Statistical Modeling

40% F. Exam
20% Rote
20% Project
20% Mid Course

Rajan Sir :-

→ Teaches beyond stats. and use the statistical ideas to build models (50% of data science).

Diff Equation
Integration
1st & Second order D.E
Vectors
Matrix & determinants (Important)

} Basic topics for this module.

Co-relation * Coefficient.

To know the relation between two variable we use "Co-variance"

$$\text{Variance}(\sigma^2) = \frac{1}{n-1} \sum_{i=1}^{n} (x-\bar{x})^2 \quad \text{(for 1 variable)}$$

$$\text{Co-variance} = S^2_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x-\bar{x})(y-\bar{y})$$

→ Co-variance ~~is~~ There will be a positive relationship if both x & y increases or decreases together ($S^2_{xy} = +ve$)

→ There will be negative or opposite relationship if both x & y changes oppositely ($S^2_{xy} = -ve$)

→ There is no relationship if x changes, then y does not change. ($S^2_{xy} = 0$)

(*) The problem is the covariance value b/w two variable 'x' & 'y' changes if the units of 'x' & 'y' are changed, so we cannot predict how strong is the relationship b/w the two variables with diff values of Co-variance

Therefore,

we use standard score for each variab and calculation the co-variance.

This co-variance in terms of z-sc is called as Co-relation Coefficient (γ).

$$\gamma = \frac{1}{n-1} \sum_{i=1}^{n} (z_x)(z_y) \qquad -1 \geq \gamma \geq 1$$

$\gamma \approx 1 \longrightarrow$ Strong positive relationship

$\gamma \approx -1 \longrightarrow$ Strong negative relationship

$\gamma \approx 0 \longrightarrow$ no relationship

Simple linear Regression :-

* The variable which we want to predict on, is called Predicted variable / Outcome variable / Target variable.

* The variables which are used to predict the Predicted variable is called Predictor variable / Regressor variable.

$$Y = mx + c$$

Since we use only one variable to predict the Predicted variable
∴ Simple.

The ↑ value of m & c will only results in the linear change i.e.
∴ linear

modification of

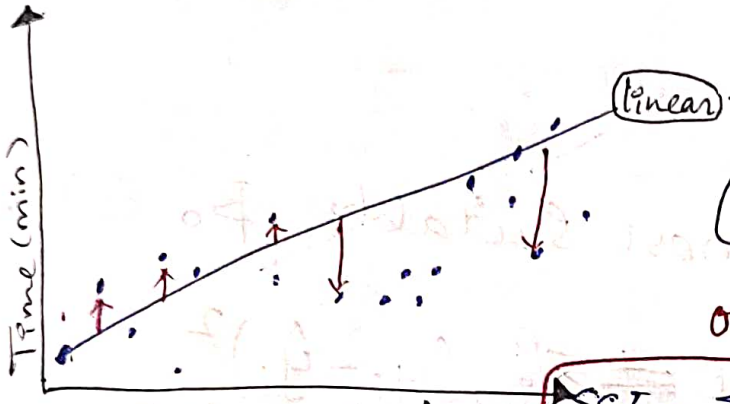
The performance of the model line when compared to simple mean line

$\hat{Y}$ → predicted value

$\bar{Y}$ → mean value

R^2 → Says how better is the model.

SST
Sum of Squared Total errors



SSE (Residual errors)
Sum of squared Errors of predtn
or Sum of Squared Residual Errors

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

(*) The best line would be the line with the least SSE this can be achived by systematically changing m & C

In Data Science field we ~~used~~ use β_0 & β_1 in place of m & C.

$$\Rightarrow \hat{y} = \beta_0 + \beta_1 x \longrightarrow \text{prediction}$$

$$y = \beta_0 + \beta_1 x + \varepsilon \longrightarrow \text{Data model measured value}$$

ε → Errors.

We assume that the errors are coming from Normal distribution with mean zero, with some S.dev

(*) The error distribution is independent of the value of x; this is called Homoscedasticity.

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

term = ε

$$\beta_1 \text{ (slope)} = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2} = \frac{S_{xy}^2}{S_x^2} = \frac{\gamma S_y}{S_x}$$

S_{xy}^2 : interms of covariance

$\frac{\gamma S_y}{S_x}$: interms of correlation

$$\beta_0 = \bar{y} = \beta_0 + \beta_1 \bar{x}$$

How to get the most suitable β_0 & β_1 value

$$\Rightarrow \text{since } SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\Rightarrow SSE = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x))^2$$

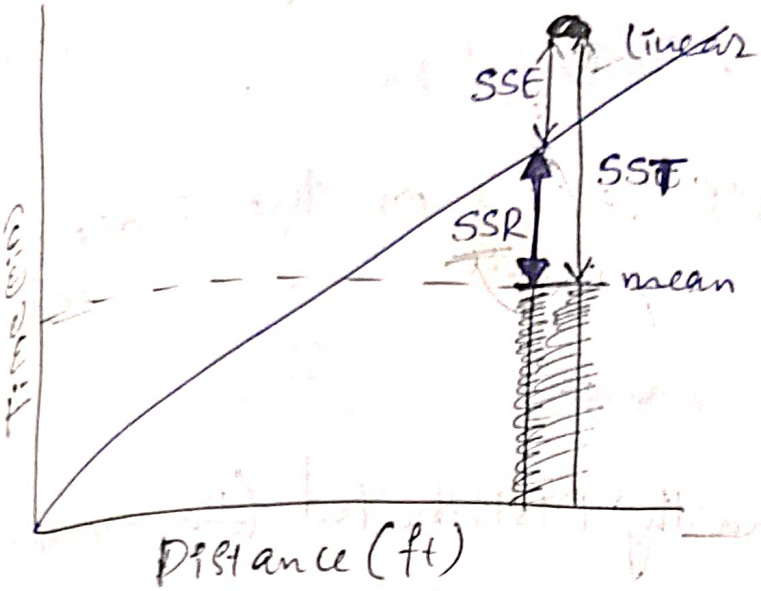
$$\frac{d(SSE)}{d\beta_0} = 0 \quad ; \quad \frac{d(SSE)}{d\beta_1} = 0$$

diff SSE w.r.t β_0 ; diff SSE w.r.t β_1

(*) β_0 has no importance if the X doesnot have '0' as the data in it.

(make y intercept = 0 during this situation while drawing trend line)

SSR :- Sum of Squared Regressions Errors



$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

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

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

$$SST = SSE + SSR$$

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

$$\frac{SSR}{SST} = 1 - \frac{SSE}{SST} = R^2$$

↳ What proportion of errors that has been rectified by the model

R^2 Close to 1 = Good model

R^2 → How much of the total error is explained by the model

SSR → Explained errors

SSE → Unexplained errors.

∴ Total err. = Ex. err + Unex. err

$$SST = Ex.err + Unex.err$$

To quantify the model is good or not

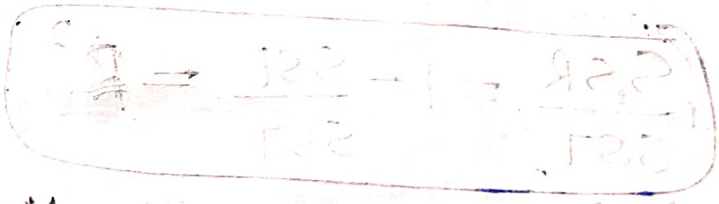
(1) R^2 value.

(2) Hypothesis testing on the slope of the model

Assumptions

* Errors are Normally Distributed. (~~Homoscedasticity~~)
* ~~Errors~~ Every measurment is independent of every other measurment
* Homoscedasticity
* No Influential outlayers.

'R' Studio :-



dtLR ← lm(dt$time ~ dt$distance)

↳ Linear model.
built time regressed upon distance.

NOW,
dtLR shows the Co-efficients :-
i.e β_0 & β_1 values

Summary(dtLR) :-

→ Residuals :- Test whether it is skewed or not

→ Co-efficients :- Estimate | std err | t-value | Pvalue

→ Residual std error :-
6.909 on 23 dof

→ Multiple R-squared :-

→ F-statistics

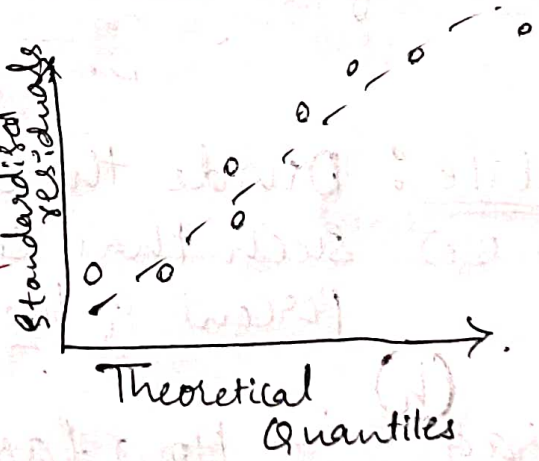
with this we accept or reject null hypothesis

plot(dtLR) :-

→ Residuals vs Fitted (This shows if there is a relationship in the ~~wast~~ diff measurment. if relationship exists, then the ② assumptions will fail

→ Normal Q-Q

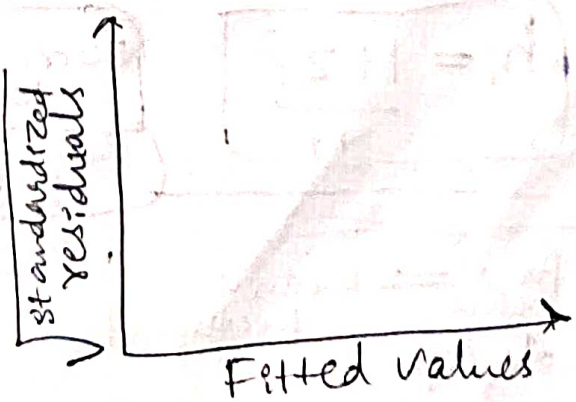
The residues must be close to Normal Q-Q line



→ Scale-Location

(Checks for Homoscedasticity)

if H.sce exists then value of errors must be same.



→ Residuals Vs Leverage

(*) Any data point that has an Cook's-d value > 1 is considered as influential outlayers.
└→ we must remove that data point and rebuilt the model.

(*) Any data point that has an Cook's-d value b/w 0.5 & 1, those are called suspected outlayers → we sud check by removing this, if there any change in the model.

Quartile: Divide the entire data into (4 - sub-ranges) such that each sub-range contains 25% of the data.

Quantile: (k-quantile) Divide the entire data into k sub-ranges such that each subrange contains 100/k percent of the data.

Leverage (h): How far away is the predictor variable from the mean of predictor variable

$$h_i = \frac{1 + z_i^2}{n}$$

Sum of leverage = mean of leverages × Sample size

Sum of leverage ⇔ Number of parameters

$$\therefore \frac{\text{No of parameters}}{\text{Sample size}} = \text{mean of leverages}$$

Example

$\frac{h}{mean(h)} > 2$; considered as outlayer

* Therefore leverage tells how far the x value far from mean of x

∴ Farther the x from mean(x), Higher the leverage

* Therefore if the leverage values or $\frac{h}{mean(h)}$ values for each x value is greater than 2, then we consider that point as influential outlayer

Studentize Residual :-

$$std\ res_i = \frac{e_i}{\sqrt{MSE(1-h_i)}}; \quad 2 < std\ res_i < -2$$

Std residual tells, how the errors of predictions (y-value) influenence.

Combining both of Leverage and Std residual we get (Cook's D)

→ Measures overall influence of an observation by seeing the regression co-efficients.

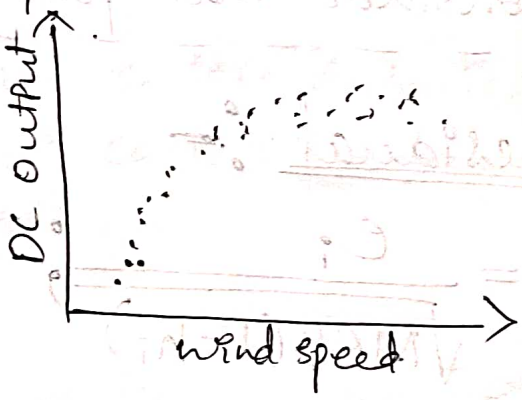
$$D_i = \frac{1}{p}(std\ res_i)^2\left(\frac{h_i}{1-h_i}\right)$$

⟹ Tukey's four Quadrant approach

(*) If the assumed linear relationship is actually non-linear for the model, then we will do non-linear transformation i.e $x^2, x^3, \log(x)$ $y^2, y^3, \log(y)$, etc, to get a linear relationship.

↳ How to know that there is a non-linearity?

→ The Scatter Plot will be having a pattern as shown below.

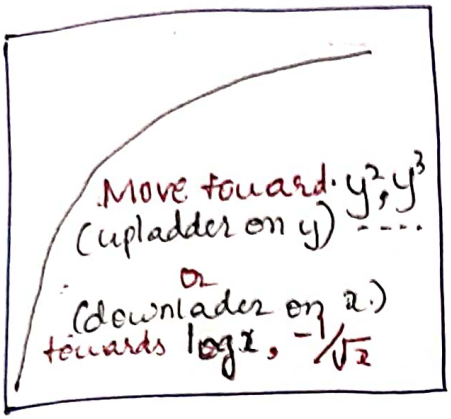


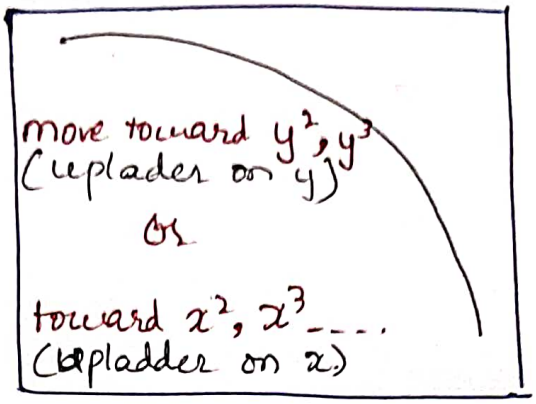
Tukey's ladder

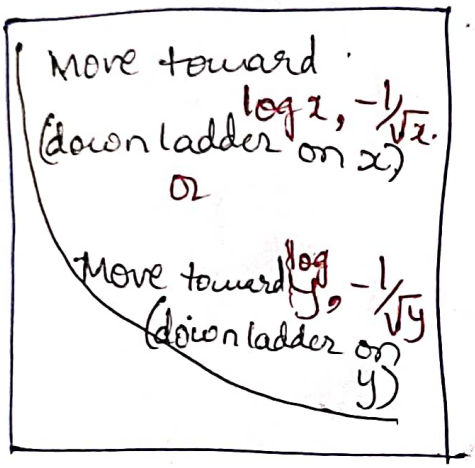
Ladder for x.		
Up ladder	Neutral	Down ladder
... x^4, x^3, x^2, x	$\sqrt{x}, x, \log x$	$\frac{-1}{\sqrt{x}}, \frac{-1}{x}, \frac{-1}{x^3}, \ldots -\log x$

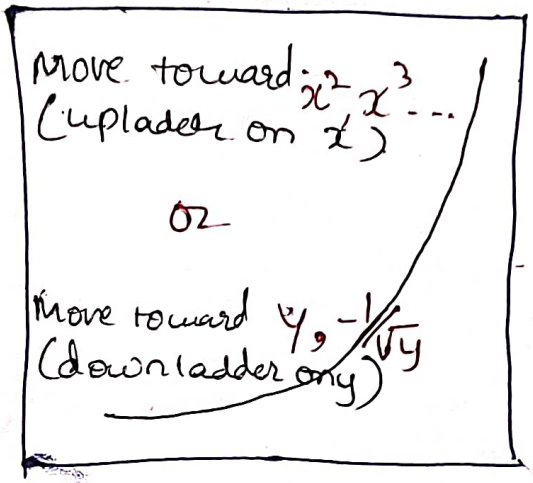
Ladder for y.		
up ladder	Neutral	Down ladder
$\ldots, y^4, y^3, y^2, y$	$\sqrt{y}, y, \log y$	$\frac{-1}{\sqrt{y}}, \frac{-1}{y}, \frac{-1}{y^2}, \ldots \log y$

(*) First do transformation of x.


Move toward y^2, y^3 ....
(upladder on y)
Or
(downladder on x)
towards $\log x, -1/\sqrt{x}$


move toward y^2, y^3
(upladder on y)
Or
toward x^2, x^3 ....
(upladder on x)


Move toward
$\log x, -1/\sqrt{x}$
(downladder on x)
Or
Move toward $\log y, -1/\sqrt{y}$
(downladder on y)


Move toward x^2, x^3 ...
(upladder on x)
Or
Move toward $y, -1/\sqrt{y}$
(downladder on y)

Multiple Linear Regression :-

(*) When there is more than one Regressor/X-variable.

↳ Some or all variables could be Categorical

↳ The variables could be interacting among themselves

↳ The variable could dependent/ on related each other

→ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots\dots \beta_n x_n + \varepsilon$,

→ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \dots + \beta_n x_n + \varepsilon$

↳ Interaction term

For only x_1, x_2 variable we have only one

→ $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$

(+) Always start with model putting all Variables and only pairwise interaction terms.

→ To address that the variables are related to each other, we must look at the co-relation b/w them.

if some of the variables be related to each other.

for eg :- x_1 is related to x_2, x_4, x_5 & x_7

These variables are multicollinear

* Multicollinearity :-

$$x_1 = \gamma_1 + \gamma_{12} x_2 + \gamma_{13} x_3 + \cdots + \gamma_{1n} x_n + \varepsilon$$

R_1^2

$$VIF = \frac{1}{1 - R_1^2}$$

$$x_K = \gamma_K + \gamma_{K1} x_1 + \gamma_{K2} x_2 + \cdots + \gamma_{KK-1} x_{K-1} + \gamma_{KK+1} x_{K+1} + \cdots + \gamma_{Kn} x_n + \varepsilon.$$

$R_{K.}^2$

$$VIF = \frac{1}{1 - R_K^2}$$

* Variance Inflation Factor :-

$$VIF = \frac{1}{1 - R_K^2}$$

R^2	$VIF = \frac{1}{1-R^2}$
0.5	2
0.75	4
0.8	5
0.9	10
0.95	20
0.99	100

The Staindard VIF is greater than 4

$VIF > 4$

If theses are the equation., i.e.:-

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \text{-----} + \beta_n x_n + \varepsilon$$

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{----} + \beta_n x_n.$$

$$SSE = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2 \Rightarrow \sum_{i=0}^{n} (y_i - (\beta_0 + \beta_1 x_1 + \text{----} + \beta_n x_n)^2$$

$$\frac{d(SSE)}{d\beta_k} = 0$$

We will get k equations :-
This is too much of calculus.
Since we have large number of variables.

∴ we move to Numerical Technique

Numerical technique :-

- It involves the assumption of values of $\beta_0, \beta_1, \beta_2 \ldots$ to calculate SSE.
- Now, Change all the β-values by some amount and see that SSE changes or not
- If it decreases, keep Changing the value until it reach to minimum value and starts to increase from there
- ∴ The SSE with min value is considered as our model.
- The Changing of these values is done by using Gradient Decent method.

R-Studio :-

```
ftLM <- lm(ft$A ~ ft$B + ft$C +
                  ft$D + ft$E.)
```

Summary (ftLM) :-

→ Residuals :-

Test weather it is skewed or not.

→ Co-efficients :-

	Estimate	Sd. Error	t value	Pr Pvalue.
Intercept	31.6084	7.1051	4.449	0.006.
B	7.0676	1.0031	7.046	0.07
C	−0.4201	0.2413	−1.741	0.14216
D	−0.2375	0.5086	−0.467	0.66018
E	−0.7936	0.2977	−2.666	0.04658

we cannot reject null hypothesis (i.e, Co-eff = 0)

→ Multiple R-Squared value : 0.9186

Adjusted R-Squared :- 0.8535

F-statistic : 14.11 on 4 and 5 DF

P-value : 0.006232

B-Values (Co-eff) are not significant

↓

Since P value is 0.14 & 0.60, that means we cannot reject null hypothesis

This could be happening because C & D might be related to each other.

we must Check the relation using VIF.

* Vif(ftLM) ← Library(car)

→ B C D E

1.031045 1.616 1.415 1.209.

Now remove the variable with highest VIF and rebuilt the model. till you get relevant values.

ftLM2 ← lm(A ~ B + D + E).

* Now check the assumptions are true!

→ plot(ftLM2).

* By doing these steps, still we are not getting good model!, then.

we must check the variable interaction terms.

including interaction terms in r-studio is as follows:-

ftLMint ← lm(A ~ B + D + B:D)

: → Colon

B:D → Interaction term

Categorical Variable :-

(*) If there are categorical variables such as gender, geographic region, occupation, marital status, etc in the regression analysis,

we need to insert n-1 dummy variable into the regression analysis replacing the category

↳ In R-Studio

import

library(fastDummies)

↳ Now select the column which is categorical and write the code to remove it as follows

⇒ SalGen ← dummy_cols(Data, select_columns =C("Gender")), remove_selected_columns = TRUE)

→ It will create 'n' number of Gender Column (i.e Gender_male & Gender_Female)

Since we need only 'n-1' number of Column (i.e either Gen_male / Gen_Female)

∴ we remove any only column by Subsetting

⇒ SalGen ← subset(Data, select = -C(Gen_male)).

(*) Now we must consider these categorical variables as continous variable.

Linear Regression — Summary.

→ The aim is to build a model of a numeri-cal variable (Continuous data) using one or more other related variables.

→ Relationship b/w variables is established by calculating the co-variance and/or co-relation co-efficient

→ The variables that are used to build the model are called predictor/regressors variable

→ The variable that is modeled is called the predicted variable / outcome variable.

→ The best possible prediction for the outcome variable is only its mean value, The performance of this "no-regressor" model is quantified by SST or Mean Squared Total (MST.)

→ When there is only one regressor variable, the model is called SLR model.

↳ The reg var may appear in either linear or non-linear form in SLR.

→ Non-linearities can be identified using Tukey's four quadrant approach and built into the model using Tukey's ladder

→ An SLR is written in the form $y = mx + c$

↳ 'x' variable may appear in any non-linear form.

→ The model building exercise involves finding the same value of 'm' and 'c' that minimise the "Sum of squared Errors" (SSE)

→ The performance of the SLR is quantified by using
  ↳ R-squared
  ↳ The hypothesis testing on the slope (m)

→ $R^2 = \frac{SSR}{SST}$ (It denotes the proportion of total error explained by model

---

→ When the number of regressor variables is more than 1, we have "multiple linear regression" (MLR)

→ In MLR, the regressor variables might be correlated among themselves.
  ↳ Correlation among regressor variables may be studied using "Variance Inflation Factor" (VIF)
  ***↳ The variables with VIF values greater than 4 are considered to be multicollinear with other variables and may be dropped from the model building exercise.

→ Each categorical regressor variable is replaced by n-1 dummy variables, where n is the number of categorical in that variable.

→ Interaction among the regressor variables are modelled as product of the variables.

→ R-Squared value in MLR is defined as.

$$1 - \frac{MSE}{MST}$$

↳ Mean errors are determined using appropriate DOF.

↳ MSE is obtained by dividing SSE by $n-k-1$ (DOF) where 'k' is the number of regressor variables used in the model. $MSE = \frac{SSE}{n-k-1}$

↳ MST is obtained by dividing SST by $n-1$.

$$MST = \frac{SST}{n-1}$$

→ The decision to reject or not reject the model is done by calculating the F-statistic, which is defined as the ratio of MSR to MSE

↳ MSR is obtained by dividing SSR by k

$$MSR = \frac{SSR}{k}$$

→ When the model is not rejected, the hypothesis testing on each regressor variable is done using t-test.

→ When one or more reg.var is found to be not statistically significant from zero, the model is rebuilt by dropping that/those variables.

→ In MLR, it is better to start with the most complex model that has all the variables and all "pairs" of interactions.

↳ The nonsignificant variables from this are dropped and a new model is built

↳ The process is repeated until the remaining regressors are significant.

Next. # Classification Models

The outcome could be either Continous or Categorical

what if Continous → linear regression
&
Categorical (Binary) → Next topic: logistic regress

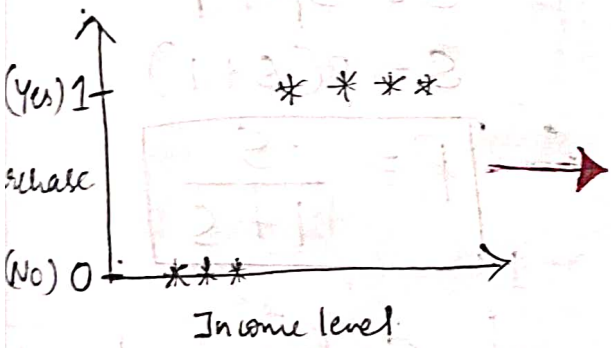
Logistic Regression.

↳ Binary classification technique

Binary classification:-

↳ outcome variable is either 0 (No) or 1 (Yes)

↳ Regressor variable are numerical or categorical

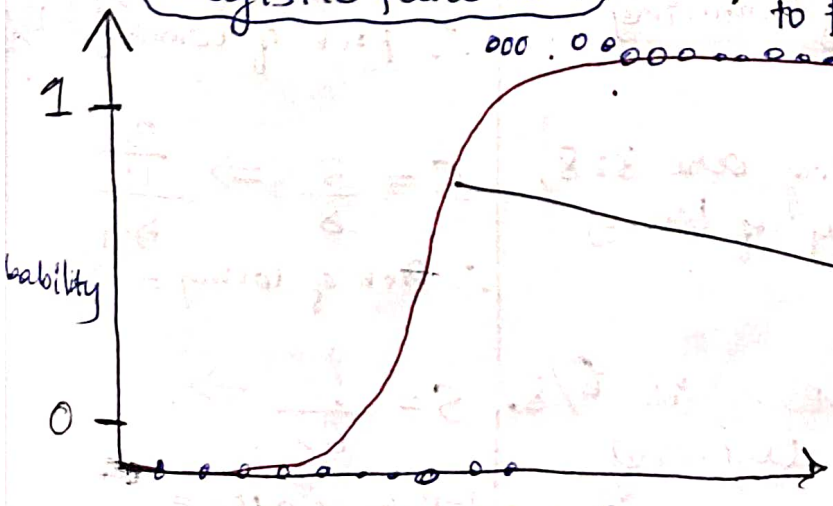


For this kind of data (Binary) if treat the outputs as continous and build a linear regression model, then that would not give exact prediction.

(1) We find the probability values and get the Sigmoid /ogive / psychometric curve as shown below.

Logistic function

↳ once we get this curve, we have to find the corresponding 'X' value for the required probability



$$f(x) = p = \frac{1}{1+e^{-u}} = \frac{e^{u}}{1+e^{u}}$$

$$u = \beta_0 + \beta_1 x_1$$

Logistic - Sigmoid - Ogive - Psychometric function

Mathematically

$$f(x) = p = \frac{1}{1+e^{-u}} = \frac{e^{u}}{1+e^{u}} = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2,\ldots\ldots\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\ldots\ldots\beta_n x_n}} \quad \text{--- (1)}$$

Since this is extremely non-linear function.

∴ we define the probability in terms of odds.

→ we introduce a term called **Odds (S)**

$$\boxed{S = \text{Odds} = \frac{p}{1-p}} \Rightarrow S(1-p) = p$$

$$S - Sp = p$$

$$S = Sp + p$$

$$S = p(S+1)$$

$$\therefore \boxed{p = \frac{S}{1+S}}$$

(*) If the prob of winning is 6/12. What are the odds of winning →

$$S = \frac{6/12}{1-6/12} = \frac{1}{1} = 1:1$$

(*) If the odds of winning are 13:2 what is the prob of winning

$$S = \frac{13}{2} \Rightarrow \frac{13}{15}$$

$$\frac{2}{15}$$

∴ Prob of winning = 13/15

(*) If the odds of winning are 3:8, what is the probability of losing

$$S = \frac{3}{8} \Rightarrow \frac{3}{11}$$

$$\frac{8}{11}$$

∴ Prob of losing = 8/11

(*) If the probability of losing is 6/8, what are the odds of winning.

$$S = \frac{?}{6/8} \Rightarrow$$

$1-p = 6/8 \therefore p = 1 - 6/8 =$

Since from (1)

$$S = \text{odds} = \frac{e^{\beta_0 + \beta_1 x + \beta_2 x_2 \ldots \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n}}$$

$$1 - \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n}}{1 + e^{\beta_0 + \beta_1 x + \ldots + \beta_n x_n}}$$

$$\therefore S = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 \ldots + \beta_n x_n}$$

Since the above equation is non linear we convert into linear

apply log on b.s

$$\log(S) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

→ Solve for Betas

↳ Logit function.

Probability (P)	odds (S)	log (S)
0	0	undefined
0.10	1/9	−2.19
0.25	1/3	−1.09
0.5	1	0
0.75	3	1.09
0.9	9	2.19
0.95	19	2.94
1	∞	inf

(2) Now we must set the boundary.

$$\therefore \boxed{\log(s) = 0 = \beta_0 + \beta_1 x_1 + \beta_2 x + \cdots + \beta_n x_n}$$

↳ Boundary

In 1-dimensional;

$$\log(s) = \beta_0 + \beta_1 x_1 = 0$$

$$\Rightarrow \boxed{x_1 = \frac{\log(s) - \beta_0}{\beta_1}} \rightarrow \text{This will be the boundary}$$

In 2-dimensional;

$$\log(s) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0$$

$$\Rightarrow \boxed{x_2 = \log\left(-\frac{\beta_0}{\beta_2} - \frac{\beta_1}{\beta_2} x_1\right)} \rightarrow \text{This is the boundary}$$

This is of the form $y = mx + c$

$$\therefore \boxed{m = -\frac{\beta_1}{\beta_2}} \qquad \boxed{c = -\frac{\beta_0}{\beta_2}}$$

(*) In linear regression for finding the values of β_0 & β_1 we used minimized SSE.

i.e, for least value of SSE we take/find the corresponding β_0 & β_1 value.

⟶ But SSE doesnot work for logistic regression Problems. Since we are solving for probability which is highly non-linear and has lot of minima and maxima.

∴ we introduce a function which deals with probability ⟶ log likelihood function

Log Likelihood

$$\log(\text{likelihood}) = \sum_{i=1}^{n} [y_i \ln \hat{y}_i + (1-y_i) \ln(1-\hat{y}_i)]$$

(Cost Function)

here.

$y_i \longrightarrow p_i \rightarrow$ Probability of the ith point to be in the category 1

$\hat{y}_i \longrightarrow \hat{p}_i \rightarrow$ predicted probability of the ith point to be in the category 1.

* The accuracy will be high when the.

$y_i \approx \hat{y}_i$ (The prediction match with actuality)

we get $\hat{y}_i$ from.

$$\log(s) = \beta_0 + \beta_1 x_1 + \beta_2 x + \cdots + \beta_n x_n$$

(i) take anti log of this, we get odds

(ii) get the probability from odds, which is predicted probability.

* The cost function is used to find the β_0 & β_1 .... values by using gradient decent method.

R-Studio :-

```
flresLgR <- glm(flres$A ~ flres$B,
                family = 'binomial')
Summary(flresLgR)
```

→ Deviance Residuals:

Min.	1Q	Median	3Q	Ma
-1.95015	-0.321	-0.05335	0.265	1.72

model { Co-efficients:

	Estimate	Std. Error	Z value	Pr(>|z|) p-Valu
Intercept	-20.40782	4.52332	-4.512	6.43e-0
flres$B	0.4259	0.09482	4.492	7.09e-0

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 123.156 on 91 DOF

Residual deviance: 49.937 on 90 DOF.

Number of Fisher Scoring iterations: 7

① Since the p-values are lesser than 0.05
∴ cannot reject the model (since we reject the null hypo that intersept & $B_1 = 0$).

② we donot get F-value in logistic Reg

(*) Since we know that F-statistic is.

$$F\text{-stat} = \frac{M.E.E}{M.U.E.E}$$

∴ we donot have any explained/unexplained errors in log.reg therefore there is no F-stat.

(*) Therefore we look at Null deviance and Residual deviance.

(*) Since we got the β_0 & β_1 values from the Summary of model.

i.e

(1) $\log(s) = -20.40782 + 0.4259 * Age.$

and loglikelihood can be found by dividing the Residual deviance by -2

i.e $\log(\text{likelihood}) = \frac{49.937}{-2} = -24.97.$

∴ Residual deviance $= -2 \times \log \text{likelihood}$ with model

Null deviance $= -2 \times \log \text{likelihood}$ without model

(*) ∴ we treat Residual deviance as explained (SSR) errors and Null dev as total errors. SST

∴ Pseudo R^2 (McFadden's R^2) $= 1 - \frac{\text{Residual Deviance}}{\text{Null Deviance}} \approx 1 - \frac{SSE}{SST}$

Most often we donot use Pseudo R^2, since it tells us that how close is P_i to $\hat{P}_i$ but we are not worried about that (But interested in categorising the data into yes or no)

We can find the 'x' value (Age) for the corresponding probability → Boundary

$$\log(s) = -20.41 + 0.43 * Age$$

if $p = 0.5$ boundray of Age = ?

↳ for $p = 0.5$ $S = 1$

$$\log(s) = 0$$

$$\therefore -20.41 + 0.43 * Age = 0$$

$$\Rightarrow \boxed{Age = 47.9}$$

i.e above 47.9 age will say yes is the model prediction for $p = 0.5$

Also from (1)

For age of 'X', which categorey does he falls.

Eg Age = 50

$$\log(s) = -20.41 + 0.43 * 50$$

$$\log(s) = 0.89$$

$$S = e^{0.89}$$

$$S = 2.43$$

$$\therefore p = \frac{S}{1+S} = 0.71$$

For this model what kind of errors are possible is given below. by Confusion Matrix or truth table.

Confusion Matrix		Prediction.	
		Class 1 (S)+	Class 0 (F)-
Actual	Class 1 (Success) +	Predicted to be +ve which is true.	Predicted to be -ve which is false.
	Class 0. (Failure) -	Predicted to be positive which is false	Predicted to be -ve which is true.

Confusion Matrix		Prediction.	
		Positive	negative.
Actual	Positive	True positive (n_1)	False negative (n_2)
	negative.	False Positive (n_3)	True Negative (n_4)

$$n_1 + n_2 + n_3 + n_4 = N \text{ (Total persons)}$$

$$\text{Accuracy} = \frac{n_1 + n_4}{N}$$

$$\text{Sensitivity (Recall)} = \frac{\text{Correctly identified to be +ve (True Positive)}}{\text{Actually Positive}} = \frac{n_1}{n_1 + n_2}$$

$$\text{Specificity} = \frac{\text{Correctly identified to be -ve (True negative)}}{\text{Actually Negative}} = \frac{n_4}{n_3 + n_4}$$

$$\text{Precision} = \frac{\text{Correctly identified to be +ve (True Positive)}}{\text{Predicted Positive}} = \frac{n_1}{n_1 + n_3}$$

According to the which type of question we are solving, we must consider the Accuracy/Precision/Specificity/Sensitivity accordingly.

In case of unbalanced data, accuracy is not good quantity to consider, ∴ we consider a new metric called Kappa Metric.

Kappa Metric :-

↳ This says that "How well the model d better than random classification"

$$Kappa = \frac{\text{total Accuracy} - \text{random Accura}}{1 - \text{random Accuracy}}$$

$$\text{totalAccuracy} = \frac{\text{Correct Predictions}}{\text{Total.}}$$

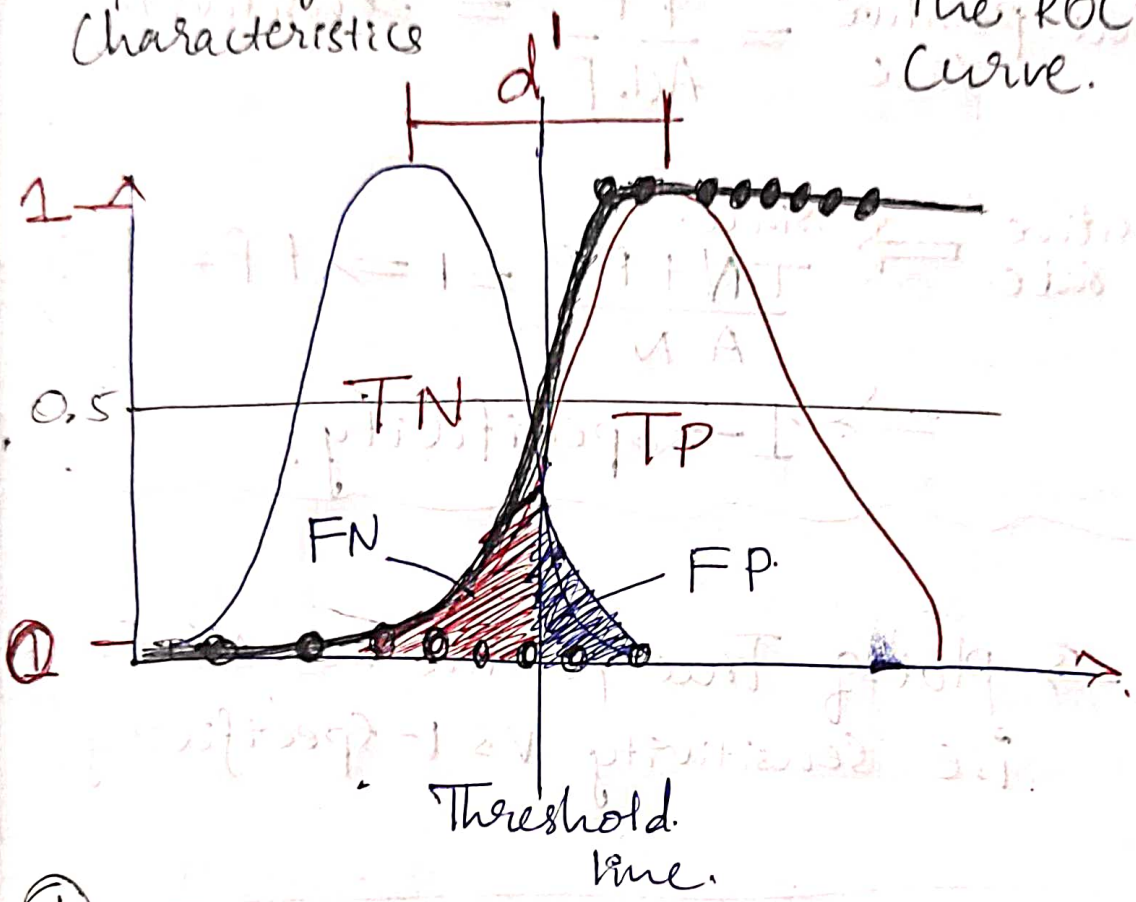
$$\text{random Accuracy} = \frac{\text{Actual False}}{\text{Total}} * \frac{\text{Pred. Fal}}{\text{Total}} + \frac{\text{Actual True}}{\text{Total}} * \frac{\text{Pred. Tru}}{\text{Total}}$$

Kappa Value.	
<0	NO Agreement
0-0.2	Slight
0.21 to 0.4	Fair
0.4 to 0.6	Moderate
0.6 to 0.8	Substantial
0.8 to 1	Almost Perfect.

ROC Curves and AUC :-

Receiver operating Characteristics

Area under the ROC Curve.



(*) If we move the threshold i.e if we move to high value, then True positeve and False positive decreases and if we move the threshold to low value. True Positive and False positive also increases.

(*) The outcome is TP & FP Changes together if we change the threshold value.

(*) The actual model's worthiness is measured using "d'" (Discriminability of the model)

(*) from this we can find the correct threshold value.

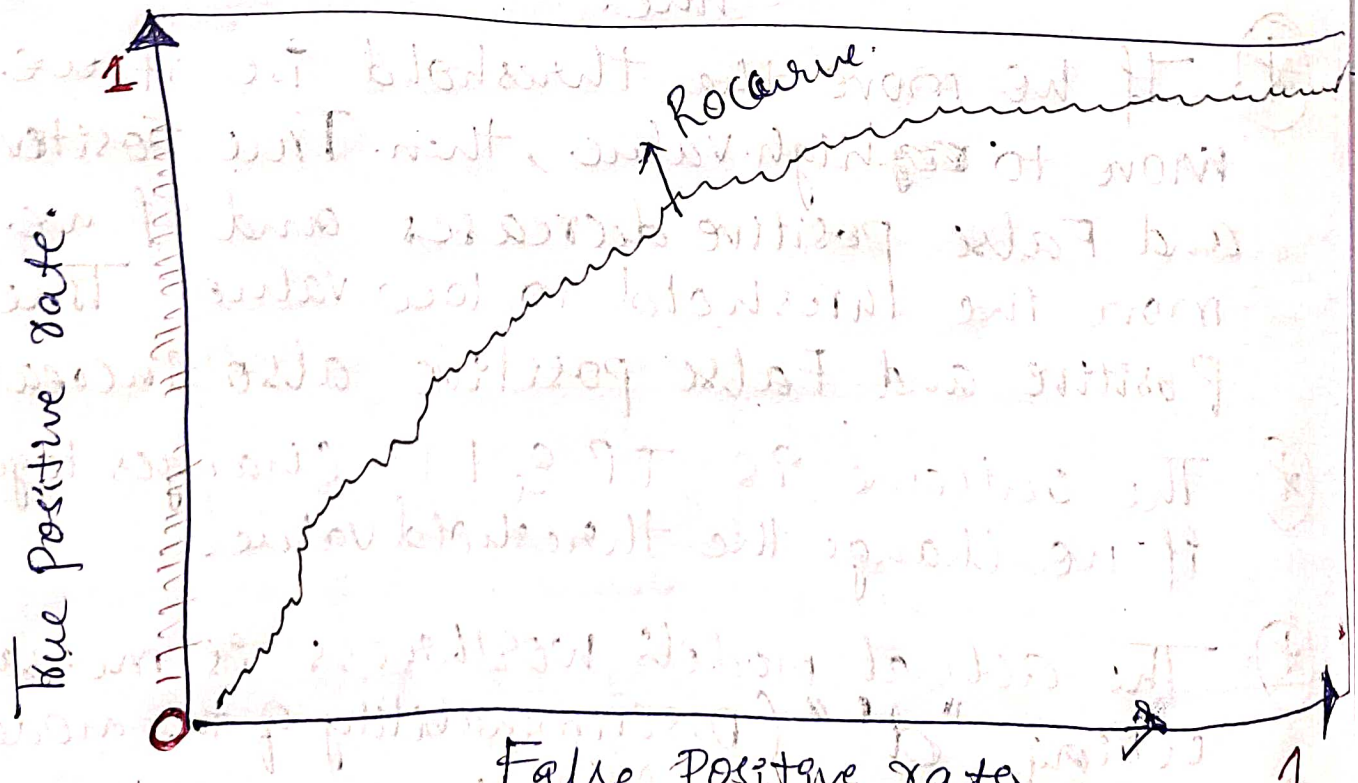
④ The graph of $\underbrace{1-\text{Specificity}}_{\text{X axis}}$ Eg. $\underbrace{\text{Sensitivity}}_{\text{Y axis}}$ is called as ROC curve.

True positive rate $= \frac{TP}{Act.\ P} = \text{Sensitivity}$.

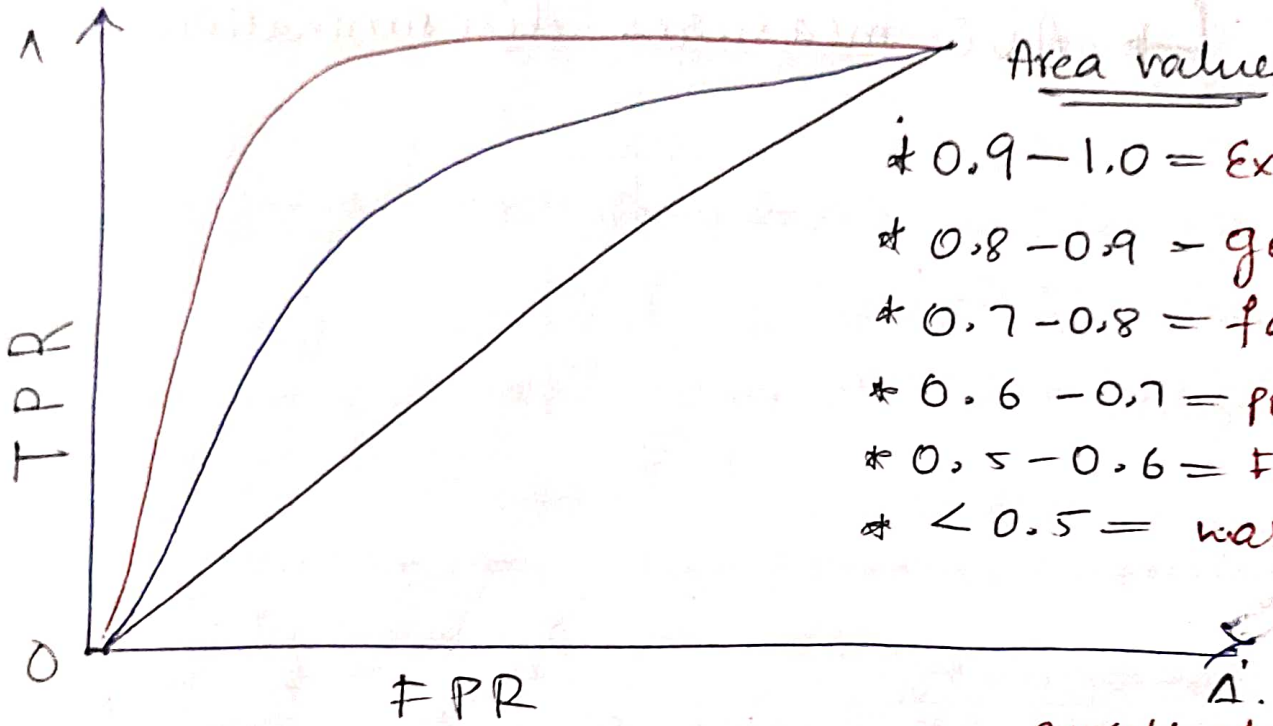
False positive rate $\Rightarrow$ Since $\frac{TN + FP}{AN} = 1 \Rightarrow FP = 1 - \frac{TN}{AN}$

$\Rightarrow 1 - \text{Specificity}$.

∴ ROC $\rightarrow$ Plot of True positive rate vs False Positive rate
i.e Sensitivity vs 1-Specificity.



* The area in the plot is one. Since 1 x 1 (x axis, y axis)
* Area under ROC curve tells how good is the classification (i.e Discriminability)
* larger the area better is Discriminability

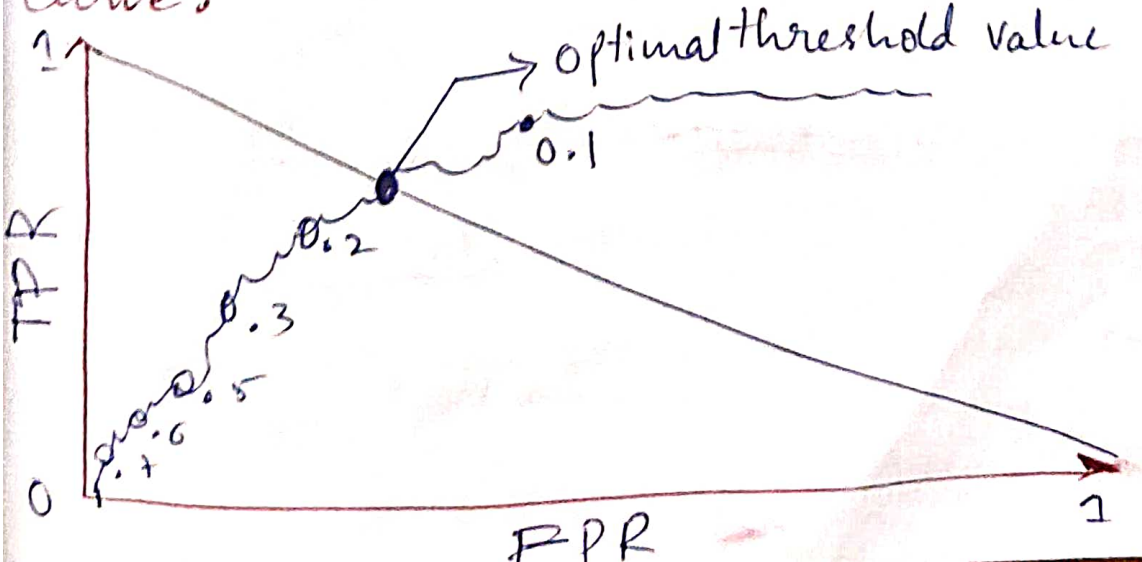


Area values
* 0.9 – 1.0 = Excellent
* 0.8 – 0.9 = good
* 0.7 – 0.8 = fair
* 0.6 – 0.7 = poor
* 0.5 – 0.6 = Fail
* < 0.5 = waste.

— excellent
— good
— worthless

Final Question is how do we pick the best threshold for the model?

By drawing a diagonal from left to right and mark the point where diagonal cuts the ROC Curve.



(+) ROC will help to get the best threshold for model.

(+) AUC will help to determine LRM is good or not; higher the AUC; better the model

↳ AUC measures discrimination

Naïve Bayes :-

) Naive Bayes is used for multi-class classificat"
) It is the application of Bayes theorem.

Naive Bayes Algorithm :-

) Naive Bayes :- Computes $P(y_i|x)$ by using Bayes theorem and instead computes the inverse conditional probability $P(x|y_i)$.

) A simple classifier that performs surprisingly well on a large class of problems

) This type of methods are called Generative learning Models.

To find → $P(A|B) = ?$

Given → $P(B|A)$

B.K.t $P(A|B) = \frac{P(A)}{P(B)} \times P(B|A)$

$P(A)$: marginal probability
$P(B)$: marginal probability
$P(B|A)$: Cond prob

This is Bayes theorem.

) But In Naive Bayes classification method we proceed with the above step along with the "Naive Assumption". for eg → to find $P(A|B, C)$

) He we assume that B & C are independent of each other and we rewrite as

Example :-

What is the probability of winning the match when the Indian cricket team wins the toss & chooses to field.

Soln Event A : Toss : $A = \{W, L\}$

Event B : Choice : $B = \{Field, Bat\}$

Event C : Match Outcome : $C = \{W, L, D\}$

To find

$$P(C=W \mid A=W, B=Field) = ?$$

"We make an assumption that Event B and Event A are independent". This assumption is called Naive Assumption.

for eg:

$$P(C=W \mid A=W, B=Field) = P(C=W \mid A=W) \times P(C=W \mid B=Field)$$

$$\therefore P(C=W \mid A=W, B=Field) = P(A=W, B=Field \mid C=W) \times \frac{P(C=W)}{P(A=W, B=Field)}$$

$$\Rightarrow P(A=W \mid C=W) \times P(B=Field \mid C=W) \times \frac{P(C=W)}{P(A=W) \times P(B=F)}$$

$\Rightarrow$ The above eqn is final eqn and solve for it.

R-Studio:-

```
library(e1071)
nb_model <- naiveBayes(Class ~ ., data=
                       trainHouseVotes84)
```

It gives all the prob values which must be used in the equation.

---

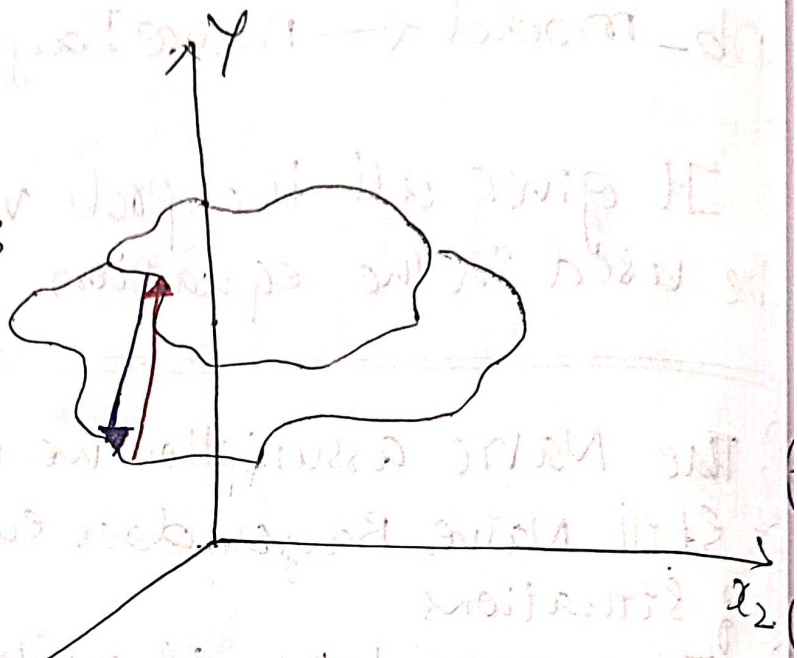
1) The Naive assumption we makes is never true.
2) Still Naive Bayes does surprisingly well in a lot of situations
3) It works best if predictor variables are Categorical
4) It is a parameter-free model

Another important idea is "When we have Continuous data we Create some bins with some conditions and make it Catogorical, then we Solve using Naive Bayes algorithm".

Gradient Descent :—

$$y = f(x_1, x_2)$$

Let, $y = 3x_1^2 + 4x_2 - 3$



Gradient is given as

$$\vec{\nabla}_y = \frac{\partial y}{\partial x_1}\hat{x}_1 + \frac{\partial y}{\partial x_2}\hat{x}_2$$

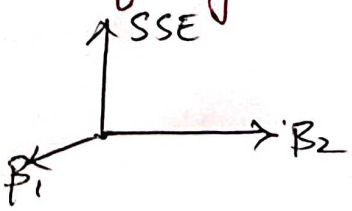
$(\vec{\nabla}_y)$

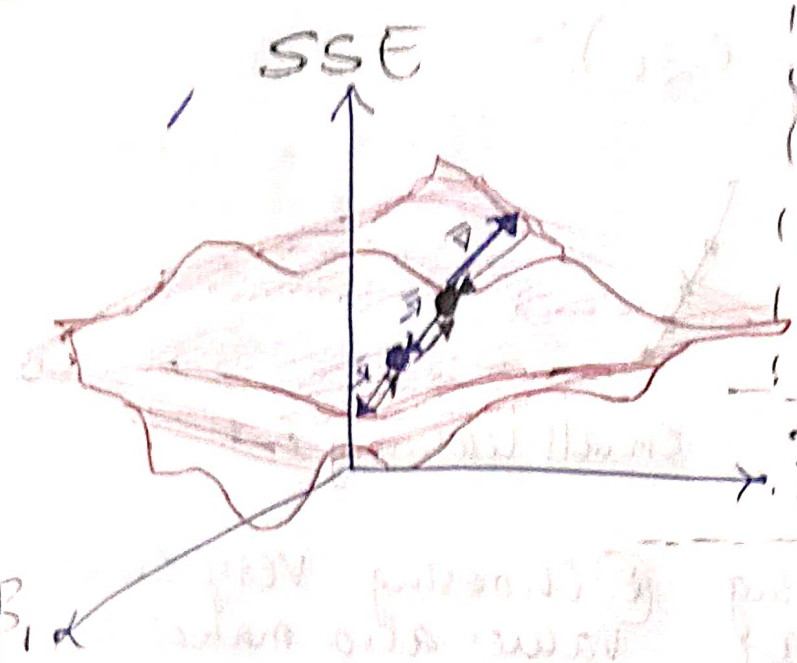
→ Gradient is always point in the direction of steepest increase in y.

(*) Since we are intrested in the steepest decrease in y ∴ Steepest decrease is $-\vec{\nabla}_y$

(*) In regression problem our aim was to minimise SSE w.r.t. β values.

∴ In the above figure instead of x_1 & x_2 we have β_1 & β_2 and instead of y we have SSE





(*) Start with a assumed β-value

(*) find the gradient (∇) and $-\nabla$

(*) In opp direction. Take a small step β_2, i.e change β_1 or β_2 by small amount. This change is called as (step size)

(*) Now we reach to a new point and find the $-\nabla$ once again.

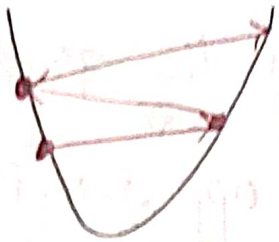
(*) Every time we reach the new point calculate the sum of squared errors and observe that the SSE is lower than the previous then. Continue doing this process, by the time SSE starts to increase, stop going beyond and consider the lowest SSE and corresponding β_1 & β_2 values.

$$\theta_0^{(t+1)} = \theta_0^{(t)} - \left(\alpha \cdot \frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0}\right)$$

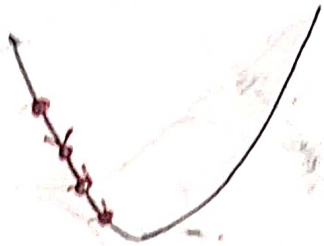
- $\theta_0^{(t+1)}$: New point
- $\theta_0^{(t)}$: Original point
- α: Learning Rate (Step size)
- $\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_0}$: Slope of steepest descent (direction of fast

(*) Setting the step size is an art and it depends on the problem given.

Learning Rate (Step size) :-



Big learning rate



Small learning rate

(*) Choosing a big learning rate cannot help to get exact minimum point.

(*) Choosing very small value also makes lot of time to get min point

∴ Choosing a step size is an art and comes by practice.

Optimization Algorithms :-

(1) Derivative Based Algorithms (Gradient Based)
- ↳ Steepest descent method.
- ↳ Newton-Raphson method.
- ↳ Conjugate gradient etc.

(2) Derivative-free methods
- ↳ Random search method
- ↳ Genetic algorithm
- ↳ Simulated annealing etc.