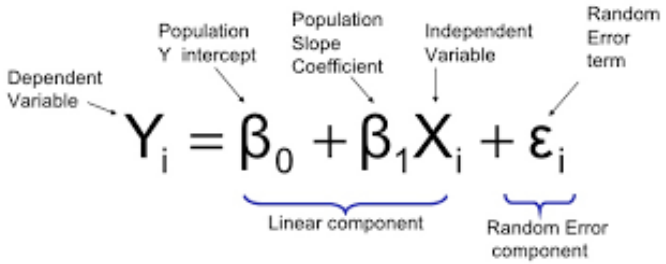
Linear Regression

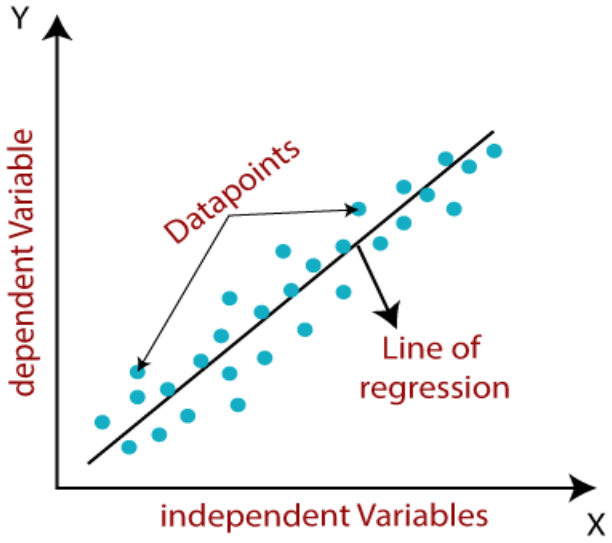
What is Regression?

- Regression searches for relationships among variables.
- For example, you can observe several employees of some company and try to understand how their salaries depend on their features, such as experience, education level, role, city of employment, and so on.
- In Regression analysis you need to find a **function that maps some features or variables to others** sufficiently well.
- The dependent features are called the **dependent variables, outputs**, or **responses**. The independent features are called the **independent variables**, **inputs, regressors**, or **predictors**.
- Regression problems usually have **one continuous and unbounded dependent variable**. The **inputs, however, can be continuous, discrete, or even categorical data** such as gender, nationality, or brand.
- It's a common practice to denote the outputs with $\boldsymbol{y}$ and the inputs with $\boldsymbol{x}$. If there are two or more independent variables, then they can be represented as the vector $\mathbf{x} = (x_1, \ldots, x_r)$, where r is the number of inputs.

Linear Regression

- Mathematically, we can represent a linear regression as:





Types of Linear Regression

Linear regression can be further divided into two types of the algorithm:

- **Simple Linear Regression:**
  If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.

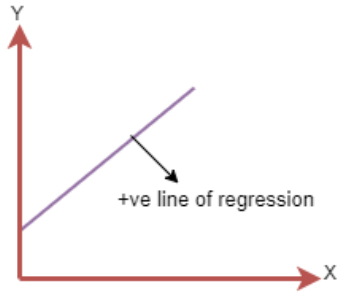
- **Multiple Linear Regression:**
  If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Linear Regression line

A linear line showing the relationship between the dependent and independent variables is called a regression line. A regression line can show two types of relationship:

Positive Linear Relationship:

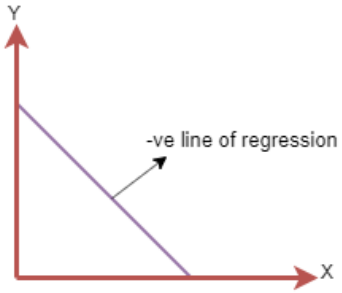
If the dependent variable increases on the Y-axis and independent variable increases on X-axis, then such a relationship is termed as a Positive linear relationship.



The line equation will be: $Y= a_0+a_1x$

Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and independent variable increases on the X-axis, then such a relationship is called a negative linear relationship.



The line of equation will be: $Y= -a_0+a_1x$

Linear Regression line

Linear Regression (LR) means simply **finding the best fitting line** that **explains the variability between the dependent** and **independent features** very well.

It describes the linear relationship between independent and dependent features, and in linear regression, the algorithm **predicts the continuous features**(e.g. Salary, Price), rather than deal with the categorical features (e.g. cat, dog).

Simple Linear Regression

Simple Linear Regression uses the slope-intercept (weight-bias) form, where our model needs to find the optimal value for both slope and intercept.

So with the optimal values, the model can find the variability between the independent and dependent features and produce accurate results.

In simple linear regression, the **model takes a single independent** and dependent variable.

$$y = b_o + b_1 x$$

Here, y and x are the dependent variables, and independent variables respectively. b1 is slope and b0 is y-intercept respectively.

b1 tells, for one unit of increase in x, How many units does it increase in y.

b0 means, What is the value of y when the x is zero.

How the Model will Select the Best Fit Line?

- First, our model will try a bunch of different straight lines from that it finds the optimal line that predicts our data points well.
- For finding the best fit line our model uses the cost function.
- In machine learning, every algorithm has a cost function, and in simple linear regression, the goal of our algorithm is to find a minimal value for the cost function.
- In linear regression (LR), we have many cost functions, but mostly used cost function is MSE(Mean Squared Error). It is also known as a Least Squared Method.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

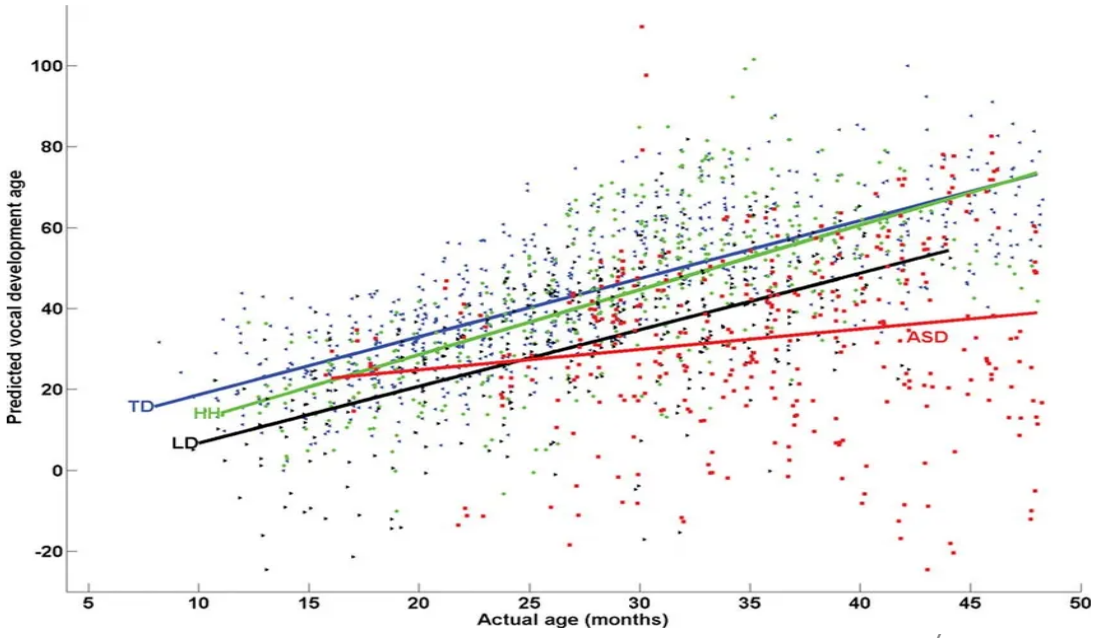
Yi − Actual value,
$\hat{Y}_i$ − Predicted value,
n − number of records.

Loss Function

It is a calculation of loss for single training data.

Cost Function

It is a calculation of average loss over the entire dataset.



How the Model will Select the Best Fit Line?

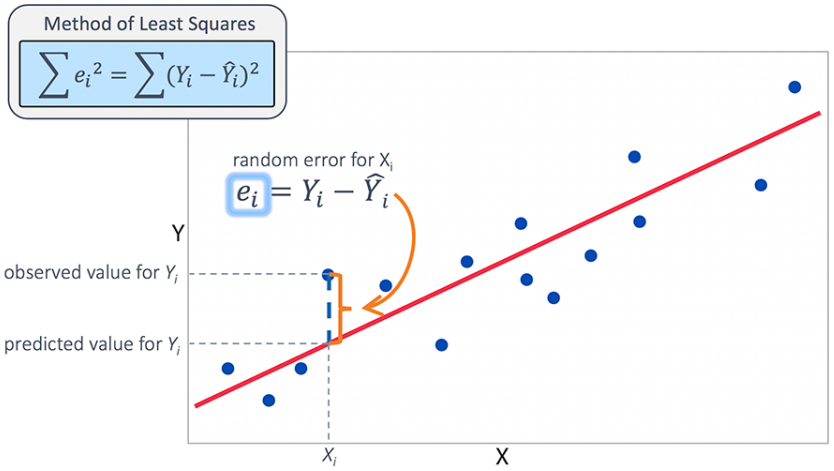
- In the below diagram blue data points are representing the actual values from training data, a red line(vector) is the predicted value for that actual blue data point.
- We can notice a random error **(actual value-predicted value),**model is trying to **minimize the error between the actual and predicted value** which is also called as **residuals.**
- We need a model, which makes the prediction very well. So our model will find the loss between all the actual and predicted values respectively and it selects the line which has an average error of all points lower.

Steps

1) Our model will fit all possible lines and find an overall average error between the actual and predicted values for each line respectively.
2) Selects the line which has the lowest overall error and that will be the best fit line.
3) This method of finding the best line is called as **method of least square**.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x};$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$



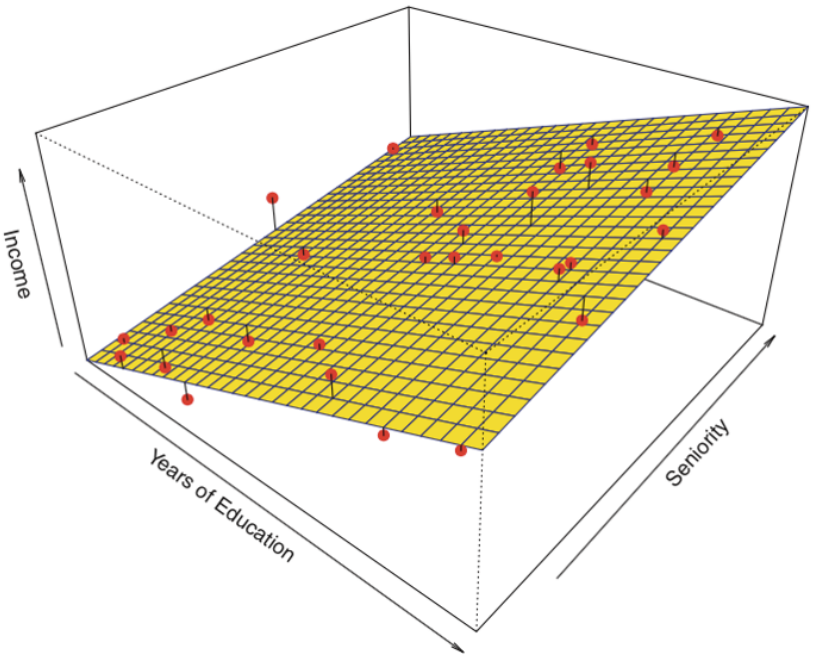
Multiple linear Regression

- In multiple linear regression instead of having a single independent variable, the model has multiple independent variables to predict the dependent variable.

$$y = b_o + b_1x_1 + b_2x_2 + b_3x_3 \ldots . + b_nx_n$$

where bo is the y-intercept, b1,b2,b3,b4…,bn are slopes of the independent variables x1,x2,x3,x4…,xn and y is the dependent variable.

- Here instead of finding a line, our model will find the best plane in in n-Dimension, our model will find the best hyperplane.
- In multiple linear regression the residuals are computed as difference between the actual value and the hyperplane **(Predicted value).**



Evaluation metrics in Linear Regression

- The metrics used for regression are different from the classification metrics.
- It means we cannot use the metrics which we used in classification to evaluate a regression model; instead, the performance of a Regression model is reported as errors in the prediction.

Following are the popular metrics that are used to evaluate the performance of Regression models.

- Mean Absolute Error
- Mean Squared Error
- R2 Score
- Adjusted R2

Evaluation metrics in Linear Regression

Mean Absolute Error (MAE)

- Mean Absolute Error or MAE is one of the simplest metrics, which measures the absolute difference between actual and predicted values, where absolute means taking a number as Positive.
- In order to find the absolute error for the complete dataset, we need to find the mean absolute of the complete dataset.
- The formula for MAE is

$$\text{MAE} = \frac{\sum|Y - Y'|}{N}$$

Here,
Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points.

Mean Squared Error

Mean Squared error or MSE is one of the most suitable metrics for Regression evaluation. It measures the average of the Squared difference between predicted values and the actual value given by the model.
Due to squared differences, it penalizes small errors also, and hence it leads to over-estimation of how bad the model is.
The formula for calculating MSE is given below:

$$\text{MSE} = \frac{\sum(Y - Y')^2}{N}$$

Here,
Y is the Actual outcome, Y' is the predicted outcome, and N is the total number of data points.

Evaluation metrics in Linear Regression

R Squared Score

- R squared error is also known as **Coefficient of Determination**, which is another popular metric used for Regression model evaluation. The R-squared metric enables us to compare our model with a constant baseline to determine the performance of the model.
- Formula for R-Squared is

$$R^2 = \frac{SSR}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

- In linear regression :

Total sum of square (SST) = Regression sum of square (SSR) + Error Sum of square (SSE)

So **R-squared is ratio of Regression sum of square to Total sum of square** which tells us about **how much of total variance is explained by the linear regression model**.

- Example: If R-squared for our model is 0.84 then we will say that **about 84% of the total variation of our dataset is explained by our linear regression model.**

Evaluation metrics in Linear Regression

Adjusted R Squared

- Adjusted R squared, as the name suggests, is the improved version of R squared error. R square has a limitation of improvement of a score on increasing the terms, even though the model is not improving, and it may mislead the data scientists.
- To overcome the issue of R square, adjusted R squared is used, which will always show a lower value than R^2. It is because it adjusts the values of increasing predictors and only shows improvement if there is a real improvement.
- We can calculate the adjusted R squared as follows:

$$R_a^2 = 1-\left[\left(\frac{n-1}{n-k-1}\right)\times\left(1-R^2\right)\right]$$

Here,
n is the number of observations
k denotes the number of independent variables
and Ra2 denotes the adjusted R2

Assumptions of Linear Regression

There are few assumptions that must be fulfilled before jumping into the regression analysis. Some of those are very critical for model's evaluation.

- Normality
- Multicollinearity
- Homoscedasticity
- Autocorrelation

Normality

This assumption states that the residuals from the model is normally distributed.

After determining the model parameters, it is good to check the distribution of the residuals.

Apart from the visual of the distribution, one should check the **Q-Q plot** for better understanding of the distribution.

Assumptions of Linear Regression

Multicollinearity

Multicollinearity is observed when two or more independent variables are correlated to one another.

If that is the case, the model's estimation of the coefficients will be systematically wrong. One can check Variance Inflation Factor (VIF) to determine the variables which are highly correlated and potentially drop those variables from the model.

$$VIF = \frac{1}{1 - R^2}$$

If the variables have high correlation, VIF value shoots up. Typically **VIF value >5 indicates the presence of multicollinearity.**

How VIF is Calculated?

To calculate the VIFs, all independent variables become a dependent variable iteratively.

Each model produces an R-squared value indicating the percentage of the variance in the individual IV that the set of IVs explains.

Consequently, higher R-squared values indicate higher degrees of multicollinearity. VIF calculations use these R-squared values.

- X1 ⇐ X2, X3, X4
- X2 ⇐ X1, X3, X4
- X3 ⇐ X1, X2, X4
- X4 ⇐ X1, X2, X3

Assumptions of Linear Regression

Remedies for Multicollinearity

If Multicollinearity is present in the data then our linear regression model is not a valid model and we cannot rely on the results which are produced from the model.

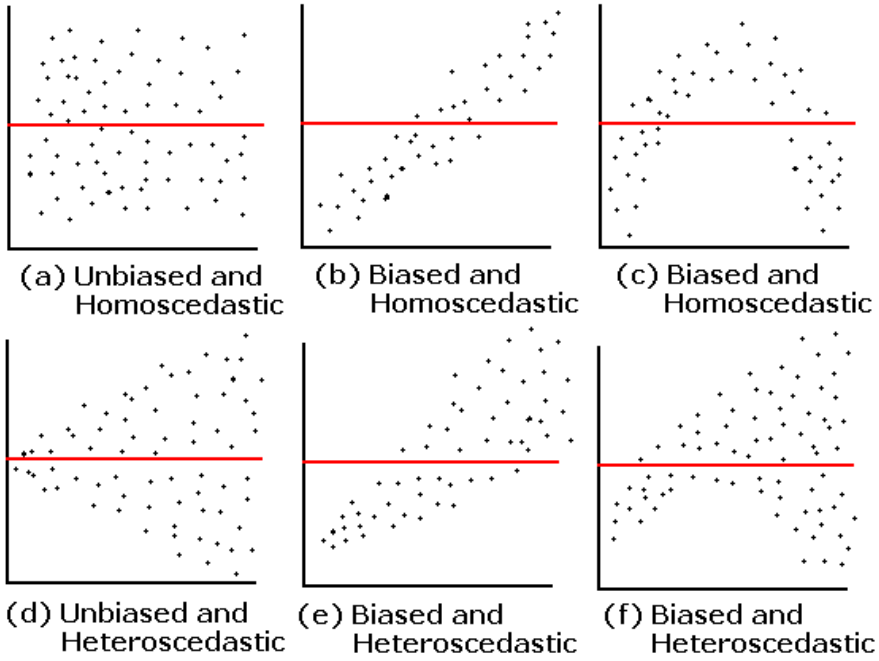
The following methods can help us to deal with multicollinearity :

1) Remove the variable with very high VIF value.
2) Apply Ridge regression (L2 Regularization)
3) Transform features using Principal component analysis (PCA).

Assumptions of Linear Regression

Homoscedasticity

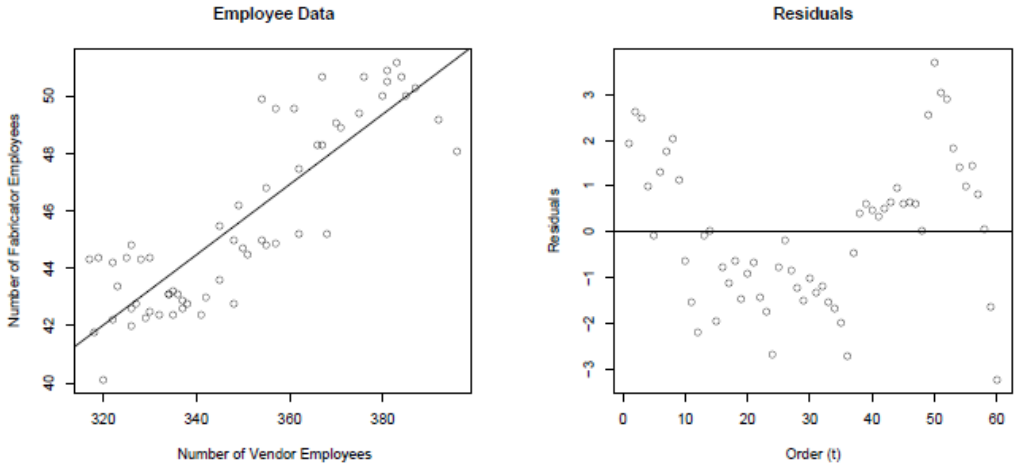
- It refers to a condition that the **variance of the error(residuals)** of the model **should be constant** for all the independent features.
- When we **plot the errors against independent features it should be constant over time**, if not then it is considered as a Heteroscedasticity. Where the errors are not constant over time and form a funnel shape.



Assumptions of Linear Regression

Autocorrelation

- Autocorrelation occurs when the residuals are not independent from each other.
- In other words when the value of y(x+1) is not independent from the value of y(x).



(a) (b)

- **Durbin-Watson's d** tests the null hypothesis that the residuals are not linearly auto-correlated. While d can assume values between **0 and 4**, values around 2 indicate no autocorrelation. As a rule of thumb **values of 1.5 < d < 2.5** show that **there is no auto-correlation** in the data.