STATISTICS FOR DATA SCIENCE

K SRI HARSHITHA

S20160020133

PROBLEM STATEMENT

There is a growing problem to represent and analyse large experimental datasets in many emerging fields of science. One such field is applied superconductivity. The dataset contains 81 features extracted from 21263 superconductores along with the critical temperature in the 82nd column. Our aim is to analyze the data and predict the critical temperature.

ABSTRACT

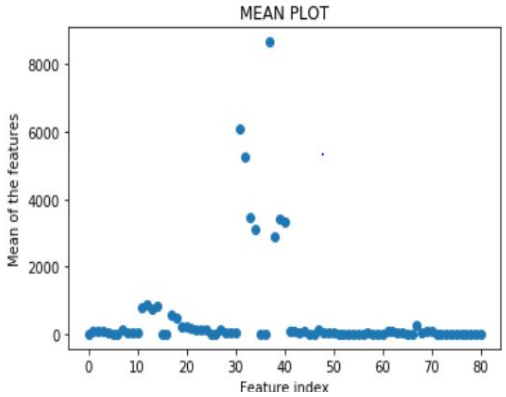
Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical (dummy coded as appropriate).There are 3 major uses for multiple linear regression analysis. First, it might be used to identify the strength of the effect that the independent variables have on a dependent variable. Second, it can be used to forecast effects or impacts of changes. Third, multiple linear regression analysis predicts trends and future values. The regression model is fit into the SuperCon data set and evaluated based on the adjusted R^2 . To overcome heteroscedasticity, Box-Cox method is used for better results. The correlation of the variables is determined.It is observed that there are many related variables. Hence, Principal Component Analysis and Factor analysis are applied to reduce the dimensions of the data set.

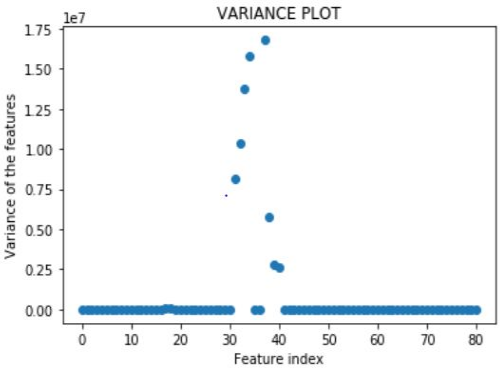
PROCEDURE

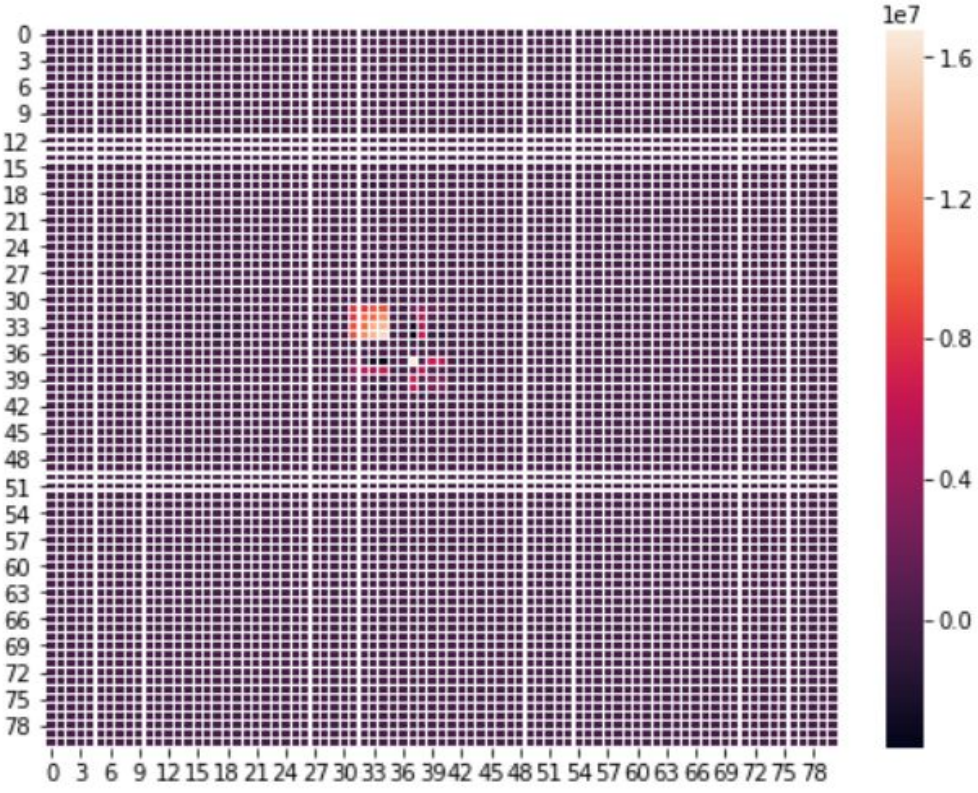
1. Modelling Multi-Linear Regression

Before fitting the regression model,the data is completely analysed. Mean, variance, covariance and correlation of the data is calculated and plotted. To

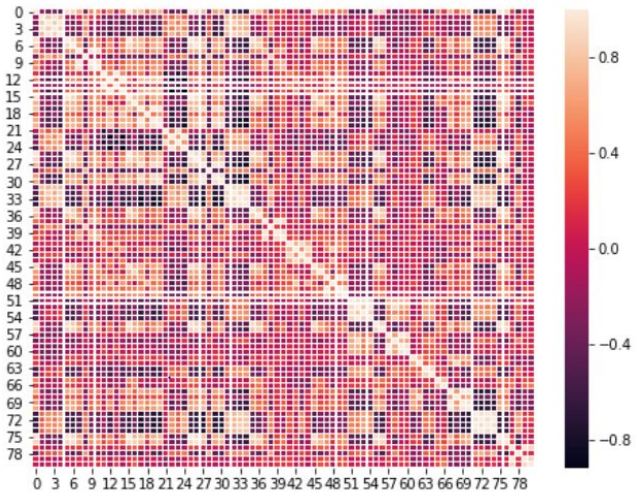
check whether the data follows normal distribution or not, histograms, Q-Q plots are plotted for few random feature vectors. The results showed that data does not follow normal distribution. Later the train set is fit to the regression model and the results are observed.







COVARIANCE



CORRELATION

2. Model Adequacy Check

GOODNESS OF FIT

Linear regression calculates an equation that minimizes the distance between the fitted line and all of the data points. Technically, ordinary least squares (OLS) regression minimizes the sum of the squared residuals. In general, a model fits the data well if the differences between the observed values and the model's predicted values are small and unbiased. R-squared is a statistical measure of how close the data are to the fitted regression line. Generally the value is greater than or equal to 0.9. Goodness of Fit got OLS model is 0.868.

R-SQUARED	0.869
Adj. R-SQUARED	0.868
F-STATISTIC	1558.
Prob (F-STATISTIC)	0.00
Df MODEL	81

TEST OF INDIVIDUAL PARAMETERS

Adding a significant variable to a regression model makes the model more effective,

while adding an unimportant variable may make the model worse. The hypothesis statements to test the significance of a particular regression coefficient, β_j , are:

$$H0 : \beta = 0$$

$$H1 : \beta != 0$$

According to the p values the the null hypothesis is either accepted or rejected. If the p value is greater than the significant value (0.05) then the featured is removed. On doing this the number of significant variable shave reduced to 70. On fitting to OSL method the R^2 remained unchanged i.e., 0.868.

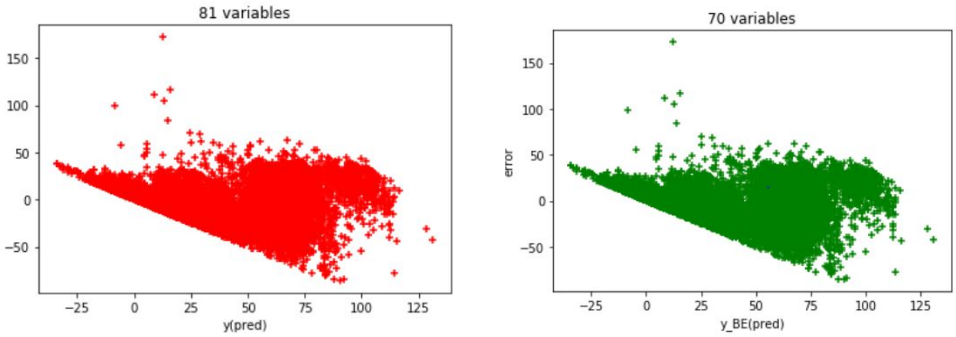
R-SQUARED	0.869
Adj. R-SQUARED	0.868
F-STATISTIC	1802.
Prob (F-STATISTIC)	0.00
Df MODEL	70

TEST OF ASSUMPTIONS

Test of assumptions has been done for both the models,i.e., before and after removing unimportant feature vectors. The errors or the residuals are calculated by taking the difference between predicted values and ground truth values.

1. **HOMOSCEDASTICITY**

   Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables. Heteroscedasticity is present when the size of the error term differs across values of an independent variable. In both the cases the graph looked like funnel shape concluding that they are heteroscedastic.

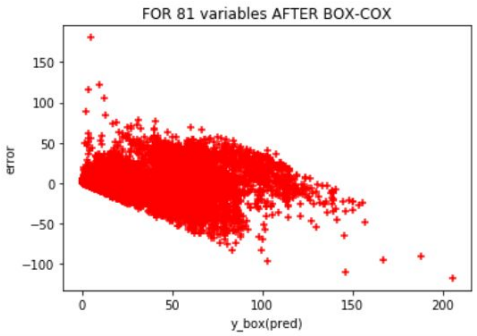


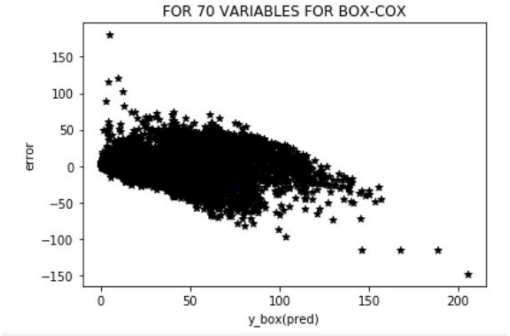
To overcome heteroscedasticity we use BOX-COX method/ the best lambda value is 0.242333. After the transformation, the regression model is fit and goodness of fit is measured.But remember that during prediction, we will have to bring back predicted Y values to original space. Adjusted R squared = 0.934.

R-SQUARED	0.935
Adj. R-SQUARED	0.935
F-STATISTIC	3386.
Prob (F-STATISTIC)	0.00
Df MODEL	81

Now the graph is plotted for the errors and the predicted y values of the new regression model.This time the graph showed homoscedasticity.

R-SQUARED	0.934
Adj. R-SQUARED	0.934
F-STATISTIC	3852.
Prob (F-STATISTIC)	0.00
Df MODEL	70



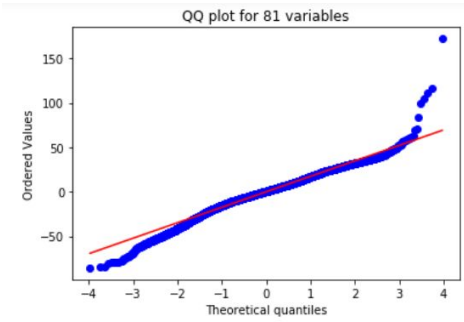


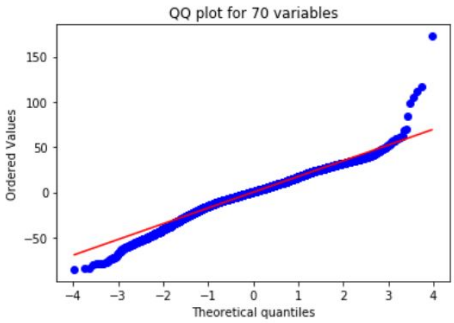
2. **RESIDUAL ARE UNCORRELATED**

   We test this assumption using DARWIN WATSON test.

3. **ASSUMPTION OF NORMAL DISTRIBUTION**

   We can test this assumptions using Q-Q plot.





3. Model Diagnosis

Cook's distance is calculated to detect the influential points. If the cook's distance is D > 1, then the point is considered to be influential. But for our data set there are no such influential points

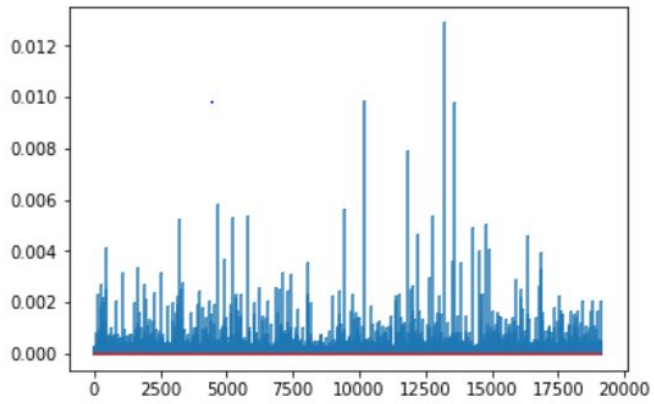
4. PCA AND MODEL ADEQUACY CHECK

Bartlett's Sphericity Test:

$$H0 : \rho = I$$

$$H1 : \rho \; != I$$

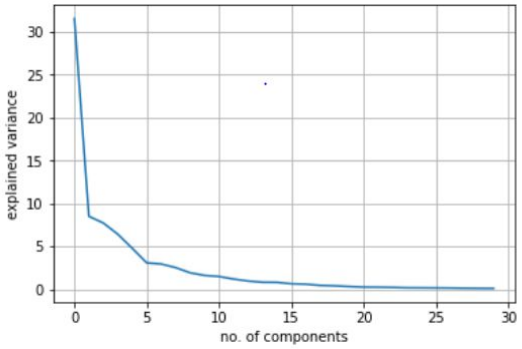
The null hypothesis is that the population correlation matrix is identity matrix or

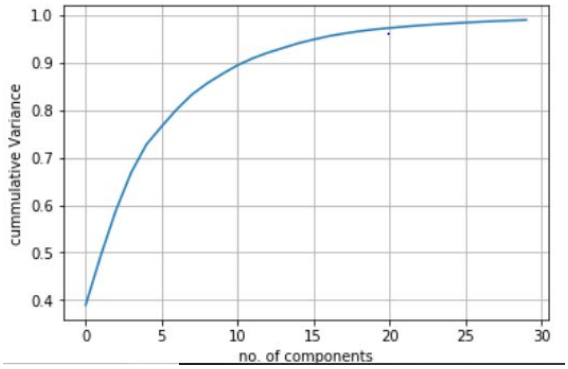


that the covariance matrix is diagonal one. But when the covariance was calculated, it is seen that it is not an identity matrix. Hence we reject the null hypothesis. Hence PCA is implemented. PCA model is fit and **explained variance** and **cumulative variance** graphs are plotted. From variance explained graph, we can observe that the out of 81 components 18 components can be considered since there exists too less variance from 18-81.

In cumulative variance explained the number of components are selected in a such way that max variance is explained by least possible number of components. Here also we observe that components are considered.

Average root method : the mean of the eigenvalues is calculated and those PC's are selected whose eigenvalues are greater than the mean eigenvalue. For our data set we get 19 such PC's

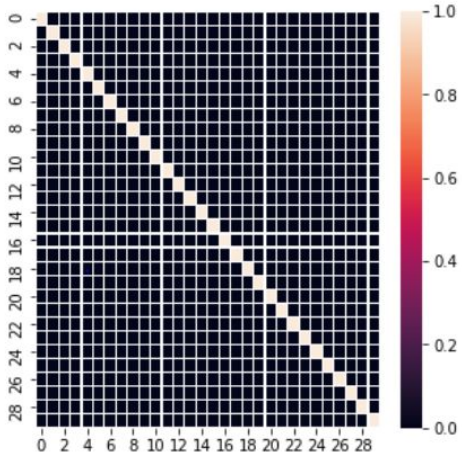




Goodness Of Fit :

The goodness of fit of the model is calculated by varying the number of components

(18 and 19). But in both the cases the goodness of fit is bad.



R-SQUARED	0.689
Adj. R-SQUARED	0.688
F-STATISTIC	2114.
Prob (F-STATISTIC)	0.00
Df MODEL	20

CONCLUSION

Since the dataset is heteroscedastic, firstly BOX-COX transformation is applied. After the transformation the model is fitted to the data and the goodness of fit is measured. We get a value of 0.933 adjusted R^2 value. From covariance matrix we also observed that there are many correlated components. Hence we applied PCA to reduce the dimension. Model adequacy test are done to determine the PC's to be considered. The goodness of fit is measured, the accuracy was bad.