2009	2019
$Y = \beta X + \varepsilon$	$Y = \beta X + \varepsilon$
STATISTICS	MACHINE LEARNING

\# 10 YEARS CHALLENGE

LINEAR REGRESSION

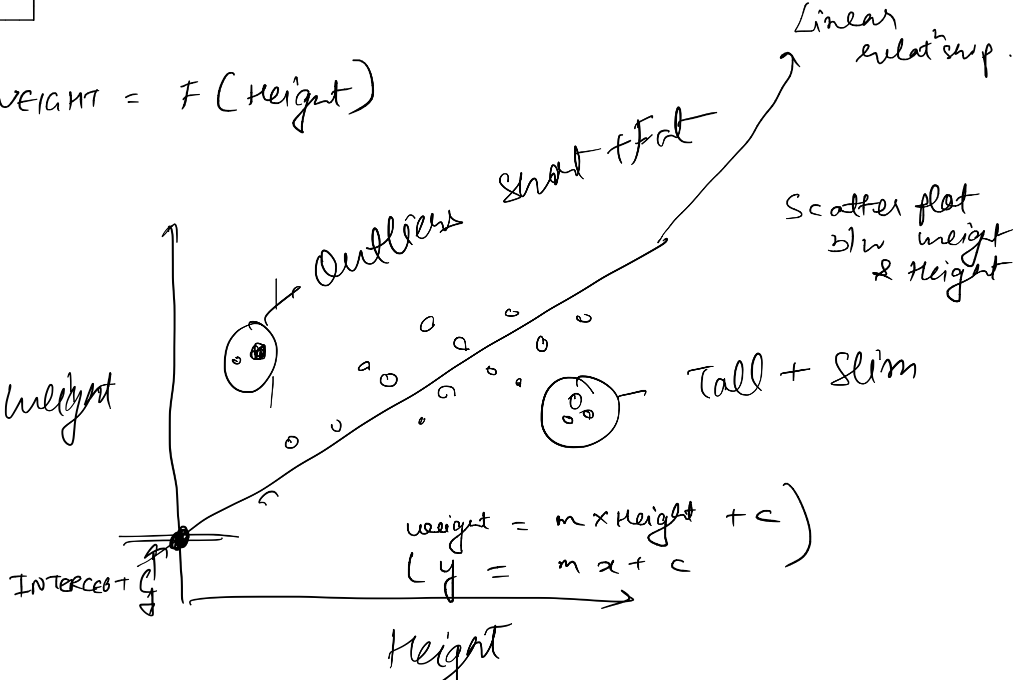
Linear Regression → o/p is continuous in nature

$$F = m * a$$
$$V = u + at$$

Some linear sytems

TASK: PREDICT WEIGHT = F(Height)

Height	weight
3	25
5	8
,	,
,	,



weight = m × Height + c
$$(y = mx + c)$$

WRITING THE EXPRESSION IN MATRIX FORM

$$weight_i = 1*\theta_0 + height_i*\theta_1$$

$$\begin{array}{lcl}
weight_1 & = & 1*\theta_0 + height_1*\theta_1 \\
weight_2 & & \vdots \\
\vdots & & \vdots \\
weight_N & & 1*\theta_0 + height_N*\theta_1
\end{array}$$

weight

N examples

$$\underbrace{\begin{bmatrix} \\ \\ \\ \end{bmatrix}}_{N\times 1} = \underbrace{\begin{bmatrix} 1 & height_1 \\ \vdots & \vdots \\ 1 & height_N \end{bmatrix}}_{N\times 2} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}}_{2\times 1}$$

$\theta_0 \rightarrow$ Bias

$\theta_1 \rightarrow$ Intercept

Previous example was $y = F(x)$ where x is one-dimensional.

Examples in multiple dimensions

IIT GN water demand = F (# occupants, Temperature)

Demand = Base demand + $K_1 \times$ # Occupants + $K_2 \times$ Temp.

we expect

Demand ↑ if # occupants ↑ ⟹ K_1 likely positive

Demand ↑ if Temp ↑ ⟹ K_2 likely positive

$\underbrace{\text{Base demand}}_{\text{Bias}}$ is demand <u>independent</u> of temp. and # occupants.

MORE GENERALLY

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}_{N \times 1} = \begin{array}{c} \\ \overbrace{\hphantom{x_{11} \quad x_{12} \quad \cdots}}^{M \text{ features}} \end{array} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots \\ 1 & x_{21} & \cdots & \\ 1 & & & \\ 1 & x_{N1} & & x_{NM} \end{bmatrix}_{(N \times (M+1))} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_M \end{bmatrix}_{(M+1 \times 1)}$$

$$y_{N \times 1} \approx X_{N \times (M+1)} \, \theta_{(M+1) \times 1}$$

UNKNOWN

KNOWN

N knowns (or N equations)

M unknowns (or M parameters / variables)

TRIVIAL CASE (Back to the weight example)

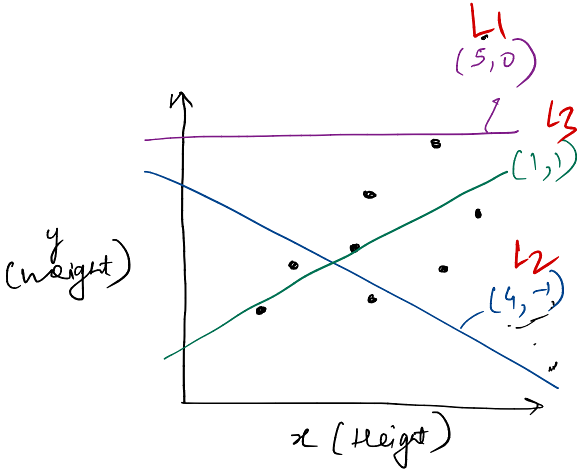
$$y_{N\times 1} \approx X_{N\times(m+1)}\,\theta_{(m+1)\times 1}$$

$m = 1$

$$\therefore \quad y = \theta_1 x + \theta_0$$

(same form as $y = mx + c$

For different θ_0 & θ_1 different relationship can be learnt.

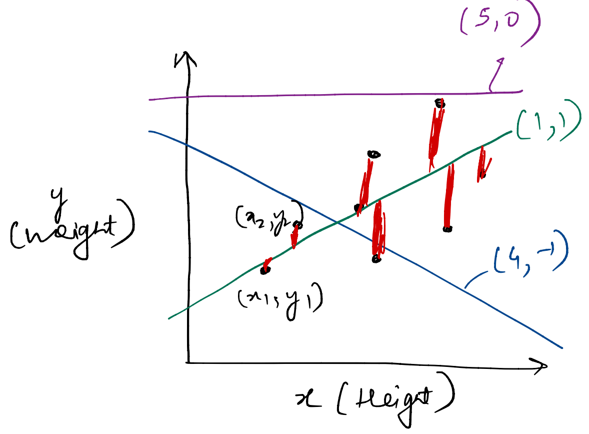


3 examples

θ_0	θ_1
5	0
4	-1
1	1

} which is best?

For different θ_0 & θ_1 different relationship can be learnt.



3 (of ∞ possible parameters)

θ_0	θ_1
5	0
4	-1
1	1

} which is best?

CHOOSE θ_0, θ_1 s.t.

$|$error $(e_i)|$ or $(e_i)^2$ are reduced, where

$$e_i = y_i - \hat{y}_i$$

Residual G.T. Pred.

$$e_i = y_i - (\theta_0 + \theta_1 * x_i)$$

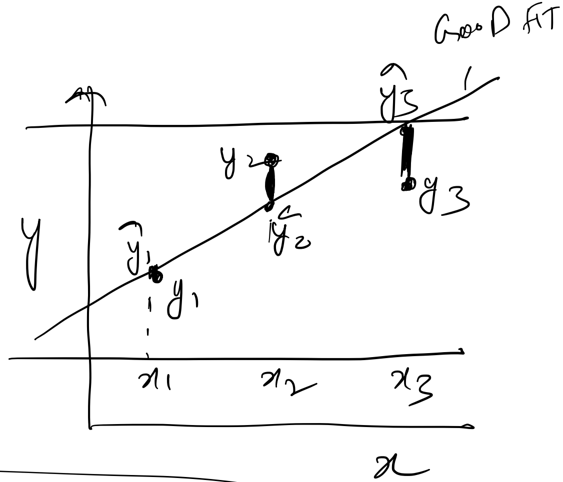
θ_0, θ_1 : Same value for all examples

e_i : Variable residual per example

For Good Fit

$|e_1|, |e_2|, |e_3| \ldots$

all should be small



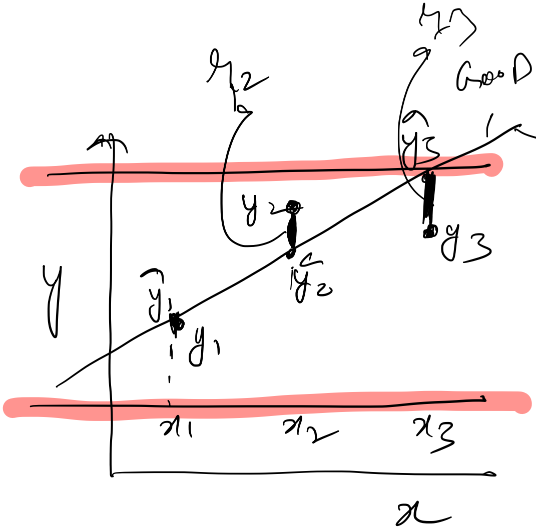
Minimize $(e_1)^2 + (e_2)^2 + \ldots (e_N)^2$ L_2

Or

$|e_1| + |e_2| + |e_N|$ L_1

TRAIN DATA

x_1	y_1
x_2	y_2
x_3	y_3



$mx + c$

$X\theta$

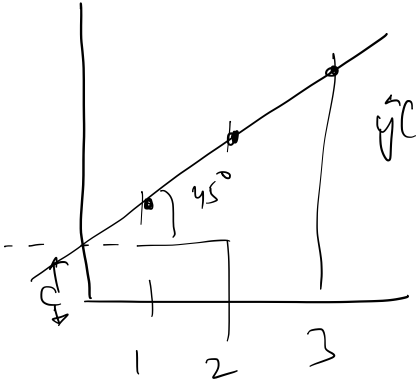
$\theta_0 + x_1 * \theta_1$

$$y = X\theta + e$$

I/p data, Bias term

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}$$

$$\hat{y} = (mx + c)$$



$\hat{y} = 1 + x$

$\hat{y}(x) = c + mx$

$\hat{y}(1) = 2 = y(1)$

$\hat{y}(2) = 3 = y(2)$

$e_1 = \hat{y}(1) - y(1) = 0$

NORMAL EQN

$$\underset{\sim}{y} \approx X\theta \;(+\,\epsilon)$$

Vector

Matrix

To learn : θ

Objective: Minimize $\epsilon_1^2 + \dots, \epsilon_N^2$

$$\epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{bmatrix} \qquad \epsilon^T = [\epsilon_1, \dots \quad \epsilon_N]$$

Equivalent to.

Minimize $\epsilon^T \epsilon$

$X : N \times (M+1)$

$\theta : (M+1) \times 1$

$$e_y = y - X\theta$$

$$e_y^T = (y - X\theta)^T = y^T - \theta^T X^T$$

$$e_y^T e_y = (y^T - \theta^T X^T)(y - X\theta)$$

$$e_y^T e_y = \underbrace{y^T}_{1\times N} \underbrace{y}_{N\times 1} - \underbrace{\theta^T}_{1\times(M+1)} \underbrace{X^T}_{(M+1)\times N} \underbrace{y}_{N\times 1} - y^T X\theta + \theta^T X^T X\theta$$

Both these are the same

$$= y^T y - 2y^T X\theta + \theta^T X^T X\theta$$

Objective minimize $e_y^T e_y = y^T y - 2y^T X\theta + \theta^T X^T X\theta$

Objective minimize $e_y^T e_y$ w.r.t. θ $= \overset{①}{y^T y} - \overset{②}{2y^T X\theta} + \theta^T X^T X \theta$

$$\boxed{\frac{\partial e_y^T e_y}{\partial \theta} = 0}$$

① $\dfrac{\partial}{\partial \theta} y^T y = \vec{0}$

② $\dfrac{\partial}{\partial \theta} (\underbrace{-2y^T X}_{A}\theta) = A^T = (-2y^T X)^T$

$= -2X^T y$

③ $$\frac{\partial}{\partial \theta}\left(\theta^T \underbrace{X^T X}\theta\right) = 2X^T X \theta$$

$$0 = -2X^T y + 2X^T X \theta$$

$$X^T y = X^T X \theta \quad \ldots\ldots$$

$$\boxed{(X^T X)^{-1} X^T y = \theta}$$

NORMAL EQ^N

Relationship b/w # variables & # examples
(M) (N)

(ii) $N < M$

example

$$weight_i = \theta_0 + height_i \times \theta_1 + Age_i \times \theta_2$$

$N = 2$
$m = 3$

$$\begin{bmatrix} 30 \\ 40 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 30 \\ 1 & 5 & 20 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$$

$$30 = \theta_0 + 6\theta_1 + 30\theta_2$$
$$40 = \theta_0 + 5\theta_1 + 20\theta_2$$

$$-10 = -\theta_1 - 10\theta_2$$

How many solⁿ ? → Infinite !!

under-determined system : $e_i = 0 \ \forall i$

(ii) $N > M$

over-determined

Sum of squared residuals > 0

Question 1

we had

$$X^T y = X^T X \theta$$

can we do?

$$\underbrace{(X^T)^{-1} X^T}_{I} y = \underbrace{(X^T)^{-1} X^T}_{I} X \theta$$

$$y = X\theta$$

$$X^{-1} y = \theta$$

X may not be a square matrix

X^{-1} may not exist.

Left inverse

$$HX = I_{M+1}$$

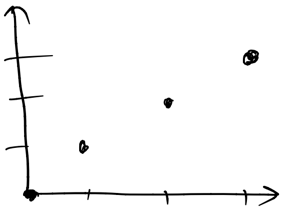
$$X_{N\times(M+1)}$$

Right inverse

$$XH = I_N$$

WORKED OUT EXAMPLE

x	y
0	0
1	1
2	2
3	3



Find θ_0 & θ_1

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix}$$

$$X^T Y = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

$$|X^T X| = 14 \times 4 - 6 \times 6 = 20$$

$$(X^T X)^{-1} = \frac{\operatorname{adj}(X^T X)}{|X^T X|} = \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}$$

$$\operatorname{adj}(X^T X) = \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \qquad \therefore \quad (X^T X)^{-1} = \frac{1}{20}\begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}$$

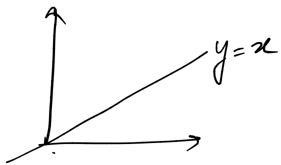
$$X^T y = \underset{2\times 4}{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 \end{bmatrix}} \underset{4\times 1}{\begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix}} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

$$\theta = (X^T X)^{-1}(X^T y) = \frac{1}{20}\begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix}\begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

$$= \frac{1}{20}\begin{bmatrix} 14\times 6 - 6\times 14 \\ -6\times 6 + 4\times 14 \end{bmatrix} = \frac{1}{20}\begin{bmatrix} 0 \\ 20 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

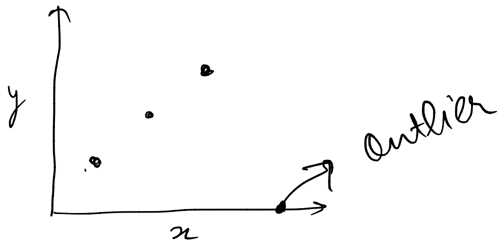
$$= \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix}$$

$$\Rightarrow \quad y = 0 + x$$



CLASS EXERCISE : learn θ_0 & θ_1 (EFFECT OF OUTLIER)

x	y
1	1
2	2
3	3
4	0



$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \qquad X^T = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 4 & 10 \\ 10 & 30 \end{bmatrix}$$

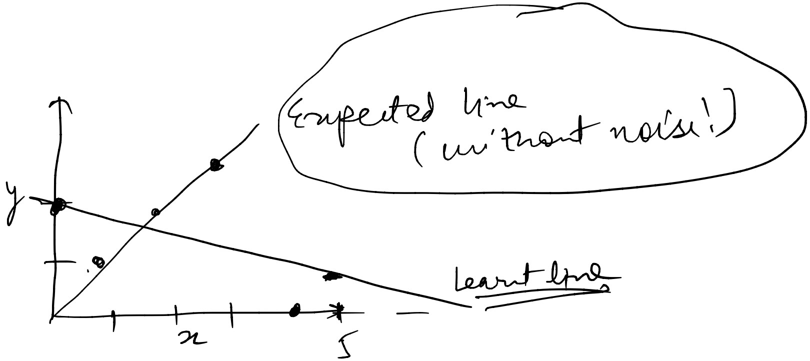
$$|(X^T X)| = 20 \qquad (X^T X)^{-1} = \frac{1}{20} \begin{bmatrix} 30 & -10 \\ -10 & 4 \end{bmatrix}$$

$$X^T Y = \underset{2 \times 4}{\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 \end{bmatrix}} \underset{4 \times 1}{\begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \end{bmatrix}} = \begin{bmatrix} 6 \\ 14 \end{bmatrix}$$

$$(X^T X)^{-1}(X^T y) = \frac{1}{20}\begin{bmatrix}30 & -10\\ -10 & 4\end{bmatrix}\begin{bmatrix}6\\ 14\end{bmatrix} = \frac{1}{20}\begin{bmatrix}180-140\\ 56-60\end{bmatrix} = \frac{1}{20}\begin{bmatrix}40\\ -4\end{bmatrix}$$

$$= \begin{bmatrix}2\\ -1/5\end{bmatrix}$$

$$y = 2 - \frac{1}{5}x$$



NEED A WAY TO HANDLE NOISE / OUTLIERS?

CLASS EXERCISE : SOLVE LINEAR REGRESSION

$$\begin{array}{ccc} x_1 & x_2 & y \\ \end{array}$$
$$\left[\begin{array}{cc|c} 1 & 2 & 4 \\ 2 & 4 & 6 \\ 3 & 6 & 8 \end{array}\right]$$

$$\left\{\begin{array}{l} y = 2x_1 + 2 \\ y = x_2 + 2 \\ y = \frac{x_2}{2} + x_1 + 2 \end{array}\right\}$$

$$X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 4 \\ 1 & 3 & 6 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 12 \\ 6 & 14 & 28 \\ 12 & 28 & 56 \end{bmatrix}$$

$$|X^T X| = 3(14 \times 56 - 28 \times 28) - 6(6 \times 56 - 28 \times 12) + 12(6 \times 28 - 14 \times 12)$$
$$= 0$$

$$\boxed{(X^T X)^{-1} \text{ does not exist}}$$

$\Rightarrow$ Matrix X is not full rank

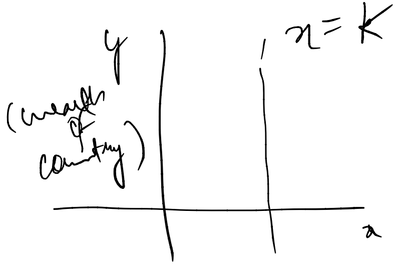
Multi-collinearity: One (or more) predictor variable / feature in X can be expressed as linear combinations of others

SOLVE (Many ways)

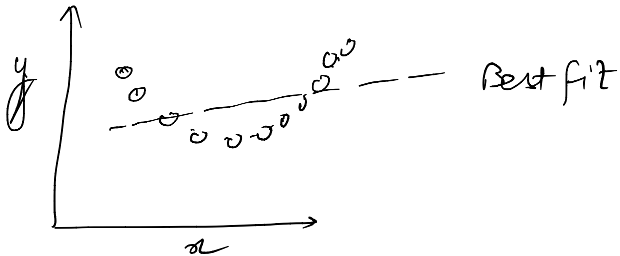
1. Regularize
2. Drop variables
3. Use different subsets of data
4. Avoid dummy variable trap

SOME THINGS TO KNOW ABOUT LINEAR REGRESSION

1) when $x = K$



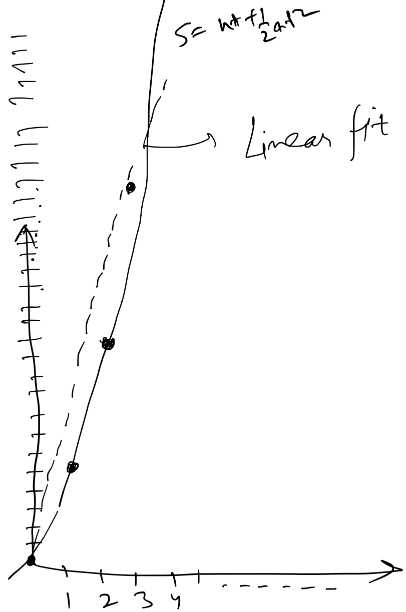
2) when relationship is non-linear



MODELING NON-LINEARITIES

Dataset

t	s
0	0
1	6
2	14
3	24
4	36



TRANSFORM

t	t^2	s
0	0	0
1	1	6
2	4	14
3	9	24
4	16	36

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \end{bmatrix}$$

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \\ 0 & 1 & 4 & 9 & 16 \end{bmatrix}$$

MODELING INTERACTIONS

$$y = \theta_0 + \theta_1 * x_1 + \cdots \theta_m * x_m$$

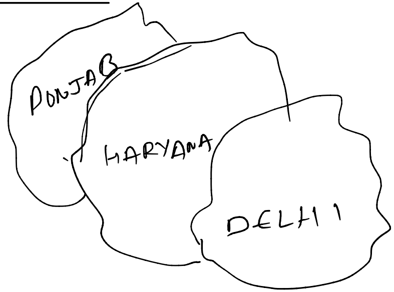
If x_1 increases by 1 unit, y increases by θ_1 units

<u>irrespective</u> of $x_2, \ldots x_m$

$$y = \theta_0 + \theta_1 * x_1 + \theta_2 * \underbrace{x_1 * x_2}_{\text{Interact}^n \text{ term}} + \theta_3 * x_2$$

$$= \theta_0 + \theta_1 * x_1 + \theta_2 * x_2' + \theta_3 * x_2 \quad (\text{Still linear})$$

DUMMY VARIABLES



POLLUTION in DELHI $= \theta_0 + \theta_1 \times$ #VEHICLES $+ \theta_2 \times$ WIND SPEED $+ \theta_3 \times$ WIND DIRECTION

(N, E, W, S)

Can we encode wind direction as: {N: 0, E: 1, W: 2, S: 3} ?

No: Implies $S > W > E > N$
(gives some ordering b/w them)

	Is it N?	Is it E?	Is it W?
N	1	0	0
E	0	1	0
W	0	0	1
S	0	0	0

ENCODING
(USES N-1 variables for N classes)

why not?

	Is N	Is E	Is W	Is S
N	1	0	0	0
E	0	1	0	0
W	0	0	1	0
S	0	0	0	1

$$x_5 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} - x_2 - x_3 - x_4$$

<u>Dataset</u>

wind direcⁿ	pollⁿ
N (1000)	1
E (0100)	2
W (0010)	3
S (0001)	4

$$\begin{array}{c} \\ X = \end{array} \begin{array}{c} x \quad x_2 \quad x_3 \quad x_4 \quad x_5 \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array} \begin{array}{l} N \\ E \\ W \\ S \end{array}$$

Multi-collinearity!!

WHY DUMMY IS ONE-HOT ENCODING & NOT BINARY

BINARY

			vs						
N	—	0 0		N —	1	0	0	0
E	—	0 1		E —	0	1	0	0
W	—	1 0		W —	0	0	1	0
S	—	1 1		S —	0	0	0	0

W & S are related by high bit

THIS INTRODUCES DEPENDECIES B/W THEM

CAN CONFUSE CLASSIFIERS, ETC...

INTERPRETING DUMMY VARIABLES

Gender	Height
F	
M	
F	
F	
M	
M	

$\Longrightarrow$

is F?	Height
1	
0	
1	
1	
0	
0	

$$h_i = \theta_0 + \theta_1 \times (\text{is}F?) + \epsilon_i$$

$$h_i = \begin{cases} \theta_0 + \theta_1 + \epsilon_i & : \text{if female} \\ \theta_0 + \epsilon_i & : \text{if male} \end{cases}$$

$$\begin{cases} \theta_0 = \text{Avg. height of male} \\ \theta_0 + \theta_1 = \text{Avg. height of female} \end{cases}$$

$\theta_1 =$ Difference in avg. of female and male heights

	Is it N?	Is it E?	Is it W?	Poll^n
N	1	0	0	4
E	0	1	0	1
W	0	0	1	1
S	0	0	0	1

$$\text{Pollution} = \theta_0 + \theta_1 (\text{Is it N}) + \theta_2 (\text{Is it E}) + \theta_3 \times (\text{Is it W})$$

$$P_i = \begin{cases} \theta_0 + \epsilon_i & \text{: if wind = South} \\ \theta_0 + \theta_1 & \text{: } \ldots = \text{North} \\ & \vdots \end{cases}$$

θ_0 : South Avg. Pollut^n

θ_1 = Difference in avg. of north − South . . .

θ_2 = − −

θ_3 = − − .

Alternative Parameter Estimation (for Linear Reg. in 2 variables)

$$\hat{y}_i = \theta_0 + \theta_1 x_i \quad \text{--- (1)}$$

$$e_i = y_i - \hat{y}_i = \underline{y_i - \theta_0 - \theta_1 x_i}$$

$$\underline{\sum e_i^2} = \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_i)^2$$

Derivative $$\frac{\partial \sum e_i^2}{\partial \theta_0} = 2 \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_i)(-1) = 0$$

$$\Rightarrow \sum_{i=1}^{N} y_i - \theta_1 \sum_{i=1}^{N} x_i - N\theta_0 = 0$$

$$\Rightarrow \boxed{\theta_0 = \frac{\sum_{i=1}^{N} y_i - \theta_1 \sum_{i=1}^{N} x_i}{N} = \bar{y} - \theta_1 \bar{x}}$$

$$\sum e_{y_1}^2 = \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_i)^2$$

Differentiate: $\frac{\partial \sum e_i^2}{\partial \theta_1} = 0 \Rightarrow 2 \sum_{i=1}^{N} (y_i - \theta_0 - \theta_1 x_i)(-x_i) = 0$

$$\Rightarrow \sum (x_i y_i - \theta_0 x_i - \theta_1 x_i^2) = 0$$

$$\Rightarrow \sum \theta_1 x_i^2 = \sum x_i y_i - \sum \theta_0 x_i$$

$$\Rightarrow \sum \theta_1 x_i^2 = \sum x_i y_i - \sum x_i (\bar{y} - \theta_1 \bar{x})$$

$$\Rightarrow \sum \theta_1 x_i^2 = \sum x_i y_i - \bar{y} \sum x_i + \theta_1 \bar{x} \sum x_i$$

$$\Rightarrow \sum x_i y_i - \sum x_i \bar{y} = \theta_1 \left(-\bar{x} \sum x_i + \sum x_i^2\right)$$

$$\Rightarrow \sum x_i y_i - \sum x_i \bar{y} = \theta_1 \left(-\bar{x} \sum x_i + \sum x_i^2\right)$$

$$\theta_1 = \frac{\sum x_i y_i - \sum x_i \bar{y}}{\sum x_i^2 - \bar{x} \sum x_i}$$

$$\theta_1 = \frac{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Variance}(x)}$$