Composition of concrete and its influence on compressive strength

Filipe P. de Farias
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
filipepfarias@fisica.ufc.br

Yvo J. M. Sales
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
yvo@gtel.ufc.br

***Abstract*—The compressive strength of a concrete is a important property since it impacts directly on its applications. The classical approach to obtain the compressive strength of a specific concrete mixture is to submit a sample to a test on a hydraulic press. However, it takes time to perform this type of test since it is necessary to wait the sample to cure. In this work, we try to find a regression model to accurately estimate the compressive strength of a concrete mixture from the concentration of its components.**

***Index Terms*—Regression models, concrete compressive strength, machine-learning, partial least squares.**

I. Introduction

A material formed by aggregates bonded together by a fluid material that hardens over time has been used by humans for construction since many years ago [3]. Nowadays this material is known as concrete and it's widely used in the construction field. The aggregates used in the mixture of the concrete affect directly its compressive strength which highly impacts its applications. For instance, in general, the concrete for columns or beams needs to have a greater compressive strength than the one for pavement. On the previous work, we have made a statistical analysis on a dataset extracted from the UCI Machine Learning Repository (University of California, Irvine) [4] that collects information about the concentration of some aggregates used to form different mixtures of concretes. In this work, we try to find a regression model to estimate the relation between the concentration of those aggregates and the strength of the concrete mixture. The goal is that such model could be a good replacement to tests of samples on hydraulic press.

This work is divided as follows. A description of the data is given in Section II resulted from the previous work with the addition of the regressor (concrete compressive strength). Section III brings a brief introduction to the regression models that will be used to fit the data. In the sequel, we present and discuss the results in Section IV. Finally, the conclusions and considerations are exposed in Section V.

II. Data description

The composition of each one of the N concrete samples is given by the concentrations (kg/m^3) of D components: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate and Fine Aggregate. The cement is what binds the elements of the concrete together. Indeed his technical name in the literature is *binder* [8]. The other components as blast furnace slag and fly ash, the outcomes of another industrial process reused in the concrete mixture, they have the role of increase the chemical hardness of the concrete, i.e. in a microscopic level. The water is responsible for react with the cement resulting in the cement stone. The superplasticizer gives fluid characteristics to the concrete aiming to better fill the mold and decrease the use of water. The coarse and fine aggregates give some macroscopical mechanical resistance to the concrete but can reduce its compressive strength if bad applied. Their major role is to occupy the spaces in the mold reducing the use of cement. The output is the concrete compressive strength which is measure in the stress test where a force its applied to a sample using a hydraulic press. When the sample reaches rupture the pressure, force per area of the sample, is observed.

TABLE I
Data description

Label	Component	Unit
D_1	Cement	kg/m^3
D_2	Blast Furnace Slag	kg/m^3
D_3	Fly Ash	kg/m^3
D_4	Water	kg/m^3
D_5	Superplasticizer	kg/m^3
D_6	Coarse Aggregate	kg/m^3
D_7	Fine Aggregate	kg/m^3
D_8	Age	days
$\mathbf{D_9}$	**Compressive strength**	**MPa**
Total	$N = 1030$ samples	

The observations are the measured compressive strengths of each sample and, as the predictors $D_1 - D_7$, are continuous. The Age (D_8) of the concrete is extremely discrete. All the data was normalised, by centering at the mean and scaling by the standard deviation to avoid any of the methods to be sensitive to different scales. At Fig. 1 we note the strongest positive correlation is between the strength and the cement component ($D_9 \times D_1$). Another important factors are the presence of the superplasticizer ($D_9 \times D_5$) and the age that

represents the time of cure ($D_9 \times D_8$). In all these components is noted a subtle correlation. And the most important fact that can be observed is the decrease of the necessity of water when the superplasticizer is used ($D_5 \times D_4$), which was its proposal in the first place.

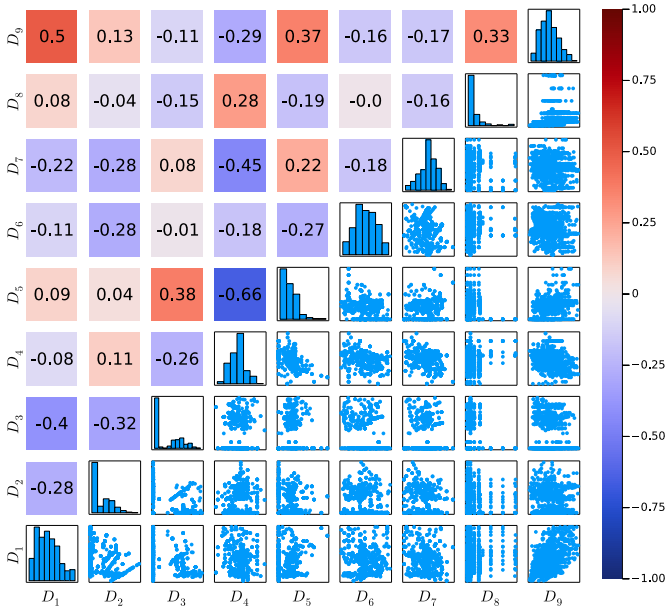

Fig. 1. Pairwise scatter, correlation and histogram plots of each concrete component.

III. Regression Models

Regression models try to find relations between the *independent variables* and the *dependent variables*, which are named, respectively, predictors and outcomes in this work. These relations can occur in different forms. The simplest one is the linear relationship, which is when the curve predictors *vs* outcomes, in the case that both are one-dimensional, forms a simple line and, in the general case, a hyperplane. In what follows, we formulate the *Linear regression*, which is a subclass of regressions dedicated to find a linear model to explain the relation between the predictors and the outcomes.

A. Linear regression

$$Y = \begin{bmatrix} \beta_0 & \boldsymbol{\beta}^\top \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ X \end{bmatrix} + \varepsilon \tag{1}$$

This method tries to find the linear regression between the predictors and outcomes by fitting a line with linear coefficient $\boldsymbol{\beta}$ and angular coefficient $\boldsymbol{\beta}$, defined in Eq. (1), through the data. $\boldsymbol{\beta}$ is the size such as the number of predictors in X, in the way we can define the vector $\boldsymbol{\beta} = [\beta_0, ..., \beta_D]^\top$. Wrapping up $[\beta_0 \ \boldsymbol{\beta}]^\top$ is a $1 \times (D+1)$, $\mathbf{x} = [\mathbf{1} \ X]^\top$ is a $(N+1) \times D$ matrix and Y is a $N \times 1$ matrix. For limitations on the implementation, it was adopted the label D_9 for the outcome, then the dimension D of the matrices must be considered without the outcome component, that is 8 components.

The fitting, i.e. finding the values for each of the linear and angular coefficients, is done by minimising a cost function that can take different forms. Each cost function yields to different optimal parameters and two of them are described in the following, the *Ordinary least squares* and *L_2-penalized least squares*. An interesting fact to observe is when the predictors correlates with the outcome, we can observe the error will be small. This is because the correlation evaluates linear variations between the variables as such as the linear regression.

The *Ordinary least squares* defines the cost function to find the optimal parameters for Eq. (1) as

$$L(Y, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2. \tag{2}$$

This cost function represents a quadratic distance between the model, i.e. the linear combination of the coefficients and the data. We have the minimal distance, we obtain the best model. This minimisation can be done by differentiating Eq. (2) w.r.t. the β's giving the coefficients $\hat{\boldsymbol{\beta}}$ of the best model by

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top Y.$$

The *L_2-penalized least squares* modifies Eq. (2) adding a term that penalize large values of the parameters, yielding to

$$L(\mathbf{y}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{3}$$

where λ is the penalisation coefficient, a tuning parameter. Different from ordinary least squares, we find λ such the model will be the best one with *cross-validation*. The values of the coefficients for a given λ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{x}^\top \mathbf{x} + \lambda I)^{-1} \mathbf{x}^\top Y.$$

B. Partial least squares

The idea of the partial least squares (PLS) is in the same line of the previous linear regression of Eq. (2). But when $\mathbf{x}$ is *ill-conditioned*, i.e. the features are strongly correlated [5]. In terms of the linear algebra theory, our matrix can have lines (rows or columns) that are scaled versions of each other hence it is rank deficient.

Now what is done is to project the predictors into a lower-dimensional predictors space, in such way that the resulting "new predictors" are a linear combination of the original ones. Then we do the regression using this new predictors. This projection is done also considering the information of Y and not just obtaining the space which $\mathbf{x}$ has its variance maximised as in *principal component regression* (PCR). In PLS we need to find $\mathbf{w}$ and $\mathbf{c}$ in Eq. (4) to create a linear combination to consider that information of Y in a such way that we obtain the maximum covariance. The process of finding these matrices is iterative and described by Indahl [6].

$$\mathbf{t} = \mathbf{x}\mathbf{w} \text{ and } \mathbf{u} = Y\mathbf{c} \text{ with } \mathbf{w}^\top \mathbf{w} = \mathbf{t}^\top \mathbf{t} = 1. \tag{4}$$

IV. Results

In this section, we present the results of fitting the *Ordinary least squares*, the *L_2-penalized least squares* and the *Partial least squares* in the data. Each one of the methods was subjected to 70% Train / 30% Test and 5-fold cross-validation strategies. In the first, 70% of the data was used to train the models and the rest for test. In the 5-fold the data was folded into 5 groups equally divided and one of them was selected as test group and the other as training group. This step was repeated until all of the 5 groups be used for test.

TABLE II
Ordinary Least Squares Regression Summary

CV	RMSE	R^2
70% Train / 30% Test	0.614014	0.612271
1-st fold	0.614187	0.634496
2-dn fold	0.647336	0.597003
3-rd fold	0.5952	0.599484
4-th fold	0.686377	0.530463
5-th fold	**0.582266**	**0.661407**

TABLE III
L_2-penalized Least Squares Regression Summary

CV	RMSE	R^2
70% Train / 30% Test	0.599004	0.640309
1-st fold	0.613885	0.634857
2-dn fold	0.647344	0.596994
3-rd fold	0.594722	0.600127
4-th fold	0.68599	0.530993
5-th fold	**0.582357**	**0.661302**

Tables II-V summarize the values of RMSE and R^2 resulted from the fitting of the data by the linear models for the 70% Train / 30% Test and 5-fold cross-validation strategies. In all three tables, the smallest value of RMSE and the largest value of R^2 are highlighted.

Since the values of R^2 are approximately at most 66%, none of the models could explain more than this value of the variance of the data. This result suggests that the relation between the predictors and the outcome is not linear, thus a non-linear model such as neural networks, as used by [4], could yield to a better fitting.

Although the values of RMSE for the applied linear models are small, we should keep in mind that the data is normalised with mean 0 and variance 1. Therefore, in this situation an RMSE of 0.58 implies to an error of more than half the variance. However, the models with the smallest values of RMSE explained better the variance of the data. Figure 2 and Figure 3 show the values of the angular coefficients of each predictor obtained, respectively, after the 70% Train / 30% Test and 5-fold cross-validation strategies. Notice the highest value of the coefficient is of the Cement D_1 and the smaller one the Water D_4. To point some real world meaningful relation to this, some previous study on the specifics of theses components is needed, but some literatures [1] refers at least to the cement as the most important compound to the strength of

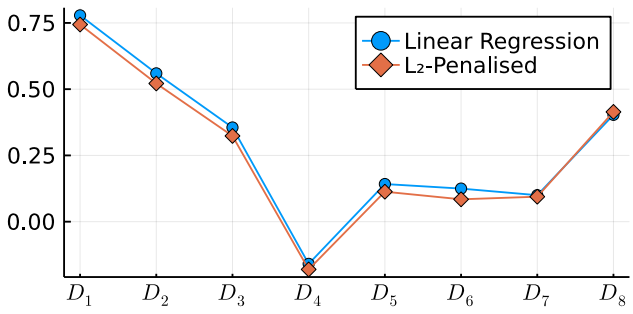

Fig. 2. Values of the angular coefficients of each predictor obtained after 30%/70% cross-validation strategy.

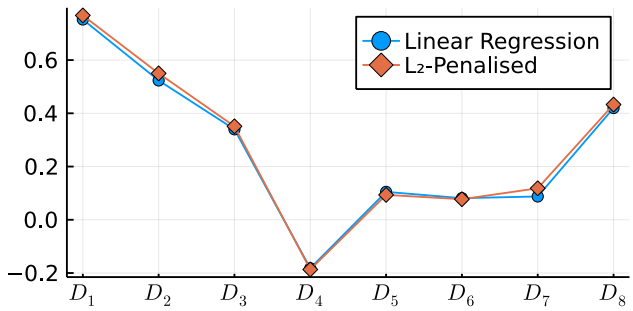

Fig. 3. Values of the angular coefficients of each predictor obtained after 5-fold cross-validation strategy.

the concrete. But some knowledge needs to be applied because this increase in the amount of cement is bounded, then this do not always leads to a better strength value [4].

For the L_2-penalized model, the λ founded was 0.71968 for the minimum RMSE. The values of RMSE for the λ values evaluated are on Fig. 4 in which we note the model looses its flexibility with the increase of λ leading to a greater error.

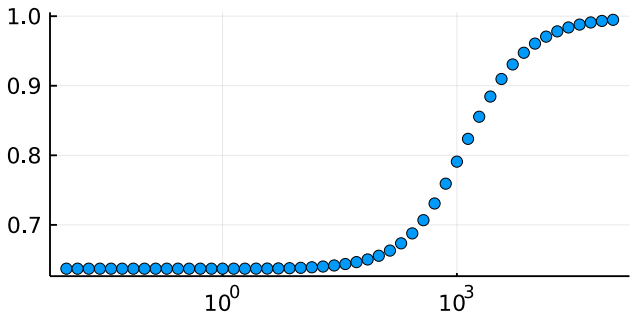

Fig. 4. RMSE (y-axis) values for λ (x-axis) parameter of the L_2-penalized least squares.

The same intuition can not be done for the PLS results in the Table IV given the space of its regression is another from the one the same as the compounds of the concrete are known. But a similar result to the previous work in which the explained variance is more disperse in the other principal components is

TABLE IV
VARIANCE EXPLAINED OF TH 5-TH FOLD SUMMARY

Principal Components	Variance Explained (%)
PC1	49.44
PC2	16.9
PC3	10.22
PC4	8.15
PC5	6.64
PC6	5.84
PC7	2.26
PC8	0.55

noticed.

TABLE V
PARTIAL LEAST SQUARES REGRESSION SUMMARY

CV	RMSE	R^2
70% Train / 30% Test	0.599264	0.639996
1-st fold	0.614106	0.634593
2-dn fold	0.646593	0.597928
3-rd fold	0.594709	0.600143
4-th fold	0.6859	0.531116
5-th fold	**0.58221**	**0.661472**

The same analysis made for the summaries of the previous models can be applied to the PLS which the results are in Table IV. The model was trained under the same cross-validation strategies yielding to the best model found with the 8 PC's used.

V. CONCLUSIONS

The database of concrete is not easy to analyse if there is no previous knowledge about the problem of the components mixture, as was noticed in the previous work and a more deep literature research was needed. Although the models worked, the authors were not capable to determine if the errors found imply in safe conditions for the concrete, neither is the objective of this work. What we can conclude is that the model is not linear given the R^2 statistics and the original work of the data [4].

REFERENCES

[1] ACI Manual of Concrete Practice 2000, Part 1: Materials and General Properties of Concrete. American Concrete Institute. Farmington Hills, MI.

[2] Tibshirani, Robert, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Germany, Springer New York, 2009.

[3] Mindess, S., and Young, J.F. Concrete. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[4] I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)

[5] Abdi, Hervé, "Partial least squares regression and projection on latent structure regression (PLS Regression)," WIREs Computational Statistics, 2010-01-01 2(1): 97-106.

[6] Ulf G. Indahl, "The geometry of PLS1 explained properly: 10 key notes on mathematical properties of and some alternative algorithmic approaches to PLS1 modelling," Journal of Chemometrics, 2013.

[7] I-Cheng Yeh, "Prediction of Strength of Fly Ash and Slag Concrete By The Use of Artificial Neural Networks," Journal of the Chinese Institute of Civil and Hydraulic Engineering, Vol. 15, No. 4, pp. 659-663 (2003).

[8] Khasanov, Irmuhamedova, et al, "Theoretical foundations of the structure formation of cement stone and concrete," IOP Conf. Series: Materials Science and Engineering 869 (2020)