Date: 24/08/2024

Session 5: Regression & Classification Model

Topic - SLR, MLR, PLR, Logistic Regression, DT, RF

Simple Linear Regression

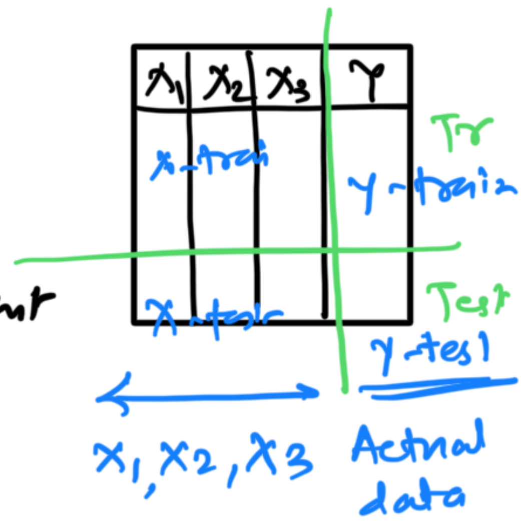
SLR Equation → $Y = \underbrace{m}_{\text{slope}}x + \underbrace{c}_{\text{Y intercept}}$

$$Y = \underbrace{\boxed{\beta_0}}_{\text{Y-intercept}} + \underbrace{\boxed{\beta_1}}_{\text{Coefficient}} \underbrace{X}_{\text{feature}}$$

X → 20

Dependent Variable

Independent Variable (feature)



Multiple Linear Regression

- Uses more than one Independent variable (x_1, x_2, x_3) to predict a dependent variable

Equation for MLR

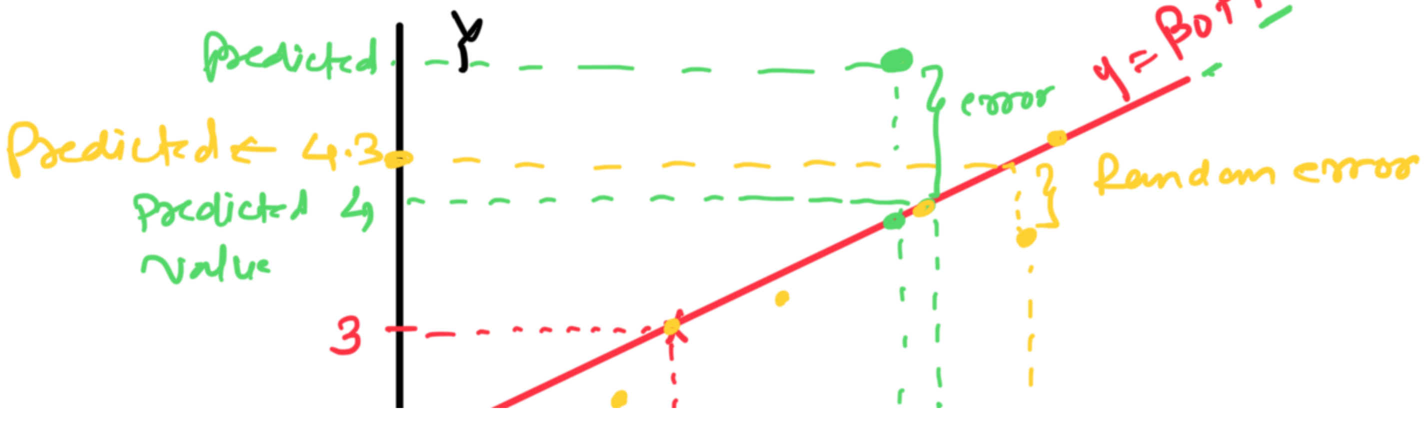
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$\beta_0 = 0.5$

$\beta_1 = 0.7$

$\beta_2 = 2.7$ → 2.7 → Most imp feature

$\beta_3 = 1.8$



Dependent Variable

Independent Variable

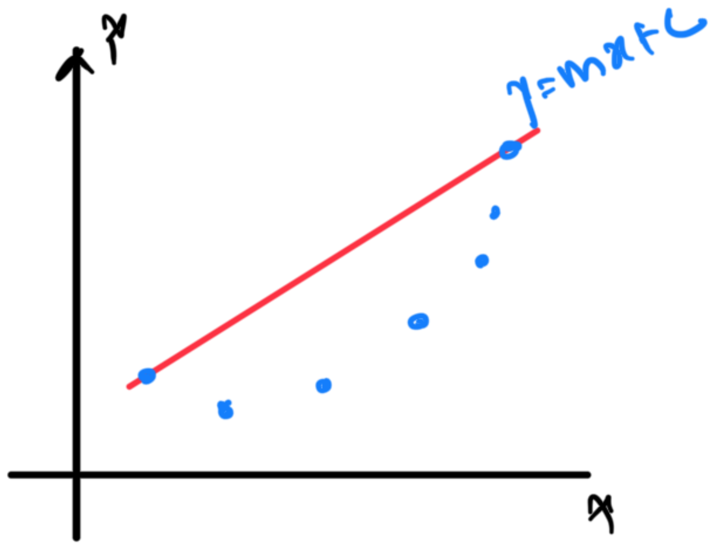
x_1, x_2, x_3

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots \epsilon$$

$\frac{\text{slop}}{\text{coefficient}}$

Error term

Polynomial Regression



SLR

Non-linear relationship

Polynomial Regression

SLR — $y = \beta_0 + \beta_1 X$

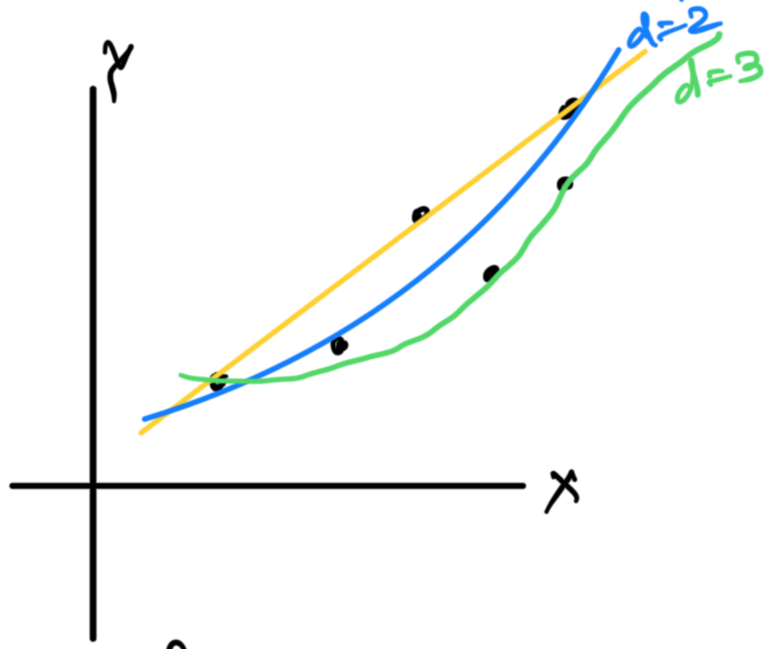
MLR — $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

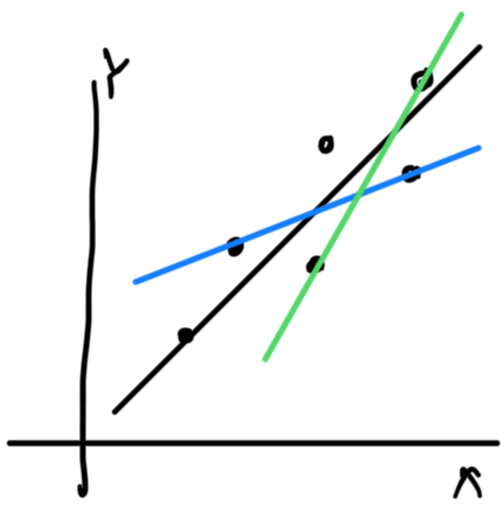
PLR — $y = \beta_0 x^0 + \beta_1 x_1^1 + \beta_2 x_2^2 + \beta_3 x_3^3 + \beta_4 x_4^4$

$x^0 = 1$ degree $= 1$

degree = 2

degree = h





Best fit line – Minimize the error between predicted value and actual value

$$\text{Error} \le \epsilon \quad \begin{cases} y = \text{Actual value (y\_test)} \\ \hat{y} = \text{Predicted value (Model output)} \end{cases}$$

– draw scatterplot & identify the best fit line.

1. R-Square Method

- statistical method that determines the goodness of best fit
- measure the strength of the relationship between the DV and IV
- Calculated 0 — 100 %

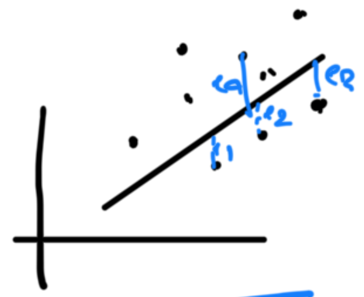
2. Residuals (Regression Error)

- Errors in regression represents the difference of the observed data point from the predicted data point in regression line.

Residuals = Actual $y\,(y_i)$ − Predicted $y\,(\hat{y}_i)$

3. Root Mean Square Error (RMSE)

- RMSE represents the Standard deviation of the residuals. It gives an estimate of the spread of observed data points across the predicted regression line



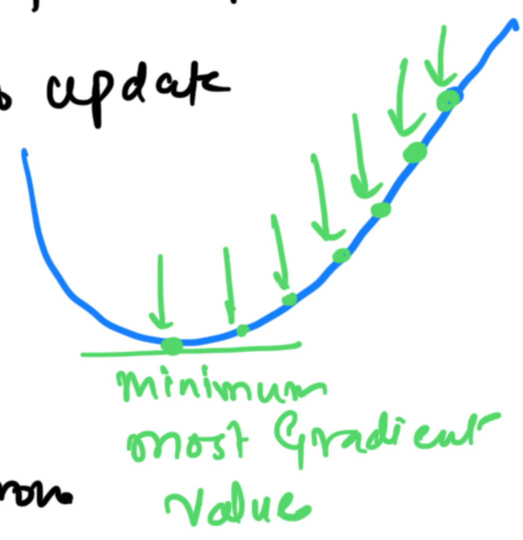
$$RMSE = \sqrt{\frac{e_1^2 + e^2 + e^3 + \dots}{n}}$$

To improve the performance of Regression

Technique — Gradient Descent

purpose - Minimize the MSE by calculating the gradient of the Cost f^n (Eq^n)

process - Iterative approach to update the coefficients to reduce the cost function



Model Performance → ↑ improve

- Goodness of fit - (Best fit line)
- R - Square → Measure the strength of the relationship bet^n DV and IV
- Values ranges from 0 — 100%
- Higher values indicates better model fit.
- It is also called as coefficient of determinat^n

Assumption of Linear Regression

1. **Linear Relationship** – Assume a linear relationship between features and the target.

2. **Multicollinearity** – Assume little or no multicollinearity between features

3. **Homoscedasticity** – Assume that the error term is the same for all values of the IV

4. **Normal distribution of Error Term** – Assume error term follow a normal distribution.

5. **No Autocorrelation**: Assume no correlation in error term.
   It reduces Model accuracy

Additional Regression Model

1. Ridge Model → L1 Regularization
2. Lasso Model → L2 Regularization
3. Elastic Net Model → (L1 + L2) Regularization

Regularization Techniques -

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

y_i → Actual y_i

$\hat{y}_i$ → Predicted y_i

Regularization

Regularisation Term

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{n}(y_i - \hat{y}_i) + \boxed{\boxed{\lambda}\sum_{j=1}^{n}\theta_j^2}$$

Regulization Parameter

Lasso Regression

$$J(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2 + \lambda\sum_{j=1}^{n}|w_j|$$

Ridge Regression

$$J(\theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2}_{\text{Loss function}} + \boxed{\lambda}\sum_{j=1}^{n}(|w_j|)^2$$