MyDataCafe
Upskill Yourself

All you need to Know about Linear Regression

Linear Regression

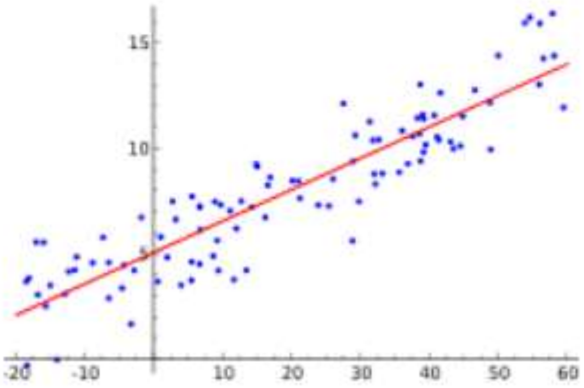
We have seen equation like below in maths classes. y is the output we want. x is the input variable. c = constant and a is the slope of the line.

y = c + ax
c = constant
a = slope

The output varies linearly based upon the input. y is the output which is determined by input x. How much value of x has impact on y is determined by “a”. In the two dimensional graph having axis ‘x’ and ‘y’ , ‘a’ is the slope of the line. ‘c’ is the constant (value of y when x is zero).

Historical Data

Lets say we have good amount of historical data wherein values of input ‘x’ and output/result ‘y’ are provided. We also know that change in x impacts y somewhat linearly. It means if we take the values of these ‘x’s and ‘y’s and plot of graph then we get a graph like this.
Not exactly same but somewhat like this wherein there seems to be clutter of points going in one linear direction.



What do we want to do

We want to look at the historical data and find the values of ‘c’ and ‘a’. Lets call ‘c’ as the constant and ‘a’ as the weight. If we are able to figure out the values of c and a then we can predict the value of y for any new x.

How to do

As you can see from the graph, although the dots form a linear pattern but they are not in same line. So we cant know the 100% perfect value of c and a. But what we can do it to find the best fit line, shown in red above.

This means our output will be approximate. There might be slight deviation from actual value. But it works fine for most business applications.

How in details

Look at the dots in the graph. What if we figure out a line whose distances from each dot in the graph is optimal/minimal.
This would mean 'best fit line' as shown in red in the pic above. Our objective is to draw the red line in the graph above.

So, we draw a random line on the graph for some random value of c and a. Lets say we keep c and a both 1 (c=1, a=1) and draw the
line on the graph for each (well at least 2) x. Based upon values of x this line might end up in one of the following positions

- On leftish side of the dots towards y axis. More vertical.
- On rightish side of dots towards x axis. More horizontal.
- Somewhere between dots but still not best fit line.

Since this was a random line, we need a mechanism to move this line iteratively and slowly towards the place where it best fits
the sample data (dots in the graph).

So, effectively we need

- To find if its best fit line or not
- If its not best fit line then move it towards the best fit line. It means we will have to change the value of c and a.
- How much values of c and a we need to change and in which direction? We will use combination of gradient descent with least square method to achieve these objectives. These are explained below.

Mathematics

For each item in the sample data (called training set too), get the value of y from our estimated line (c=1, a=1). Lets call it h(y).

- Also, we have y which is real value for each sample data.
- We get the difference between approximated h(y) and y as h(y)—y. Square this difference up. So for one sample we have (h(y)—y) ^ 2
- Do it for all data points (rows). Lets say we have sample size as 'm', we get the squares of differences for each sample size, sum it up (summation from 1 to m). We now have our cost function. We need to minimize this cost function. It means we need to minimize the distance of our line with sample data (dots) to get best fit line.
- But we have 2 variables c and a which we need to keep changing to get at best fit line.

- For initial combination of c and a, we need to find how much to move and then move it. Once we have new line for new values of c and a then do the distance calculation with each point again and keep doing it till we find that we are not moving much e.g. moving quite less.
- Since we need to change both c and a independently, we will use partial differentiation. So we will get the derivative of above cost function wrt c and then wrt a. Solving these 2 we can get the values of c and a.

For a Linear Regression Line the Sum of All residuals is always 0 as some errors are positive (where we predict more than the Actual) and some are negative (where we predict less than the Actual).

Key Terms Linear Regression

Response:

The variable we are trying to predict. Also known as Dependent variable, Y variable, target, Outcome.

Independent Variable

The variable used to predict the response. Also known as X variable, Feature, attribute.

Record

A row of a dataframe or more logically a DV and IV's for a specific individual or case. example in mtcars each row denotes a car. so, each row is a record.

Intercept

The intercept of the regression line. i.e. The predicted value when x=0

β 0 or b0

Regression Co-efficient

The slope of the regression line. For unit increase in x how much does the y (DV) variable increase.

$\beta 1$, b1

Fitted values

The estimates obtained from the regression line. Ý (Y hat). These are the predicted values.

Residuals/ Errors

The difference between the fitted/ predicted and the actual values

Least Squares

The method of fitting a regression line by minimizing the sum of squared errors/residuals. Also known as OLS

HAT NOTATION (FITTED vs THE KNOWN VALUES)

the hat notation is used to differentiate between estimates and known values. So b-hat is an estimate (approximation) of unknown value b. Why do we differentiate between the estimate and the True value. Because the estimate has uncertainty, whereas True value is fixed.

#=====================

The regression equation (model) models the relationship between a response variable y and a predictor variable X as a line. A regression model yields fitted values and residuals.i.e. Predictions of the response and the errors of the predictions respectively.

Regression Model Evaluation Parameters

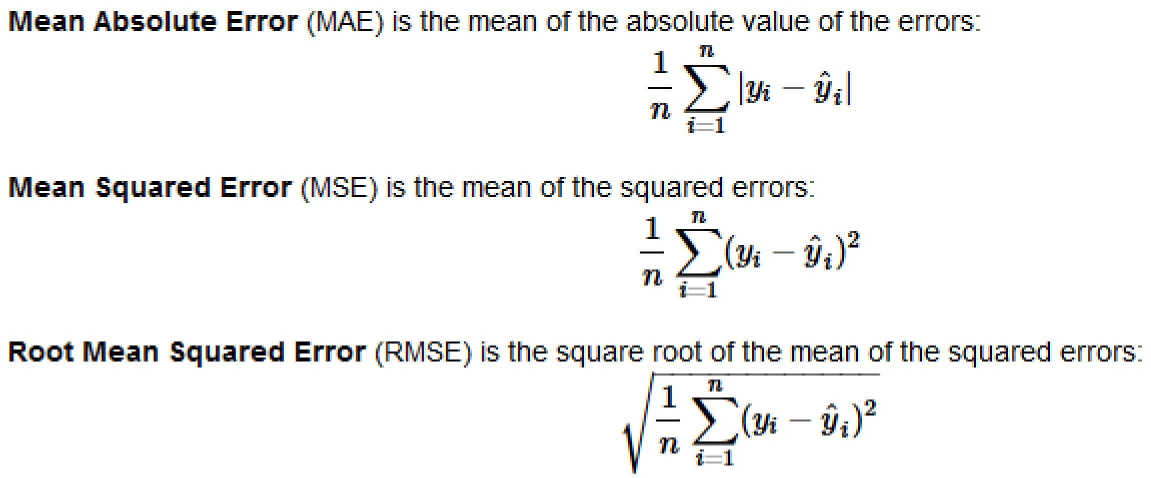
Root Mean Squared Error:

The square root of the average squared error of the regression.

R Squared/ Co-efficient of determination

It is the proportion of variance in the observed data that is explained by the model, or the reduction in error over the Baseline/null model. *The null model just predicts the mean of the observed response, and thus it has an intercept and no slope.*

Apart from MAPE, these error parameters are also used.



Mean Absolute Error (MAE) is the mean of the absolute value of the errors:

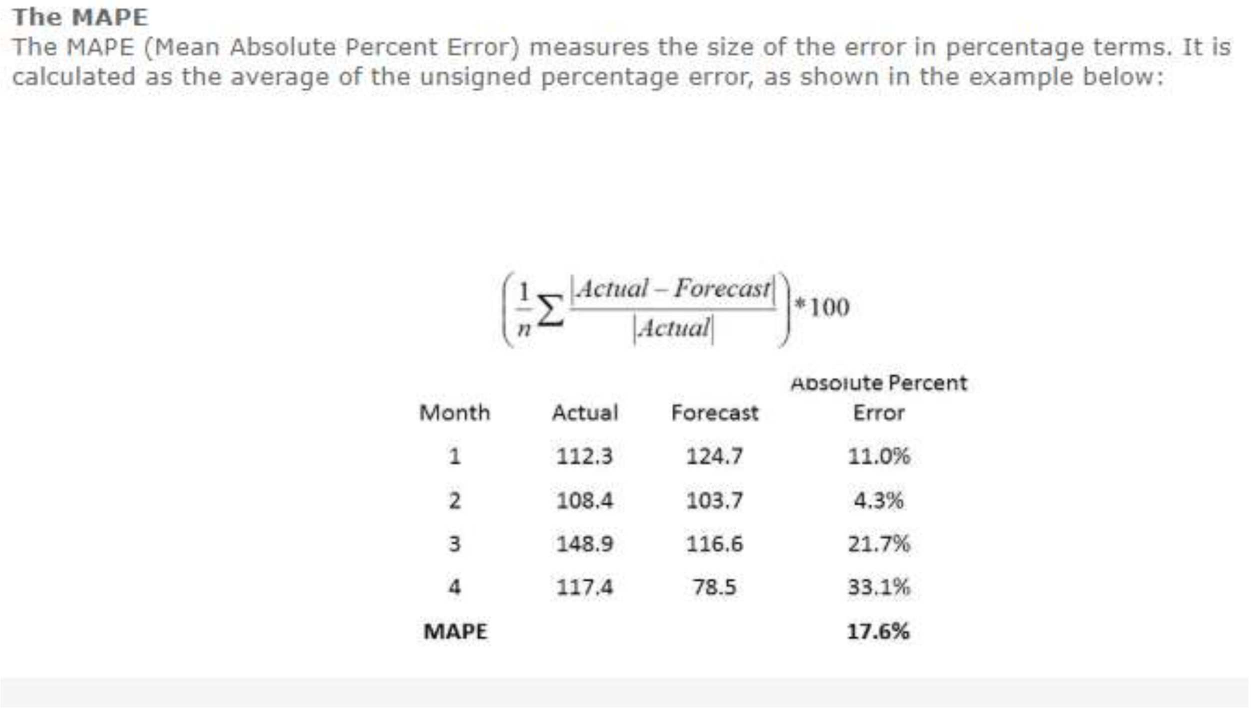
$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error (MSE) is the mean of the squared errors:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error (RMSE) is the square root of the mean of the squared errors:

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$



The MAPE

The MAPE (Mean Absolute Percent Error) measures the size of the error in percentage terms. It is calculated as the average of the unsigned percentage error, as shown in the example below:

$$\left(\frac{1}{n}\sum\frac{|Actual - Forecast|}{|Actual|}\right)*100$$

Month	Actual	Forecast	Absolute Percent Error
1	112.3	124.7	11.0%
2	108.4	103.7	4.3%
3	148.9	116.6	21.7%
4	117.4	78.5	33.1%
MAPE			**17.6%**

A baseline Model is one which always predicts the average.

Key Limitations of R-squared

A high R-squared does not necessarily indicate that the model has a good fit. It has to be validated by calculating the Error parameters also. A model with High R2 and Low Error can be said to be a good model. R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data!

Interpret Co-efficients:

How to Interpret the Coefficients for Linear Regression?

The sign of a regression coefficient tells you whether there is a positive or negative correlation between each independent variable & the dependent variable.

A positive coefficient indicates that as the value of the independent variable increases, the value of the dependent variable also tends to increase. A negative coefficient suggests that as the independent variable increases, the dependent variable tends to decrease.

The coefficient value signifies how much the mean of the dependent variable changes given a one-unit shift in the independent variable while holding other variables in the model constant. This property of holding the other variables constant is crucial because it allows you to assess the effect of each variable in isolation from the others.

Interpret P-values

Regression analysis generates an equation to describe the statistical relationship between one or more predictor variables and the response variable.

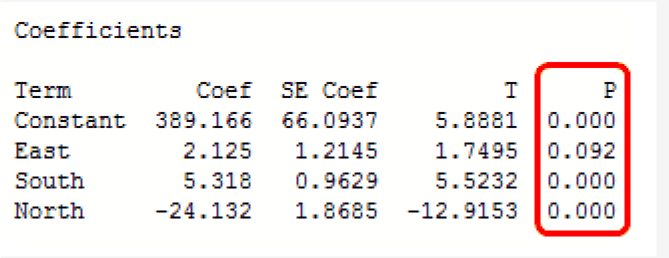
How to Interpret the P-Values in Linear Regression Analysis?

The p-value for each term tests the null hypothesis that the coefficient is equal to zero (no effect). A low p-value (< 0.05) indicates that you can reject the null hypothesis. In other words, a predictor that has a low p-value is likely to be a meaningful addition to your model because changes in the predictor's value are related to changes in the response variable.

Conversely, a larger (insignificant) p-value suggests that changes in the predictor are not associated with changes in the response.

In the output below, we can see that the predictor variables --> South and North are significant because both of their p-values are 0.000. However, the p-value for East (0.092) is greater than the common alpha level of 0.05, which indicates that it is not statistically significant.(or not important)

Typically, you use the coefficient p-values to determine which terms to keep in the regression model. In the model above, we should consider removing East.



```
Coefficients

Term          Coef  SE Coef         T      P
Constant   389.166  66.0937    5.8881  0.000
East         2.125   1.2145    1.7495  0.092
South        5.318   0.9629    5.5232  0.000
North      -24.132   1.8685  -12.9153  0.000
```

Adjusted R2

R^2 shows how well terms (data points) fit a line. **Adjusted R^2** also indicates how well terms fit a curve or line, but adjusts for the number of terms/variables in a model. If you add more and more **useless** variables to a model, adjusted r-squared will decrease. If you add more**useful** variables, adjusted r-squared will increase.
Adjusted R^2 will always be less than or equal to R^2.

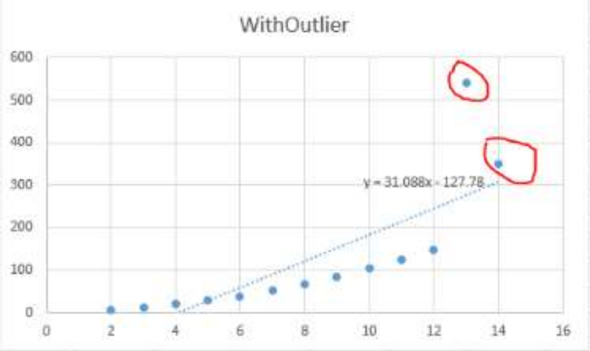
Effect of Missing Values & Outliers on Linear Regression

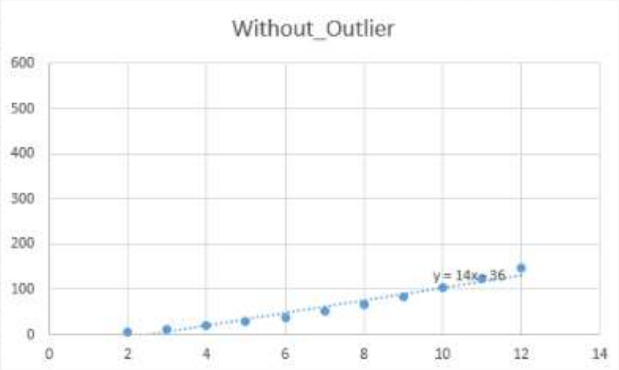
Missing Values:

A linear regression Model will not predict for a row if it has NA (missing values) in any of the predictor variables. Hence, it is important to fill all Missing values before we build a LR Model.

Outliers:

Outliers can have a dramatic impact on linear regression. It can change the model equation completely i.e. bad prediction or estimation. Look at the below scatter plot and linear equation with or without outlier.





Look at the both snapshots, equation parameters changing a lot. More info here

https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/

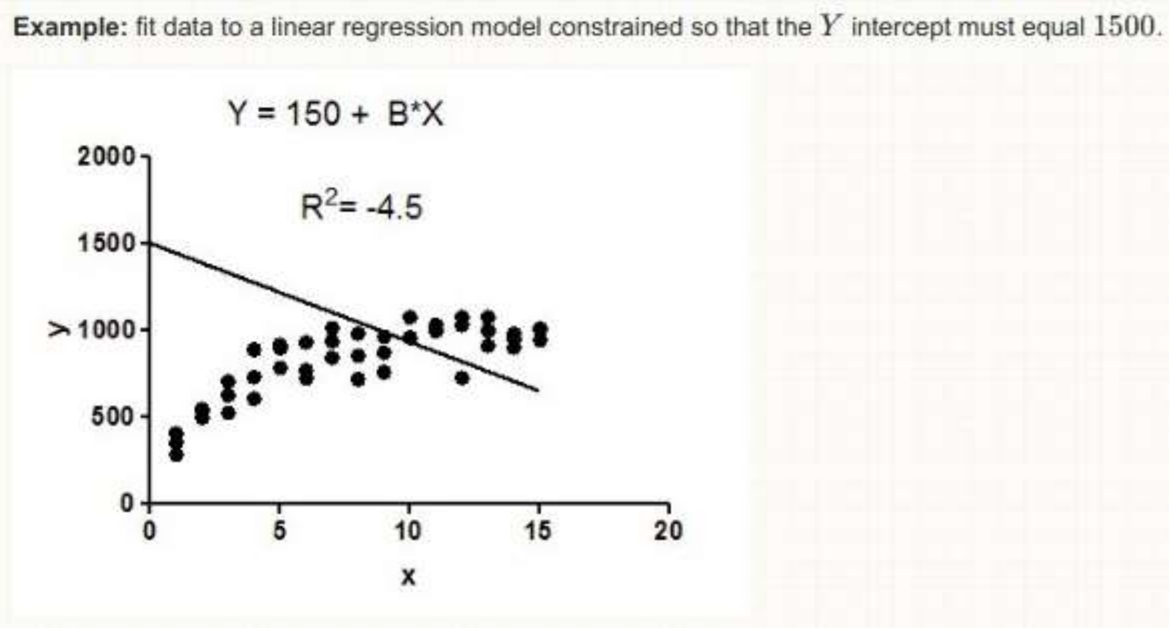
For the purpose of fitting a regression that reliably predicts the future data, identifying outliers is only useful in smaller data-sets. For regressions involving many records, it is unlikely that one observation will carry sufficient weight to cause extreme influence on the fitted equation.

For the purpose of anomaly detection, though, identifying influential observations can be very useful.

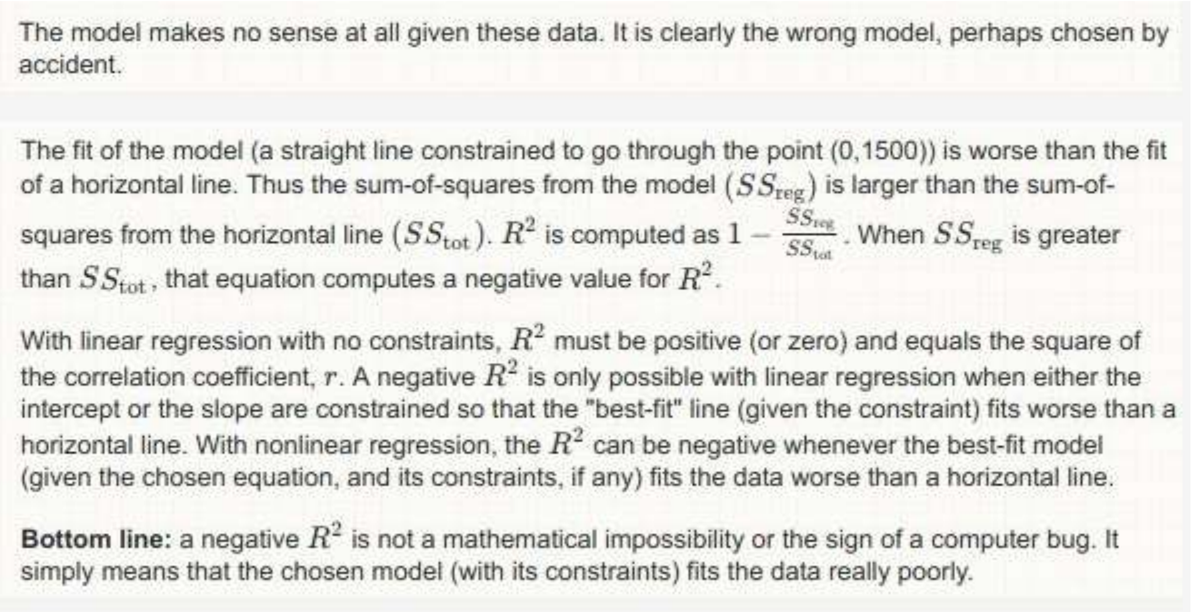
Read more about outliers at this amazing blog:

https://www.analyticsvidhya.com/blog/2016/01/guide-data-exploration/?utm_source=outlierdetectionpyod&utm_medium=blog

Negative R2 Is it possible?



Example: fit data to a linear regression model constrained so that the Y intercept must equal 1500.



The model makes no sense at all given these data. It is clearly the wrong model, perhaps chosen by accident.

The fit of the model (a straight line constrained to go through the point (0,1500)) is worse than the fit of a horizontal line. Thus the sum-of-squares from the model (SS_{reg}) is larger than the sum-of-squares from the horizontal line (SS_{tot}). R^2 is computed as $1 - \frac{SS_{\text{reg}}}{SS_{\text{tot}}}$. When SS_{reg} is greater than SS_{tot}, that equation computes a negative value for R^2.

With linear regression with no constraints, R^2 must be positive (or zero) and equals the square of the correlation coefficient, r. A negative R^2 is only possible with linear regression when either the intercept or the slope are constrained so that the "best-fit" line (given the constraint) fits worse than a horizontal line. With nonlinear regression, the R^2 can be negative whenever the best-fit model (given the chosen equation, and its constraints, if any) fits the data worse than a horizontal line.

Bottom line: a negative R^2 is not a mathematical impossibility or the sign of a computer bug. It simply means that the chosen model (with its constraints) fits the data really poorly.

Assumptions of Linear Regression

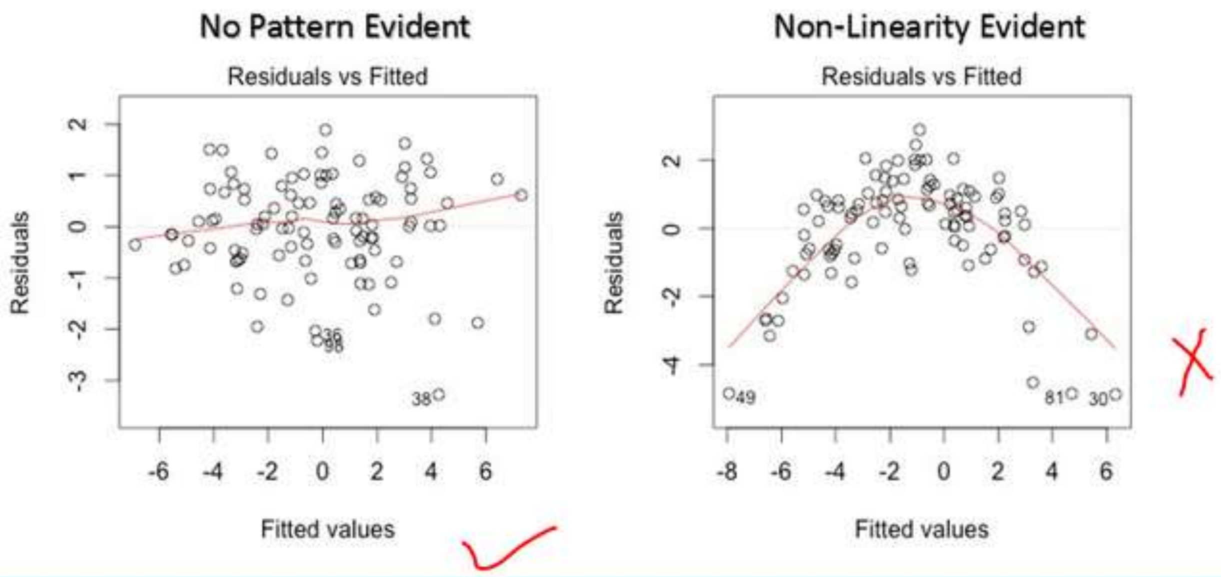
Assumptions of Linear Regression

Regression is a parametric approach. ‘Parametric’ means it makes assumptions about data for the purpose of analysis. Due to its parametric side, regression is restrictive in nature. It 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.

Assumption 1

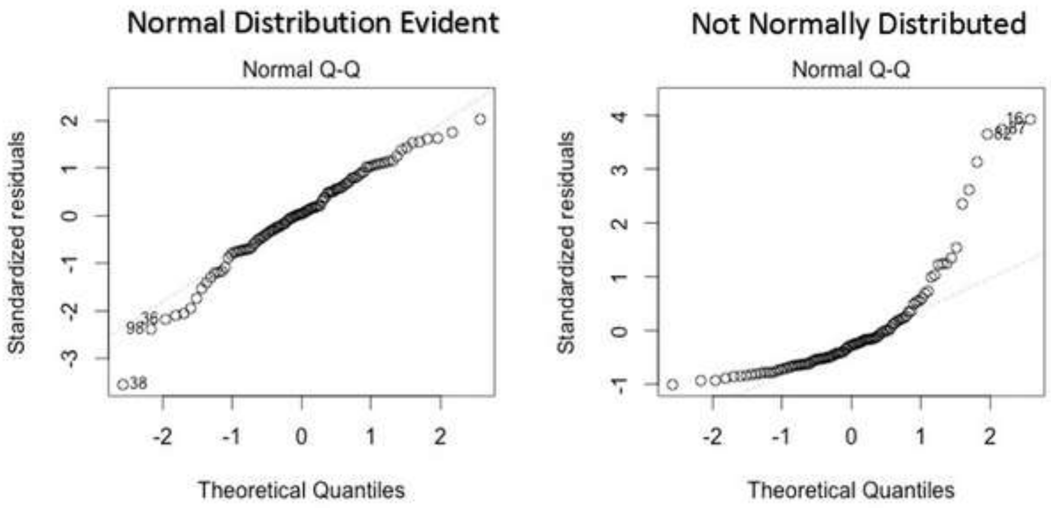
There should be a linear relationship between dependent (response) variable and independent (predictor) variable(s). A linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of X.

Look for residual vs fitted value plots. If there exist any pattern (the red line if approximately horizontal shows no 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.



Assumption 2

Normality of residuals: We draw a histogram of the residuals, and then examine the normality of the residuals. If the residuals are not skewed, that means that the assumption is satisfied. This assumption can best be checked with a histogram or a Q-Q-Plot. *When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.*



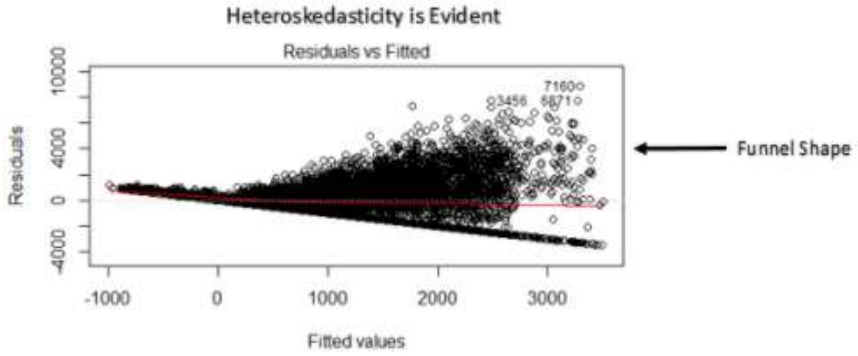
Assumption 3

Linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other. Multicollinearity may be tested using co-elation matrix among all Quantative IV's or by using VIF.

The simplest way to address the problem is to remove one of the IVs which is highly co-related with another. Otherwise we can use PCA.

Assumption 4

The error terms must have constant variance. This phenomenon is known as homoskedasticity. The presence of non-constant variance is referred to 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. *You can look at residual vs fitted values plot. If heteroskedasticity exists, the plot would exhibit a funnel shape pattern*



StepWise Regression (Optional)

Stepwise regression is a semi-automated process of building a model by successively adding or removing variables based solely on the t-statistics/*P value* of their estimated coefficients. It is especially useful for sifting through large numbers of potential independent variables and/or fine-tuning a model by poking variables in or out. Improperly used, it may converge on a poor model while giving you a false sense of security. This includes the following types-

Forward Stepwise Selection:

- Start with a null model. The null model has no predictors, just one intercept (The mean over Y).
- Fit p simple linear regression models, each with one of the variables in and the intercept. So basically, you just search through all the single-variable models the best one (the one that results in the lowest residual sum of squares). You pick and fix this one in the model.
- Now search through the remaining p minus 1 variables and find out which variable should be added to the current model to best improve the residual sum of squares.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Backward Stepwise Elimination:

Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

- Start with all variables in the model.
- Remove the variable with the largest p-value | that is, the variable that is the least statistically significant.
- The new (p - 1)-variable model is

then re-built, and the variable with the largest p-value is removed again.

- Continue until a stopping rule is reached. For instance, we may stop when all remaining variables have a significant p-value de fined by some significance threshold.

Bidirectional Elimination:

Bidirectional elimination, a combination of the above, testing at each step for variables to be included or excluded.

https://en.wikipedia.org/wiki/Stepwise_regression