**Machine Learning : 'Regression' - Day 2**

* Published on Published onFebruary 25, 2018

[[](https://www.linkedin.com/in/panchalshivam/)](https://www.linkedin.com/in/panchalshivam/)

**[Shivam Panchal](https://www.linkedin.com/in/panchalshivam/)**

 2

*  5

[Write an article](https://www.linkedin.com/post/new)

Welcome to the post, I will not bore you **much** with the theory behind, I will try to put it as easy as possible.

In this post, we will get deep into applicability of regressive model unlike to the first post which was more likely a beginner introduction to Regression and we will also see **how do we interpret a regression model**.

Well, if you haven't see the first post, here is the link for you, please do check.

The main objective of the post will also include the understanding of the **various assumptions in Regression model**. Moreover, we will also discuss what if these assumptions get violated, how do you build your linear regression model then.

Believe me, regression is not just fitting a line to the predicted values or defining it as an equation (y = m\*x + b) like I did in the previous post, there is so much into it. That’s the main idea of this post. I had to recollect my thoughts on these things again today for this post.

One more thing, I would like to mention is that, regression is considered to be the simplest algorithm in ML. When we start playing with ML, most of us start with regression, but it is not understood well by the beginners, I was asked a few questions in one of my initial technical interviews when I applied to a job sometime back. I wasn't comfortable that time, It is important and these things describe how well you understand the math behind ML.

So make sure, Machine Learning is not just loading classes from sci-kit learn and fitting data and predicting targets, it is something more than that.

So, let’s begin the Day 2nd!!



\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Regression is a parametric approach** by saying it a ‘Parametric’ approach, I mean it is going to make assumptions about your data for the purpose of analysis. And, because of this reason, it has some limited uses and other regression techniques like tree based regression and deep nets are used practically. **Linear Regression surely fails to deliver good results with data sets which doesn’t fulfill its assumptions.** Therefore, for a successful regression analysis, it’s essential to validate these assumptions.

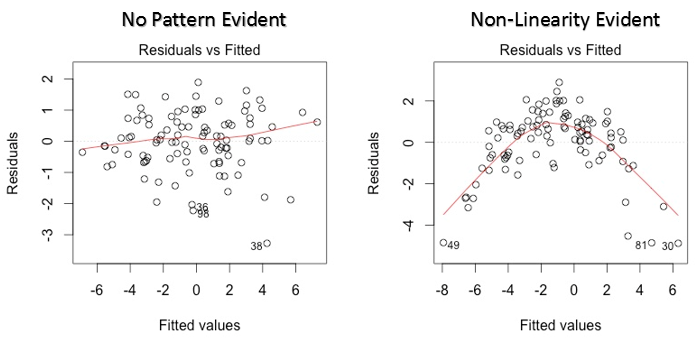
So, how would you check if your Mr. Data Set follows all regression assumptions?

So, let’s look at the important assumptions in regression analysis one by one.

**1.Relationship:** There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set and it will be counted as a mistake of your life.

**Solution:** To solve this, you should look for **residual vs fitted** value plots . Also, you can include polynomial terms (X, X², X³) in your model to capture the non-linear effect.

***Residual vs Fitted Values -*** This scatter plot shows the distribution of residuals (errors) vs fitted values (predicted values). It is one of the most important plot which everyone must learn. It reveals various useful insights including outliers. The outliers in this plot are labeled by their observation number which make them easy to detect.



There are two major things which you should learn:

If there exist any pattern (may be, a parabolic shape) in this plot as you can see above, consider it as signs of non-linearity in the data. It means that the model doesn’t capture non-linear effects. Introduce the non-linearity by doing transformations. You can do a non-linear transformation of predictors such as log (X), √X or X² transform the dependent variable.

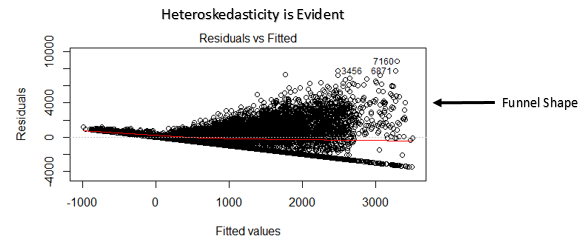
**2. Auto Correlation:** No correlation between the residual (error) terms. The presence of correlation in error terms affects your model’s accuracy.

**Solution:** Look for Durbin – Watson (DW) statistic. It must lie between 0 and 4. If DW = 2, implies no autocorrelation, 0 < DW < 2 implies positive autocorrelation while 2 < DW < 4 indicates negative autocorrelation.

**3.** **Multicollinearity:** *The independent variables should not be correlated.* It is said to be existing when the independent variables are found to be moderately or highly correlated. It becomes difficult to find out which variable is actually contributing to predict the response variable, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of coefficients.That’s not good!

**Solution:** You can use scatter plot to visualize correlation effect among variables. Define a correlations threshold in your mind and say good bye to the highly correlated variables.

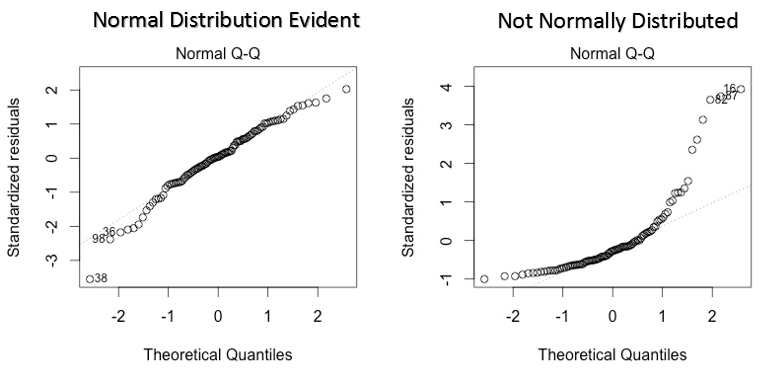
**4. Heteroscedasticity :** The error terms must have constant variance. This phenomenon is known as homoscedasticity. The presence of non-constant variance is referred to heteroscedasticity. Generally, it arises in presence of outliers or extreme values. And then these values get too much weight, thereby affecting the model’s performance.



**Solution**: Put your eyes again on the residual vs fitted values plot. If heteroscedasticity exists, the plot would exhibit a funnel shape pattern. If a funnel shape is evident in the plot, consider it as the signs of non-constant variance i.e. heteroscedasticity. To overcome heteroskedasticity, a possible way is to transform the response variable such as log(Y) or √Y. Also, you can use weighted least square method to tackle heteroscedasticity.

**5. Normal Distribution of error terms:** The error terms must be normally distributed. If the error terms (residuals) are non- normally distributed, confidence intervals may become too wide or narrow. Model deteriorates.

**Solution:** Check again the QQ plot. Your residuals should incline the mod line, i.e. 45 degree line there, if it is not aligned, there is not normal distribution in your residuals, your model has some problems, perform transformations and optimise your QQ plot.



**Normal Q-Q Plot** - It is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. Absence of normality in the errors can be seen with deviation in the straight line. If the errors are not normally distributed, non – linear transformation of the variables (response or predictors) can bring improvement in the model.

I hope, I was able to explain this properly! Feel free to criticize me in the comments section below.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Implementation in Python!

Link to the GitHub Repo!

Let's summarize, what we discussed today!

1. Detecting co linearity (scatterplots)
2. Diagnosing model fit (analyze model)
3. Assumptions of linear regression
4. Transforming features to fit non-linear relationships (scale your data, transform your data, using log, or powers)
5. Interaction terms (in your formula, use y ~ x1 + x2\*x3 structure to introduce interactions)

And so much more!

Notes

You could certainly go very deep into linear regression, and learn how to apply it really, really well It's an excellent way to start your modeling process when working a regression problem However, it is limited by the fact that it can only make good predictions if there is a linear relationship between the features and the response, which is why more complex methods (with higher variance and lower bias) will often outperform linear regression