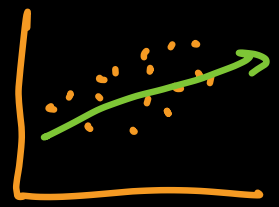
Regression Analysis

predicting future value

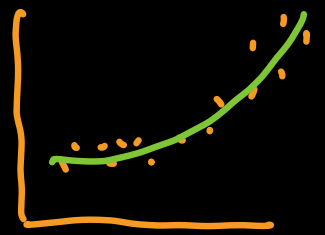
Regression Analysis

- Regression analysis is a fundamental statistical technique used to examine relationships between variables and make predictions
- It models the relationship between a dependent variable (outcome) and one or more independent variables (predictors) by fitting a mathematical equation to observed data
  - collected data
- Applications
  - **Business**: Forecasting sales, pricing strategy, demand analysis
  - **Healthcare**: Predicting disease outcomes, analyzing treatment effects
  - **Economics**: Studying the impact of interest rates or unemployment
  - **Machine Learning**: Basis for algorithms in predictive modeling

Types of regression



- **Linear Regression** → best fit Regression Line
  - Models the relationship using a straight line.
  - Used when the dependent variable is continuous and the relationship is linear
- **Multiple Linear Regression**
  - Involves two or more independent variables.



- **Polynomial Regression**
  - Used when the relationship between variables is nonlinear
  - Includes powers of the independent variable
- **Logistic Regression** → classification
  - Used when the dependent variable is binary (e.g., Yes/No, 0/1)
  - Predicts the probability of an outcome
- **Ridge, Lasso, and ElasticNet Regression**
  - Regularized versions of linear regression to avoid overfitting

Linear Regression

Graphical

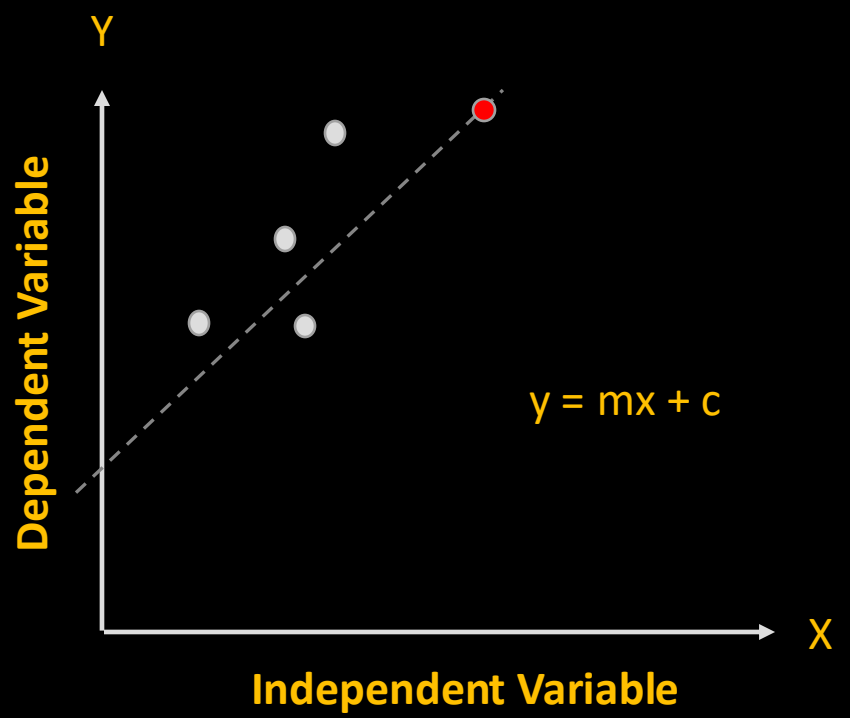
Algebric

Overview

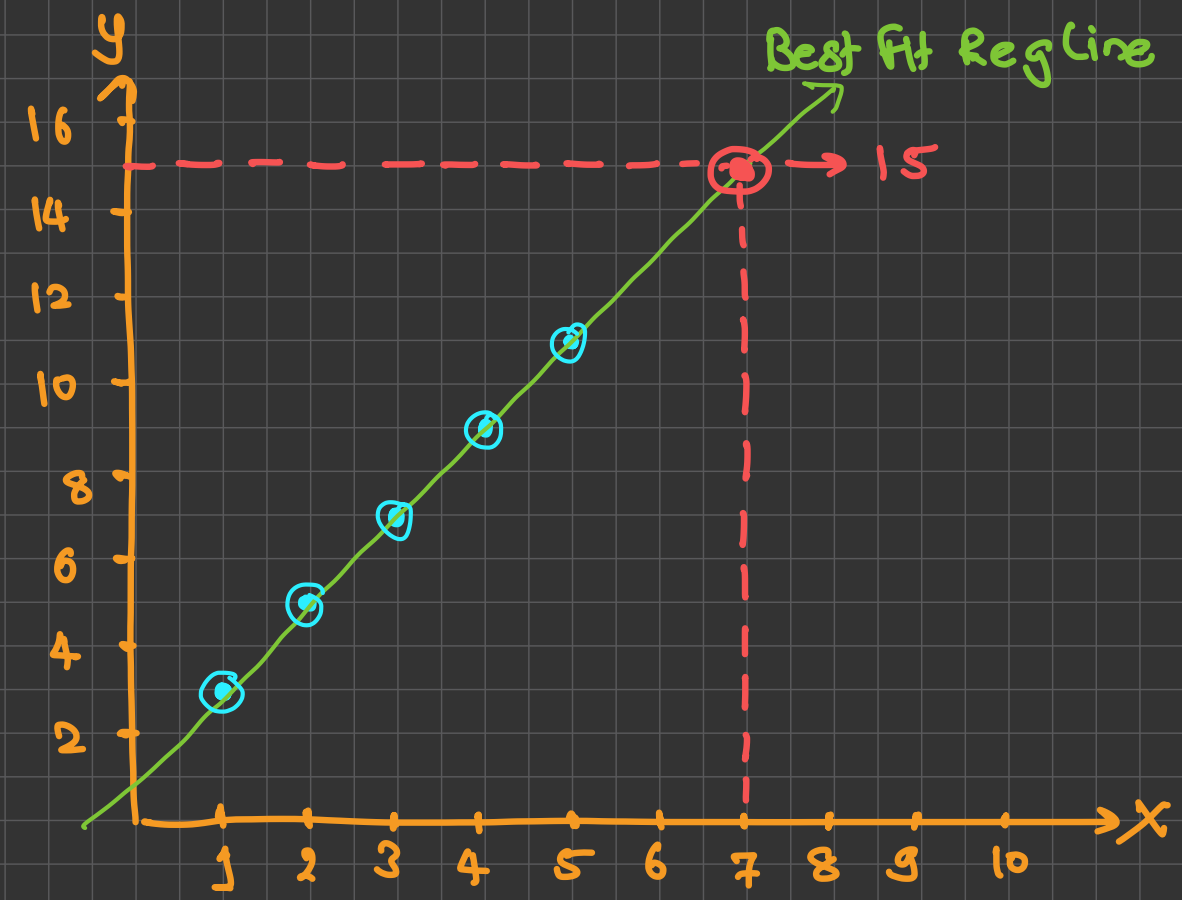
- Linear regression is a fundamental statistical technique used to model the **relationship between a dependent variable (Y)** and **one independent variable (X)** by fitting a straight line through the data points
- It helps in **predicting the value of Y** for a given value of X
- **Linear Regression Equation**

  $Y=a+bX+eY = a + bX + eY=a+bX+e$

  - **Y**: Dependent variable (what we want to predict)
  - **X**: Independent variable (predictor)
  - **a**: Intercept (value of Y when X = 0)
  - **b**: Slope (change in Y for a one-unit change in X)
  - **e**: Error term (difference between actual and predicted values)



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

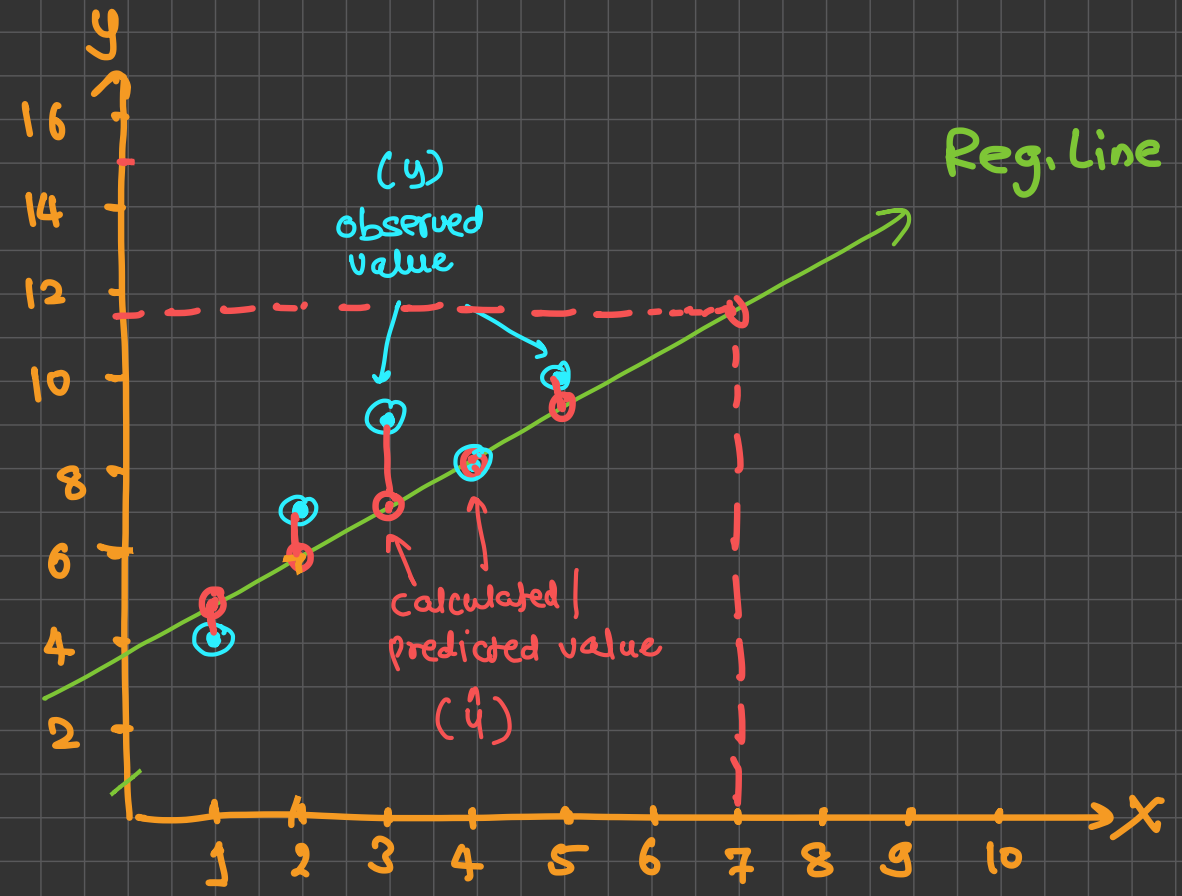


$x = 7,\ y = 15$

$$y = 2x + 1$$

x	y	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$
1	4	5	−1	1
2	7	6	1	1
3	9	7	2	4
4	8	8	0	0
5	10	9	1	1
7	?			7
	⇓			↑
	12			

$y = mx + c$



sum of squared Errors : $\Sigma(y-\hat{y})^2$
(SSE)

↳ ordinary lowest squared Errors

↳ Best fit line = lowest SSE

Assumptions of Linear Regression

- **Linearity**
  - The relationship between X and Y is linear
- **Independence**
  - Observations are independent of each other
- **Homoscedasticity**
  - Constant variance of errors across all levels of X
- **Normality**
  - Residuals (errors) are normally distributed

Polynomial Regression

Introduction

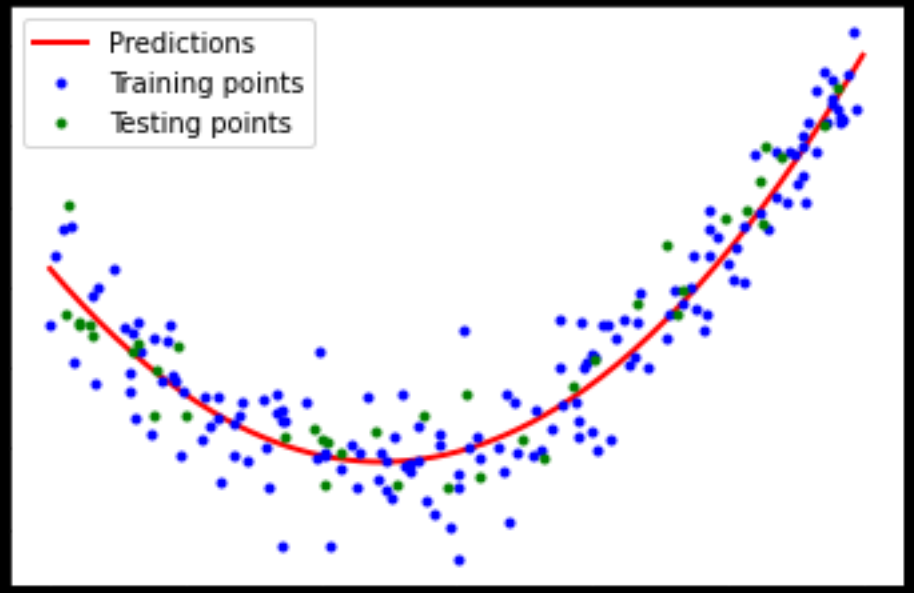
- Polynomial regression is a form of Linear regression where only due to the Non-linear relationship between dependent and independent variables we add some polynomial terms to linear regression to convert it into Polynomial regression
- Suppose we have X as Independent data and Y as dependent data. Before feeding data to a mode in pre-processing 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.

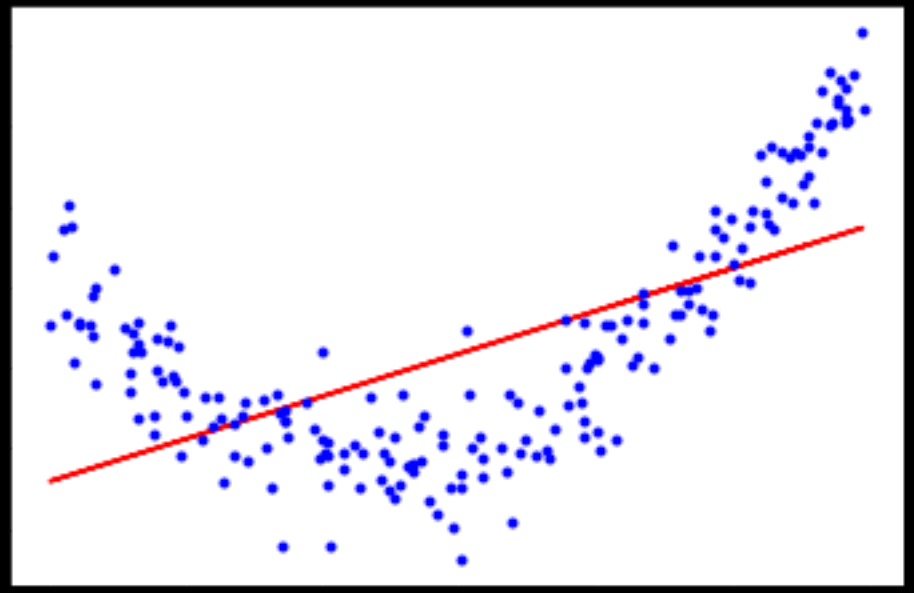
$$y = a_0 + a_1x_1 + a_2x_1^2 + \ldots + a_nx_1^n$$

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 variable 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