Machine Learning

Lecture 17: Regression

COURSE CODE: CSE451

2023

Course Teacher

Dr. Mrinal Kanti Baowaly

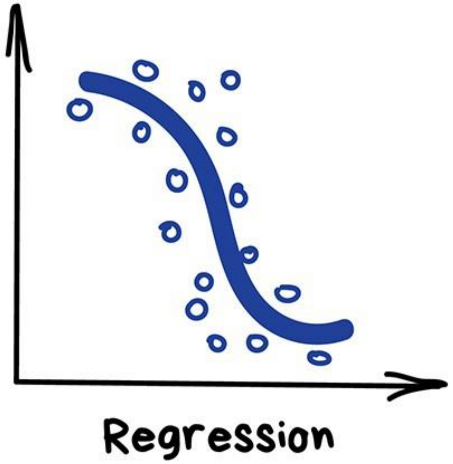
Associate Professor

Department of Computer Science and Engineering, Bangabandhu Sheikh Mujibur Rahman Science and Technology University, Bangladesh.

Email: mkbaowaly@gmail.com

Regression Analysis

- Supervised learning that predicts a real/continuous value such as income, price, height, weight, scores or probability etc.
- Investigates the relationship between a dependent (target) and independent variable (s) (predictor)
- Why do we use Regression Analysis?
  - ✓Forecasting
  - ✓Demand and sales volume analysis
  - ✓Time series modelling
  - ✓Medical diagnosis etc.



Source: AnalyticsVidhya

Types of Regression Models

- Linear Regression
- Polynomial Regression
- Logistics Regression

Linear Regression

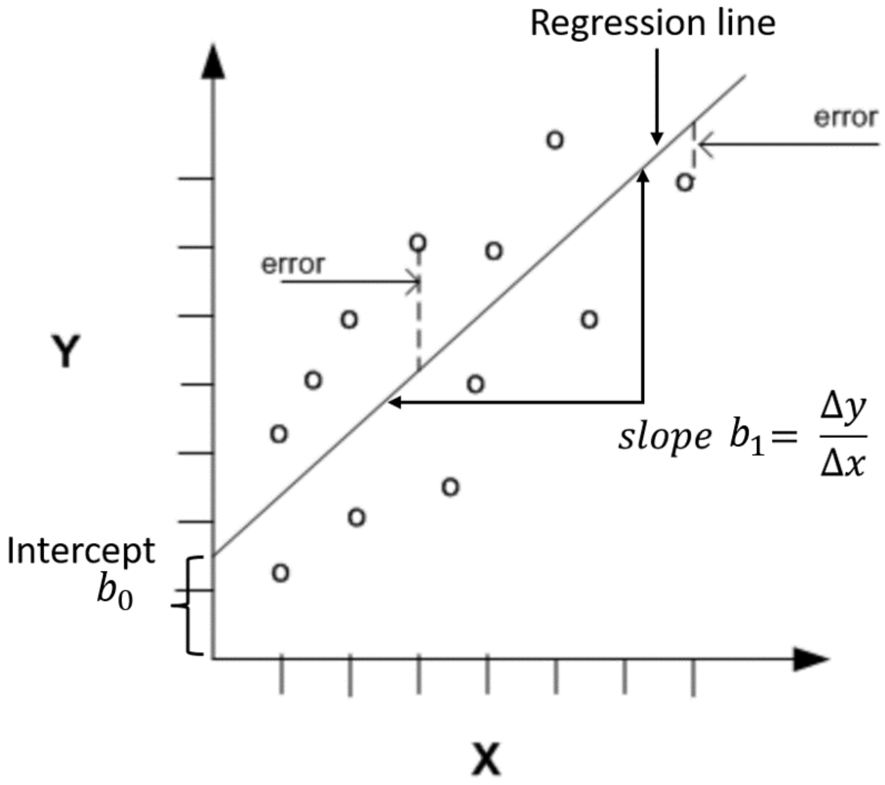
- **Linear regression** is a linear model, e.g. a model that establishes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x) using a best fit straight line (also known as regression line).
- Linear Regression is very sensitive to outliers. It can terribly affect the regression line and eventually the forecasted values.
- **Simple linear regression**: When the method has a single input variable (x)
- **Multiple linear regression:** When the method has multiple input variables (x)

Simple Linear Regression

- Simple linear regression is useful for finding relationship between two variables. One is predictor or independent variable (x) and other is response or dependent variable (y)
- The core idea is to obtain a straight line that best fits the data. The line can be represented by an equation:

$$y = b_0 + b_1 x$$

where *y* is the dependent variable, x is the independent variable, b_1 is the slope of the line and b_0 is y-intercept (constant).



How to Solve Linear Regression Line

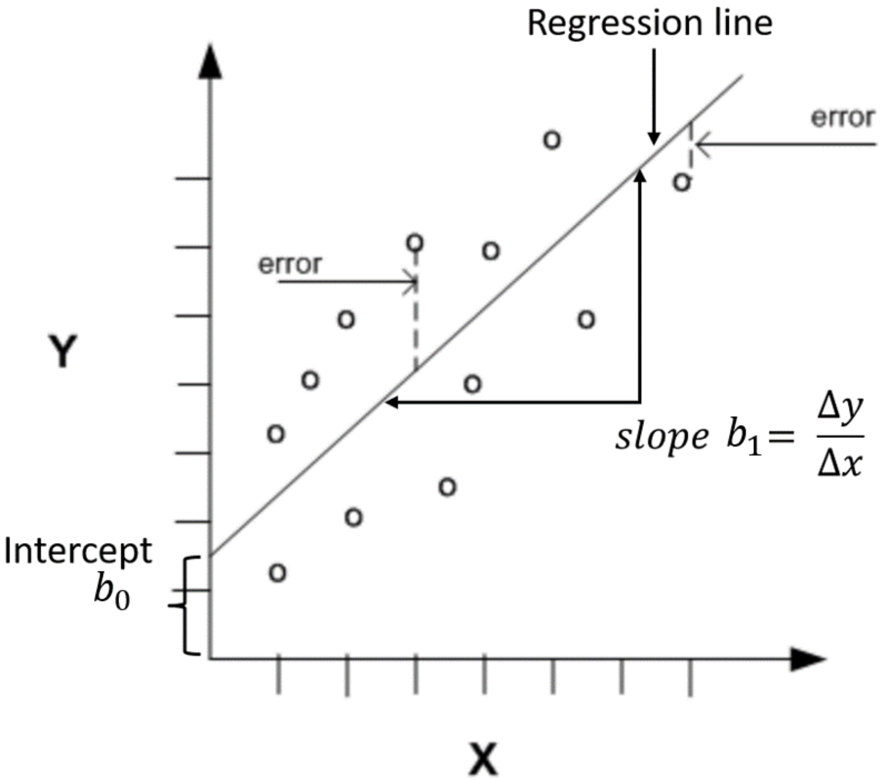
- The regression line can be represented by an equation:

$$y = b_0 + b_1 x$$

b_0 and b_1 can be calculated by using the following formula.

$$b_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$



Source: Link

Simple Linear Regression – An Example

- Find linear regression equation for the following two sets of data:

x:	2	4	6	8
y:	3	7	5	10

- Solution:

x	y	x^2	xy
2	3	4	6
4	7	16	28
6	5	36	30
8	10	64	80
$\sum x$ = 20	$\sum y$ = 25	$\sum x^2$ = 120	$\sum xy$ = 144

$$b_0 = \frac{\sum y \sum x^2 - \sum x \sum xy}{n(\sum x^2) - (\sum x)^2}$$

$$= \frac{25 \times 120 - 20 \times 144}{4(120) - 400} = 1.5$$

$$b_1 = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2}$$

$$= \frac{4 \times 144 - 20 \times 25}{4 \times 120 - 400} = 0.95$$

The equation: $y = b_0 + b_1 x$ $\quad y = 1.5 + 0.95x$

Multiple Linear Regression (MLR)

- Multiple linear regression (MLR), also known simply as multiple regression, is a machine learning technique that uses several input variables to predict the outcome of a response variable.
- The goal of multiple linear regression (MLR) is to model the linear relationship between the input (independent) variables and response (dependent) variable by fitting a plane or a hyper-plane to training data. The plane can be represented by an equation as follows, given n input variables:

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \ \ldots\ldots + b_n x_{in}\ ; for\ i = 1, 2, \ldots\ldots n$$

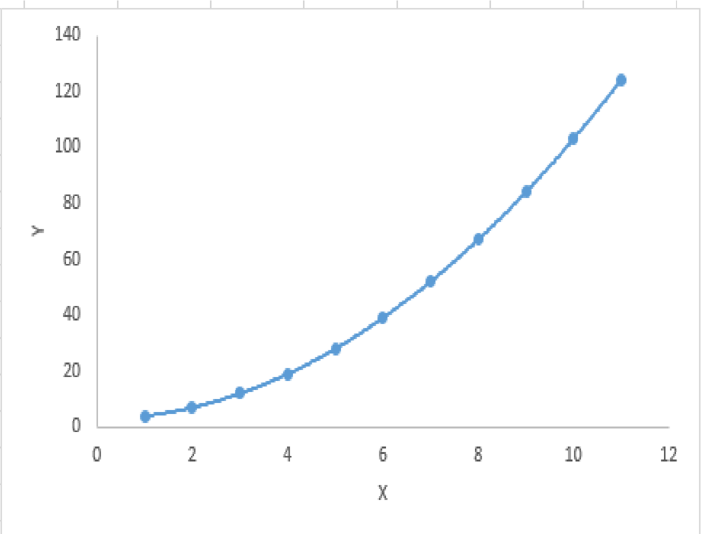
where y_i is the dependent variable, x_i is the independent variable, b_i is the slope for each input variables and b_0 is y-intercept (constant).

Polynomial Regression

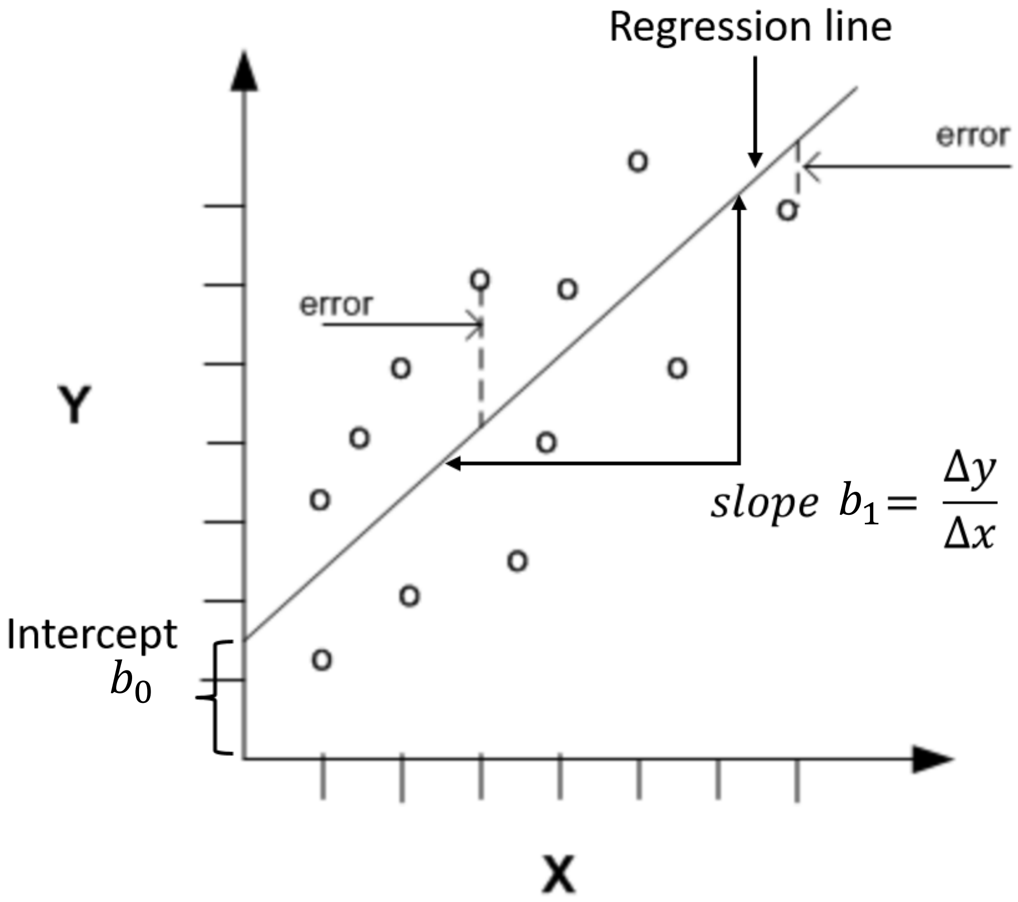
- A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation:

$$y = b_0 + b_1 x^2$$

- In this regression, the best fit line is not a straight line. It is rather a curve that fits into the data points.



How to evaluate regression model



$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

Where,

$\hat{y}$ − *predicted value of y*

$\bar{y}$ − *mean value of y*

Source: AnalyticsVidhya

Example of calculating evaluation metrics

- Original values: -2, 1, -3, 2, 3, 5, 4, 6, 5, 6, 7
- Predicted values: -1, -1, -2, 2, 3, 4, 4, 5, 5, 7, 7
- Find:
  - ✓MAE (Mean absolute error) : 0.6363636
  - ✓MSE (Mean Squared Error): 0.8181818
  - ✓RMSE (Root Mean Squared Error): 0.904534
  - ✓R^2 (R-squared) Score: 0.9173623
- Lab Work: Use library function as well as the formulas manually to calculate these metrics

How to obtain best-fit regression line

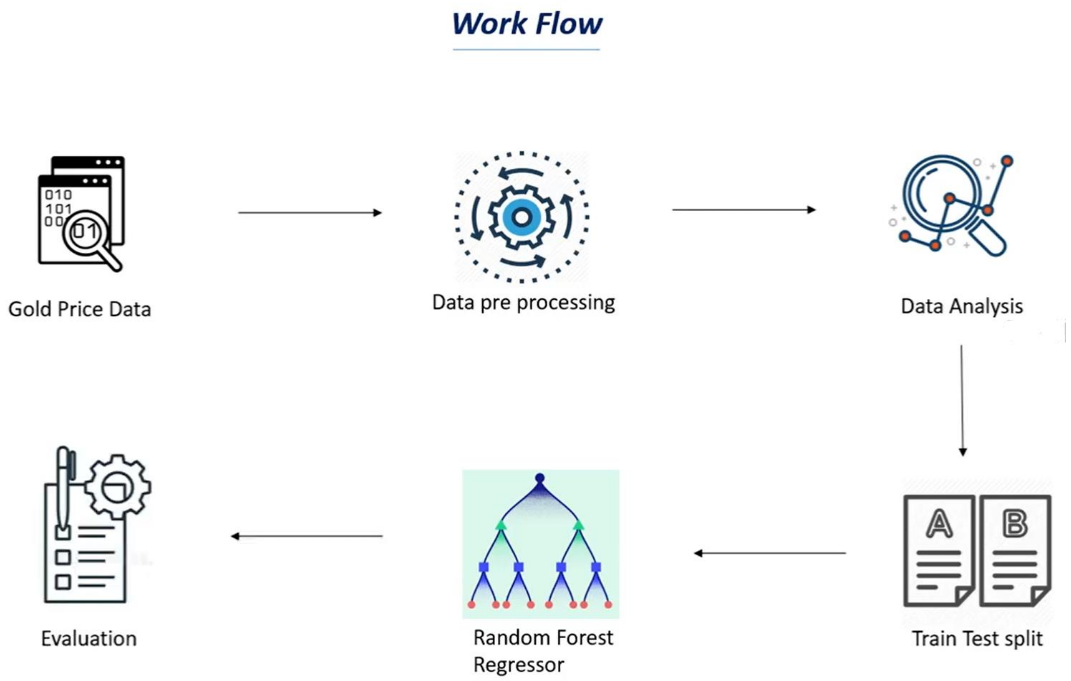
- **Cost Function (Loss Function)**: By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the coefficient values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

  Cost function of Linear Regression may be MAE (Mean absolute error), MSE (Mean Squared Error), RMSE (Root Mean Squared Error), R2 (R-squared) Score etc.

- **Gradient Descent:** This is a process of optimizing the values of the coefficients by iteratively minimizing the cost function of your model. The idea is to start with random values for each coefficient and then iteratively update the values until reaching minimum cost (error).

An Example Project on Regression

- Building A Gold Price Prediction (GLD) Model Using Machine Learning



Date	SPX	GLD	USO	SLV	EUR/USD
1/2/2008	1447.16	84.86	78.47	15.18	1.47
1/10/2008	1420.32	88.25	74.01	16.06	1.48
1/17/2008	1333.25	86.50	71.02	15.71	1.46
1/28/2008	1353.95	91.75	72.34	16.54	1.478
2/26/2008	1381.29	93.70	80.09	18.60	1.49
3/18/2008	1330.73	96.50	85.80	19.37	1.56

Dataset Link

Prediction: AV, Kaggle

Linear Regression vs Logistic Regression

- Linear Regression is used for solving Regression problem. It is used to predict the continuous dependent variable using a given set of independent variables.
- Logistic regression is used for solving Classification problems. It is used to predict the categorical dependent variable using a given set of independent variables.

Detail: JavaTpoint

Study Materials of Regression

7 Regression Techniques you should know!

Know The Best Evaluation Metrics for Your Regression Model

Linear Regression for Machine Learning

Linear Regression in Machine Learning, Simple Linear Regression, Multiple Linear Regression, Polynomial Regression