# Linear Regression

## Supervised Learning

# Agenda

- Regression
- What is linear regression?
- Regression use case
- Linear regression line
- Error term and calculating the error term
- Which line best fits our data?
- Methods to get the best fit line
  - Ordinary least squares method
  - Gradient descent
- Error Analysis
- Assumptions of Linear regression
- Advantages and disadvantages
- Summary

# Regression

- Regression generally means “**stepping back towards the average**”.
- Regression analysis is also defined as the **measure of average relationship** between two or more variables.
- The predicted variable is known as **dependent/target/response.**
- The variable(s) which is used for prediction is known as **independent/explanatory/regressor.**

# What is linear regression?

- Linear regression is one of the supervised learning algorithms that are used to **identify a relationship between two or more variables.**

- This relationship can be used to predict values for one variable, if value(s) of other variable(s) is given.

- A simple linear regression model (also called **bivariate regression**) has one independent variable X that has a linear relationship with the dependent variable Y.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Here $\beta_0$ and $\beta_1$ are the parameters of the linear regression model.

# Regression: Use case

- Let us consider impact of a single variable for now.


We say, that only Age decides what the Blood pressure of a person should be.

# Regression: Use case

Let us consider the following data:

- The task is to predict the blood pressure when the age is 40.

| Age in years (x) | Blood pressure (y) |
|---|---|
| 25 | 120 |
| 36 | 135 |
| 68 | 143 |
| 55 | 139 |
| 49 | 120 |
| 72 | 165 |
| 40 | ? |

# Linear Regression line

$$y = \beta_0 + \beta_1 x + \varepsilon$$

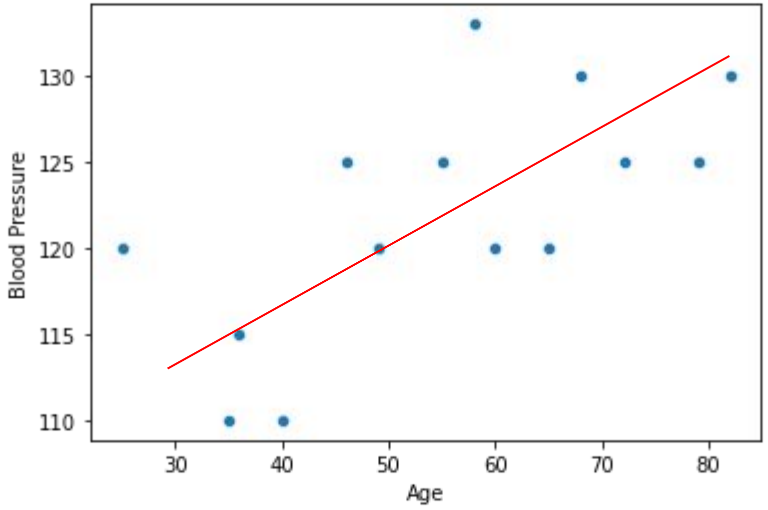
$y$ = set of values taken by dependent variable Y

$x$ = set of values taken by independent variable X

$\beta_0$ = y intercept

$\beta_1$ = slope

$\varepsilon$ = random error component



# Linear Regression line

In context with our example,

$$\textbf{Blood Pressure} = \beta_0 + \beta_1 \textbf{ Age} + \varepsilon$$

y = set of values taken by dependent variable, blood pressure

x = set of values taken by independent variable, age

$\beta_0$ = Blood pressure value where the best fit line cuts the Y - axis (blood pressure)

$\beta_1$ = beta coefficient for age

$\varepsilon$ = random error component

# What is error term?

- Error term also called residual represents the distance of the observed value from the value predicted by regression line.
- In our example,

  Error term = Actual blood pressure - Predicted blood pressure

  for each observation

# Calculating the error term

- Equation of regression line is given by, $\mathbf{y = \beta_0 + \beta_1 x + \varepsilon}$

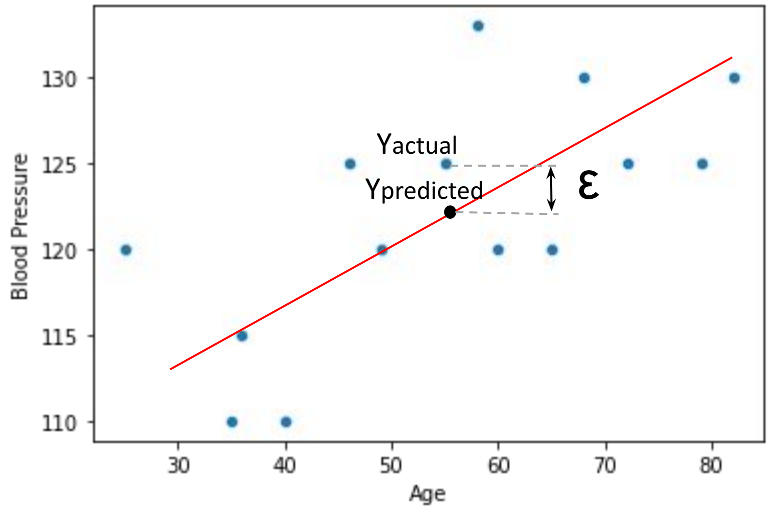
- Error term can be calculated as, $\mathbf{\varepsilon = y - (\beta_0 + \beta_1 x)}$

- We have an error term for every observation in the data.

  $\mathbf{\varepsilon_i = y_{actual} - y_{predicted}}$

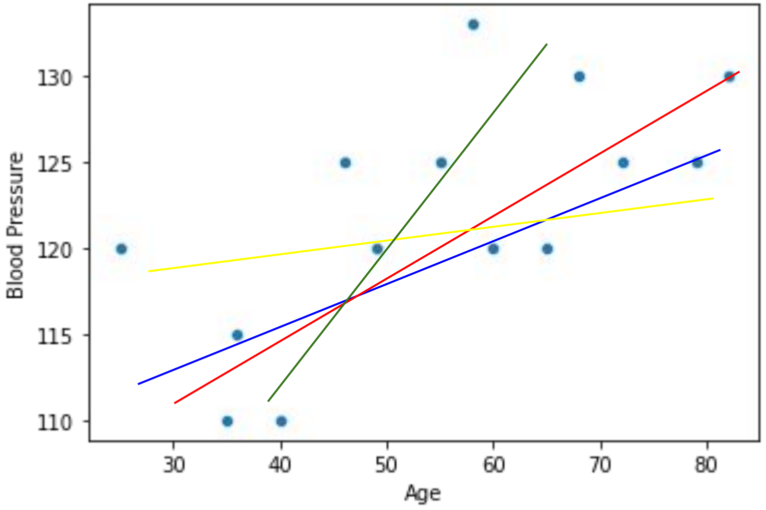
- Squared error : $\mathbf{\varepsilon_i^2 = (y_{actual} - y_{predicted})^2}$

- Sum of squared errors = $\sum \mathbf{\varepsilon_i^2}$



# Which line best fits our data?

- The regression line which best explains the trend in the data is the best fit line
- The line with the least error will be chosen as the best fitting line



# Methods to get the best fit line

- Two methods can be used to find the best fit line

# Ordinary Least Squares method

- The ordinary least square method is used to find the best fit line for given data.
- This method aims at minimizing the sum of squares of the error terms, that is, it determines those values of $\beta_0$ **and $\beta_1$ at which the error terms are minimum.**

$$\min \sum_{i=1}^{n} (yi - (\beta_0 + \beta_i\, xi))^2$$

# Math behind OLS

- We have seen that the error term $\boldsymbol{\varepsilon = y - (\beta_0 + \beta_1 x)}$
- The OLS method minimizes E = $\sum \boldsymbol{\varepsilon^2 =} \sum(\boldsymbol{y - (\beta_0 + \beta_1 x))^2}$
- To minimize the error we take partial derivatives with respect to $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ and equate them to zero.

$$\delta E/\delta\beta_0 = 0$$

$$\delta E/\delta\beta_1 = 0$$

- So we get two equations with two unknowns, $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$

# Math behind OLS

- So we get:

$$\delta E/\delta \boldsymbol{\beta_0} = \sum 2\,(\mathbf{y} - \boldsymbol{\beta_0} - \boldsymbol{\beta_1}\mathbf{x})\,(-1) = 0$$

$$\delta E/\delta \boldsymbol{\beta}_1 = \sum 2\,(\mathbf{y} - \boldsymbol{\beta_0} - \boldsymbol{\beta_1}\mathbf{x}))\,(-x_1) = 0$$

- Expanding these equations, we get $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ as:

$$\boxed{\boldsymbol{\beta_0} = \bar{\mathbf{y}} - \boldsymbol{\beta_1}\bar{\mathbf{x}}}$$

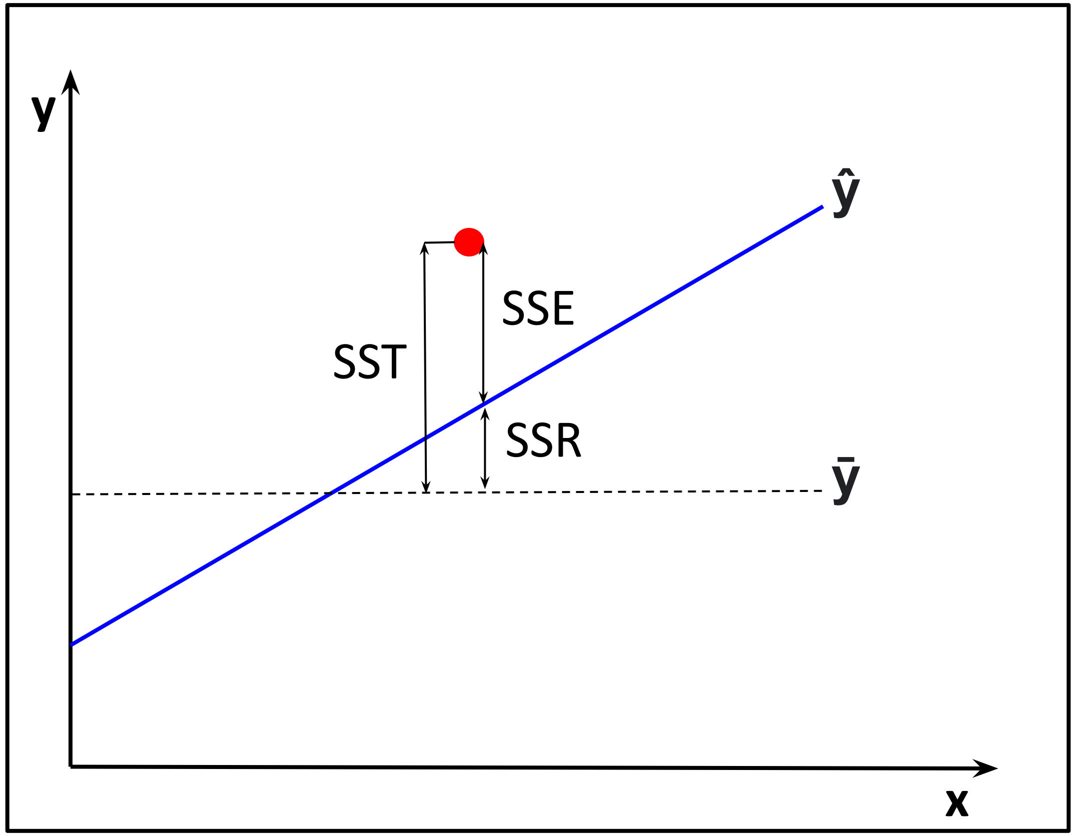
$$\boxed{\boldsymbol{\beta_1} = \mathbf{Cov(X,Y) / Var(X)}}$$

Here, $\bar{\mathbf{y}}$ is sample mean of y and $\bar{\mathbf{x}}$ is sample mean of x

# Interpretation of β coefficients

- $\beta_1$ gives the amount of change in response variable per unit change in predictor variable.
- $\beta_0$ is the y intercept which means when X=0, Y is $\beta_0$.
- β's have an associated p value, which is used to assess its significance in prediction of response variable.
- Depending on whether β's take a positive value k or - k the response variable increases or decreases respectively by k units for every one unit increment in a predictor variable, keeping all other predictor variables constant.

# Measures of variation




Sum of squared error (SSE):

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

Sum of squared total (SST):

$\Sigma(y_i - \bar{y})^2$

Sum of squared regression (SSR):

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

**yi** = observed values of y
**ŷi** = predicted values of y
**ȳ** = mean value of variable y

# R-squared

- **$R^2$** also called the **coefficient of determination** gives total percentage of variation in Y that is explained by predictor variable.

$$R^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SSR}{SST}$$

$$0 \leq R^2 \leq 1$$

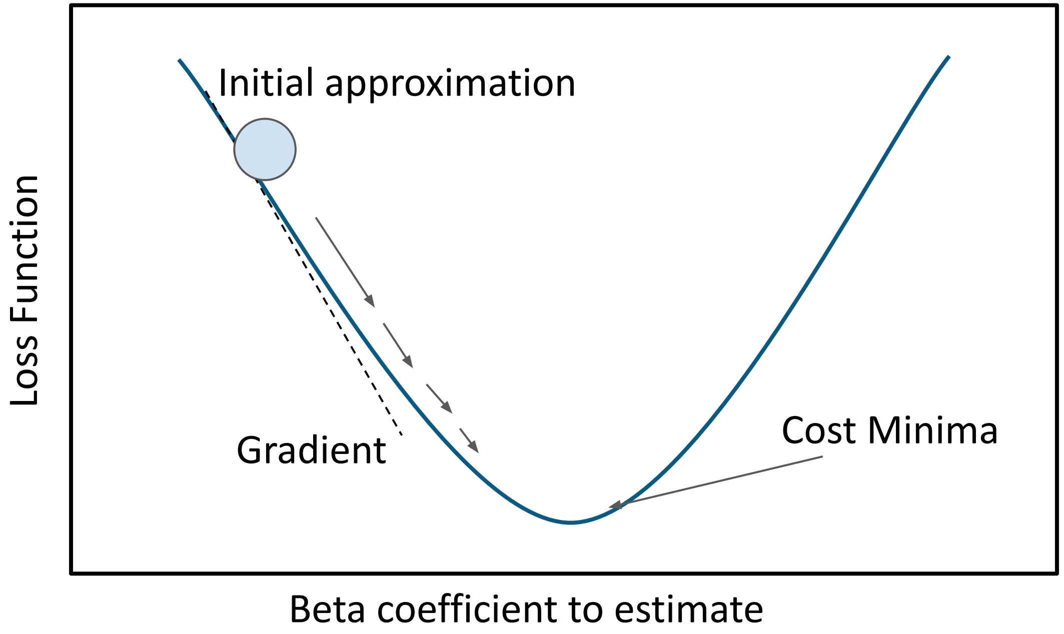
- $R^2$ assumes that all the independent variables explain the variation in the dependent variable
- For simple linear regression, the squared correlation between the response variable Y and independent variable X is the $R^2$ value

# Gradient descent

- Using the OLS method, we get the estimates of parameters of the linear regression model by minimizing the sum of the square of errors.
- The gradient descent is an optimization technique which finds the β parameters such that the error term is minimum.
- Computation speed for higher data dimension is more if parameter were to be obtained using the OLS method whereas the gradient descent does it faster.

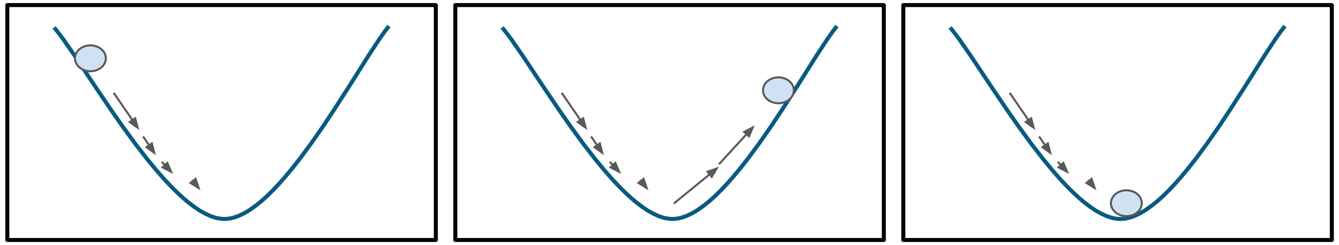
# Gradient descent

- An error function, also known as a loss function is used to calculate the cost associated with the deviation of observed data from predicted data.
- It is an iterative method which converges to the optimum solution.
- The estimates of the parameter are updated at every iteration.



# Gradient descent

- Consider a ball rolling down the slope as shown below
- Any position on the slope is the loss of the current values of the coefficients (cost)
- The bottom of the slope where the cost function is minimum
- The objective is to find lowest point in the cost function by continuously trying different values of the parameters
- Repeating this process numerous times, the best parameters are such that the cost is minimum



# Error analysis

Some of the Model evaluation metrics in regression are:

- Mean Squared Error (MSE):

$$MSE = 1/n \sum_{i=1}^{n} (yi - \hat{y}i)^2$$

- Root mean Squared Error(RMSE):

$$RMSE = \sqrt{1/n \sum_{i=1}^{n} (yi - \hat{y}i)^2}$$

- Mean absolute Error(MAE):

$$MAE = 1/n \sum_{i=1}^{n} |yi - \hat{y}i|$$

- Mean absolute Percentage Error(MAPE):

$$MAPE = 100(1/n) \sum_{i=1}^{n} |yi - \hat{y}i/yi|$$

# Assumptions of Linear Regression

- The dependent variable must be numeric.
- Linear relationship between dependent and independent variables.
- Predictors must not show multicollinearity.
- Independence of observations should exist (absence of autocorrelation).
- The error terms should be homoscedastic.
- The error terms must follow normal distribution.

# Advantages & disadvantages

- **Advantages**
  - Simple to implement
  - It’s easier to interpret the output coefficients.

- **Disadvantages**
  - It makes an assumption that the dependent and independent variables will have a linear relationship between them.
  - It is sensitive to outliers
  - Assumes independence between attributes.

# Summary

In this module we discussed:

- Regression analysis and what is linear regression?
- Regression use case
- Linear regression line
- Error term and calculating the error term
- Which line best fits our data?
- Methods to get the best fit line: ordinary least squares and gradient descent.
- Error analysis and assumptions of Linear Regression.
- Advantages and disadvantages.

# Thank you!

Happy Learning :)