MATLAB project report - group 19

Tommaso Borelli (251831), Francesco Jin (273941),

Maria Chiara Lischi (271281), Nadeer Salem (274491)

December 9, 2022

Contents

Abstract

We consider measurements of 100 sample of concrete with the aim to predict its compressive strength using the measurements of the ingredients that constitue it: cement, water, coarse aggregate and fine aggregate.

1 Task 1

Implementation and analysis of four linear models: one for each regressor

At first, we tried to predict the strenght of concrete using as regressors the single components, seeking for a linear relationship. The result of this first attempt are the following scatter plots, from which is clear that there is no linearity between the response variable and each of the regressors. The only regressor that seems to have a remote linear relationship with the strenght is the cement, but it cannot be considered as a good model.

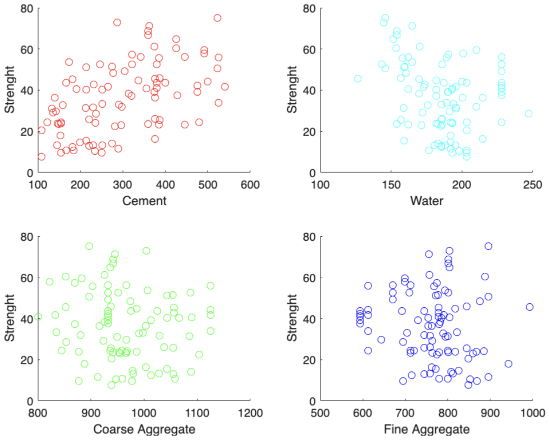

Figure 1: regressors vs regressand

Those results are actually not surprising, because of to the nature of the dataset. It is reasonable to the the strenght we aim to forecast is predictable basing the proportions of the single ingredients in the sample, more than on the quantity of a single component.

Below, the regression line for each of the four linear model analysed. As mentioned before, it is clear that single linear models are not the best ones to predict the strenght of the concrete sample.

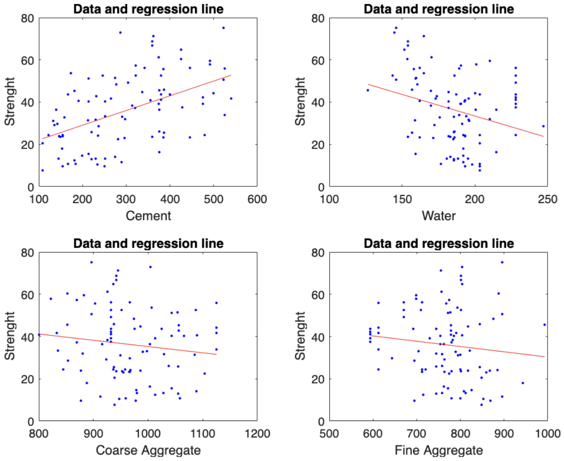

Figure 2: regression lines

This can also be seen from the errors by which we evaluated our models:

- Model with cement as regressor:

least-squares error:	141.782
residual sum of squares:	20102.107
mean square error:	211.601
root mean square error:	14.547

- Model with water as regressor:

least-squares error:	155.816
residual sum of squares:	24278.571
mean square error:	255.564
root mean square error:	15.986

- Model with coarse aggregate as regressor:

least-squares error:	161.429
residual sum of squares:	26059.232
mean square error:	274.308
root mean square error:	16.562

- Model with fine aggregate as regressor:

least-squares error:	161.598
residual sum of squares:	26113.942
mean square error:	274.884
root mean square error:	16.580

From the comparison of these parameters, it emerges that coarse aggregate and fine aggregate presents very similar (high) errors, while water and cement seem to perform a little better. The model using cement as regressor, in particular, is the best one overall, with relatively low errors with respect to the others. Water is present with the same negative slope as the aggregates, but with a relatively lower lse, and is the second best model.
It is clear that our models are not reliable for forecasting the strenght. This is also confirmed by the fact that the predictions of the models made on the test data are far from being considered reliable for each of the regressors.

2 Task 2

Implementation and analysis of a multiple linear model considering all the regressors

We want to consider now a multiple linear model, having as regressors all the four predictor variables. From the implementation of the model, the values we obtained as coefficients are:

- $\beta_0 = 237.7899$
- $\beta_1 = 0.0423$
- $\beta_2 = -0.4218$
- $\beta_3 = -0.0719$
- $\beta_4 = -0.0870$

Our β_0 represents the intercept of the line with the y-axis. For the slope of each predictors we observe that the only one with a positive value is the cement, meaning that for a one unit increase in cubic meters of cement, there will correspond a 0.042 increase in the streght of the sample. Instead, all the others regressors have a negative sign of the slope, meaning that for each unit increase of each of them, correspond a decrease in the overall strenght.

Implementation of a multiple linear model considering three of the regressors

Looking at the linear models computed in task 1, we decided to propose a linear model with three predictors. We want to use as regressors:

- the cement, that is the best regressor found so far, and has a strong impact on the response ($\beta_1 = 0.0423$)
- the water, since is the second best regressor and also have a strong impact in the response ($\beta_2 = -0.4218$)
- the coarse aggregate. The reasons for which we prefer the coarse aggregate to the fine aggregate are the lower RMSE (16.56 rather than 16.58), the fact that it forecsted better the strenght of the test data, and that it has a lesser negative effect on our current Multiple linear regression model.

Implementation and analysis of others non-linear models

Analysing the four scatter plots that we have obtained in task 1 emerged that there is no linear relationship between the predictors considered alone and the response variable. We want now to approximate the data with some polynomial models, using as predictors the cement and then the water, which have proved to be the best regressors over the four available.
For the first models, that use as regressor the cement, we started by sorting in ascending order the dataset accordingly to the cement values. We then performed several polynomial regressions using the "polyfit" function with different degree of polynomial fit: precisely 2, 3 and 4. Once we had the coefficients for the different fits, we fitted those values to the cement variable and computed the R^2 for each of them. The model presenting the best R^2 value is unsurprisingly the one with the highest fitting degree (0.2665). Above the graph representing the polynomial model of degree 4 with cement as regressor:

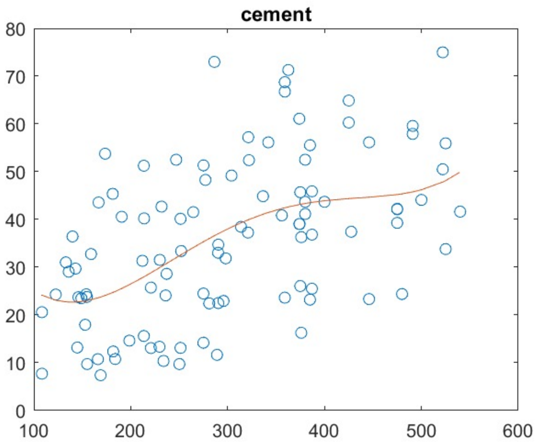

Figure 3: cement polynomial

For the other models, using water as regressor, we performed the analogous operations, finding that we had a better R^2 (0.2706) than before for the model with the highest degree polynomial (also this time, 4). Above the graph representing the polynomial model of degree 4 with water as regressor:

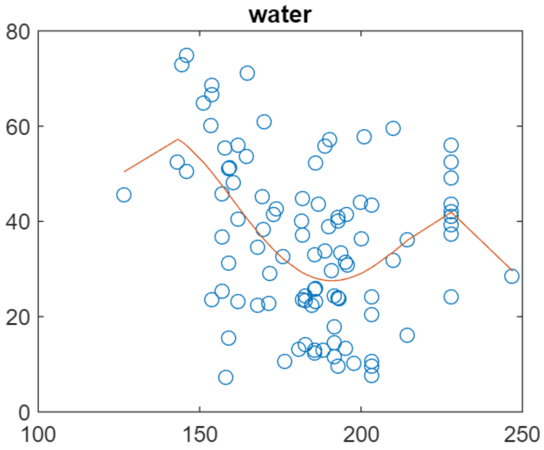

Figure 4: water polynomial

The best model, according to the root mean square error is the multivariate linear regression which considers all the four regressors with an $RMSE$ of 12.7474, followed by the multivariate linear regression considering all but the fine aggregate, with an $RMSE$ of 13.5245. Coming last, in order we have the water polynomial regression ($RMSE = 14.2019$) and the cement polynomial regression ($RMSE = 21.0045$).

3 Task 3

We want now to analyse the dataset features according to the regressors.
The following graphs represent the proportion of total sample variance represented by each eigenvalue. We calculated that the first eigenvalue covers 53% of the sample variance, and together with the second one a total of 82% of the sample variance is covered.

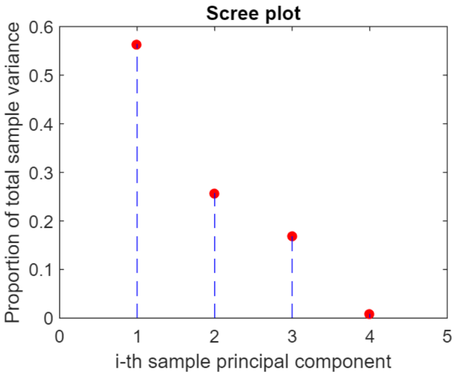

Figure 5: scree plots

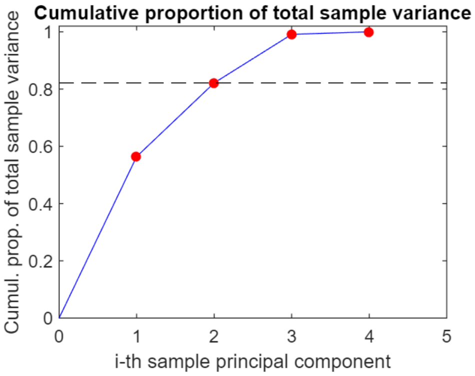

Figure 6: cumulative proportion of total sample variance

It is enough to use only the first two sample principal components because they describe more than the 82% of the dataset's variability (which is an high percentage considering that in the principal component analysis we want to balance information gain with the use of least amount of variables). For this reason, it makes sense to use the first two components for the loadings.
The first loading is $y_1 = -0.92_{x_1} - 0.03_{x_2} + 0.02_{x_3} + 0.38_{x_4}$. First we have noticed that the first two elements (cement and water) have negative sign whereas the aggregates are positive; also, the cement and the fine aggregate have a stronger impact with respect to the others.
The second loading is $y_2 = -0.27_{x_1} - 0.09_{x_2} + 0.69_{x_3} + 0.66_{x_4}$. In this second loading we have noticed that the cement and fine aggregate have negative sign whereas water and coarse aggregate are positive; also, in this case the components having the strongest impact are the two aggregates.
The scores of all the principal components are reported below:

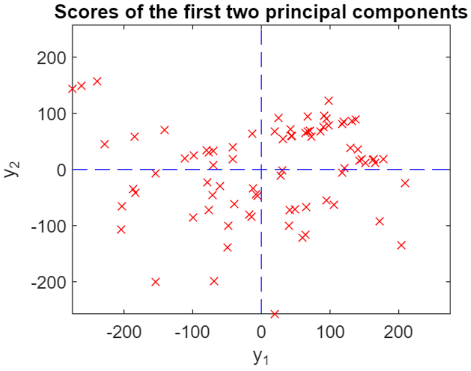

Figure 7: sample principal components

We notice that:

- Points in the first quadrant have both first and second principal component positive.
- Points that lie in the second quadrant have negative coordinate of the first sample principal component and positive coordinate for the second one.
- Points that lie in the third quadrant have negative coordinates of both the sample principal components.

- Points that lie in the fourth quadrant have positive coordinate of the first sample principal component and negative coordinate for the second one.