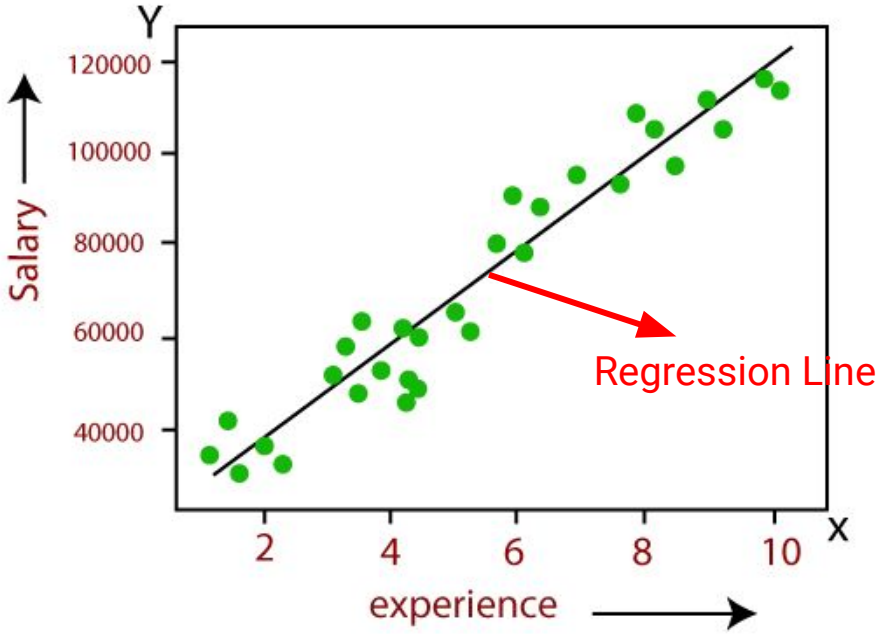
Regression

Dipon Talukder
Assessment Developer (Data and AI), Workera.ai
Research Assistant, Universal Machine.io
Adjunct Faculty, East Delta University

Supervised Learning



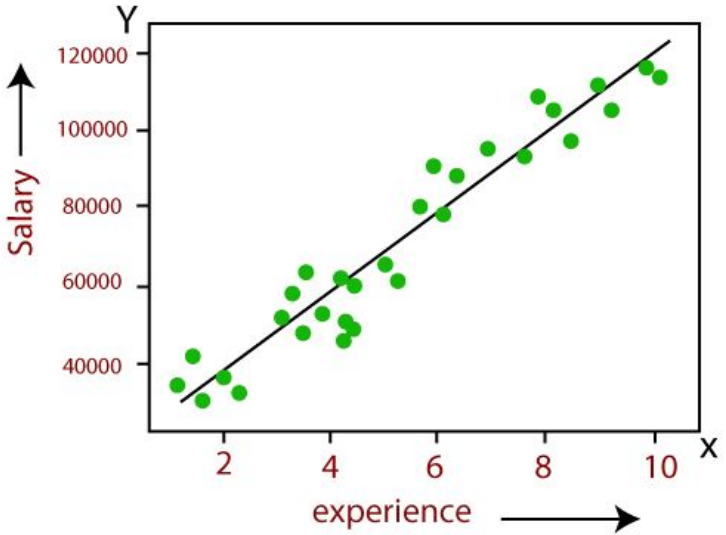
- Predicting numbers
- This particular type of problem is called Regression.
- Algorithms: Linear regression, Polynomial regression

What is Regression???

- Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables.
- More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.
- It predicts continuous/real values such as temperature, age, salary, price, etc.

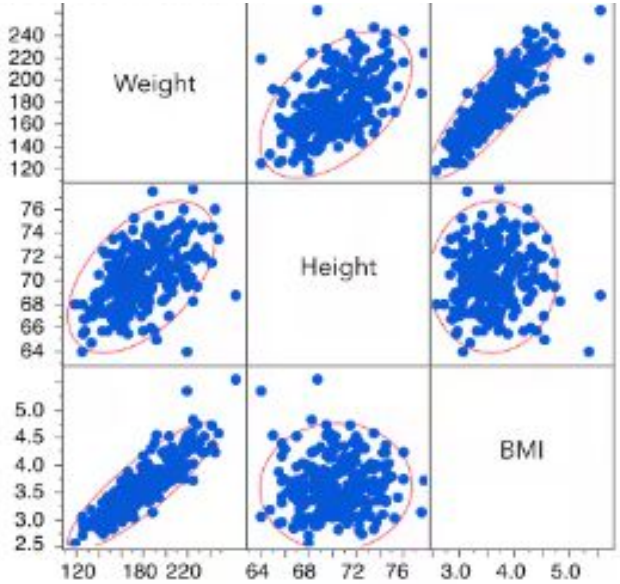
Regression Terminologies

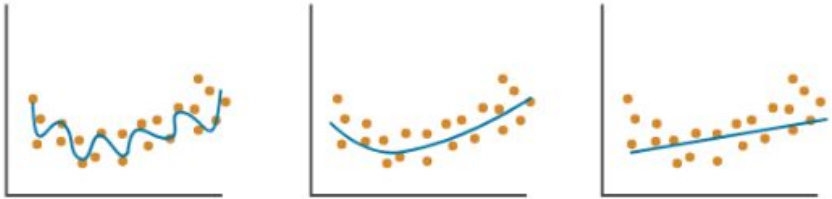
- **Dependent Variable:** The main factor in Regression analysis which we want to predict or understand is called the dependent variable. It is also called target variable.
- **Independent Variable:** The factors which affect the dependent variables or which are used to predict the values of the dependent variables are called independent variable, also called as a predictor.
- **Outliers:** Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided.



Regression Terminologies

- **Multicollinearity:** If the independent variables are highly correlated with each other than other variables, then such condition is called Multicollinearity. It should not be present in the dataset, because it creates problem while ranking the most affecting variable.
- **Underfitting and Overfitting:** If our algorithm works well with the training dataset but not well with test dataset, then such problem is called Overfitting. And if our algorithm does not perform well even with training dataset, then such problem is called underfitting.

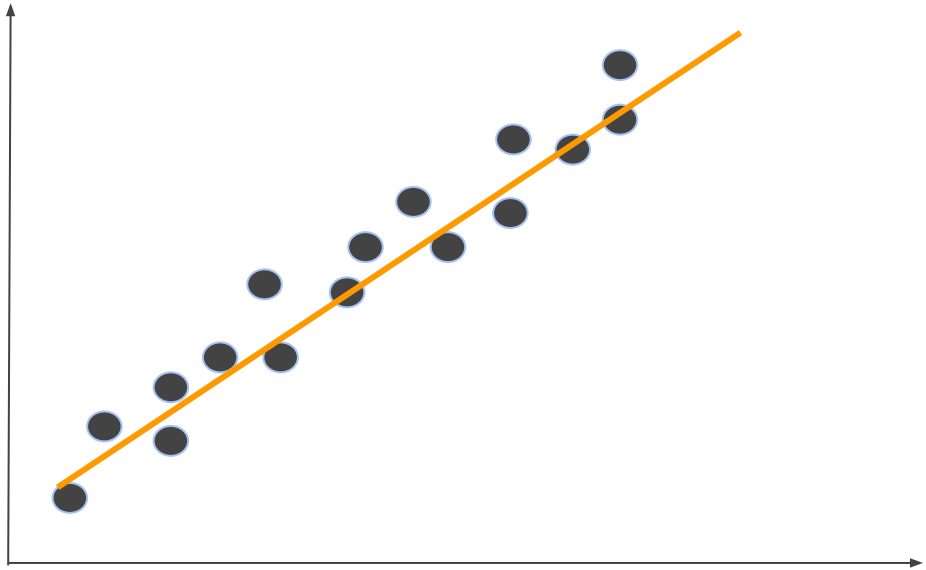




Regression Algorithms

- Linear Regression
- Logistic Regression
- Polynomial Regression
- Support Vector Regression
- Decision Tree Regression
- Random Forest Regression

Linear Regression



$$\hat{Y} = \theta_1 + \theta_2 X$$

Target:

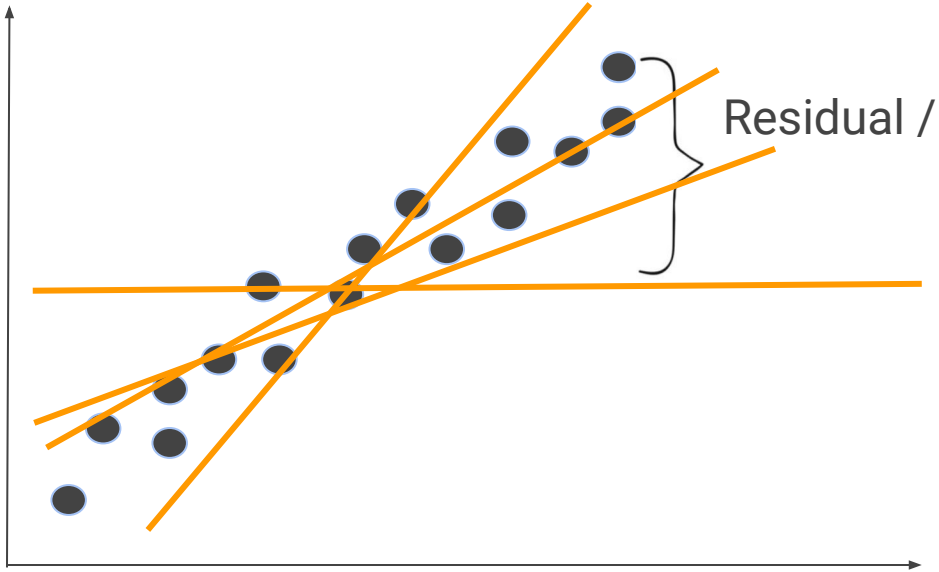
Finding the best fit line (values of Θ_1 and Θ_2) such that

$$MSE = 1/N \sum (Y - Y')^2$$

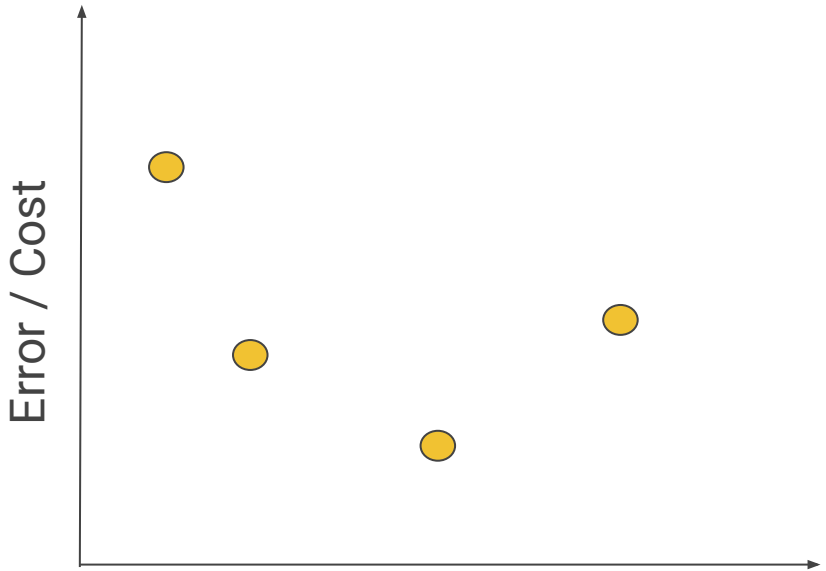
is the lowest.

But HOW?????

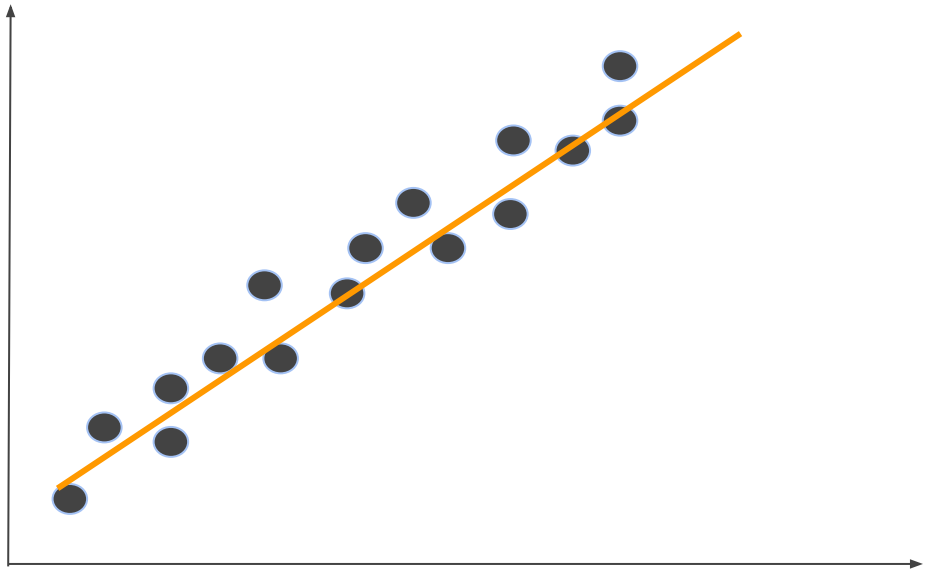
Linear Regression (Manual Approach)



$$\hat{Y} = \theta_1 + \theta_2 X$$



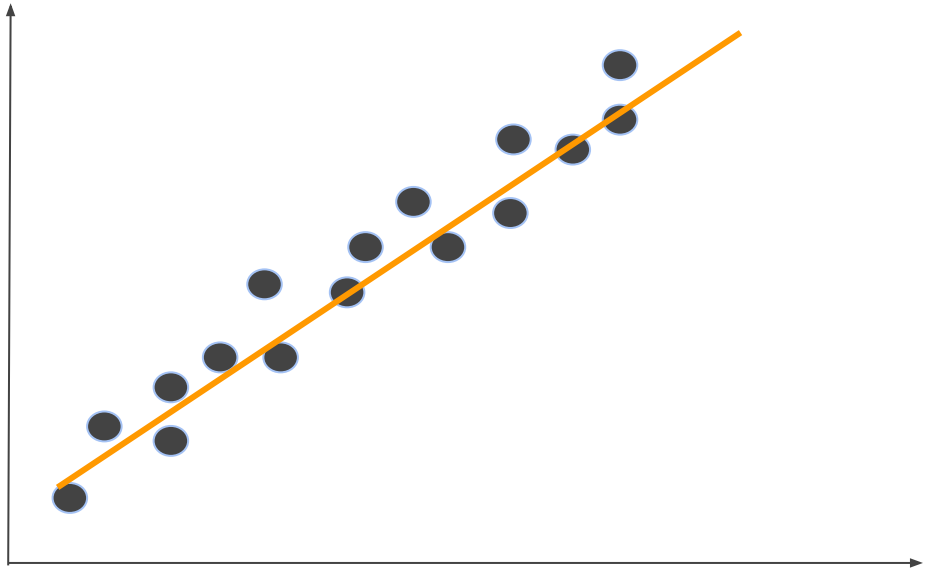
Linear Regression



$$\hat{Y} = \theta_1 + \theta_2 X$$

- Same Same but different approach like previous
- Several ways to find the values of Θ_1 and Θ_2 .
- Popular ones are:
  - ✅ Ordinary Least Squares (OLS)
  - Gradient Descent
  - Least Absolute Deviations

Linear Regression (Ordinary Least Squares (OLS))



$$\hat{Y} = \theta_1 + \theta_2 X$$

- $\theta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- $\theta_2 = \bar{y} - \theta_1 \bar{x}$

NO NEED TO REMEMBER THESE FORMULAS!

Polynomial Regression

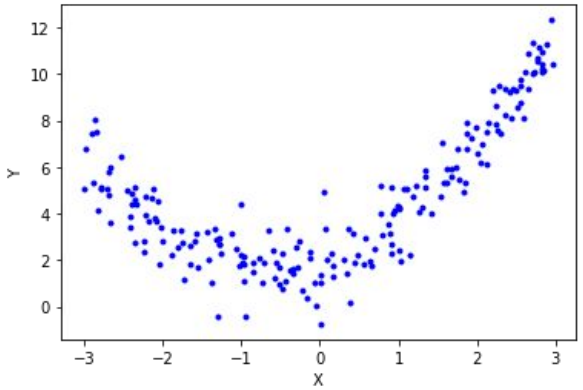
Let’s play a little with equations here:

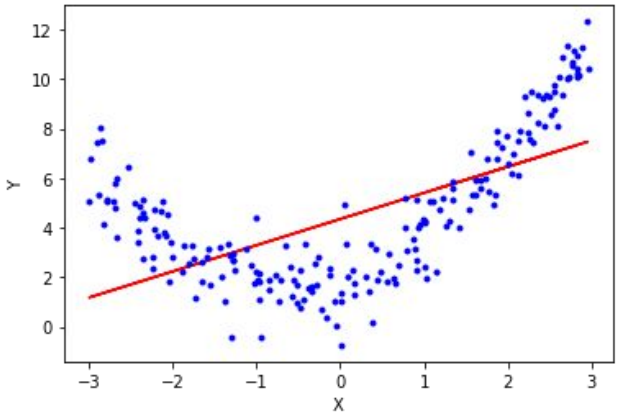
https://www.desmos.com/calculator

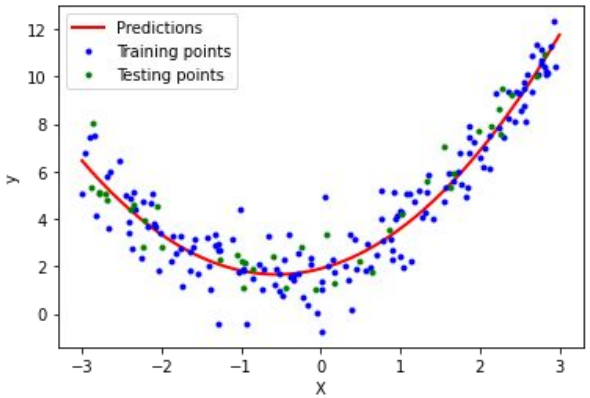
Try: ax^2+bx+c and others

Polynomial Regression

- Helps to build a model over non-linear data

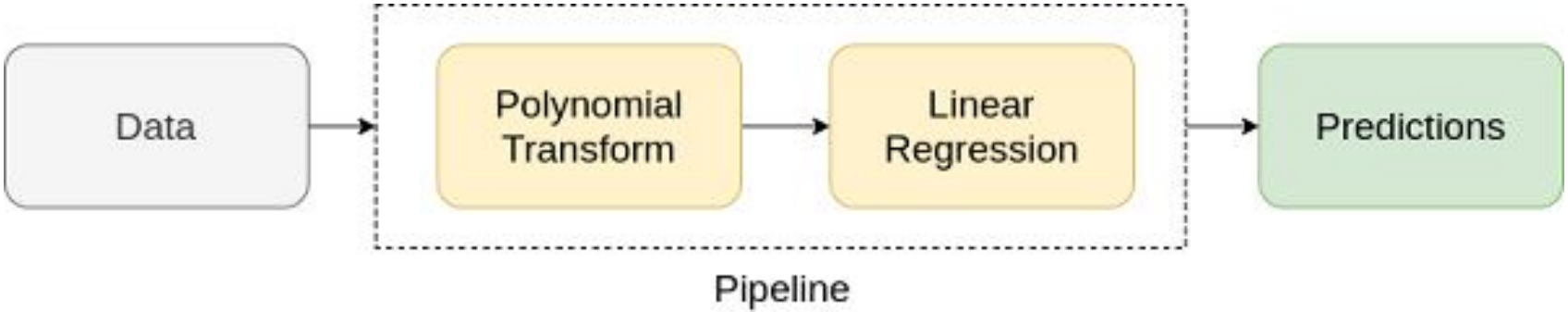






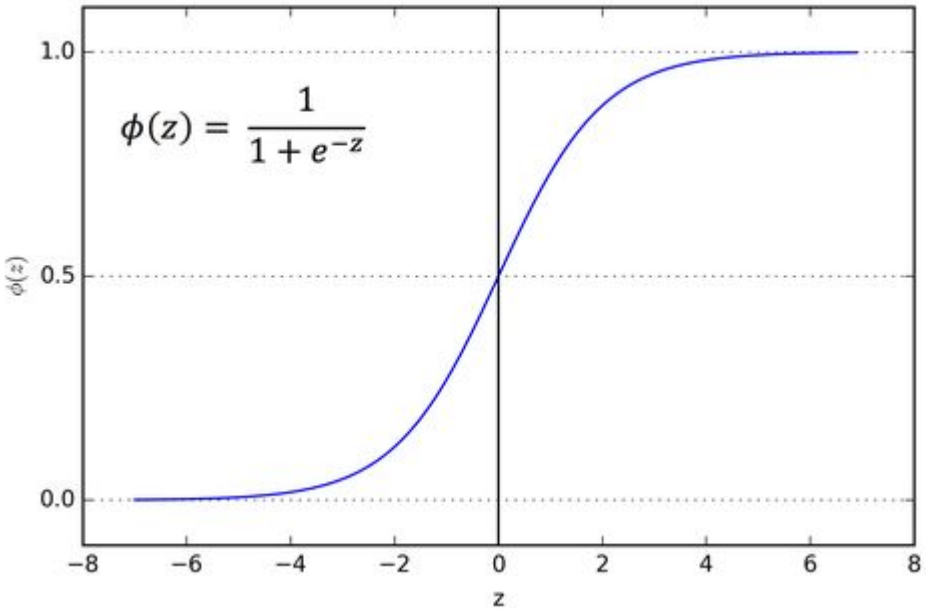
Polynomial Regression

- Polynomial regression is just adding an extra term of product. It is a preprocessing!



Logistic Regression

- Used for classification tasks
- Leverages the properties of the sigmoid function.
- Input of the sigmoid function is the output of logistic regression ($z = \Theta_1 x + \Theta_2$).



R2 Score Metric