LINEAR REG

Linear regression is a linear model i.e. a model that assumes a linear relationship between the input variables (x) and the single output variable (y).

When there is a single input variable (x), the method is referred to as simple linear regression. When there are multiple input variables, literature from statistics often refers to the method as multiple linear regression.

Form of Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

y is the response

β values are called the model coefficients. These values are “learned” during the model fitting/training step.

β0 is the intercept

β1 is the coefficient for X1 (the first feature)

βn is the coefficient for Xn (the nth feature)

How linear regression works

The way Linear Regression works is by trying to find the weights (namely, W0 and W1) that lead to the best-fitting line for the input data (i.e. X features) we have. The best-fitting line is determined in terms of lowest cost. In the context of Linear Regression, training is basically finding those weights and plugging them into the straight line function so that we have best-fit line (with W0, W1 minimizing the cost). Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

Important Terms

Bias

It is the intercept where our line intercepts the y-axis.

It is an additional parameter which is used to adjust the output along with the weighted sum of the inputs.

It is a constant which helps the model in a way that it can fit best for the given data.

Model Coefficients/Parameters

When training a linear regression model, we are trying to find out the coefficients for the linear function that best describe the input variables.

How Do We Estimate the Coefficients?

For that task there's a mathematical algorithm called Gradient Descent.

Start with some values of the coefficients/parameters, e.g. $\beta 0=0$, $\beta 1=0$. Keep changing B0 and B1 to reduce the J(B0, B1) until we hopefully end up at a minimum.

Predictor Relevance

Now that you have coefficients, how can you tell if they are relevant to predict your target? The best way is to find the p-value. The p-value is used to quantify statistical significance as it allows to tell whether the null hypothesis is to be rejected or not.

For any modelling task, the hypothesis is that there is some correlation between the features and the target. The null hypothesis is therefore the opposite i.e. there is no correlation between the features and the target. So, finding the p-value for each coefficient will tell if the variable is statistically significant to predict the target.

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. As a general rule of thumb, if the p-value is less than 0.05 then there is a strong relationship between the variable and the target. 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 p-value suggests there is no relationship between the variable and the target. It also means that changes in the predictor are not associated with changes in the response.

Residuals

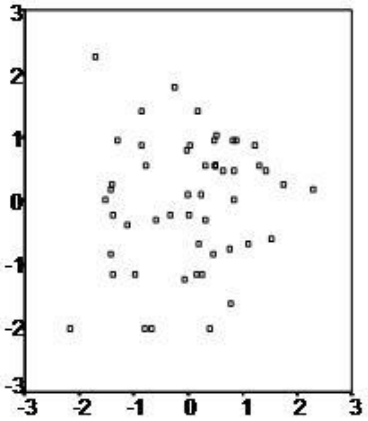
The cost function defines a cost based on the distance between true target and predicted target (shown in the graph as lines between sample points and the regression line), also known as the residual. If the particular line is far from all the points, the residuals will be higher, and so will the cost function. If a line is close to the points, the residuals will be small, and hence the cost function. Further using Gradient Descent, the process of finding the best model out of the many possible models is called optimization.

Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables.

Assumptions of Linear Regression

1. There should be a linear Relationship between the features and target:

According to this assumption there is linear relationship between the features and target. Linear regression captures only linear relationship. 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. This can be validated by plotting scatter plots between each feature and the target and by correlation matrix.



The above example depicts a case where no and little linearity is present.

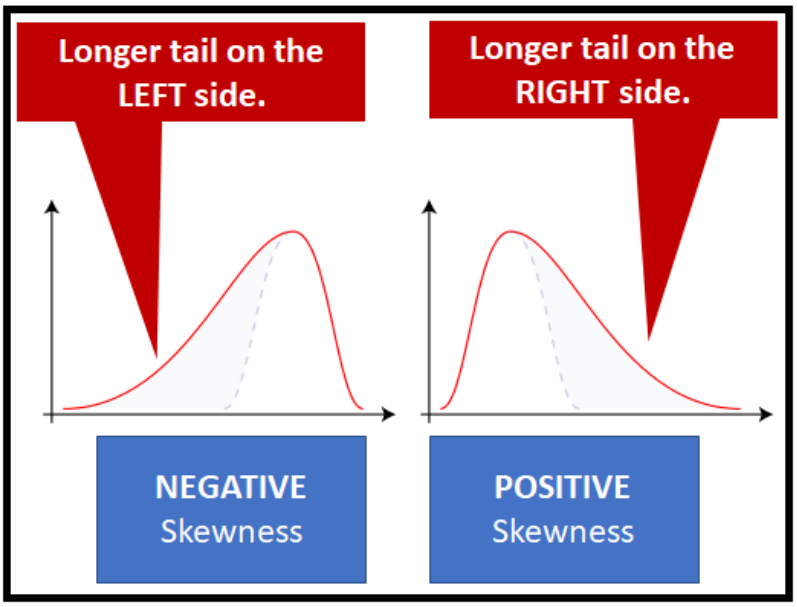
2. Target variable should follow a normal distribution

The next assumption is that the variables follow a normal distribution i.e. it must not be skewed. In order words, we want to make sure that for each x value, y is a random variable following a normal distribution and its mean lies on the regression line.

A histogram or scatter plot can be used to check normalization.

Skewness is another way to check normality of any variable. Skewness of the normal distribution is zero.

$$\textbf{Skewness} = \frac{\mathbf{3}\,(\boldsymbol{Mean} - \boldsymbol{Median})}{\boldsymbol{Standard\ Deviation}}$$



For a positive skewness mean > median > mode. To reduce positive/right skewness, take roots or logarithms or reciprocals (roots are weakest).

For a negative skewness mean < median < mode. To reduce negative/left skewness, take squares or cubes or higher powers.

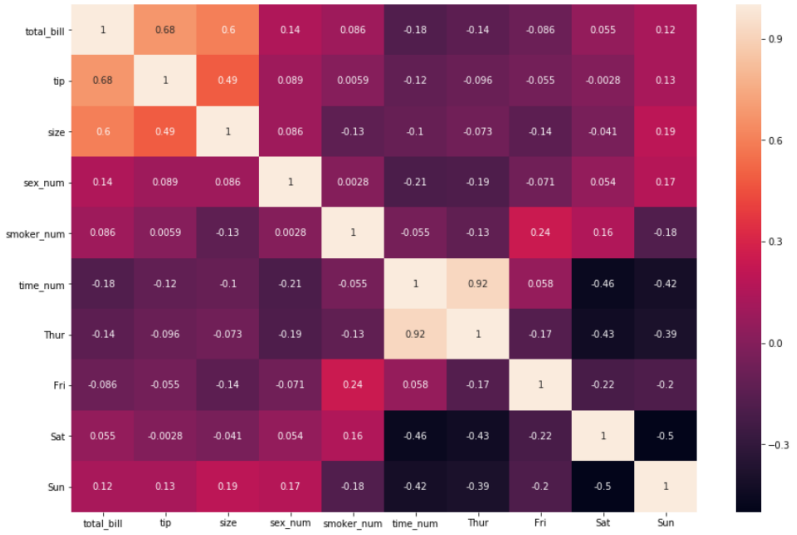
If the skewness is between -0.5 to +0.5 then we can say data is fairly symmetrical. If the skewness is between -1 to -0.5 or 0.5 to 1 then data is moderately skewed. And if the skewness is less than -1 and greater than +1 then our data is heavily skewed.

3.There should be little or no multicollinearity between the features

Multicollinearity is a state of very high inter-correlations or inter-associations among the independent variables. It is therefore a type of disturbance in the data if present weakens the statistical power of the regression model.

For Numeric columns, you can also use VIF factor. VIF value = 1 suggests no collinearity whereas a value of >1 and <5 implies moderately collinearity whereas a value of >10 implies highly correlated. A VIF between 5 and 10 indicates high correlation that may be problematic. And if the VIF goes above 10, you can assume that the regression coefficients are poorly estimated due to multicollinearity.

Heat maps (correlation matrix) can be used for identifying highly correlated features.



For Categorical columns, you can use Chi square test. Chi-square test is used to find the statistical significance between two categorical variables.

Chi square test will give you a list of four variables

The test statistic - chi2

The p-value of the test - p

Degrees of freedom - dof

The expected frequencies, based on the marginal sums of the table – expected

If the P-value is less than the significance level (0.05), we conclude that there is a relationship between the two variables

Why removing highly correlated features is important?

The interpretation of a regression coefficient is that it represents the mean change in the target for each unit change in a feature when you hold all of the other features constant. However, when features are correlated, changes in one feature in turn shifts another feature/features. The stronger the correlation, the more difficult it is to change one feature without changing another. It becomes difficult for the model to estimate the relationship between each feature and the target independently because the features tend to change in unison.

4. The errors must have constant variance

Variance of error terms should be similar across the values of the independent variables

Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables. There should be no clear pattern in the distribution and if there is a specific pattern, the data is heteroscedastic.

A scatter plot of residual values vs predicted values is a good way to check for homoscedasticity. We can plot the residuals versus each of the predicting variables to look for independence assumption. If the residuals are distributed uniformly randomly around the zero x-axes and do not form specific clusters, then the assumption holds true.

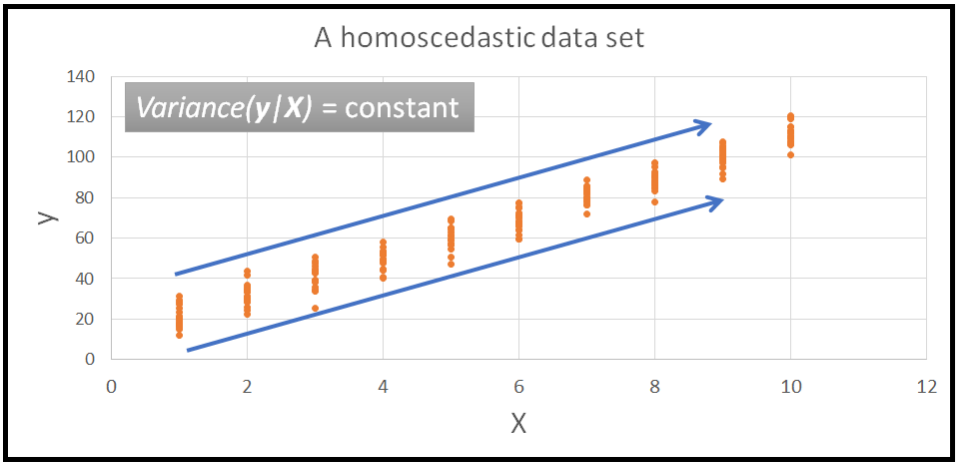
When your data is heteroscedastic:

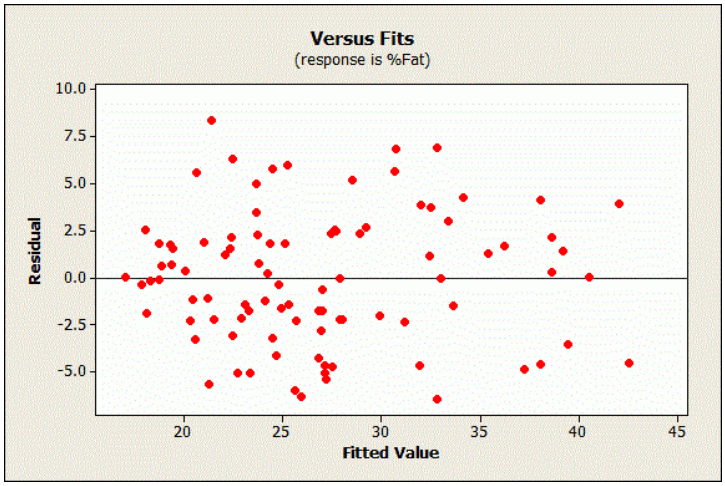
Variance(y|X) = f(X)

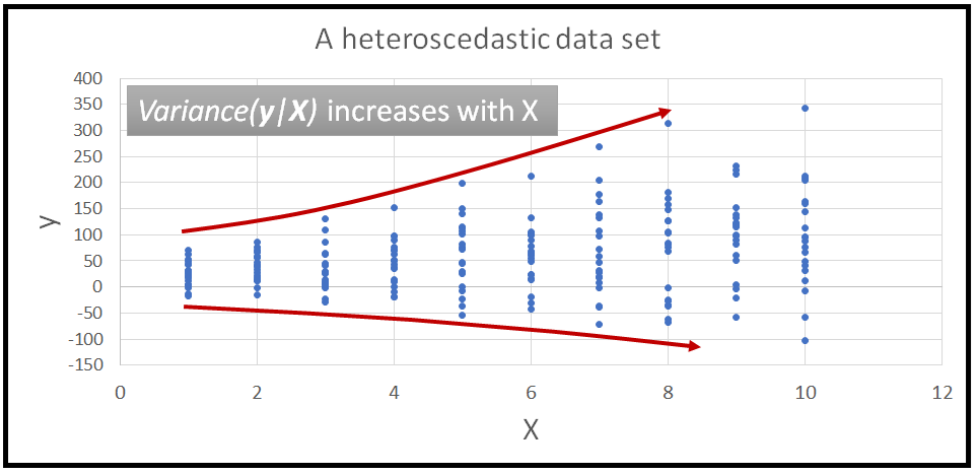
Where f is some function of X.

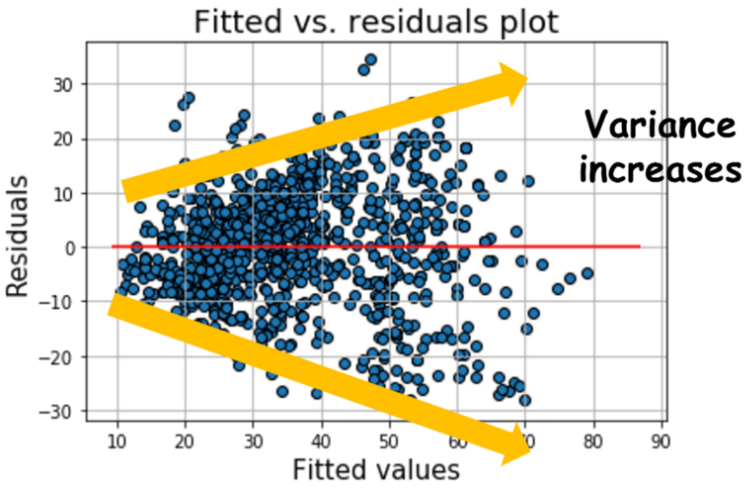
The opposite of heteroscedasticity is homoscedasticity where the variance is constant

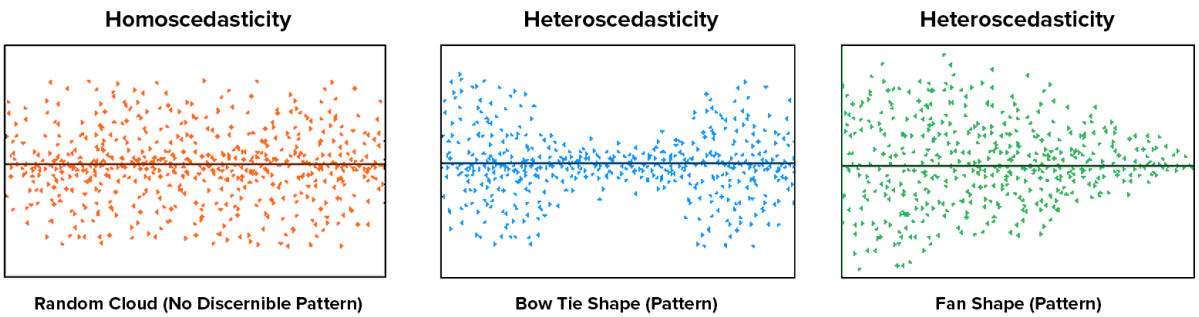
Variance(y|X) = σ^2 (a constant value)




Versus Fits
(response is %Fat)
Residual
Fitted Value
10.0
7.5
5.0
2.5
0.0
-2.5
-5.0
20
25
30
35
40
45


A heteroscedastic data set
Variance(y|X) increases with X
y
X
400
350
300
250
200
150
100
50
0
-50
-100
-150
0
2
4
6
8
10
12


Fitted vs. residuals plot
Variance
increases
Residuals
Fitted values
30
20
10
0
−10
−20
−30
10
20
30
40
50
80
90



What Causes Heteroscedasticity?

Heteroscedasticity occurs more often in datasets that have a large range between the largest and smallest observed values. While there are numerous reasons why heteroscedasticity can exist, a common explanation is that the error variance changes proportionally with a factor. This factor might be a variable in the model.

A simple bivariate example can help to illustrate heteroscedasticity: Imagine we have data on family income and spending on luxury items. Using bivariate regression, we use family income to predict luxury spending. As expected, there is a strong, positive association between income and spending. Upon examining the residuals, we detect a problem as the residuals are very small for low values of family income (almost all families with low incomes don't spend much on luxury items) while there is great variation in the size of the residuals for wealthier families (some families spend a great deal on luxury items while some are more moderate in their luxury spending). This situation represents heteroscedasticity because the size of the error varies across values of the independent variable. Examining a scatterplot of the residuals against the predicted values of the dependent variable would show a classic cone-shaped pattern of heteroscedasticity

How to fix the problem?

-Transform the dependent variable using one of the variance stabilizing transformations such as logarithmic or square transformation. Log-transform the y variable to 'dampen down' some of the heteroscedasticity, then build an OLSR model for log(y)

- Use a Weighted Least Squares (WLS) or a Generalized Least Squares (GLS) model as these two models that do not assume a homoscedastic variance.

5. There should be little or no autocorrelation in the residuals

Autocorrelation occurs when the residual errors are dependent on each other. 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. You can test the linear regression model for autocorrelation with the Durbin-Watson test. Durbin-Watson's d tests the null hypothesis that the residuals are not linearly auto-correlated.

Some notes on the Durbin-Watson test:

the test statistic always has a value between 0 and 4

value of 2 means that there is no autocorrelation in the sample

values < 2 indicate positive autocorrelation

values > 2 indicate negative autocorrelation

The solution is fitting a time series/forecasting model. Another solution could be to try a model with better parameters and better variables

6. There should be normality of errors i.e. the residuals should follow a normal distribution.

A histogram or distribution plot or probability plot can be used to check normalization.

Skewness is another way to check normality of any variable.

One of the solution to overcome this issue is to try and build a model with better parameters and better variables.

PROBLEM WITH PARAMETRIC MODELS

Parametric models are those which assume the form of the data i.e. make certain assumptions about the data.

Non parametric models don't make such assumptions. They are also called free models.

Parametric models like linear regression, logistics regression have high bias and low variance problems. The reason being non parametric models make a lot of assumptions which ultimately make models biased. Non parametric models like decision tree, random forests have low bias and high variance problems.