LAB 2 PROPOSAL

Amit Gattadahalli, Rebecca Hile, and Tres Pimentel

1. Context

Concrete is the most popular building material in the world. It is produced using construction aggregate (e.g. gravel, sand) and cement. Using different ratios of these aggregate types and cement in production can provide different benefits to the concrete mix - strength, affordability, or a reduction in emissions. We are interested in assessing what kind of mixture can be used to produce concrete solutions of appropriate strength given the goals of a project.

2. Research Question

We will begin our exploratory analysis with the following research question in mind:

"How do individual components and ratios of components in a concrete mixture positively or negatively impact the mixture's compressive strength? "

Based on our findings we will use an iterative process to determine the appropriate question to ask our data via regression analysis.

3. Dataset

This dataset contains 1030 observations, where each observation contains data relevant to a unique concrete sample. These samples are independent, meaning that one sample does not inform on data regarding another sample. Each sample contains 8 total features, 7 of which are various concrete components measured in kg/m^3 while the final feature is age of the sample measured in days. Each sample also contains a numeric outcome of concrete compressive strength measured in megapascals (MPa).

4. Plan of Action

We will be attempting to explore the explanatory relationship between the different mixture components and the compressive strength of the concrete. The approach will be to evaluate the individual components, the ratios of the components, and the age of the mixture to see if there is a statistically significant relationship between specific factors and strength. The age variable, measured in days, will be translated into a categorical variable to see if the age of the concrete impacts its compressive strength, holding all other variables constant. Additional feature engineering will include creating unique pairwise ratios among the concrete components, as well as including the natural logarithm of strictly positive components to account for potential nonlinearity. During the model development phase, we will begin by splitting 30% of the data into an exploratory set and 70% into a test set. On the explanatory dataset, we will conduct forward and backward stepwise approaches to identify a subset of statistically significant features while preserving model parsimony. A backwards approach will start with a full model and sequentially remove insignificant predictors, while a forwards approach will start with an empty model and sequentially add significant predictors. If these two approaches do not converge to the same solution, adjusted R^2 will be used to differentiate between the two resulting models to choose a final model whose structure will serve as the basis for our causal theory. We will apply our resulting linear model towards our test data to see how well it predicts the outcome variable. The research question will be explored via exploratory and regression analysis by evaluating individual coefficients and partial derivatives of the final model to understand how individual ingredients, ratios, and age contribute to the strength of a final concrete solution.