

## 6 Assumptions of Linear Regression :Plots and Solutions

Home > Algorithm > 6 Assumptions of Linear Regression :Plots and Solutions



avcontentteam

24 Sep, 2023 • 8 min read

# B2BESTADFORMATS

Regression analysis is crucial in [predictive modeling](#), but merely running a line of code or looking at  $R^2$  and MSE values isn't enough. In R, the `plot()` function generates four plots that reveal valuable insights about the data. Unfortunately, many beginners fail to interpret these plots. This article explains important regression assumptions, fixes for violations, and the significance of these plots. Understanding these concepts can greatly enhance your [regression models](#). Read on to learn all about the assumptions of [linear regression](#) and polynomial regression.

All models are wrong, but some are useful

*George Box*

[Table of contents](#)

Reading list

## What are Assumptions in Regression?

## 6 Assumptions of Linear Regression :Plots and Solutions



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 (validate) if a data set follows all regression assumptions? You check it using the regression plots (explained below) along with some statistical test.

### What are Assumptions of Linear Regression?

Violations of assumptions of linear regression can lead to biased or inefficient estimates, and it is important to assess and address these violations for accurate and reliable regression results.

6 Assumptions of linear regression include:

1. **Linearity:** The relationship between the dependent and independent variables is linear.
2. **Independence:** The observations are independent of each other.
3. **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
4. **Normality:** The errors follow a normal distribution.
5. **No multicollinearity:** The independent variables are not highly correlated with each other.

## 6 Assumptions of Linear Regression :Plots and Solutions



### IMPORTANT ASSUMPTIONS IN REGRESSION Analysis

Let's look at the important assumptions in regression analysis:

The banner has a purple background. At the top left is the DataHour logo with an upward arrow icon. In the center, the title 'Significance of Vector Databases in Gen AI' is displayed in white. To the right is a circular profile picture of a man with a mustache. Below the title, the speaker's name 'Shanthababu Pandian' and title 'Solution Architect' are shown. At the bottom left, there are icons for a calendar and a clock, with the date '30 January 2024' and time '6:00 PM - 7:00 PM IST'. At the bottom right is the JPOWER4 logo.

1. There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in  $X^1$  is constant, regardless of the value of  $X^1$ . An additive relationship suggests that the effect of  $X^1$  on Y is independent of other variables.
2. There should be no correlation between the residual (error) terms. Absence of this phenomenon is known as Autocorrelation.
3. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
4. The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to heteroskedasticity.
5. The error terms must be normally distributed.

## 6 Assumptions of Linear Regression :Plots and Solutions



Let's dive into specific assumptions and learn about their outcomes (if violated):

### 1. Linear and Additive

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.

**How to check:** Look for residual vs fitted value plots (explained below). Also, you can include polynomial terms ( $X$ ,  $X^2$ ,  $X^3$ ) in your model to capture the non-linear effect.

### 2. Autocorrelation

The presence of correlation in error terms drastically reduces model's accuracy. This usually occurs in time series models where the next instant is dependent on previous instant. If the error terms are correlated, the estimated standard errors tend to underestimate the true standard error.

If this happens, it causes confidence intervals and prediction intervals to be narrower. Narrower confidence interval means that a 95% confidence interval would have lesser probability than 0.95 that it would contain the actual value of coefficients. Let's understand narrow prediction intervals with an example:

## 6 Assumptions of Linear Regression :Plots and Solutions



prediction interval narrows down to (13.82, 16.22) from (12.94, 17.10).

Also, lower standard errors would cause the associated p-values to be lower than actual. This will make us incorrectly conclude a parameter to be statistically significant.

**How to check:** 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. Also, you can see residual vs time plot and look for the seasonal or correlated pattern in residual values.

### 3. Multicollinearity

It occurs when the independent variables show moderate to high correlation. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

Another point, with presence of correlated predictors, the standard errors tend to increase. And, with large standard errors, the confidence interval becomes wider leading to less precise estimates of slope parameters.

Additionally, when predictors are correlated, the estimated regression coefficient of a correlated variable depends on the presence of other predictors in the model. If this happens, you'll end up with an

## 6 Assumptions of Linear Regression :Plots and Solutions



not good!

**How to check:** You can use scatter plot to visualize correlation effect among variables. Also, you can also use VIF factor. VIF value  $\leq 4$  suggests no multicollinearity whereas a value of  $\geq 10$  implies serious multicollinearity. Above all, a correlation table should also solve the purpose.

### 4. Heteroskedasticity

The presence of non-constant variance in the error terms results in heteroskedasticity. Generally, non-constant variance arises in presence of outliers or extreme leverage values. Look like, these values get too much weight, thereby disproportionately influences the model's performance. When this phenomenon occurs, the confidence interval for out of sample prediction tends to be unrealistically wide or narrow.

**How to check:** You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern (shown in next section). Also, you can use Breusch-Pagan / Cook – Weisberg test or White general test to detect this phenomenon.

### 5. Normal Distribution of error terms

If the error terms are non- normally distributed, confidence intervals may become too wide or narrow. Once confidence interval becomes

## 6 Assumptions of Linear Regression :Plots and Solutions



studied closely to make a better model.

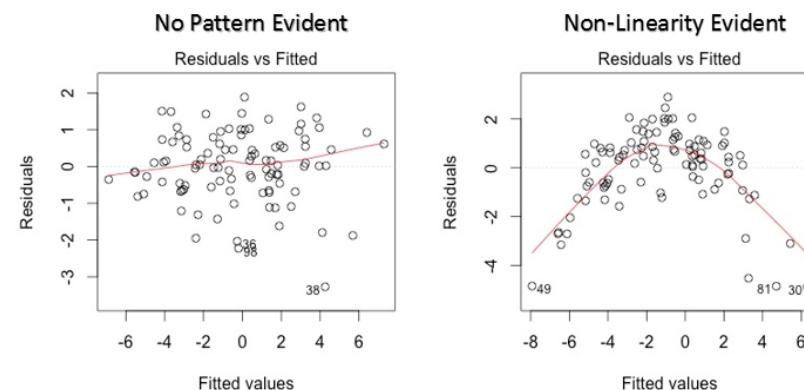
**How to check:** You can look at QQ plot (shown below). You can also perform statistical tests of normality such as Kolmogorov-Smirnov test, Shapiro-Wilk test.

### Interpretation of Regression Plots

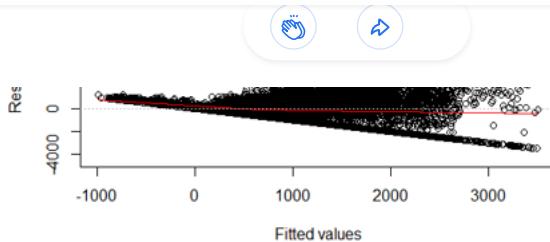
Now, we know all about important regression assumptions and the methods take care of them in case of violation.

But that's not the end. Now, you should know the solutions also to tackle the violation of these assumptions. In this section, I've explained the 4 regression plots along with the methods to overcome limitations on assumptions.

#### 1. Residual vs Fitted Values



## 6 Assumptions of Linear Regression :Plots and Solutions



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.

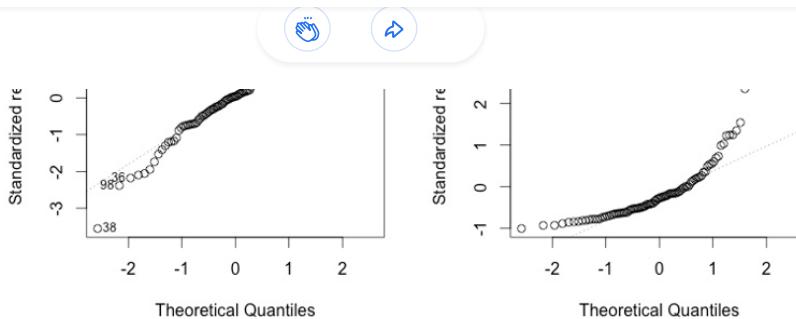
There are two major things which you should learn:

1. If there exist any pattern (may be, a parabolic shape) in this plot, consider it as signs of non-linearity in the data. It means that the model doesn't capture non-linear effects.
2. If a funnel shape is evident in the plot, consider it as the signs of non constant variance i.e. heteroskedasticity.

**Solution:** To overcome the issue of non-linearity, you can do a non linear transformation of predictors such as  $\log(X)$ ,  $\sqrt{X}$  or  $X^2$  transform the dependent variable. To overcome heteroskedasticity, a possible way is to transform the response variable such as  $\log(Y)$  or  $\sqrt{Y}$ . Also, you can use weighted least square method to tackle heteroskedasticity.

### 2. Normal Q-Q Plot

## 6 Assumptions of Linear Regression :Plots and Solutions



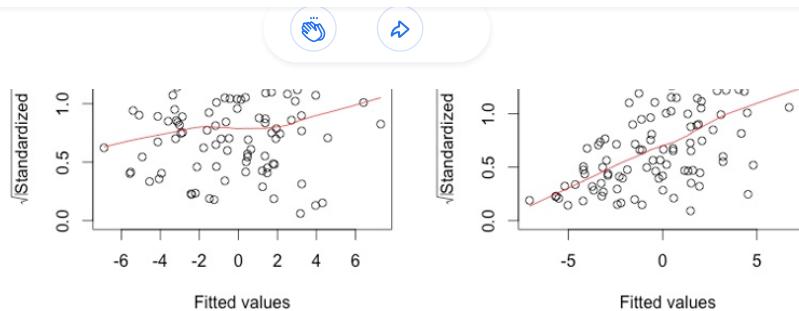
This q-q or quantile-quantile is a scatter plot which helps us validate the assumption of normal distribution in a data set. Using this plot we can infer if the data comes from a normal distribution. If yes, the plot would show fairly straight line. The straight lines shows the absence of normality in the errors.

If you are wondering what is a 'quantile', here's a simple definition:  
Think of quantiles as points in your data below which a certain proportion of data falls. Quantile is often referred to as percentiles.  
For example: when we say the value of 50th percentile is 120, it means half of the data lies below 120.

**Solution:** If the errors are not normally distributed, non – linear transformation of the variables (response or predictors) can bring improvement in the model.

### 3. Scale Location Plot

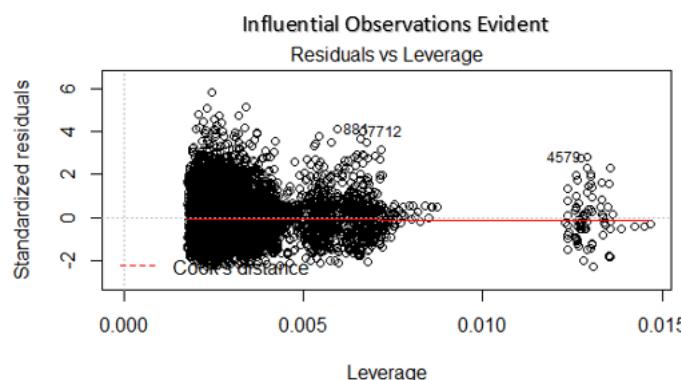
## 6 Assumptions of Linear Regression :Plots and Solutions



This plot is also detects homoskedasticity (assumption of equal variance). It shows how the residual are spread along the range of predictors. It's similar to residual vs fitted value plot except it uses standardized residual values. Ideally, there should be no discernible pattern in the plot. This would imply that errors are normally distributed. But, in case, if the plot shows any discernible pattern (probably a funnel shape), it would imply non-normal distribution of errors.

**Solution:** Follow the solution for heteroskedasticity given in plot 1.

### 4. Residuals vs Leverage Plot



## 6 Assumptions of Linear Regression :Plots and Solutions



removing such points from the model can completely change the model statistics.

But, can these influential observations be treated as outliers? Only data can answer this question.. Therefore, in this plot, the large values marked by cook's distance might require further investigation.

**Solution:** For influential observations which are nothing but outliers, if not many, you can remove those rows. Alternatively, you can scale down the outlier observation with maximum value in data or else treat those values as missing values.

**Case Study:** [How I improved my regression model using log transformation](#)

## Conclusion

In conclusion, understanding and acknowledging the assumptions of linear regression is vital for accurate and reliable analysis. By recognizing the regression assumptions, we can ensure the validity of our models and interpret the results effectively. It is essential to assess the assumptions and address any violations to enhance the reliability of our findings. Adhering to these assumptions allows us to make informed decisions and draw meaningful insights from linear regression analyses in various data science applications.

## Frequently Asked Questions

## 6 Assumptions of Linear Regression :Plots and Solutions



no endogeneity, ensuring valid and reliable regression results.

### **Q2. What are the 4 assumptions for regression analysis?**

A. Regression analysis relies on the assumptions of linearity, independence, homoscedasticity, and normality to interpret and validate the model.

### **Q3. What is assumption 5 econometrics?**

A. Assumption 5 in econometrics refers to no multicollinearity, where highly correlated independent variables can lead to unstable parameter estimates.

### **Q4. What are the assumptions of regression error?**

A. The assumptions of regression error include independence, zero mean, constant variance, and normality, ensuring adherence to the regression model's assumptions.

[advanced regression](#)

[heteroskedasticity](#)

[Homoskedasticity](#)

[interpretation regression plots](#)

[Multiple Regression](#)

[ordinary least square](#)

[regression](#)

[residual plots](#)

[Residual sum of squares](#)

[residual vs leverage plot](#)

[scale location plot](#)

[total sum of squares](#)



[avcontentteam](#)

24 Sep 2023

## 6 Assumptions of Linear Regression :Plots and Solutions



Machine Learning

### Related Articles



#### Diagnosing residual plots in linear regression models

This article lays out how to validate assumptions in a linear regression model. ...

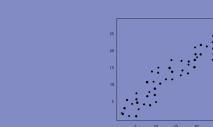
Business Analytics

linear regression



Tavish Sriv...

25 Jun, 2020 • 5 min read



#### All you need to know about your first Machine Learning model –...

Linear Regression is the supervised ML model in which the model finds the best f...

blogathon

linear regression

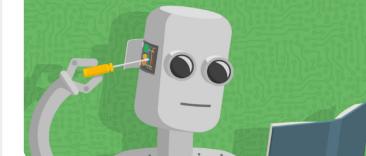


Deepanshi

20 Jul, 2023 • 7 min read

## LINEAR

## REGRESSION



#### 25 Questions to Test Your Skills on Linear Regression Algorithm

Here, we will discuss important questions on the Linear Regression Algorithm, hel...

blogathon

data science interview



CHIRAG GOYAL

22 Jun, 2022 • 13 min read

## 6 Assumptions of Linear Regression :Plots and Solutions



### Scribe, Shine, Succeed →

Write, captivate, and earn accolades and rewards for your work

- ✓ Reach a Global Audience
- ✓ Get Expert Feedback
- ✓ Build Your Brand & Audience
- ✓ Cash In on Your Knowledge
- ✓ Join a Thriving Community
- ✓ Level Up Your Data Science Game



CHIRAG GOYAL

87



Barney Darlington

5



Suvojit Hore

9

Company	Discover	Learn	Engage	Contribute	Enterprise
About Us	Blogs	Free courses	Community	Contribute & win	Our offerings
Contact Us	Expert session	Learning path	Hackathons		Case studies

## 6 Assumptions of Linear Regression :Plots and Solutions



Become an  
instructor

Download App



---

[Terms & conditions](#) • [Refund Policy](#) • [Privacy Policy](#) • [Cookies Policy](#) © Analytics Vidhya 2023. All rights reserved.