Linear Regression: It is an approach for predicting a quantitative dependent variable y on the basis of many independent predictors(x1,x2,x3…xn).It assumes there is a linear relationship between x and y.

In our project we applied linear regression two times, one at base model and one after backward elimnination.We found that the R2 score was very low on base model as well as after applying backward elimination.

R2\_score on base model: 8.0%

R2\_score after backward elimination: -0.0033%

This is because our data doesn’t follow assumptions of Linear Regression as shown below:

Summary of Assumptions:

1. For the residual plot there is a clear pattern between residuals and target column (shares).Therefore this assumption is not fulfilled.
2. For checking the normality of the data we applied Shapiro’s test and found out that our p value was 0.0.Thus we reject our null hypothesis and accept the alternate hypothesis which means our data is not normal. This assumption was also not fulfilled.
3. For checking the hetroscedasticity we applied the goldfeld quandt test.Here we got our p value as 9.027698342437254e-13,which means our data is hetroscedasticity and shows an increasing pattern
4. For checking the auto correlation in errors, we applied the Durbin-Watson test and it returned the value of 2.007.Therefore there is almost no correlation between in errors. This assumption thus was fulfilled.
5. Data appears to be non-linear in nature as the rainbow test p value is 0.0026.

Thus we see that our model doesn’t meet many of the assumptions of the linear Regression. Therefore it is not preferable to apply it on this dataset.