# Linear Regression

## Epoch IIT Hyderabad

Himani Agrawal
MA22BTECH11008

## Contents

### Introduction

Linear regression is a statistical method used to model the relationship between two variables. It involves finding the line of best fit that describes the relationship between the two variables. The line of best fit is a straight line that minimizes the distance between the observed values and the predicted values.

### Assumptions of linear regression

1. Linearity: The relationship between the independent variables (predictors) and the dependent variable (response) is assumed to be linear. This means that changes in the response variable are proportional to changes in the predictor variables.
2. Independence: The residuals (the differences between the actual and predicted values) should be independent of each other. In other words, the error terms should not exhibit any patterns or trends.
3. Homoscedasticity: Also known as constant variance, this assumption states that the variance of the residuals should be constant across all levels of the predictor variables. In simpler terms, the spread of the residuals should be roughly the same throughout the range of predicted values.
4. Normality: The residuals should be normally distributed. This assumption is important for hypothesis testing, confidence intervals, and other statistical inferences.
5. No Multicollinearity: The predictor variables should not be highly correlated with each other. High multicollinearity can make it difficult to isolate the individual effects of each predictor on the response variable.

## Mathematical Formula

The value to be predicted is given by:

Y(pred) = b0 + b1*x

where, Y = Dependent variable, bo = constant/Intercept, b1 = Slope/Intercept, X = Independent variable.

the difference between the observed value of the dependent variable($y_i$) and the predicted value(predicted) is called the residuals.

$$\varepsilon_i = Y(pred) - Y_i$$

R-Squared is a number that explains the amount of variation that is explained/captured by the developed model. It always ranges between 0 & 1 . Overall, the higher the value of R-squared, the better the model fits the data.

R^2 = 1 – ( RSS/TSS )

Where, RSS is residual sum square given by $RSS=\sum(y_i-b_0-b_1x_i)^2$ and;

TSS is total sum of square given by $TSS = \sum (y_i - \bar{y_i})^2$

## optimization

Gradient Descent is one of the optimization algorithms that optimize the objective function to reach the optimal minimal solution. To find the optimum solution we need to reduce the MSE(Mean squared error) for all data points. This is done by updating the values of B0 and B1 iteratively until we get an optimal solution.