Arshdeep Singh

**1.Introduction**

This paper analyzes the effect of ingredients and time on concrete strength. The ingredients in question will be cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate, and we will be using them on varying days. We want to find out what variables have the most substantial effect on the strength of concrete. The independent variables will be the data, and the ingredients are cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate. The dependent Variable will be the strength of the concrete. We will analyze the data and create a model showing how each Variable affects the concrete strength.

**2. Exploratory Analysis**

First, we must view the data in its raw form. We do this to understand what the data is showing; essentially, we use this stage to find specific issues with the data. This step lets us clean the data; we can see apparent issues, find patterns that can help us understand the data points better, clean the data overall, and prepare it for a complete analysis.

The first thing we will do is look at the normality of the dependent Variable. We can see from the Histogram Shown in Appendix 2.1 that we have some normality, but I would like to create a better version if possible. Using the Log Transformation, we can see the new Histogram shown in Appendix 2.2, which ends up being worse for normality. So we ignore this transformation for now.

The next thing I did was separate the Day variable into four different variables; dividing them into four distinct quarters of the age allows me to have a better understanding of the data as I can comprehend how the various times of the age can affect the strength of the concrete if I had left the data into the 365 count, I would have had data results that could have been misleading.

Next, we must look at the correlation values. This will show us how strongly concrete strength is correlated by the ingredients and Age. Looking at appendix (2.3) we can see the correlation matrix which shows that all the variables have a correlation value of under .5, with the strongest correlation being .49 with Cement compound. We have a lot of negative values, which means that those variables have a negative relationship based on the output. The LOP which is the transformed Variable I created is showing similar results. Low Correlation values show lower linearity. This is also shown by appendix 2.4 which is a scatterplot Matrix, all the plots in the matrix show little to no linearity, the only one that could be seen would be Strength being slightly correlation with the concrete compound. Thus, most of the variables are not that linear, they have weak linearity/correlation with the strongest one being Cement Compound. This means that the variables don’t really have a large effect on the strength on the concrete.

Next looking at descriptives from appendix 2.5, we can use this to determine if we need to do transformations or adjust something in the data. This is where we realize that we need to transform the age variable into 4 different sections, the age variable is a categorical variable. This will allow us to get a better understanding of the data and create better models.

3. Interaction Variables

Using interaction variables, we can find out if certain variables which are related can affect the other Variable, by combining certain variables we can then further our understanding the independent variables effect on the response variable.

The first interaction that I created was the combination of water and Coarse aggregate. This was chosen as the mixture of water to aggregate could help us strengthen the concrete. After running a Model, I realized that this combination was not necessary as this combination had a p Val was insignificant, the data point that was created did not offer any new information that could help us understand the data better. The variables in the model are all ingredients aside from the age of the concrete, the one Variable that we could use to test interaction would be water with other ingredients, but this would lead to a effect on the other ingredients causing a result that we are not looking for, this would create a negative effect on the concrete, the water seems to be balanced already.

4. Transformations of variables

After creating a histogram of the response variable (appendix 2.1) I realized that the data was overall under the bell curve and showing normality, but in hopes to create a better variable that was more Normal, I decided to transform the Variable using the LOG method. The transformation was not the best, it showed wonky results. So, this transformation was not done, and I kept the original Variable of Concrete strength. The other variables don’t need to be transformed as they do not show issues. Another transformation that needed to be done was the age of the concrete, having 365 values means the data needs to be split, so using a dummy variable of Q we can set the value equal to 0 to represent the first quarter of the year and keep going and end up splitting the year into 4 quarters.

5. Collinearity

Collinearity means that two or more independent variables are affecting each other, for example Height and weight are two variables which would probably show collinearity. For this data set, logically there are no variables which would have this issue. And after running a model to observe the VIF (appendix 5.1) we can see that none of the variables have a significant VIF (over 10) which is the indicator of multicollinearity. . There is some Collinearity, but it is under 10, which is our threshold, meaning that the effect isn’t strong enough for us to do something about it.

6. Model selection process

We have a lot of independent variables, and we need to ensure that the model that we produce is the best fit. The best fit model is decided by the models ADJR^2, the highest Value of ADJR^2 shows the best model, this would mean that it explains the variation in the response variable the best. After running a few models, that all the methods result in the same model variables and ADJR^2 values of .53 shown in appendix 6.1. This Values means that the model can explain 53% of the variant in the response variable. The starting ADJR^2 was .24, which meant that before our adjustment, the model could only explain 24% of the variance in the response variable. The backward method removes the fine aggregate compound variable and the Coarse aggregate compound variable, as this method deems them insignificant to the model.

Another thing that I decided to do was to run a selection method on the previously mention transformed response variable which was the LOG (Concrete strength). Although the transformation is not used in the actual modeling, it can be used for comparison to the transformed Variable, this can give us more information to the response variable. After running a backward selection method using the transformed Variable as the response variable, I got a ADJR^2 of .46 shown in appendix 6.2, which means that our model now explain 46% of the variation in the response variable. This is a downgrade from the untransformed Variable, the lower ADJR^2 tells us that this is a worser model than the untransformed variable model in appendix (6.1). This model with the transformed Variable removes the same variables of Coarse aggregate compound and the fine aggregate compound. The resulting ADJR^2 shows that this is a worser model compared to the First model. We will continue to ignore the Transformed variable.

7. Influential and outliers Points

Influential points are points that effect the slope of the model heavily, and outliers are points that are far out from the slope line. I set my parameter for removing points as the point having to be both an influential point and an outlier. If a point is influential but not an outlier, it is needed for the slope as it not being an outlier means a crucial point that effects the slope. A point being only an outlier means that even though the point isn’t close to the data trend, it doesn’t affect the slope in a very meaningful way. Based on this reasoning, I only removed a single point, as only that point meets the criterion for removal. The point that I removed as number 382 shown in appendix 7.1. Removing this point didn’t affect the model in a meaning full way as the total number of OBS are large. However, the point should still be removed.

8. Assumptions

Next, we look to verify that the assumption is met for the regression. These Assumptions are Linearity, Independence, Homoscedasticity, and normality. These assumptions need to be met for the model to be reliable and valid. Starting at the Linearity assumption, based on the scatterplot appendix (2.4) we can see that most of the data is not linear, the only one that seems to show linearity would be cement compound. So I would say that Linearity is violated here. Next looking at Homoscedasticity, or constant variance, we will look at appendix (8.1), the plot shows a certain pattern, which means that there is a pattern in the difference between predicted and actual values, this means that the constant variance clause is violated. This also means that independence is violated, meaning as there is a pattern being shown in appendix 8.1. lastly looking at normality we look at the CDF of studentized residuals plots (appendix 8.2) this show a kink line that means it also violates the model assumption. Our model actually violates all of our assumptions, this is a issue as we have already cleaned the data and have set up the correct model.

9. Training models

Another thing that can be done is Training and testing the data. We can first train the data based on a split of data we decide, I chose a 70/30, this would mean 70 Percent of the data is being used to train the model and 30 is being used to test it. By training a model, we can access the performance of said model to gauge the accuracy of the model. The total amount of observations in the data was 1030, after the split, we had 308 in one data set which was the testing, and 722 in the training set. After running the test on the trained model, we can see a correlation value of .71 shown in appendix 9.1, this refers to the correlation between the actual value of concrete strength and the trained model value. This means that there is moderate correlation between the trained model and the actual values, this model is not the best, but it still is catching a lot of the predictor’s effects., overall, the model is usable even if its correlation isn’t the strongest, it is still showing a strong mediocre correlation.

10.Final model

The final model after analysis Multicollinearity, addressing outliers and influential points, using a model selection method, and training and testing, was shown to have the variables Cement component, Blast Furnace Slag component, Fly Ash component, Water component, Superplasticizer, Q (How old the concrete was incrementing in 3 months). The final model equation was Concrete strength = 22.98+.11(Concrete Compound KG) +.083(Blast furnace slag) +.064(Fly Ash) =.165(Water) +.265(Superplasticizer) +9.39(Q) shown in appendix 10.1. The ADJR^2 is at .54, and the RMSE is at 11.37 with a F value of 197 at a significant P Value, see appendix 10.1. With the ADJR^2 we can say that our model can explain 54% of variation in concrete strength. We can use this model to create prediction intervals, and from the table shown in appendix 10.2, we can see that our prediction intervals are not the best, as our model Is showing large errors, the lowest predictions Std Error Mean Predict is .58, this is showing large errors in our model, with the largest being 2.1. Our residuals are not consistent and are all over the place, this means that our model is not the best at concisely predicting the values of concrete strength.

Looking at appendix 10.3 and observing Standardized estimates, we can find the strongest predictors for the model. We can see that the order from strongest to weakest is Cement compound, Blast furnace slag, Q(Quarter of year) ,Fly ash, water, and then Superplasticizer.

11. Improvements to model

We can improve the model by having more data to run the training on, this doesn’t mean to train more for the same data i.e. a different split but having more raw data. If we use a higher split for the same data, we can get a better correlation value, but that would mean overfitting the data and that would lead to misleading results. If we have more data, we can catch the predictors effects better. Another way we can improve the data is by creating a better dummy variable for the Age variable, logically speaking age should lead to a negative effect as the concrete gets older and older, not stronger and stronger like in our model. If we are to split the Age variable into 6 months and 1 year, we could see better results.

APPENDIX SECTION   
 **2. Exploratory Analysis**

2.1) this shows the Histogram for Concrete strength

A graph with a blue line

Description automatically generated

2.2) This shows the Histogram for Concrete Strength transformed

A graph with a line going up

Description automatically generated2.3) This shows the Correlation values for the Variables involved in the data set.

A screenshot of a computer

Description automatically generated

2.4) This shows the Scatterplot matrix relating to the Variables in the data set.

A screenshot of a scatter chart

Description automatically generated

2.5) This chart shows the descriptives of the data set

A screenshot of a computer

Description automatically generated

**5. Multicollinearity**

5.1) This shows the VIF chart used for Multicollinearity

A screenshot of a computer

Description automatically generated

6. Model selection process

6.1) This is the chart that shows the ADJR^2 values and removal of variables.

**A screenshot of a computer

Description automatically generated**

6.2) This is the chart that shows the ADJR^2 values and removal of variables of the Transformed variable.

A screenshot of a computer

Description automatically generated

**7. Influential and outliers Points**

7.1) This is a visual of OBS 382 being both a outlier and influential point.



**8.) Assumptions**

8.1 This chart shows the residual plot for the model.

A group of blue dots

Description automatically generated

8.2 This chart shows the CDF of studentized residual against Normal Cumulative distribution.

A graph showing a growth

Description automatically generated with medium confidence

**9. Training and testing models**

9.1) this shows the correlation between Yhat and compressive strength.

**A screenshot of a computer

Description automatically generated**

**10. Final model**

10.1) This shows the final model table after all adjustments

A screenshot of a computer

Description automatically generated

10.2) This shows the table that provides the prediction and confidence intervals for a few OBS

A table with numbers and a few black text

Description automatically generated with medium confidence

10.3) This is a table used from accessing the strength of each predictor.

A screenshot of a computer

Description automatically generated

Future works:

I would in the future, leave the Age variable as is, I misunderstood the Variable and decided to transform, when in reality, if I had left it in its 365 form, I would have a better variable that was more accurate. Another thing that I should have done was to use the selection method after splitting data with the train set; this would have resulted in a more accurate train and model performance.

**­**