Dimensionality Reduction :-

Ex:-

Roll NO Name Mobile Number

Date - 20/10/22 UNIT-3 Model Development

Regression :-

→ Regression analysis is a ~~stat~~ statistical Method to model the relationship between a dependent and independent variable with one or more independent variables.

→ Regression analysis reveals average relationship between two variables and this makes possible estimation / Prediction.

→ Regression used for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.

→ Ex:- Prediction of rain using temp. and other factor

Determining Market trends

Prediction of road accidents due to rash driving.

Simple Linear Regression :-

It is a type of Regression algorithms that models the relationship between a dependent variable and a single independent variable.

↳ The key point of Simple Linear Regression is that the dependent variable must be a continuous/ real value.

↳ ex:- Model relationship between income and expenditure.

Forecasting new observation.

Least Square Method

$$\Sigma (Y - Y_c)^2$$

Regression equation of y on x. $\boxed{Y = a + bx}$

formula:- $\boxed{\begin{aligned} \Sigma y &= Na + b\Sigma x \\ \Sigma xy &= a\Sigma x + b\Sigma x^2 \end{aligned}}$

N = no. of observation.

Example-1

(in) (X) over	(d) (Y) run	xy	x^2
1	2	2	1
2	5	10	4
3	8	24	9
4	12	48	16
5	14	70	25
6	18	108	36
$\Sigma x = 21$	$\Sigma y = 59$	$\Sigma xy = 262$	$\Sigma x^2 = 91$

$\Sigma y = na + b\Sigma x$

$\Rightarrow 59 = 6 \times a + b \times 21$

$\Rightarrow 59 = 6a + 21b$ — (1)

Multiply 7 on eq (1)

$7(6a + 21b) = 7 \times 59$

$\Rightarrow 42a + 147b = 413$ — (3)

$\Sigma xy = a\Sigma x + b\Sigma x^2$

$\Rightarrow 262 = a \times 21 + b \times 91$

$\Rightarrow 262 = 21a + 91b$ — (2)

multiply 2 on eqn (2)

$2(21a + 91b) = 2 \times 262$

$\Rightarrow 42a + 182b = 524$ — (4)

Subtract eqn (3) from eqn (4), we get

$42a + 182b - 42a - 147b = 524 - 413$

$\Rightarrow 35b = 111$

$\Rightarrow b = \frac{111}{35}$

$\Rightarrow \boxed{b = 3.17}$

$y = a + bx$

$\Rightarrow \boxed{y = -1.26 + 3.17x}$

put this on eqn (1)

$6a + 21b = 59$

$\Rightarrow 6 \times a + 21 \times 3.17 = 59$

$\Rightarrow 6a + 66.57 = 59$

$\Rightarrow 6a = -7.57$

$\Rightarrow a = -\frac{7.57}{6} = -1.26.$

Q.

(x) over	(y) Runs	xy	y^2
1	2	2	4
2	5	10	25
3	8	24	64
4	12	48	144
5	14	90	196
6	18	108	324
21	59	282	757

$$\Sigma x = na + b\Sigma y$$

$$\Rightarrow 21 = 6a + 59b \quad \text{---} (1)$$

$$\Sigma xy = a\Sigma y + b\Sigma y^2$$

$$\Rightarrow 282 = 59a + 757b \quad \text{---} (2)$$

$$59(6a + 59b) = 21 \times 59$$

$$\Rightarrow 354a + 3481b = 1239 \quad \text{---} (3)$$

$$6(59a + 757b) = 6 \times 282$$

$$\Rightarrow 354a + 4542b = 1692 \quad \text{---} (4)$$

$35a + 4547b - 34.95 - 3481b = 1572 - 1239$

$\Rightarrow 1066b = 333$

$\Rightarrow b = \frac{333}{1066}$

$\Rightarrow \boxed{b = 0.3}$

$6a + 59b = 21$

$\Rightarrow 6a + 59 \times 0.31 = 21$

$\Rightarrow 6a + 18.29 = 21$

$\Rightarrow 6a = 2.71$

$\Rightarrow a = \frac{2.71}{6}$

$\Rightarrow \boxed{a = 0.45}$

$X = a + by$

$\Rightarrow X = 0.41 + 0.31y$

$X = 0.41 + 0.31 \times 36$

$\Rightarrow X = 0.417$

$\Rightarrow \boxed{X = 11.57}$

Utility of Regression :-

1) Nature of Relationship
2) Estimation of Relationship.
3) Prediction
4) Useful in economic and Business Research.

Multiple Regression :-

→ It considering the values of the available multiple independent variables and predicting the value of one dependent variables.

→ The variables considered for the model should be

$\Sigma y =$

Date - 27/10/22

Q) ~~Structure~~

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$\Sigma y x_1 = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Sigma y x_2 = a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2.$$

} Multiple Regression

Stu-Name	Marks	Liveclass	Book
A	8	3	4
B	9	4	5
C	7	3	3
D	10	5	5
E	6	2	3

Name	Y	x_1	x_2	$Y x x_1$	$Y x x_2$	$x_1 x x_2$	Y^2	x_1^2	x_2^2
A	8	3	4	24	32	12	64	9	16
B	9	4	5	36	45	20	81	16	25
C	7	3	3	21	21	9	49	9	9
D	10	5	5	50	50	25	100	25	25
E	6	2	3	12	18	6	36	4	9
	$\Sigma=40$	17		$\Sigma=143$	$\Sigma=166$	$\Sigma=72$	$\Sigma=63$		$\Sigma=84$

$$\Sigma y = na + b_1 \Sigma x_1 + b_2 \Sigma x_2$$

$$40 = 5a + 17b_1 + 20b_2 \quad \text{——} (1)$$

$$\Sigma y x_1 = a \Sigma x_1 + b_1 \Sigma x_1^2 + b_2 \Sigma x_1 x_2$$

$$\Rightarrow 143 = 17a + 63b_1 + 72b_2 \quad \text{——} (2)$$

$$\Sigma y x_2 = a \Sigma x_2 + b_1 \Sigma x_1 x_2 + b_2 \Sigma x_2^2$$

$$\Rightarrow 166 = 20a + 72b_1 + 84b_2 \quad \text{——} (3)$$

$a = 2.52$

$b_1 = 0.833$

$b_2 = 0.66$

$$\boxed{Y = 2.52 + 0.833 x_1 + 0.66 x_2}$$

Regression Error analysis :-

a)

X	Y
6	7
2	9
8	5
4	6
10	8

y on x : $y_c = a + bx$

x on y : $x_c = a + by$

X	Y	X_c	Y_c	$X - X_c$	$(X-X_c)^2$	$Y - Y_c$	$(Y-Y_c)^2$
6	7	6	7	0	0	0	0
2	9	7.6	4.8	-5.6	31.36	4.2	17.64
8	5	7.2	6.7	0.8	0.64	-1.7	2.89
4	6	6.6	7.3	-2.6	6.76	-1.3	1.69
10	8	5.4	6.4	4.6	21.16	1.6	2.56
					59.92		24.78

Reg eqn y on x $\quad y = 7.9 - 0.15x$

eq eqn x on y $\quad x = 10.2 - 0.6y$

$y_c = 7.9 - 0.15x$

$\Rightarrow y_c = 7.9 - 0.15 \times 6$

$\Rightarrow y_c = 7$

$$S_{ny} = \sqrt{\frac{\Sigma (x - x_c)^2}{N}}$$

$= \sqrt{11.984}$

$= 3.46$

$$S_{yx} = \sqrt{\frac{\Sigma (y - y_c)^2}{n}}$$

$$= \sqrt{\frac{24.78}{5}}$$

$$= \sqrt{4.946}$$

$$= 2.223$$

Date - 28/10/22

Matrix formulation for Multiple Regression Model :-

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

Q) find the coefficient of Regression in matrix form from the data below

y	9	10	13	14	16
x_1	1	3	4	6	7
x_2	10	14	15	18	20

$\beta_0 = 1$

$$y = \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix}$$

$$\beta = (X^T X)^{-1} X^T y = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 1 & 1 & 10 \\ 1 & 3 & 14 \\ 1 & 4 & 15 \\ 1 & 6 & 18 \\ 1 & 7 & 20 \end{bmatrix}$$

$$= \begin{bmatrix} 1+1+1+1+1 & 1+3+4+6+7 & 10+14+15+18+20 \\ 1+3+4+6+7 & 1+9+16+36+49 & 10+42+60+108+140 \\ 10+14+15+18+20 & 10+42+60+108+140 & 100+196+225+324+400 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 21 & 77 \\ 21 & 111 & 360 \\ 77 & 360 & 1245 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 3 & 4 & 6 & 7 \\ 10 & 14 & 15 & 18 & 20 \end{bmatrix} \begin{bmatrix} 9 \\ 10 \\ 13 \\ 14 \\ 16 \end{bmatrix}$$

$$= \begin{bmatrix} 9+10+13+14+16 \\ 9+30+52+84+112 \\ 90+140+195+252+320 \end{bmatrix}$$

$$= \begin{bmatrix} 162 \\ 287 \\ 997 \end{bmatrix}$$

$$A^{-1} = \frac{1}{|A|} \cdot \text{Adj } A$$

$$\det(X^T X) = \begin{bmatrix} 5 & 21 & 77 \\ 21 & 111 & 360 \\ 77 & 360 & 1245 \end{bmatrix}$$

$$= 5\left[111 \times 1245 + 360 \times 360\right] - 21\left[\right.$$

$$= 5\left(138375 + 129600\right) - 21\left(26145 + 27720\right) + 77\left(75600 - 79541\right)$$

~~714375 = 138195 + 1240239~~

$|X| = \det(X^T X)$

adj of $(X^T X)$ ⇒ co efficient of $(X^T X)$

$$\begin{bmatrix} \begin{bmatrix} 111 & 360 \\ 360 & 1245 \end{bmatrix} - \begin{bmatrix} 21 & 360 \\ 77 & 1245 \end{bmatrix} + \begin{bmatrix} 21 & 111 \\ 77 & 360 \end{bmatrix} \\ \begin{bmatrix} 111 & 360 \\ 360 & 1245 \end{bmatrix} + \begin{bmatrix} 21 & 360 \\ 77 & 1245 \end{bmatrix} - \begin{bmatrix} 21 & 111 \\ 77 & 360 \end{bmatrix} \\ \begin{bmatrix} 111 & 360 \\ 360 & 1248 \end{bmatrix} + \begin{bmatrix} 21 & 360 \\ 77 & 360 \end{bmatrix} - \begin{bmatrix} 5 & 21 \\ 21 & 111 \end{bmatrix} \end{bmatrix}$$

$$\begin{bmatrix} (138195 - 129600) - (26145 - 27720) + (7560 - 8547 \\ -(138195 - 129600) + (26145 - 27720) - (7560 - 854 \end{bmatrix}$$

$$= \begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 298 & -183 \\ -987 & -183 & 114 \end{bmatrix} = \text{adj } A$$

$$B = (X^T X)^{-1} X^T Y = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 8595 & 1575 & -987 \\ 1575 & 296 & -183 \\ -987 & -185 & 114 \end{bmatrix} * \begin{bmatrix} 63 \\ 287 \\ 947 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 532890 + 452025 - 984039 \\ 97650 + 84952 - 182451 \\ -61194 - 52521 + 113658 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \frac{1}{51} \begin{bmatrix} 876 \\ 151 \\ -57 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 17.17 \\ 2.96 \\ -1.11 \end{bmatrix} = \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

$$\boxed{y_i = 17.17 + 2.96 x_1 - 1.11 x_2}$$

Date- 29/10/22

Model Evalution and Selection :-

- Metrics for performance Evalution
- Methods for performance Evalution
- Methods for Model comparison
- Model selection.

Confusion Matrix :-

A table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known

Ex

Actual class	Predicted class: yes	Predicted class: no	
yes	40 TP	20 FN	60 (P)
no	10 FP	90 TN	100 (N)
			T = P + N = 160

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Date- 31/10/22

True positive (TP)

↳ Positive tupples that were correctly level by the classified.

↳ negative tupples that were correctly leveled by the classifien.

False positive (FP)

↳ negative tupples that were incorrectly level by the positive.

↳ Positive tupples that were mislebeled as negative.

Performance matrices :—

1. Accuracy :-

↳ It is also called as Recognition Rate.

↳ Percentage of test set tupples that are correctly classified.

↳ $\frac{TP + TN}{P + N}$

2. Error Rate :-

↳ 1 - Accuracy

↳ It is also called as Misclassification Rate

↳ $\frac{FP+FN}{P+N} = \frac{10+20}{160} = 0.19$

3. Calculate sensitivity :—

↳ It also refers to the TP recognition rate.

↳ The proportion of positive tuppple that are currently identified.

↳ sensitivity = $\frac{TP}{P}$

↳ specificity :—

↳ It refers to the TN recognition rate.

↳ The proportion of negative tupples that are currently identified.

↳ specificity = $\frac{TN}{N}$

5. Precision :—

↳ It is called measure of exactness. It means what % of tupples is truely positive.

↳ It lies between 0 to 1.

↳ precision = $\frac{TP}{TP+fp}$

6. Recall :-

↳ It is a measure of completeness.

↳ what % of positive tupples are level as Positive.

↳ $\frac{TP}{TP + FN}$

7. F-Score / F-Measure :-

$$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{1.056}{1.46} = 0.72$$

↳ It is the harmonic mean precesion & Recall.

↳ It gives same weightage to precesion & Recall.

↳ It is also called as F_1 score.

Date - 05/11/22 Measures of insample Evalution :-

Loss function

A function that calculate loss of 1 Data point is known as Loss function.

↳ $(y_i - \hat{y}_i)^2$

y_i = Actual value

$\hat{y}_i$ = Predicted value.

cost function

A function that calculate Loss of entire data is known as Cost function.

↳ Mean sq. error = $\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

n = sample size.

Mean Absolute Error

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

Lower the MAE

higher the accuracy

Arithmatic avg of absolute Errors.

Mean Bias Error

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

A positive Bias Means the error from the data is over estimated

A negative bias means the error is underesti.

Relative Absolute Error

$$RAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}|}$$

where, $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \rightarrow$ Mean of actual values.

Result - 0 - 1

$0 \rightarrow$ good model.

Mean absolute percentage error

$$MAPE = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right| * 100\%$$

Root Mean square error

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

RMSE = 0 → Perfect Model

Relative squared error :-

$$RSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ → Mean of actual value.

Relative Root Mean squared error :-

$$RRMSE = \sqrt{\frac{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (\hat{y}_i)^2}}$$

Excellent when RRMSE < 10%.
good 10% to 20%.
Fair 20 to 30%
Poor > 30%.

Root Mean squared Logarithmic Error :-

$$RMSLE = \sqrt{\log(y_i + 1) - \log(\hat{y}_i + 1)^2}$$

Coefficient of Determination (R squared) :-

- It is used to analyze how differences one variable can be explained by a difference in a second variable.
- Its Ranges from 0 to 1 (0% to 100%)
- If $R^2 = 0$, then dependent variable can't be predicted from the independent variable.
- If $R^2 = 1$, then the dependent variable can be predicted from the independent variable.
- If R^2 = between 0 and 1, then the dependent variable can be predicted.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Formula-2

$$R^2 = 1 - (RSS/TSS)$$

RSS ↓ Residual sum square
TSS ↓ Total sum squa.

Date - 07/11/22

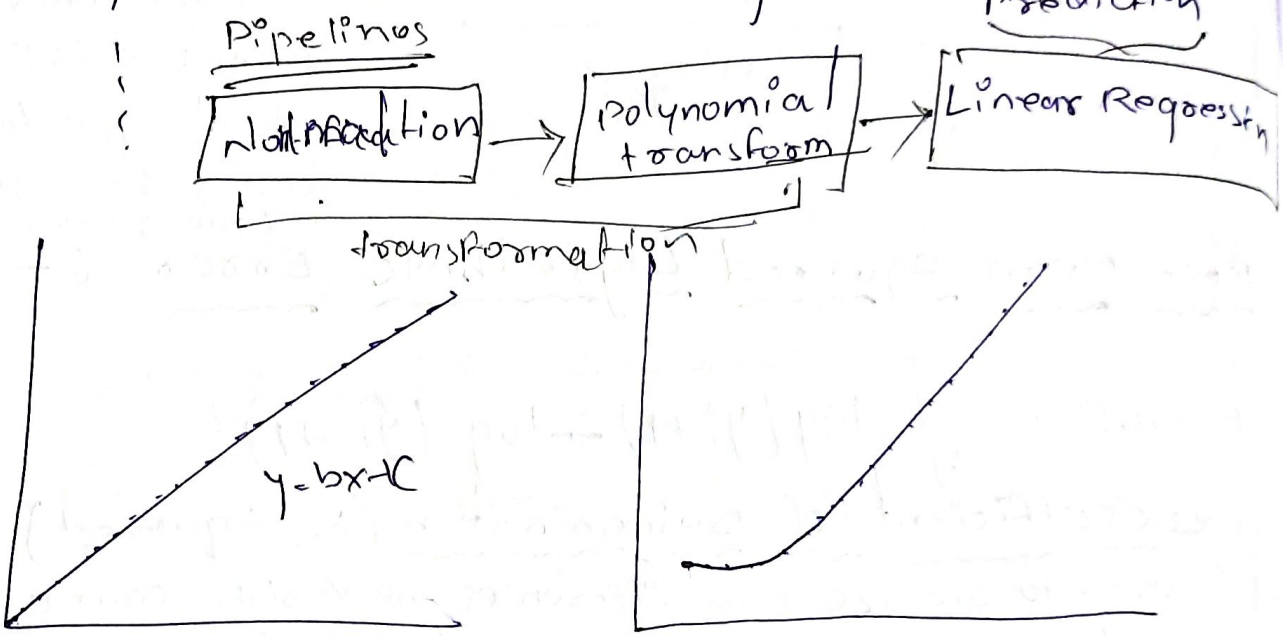
Polynomial Regression :-

↳ Polynomial Regression is a Regression algorithm that models a relationship bet'n a dependent(y) and independent variable(x) as n^{th} degree polynomial

if y = constant - 0-degree.

$y = bx^1 + c$ - 1-degree

$y = ax^2 + bx + c$ - 2-degree polynomial



$y = b_0 + b_1 x_1 + b_2 x_1^2 + b_3 x_1^3 \ldots b_n$

↳ Polynomial Regression fits a non linear relationship.

→ Here the data points are arranged non-linear fashion.

If we apply a linear model or simple RG in a linear data set, then it provides us a good result but if we apply the same model without any modification on a non linear data set, then it will produce a drastic output. due to which loss function will increase, the error rate will be high and accuracy will be decreased. so for such cases we need the polynomial regression model.

Example :-

x	y	$y_i x_i$		x^2	x^3	x^4
3	9.5	7.5	22.5	9	27	
4	3.2	12.8	51.2	16	64	
5	3.8	19	95	25	125	
6	6.5	39	234	36	216	
7	19.5	80.5	563.5	49	343	
25	22	158.8	966.5	135	775	4659

$$\Sigma y_i = n a_0 + a_1 \Sigma x_i + a_2 (\Sigma x_i^2)$$

$$\Sigma y_i x_i = a_0 (\Sigma x_i) + a_1 (\Sigma x_i^2) + a_2 (\Sigma x_i^3)$$

$$\Sigma y_i x_i^2 = a_0 (x_i^2) + a_1 (\Sigma x_i^3) + a_2 (\Sigma x_i^4)$$

$$\Rightarrow 22 = 5 a_0 + 25 a_1 + 135 a_2$$

$$\Rightarrow 158.5 = 25 a_0 + 135 a_1 + 775 a_2$$

$$\Rightarrow 966.2 = 135 a_0 + 775 a_1 + 4659 a_2$$

$a_0 = 12.42$

$a_1 = -5.51$

$a_2 = 0.76$

$y = a_0 + a_1 x + a_2 x^2 - - - a_n x_i^2$

$y = 12.42 - 5.51 x + 0.76 x^2$

Date- 10/11/23

Residual plots for regression model validation ?

A Residual is a measure of how far away a point is vertically from the regression line. Simply, it is the error between a predicted value and the observed actual value.

↳ Residual (ϵ) = $y - \hat{y}$

Residual plot

A typical residual plot has the residual values on the y-axis and the independent variable on the X-axis.

Residual plot Analysis :-

The most important assumption of a linear regress. model is that the errors are independent and normally distributed.

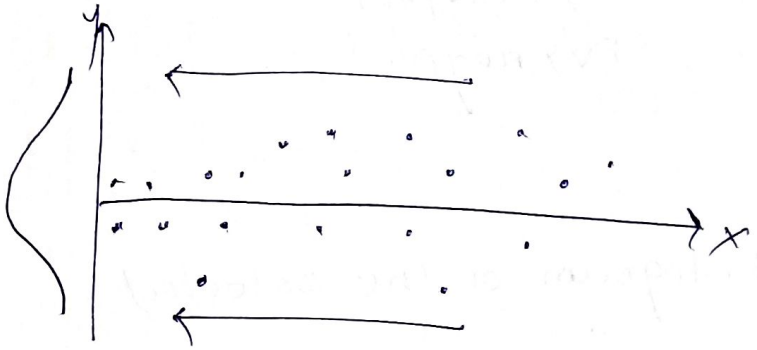
Response = Deterministic + Stochastic

Characteristics of Good Residual plots :-

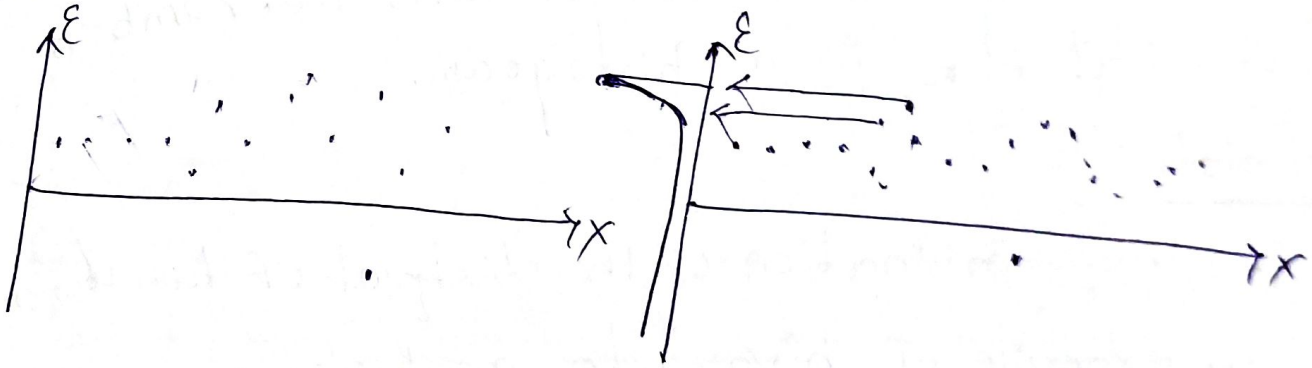
↳ It has a high density of point close to the origin and a low density of point away from the origin.

↳ It is symmetoic about the origin.

Good Residual plot



↳ Residual errors are approximately distributed in the same manner.



Bad Residual plot

Distribution plot :-

~~These plot~~

↳ seaborn is a python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics.

↳ These plot help us to visualize the distribution of data. we can use these plots to understand the mean, median, range, variance, deviation, etc of the data.

↳ It is of 4 types of plots :- i) joinplot
ii) distplot
iii) Pairsplot
iv) rugplot

a) Distplot

↳ Dist plot gives us the histogram of the selected continuous variable.

↳ It is an example of a univariate analysis.

↳ We can change the number of bins i.e. number of vertical bars in a histogram.

b) Join plot

→ It is the combination of the distplot of two vari

→ It is an example of bivariate analysis.

Pair plot :-

↳ It takes all the numerical attributes of the data and plot pairwise scatter plot for two different variables and histograms from the same variables.

Rugplot :-

↳ It draw a dash mark instead of a uniform distribution as in distplot.

↳ It is an example of a univariate analysis.

~~Pipelines~~ :-

Prediction and Decision making :-

↳ Do the predicted values make sense

↳ Visualization

↳ numerical measures for evaluation.

↳ comparing Models

Date-12/11/22

↳ Mean square Error for a multiple Linear Regression model will be smaller than the Mean square Error for a simple Linear Regression model, since the errors of the data will decrease when more variables are included in the model.

↳ Polynomial regression will also have a smaller Mean Square Error than the linear regular regression.