BT-307 (Post midsem)

End sem → 30 Quiz → 10, Lab → 5, Attendance → 5

Probability, Linear Algebra ⇒ Imp topics req

Exam ⇒ Mathematical (Derivation / Questions)

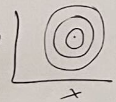
Lec 1 (4 Mar)

Covariance ⇒ measure of association of 2 variables X & Y.

$$\text{cov}(X,Y) = E\left[(X - E(X))(Y - E(Y))\right]$$

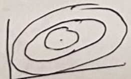
$$= E(XY) - E(X)E(Y)$$

If $\text{cov}(X,Y) = 0$, graph will be concentric circles



If $\text{cov}(X,Y) > 0$, i.e. if X ↑ Y ↑ & if X ↓, Y ↓

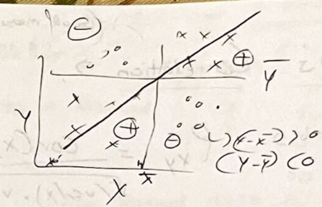
Graph will be ellipse



X	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n
$\bar{x}$	$\bar{y}$

S^2_{XY} ⇒ Covariance of X & Y collected from sample.

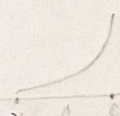
$$\boxed{S^2_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)}}$$



if S^2_{XY} is positive, X & Y are positively associated

if $S^2_{XY} < 0$, X & Y have -ve correlation

$$-\infty \le \text{cov}(X,Y) \le +\infty \text{ (infinity)}$$



if $X \perp Y$ (X is independent of Y), $\text{cov}(X,Y) = 0$

but if $\text{cov}(X,Y) = 0$, it doesn't necessarily mean that X is dependent on Y,

Weakness of Covariance ⇒

1. For 3 variables X, Y & Z, they can have different units. Because of which, we are unable to compare the pair and get which pair has more association than the other.

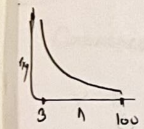
So, instead of covariance, we use Pearson's Correlation. as it standardizes the units i.e if x changes linearly Y changes linearly — Association

• Pearson's Correlation ⇒

if we are for prob ⇒ $\rho_{xy} = \frac{Cov(x,y)}{\sqrt{(var(X)\cdot var(Y))}}$

for sample data, $r_{xy} = \frac{S^2_{xy}}{S_x S_y}$ { S_x & S_y are std. deviation of x & y }

depends on sample size



$-1 \le \rho_{xy} \le +1$

[0 when independent]

If your samples have no-correlation, it will still have some Pearson's correlation value just because it has small n (if n = 2, r_{xy} is always 1). As ~~ρ_{xy}~~ r_{xy} heavily depends on n (sample size), so we do t-test to find out if null hypothesis is true or false. ~~Then~~ If null hypothesis is rejected, then the value calculated is accepted & if null hypothesis is ~~expe~~ accepted, then we can for sure say that there is no correlation.

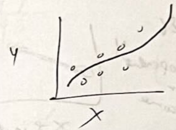
$H_0 : \rho_{xy} = 0$

Spearman's Correlation ⇒ for Non-linear & Monotonic situation.

R(x)	R(y)		X	Y	
2	2		$40 = x_1$	y_1	4
3	3		$32 = x_2$	y_2	6
1	4		25 = ⋮	⋮	8
4	1		50		2

Convert the given data to a Ranked data.

suppose ⇒



Use Ranked data to cal. Pearson's Correlation $\boxed{\rho_{xy} = \frac{Cov(R(x), R(y))}{\sigma R(x) \cdot \sigma R(y)}}$

∗ ⇒ Association DOESN'T Lead to Causation.

Lec 2 (5 Mar)

$\vec{u} = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}$ $v = \begin{bmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{bmatrix}$ $S^2_{xy} = \frac{\vec{u}\cdot\vec{v}}{n}$

X	Y
x_1	y_1
x_2	y_2
⋮	⋮
x_n	y_n
$\bar{X}$	$\bar{Y}$

$\vec{u}^T \vec{u} = \vec{u}\cdot\vec{u} = \|\vec{u}\|^2$

$S_x^2 = \frac{\|\vec{u}\|^2}{n}$ $S_y^2 = \frac{\|\vec{v}\|^2}{n}$

$r_{xy} = \frac{S^2_{xy}}{S_x S_y} = \frac{\vec{u}\cdot\vec{v}}{\|\vec{u}\|\|\vec{v}\|} = \cos\theta$

Also $|\cos\theta| \le 1$ ∴ $|\vec{u}\cdot\vec{v}| \le \|\vec{u}\|\|\vec{v}\|$

when it also equal to 1, ~~it~~ ∴ u & v are linearly dependent. → $\vec{x} = c\vec{y}$

if there is linearity, value will be +1/−1.

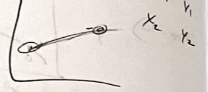
Covariance matrix = $\begin{vmatrix} S_x^2 & S_{xy} \\ S_{xy}^2 & S_y^2 \end{vmatrix}$

∗ Regression doesn't have unique sol^n.

After we find correlation, we have to see if it can be written as $\underline{Y = f(x)}$ → simplest model → Linear → $\underline{Y = ax + b}$

g - Relation bet^n covariance & corr^n.

for 1 data pt = 2, we can easily get a & b.



if n — = 3, it becomes overdetermined system. Which to choose?

Y → dependent or outcome

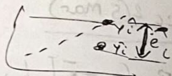
X → independent or predictor

do optimization of sol^n ⇒ $\hat{a}$ → estimated value & $\hat{b}$ → est. value

$$\hat{y}_i = \hat{a} x_i + \hat{b}$$

$\neq$

y_i



error/deviation ← $e_i = (y_i - \hat{y}_i)^2$ ← makes it convex

Residual sum of square $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

Graph of RSS → only single minima

Choose a & b when RSS is min.

Do partial derivative wrt. a & b separately, we get 2 eq^{ns} & set them = 0 & solve to get.

By solving →

$$\boxed{\hat{a} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}} = \frac{Cov(x,y)}{Sx^2} \text{ or } \frac{cov(x,y)}{var(x)}$$

$$\boxed{\hat{b} = \bar{y} - \hat{a}\bar{x}}$$

$y_1 = ax_1 + b$

$y_2 = ax_2 + b$

$$\Rightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

$$\vec{Y} = X \vec{B}$$

eg → $y = 2x + 1$

x	y
1	3
2	5

$$\begin{bmatrix} 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$a \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_{\vec{x_{c1}}} + b \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{\vec{x_{c2}}} = \underbrace{\begin{bmatrix} \quad \end{bmatrix}}_{\vec{v_n}}$$

$\vec{v_n}$ is in the vector space of $\vec{x_{c1}}$ & $\vec{x_{c2}}$

∴ $\vec{Y}$ is in column space of X as it can be generated by the linear comb^n of columns of X, coeff. matrix.

$A = \begin{bmatrix} 1 & 4 \\ 2 & 8 \end{bmatrix}$ with columns $\vec{c_1}$, $\vec{c_2}$

$$\boxed{\vec{q} = x\vec{c_1} + y\vec{c_2}}$$

∴ q is in col_A (column space of A)

$$\boxed{B = X^{-1}\vec{Y}}$$

only cond^n is that X should be invertible.

$\vec{y} = a\vec{c_1} + b\vec{c_2}$

∴ $\vec{Y}$ is in column space of X. (col_X)

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_N & 1 \end{bmatrix}_{N \times 2 \, (N > 2)} \begin{bmatrix} a \\ b \end{bmatrix}$$

$\vec{Y}$ $\quad X \quad \vec{B}$

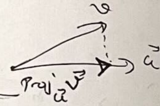
if $\vec{Y} = a\vec{c_1} + a\vec{c_2}$ can be d

As it is non invertible no value of a & b which can satisfy the cond^n & $\vec{Y}$ is not in

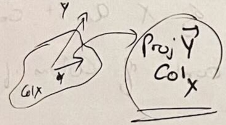
∴ Column space of X. ∴ $\vec{Y}$ is outside the ~~column~~ col X.

∴ We have to take the project^n of $\vec{Y}$ on ~~the~~ all the column of X,

$\dfrac{\vec{v} \cdot \vec{u}}{\quad}$ ⟸ $Proj_{\vec{u}} \vec{v}$



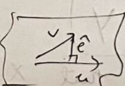
As it is the closest to the answer.

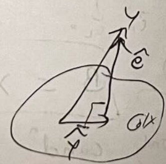


$Proj\, \vec{Y}$ Col_X

$$\hat{Y} = X \cdot \vec{y} = x^T \cdot y^T$$

$$\hat{e} = \vec{Y} - \vec{\hat{Y}}$$





$$\hat{e} \perp \vec{u}$$

here, $\hat{e} \perp X$ ∴ $\underline{X \cdot \hat{e} = 0}$

$$X^T \cdot \hat{e} = 0$$

$$X^T (Y - \hat{Y}) = 0$$

$$X^T (Y - X\vec{B}) = 0$$

$$\cancel{X^T X} \; X^T X \vec{B} = X^T Y$$

$$\cancel{\quad} \; \boxed{\vec{B} = (X^T X)^{-1} X^T Y}$$

∴ For $Y = ax + b$ where no. of data pts is greater than 2, to find B vector →

$$\boxed{\vec{B} = (X^T X)^{-1} X^T Y}$$

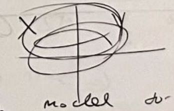
Hat matrix

Peudo inverse of the X matrix

Mean of est. value of ~~(~~ Y $(\hat{Y})$ will be equal to mean of Y.

& sum of error of each y is equal to zero.

$\hat{}$ Lec (15/3)



X	Y
x	

for Alternate model for $\hat{Y} = X\vec{B}$

Bad model ⇒ Y doesn't depend on X.

i.e $Y = C$ (const.)

∈ ~~will show~~ best value for $C = \overline{Y}$ ~~(const)~~ (estimate)

To calculate the difference, RMS error is used, i.e. error will be compared

Error : $Y = aX + b \Rightarrow \boxed{RSS = \Sigma (y_i - \hat{y}_i)^2}$

for $\boxed{Y = c = \bar{Y}} \Rightarrow \boxed{TSS \Rightarrow \Sigma (y_i - \bar{y})^2}$

Total sum of squares

To calculate diff $\Rightarrow \boxed{R^2 = \dfrac{TSS - RSS}{TSS}}$

coeff. of determination

tells how much the model is good wrt $(Y = \bar{Y})$

$R^2 = 1 - \dfrac{RSS}{TSS}$

If there is an intercept

$\Rightarrow TSS = RSS + ESS$

$\boxed{ESS = \Sigma (\hat{y}_i - \bar{y})^2}$ variation of estimates from mean

$\boxed{R^2 = ESS/TSS = \dfrac{\Sigma (\hat{y}_i - \bar{y})^2/n}{\Sigma (y_i - \bar{y})^2/n}}$

variation of real values from mean

$0 \leq R^2 \leq 1$

for R^2 to be (−ve), RSS has to be greater than TSS.

possible analytically

$R^2 = 0.4$

$R^2 = 0.8 \Rightarrow Y = ax + b$ is better than $Y = \bar{Y}$.

After calc. R^2, we have to see how correct it is by generalizing it.

few data pts, but doesn't imply for complete dataset

→ Statistics

$x_1, x_2, \ldots, x_n$

$\bar{x} \rightarrow \mu$

how good sample statistic represents the population? sample statistics pop^n params

$Y = a x + b$

$\downarrow$

$Y = \alpha x + \beta$ → parameter

$$y_i = \alpha x_i + \beta + \varepsilon_i \rightarrow \text{noise}$$

$$\varepsilon_i \sim N(0, \sigma^2) \rightarrow \text{normal distribution}$$

t-test for regression coefficients : alternative

for $a \Rightarrow$ we have to make Null & alt. hypothesis at the population level.

$H_0 : \alpha = 0 \; ; \; H_1 : \alpha \neq 0.$

$t = \dfrac{\text{sample stat}^{(a)} - \text{population parameter}(\alpha)}{\text{SE of sample stat}}$

if H_0 is true $\Rightarrow t = \dfrac{a - 0}{SE(a)}$

no of pts

→ here, we divide by $(n - 2) \rightarrow$ df of t.

as regression problem starts when samples are > 2, (2 pts only defines a line when the data pt. is anchored/fixed)

$df \rightarrow (n-1)$

$P(t \mid H_0 = T) >$ cut-off {significant level}

the H_0 is correct.

Catch → (Sample stat − Pop. para) can be (Pop. para − Sample stat.)

When it is the case, do 2 sided t-test.

1) → Do linear regression & find a & b

2) → Find R^2

3) → Do t-test for a & b.

If $b=0$, & $a \neq 0$ cases, then we have to do the above steps again with $b=0$ to check the $\hat{a}$ value is true/not.

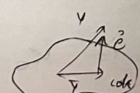
Lec (18/3/2024)

TSS = RSS + ESS

$$\hat{e} \cdot \hat{Y} = 0$$

$$(Y-\hat{Y})\,\hat{Y} = 0$$

$$\boxed{Y^T\hat{Y} = \hat{Y}^T\hat{Y}}$$



e is orthogonal to Col X & ∴ e is orthogonal to the column vector of 1. ∴

$$\hat{e}\cdot\vec{1} = 0$$

$$\hat{e}\cdot(\alpha\vec{1}) = 0$$

$$\hat{e}\,(\bar{Y}\cdot 1) = 0$$

$$(Y-\hat{Y})^T\,\bar{Y} = 0$$

$$\boxed{Y^T\bar{Y} = \hat{Y}^T\bar{Y}}$$

$$TSS = \Sigma(Y_i-\bar{Y})^2$$
$$= (Y-\bar{Y})^T(Y-\bar{Y})$$
$$= Y^TY - 2\bar{Y}^TY + \bar{Y}^T\bar{Y}$$

$$RSS + ESS = \Sigma(Y_i-\hat{Y}_i)^2 + \Sigma(\hat{Y}_i-\bar{Y})^2$$
$$= (Y-\hat{Y})^T(Y-\hat{Y}) + (\hat{Y}-\bar{Y})^T(\hat{Y}-\bar{Y})$$
$$= Y^TY - \cancel{2\hat{Y}^TY} + \cancel{\hat{Y}^T\hat{Y}} + \cancel{\hat{Y}^T\hat{Y}} - 2\hat{Y}^T\bar{Y} + \bar{Y}^T\bar{Y}$$
$$= Y^TY - 2Y^TY + \bar{Y}^T\bar{Y}$$

$$\therefore \boxed{TSS = RSS + ESS}$$

2 input variables affecting

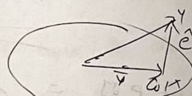
Multiple Regression (only single variable changes)

x_1	x_2	Y
x_{11}	x_{21}	y_1
⋮	⋮	⋮
x_{1n}	x_{2n}	y_n

Simplest model ⇒ $y = a_1x_1 + a_2x_2 + b$

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_{11} & x_{21} & 1 \\ x_{12} & x_{22} & 1 \\ \vdots & \vdots & \vdots \\ x_{1n} & x_{2n} & 1 \end{bmatrix} \quad B = \begin{bmatrix} a_1 \\ a_2 \\ b \end{bmatrix}$$

$$B = (X^TX)^{-1}X^Ty$$



$$\hat{Y} = XB$$

If x_1 depends on x_2 linearly, then $|X|$ will be 0. & inverse can't be calculated. This is called multilinearity problem.

~~If~~ I have multilinearity problem then $(X^TX)^{-1}$ goes very high which leads to variance being v. high.

$$X_j = P_1X_1 + P_2X_2 + \dots + P_{j-1}X_{j-1} + P_{j+1}X_{j+1} + \dots + P_kX_k + q$$

Auxilliary Regression

R_j^2 → $1 - \frac{RSS}{TSS}$

cal. R^2, if close to 1 & good model. i.e X_j has linear dependency & ∴ it has to be removed. To determine do reg. R_j

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

(Variance Inflation factor) x_j

if $VIF > 10$, then there must be linear dependency i.e $R_j^2 > 0.9$

Lec (22/3)

Multiple Linear Regression

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad B = \begin{bmatrix} a_1 \\ a_2 \\ b \end{bmatrix}$$

$$X = \begin{bmatrix} x_{11} & x_{21} & 1 \\ \vdots & \vdots & \vdots \\ x_{1N} & x_{2N} & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} a_1 \\ a_2 \\ b \end{bmatrix} = (X^TX)^{-1}X^TY$$

Y	x_1	x_2
Y_1	x_{11}	x_{21}
⋮	⋮	⋮
Y_N	x_{1N}	x_{2N}

$Y = a_1x_1 + a_2x_2 + b$

- Multi-colinearity problem { Var(Param) ↑↑ ; t-test → wrong }

To check dependency of x_1 & x_2, cal. Pearson coeff / by Aux. Regression

We can also cal. R^2 for the model to check it & wrt $y = \bar{y}$.

Here, there are 2 R^2, multiple R^2 & adjusted R^2 $\{R^2 = 1 - \frac{RSS}{TSS}\}$

$y = a_1x_1 + b \rightarrow RSS_1 = 26$

$y = a_2x_2 + b \rightarrow RSS_2 = 13$

$y = a_1x_1 + a_2x_2 + b$

Algorithm always tries to get the lowest pt. of that variable. So, in 3rd model, it ~~...~~ ($a_1 \neq 0$) introduces causes RSS ↑, it will force it to zero

When more ~~terms~~ independent predictors are introduced in the model, the ~~model might~~ data fitting will ↑, but its predicting power ↓. So not good.

$$\boxed{\text{Adjusted } R^2 = 1 - \frac{\frac{RSS}{(n-k-1)}}{\frac{TSS}{(n-1)}}}$$

[k → no. of indept predictors here (x_1 & x_2 i.e. 2)]

For multiple data pts, pairwise t-test are not ideal {i.e. are problematic.

F-test

Small model ⇒ RSS_S, $RSS_S > RSS_L$

Large model ⇒ RSS_L

$$\left(\frac{RSS_S - RSS_L}{RSS_L}\right)$$

$$\boxed{y = a_1x_1 + a_2x_2 + b}$$

$a_1 \to \alpha_1$, $a_2 \to \alpha_2$, $b \to \beta$

α_1: population level parameter.

H_0 : all $\alpha = 0$ {no dependency}

H_1 : at least one $\alpha \neq 0$ {at least 1 dependency}

$H_0 \Rightarrow y - \bar{y}$

$RSS_{H_0} = \Sigma(y_i - \bar{y})^2$

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

$\therefore \frac{RSS_S - RSS_L}{RSS_L} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y_i - \hat{y}_i)^2}{\Sigma(y_i - \hat{y}_i)^2}$

$= \frac{TSS_L - RSS_L}{RSS_L} = \frac{ESS_L}{RSS_L}$

$$\frac{ESS_L}{RSS_L} \Rightarrow \left.\frac{\frac{ESS_L}{(K)}}{\frac{RSS_L}{(n-K-1)}}\right\} \xrightarrow{\text{normalization by dof}} = \text{F - value}$$

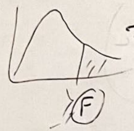
↓

F - distribution

(parameters)

$(K, n-K-1)$

F-distrib^n ⇒



⇒ if F-value is large, i.e. p value is small, it means that the small model of (H_0: all $\alpha = 0$) is ~~rejected~~ bad

∴ null ~~H~~ hypothesis is rejected.

if p value is big, F-value is small

i.e the smalle model is not bad

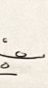
∴ We accept null hypo.

When null hypo. is rejected, i.e (they are non zero), we then do independent t-test to find the $a_1, a_2, \ldots b$

Lec (26/03)

Non Linear regression:

eg:- $\frac{dx}{dt} = -Kx \Rightarrow \frac{x}{x_0} = e^{-xt}$ y.



If non linear eq^n, see if we can linearise the eq^n

$y = e^{-Kt} \Rightarrow \ln y = -Kt \Rightarrow P = -K t$

This is called Linearizing Non Linear eq^n then you do Linear Regression.

For eq^n which can't be linearised (eg ⇒ $y = ae^{-K_1 x} + be^{-K_2 x}$)

we do Non Linear least square

$$\min\left[\Sigma (y_i - \hat{y}_i)^2\right]$$

$$y_i = e^{-Kt_i}$$

keep changing K to find the least value of this.

→ max. ~~the~~ Likelihood.

$Y = a_1 x + a_2 x^2 + \ldots b$ { Polynomial Regression }

⇓

$Y = a_1 x + a_2 w + a_3 z + \ldots b$ { multiple linear regression }

Y	x	w (x^2)	z (x^3)

x, w, z are not linearly dependent.

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} x_1 & x_1^2 & \cdots & x_1^3 & 1 \\ \vdots & \vdots & & \vdots & \vdots \\ x_n & x_n^2 & & x_n^3 & 1 \end{bmatrix}$$

$$B = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b \end{bmatrix} = (X^T X)^{-1} X^T Y$$

Convert data to matrices $(Y, X \& B)$ {eq. in eqn}

To use supervised learning, You should have labelled data.

Classification / clustering → we see which data pts. are close to each other without ~~Unsu~~ any filter (Unsupervised learning)

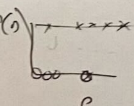
↓

Labelled Data → Trained → Classifier

Lec (27/3)

$D = f(P)$

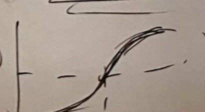
$P(D) = a.P + b$



P	D
	Y
	Y
	1
	1
	N

Logistic Classifier & Logistic Regression

$$P(D) = \frac{1}{1+e^{-(a+bx)}}$$



$$\boxed{P(D) = \frac{1}{1+e^{-(a+bx)}}}$$

$$a + bx = \ln\left(\frac{P(D)}{1-P(D)}\right) \rightarrow \text{odd of disease}$$

$$\theta = \{a, b\}$$

$$y = \frac{1}{1+e^{-(a+bx)}}$$

Posterior prob. of x ← $P(x/y) = \dfrac{P(y/x) \cdot P(x)}{P(y)}$ ← Prior

$d \rightarrow$ data

$\theta \rightarrow$ arbitrary value of a & b

$$P(\theta = Y / d = Y) = \frac{P(d = Y | \theta = Y)\, P(\theta = Y)}{P(d = Y)}$$

$$P(\theta | d) = \frac{P(d|\theta) \cdot P(\theta)}{P(d)}$$

$P(d)$ → probability of the data is correct.

We have a belief / prob. of θ ($P(\theta)$). & we see the data (d). Then we calculate the prob. of θ being correct given that we have seen the data ($P(\theta/d)$)

Lec (1/4)

Logistic Regression → classifier

Every classifier starts with labelled data.

We try to fit a sigmoidal curve.

$$\boxed{y = \frac{1}{1+e^{-(a+bx)}}} \rightarrow \text{Logistic eq}^n$$

$$P(D) = \frac{1}{1+e^{-(a+bP)}}$$

Lec (5/4/2023)

Estimating a & b in logistic eqn then to data is done by maximum Likelihood method.

$\left\{ \theta \Rightarrow \text{set of Parameters} \right.$

Baysian thm. $\Rightarrow$ $P(Y/X) = \dfrac{P(X/Y) \cdot P(Y)}{P(X)}$

We are Maximizing the probability of ~~data~~ θ given that we ~~with given θ~~ observed the data $\Rightarrow P(\theta \mid D)$

$\cancel{P(D \mid \theta) = P(\theta \mid D) \cdot P(D)}$

Posterior prob. $\leftarrow$ $P(\theta \mid D) = \dfrac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$

Prior { Probability of all θ will be most likely }

Likelihood of θ (Prob. of Data given θ)

$P(D)$ $\to$ Marginal

$[P(D \mid \theta) = L(\theta)$ $\{P(D) = P(D \mid \theta_1) \cdot P(\theta_1) + P(D \mid \theta_2) \cdot P(\theta_2) + \ldots$

For maximizing $P(\theta \mid D)$, we have to Maximize $P(D \mid \theta)$ i.e the Likelihood (coz. $P(\theta)$ & $P(D)$ are const.)

ex. Scenario $\Rightarrow$ 10 coin tosses, 4 heads.

$D = \begin{matrix} n = 10 \\ h = 4 \end{matrix}$, $\theta = \{P_H\}$

$P(D \mid \theta) \Rightarrow$ Prob. of getting D with given P_H (by binomial)

$\therefore P(D \mid \theta) = {}^{n}C_{h} \, P_H^{\,h} \, (1 - P_H)^{n-h}$

Likelihood of θ, we need to find θ which will maximize the prob.

D

Y	X
0	x_1
1	x_2
⋮	⋮
y_i	x_i
	⋮
	x_n

$y_i = 0 \text{ or } 1$

$\theta = \{a, b\}$

$\max(P(D \mid \theta))$

Probability of y_i in Class 1 $= P(x_i) = P_i$

$= \dfrac{1}{1 + e^{-(a + b x_i)}}$

$P(y_i \text{ in class } 0) = 1 - P_i$

then $\to$

$x_1 \Rightarrow 1 \longrightarrow P(x = x_1) / \sigma(x_1)$

$x_2 \to 1 \longrightarrow \sigma(x_2)$

$x_3 \to 0 \longleftarrow 1 - \sigma(x_3)$

$\boxed{\prod_{i=1}^{n} \sigma(x_i)^{y_i} \left[1 - \sigma(x_i)\right]^{1 - y_i}}$ $\to$ for all ~~data~~ n

$$= P(D|\theta) = P(D|a,b)$$

Now, we have to maximize this value

major problem in it is that the value will be very small & it will be difficult to compute for the machine.

Therefore we take $\ln$.

$$\therefore \ \underline{\ln(L(\theta))}$$

& Max of $\ln(L(\theta))$ will be max of $\boxed{L(\theta)}$

Algorithm to use $\Rightarrow$ Gradient Descent or Ascent.