x	y
1	1
2	4
3	9
4	16

$y = x^2$

↳ $y = \underset{\beta_1}{\underset{\uparrow}{1}} \times x^2 + \underset{\beta_0}{\underset{\uparrow}{0}}$

Regression Analysis

prediction of descrete dependent variable

find out causal relationship

What is regression analysis

- Linear regression is a basic and commonly used type of predictive analysis

→ future / prediction / forecast

- The dictionary meaning of the word Regression is ‘Stepping back’ or ‘Going back’
- Set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables

→ creating a formula → model ←

- It attempts to establish the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting
- The overall idea of regression is to examine two things

independent = predictor

  - does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables

$$y = f(x)$$

dependent

independent / predictor

$$y = mx + c \rightarrow y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_0$$

constant

intercept

beta estimators (coefficients)

Correlation vs Regression

direction

strength

No estimation

- Correlation is a statistical measure which determines co-relationship or association of two variables while Regression describes how an independent variable is numerically related to the dependent variable

↳ estimation of relationship → formula

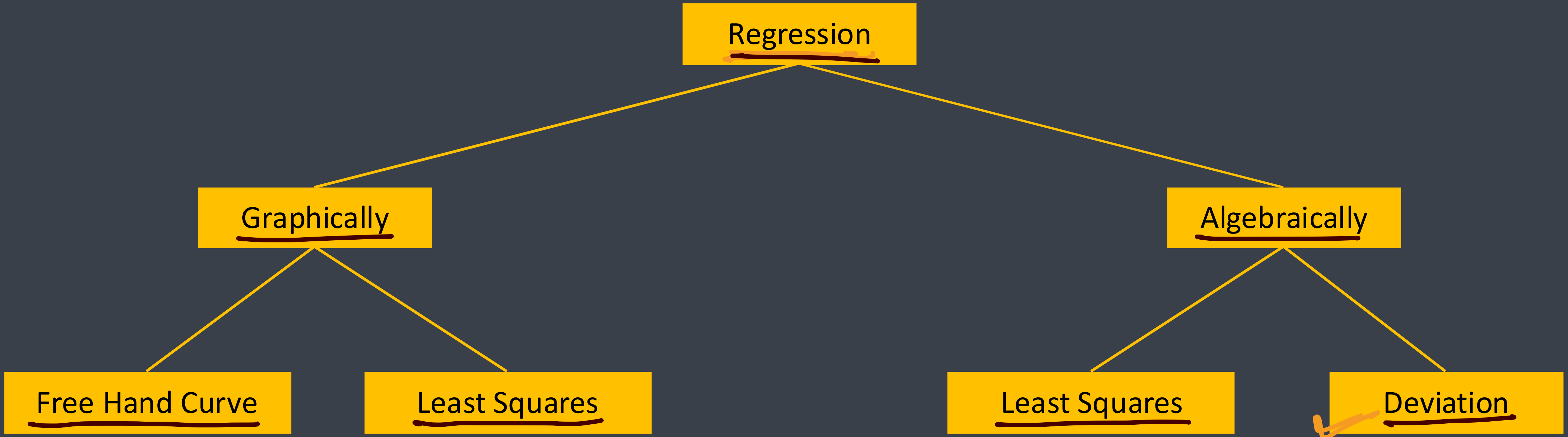
- Correlation is used to represent linear relationship between two variables while regression is used to fit a best line and estimate one variable on the basis of another variable.
- Cov(x, y) = Cov(y, x) , $cor(x,y) = cor(y,x)$
- but reg(x,y) != reg(y, x)

Applications of regression analysis

↪ formula / model

- It helps in the formulation and determination of functional relationship between two or more variables
- It helps in establishing a cause and effect relationship between two variables in economics and business research
- It helps in predicting and estimating the value of dependent variable as price production sales etc
- It helps to measure the variability or spread of values of a dependent variable with respect to the regression line *
- In the field of business regression is widely used by businessmen in
  - Predicting future production
  - Investment analysis
  - Forecasting on sales etc.

Methods of studying regression



Regression

Graphically

Algebraically

Free Hand Curve

Least Squares

OLS method

Least Squares

OLS method

Deviation

Types of regression

- ✓ Linear Regression
- ✓ Polynomial Regression
- ✓ Logistic Regression
- ✓ Ridge Regression
- ✓ Lasso Regression
- ✓ Elastic Net Regression
- ✓ Support Vector Regression
- Quantile Regression
- Principle Component Regression
- Partial Least Square Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Cox Regression

Least Square Method

- A form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable
- The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied
- It aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model
- It begins with a set of data points to be plotted on an x- and y-axis graph
- An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

Regression Equation

x y

Adv	Sales
100	500
90	400
80	450
95	510
150	??

observed

Sales depends on advertisement
Y depends on X [Y on X]

$$\underline{(Y - \bar{Y})} = \underline{b_{yx}}\,\underline{(X - \bar{X})}$$

$\bar{y}$ = mean of y

$\bar{x}$ = mean of x

$$b_{yx} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - (\sum X)^2}$$

$$b_{yx} = r\frac{\sigma y}{\sigma x}$$

$$b_{yx} = \frac{cov(x,y)}{(\sigma x)^2}$$

$x = 150$

$$y - \bar{y} = b_{yx}(150 - \bar{x})$$

$\bar{x} = 91.25$

$\bar{y} = 465$

$$y - 465 = 3.31 \times (150 - 91.25)$$

$$y = 3.31 \times 58.75 + 465$$

$$y = 659.46$$

x	y	$x \times y$	x^2
100	500	50000	10000
90	400	36000	8100
80	450	36000	6400
95	510	48450	9025
365	1860	170450	33525
Σx	Σx	Σxy	Σx^2

$$(\Sigma x)^2 = \underline{133225}$$

$$b_{yx} = \frac{n\Sigma xy - \Sigma x \, \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{4 \times 170450 - 365 \times 1860}{4 \times 33525 - 133225}$$

$$= \frac{681800 - 678900}{134100 - 133225}$$

$$b_{yx} = \frac{2900}{875} = \underline{\underline{3.31}}$$

Regression Equation

X	Y	xy	x^2
3	11	33	9
4	12	48	16
8	9	72	64
7	3	21	49
2	5	10	4
24	40	184	142

What likely to be the value of Y if X = 10

$\bar{x} = 4.8$

$\bar{y} = 8$

$$y - \bar{y} = byx\,(x - \bar{x}), \quad byx = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2}$$

$$byx = \frac{5\times184 - 24\times40}{5\times142 - 576} = \frac{920-960}{710-576} = \frac{-40}{134} = -0.29$$

$$y - 8 = -0.29\times(10-4.8) = -0.29\times5.2$$

$$y = 8 - 1.50 \quad = 6.5$$

$$\boxed{\text{for } x = 10,\ y = 6.5}$$

straight line → $y = mx + c$

Linear Regression

Simple

1 dependent + 1 independent

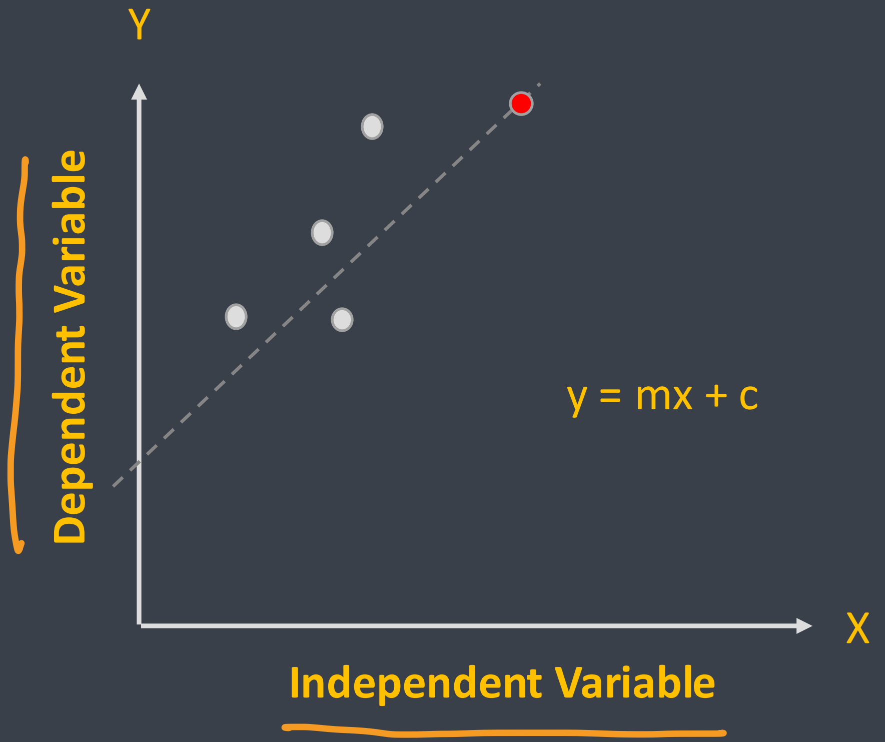
multiple

1 dependent + 2 or more independent

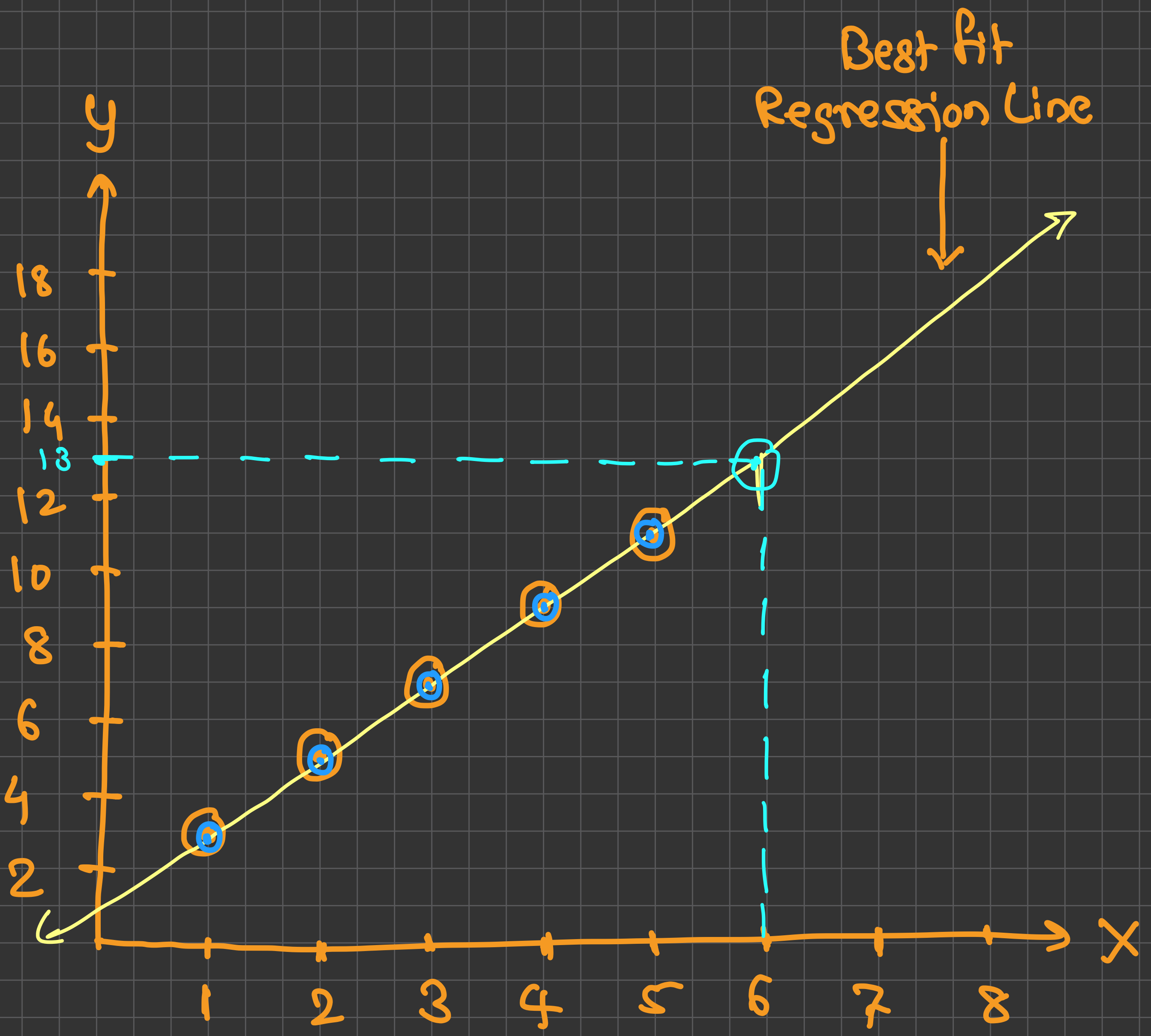
Overview

- The data in Linear Regression is modelled using a straight line
- It is used with continuous variable
- It gives a future value as an output → prediction
- To calculate accuracy following methods are used
  - R-squared
  - Adjusted R-squared

- MAE
- MSE
- RMSE



x	y
1	3
2	5
3	7
4	9
5	11
6	? 13



OLS

x	y	$\hat{y}$	Error $y-\hat{y}$	$(y-\hat{y})^2$
1	3	3	0	0
2	4	5	-1	1
3	2	6	-4	16
4	11	8	3	9
5	10	9	1	1
				$\boxed{27}$ SSE

observed data

true data

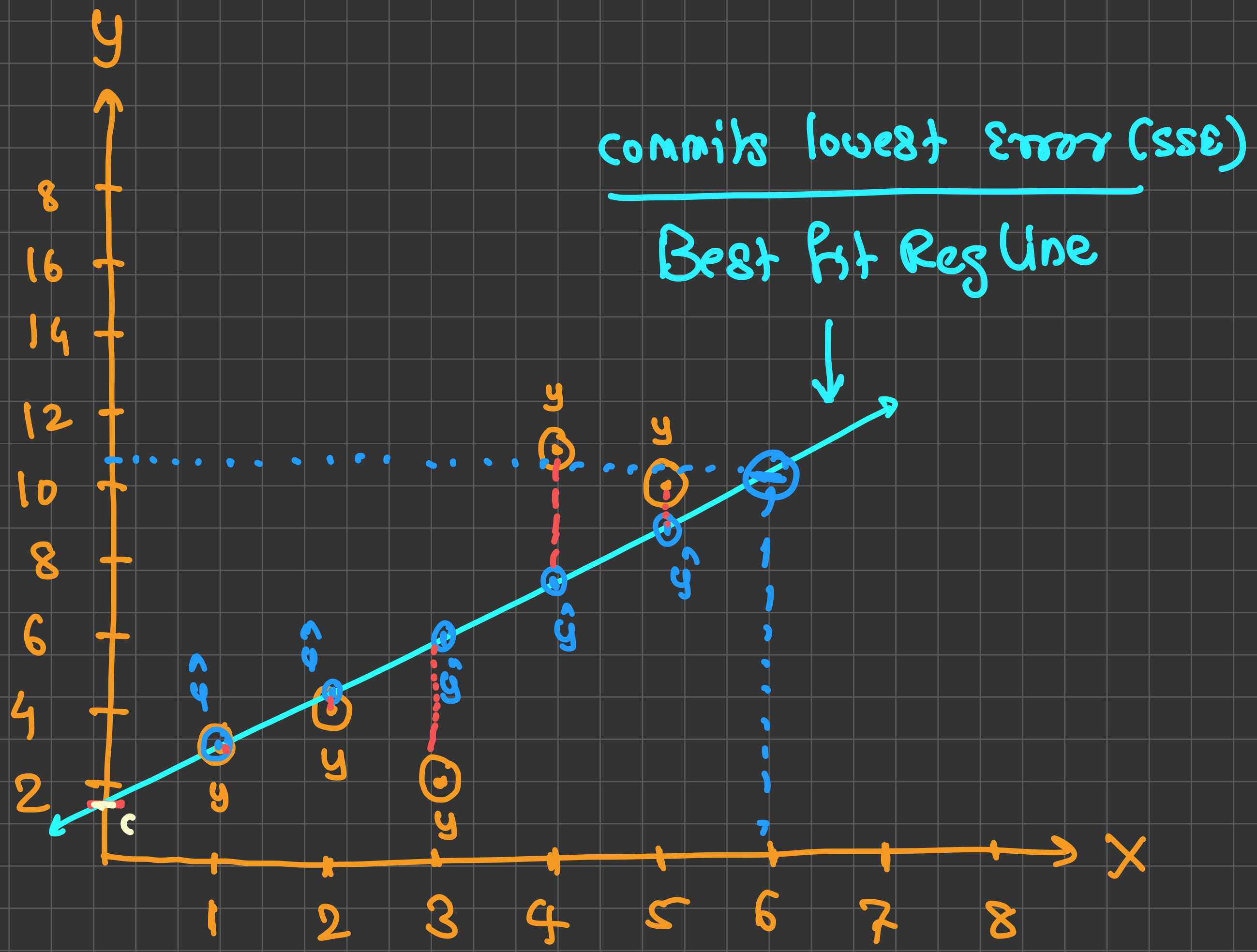
y = observed value

$\hat{y}$ = Expected / calculated

Error = $y - \hat{y}$

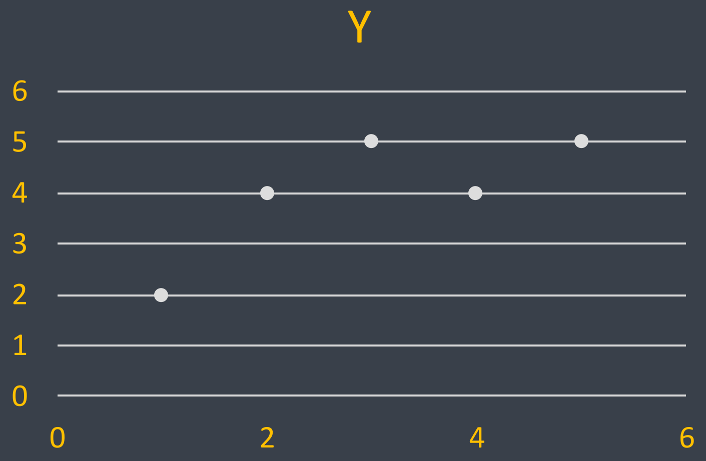
SSE = Sum of Squared Error

$= \Sigma(y-\hat{y})^2$



Least Square Method

X	Y
1	2
2	4
3	5
4	4
5	5



X	Y	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})^2$	$(X - \bar{X})\,(Y - \bar{Y})$
1	2				
2	4				
3	5				
4	4				
5	5				

$$\underline{m} = \frac{\sum (X - \bar{X})\,(Y - \bar{Y})}{\sum (X - \bar{X})^2}$$

$$y = \underline{m}x + c$$

Polynomial Regression

Introduction

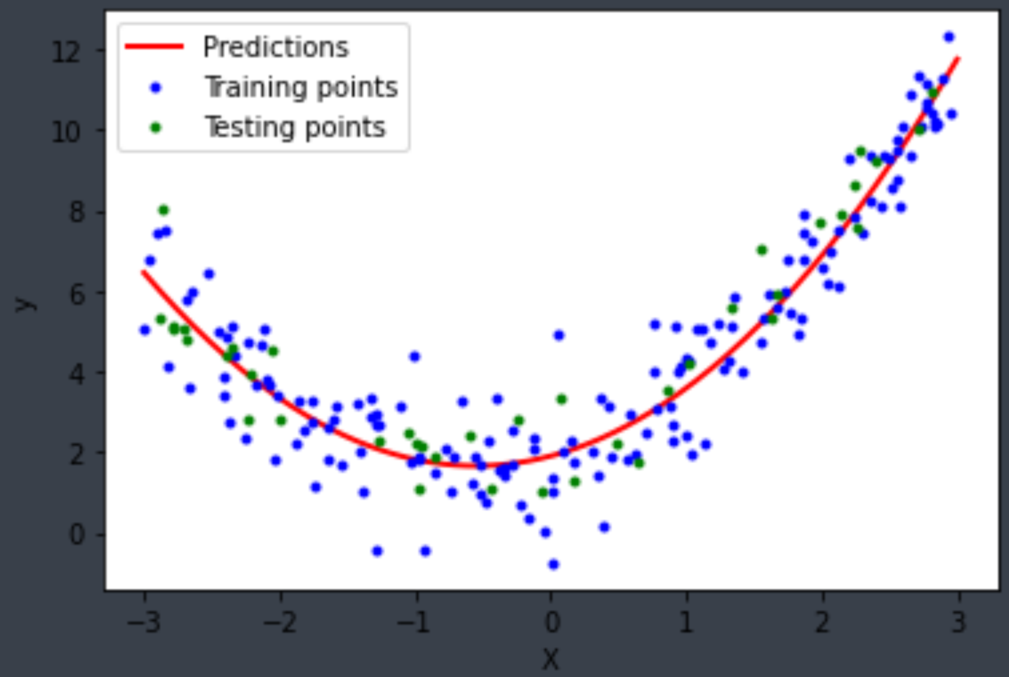
- Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression
- Suppose we have X as Independent data and Y as dependent data. Before feeding data to a mode in preprocessing stage we convert the input variables into polynomial terms using some degree
- Consider an example my input value is 35 and the degree of a polynomial is 2 so I will find 35 power 0, 35 power 1, and 35 power 2 And this helps to interpret the non-linear relationship in data.
  The equation of polynomial becomes something like this.

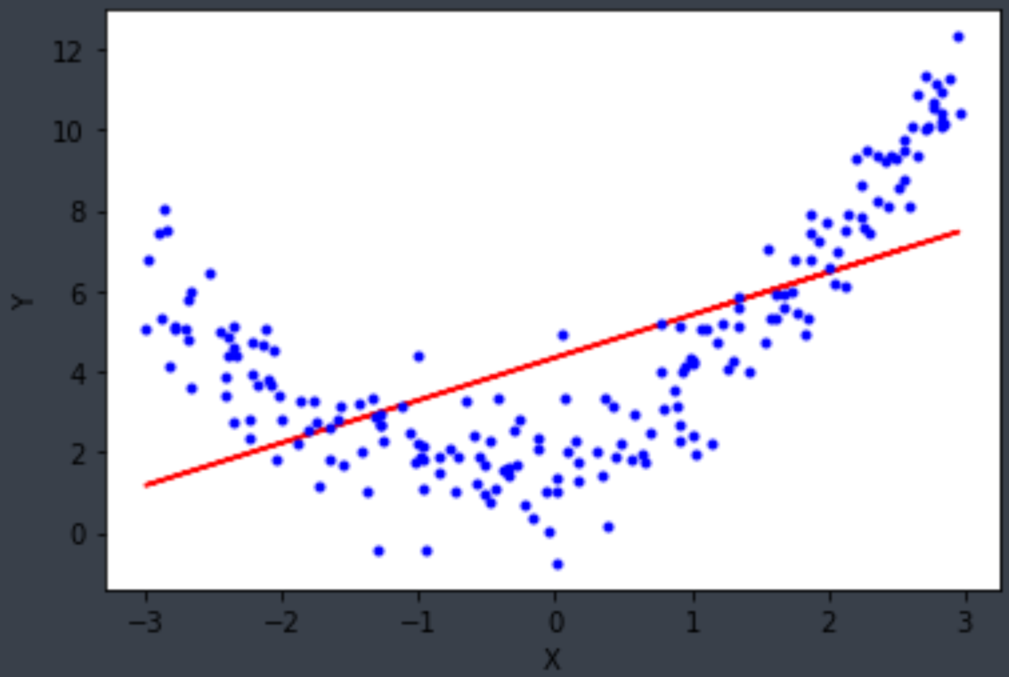
$$y = a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n$$

Introduction

- The degree of order which to use is a Hyperparameter, and we need to choose it wisely. But using a high degree of polynomial tries to overfit the data and for smaller values of degree, the model tries to underfit so we need to find the optimum value of a degree
- If you see the equation of polynomial regression carefully, then we can see that we are trying to estimate the relationship between coefficients and y

Polynomial vs Simple Linear





Stepwise Regression

Introduction

- Stepwise regression is a variable selection procedure for independent variables
- It consists of a series of steps designed to find the most useful x-variables to include in a regression model
- At each step of procedure, each variable is evaluated using a set criterion to see if should remain in the model or not
- Basis for selection could be
  - Choosing variable that satisfies the stipulated criterion
  - Removing the variable that least satisfies the criterion
- Example of such a criterion is t-statistic value

Approaches

- **Forward selection**
  - Begins with no variables in the model
  - Tests each variable as it is added to the model
  - Then keeps those that are deemed most statistically significant
  - Repeating the process until the results are optimal
- **Backward elimination**
  - Starts with a set of independent variables
  - Deleting one at a time
  - Then testing to see if the removed variable is statistically significant
- **Bidirectional elimination**
  - It is a combination of the first two methods that test which variables should be included or excluded

Advantages

- The ability to manage large amounts of potential predictor variables, fine-tuning the model to choose the best predictor variables from the available options
- It’s faster than other automatic model-selection methods
- Watching the order in which variables are removed or added can provide valuable information about the quality of the predictor variables

Disadvantages

- Stepwise regression often has many potential predictor variables but too little data to estimate coefficients meaningfully. Adding more data does not help much, if at all.
- If two predictor variables in the model are highly correlated, only one may make it into the model
- R-squared values are usually too high
- Adjusted r-squared values might be high, and then dip sharply as the model progresses. If this happens, identify the variables that were added or removed when this happens and adjust the model
- Predicted values and confidence intervals are too narrow
- P-values are given that do not have the correct meaning
- Regression coefficients are biased and coefficients for other variables are too high
- Collinearity is usually a major issue. Excessive collinearity may cause the program to dump predictor variables into the model.
- Some variables (especially dummy variables) may be removed from the model, when they are deemed important to be included. These can be manually added back in.

Elastic Net Regression

Introduction

- Linear regression refers to a model that assumes a linear relationship between input variables and the target variable
- With a single input variable, this relationship is a line, and with higher dimensions, this relationship can be thought of as a hyperplane that connects the input variables to the target variable
- The coefficients of the model are found via an optimization process that seeks to minimize the sum squared error between the predictions (*yhat*) and the expected target values (*y*)
- A problem with linear regression is that estimated coefficients of the model can become large, making the model sensitive to inputs and possibly unstable
- This is particularly true for problems with few observations (*samples*) or more samples (*n*) than input predictors (*p*) or variables
- Elastic net is a penalized linear regression model that includes both the L1 and L2 penalties during training

Regularization

Linear Regression

- Linear regression attempts to find a relationship between a dependent variable and one or more explanatory (or independent) variables
- Linear regression can be used for various tasks
- For example,
  - a given dataset has data about locations of houses in a state/province, their prices, their architecture, their neighborhood, etc. This dataset can be used to estimate the prices for houses (which may not have been listed yet) in that particular state. This is useful for house owners, potential buyers, and real estate agencies.
  - Stock price prediction
  - Weather forecasting
  - Predictive analysis from survey data
  - Market research studies
  - Future sales prediction and much more

R-Squared

- R-squared is a statistical measure of how close the data are to the fitted regression line
- It evaluates the scatter of the data points around the fitted regression line
- It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression
- It is the percentage of the response variable variation that is explained by a linear model
- R-squared = Explained variation / Total variation
- R-squared is always between 0 and 100%
  - 0% indicates that the model explains none of the variability of the response data around its mean
  - 100% indicates that the model explains all the variability of the response data around its mean
- In general, the higher the R-squared, the better the model fits your data

Adjusted R Squared

- R-Squared suffers from a major flaw: Its value never decreases no matter the number of variables we add to our regression model
- That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables
- This clearly does not make sense because some of the independent variables might not be useful in determining the target variable
- Adjusted R-squared deals with this issue
- The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable
- In doing so, we can determine whether adding new variables to the model actually increases the model fit

$$Adjusted\ R^2 = \{1 - [\frac{(1-R^2)(n-1)}{(n-k-1)}]\}$$

Regularization

- Regularization is a kind of regression that shrinks the coefficient estimates towards zero
- This technique discourages formation of a complex model, so as to avoid risk of overfitting

Underfitting and Overfitting

- Underfitting occurs when a model is not able to capture the underlying trend of the data
- Overfitting occurs when a model follows the trend of training data very closely but is not able to replicate the same performance on testing data
- A good fit model generalizes well and neither underfits nor overfits

Ridge Regression

- Ridge regression or **L2 regularization** brings values of coefficients near zero to enforce regularization
- Penalty is described by λ parameter
- The more the value of λ, the lesser the flexibility
- For low values of λ, the coefficients are very similar to that of a multiple linear regression model
- As λ increases, the differences between the results of Ridge model and linear regression model increase

Lasso Regression

- Lasso regression or **L1 regularization** not only brings values of coefficients near zero but to exact zero in case of weak regressors
- So, it not only shrinks coefficient estimates towards zero but also helps in feature selection
- Penalty is described by λ parameter.
- The more the value of λ, the lesser the flexibility
- For low values of λ, the coefficients are very similar to that of a multiple linear regression model
- As λ increases, the differences between the results of Lasso model and linear regression model increase