Overview

You have been hired to optimize the production process of a company that makes concrete. They provided you a dataset (Concrete.jmp) containing information on 500 different batches of concrete that have been produced in the past. The dataset contains information on the amount used for seven different ingredients, as well as the age of the concrete. The variable of interest is Concrete compressive strength (measured in megapascals).

For background information, see the following reference (especially the first three paragraphs):

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).

Part I- Draft Due April 7, 2017

Begin by analyzing the distributions of the variables in your dataset. Create boxplots and histogram plots and briefly describe the important features and characteristics of each variable's distribution. Do any of these variables exhibit features that might complicate your analysis? Use scatterplots to examine the relationship between variables. Are any variables highly correlated with others? If so, in which direction and how strong is the linear relationship? Your report should include a few paragraphs describing the variables in the dataset. Include only figures that you believe show something important.

The Abrams water-to-cement ratio (w/c) pronouncement of 1918 asserts that the lower the water to cement ratio, the stronger the concrete will be. According to Yeh (1998), this has been described as the most useful and significant advancement in the history of concrete technology. The principle also implies that ingredients other than water and cement are mostly irrelevent. For this part of the project, assume that this is true, and that only the water to cement ratio needs to be considered when modeling strength of concrete.

The company currently uses a ratio of 1.25 kg. of cement for every 1 kg. of water used. They are considering increasing this ratio to 1.5 kg. of cement per 1 kg. of water and would like to know the impact that this change will have on the strength of their concrete. Investigate the estimated impact of this change and provide a report for the company. Do you recommend this change? (Hint: you will probably want to create a new variable in the dataset representing the cement-to-water ratio.)

In addition to your recommendations,justify the appropriateness of your techniques. If you have any concerns about the validity of the assumptions behind your model, report these to the company along with your findings. Include appropriate graphics, with captions, to illustrate your conclusions and justify your assumptions.

Your report should include

- A letter to management describing your work and summarizing your most important recommendations. The letter should not exceed 1 page but should be sufficiently informative to management. Your language should be non-technical and easy for some with little statistics knowledge to understand. It should catch the company's attention and motivate them to read the rest of your report. In the final version, this letter will address both parts I and II of the project.
- A detailed report which will be read by the company's statisticans and engineering experts. This should include
  - An introduction describing your task
  - A description of all variables with appropriate plots.
  - A discussion of the simple linear regression model you used.
  - A discussion of the results of your model and interpretation of parameter estimates.
  - An analysis of the simple linear regression model assumptions, and appropriate plots, tests, and discussion.

Part II - Due April 21, 2017

The company wants to know whether Abrams' water-to-cement ratio is overly simplistic and whether ingredients other than cement and water impact the strength of concrete, and if so, by how much. If it is true that strength depends almost entirely on water-to-cement ratio, then the company can save money by not purchasing ingredients such as Blast Furnace Slag, Fly Ash, and Superplasticizer.

Using as many variables as you deem appropriate, develop a multiple regression model for strength of concrete, to investigate. Your model should be complex enough to accurately predict concrete strength (to the greatest extent possible), but simple enough to explain to the company. (Hint: You might want to consider creating new variables from the ones contained in the original dataset, as you did in Part I.) Provide the company with an interpretation of all regression coefficient in your model. If the interpretation of one or more of these coefficients does not make sense, explain why, and tell the company what it needs to know about how these variables impact the strength of concrete. Are variables other than cement and water important? If so, which are most important? Do you recommend the company save money by only using water and cement? Explain your reasoning to the company.

Your report should include a discussion of how you decided what variables to include in your model. You should assess the validity of assumptions involved in multiple regression, and should note any potential concerns about the validity of your model, such as multicollinearity, and overfitting.

In order to test the effectiveness of your model, the company provides you information on 100 more batches of concrete that have already been made, but does not tell you the hardness (Concrete_test.jmp). Using your model, predict the hardness of these batches and give a range of reasonable values for the hardness of each batch. The company will then compare your predictions to the true hardness for these batches.
Each prediction will be scored as follows:

$$\begin{aligned}\text{Score} =&(\text{Predicted} - \text{True})^2 + \text{Width of Interval(if interval contains true value)} \\ &+ 100 \text{ (if true value is not contained in interval)}\end{aligned}$$

An additional 200 points will be added to your total score for each term in your model.

Whether the company follows your advice and ultimately retains your services depends on the quality of your predictions. **For this project, the team with the three lowest scores will earn 5, 3, and 1 bonus point, respectively on the project.**

In your report to the company

- Add your most important recommendations to the recommendation letter from part I. Together, this letter should not exceed 1 page.
- Create a new section of your report for part II containing:
  - A detailed discussion on how you decided which variables to include in your multiple regression model.
  - Interpretation of multiple regression coefficients and recommendations that the company's managers (who are not statisticians or engineers) will be able to understand.
  - Appropriate checks of model assumptions and a discussion of the validity of your model.
  - A .jmp file containing your predicted values and intervals for the 100 additional batches of concrete (submit electronically).