



Linear Regression



Outline

- What is Regression?
- Regression use-case
- Types of Regression
- What is Linear Regression?
- Checking goodness of fit using R square
- Method
- Examples
- Implementation of Linear Regression using Python (scikit-learn)

What is Regression?

Regression analysis is a form of predictive modelling technique which investigates the relationship between a dependent variable and independent variable.



Use of Regression

Three major uses of Regression are:

- Determining the strength of predictions
- Forecasting an effect
- Trend forecasting



Determining the strength of predictions

Regression is used to identify the strength of the effect that the independent variables have on the dependent variable.

- What is the strength of relationship between sales and marketing spending?
- What is the relationship between age and income?



Forecasting an effect

In this case, Regression can be used to forecast effects or impact of changes, i.e., it helps us understand how much the dependent variable changes with the change in one or more independent variable.

- How much additional sales income will I get for each thousand naira spent on marketing?



Trend forecasting

Regression can be used to predict trends and future values.

It can be used to get point estimates.

- What will be the price of Bitcoin in next 6 months?



Types of Regression

- Linear Regression (Simple or Multiple).
- **Logistic Regression (THIS IS NOT REGRESSION)**
- Ridge Regression.
- Lasso Regression.
- Polynomial Regression.
- Bayesian Linear Regression.
- Partial Least Squares Regression.

What is Linear Regression?

$$Y = mX + c$$



What is Linear Regression?

In Linear Regression the data is modelled using a straight line equation.

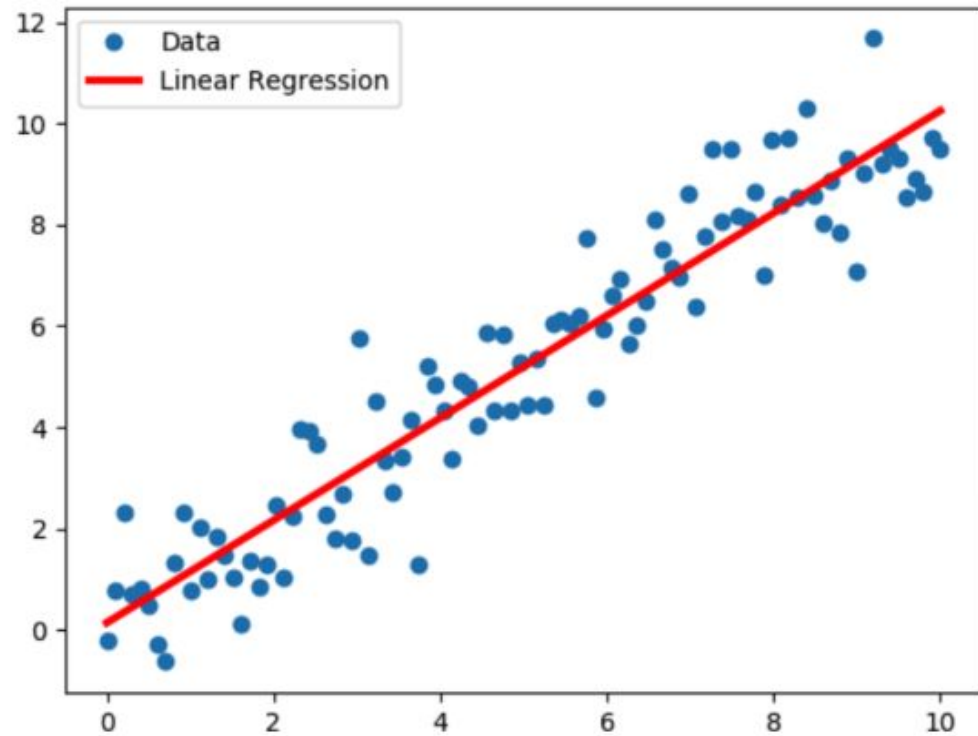
$$Y = mX + c$$

For an independent variable, X, and dependent variable, Y, we are interested with the correlation of the X and Y variable.

This means that if X and Y are continuous, every value of X has a corresponding value of Y.

Linear Regression is used with continuous variables and the output of prediction is the value of the Y variable.

The performance of the Linear Regression model is measured by R squared, etc.



Linear Regression



Assumptions of Linear Regression

- Linearity: The relationship between X and the mean of Y is linear.
- Normality or Multivariate Normality: For any fixed value of X , Y is normally distributed.
- Multicollinearity or Independence: Observations are independent of each other.
- Autocorrelation
- Homoscedasticity: The variance of residual is the same for any value of X .



Linearity

First, linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots, the following two examples depict two cases, where no and little linearity is present.



Normality

Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.



Multicollinearity

Thirdly, linear regression assumes that there is little or no multicollinearity in the data. Multicollinearity occurs when the independent variables are too highly correlated with each other.

- 1) Correlation matrix – when computing the matrix of Pearson's Bivariate Correlation among all independent variables the correlation coefficients need to be smaller than 1.
- 2) Tolerance - With $T < 0.1$ there might be multicollinearity in the data and with $T < 0.01$ there certainly is.
- 3) Variance Inflation Factor (VIF) - With $VIF > 5$ there is an indication that multicollinearity may be present; with $VIF > 10$ there is certainly multicollinearity among the variables.



Autocorrelation

Fourthly, linear regression analysis requires that there is little or no autocorrelation in the data. 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 $y(x)$.

While a scatterplot allows you to check for autocorrelations, you can test the linear regression model for autocorrelation with the Durbin-Watson test.



Homoscedasticity

The last assumption of the linear regression analysis is homoscedasticity. The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line).

If homoscedasticity is present, a non-linear correction might fix the problem.

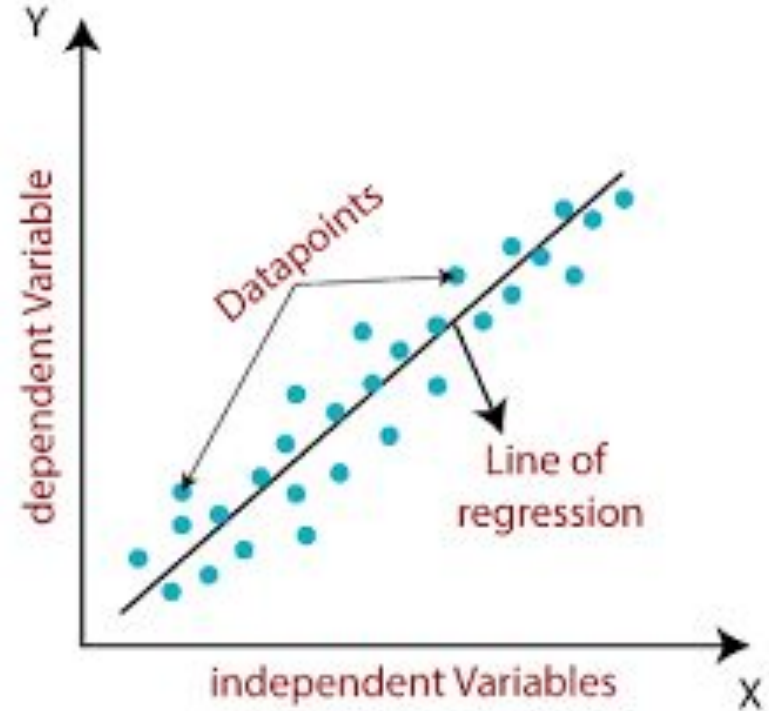
Understanding Linear Regression

Dependent variable

Independent variable

Line of regression

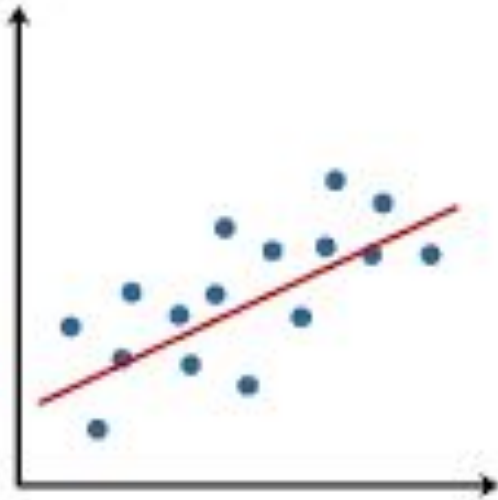
Data points



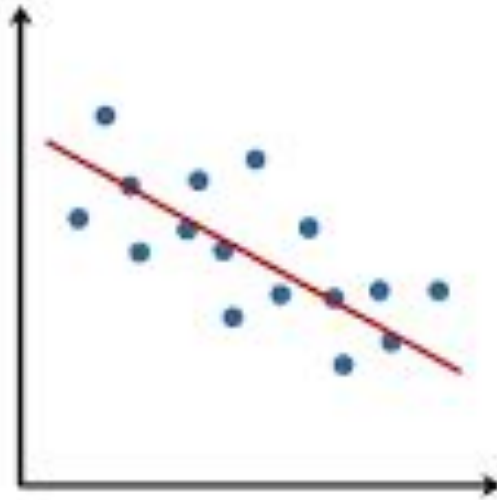


Understanding Linear Regression

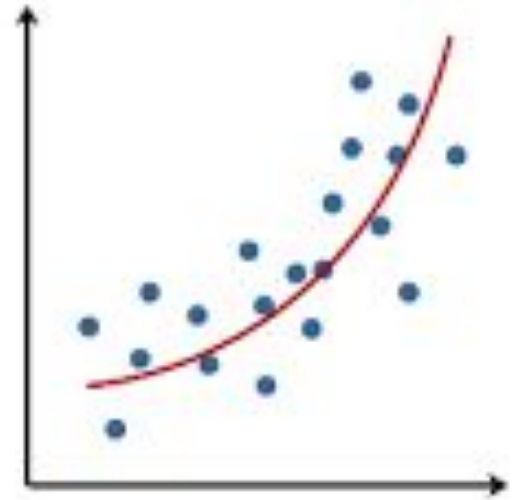
Linear



Linear

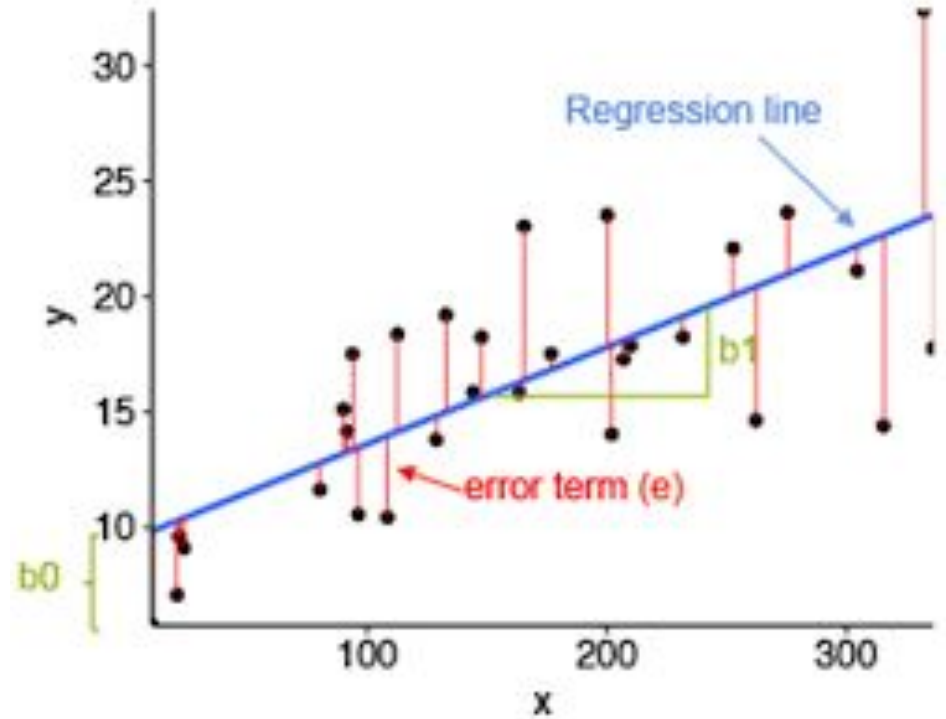


Non-Linear

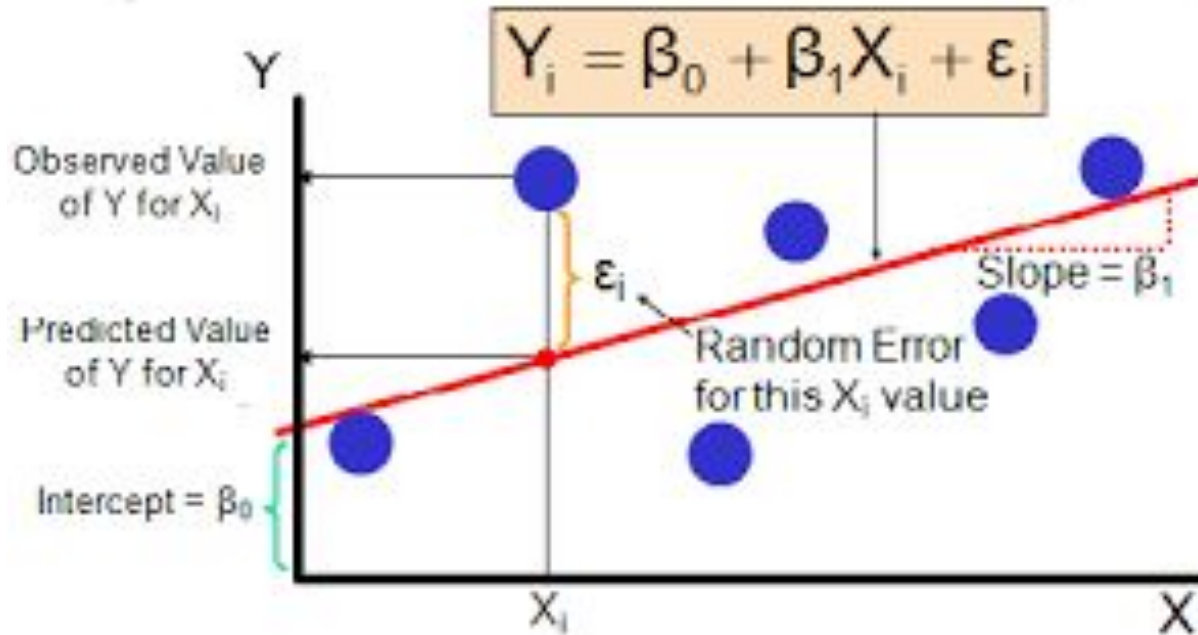


Understanding Linear Regression

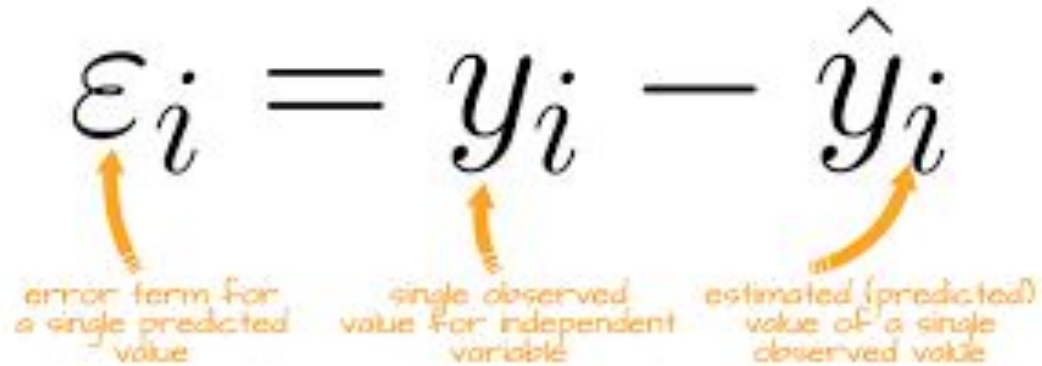
Error term (e)



Understanding Linear Regression



Understanding Linear Regression

$$\epsilon_i = y_i - \hat{y}_i$$


error term for a single predicted value

single observed value for independent variable

estimated (predicted) value of a single observed value



Example - Question 1

Suppose the relationship between the independent variable height (x) and dependent variable weight (y) is described by a simple linear regression model with true regression line

$$y = 7.5 + 0.5x$$

Q1: What is the interpretation of $m = 0.5$?



Height (x) vs Weight (y)

A1: The expected change in height associated with a 1-unit increase in weight



Example - Question 2

Suppose the relationship between the independent variable height (x) and dependent variable weight (y) is described by a simple linear regression model with true regression line

$$y = 7.5 + 0.5x$$

Q2: If $x = 20$ what is the expected value of Y?



Height (x) vs Weight (y)

$$A2: \hat{y} = 7.5 + 0.5(20) = 17.5$$



Checking goodness of fit using R square Method

R-squared is a statistical measure of how close the data are to the fitted regression line.

It is also known as **coefficient of determination** or **coefficient of multiple determination**.



Categorical independent variable

Qualitative variables are easily incorporated in regression analysis through dummy variables.

Simple example: sex can be coded as 0/1

What if my categorical variable contains three levels:

Marital status - 0/1/2 (Single/Married/Divorced)

The above example is wrong and will lead to multicollinearity.

Solution is to set up a series of dummy variable.

scikit-learn

Machine Learning in Python

Getting Started

Release Highlights for 1.2

GitHub

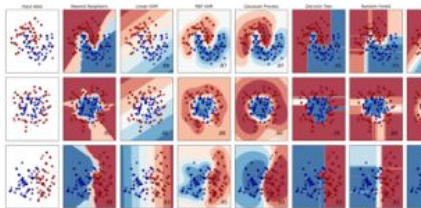
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.

Algorithms: SVM, nearest neighbors, random forest, and more...

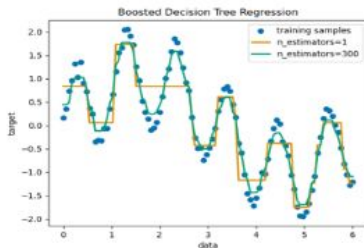


Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, nearest neighbors, random forest, and more...



Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, and more...

