Final Project

*Anirudh Sharma and Karm Veer*

Our data deals with a key component of construction, which is Concrete.

Concrete is the most important material in civil engineering, prominently used in all kinds of constructions; On a small scale for foundations in homes, or on a massive scale for building bridges and monuments.

The concrete compressive strength is a highly nonlinear function of age and ingredients. These ingredients include cement, blast furnace slag, fly ash, water, superplasticizer, coarse aggregate, and fine aggregate.

**Source:**

<https://www.kaggle.com/elikplim/concrete-compressive-strength-data-set>

what is your dependent variable? what type of data is it?

Our dependent variable is concrete compressive strength measured in Mega Pascals (MPa) which makes it numerical data.

describe your independent variable(s)

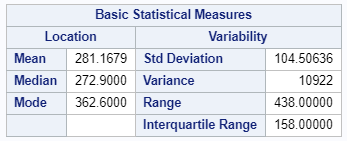
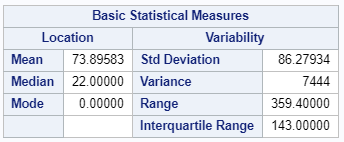
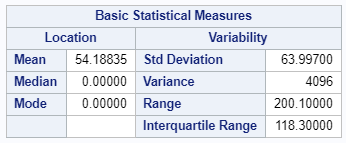
We have the following independent variables, which are all numerical, and measured in Kgs in an M3 mixture (except for age):

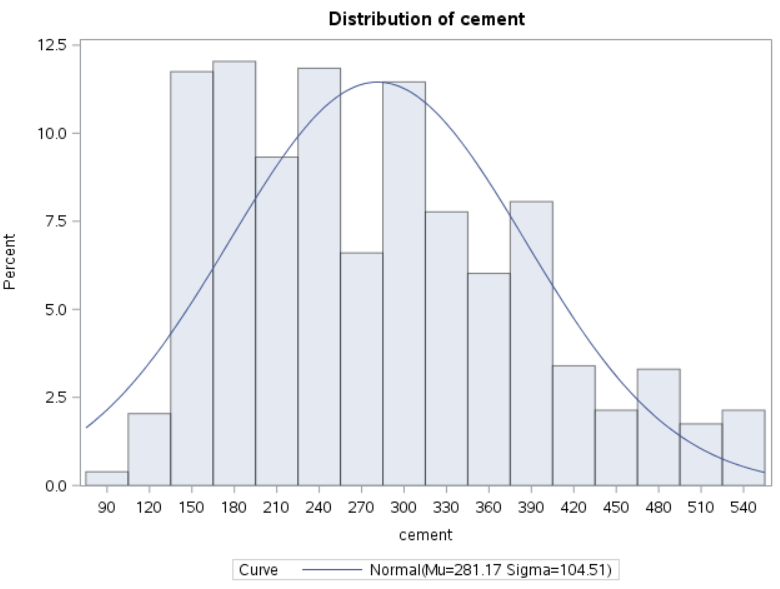
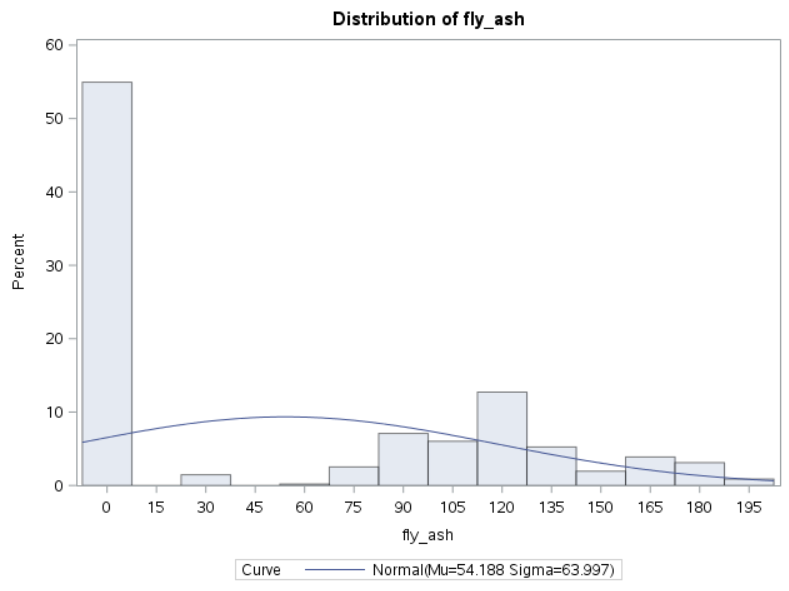
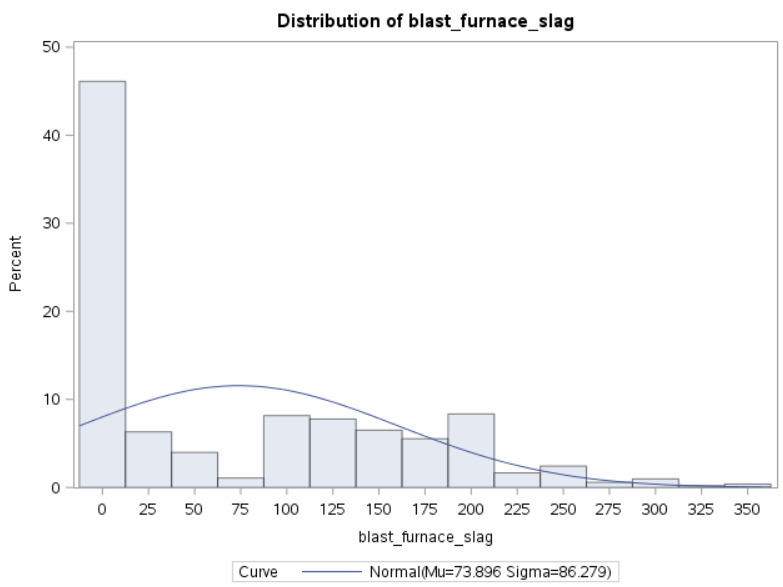
1. Cement
2. Blast Furnace Slag
3. Fly Ash
4. Water
5. Superplasticizer
6. Coarse Aggregate
7. Fine Aggregate
8. Age Measured in Days (1~365)

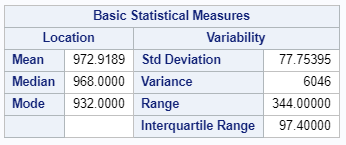
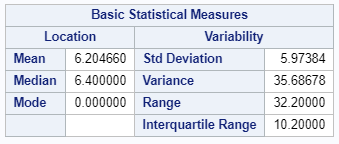
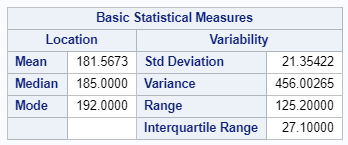
We have 1030 observations.

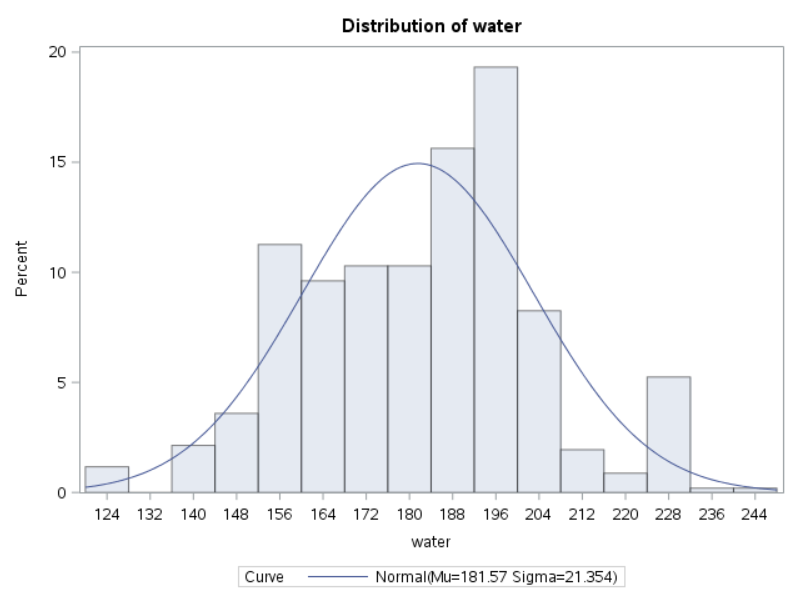
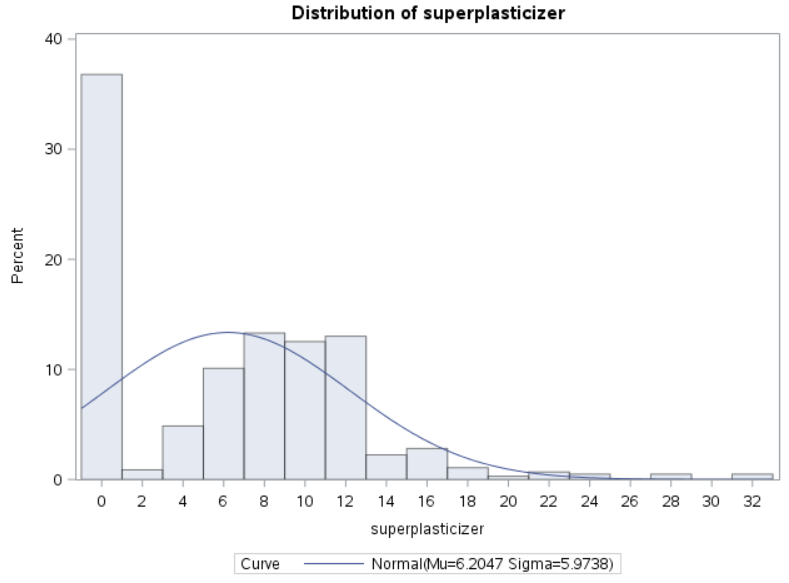
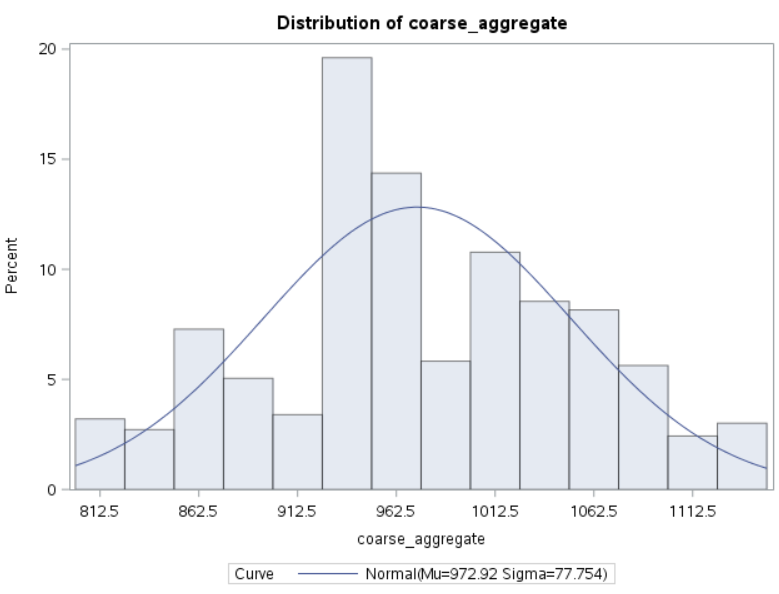
Data Summary

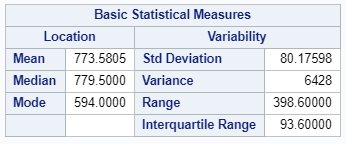
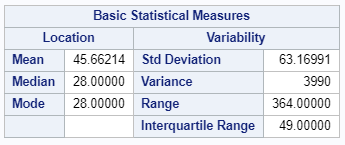
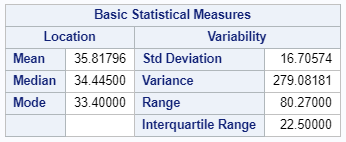
In this section we perform Univariate analysis on all the variables. All our variables are numerical and continuous. To provide a brief overview of all the variables, we have provided tabulation of statistical measures along with corresponding histograms.

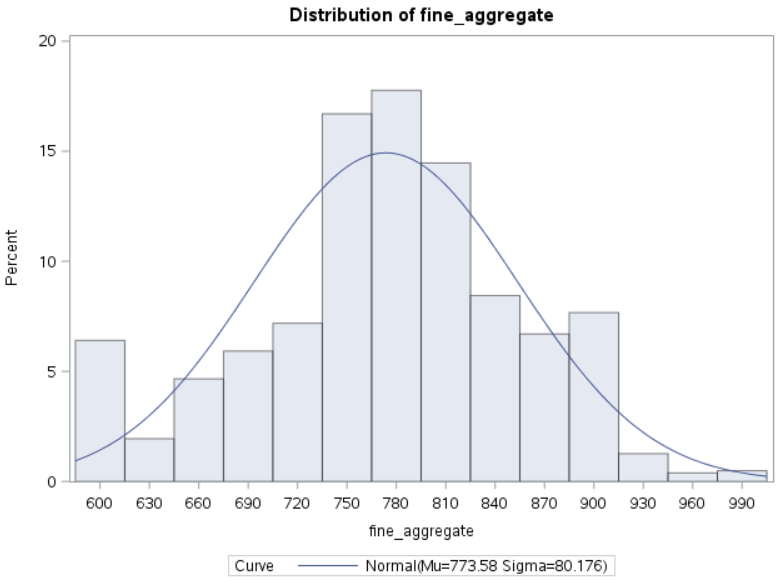
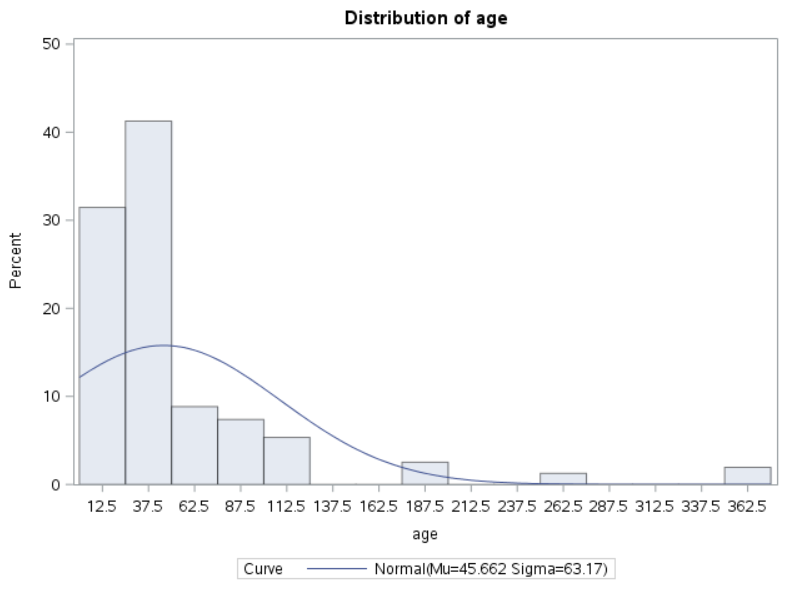
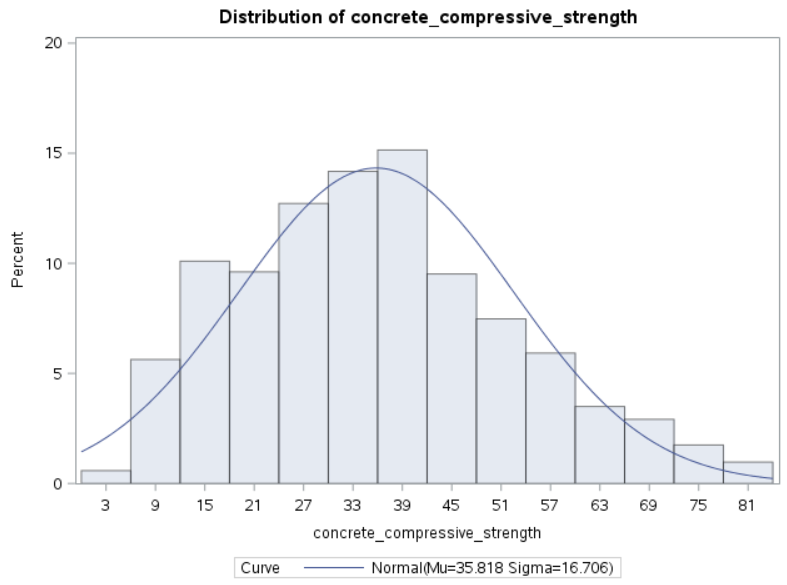
  



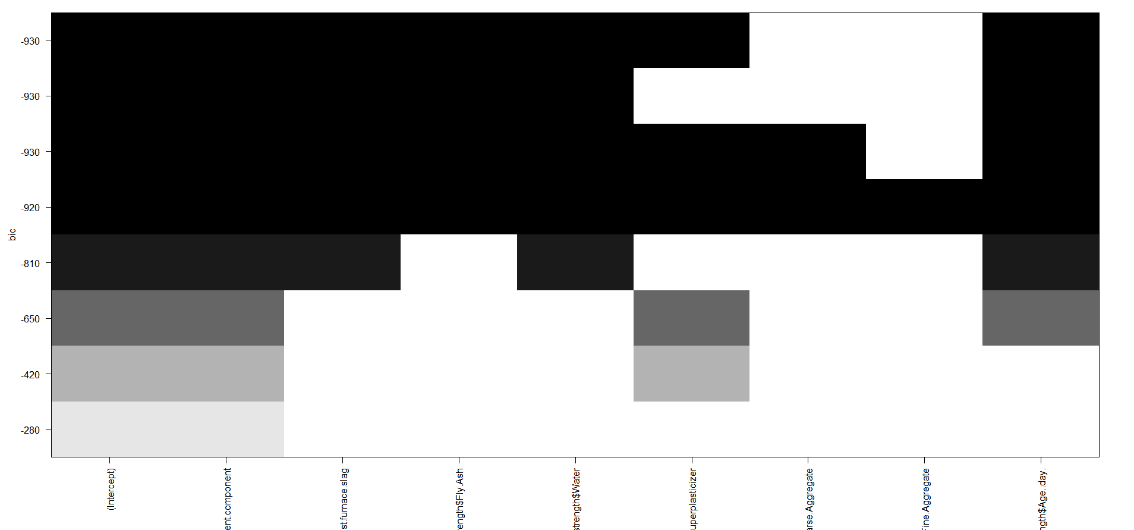
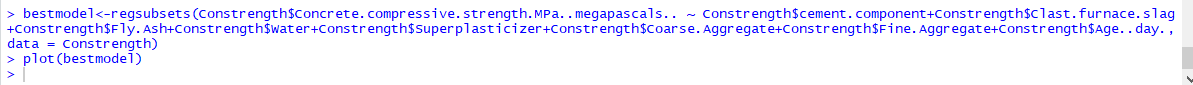
Among all the variables, only Fine aggregate and concrete compressive strength appear to be normally distributed as is evident in the univariate histograms.

Analysis

Since our dataset has a lot of covariates it is important to determine the significant covariates of them all. We did an AIC analysis on our model in R to determine the important covariate in our data set.First we need to load the data set in R and create a model



We have a significant f value but the individual p values for some of the covariate is not significant so we are doing AIC analysis to find out the best model.



Our analysis shows the course grain and fine grain aggregate is not important enough and hence we can remove this form our models.

Here is our new model



Now all the p values are small enough to be considered in our model.

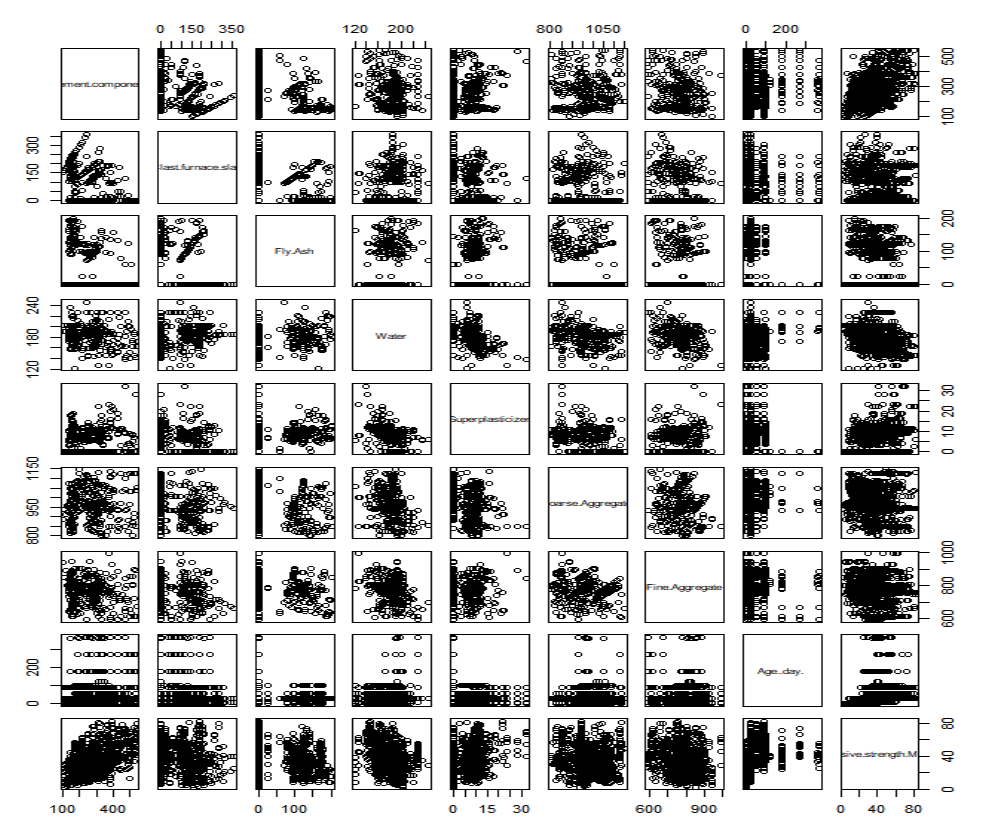
**Hypothesis 1:**

Our Hypothesis is that concrete strength does not depend on any of the covariates hence change in one of the covariates will not result in a significant change in the concrete strength

H0 : B1= B2=B3=B4=B5=0

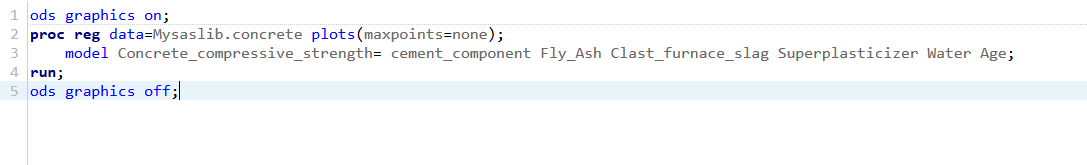
H1: At least one of the slope is non zero.

Lets check the some of the variables against one another to check if there is a relationship between the variables. R has a function pair we will use this function to plot each variable against one another this will make it easy for us to observe any relationship variables have with each other.

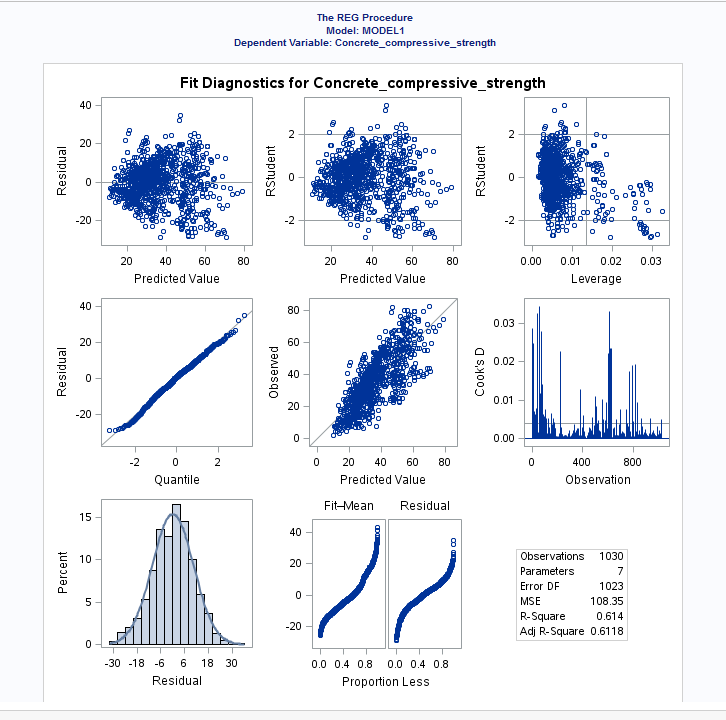
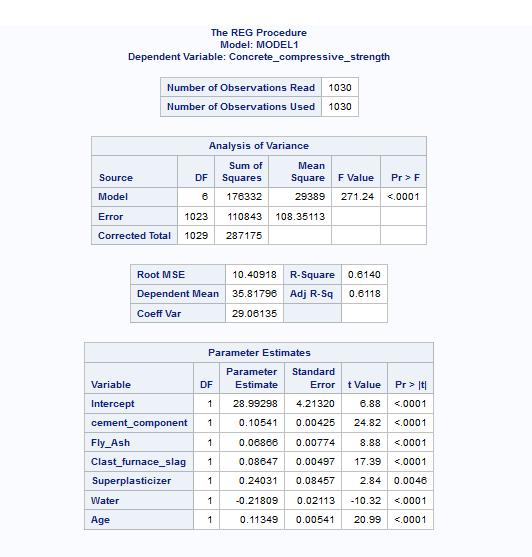


We can observe form the scatter plot that there is a pattern between cement and water a strong relation with concrete strength but where concrete has a positive relation water has a negative relation with the concrete strength suggesting adding more water will decrease the concrete strength.

SAS Code

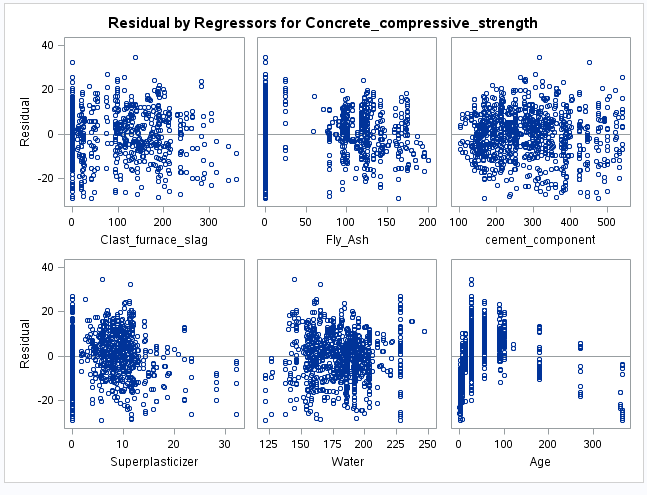


Output:



We observe here that each of the covariates have a slope significantly different from zero and with a p value way less than 0.5 which indicates that all the covariates are significant individually. F value is also small suggenting combination of all the variables are also significant. The resudual vs predicted value doesn’t has any pattern and it is evenly distributed around the zero suggesting the data is normal . We may also check the QQ plot here which follows the line perfectly which is an indication of normality of the data all these graphs suggest out model is normal and with equal variance and it is linear.

Next we will check if the graph of the residual some pattern for individual covariate.



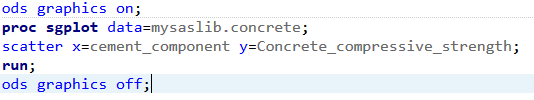
The residual plot looks evenly distributed around zero suggesting there is not a major problem here. At last after observing all the p values ,f values and graphs we may safely reject the null hypothesis.

**Hypothesis 2:**

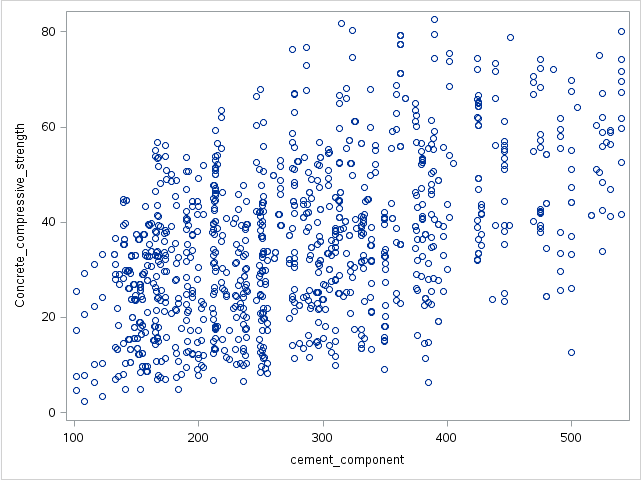
From the previous test we concluded that water and composition of cement are most significant factors in determining the compressive strength of the cement. Let us check if these two covariates alone are enough to determine the compressive strength.

H0: B1=B2=0 (slope for each covariate is zero)

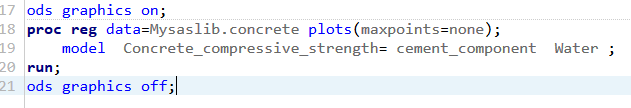
H1: At least one of the slope is non zero



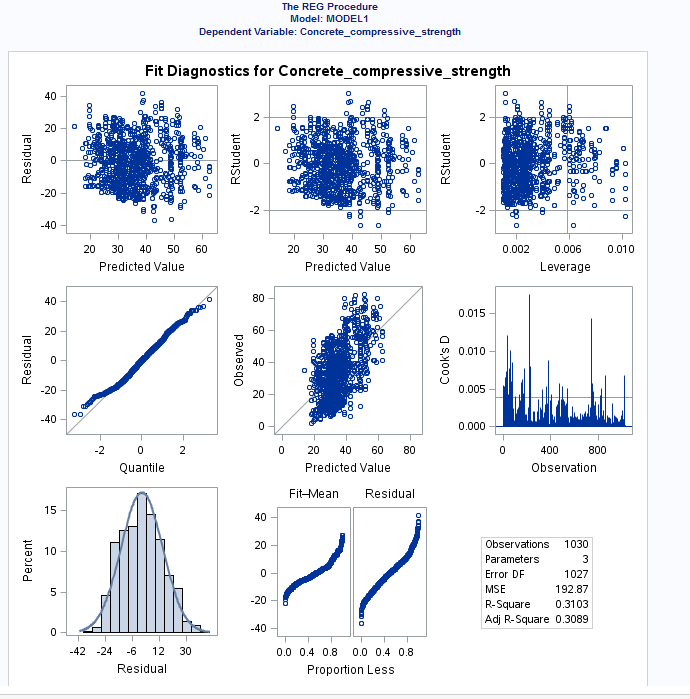
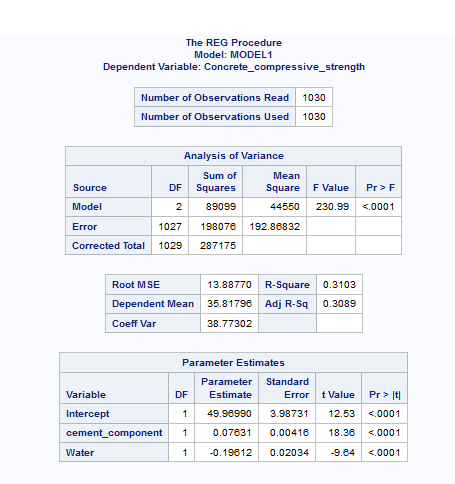
Lets check the scatter plot of the graph cement vs compressive strength.



This shows that the compressive strength increases with the increase of the cement composition in general but the relationship is not linear as there are a lot of cases there higher cement composition has a lower compressive strength and vice versa.



Output:



We have a nonzero slope for both of the variable and the slope for water in negative suggesting the compressive strength is inversely proportional to each other. P values for both water and cement composition is low which means both these variables are significant and the F values is also less than 0.05 meaning the overall model is relevant meaning the combination of both the variables are enough to change the compressive strength. QQ plot also follows the line closely suggesting normality of the model. Equal distribution champions our claim of normality.

With these observation we may reject the null hypothesis

Conclusion

We created a best model using AIC analysis and then two regression to check if out model is good enough. Our first hypothesis was that all the covariates combined are not significant enough that is change in one of the covariate will not result in a significant change in compressive strength .we chose all of the covariates from out AIC test and we got a very low p value for each individual covariate which means all the covariates are significant in the model and an overall low f value of the model suggesting the model itself is significant based on these output we decided we may reject the hypothesis. Our second hypothesis was the water and cement composition alone are not significant enough to affect the compressive strength of the model. Result of our regression shows both these covariates has a p value less than 0.05 individually and the f value is also low. We could see the qq plot and residual plot which championed the fact that data was normal. Based on the following evidences we rejected our null hypothesis.