Studying Optimal Concrete Production Using Regression Analysis

Professors - Bruno Damásio, Carolina Vasconcelos, Beatriz Sousa, Carolina Shaul

Dinis Gaspar 20221869, Dinis Fernandes 20221848, João Capitão 20221863, Luis Mendes 20221949

Introduction

[1] Concrete is one of the most crucial materials in the world. It is widely used across the globe with many different applications. For this reason, finding adequate ways to judge which components are more crucial in a mix of concrete to obtain a stronger final product is a very important task. In this project, we will show how a regression and its associated analysis can be used to gain some of these insights.

In this project we will aim to answer 2 questions specifically:

- **1)** Which controllable factors influence the Concrete compressive strength?
- **2)** What is the effect of an additional day in the Concrete compressive strength?

Data

The dataset that we will use [2] includes information on various factors associated with the production of different types of concrete as well as their compressive strength, which will be our target variable.

A summary of the variables and their units can be found below:

Feature	Description	Unit	Type
Cement	quantitative	kg in a m3 mixture	Input Variable
Blast Furnace Slag	quantitative	kg in a m3 mixture	Input Variable
Fly Ash	quantitative	kg in a m3 mixture	Input Variable
Water	quantitative	kg in a m3 mixture	Input Variable
Superplasticizer	quantitative	kg in a m3 mixture	Input Variable
Coarse Aggregate	quantitative	kg in a m3 mixture	Input Variable
Fine Aggregate	quantitative	kg in a m3 mixture	Input Variable
Age	quantitative	Day (1~365)	Input Variable
Concrete compressive strength	quantitative	MPa	Target Variable

It is worth noting that variables such as temperature and humidity, that could be relevant predictors, are not present in our data. This, however, does not create any issues as these variables are, we believe, not correlated in any way to the variables that we have.

Methodology

Firstly, we plotted all variables against our target to look for clear indicators of transformations that would be required. No transformations were obvious.

Assumptions:

Before creating any models, it is important to look at potential issues with the classical assumptions of the Multiple Linear Regression model.

By using the lm function in R we don't have to worry about **MLR1 (linearity in parameters)** , we also have no problems with **MLR2 (Random Sampling)** . After checking the correlations between our variables, we can see that we don't have perfectly correlated variables, thus respecting **MLR3 (No perfect multicollinearity)**. All potentially explanatory variables for our target are either accounted for in the model or uncorrelated from those that are included [3], this ensures **MLR4 (Zero conditional mean)**, we will **test for heteroskedacity** and deal with it as necessary **(MLR5)**. As we have a **large sample** (1030 observations) we have **asymptotic normality of residuals and estimators MLR6**. Now we can proceed with modelling.

Modeling:

We started by estimating a model with all the independent variables in our dataset. We then tested this model for heteroskedacity (doing the Breusch-PaganTest) which allowed us to conclude that we did have heteroskedacity in our model (because we obtain a p-value of 2.2e-16 in the test, rejecting h0 at 5% significance level, meaning that the variance of our residuals are not constant). With this in mind, **we will use heteroskedacity robust estimators while performing statistical tests on our model**. As we have a large sample (1030 observations), this is the only concern for us about the veracity of our tests.

The first tests we performed were t-tests and these allowed us to conclude that we have two individually insignificant variables: 'Coarse Aggregate', 'Fine Aggregate'. Before deciding to remove them or not we will perform an F-test (wald test, since it uses robust estimators) to evaluate whether they are also jointly insignificant. This test allowed us to conclude that they were, but as they are correlated with the remaining explanatory variables, they will not be removed to ensure that the zero conditional mean assumption is not violated.

With all this, in mind we will now perform a manual **RESET test** (because of heteroskedacity) using the auxiliary regression and a wald test to see if our model is defined adequately. Initially, our **model was not specified correctly** (since we obtaining a p-value of 3.537e-12), but after some transformations we were able to **not reject the null hypothesis of the RESET test at 5% significance level (obtaining a p-value of 0.1184), which means we now have a well-defined model**.

This corrected model includes all our variables, cement as a logarithmic variable, as well as a quadratic term for the age variable, as literature suggests that the impact of this variable diminishes over time.

With our final model, we re-tested all of the hypothesis tested above to see the actual outcomes, and the results were the followings.

Results

The final model:

Concrete compressive strength(estimated) = -87.63 + 28.207 **log(Cement)** + 0.084537 **Blast Furnace Slag** + 0.05926 **Fly Ash** - 0.23254 **Water** + 0.15194 **Superplasticizer** - 0.007255 **Coarse Aggregate** -0.0075929 **Fine Aggregate** + 0.35285 **Age** - 0.00081934 **Age**^{2}

$$n = 1030,\ R^2 = 0.7416$$

Significant Variables:

- **t-test:**

Using our well specified model and the robust estimators we obtain in our t-tests that **all variables were significant except for 'Fine Aggregate', 'Coarse Aggregate' and 'Superplasticizer'** (with a p-value of 0.41160, 0.36748 and 0.10598 respectively, not rejecting h0 at 5% significance level), now we need to test if they are jointly insignificant.

- **F-Test:**

This test allowed us to conclude that they were **jointly insignificant** (since we obtain a p-value of 0.09033, not rejecting h0 at 5% significance level), but as they are correlated with some of the remaining explanatory variables, they **will not be removed from the final model to ensure that the zero conditional mean assumption is not violated.**

Conclusion

We will now answer the questions we outlined at the start.

Firstly, we can conclude that out of the variables in our dataset, **Cement, Blast Furnace Slag, Fly Ash, Water and Age are controllable factors that influence the compressive strength of concrete.**

We can also conclude that it would take 215 days for the effect of adding an extra day to become negative, the return of age (in days) to start decreasing; that is because **the partial effect of the age variable influences the Concrete compressive strength by: 0.35258 - 2 * 0.00081934 * age,** in other words if age increased by 1 day, the Concrete compressive strength will increased 0.35258 - 2 * 0.00081934 * age megapascal.

References

- 1: Concrete Compressive Strength Definition.
- 2: Concrete Compressive Strength Dataset.
- 3: Concrete Compressive Strength Factors.