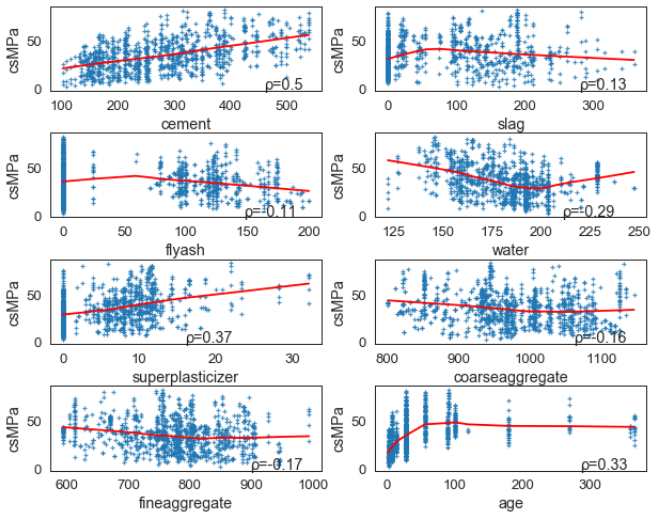

Fig. 1. Correlation between predictors and csMPa

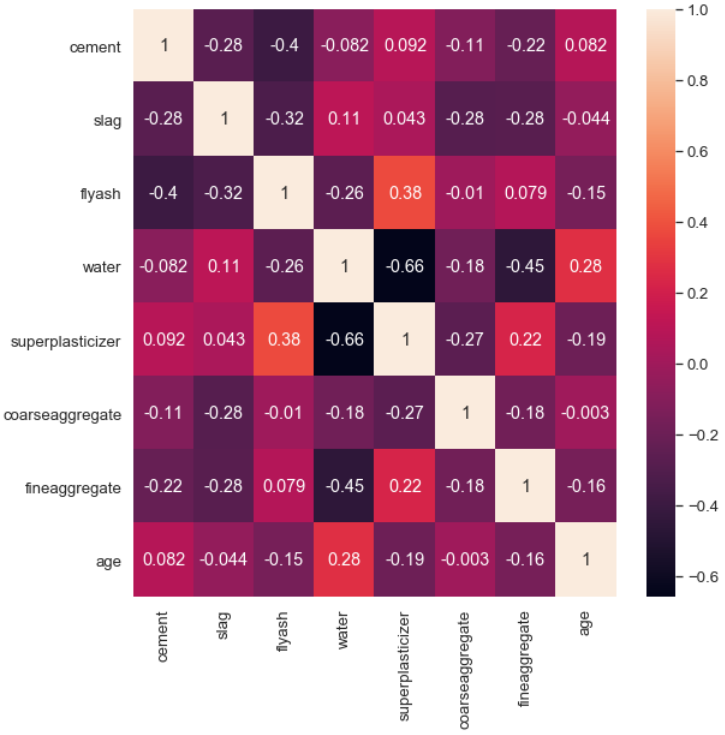

Fig. 2. Correlation matrix of the predictors

A multi-variate analysis was conducted to further explore the possible correlations. A Principal Component Analysis (PCA) of the predictors has shown that the variance explained by the PCs is spreaded out, as can be seen in figure 3. The first and second PCs correspond to a variance of approximately 0.463 and a cumulative variance above 0.95 is reached with

at least 6 PCs. This is expected since there isn't a prevalent collinearity among the predictors. In Figure 3 we can also see that there isn't a clear relationship between the first PCs and the response, and no clear pattern in between PC_1 and PC_2.

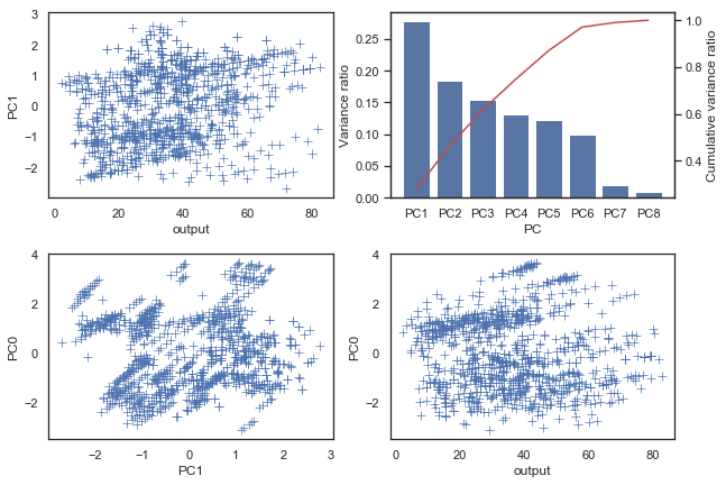

Fig. 3. Principal components of Concrete data

Three different model approaches were fit, ordinary linear regression, ridge regression and PCR and validated the results via 10-fold cross validation. The λ factor of the L_2 penalization was also validated via 10-fold cross validation and the optimal value found was 5.45. This was choosen from a 10 value sweep from 5 to 6, after finding that values out of this range did not improve the error. From the conclusions of the multivariate analysis, the number of PCs used for the PCR was 6.

A separate fitting of an OLS model was made to evaluate some general characteristics of the model. The intercept obtained as 35.75 and the three highest weights were related to age, cement and slag, being 10.08, 9.17 and 5.54 respectively. Near zero weights were set for fly ash, fine aggregate and coarse aggregate.

Figures 4 shows the RMSE and R^2 values for the 10-folds. There is a clear difference in responses for each fold, indicating some instability in the model. It is also clear that both Ridge and PCR did not make visible improvements to the linear model.

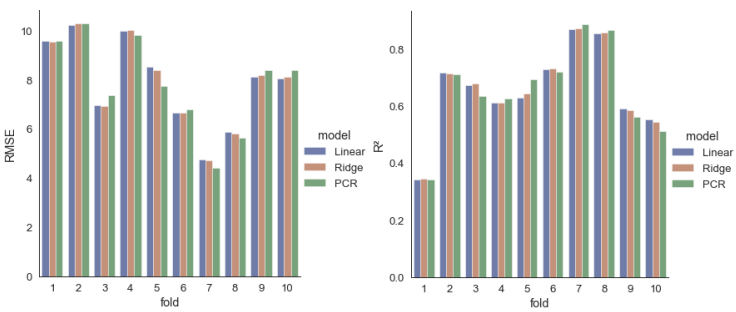

Fig. 4. RMSE and R^2 for each fold

The fold wise average RMSE and R^2 is in Table 2, and it shows that the error has diminished with Ridge and PCR, although very slightly.

TABLE II
FOLD WISE AVERAGE RMSE AND R^2

	Linear Regression	Ridge	PCR
RMSE	7.868	7.856	7.836
R^2	0.656	0.657	0.655

IV. CONCLUSION

In this study we have evaluated the use of linear regression and some of its variants for the prediction of concrete strength and have concluded that while having some predictive power, the simplicity of linear modeling has flaws in giving a satisfying accuracy.

Penalized and reduced models have reduced the overall average error slightly, this is likely due to the bias inserted by such approaches, which may allow for better generalizations. However, this approach does not seem very fruitful, as the differences were not stable through the folds.

A more promising direction for modeling such data would be to introduce more complex models that may capture the predictor relationships more completely, such as neural networks. In his study, I.-C. Yeh has not provided figures of straight forward error measurements, but his test results for R_2 obtained with neural networks range from 0.855 to 0.922, showing how complexity may improve the modeling.

REFERENCES

[1] L. Sear, J. Dews, B. Kite, F. Harris, and J. Troy, "Abrams law, air and high water-to-cement ratios," *construction and Building Materials*, vol. 10, no. 3, pp. 221–226, 1996.

[2] F. A. Oluokun, "Fly ash concrete mix design and the water-cement ratio law," *Materials Journal*, vol. 91, no. 4, pp. 362–371, 1994.

[3] I.-C. Yeh, "Modeling of strength of high-performance concrete using artificial neural networks," *Cement and Concrete research*, vol. 28, no. 12, pp. 1797–1808, 1998.

[4] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[5] M. Mulas, "Data analysis and pre-processing," November 2020.

[6] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[7] J. Friedman, T. Hastie, R. Tibshirani *et al.*, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1, no. 10.

[8] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.