**How transformations can help in removing Skewness and increase accuracy of Linear Regression?**

Linear regression is a popular algorithm of Machine learning and mainly used for predicting continuous target variable like-price, age, height and weight etc. But for fitting Linear Regression model, there are few underlying assumption,which should be followed before applying this algorithm on data. But in practice people do not care about these assumptions much and directly apply this algorithm on data. Which can cause the wrong results.

Firstly I will give a brief idea about these assumptions.

1. **Regression model should be linear in the coefficients (means change in target variable due to one unit change in independent variable should be constant). And violations of this assumption can cause serious issues for e.g. If you apply linear regression on non linear data then you will definitely get wrong results and model performance will be bad.**
2. **No multicollinearity (no correlation b/w independent variables).**
3. **Linearity and independence of residuals or no autocorrelation (we can check linearity drawing the scatter plot between residuals (x-axis) and target variable values (y-axis) and independence means no correlation in error terms).**
4. **Residuals should follow Normal distribution with zero mean and equal variance (homoscedasticity).**

***Possible approaches to check these assumptions:-***

1. For First assumption, you can draw scatter plot between fitted and normalized residuals or check predicted vs observed values plot and if there is any pattern in the plot that means data is non-linear and your model does not fit well.
2. For this assumption you can check correlation matrix or draw heat map and find out if there is any correlation between any two independent variables or not.
3. We can check third autocorrelation(independence of residuals) using Dublin Watson test

[Value of this statistic should lie between 0 and 4.

Value between 0 and 2(+ve autocorrelation)

Value =2(no autocorrelation)

Value between 2 and 4(-ve correlation)]

Other way to check this assumption is to plot residual vs fitted value and check for any pattern.

1. For draw qq-plot or perform any test of normality like- Shapiro-Wilk test or Kolmogrov test and for equality of variance.

**Now I will focus on When and why should we use transformation?**

1. To fix non-linearity of data ,You can perform some Non-linear transformations on Independent variables or target variable.
2. We can use transformation approach to handle heteroskedasticity (non-constant variance) by applying transformation on target variable.
3. If errors do not follow Normal distribution, then we can use Non-linear transformations on independent or target variable.

Now I will focus on the issue of Skewness.

**What is Skewness?**

Skewness  is a measure of symmetry or we can say it is also a measure for lack of symmetry, and sometimes this concept is used for checking lack of Normality assumption of Linear Regression.

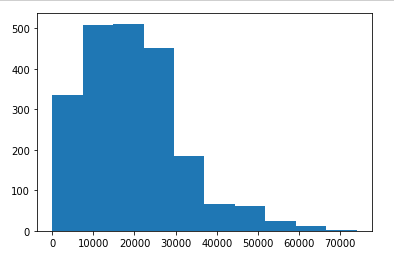
Why should we focus on Skewness ?

Sometimes we see that we did everything right for good performance our Linear Regression Model but still we did not get good accuracy. We tuned hyper parameters and still same issue i.e. bad performance of model. But maybe we are forgetting something i.e. our data is following all assumptions of Linear Regression and when we checked we find out that our data is skewed and violating the 4th assumption (Normality).Hence Skewness is a serious issue and may be the reason of bad performance of your model.

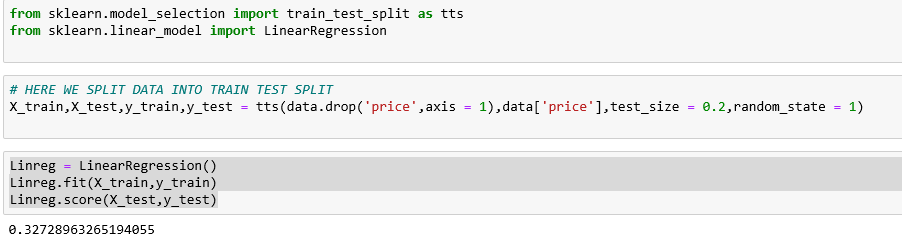
It is not always possible to address Skewness of data. But If We have a choice to reduce remove Skewness ,We should try to do it. Because we aim to achieve good accuracy.

I will give an example and give clear idea about How Skewness will affect your Linear Regression model accuracy.

Here I will use car data set to predict price. I will not use too much code here. I will attach data csv file and code file in the conclusion of the article. So first let’s see the target variable’s histogram (because my target variable is continuous, so I am using histogram here) i.e.



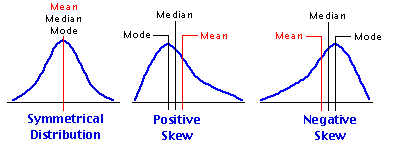
data is positively skewed and Now I will use multiple linear regression model to predict price.



The accuracy score is 32.7% means If I have 100 values to predict then 32-33 values will be correctly predicted.

So let us get the basic idea of Types of Skewness first and then we will go for the handling techniques.

Types of Skewness:-



**Positive skewed:-**

Positive skewed distribution has long tail towards the positive direction of the number line. This is also known as right skewed distribution.

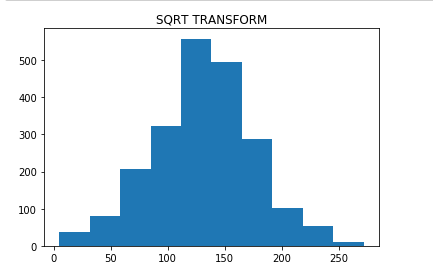
**Negatively skewed**:-

Negatively skewed distribution has long tail towards the negative direction of the number line. This is also known as negatively skewed distribution.

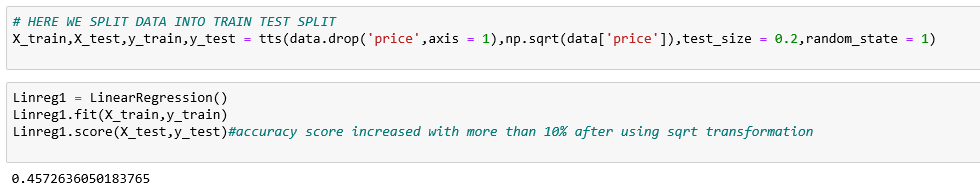
**Types of Transformations to deal with skewed data:-**

1. **SQRT TRANSFORMATION:-**

This transformation can work well on positively skewed continuous data sometimes .Here for this case this transformation is doing pretty well and data is looking alike normal data.



Let’s fit the Multiple linear regression model again and try to see if accuracy has increased or not.

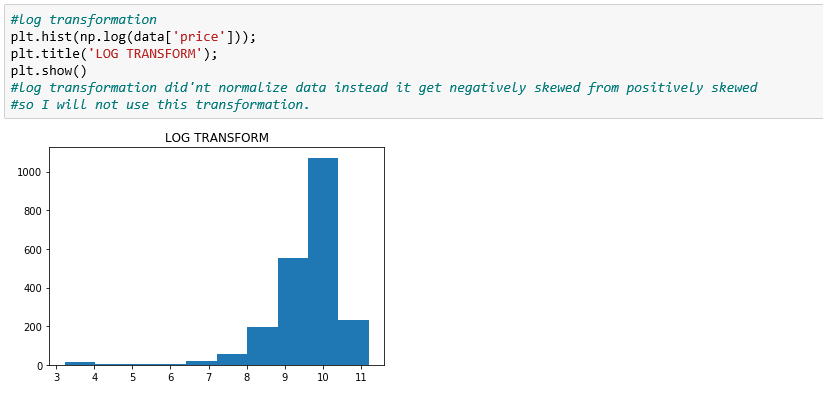


Wow accuracy has increased with more than 10%.For this transformation I have applied sqrt function from numpy on the target variable.

1. **Log Transformation:-**

log transformation is one of the most popular transformations to deal with skewed data. But people usually ignore this point that If the original data follows a log-normal distribution or approximately, then log-transformed data follows a normal or near normal distribution and does remove or reduce skewness.

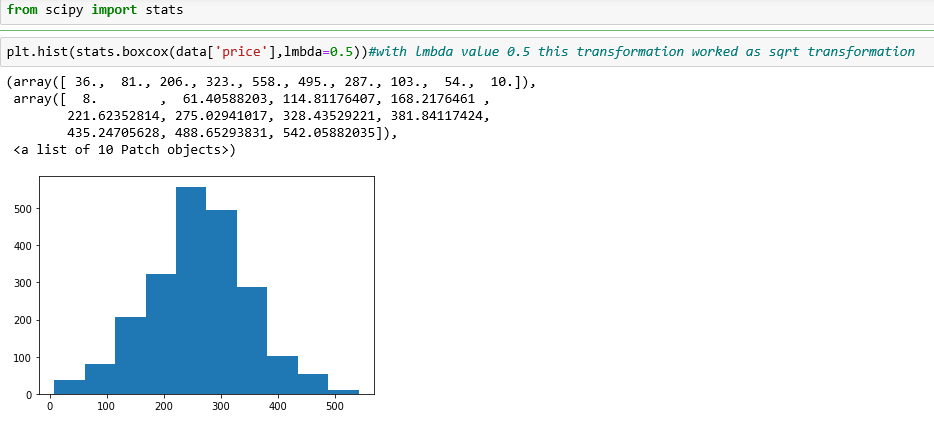
In my case this transformation doesn’t suit well ,because my original data is not following log normal curve, so there is no point in fitting the Multiple Linear Regression model using this transformation and we can check this from the histogram of log transformed data i.e.



1. **BOXCOX-TRANSFORMATION:-**

This is also a popular transformation to handle skewed data .In python You can do it using stats module from scipy.

In my case It’s working same as square root transformation.



Here [lambda= -1( reciprocal transform)

lambda*=* -0.5 ( reciprocal square root transform)

lambda*=* 0.0 ( log transformation)

lambda = 0.5 (square root transformation)

lambda = 1.0 ( no transformation)]

1. Exponential Transformation of target variable or predictors if data looks like exponential curve then this transformation can linearize your data.
2. Polynomial Transformation of target variable or predictors

**Conclusion:-**

I am writing this article because I have seen in practice pople generally forget about the assumptions of linear regression model and that lead to poor performance of linear regression model or even give wrong predictions if used for forecasting. So It is suggested that before performing this algorithm one should check all the assumptions and if any assumption is violating by data then immediately fix that violation and make sure that change should not cause of violating other assumptions.

My point to introduce Skewness was related with Normality assumption. If you have skewed data either predictors or target variables ,you should work on to remove Skewness before applying this algorithm.

**References:-**

1. <https://www.statisticshowto.com/probability-and-statistics/skewed-distribution/>
2. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4120293/>
3. https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/