**Linear Regression**

**Assumptions of Linear Regression-** Regression is a parametric approach so it makes assumptions about the data to analyze it.

1. Linearity
2. Independence
3. Homoscedasticity
4. Normality
5. No multicolinearity
6. No endogeneity
7. Relationship between independent and dependent variables is linear
8. The observations are independent of each other.
9. The variance of the errors is constant across all levels of the independent variables.
10. The errors follow a normal distribution.
11. The independent variables are not highly correlated with each other.
12. There is no relationship between the errors and the independent variables.

**Important assumptions in regression analysis-**

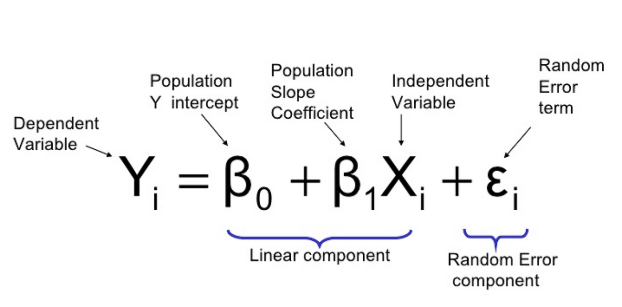
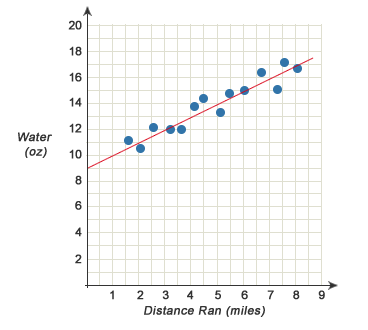
1. There should be a linear and additive relationship between the dependent variable or response (target) variable and independent variable (predictive).
2. The linear relationship suggests that a change in response Y due to one unit change in X is constant, regardless of the value of Y.
3. An additive relationship suggests that the effect of X on Y is independent of other variables.
4. There should not be any correlation between the error terms. Absence of this phenomenon is known as auto correlation.
5. The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.
6. The error terms must have constant variance. The phenomenon is known as homoscedasticity. The presence of non constant variance is refereed to hetroskedasticity.
7. The error terms must be normally distributed.

**Linear Regression-** Linear regression is a type of statistical analysis used to predict the relationship between two variables. It assumes a linear relationship between the independent variable and dependent variable and aims to find the best fitting line that describes the relationship. The line is determined by minimizing the sum of the squared difference between the predicted and actual values.

Simple linear regression- There is one independent variable and one dependent variable. The model estimates the slope and the intercept of the line of the best fit, which represents the relationship between the variables.

The slope represents the change in the dependent variable for each unit change in the independent variable, while the intercept represents the predicted value of the dependent variable when the independent variable is zero.

Linear regression is a supervised learning algorithm in ML that supports finding the linear correlation among variables. The result or output of the regression problem is a real or continuous value.

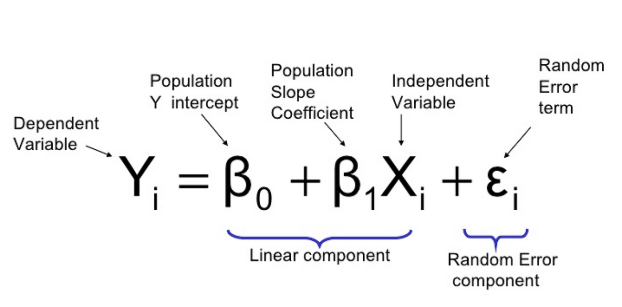


To find out the best fit line, we need to find out the values of and. Also, the best fit line is the one that has the least error value between the predicted and actual values as minimum.

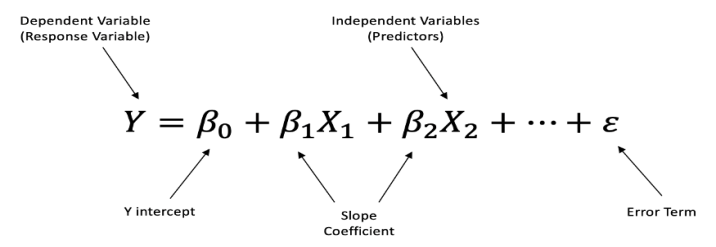
The bets fir line is the one that fits the given scatter plot in the best way. The best fit line is obtained by minimizing the residual sum of the squares.

**Multi linear regression-**

The MLR assumption is the same as SLR: it assumes that data can be represented using a linear form. The only difference in MLR is that there is just more predictors to consider.



Simple Linear Regression only has one predictor, X, a slope intercept, and an error term.



Multiple Linear Regression includes multiple predictors (X1, …, Xp), a slope intercept and an error term.

**Evaluation of linear regression-**

1. R Square/Adjusted R Square

2. Mean Square Error (MSE)/Root Mean Square Error (RMSE)

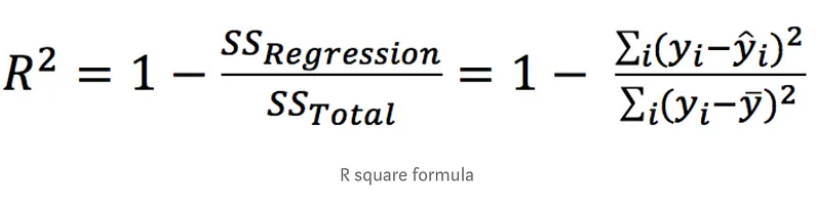
3. Mean Absolute Error (MAE)

4. Illustrate Residual of model as a normal distribution (bell shape)

5. By OLS from state models formula

**R Square/Adjusted R Square**

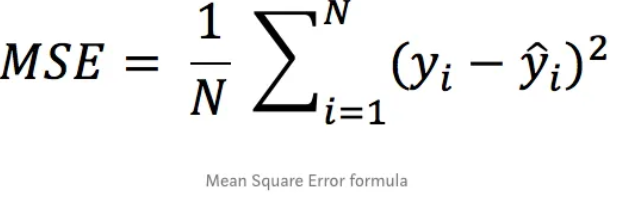
This is a first measure of regression model especially we, everybody, do during evaluation because it is easy to interpret score between 0 to 1. If we see good score like close to 1, then we assume that model is good fit. Of course , R Square is a good measure to determine how well the model fits the dependent variables. However, it does not take into consideration of overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data but performs badly for testing data.So I recommend that we have to see all perspective for better evaluation . let’s talk what is actually mean R² . R² is calculated by the sum of squared of prediction error divided by the total sum of square which replace the calculated prediction with mean. R Square value is between 0 to 1 and bigger value indicates a better fit between prediction and actual value.



When we do fine tuning to model to get better accuracy then R² Adjust help us to better understand. It happen when we add more independent features or penalize more feature due to over fitting . Then we can see different score between on these measures.

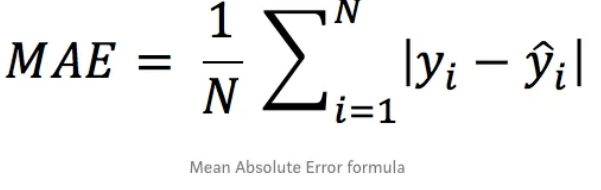
**Mean Square Error (MSE)/Root Mean Square Error(RMSE):**

While R² is a relative measure of how well the model fit dependent variables, whereas Mean Square Error is an absolute measure of the fit of model. MSE is calculated by sum of square of prediction error. Where prediction error is minus between true values and prediction values, and then it is made by square because we avoid negative error score. It’s result gives us how much deviation from actual number. It’s number might be larger number which may be like uncommon . You might be question how is error score is too big .



**Mean Absolute Error (MAE):**

This is almost same to Mean Square Error metric but only MAE take absolute error value instead of square of predicted error for avoiding negative score . However, here , we don’t need to calculate Root of MAE score . We can interpret directly the score with real values.



**Overfitting and underfitting-**

Overfitting and underfitting are common issues in machine learning, including linear regression, and refer to the model's ability to generalize to unseen data. Let's explore these concepts in the context of linear regression.

1. Underfitting:

Underfitting occurs when a model is too simple to capture the underlying patterns in the data. In linear regression, this might mean using a linear model when the relationship between the features and the target is more complex. The model's performance will be poor not only on the training data but also on new, unseen data. Signs of underfitting include a high training error and high validation/testing error.

2. Overfitting:

Overfitting happens when a model is too complex and learns the noise in the training data instead of the true underlying pattern. In linear regression, this could involve using a high-degree polynomial to fit the data, resulting in a model that fits the training data extremely well but fails to generalize to new, unseen data. Signs of overfitting include a very low training error but a significantly higher validation/testing error.

**Here are some strategies to address these issues in linear regression:**

Underfitting Solutions:

Increase Model Complexity: Use a more complex model, like adding higher-degree polynomial features or including interaction terms.

Feature Engineering: Create new features that provide additional information for the model to learn from.

Overfitting Solutions:

Regularization: Use techniques like Lasso (L1 regularization) or Ridge (L2 regularization) to penalize overly complex models and prevent overfitting.

Cross-validation: Use techniques like k-fold cross-validation to evaluate the model's performance on different subsets of the data, helping to identify overfitting.

It's crucial to strike a balance between model complexity and generalization. Techniques like cross-validation, regularization, and monitoring performance metrics (e.g., mean squared error, R-squared) on both the training and validation datasets can help in achieving an optimal model that generalizes well to unseen data.

**Effect of missing values-** Missing data creates various problems. The absence of the data reduces the statistical power which refers to the probability that the test will reject the null hypothesis when it is false. The lost data can cause bias in the estimation of parameters. It can reduce the representativeness of the samples.