Data Science

Machine Learning

8/19/2019

Analytics Stream

MADAN MARAM

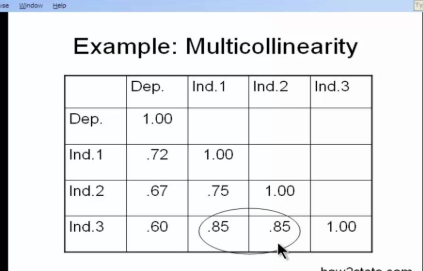
**Statistics**

**multi co-linearity**: if we have two independent variables ,like: Education, Experience .it should be related to each other is called Multi co-linearity

 multicollinearity refers to which independent variables are correlated. Multicollinearity exists when:

* One independent variable is correlated with another independent variable.
* One independent variable is correlated with a linear combination of two or more independent variables.

|  |  |  |  |
| --- | --- | --- | --- |
| Idno | Education(indvar) | Exp(indvariable) | Sal(target variable) |
| 1 | Degree | 5 | 50,000 |
| 2 | B tech | 4 | 55,000 |
| 3 | diploma | 3 | 25,000 |
| 4 | B tech | 3 | 45,000 |



**client:**

**suppose we have a data set with information about the area of house (in sqft) and its price(thousand of dollars) and our task is built a machine learning model which can predict the price given area**

**in simple linear regression we establish a relation between target variable and input variable by fitting a line known as regression line .**

**A new line can be represented by a linear equation**

**y = mx + c**

**y - dependent**

**x - independent**

**c - connstant**

**Assumptions of Linear Regression**

**Introduction**

A Regression Model is the one where we predict the target values with the help of the independent variables. Target values in Regression models are continuous and not discrete. There are different kinds of regression models. These models differ based on the relationship between the Independent variable and the dependent variable, and the number of independent variables that are being used.

One of the basic model is Linear regression.

Linear Regression does the task of predicting the dependent variable value(y) with the help of one or more independent variables (x or xi where i = 1, 2, 3 .. n). This regression

technique finds out a linear relationship between x(input) and y(output). Hence, the name “Linear regression”.

Before diving deep into to model, let’s understand when one should choose to apply this model.

**When to use Linear regression?**

Linear regression is a parametric model. Parametric models are the models that make some assumptions about the data for the purpose of analysis. We need to keep in mind the assumptions for linear regression otherwise the model fails to deliver good results with the data.

Thus, for a successful regression analysis, it is very crucial to validate these assumptions on the given dataset.

We will take a look at these assumptions.

**Assumptions of Linear regression**

There are about 8 major assumptions for Linear Regression models.

major important is 4. we can easily understand these 4 assumptions use LINE .L->Linear ,I->Independent ,N->Normal distribution , E->Equal variance of residuals

1. Linear relationship between Independent and dependent variables.

2. Number of observations should be greater than number of independent variables.

3. No multi-collinearity in independent variables.

4. The variance in the independent variables should be positive.

5. Mean of residuals should be zero.

6. No auto-correlation between the residuals.

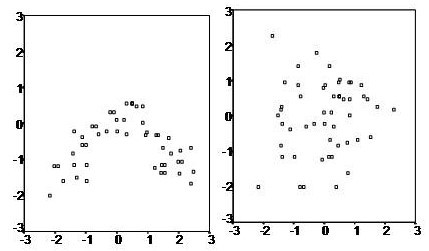
7. Residuals must be normally distributed

8. Residuals should be constant or equal variance i.e Homoscedasticity.

Let us dive deep into the assumptions.

***Assumption 1****: The relationship between Independent and Target variables should be linear.*

For Linear regression, the relationship between the independent variables needs to be linear. The linearity assumption can be tested using scatter plots.



Scatter Plots between 2 Independent Variables

Above, are the 2 cases, where there is no linearity or very little linearity is present in the data.

Mathematically, model equation should be linear in parameters.

Y = B0 + (b1\* X1) + (b2\*X2^2) + …

Although, X2 is raised to power 2, the equation is still linear in parameters.

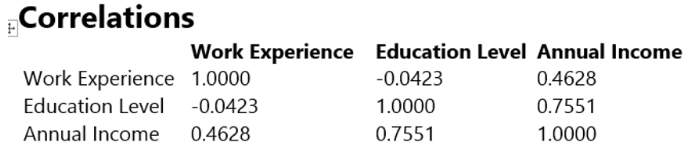
***Assumption 2:****Number of observations should be greater than number of independent variables.*

As part of regression analysis, the dataset should contain at least 20 examples for each independent variable. Most importantly, the number of data points we have in the data set should be more than the number of independent variables.

***Assumption 3:****No multi-collinearity in independent variables*.

(**multi co-linearity**: if we have two independent variables ,like: Education, Experience .it should be related to each other is called Multi co-linearity)

There should not be perfect linear relationship between the independent variables. In other words, there should be no or little multi-collinearity between the independent variables. We should always look for ways where we can minimize the linear relationship between the independent variables.



Otherwise,

1. It becomes very difficult to find out which variable is actually contributing to predict the dependent variable’s value.

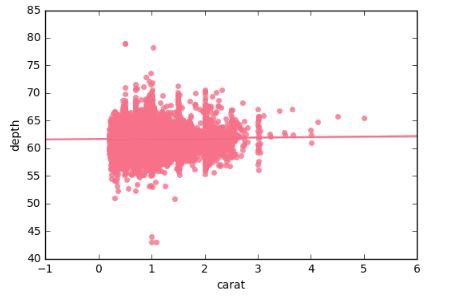
2. In presence of correlated variables, the standard error increases.

Multi-collinearity can be tested with below criteria:

a. **Scatter plot between the variables**

Very basic method is to visualize the data by plotting scatter plots between the variables. But this method becomes very difficult if you have more than 3 variables. The number of scatter plots will increase drastically and will be cumbersome for comparison and deriving insights.

Here, we have taken the Diamond dataset and plotted the scatter plots between the variables. This gives a very good understanding about the distribution of the values of the variables and the behavior of one variable with another.



Depth vs Carat

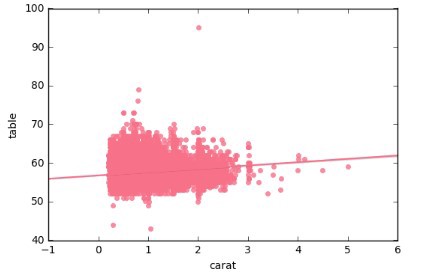


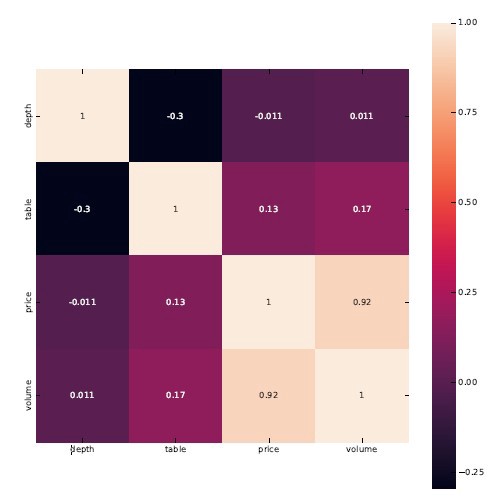
Table vs Carat

b. **Correlation matrix**

Compute the matrix of Pearson’s bivariate correlation among all the independent variables. Choose only one of the variables in the pairs where correlation coefficients are high. Values over +/- 0.75 can be considered as having high correlation coefficients.

Note: Correlation coefficient values can be negative also.

Correlation plot of the numerical variables in Diamond dataset. We can see the correlation values as well as the color coding for the correlation values. We can notice that price and volume have correlation value of 0.92. These 2 variables are highly correlated.



Correlation Matrix for Variables in Diamond Dataset

c. **Tolerance**

The tolerance measures the influence of one independent variable over other independent variables. The tolerance is calculated with initial regression analysis.

It is defined as T = (1- R-squared).

T < 0.1; indicates Presence of multicollinearity

T < 0.01; indicates Certain presence of multicollinearity.

d. **Variance Inflation Factor (VIF)**

VIF is the ratio of variance in a model with multiple terms, divided by the variance of a model with one term alone. It quantifies the severity of multicollinearity.

It is defined as VIF = 1 /(1-R2)

= 1/(Tolerance).

R2 = R-squared value for the X variable under consideration against all other X variables in the linear model. (Note: X variables are Independent variables)

VIF value greater than 4, there is an indication of multi-collinearity. The greater the VIF value, the more the presence of multi-collinearity.

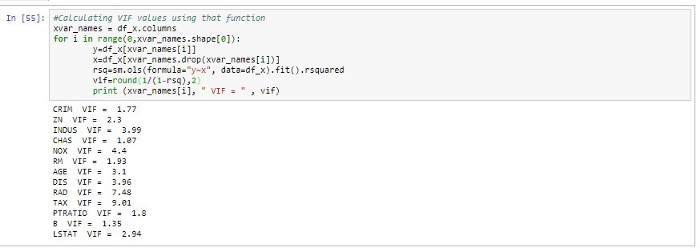
In other words, VIF is a metric that is computed for each independent variable that goes into the linear model. If the VIF value is high for a variable then the information in that variable is already explained by another one or more variable in the given model. The more the VIF value for a variable, the more is redundant is the variable.

There are 2 ways we can take action. One is, remove the variable with high VIF or see the correlation between all variables and keep only one variable from highly correlated pairs.

There is also another way to deal with multicollinearity: Centering the data i.e. deducting the mean of the variable from each value. But this approach is debatable among experts. A topic that can be taken up separately in my next article.

We have taken Boston housing dataset and performed a Linear Regression on it. We can see the VIF values for the columns at the end.





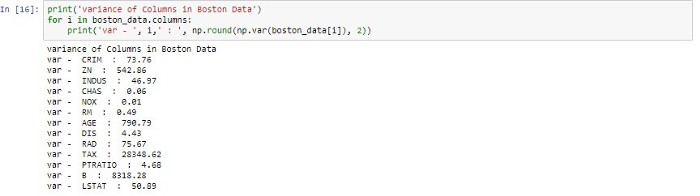
We can notice that variables RAD, TAX are very high. We can remove the variables with highest VIF value and perform the Linear Regression again. We can then, check the VIF values for variables again and see of any variable’s VIF value is greater than 4. We can iterate this till we have bunch of variables with VIF values less than 4. Then, we can say that we have eliminated multicollinearity from the model.

***Assumption 4:****The variance of the independent variables should be positive.*

The variability of the independent variables must not be the same. Also, the variance of the independent variables should be greater than zero.

So, what does this mean, if variability of the independent variables is same then the variables are related to each other. Another point, if variance is Zero for an independent variable, then all the entries for the variable is same.

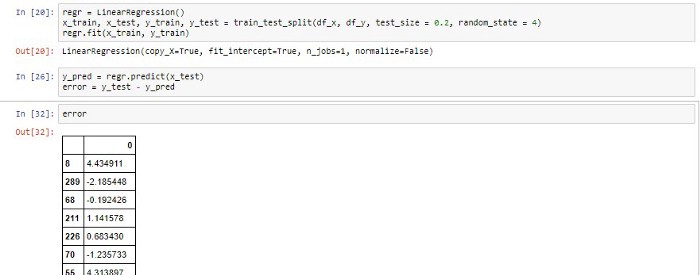
We will observe the Variance of the independent variables in Boston Housing data. We can notice that variances are different and greater than zero.



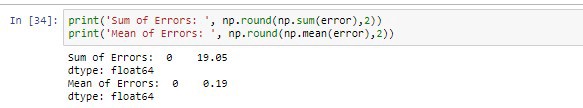
***Assumption 5:****Mean of residuals should be zero*

A residual is the vertical distance between the data point and the regression line. Each data point has a residual. They can be positive, negative or zero. Residuals are also call errors.

The sum of the residuals is always zero or near zero value. Thus, if sum is zero, mean is also zero.



Residuals are the error terms. We can see that error terms are Positive as well as Negative.



Now, the Sum of the errors is 19 and Mean is 0.19. Once we start tuning the model along with new features we can observe that sum of error (and Mean of the errors) will start approaching towards Zero.

***Assumption 6:****No auto-correlation between the residuals.*

Linear Regression requires that the residuals are have very little or no auto-correlation in the data. Auto-correlation happens when the residuals are not independent of each other. Error(i+1) term is not independent of Error(i) term. In other words, the current residual value is dependent on the previous residual value. Presence of auto-correlation drastically reduces the accuracy of the model.

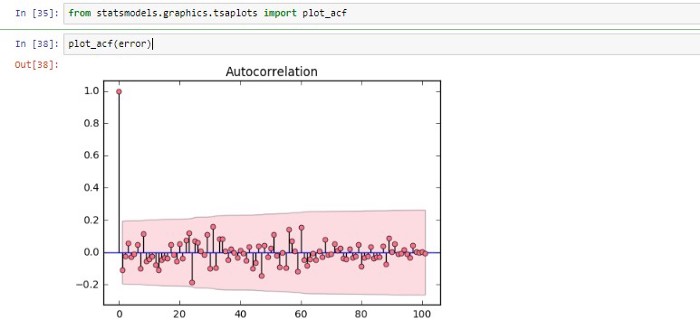
Some basic ways to check or auto-correlation:

· Use ACF plot

X-axis defines the lags of the residuals, increasing in steps of 1. 1st line in the ACF plot (read from left to right), corresponds to correlation of residual with itself (lag 0), therefore, if will always be 1. If residuals are not auto-correlated, the correlation (Y-axis) from the next line onwards (Lag 1) will drop to below the blue line (significance level).

We shall plot the ACF plot for our Residual terms (or Error terms) that we found above. As described above, 2nd vertical line onwards, all the lines are below the Horizontal Blue lines (Pink shaded area). Here, 0.2 is the significance level (blue line).

Our residual terms in the Boston Housing Price model, are not auto-correlated.



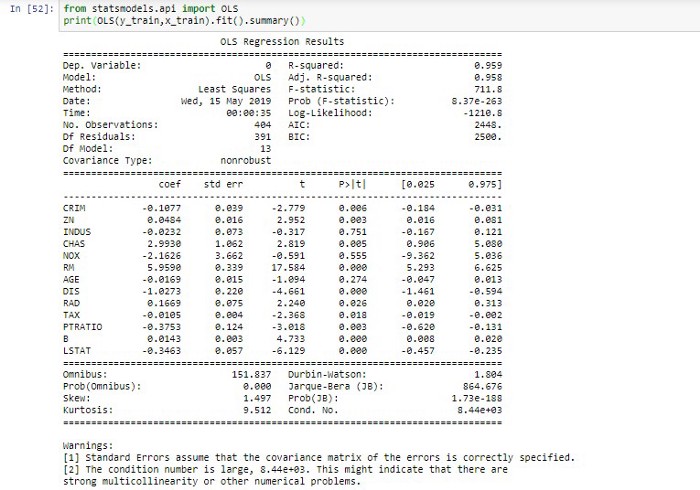
· Durbin-Watson test

Durbin-Watson’s d-test tests the Null hypothesis that residuals are not linearly auto-correlated. Here, d has range of values from 0 to 4 where values around 2 indicate no autocorrelation.

Thumb rule, 1.5 <d < 2.5 indicates no auto-correlation.

Statsmodels package in python provides the summary of the model. We will take the output of Linear Regression model (OLS) from this package for summary.

We can see the summary of our Linear Regression model of Boston housing price data. Durbin-Watson’s d-test value is 1.804 which within our thumb rule range. Thus, NO auto-correlation in the residuals.

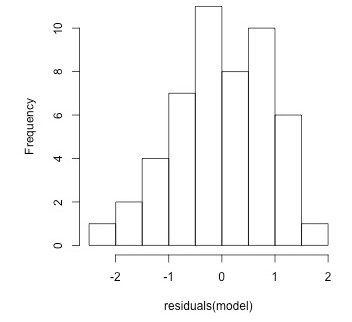


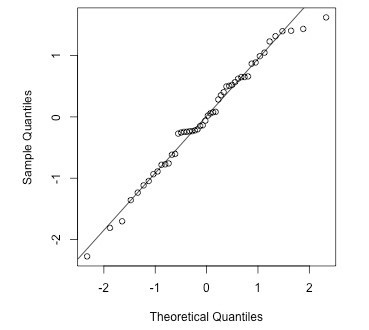
There are many other tests for finding our auto-correlation. But for testing our assumption, above 2 tests are sufficient.

We can deal with autocorrelation by introducing lag 1 of residual as an X-variable to the original model.

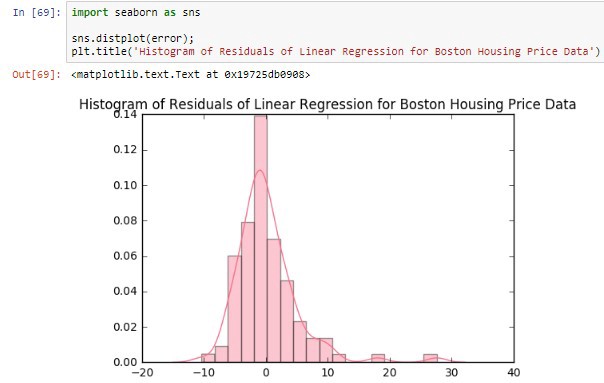
***Assumption 7:****Residuals must be normally distributed*

In a linear regression model, the error terms are normally distributed. So, we can say we are trying to understand the Normal probability plot of the errors or the residuals. If a data follows normal distribution, then a plot of theoretical percentiles of the normal distribution versus the observed sample percentiles should be approx. linear. Since our discussion is about the residuals, if the resulting plot is linear, then the residuals are normally distributed.





Let us visualize the same for our dataset.



The histogram for our Residuals (error) looks approximately Normal (Gaussian) distribution.

***Assumption 8****: Residuals should be constant or equal variance i.e Homoscedasticity*

If the variance is not constant across the error terms, then there is a case of heteroscedasticity. These non-constant variance across the error terms are due to the presence of outliers in the original data. These outliers influence the model to a huge extent.

Simple scatter plot of the residuals can highlight the presence of heteroscedasticity.

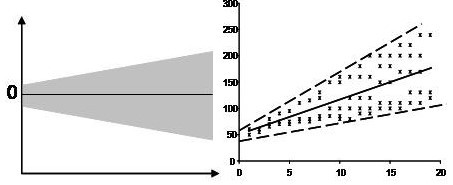
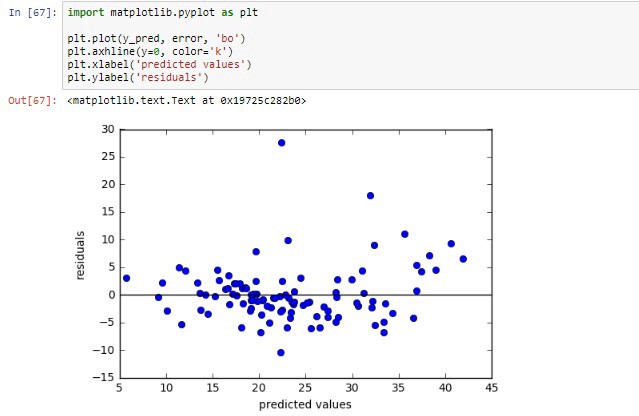


Fig 1

Visualizing the error and fitted values from our dataset.



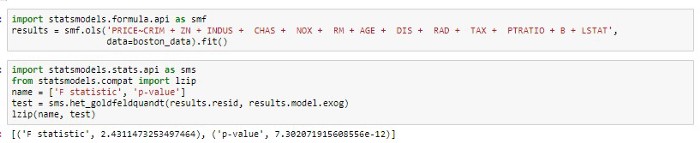
Above scatter plot shows the distribution of Residuals and Predicted Values (or Fitted Values).

There is no pattern in the plot, specifically no parabolic pattern, that means, the model has captured non-linear effects (if any) also. There is no funnel shape distribution in the plot, this means, signs of homoscedasticity. If a funnel shaped distribution was present as shown in Fig.1, then consider signs of non-constant variance i.e. Heteroscedasticity.

We can also, check this with the help of residual vs fitted plot. There are also statistical tests like Goldfeld-Quandt Test, Breusch-Pagan / Cook — Weisberg test or White general test.

Let us see one of the Test: *GoldFeld-Quandt Test*

Aim is to test whether variance is same in 2 sub-samples. Null hypothesis is that variance of 2 sub-samples are same. If the variances differ, the test rejects the null hypothesis that the variances of the errors are *not*constant. It corresponds to F-test for equality of variances. This, large values of F-values indicate that the variances are different. Let us check the same for our Boston Housing dataset.



Similarly, other tests can also be performed to check the homoscedasticity.

**Conclusion**

We can leverage the actual power of regression to its maximum extent by understanding the nuances of the assumptions and applying the solutions described above. I have tried to explain the implementations using Python.

The motive behind the article was to gain the intuition and insights of regression assumptions. Also, we have not covered all the tests as mentioned in the article. But, I guess, this helps as a starter for exploring other tests and methods.

*If you liked the article, feel free to give me claps and help others to find it.*

Also, let me know if I have missed out assumption/topic. Happy to learn and incorporate.

[**Linear regression analysis in Python**](https://medium.com/m/signin?operation=register&redirect=https%3A%2F%2Fmedium.com%2F@asutosh405%2Fassumptions-of-linear-regression-b3d94d2b297f&source=post_sidebar-----b3d94d2b297f---------------------clap_sidebar-)

youtube(<https://www.youtube.com/watch?v=NUXdtN1W1FE>)

linear regression medium app(<https://towardsdatascience.com/multiple-linear-regression-in-four-lines-of-code-b8ba26192e84>)

Gitlink([https://github.com/MadanMaram/MLProjects/tree/master/linear%20regression](https://github.com/MadanMaram/MLProjects/tree/master/linear regression))

