STATISTICS FOR DATA SCIENCE

Study on Multiple Linear Regression analysis

K Anjali Poornima

S20160020132

PROBLEM STATEMENT

Given a dataset of Superconductivity data containing 81 features extracted from 21263 superconductors along with the critical temperature in the 82nd column.The goal here is to analyze the data and predict the critical temperature based on the features extracted.

ABSTRACT

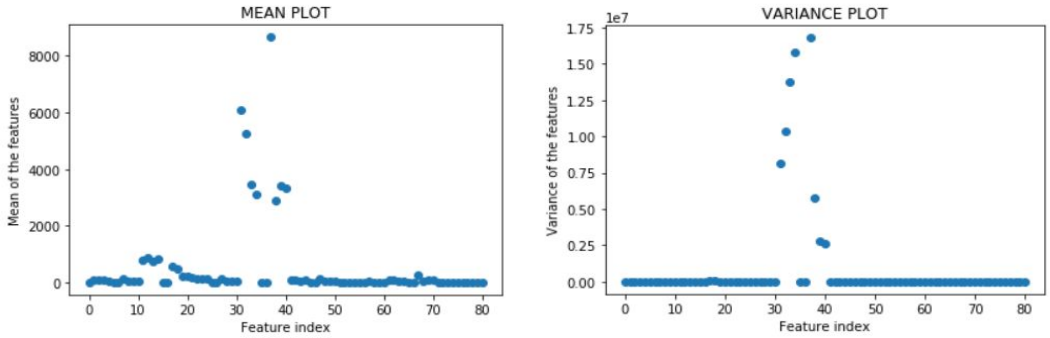
Regression analysis is a statistical technique for estimating the relationship among variables which have reason and result relation.Regression models with one dependent variable and more than one independent variables come under the category of multilinear regression.In this study, data for multilinear regression analysis is taken from SuperCon database. Assumptions of multilinear regression analysis - normality,linearity,no extreme values and missing values were examined.Then a regression model is fit into the data and is evaluated based on adjusted R square(which is less than 0.9). The correlation matrix is plotted and observed that there are many related features.Thus Principal Component Analysis(PCA) and Factor analysis are fit to reduce the dimension but the accuracy proved to be not so good(0.671). To overcome heteroscedasticity(different variability), Box-Cox method is used and this resulted in a good adjusted R square(appx. 0.93) compared to initial data(0.869).

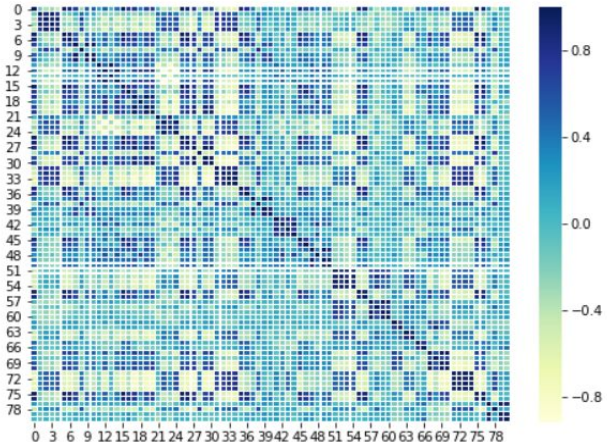
Methodology

1. Modelling multiple linear regression problem
2. Model Adequacy testing
3. Model Diagnostics
4. Implementation of PCA and Model Adequacy testing

Modelling multi-linear regression problem

- Initially the dataset is studied; mean, variance and correlation coefficients are plotted.
- From the correlation coefficient matrix we can see that the variables are not independent. Later we will use PCA and try to reduce these correlations and make them independent.





- It is verified if the data is from normal distribution. This is done manually by plotting few histograms and QQ-plots for each feature vector . But both of them concluded that the data has **not** been derived from normal distribution.
- Then, the data is split into train and test sets and the training data is fit into a linear regression model (sklearn.linear_model.LinearRegression()) and OLS(Ordinary Least Squares) model (statsmodels.regression.linear_model.OLS()) and summary is taken.

Model Adequacy Checking

1. Goodness of fit
2. Test of individual Parameters
3. Test of assumptions

Goodness of Fit:

- To evaluate the overall fit of a linear model, we use the **R-squared** value. R-squared is the **proportion of variance explained.** R-squared is between 0 and 1. But R-squared value is biased due to fact that it increases whenever we add a new pedictor to the model.
- Also If a model has too many predictors and higher order polynomials, it begins to model the random noise in the data. Due to this we will see a misleadingly high value even when the model is not a good fit.
- So, we use adjusted R-squared to over these problems.
- The **Goodness of Fit** of the about OLS model is measured as **0.868 (Adjusted R square)**

OLS Regression Results

Dep. Variable:	y	**R-squared:**	0.869
Model:	OLS	**Adj. R-squared:**	0.868
Method:	Least Squares	**F-statistic:**	1558.
Date:	Fri, 23 Nov 2018	**Prob (F-statistic):**	0.00
Time:	21:39:21	**Log-Likelihood:**	-82005.
No. Observations:	19136	**AIC:**	1.642e+05
Df Residuals:	19055	**BIC:**	1.648e+05
Df Model:	81		
Covariance Type:	nonrobust		

Test of individual Parameters:

- By applying the regression model we estimate β j's(coefficients of Xj), which estimates the significance of Xj.
- Hypothesis Testing is done for every β and according to p-values we either accept or reject the null hypothesis.
- If the p-value is greater than the significant level, those features are removed since we fail to reject null hypothesis(β j = 0). In this study the significant level is set as 0.05. The features are reduced to 70 from 81.
- The features that were rejected are :
    a. wtd_mean_Density
    b. Wtd_std_fie

c. wtd_range_Density
d. wtd_std_ThermalConductivity
e. Gmean_atomic_radius
f. wtd_entropy_ThermalConductivity
g. Wtd_entropy_atomic_mass
h. Wtd_range_atomic_mass
i. Wtd_std_atomic_mass
j. wtd_range_Valence
k. entropy_ElectronAffinity

- This is again fit to OLS method and summary is drawn and **Adjusted R Squared remained unchanged(0.868).**

OLS Regression Results

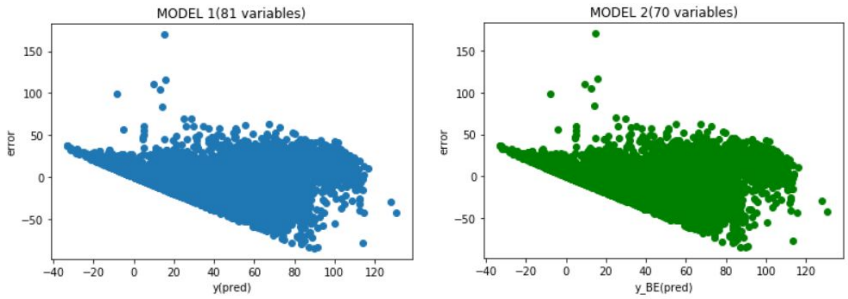
Dep. Variable:	y	**R-squared:**	0.869
Model:	OLS	**Adj. R-squared:**	0.868
Method:	Least Squares	**F-statistic:**	1802.
Date:	Fri, 23 Nov 2018	**Prob (F-statistic):**	0.00
Time:	15:21:09	**Log-Likelihood:**	-82010.
No. Observations:	19136	**AIC:**	1.642e+05
Df Residuals:	19066	**BIC:**	1.647e+05
Df Model:	70		
Covariance Type:	nonrobust		

Test of Assumptions:

- In this analysis, we have tested the assumptions for both the models mentioned above(before elimination of features(with 81 features) and after removing features(with 70 features)).
- The predictions are taken for the training data. The errors or the residuals are calculated by taking the difference between predicted values and ground truth values.

1. Assumption of Homoscedasticity

To test this assumption,we plot the predicted Y values across the errors.If the scatter plot lies between two parallel lines, they are homoscedastic.In both the cases the graph looked like funnel shape concluding that they are heteroscedastic.



REMEDY:

To overcome HETEROSCEDASTICITY, we transform Y to another domain like log Y or sq root(Y). For this we use **Box-Cox method.** We should obtain lambda(parameter for transforming Y) that gives minimum SSE.

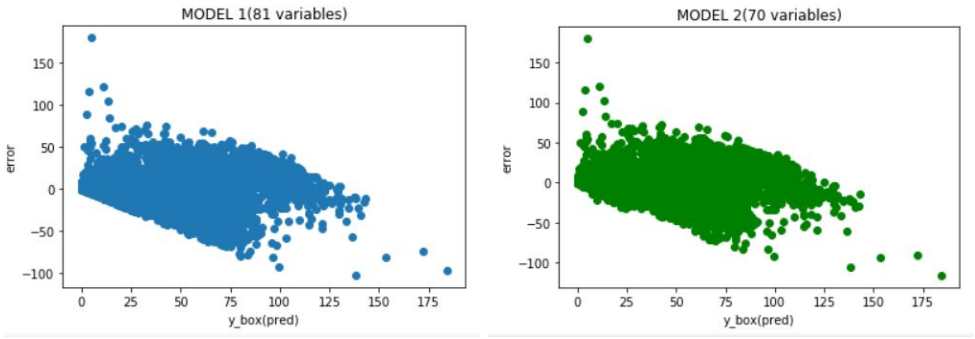
```
To overcome heteroscadacity we used boxcox method to transform y's. And the best lambda is 0.242333
```

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 = 0.933.

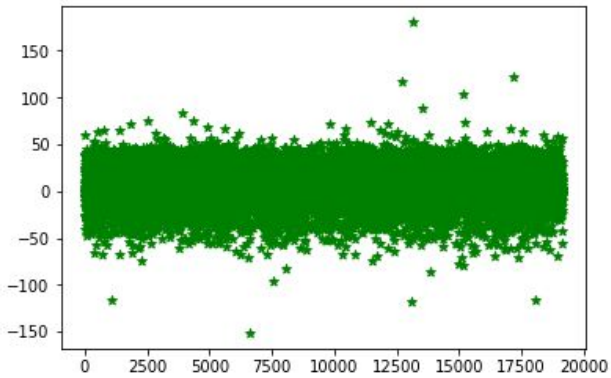
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.934
Model:                            OLS   Adj. R-squared:                  0.933
Method:                 Least Squares   F-statistic:                     3306.
Date:                Fri, 23 Nov 2018   Prob (F-statistic):               0.00
Time:                        15:58:31   Log-Likelihood:                -32885.
No. Observations:               19136   AIC:                         6.593e+04
Df Residuals:                   19055   BIC:                         6.657e+04
Df Model:                          81
Covariance Type:            nonrobust
==============================================================================
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.933
Model:                            OLS   Adj. R-squared:                  0.932
Method:                 Least Squares   F-statistic:                     3771.
Date:                Fri, 23 Nov 2018   Prob (F-statistic):               0.00
Time:                        15:59:54   Log-Likelihood:                -33018.
No. Observations:               19136   AIC:                         6.618e+04
Df Residuals:                   19066   BIC:                         6.673e+04
Df Model:                          70
Covariance Type:            nonrobust
==============================================================================
```

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



2. Residuals are uncorrelated



The residuals for each observation is plotted to interpret the correlation between them.

We test this assumptions using Darwin Watson(DW) test(**import durbin_watson from statsmodels.stats.stattools**).How to check the independence of residuals?

If DW = 2 ---> No Correlation

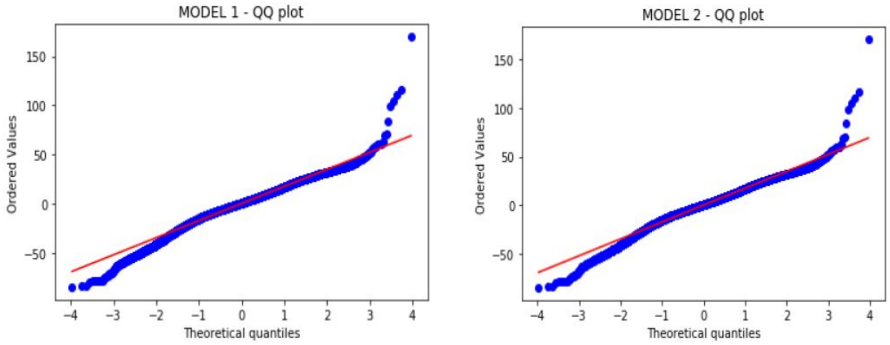
DW > 2 ---> Negative Correlation

DW < 2 ---> Positive Correlation

```
DW(Model 1) = 1.982261 nearly 2 => residuals(errors) are uncorrelated
DW(Model 2) = 1.982182 nearly 2 => residuals(errors) are uncorrelated
```

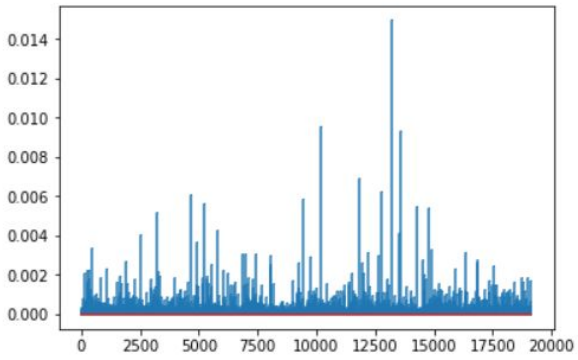
3. Assumption of normal distribution of residuals

We can test this normality using a QQ-plot(import scipy.stats.probplot). The residuals almost are aligned with the quantiles, which conclude that they are normally distributed.



Model Diagnostics

- In this module, we detect the influential points using cook's distance. Generally if cook's distance , Di >1 the point is an influential.
- In code, this is calculated using statsmodels.regression.linear_model.OLSResults.get_influence()
- There is no such point in our dataset as cook's distance satisfied the condition.

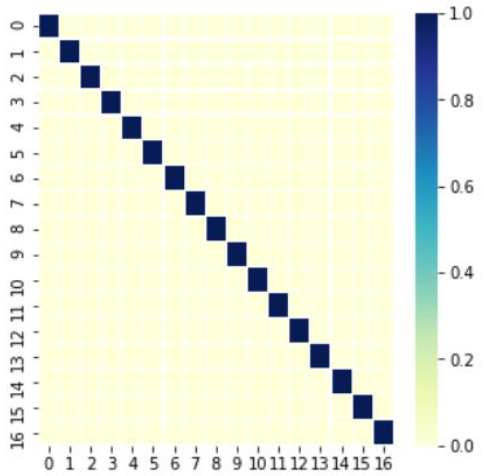


Implementation of PCA and Model Adequacy checking

- BARTLETT'S SPHERICITY TEST : It is often done prior PCA or factor analysis, tests
- whether the data comes from multivariate normal distribution with zero covariances.The null hypothesis is that the population correlation matrix is identity matrix or that the covariance matrix is diagonal one.
- But as seen before, the covariance matrix is not identity matrix, we reject null hypothesis and PCA can be implemented.
- In this analysis, we computed p-value for bartlett test using scipy.stats.bartlett. It is around 0.001 and we reject null hypothesis.

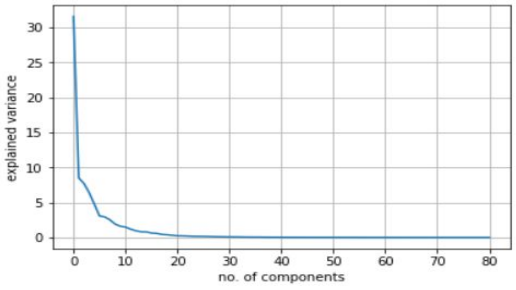
IMPLEMENTATION:

- Before we fit the PCA model, bias term is added.Bias nodes are added to increase the flexibility of the model to fit the data. Specifically, it allows the network to fit the data when all input features are equal to 0, and very likely decreases the bias of the fitted values elsewhere in the data space.
- Then ,we fit the PCA model from sklearn.decomposition.PCA to the dataset after standardizing it.
- Correlation matrix is plotted to verify if the there are still correlated principal components.
- Then the explained variance and cumulative variance graphs are plotted.



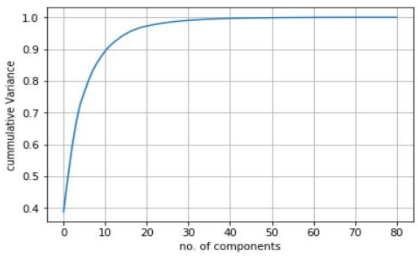
SELECTION OF No. OF COMPONENTS:

- Variance Explained: The number of PC's are selected in such a way that the components that explain too less variance are ignored. Here the no. of components that may be considered are 16.



- Cumulative Variance Explained: The number of PC's are selected in such a way

that the maximum variance is explained by the components Here also the no. of components that may be considered are 15.



- Average root method:The mean of eigen values are calculated and those eigenvalues are selected for PC's that are greater than the mean.The number of PC's selected in this way are nearly 12. Here we got mean eigenvalue as 1.00004703226. There are 12 eigenvalues that are greater than 1, hence getting 12 PC's.

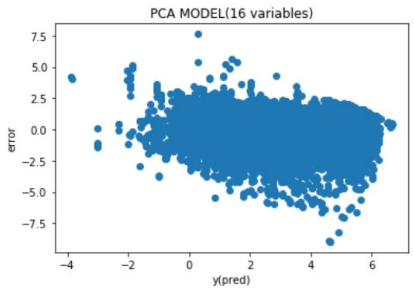
GOODNESS OF FIT:

- The goodness of fit of the model is calculated with PCA model of 16 principal components.
- The results are as shown below:

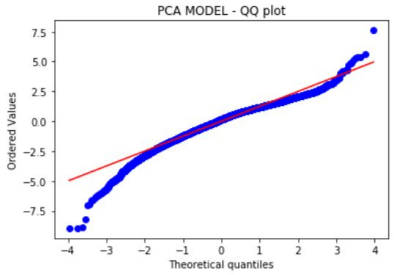
OLS Regression Results

Dep. Variable:	y	R-squared:	0.671
Model:	OLS	Adj. R-squared:	0.671
Method:	Least Squares	F-statistic:	2293.
Date:	Mon, 26 Nov 2018	Prob (F-statistic):	0.00
Time:	21:37:52	Log-Likelihood:	-31705.
No. Observations:	19136	AIC:	6.345e+04
Df Residuals:	19118	BIC:	6.359e+04
Df Model:	17		
Covariance Type:	nonrobust		

- 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.



CONCLUSION:

- **The model that fitted the given data was that applied for 81 variable data(initial dataset) after applying BOX-COX transformation on the dependent variable. The goodness of fit measured 0.933 Adjusted R square value.**
- **As there were many correlations, we have also tried to implement PCA to reduce the dimension and have done model adequacy tests to determine the number of components to be considered.**