STATISTICS FOR DATA SCIENCE

Analysis of Multi-Linear Regression

NAME :- B S N V CHAITANYA

ROLL NO :- S20160020115

Problem Statement

Superconductivity data dataset is given and we are supposed to analyze the data and predict the critical temperature.

Abstract

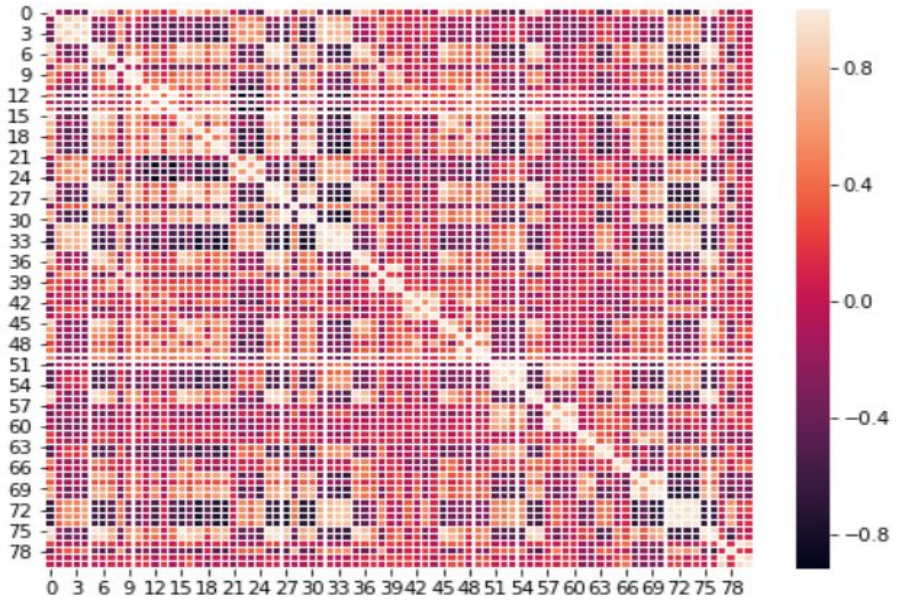
Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. Every value of the independent variable *x* is associated with a value of the dependent variable *y*. A multiple linear regression analysis is carried out to predict the values of a dependent variable,Y, given a set of p explanatory variables (x1,x2,....,xp). In this case study, dataset with 81 features and 21263 samples of superconductors is analyzed.This is done by fitting a regression model, measuring the goodness of fit, testing the assumptions and finally doing model adequacy checking. By analyzing the correlation matrix and by Bartlett's Sphericity test, Principle Component Analysis(PCA) is applied.

Multiple Linear Regression

Modelling

- ➔ Initially dataset is checked if it contains any missing values. Thankfully there are no missing values.
- ➔ Variance and Correlation coefficients of the features are plotted to get the idea of the raw data i.e., to know if there are any correlations or not. Clearly we can see from the heatmap of correlation coefficients that the variables that are supposed to be independent are not actually independent.

➔ Later we try to reduce these correlations using Principal Component Analysis.



➔ Normal distribution plots(Q-Q Plots) are plotted of different features to check if the data is coming from normal distribution. But it is observed that the data is not coming from normal distribution.

➔ The data is split into testing and training data with ratio of 0.2% as the test data; the training data is fit into a regression model and summary is taken from OLS(Ordinary Least Squares) model.

Model Adequacy Checking

Goodness Of Fit

- The goodness of fit of a statistical model describes how well it fits a set of observations. R-squared and Adjusted R-squared are a goodness-of-fit measure for linear regression models. This statistic indicates the percentage of the variance in the dependent variable that the independent variables explain collectively.
- Adjusted R-squared is used over R-squared to check the goodness of fit as R-squared has some problems.
- Every time you add a predictor to a model, the R-squared increases, even if due to chance alone. It never decreases. If a model has too many predictors and higher

order polynomials, it begins to model the random noise in the data. This condition is known as overfitting the model and it produces high R-squared values even when the model is not actually making good predictions.

- For the above model, it is observed from the OLS summary that the

R - Squared	0.869
Adjusted R -squared	0.868

Test of Individual Parameters

➔ After applying the regression model, we get the estimates of βj's that tell us the significance of each feature.

➔ To check this significance, we test each parameter using the hypothesis testing with null Hypothesis : $\beta j = 0$; alternative hypothesis : $\beta j \neq 0$; with level of significance, $\alpha = 0.05$.

➔ If the P-value > α, we fail to reject null hypothesis. After testing the individual parameters ,the features are reduced to 70 features from 81 features.

➔ The features removed are :

- wtd_mean_Density
- Wtd_std_fie
- wtd_range_Density
- wtd_std_ThermalConductivity
- Gmean_atomic_radius
- wtd_entropy_ThermalConductivity
- Wtd_entropy_atomic_mass
- Wtd_range_atomic_mass
- Wtd_std_atomic_mass
- wtd_range_Valence
- entropy_ElectronAffinity

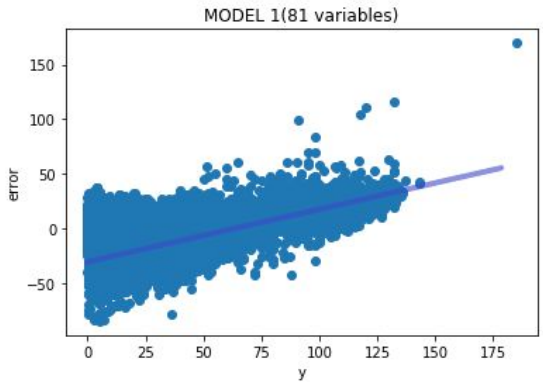
➔ After this, again OLS method is fit and the Adjusted R Squared remained the same 0.868.

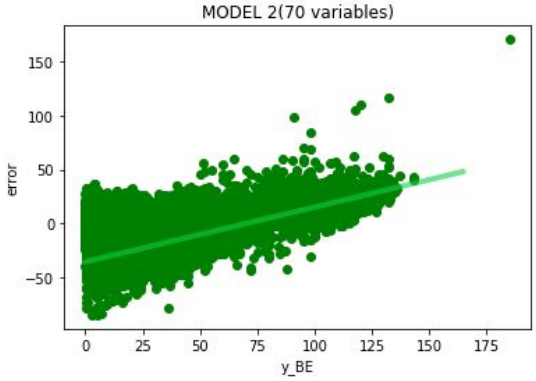
R - Squared	0.869
Adjusted R -squared	0.868

Test of Assumptions

1 . **Test of Linearity**

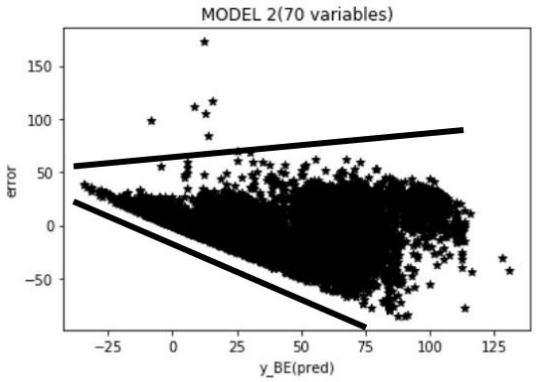
- ➔ The most important assumption that we take while fitting a regression model is that the residuals are linear. We can test this by plotting the residuals versus actual prediction plot.
- ➔ From the plot it is clear that both the models that uses 81 variables and 70 variables are linear.

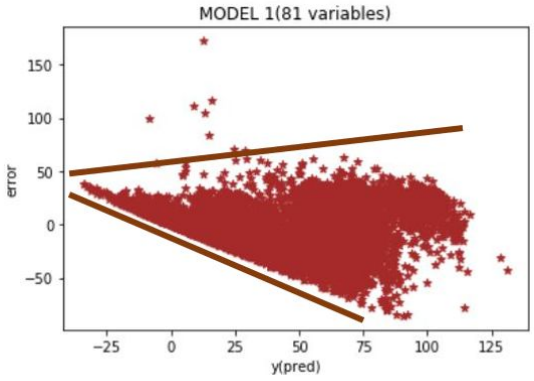




2 . **Assumption of Homoscedasticity**

- ➔ This test is done with both the models, one with 81 variables(initial dataset) and another with 70 variables after parameter checking. To check the homoscedasticity a graph is plotted with the residuals and the predicted values of training Y. The graphs has funnel shape which concludes the heteroscedasticity.





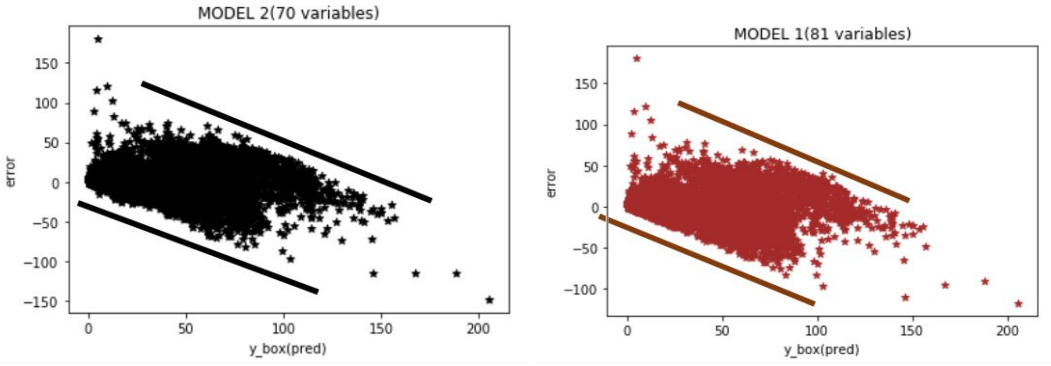
- To overcome this problem, we have to transform the Y to another domain. Commonly used method is BOX-COX method where the parameter for the transformation , λ is selected in such a way to reduce the Squared error. The λ obtained is 0.242333, that is used to transform Y.
- After the transformation the model is fit to regression model and Adjusted R squared is checked.But we should transform the predicted Y values to original domain.
- The results from the OLS summary of model 1(81 variables) are

R - Squared	0.934
Adjusted R -squared	0.933

- The results from the OLS summary of model 2(70 variables) are

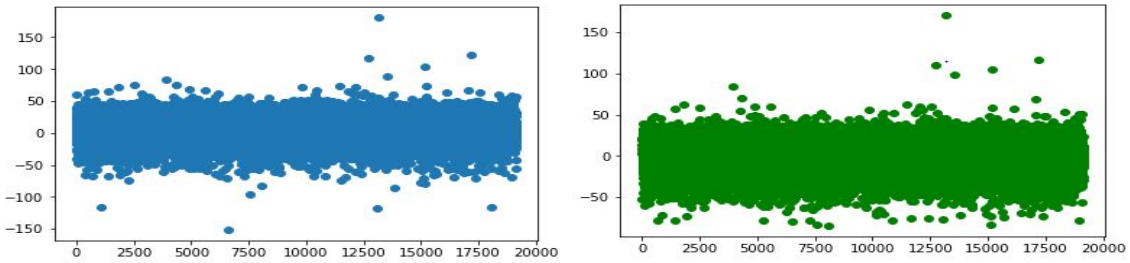
R - Squared	0.933
Adjusted R -squared	0.932

- The scatter plot now is in between two parallel lines which concludes that the residuals are homoscedastic.



3 . **Residuals are uncorrelated**

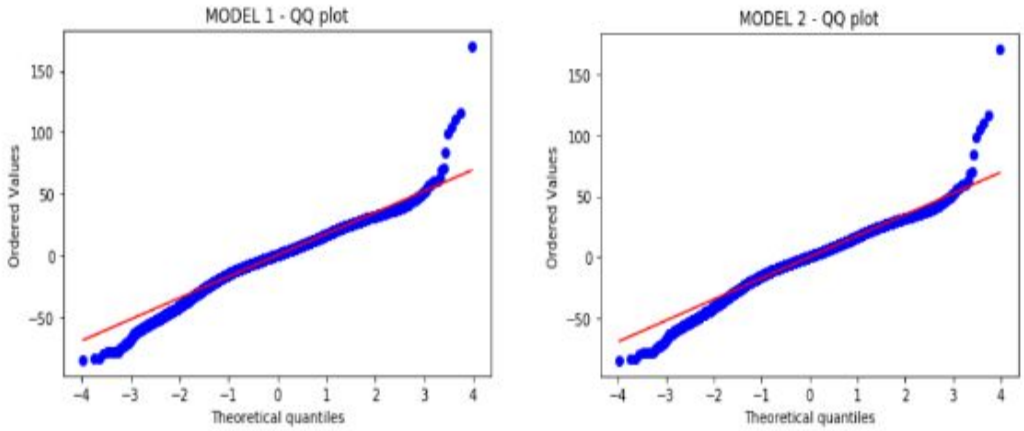
- ➔ This assumption is first checked by plotting the error for each observation to try to interpret about the correlation of the residuals. The left was obtained from the model with 81 variables and the right was the model with 70 variables.



- ➔ But the errors are not following any particular pattern. So, we won't be able to say anything about correlations between the residuals just by looking at the plot.
- ➔ So we use Darwin - Watson test where the statistic is DW = 2(1-r) .
- ➔ If DW = 2, there is no correlation.
- ➔ If Dw > 2, residuals have negative correlation
- ➔ If DW < 2 , residuals have positive correlation
- ➔ In this case study, the DW value for first and second model is 1.982 and 1.979 respectively which are nearly equal to 2, that implies that residuals are uncorrelated.

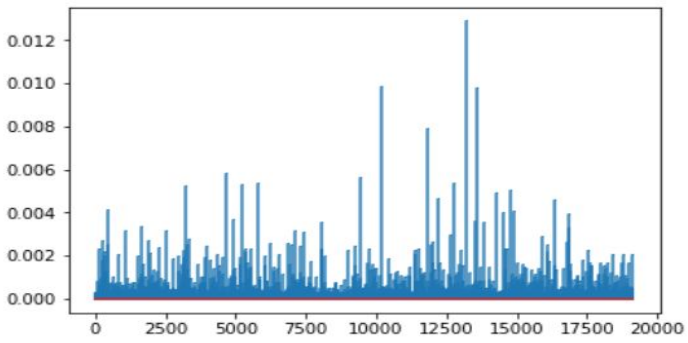
4 . **Residuals are normally distributed**

- ➔ Normality test is done using Q-Q plot.In statistics, a Q–Q plot is a probability plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- ➔ The residuals are plotted against these quantiles. The graph is linear without large deviation, thus can be concluded that the residuals are normally distributed.



Model Diagnostics

- ➔ In model diagnosis, one of the key function is to detect influential points.
- ➔ This is done using Cook's distance criteria.Generally, if Cook's distance is greater than 1, it is said to be influential point.
- ➔ In our case study,there is no such point whose cook's distance is greater than 1.Thus no influential point.



Principal Component Analysis(PCA)

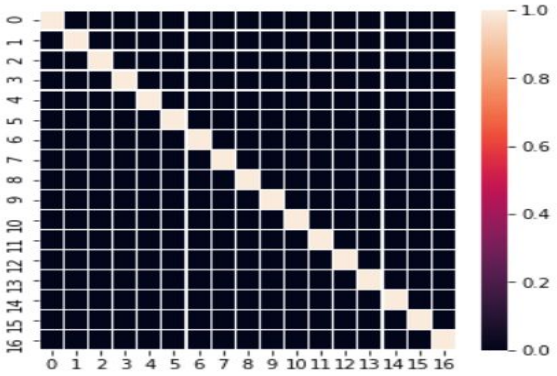
Principal Component Analysis (PCA) is a dimension reduction technique. We obtain a set of Principal Components which summarize, as well as possible, the information available in the data. The factors are linear combinations of the original variables. The approach can handle only quantitative variables.

Implementation of PCA

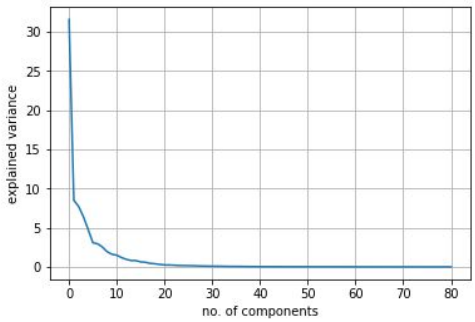
- ➔ Bartlett's Sphericity Test: Bartlett's test compares the observed correlation matrix to the identity matrix. In other words, it checks if there is a certain redundancy between the variables that we can summarize with a few number of factors.
- ➔ This is done prior to using PCA to confirm if it is of any worth to do PCA.
- ➔ This test checks if the observed correlation matrix diverges significantly from the identity matrix.
- ➔ But we have already seen that correlation matrix is not a identity matrix and we also reject the null hypothesis according to Bartlett's test with a level of significance of 0.05 and we can perform PCA efficiently.

Implementation

- ➔ We fit the PCA model to the dataset after standardizing it and graphs of explained variance and cumulative variance are plotted. According to the graphs, we select the number of components that are enough to explain most of the variance.
- ➔ Correlation matrix is plotted to verify if the there are still correlated principal components and we can also notice that the extracted PC's are uncorrelated.



- ➔ Then the explained variance and cumulative variance graphs are plotted.
- ➔ The Graph of explained variance: The number of PC to be used are selected in such a way that the components that explain too less variance are ignored. In this case the number of components derived are 16.
- ➔ The explained variance versus no.of components plot is as follows



➔ Average root method : In this method, the number of components are selected in a way where the eigenvalues are greater than the average of all the eigenvalues.In this case, we got 12 eigenvalues that are greater than average eigenvalue.

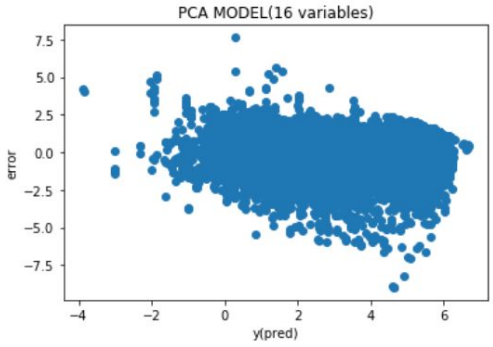
Goodness of Fit

➔ The goodness of fit is calculated with 16 Principal Components.

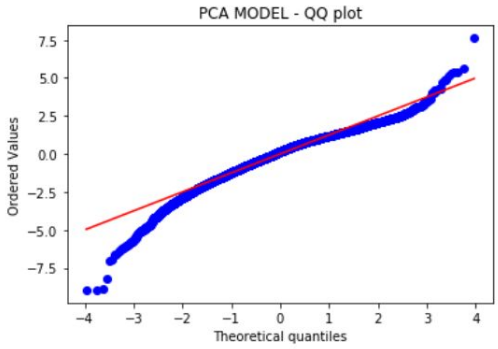
➔ The results from the OLS summary of model

R - Squared	0.671
Adjusted R -squared	0.671

➔ Graph is plotted between the residuals and predicted values of Y, which resulted in homoscedasticity.



➔ Q-Q plot is also plotted and residuals followed normal distribution.



➔ The DW value for this model is 1.991 which is also very much close to 2. So, we can say that the residuals of the model are uncorrelated.

CONCLUSION

Finally, after the analysis it resulted that the regression model with 81 variables,after applying Box-Cox transformation has the best Adjusted R-Square of 0.933 when compared to all other models implemented here. This model is also satisfying all the assumptions taken for fitting a regression model.