MACHINE LEARNING: USING OVERFITTING TO EVALUATE LINEAR REGRESSION MODEL AND NON-LINEAR REGRESSION

DONE BY: SAHITI EMANI
STUDENT ID:19556
GUIDED BY:PROF.HENRY CHANG

TABLE OF CONTENTS:

INTRODUCTION:

I. What is regression?

A regression is a statistical analysis assessing the association between two variables. It is used to find the relationship between two variables.

II. What are the types of regression?

There are two types of regression: Linear and Nonlinear.

III. What is meant by linear regression?

It is the most **widely used statistical technique**. It is a relationship between two sets of variables which results in a linear regression equation that is to make predictions about data.

IV. What is meant by non-linear regression?

It is a statistical technique that is used to describe non-linear relationships about the acquired experimental data.

LINEAR REGRESSION:

- Linear Regression is a **supervised machine learning algorithm** , we can say that the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. cat, dog).
- **SIMPLE LINEAR REGRESSION**:This uses traditional slope-intercept form, where m and b are the variables our algorithm will try to produce the most accurate predictions.
- x= represents our input data and y= represents our prediction.

y=mx+b

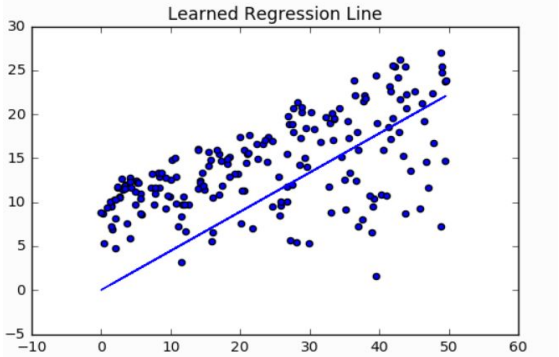
Company	Radio ($)	Sales
Amazon	37.8	22.1
Google	39.3	10.4
Facebook	45.9	18.3
Apple	41.3	18.5

A dataset is given about how much a company spends on radio advertising each year and its annual sales.

Our prediction function outputs an estimate of sales with current W and B.i.e:**Sales=Weight*Radio+Bias**

Our algorithm will try to *learn* the correct values for Weight and Bias. By the end of our training, our equation will approximate the *line of best fit*.

The Fig for above data:

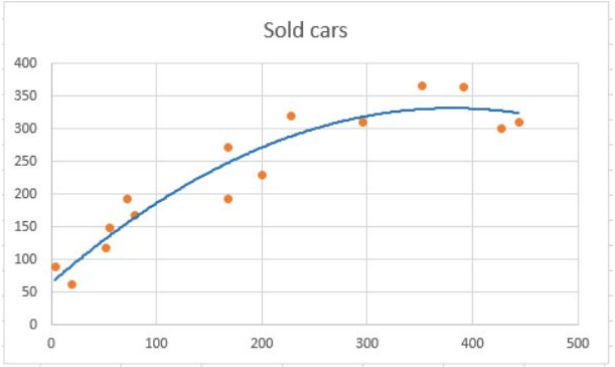


NON-LINEAR REGRESSION MODELS:

- Non-Linear regression is one type of polynomial regression.
- Nonlinear regression models are usually assumed to be parametric.
- Independent and dependent variables used in nonlinear regression should be quantitative.
- When we say a Parametric nonlinear regression models the dependent variable (also called the response) as a function of a combination of nonlinear parameters and one or more independent variables (called predictors).
- A model can be univariate (single response variable) or multivariate (multiple response variables).

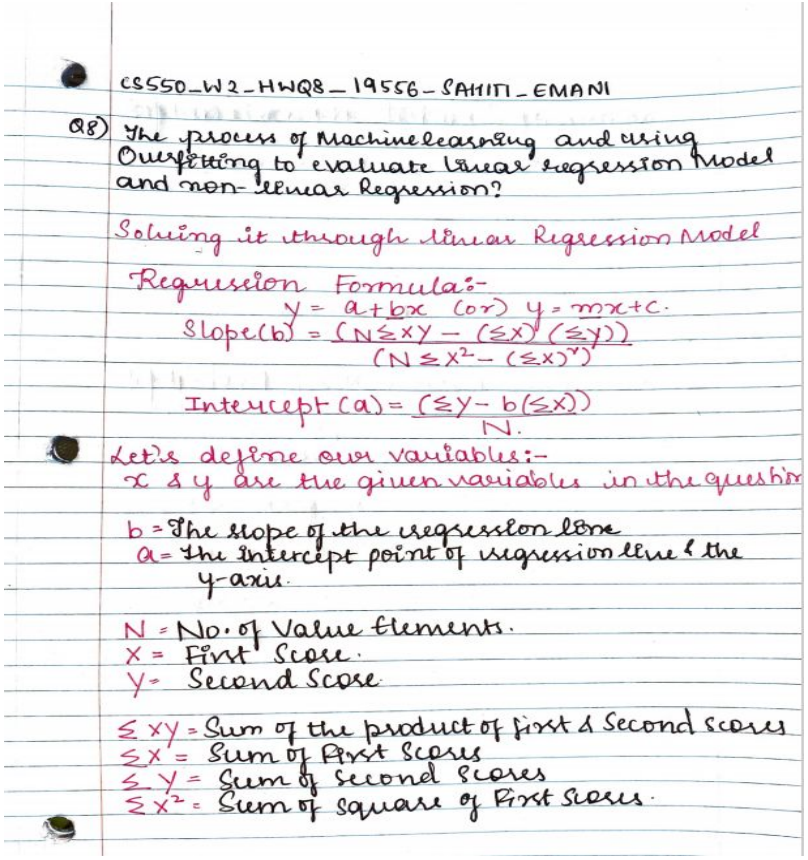
	No of weeks	No of weeks ^2	Sold cars
0	168	28224	272
1	428	183184	300
2	296	87616	311
3	392	153664	365
4	80	6400	167
5	56	3136	149
6	352	123904	366
7	444	197136	310
8	168	28224	192
9	200	40000	229
10	4	16	88
11	52	2704	118
12	20	400	62
13	228	51984	319
14	72	5184	193

Eg: A CURVILINEAR PLOT WHEN ITS A NON-LINEAR MODEL



Polynomial trend line over the scatter plot

EXAMPLE AND CALCULATION:



CS550_W2_HWQ8_19556_SAHITI_EMANI

Q8) The process of Machine learning and using Overfitting to evaluate linear regression model and non-linear Regression?

Solving it through linear Regression Model

Regression Formula:-

$y = a+bx$ (or) $y = mx+c$.

$$\text{Slope}(b) = \frac{(N\Sigma xy - (\Sigma x)(\Sigma y))}{(N\Sigma x^2 - (\Sigma x)^2)}$$

$$\text{Intercept}(a) = \frac{(\Sigma y - b(\Sigma x))}{N}$$

Let's define our variables:-

x & y are the given variables in the question

b = The slope of the regression line

a = the intercept point of regression line & the y-axis.

N = No. of Value Elements.

X = First Score.

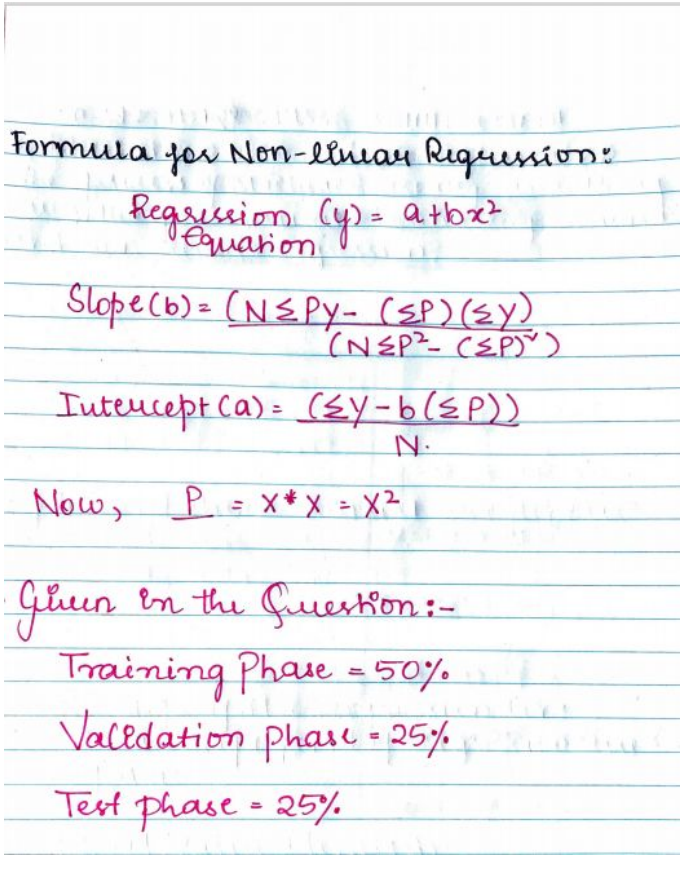
Y = Second Score

Σxy = Sum of the product of first & Second scores

Σx = Sum of First Scores

Σy = Sum of second scores

Σx^2 = Sum of square of First scores.



Formula for Non-linear Regression:

Regression Equation $(y) = a+bx^2$

$$\text{Slope}(b) = \frac{(N\Sigma Py - (\Sigma P)(\Sigma y)}{(N\Sigma P^2 - (\Sigma P)^2)}$$

$$\text{Intercept}(a) = \frac{(\Sigma y - b(\Sigma P))}{N}$$

Now, $P = X * X = X^2$

Given in the Question:-

Training Phase = 50%

Validation phase = 25%

Test phase = 25%

CONT'D:

Training Phase :

→ Calculating for Linear Regression:-

Training phase

Model – 1: Linear Regression.

x	y	$\hat{y} = a1 + b1 * x$
1	1.8	$y = 0.51 + 0.86(1) = 1.37$
2	2.4	$y = 0.51 + 0.86(2) = 2.74$
3.3	2.3	$y = 0.51 + 0.86(3.3) = 3.348$
4.3	3.8	$y = 0.51 + 0.86(4.3) = 4.208$
5.3	5.3	$y = 0.51 + 0.86(5.3) = 5.068$
1.4	1.5	$y = 0.51 + 0.86(1.4) = 1.714$
2.5	2.2	$y = 0.51 + 0.86(2.5) = 2.66$
2.8	3.8	$y = 0.51 + 0.86(2.8) = 2.918$
4.1	4.0	$y = 0.51 + 0.86(4.1) = 4.036$
5.1	5.4	$y = 0.51 + 0.86(5.1) = 4.8896$.

To find the regression equation we will first find slope, intercept & use it to form regression equation :-

Step 1 : Count the no. of values.

N = 10

Step 2 : Find x * y & x^2.

x-Value	y-Value	x * y	x * x.
1	1.8	$1 \times 1.8 = 1.8$	$1 \times 1 = 1$
2	2.4	$2 \times 2.4 = 4.8$	$2 \times 2 = 4$
3.3	2.3	$3.3 \times 2.3 = 7.59$	$3.3 \times 3.3 = 10.89$
4.3	3.8	$4.3 \times 3.8 = 16.34$	$(4.3)^2 = 18.49$
5.3	5.3	$5.3 \times 5.3 = 28.09$	$(5.3)^2 = 28.09$
1.4	1.5	$1.4 \times 1.5 = 2.1$	$(1.4)^2 = 1.96$
2.5	2.2	$2.5 \times 2.2 = 5.5$	$(2.5)^2 = 6.25$
2.8	3.8	$2.8 \times 3.8 = 10.64$	$(2.8)^2 = 7.84$
4.1	4.0	$4.1 \times 4.0 = 16.4$	$(4.1)^2 = 16.81$
5.1	5.4	$5.1 \times 5.4 = 27.54$	$(5.1)^2 = 26.01$

CONT'D:

Step 3:- Find $\Sigma x, \Sigma y, \Sigma xy, \Sigma x^2$.

$\Sigma x = 31.8$ $\quad \Sigma xy = 120.8$

$\Sigma y = 32.5$ $\quad \Sigma x^2 = 121.34$

Step 4:- Substitute the values in the given Formula:

$$\text{Slope}(b) = (N\Sigma xy - (\Sigma x)(\Sigma y))/(N\Sigma x^2 - (\Sigma x)^2)$$

$$= \frac{(10)(120.8) - (31.8)(32.5)}{(10)(121.34) - (31.8)^2}$$

$$= \frac{1208 - 1033.5}{1213.4 - 1011.24}$$

$$= \frac{174.5}{202.16} = \boxed{0.8631 = b}$$

Step 5:- Intercept (a).

$$(a) = \frac{(\Sigma y - b(\Sigma x))}{N}$$

$$= \frac{32.5 - 0.86(31.8)}{10}$$

$$\boxed{a = 0.5152 \approx 0.51}$$

Step 6:- Substitute the values of a & b in the regression formula:-

$$(y) = a + bx.$$

$$y = 0.51 + 0.86x.$$

Step 7:- Suppose if we have to calculate the approximate value of y for variable

(i) $x = 1$

$(y) = a + bx = 0.51 + 0.86(1) = 1.37$

(ii) $x = 2$

$(y) = a + bx = 0.51 + 0.86(2) = 2.231$

(iii) $x = 3.3$

$(y) = a + bx = 0.51 + 0.86(3.3) = 3.348$

(iv) $x = 4.3$

$(y) = a + bx = 0.51 + 0.86(4.3) = 4.208$

(v) $x = 5.3$

$(y) = a + bx = 0.51 + (5.3)(0.86) = 5.068$

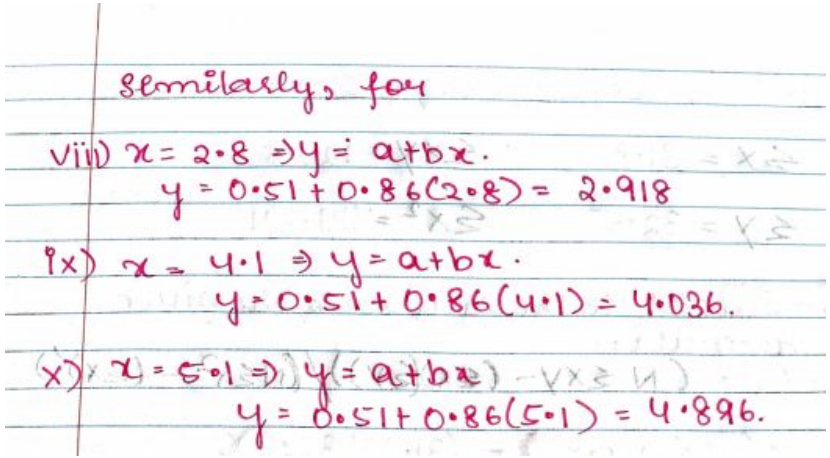
(vi) $x = 1.4$

$(y) = a + bx = 0.51 + 0.86(1.4) = 1.714$

(vii) $x = 2.5$

$(y) = a + bx = 0.51 + 0.86(2.5) = 2.66$

CONT'D:



Similarly, for

viii) $x = 2.8 \Rightarrow y = a + bx$.

$y = 0.51 + 0.86(2.8) = 2.918$

ix) $x = 4.1 \Rightarrow y = a + bx$.

$y = 0.51 + 0.86(4.1) = 4.036$.

x) $x = 5.1 \Rightarrow y = a + bx$

$y = 0.51 + 0.86(5.1) = 4.896$.

CONT'D:

Calculating the Training phase in the form of a non-linear model:-

Regression Equation: $(y) = a + bx^2$

Slope $(b) = \frac{(N \leq Py - (\leq P)(\leq y))}{(N \leq P^2 - (\leq P)^2)}$

Intercept $(a) = \frac{(\leq y - b(\leq P))}{N}$ [∵ where $P = x * x = x^2$]

TP ← Model 2 → Non-linear Regression.

x	y	$y = a2 + b2 * x^2$
1	1·8	$1.67 + 0.13(1)^2 = 1.8$
2	2·4	$1.67 + 0.13(2)^2 = 2.19$
3·3	2·3	$1.67 + 0.13(3.3)^2 = 3.0857$
4·3	3·8	$1.67 + 0.13(4.3)^2 = 4.0737$
5·3	5·3	$1.67 + 0.13(5.3)^2 = 5.3217$
1·4	1·5	$1.67 + 0.13(1.4)^2 = 1.9248$
2·5	2·2	$1.67 + 0.13(2.5)^2 = 2.4825$
2·8	3·8	$1.67 + 0.13(2.8)^2 = 2.6892$
4·1	4·0	$1.67 + 0.13(4.1)^2 = 3.8553$
5·1	5·4	$1.67 + 0.13(5.1)^2 = 5.0513$

Non-linear Method:-

Step-1: No. of Values N=10

Step-2: Find x*x

x-Values	x Values
1	1
2	4
3·3	10·89
4·3	18·49
5·3	28·09
1·4	1·96
2·5	6·25
2·8	7·84
4·1	16·81
5·1	26·01

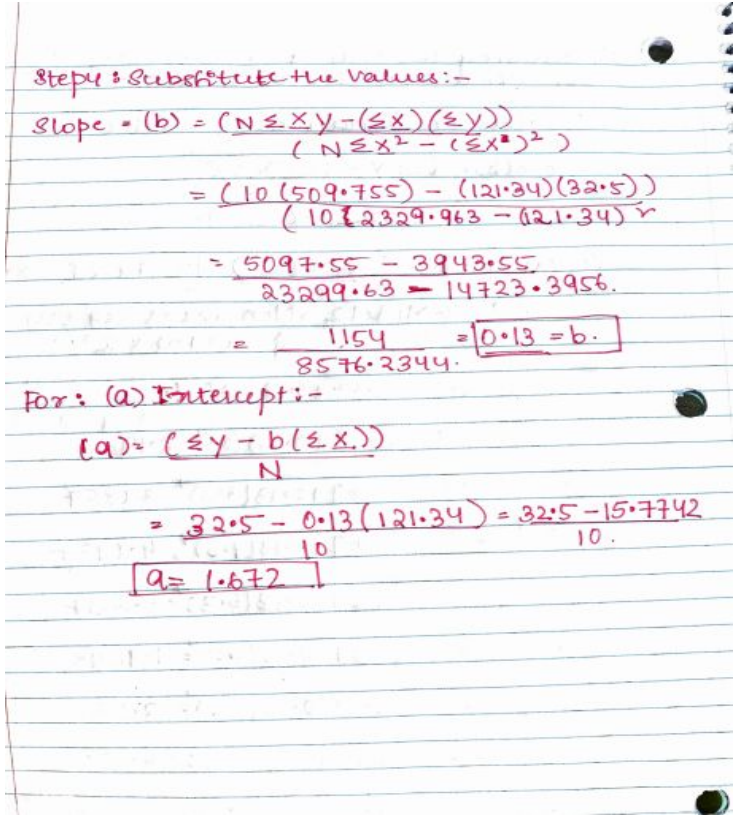
Step3: Find x*y & x^2.

x Value	y Value	x * y	x*x = x^2
1	1·8	1×1·8 = 1·8	$(1)^2 = 1$
4	2·4	4×2·4 = 9·6	$(4)^2 = 16$
10·89	2·3	10·89×2·3 = 25·04	$(10·89)^2 = 118·592$
18·49	3·8	18·49×3·8 = 70·262	$(18·49)^2 = 341·88$
28·09	5·3	28·09×5·3 = 148 877	$(28·09)^2 = 789·04$
1·96	1·5	1·96×1·5 = 2·94	$(1·96)^2 = 3·841$
6·25	2·2	6·25×2·2 = 13·75	$(6·25)^2 = 39·06$
7·84	3·8	7·84×3·8 = 29·792	$(7·84)^2 = 61·46$
16·81	4·0	16·81×4·0 = 67·24	$(16·81)^2 = 282·57$
26·01	5·4	26·01×5·4 = 140·454	$(26·01)^2 = 676·52$

$\leq x = 121.34$ $\quad \leq xy = 509.755$

$\leq y = 32.5$ $\quad \leq x^2 = 2329.963$.

CONT'D:



Step4 : Substitute the values :-

$$\text{Slope} = (b) = \frac{(N\Sigma xy - (\Sigma x)(\Sigma y))}{(N\Sigma x^2 - (\Sigma x^2)^2)}$$

$$= \frac{(10(509.755) - (121.34)(32.5))}{(10[2329.963 - (121.34)^2)}$$

$$= \frac{5097.55 - 3943.55}{23299.63 - 14723.3956}$$

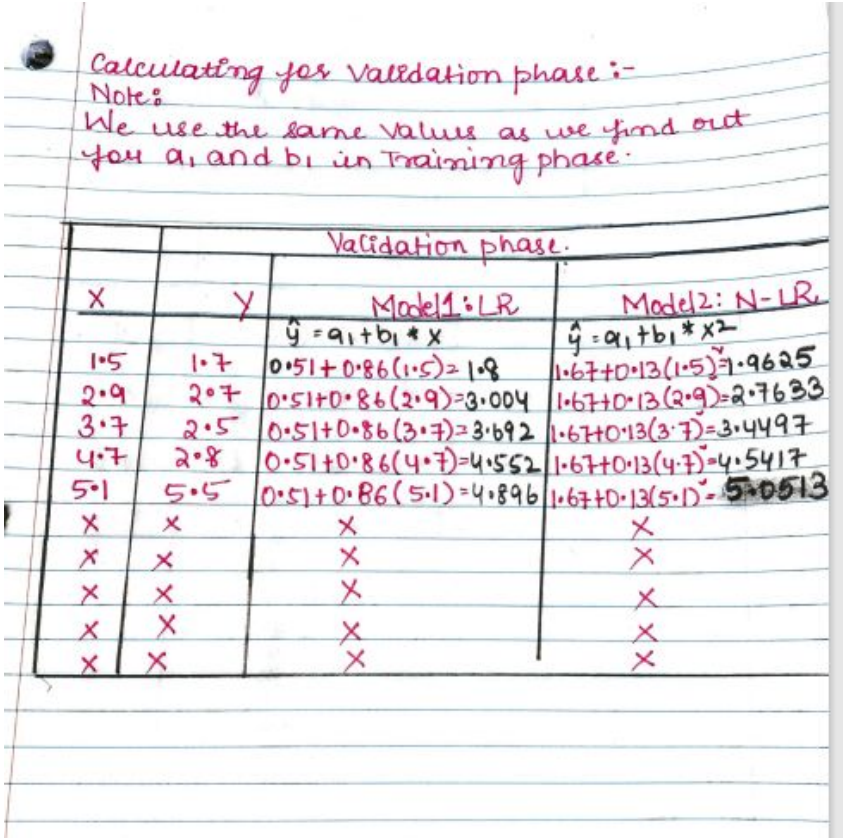
$$= \frac{1154}{8576.2344} = \boxed{0.13 = b.}$$

For : (a) Intercept :-

$$(a) = \frac{(\Sigma y - b(\Sigma x))}{N}$$

$$= \frac{32.5 - 0.13(121.34)}{10} = \frac{32.5 - 15.7742}{10}.$$

$$\boxed{a = 1.672}$$



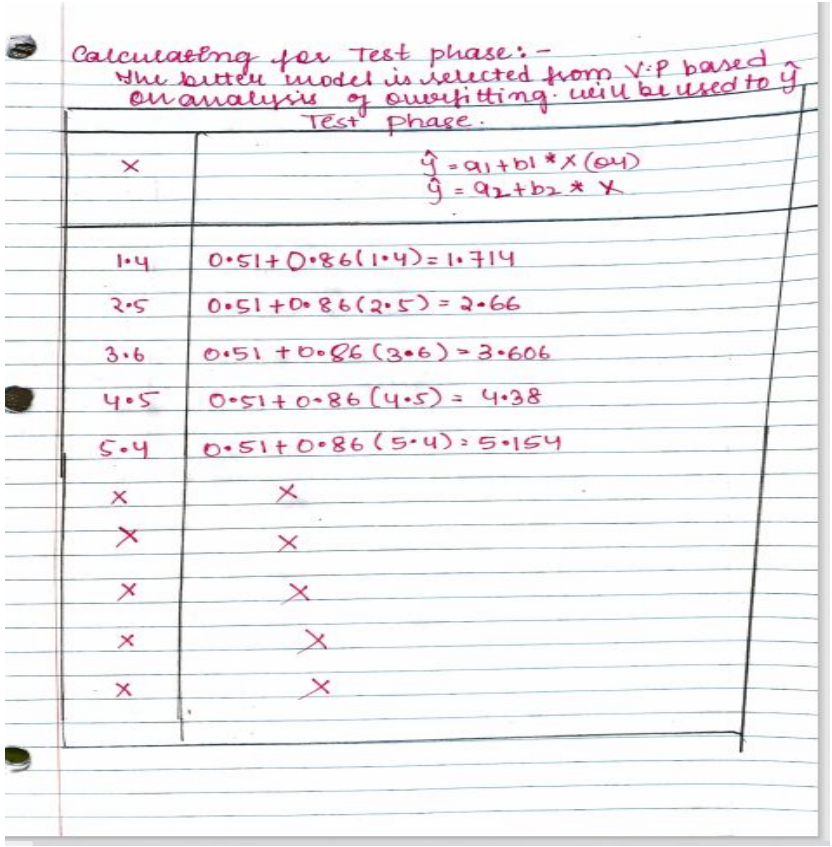
Calculating for Validation phase :-

Note:

We use the same values as we find out for a_1 and b_1 in Training phase.

		Validation phase.	
X	Y	Model 1 : LR	Model 2 : N-LR
		$\hat{y} = a_1 + b_1 * x$	$\hat{y} = a_1 + b_1 * x^2$
1.5	1.7	$0.51 + 0.86(1.5) = 1.8$	$1.67 + 0.13(1.5)^2 = 1.9625$
2.9	2.7	$0.51 + 0.86(2.9) = 3.004$	$1.67 + 0.13(2.9)^2 = 2.7633$
3.7	2.5	$0.51 + 0.86(3.7) = 3.692$	$1.67 + 0.13(3.7)^2 = 3.4497$
4.7	2.8	$0.51 + 0.86(4.7) = 4.552$	$1.67 + 0.13(4.7)^2 = 4.5417$
5.1	5.5	$0.51 + 0.86(5.1) = 4.896$	$1.67 + 0.13(5.1)^2 = 5.0513$
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X
X	X	X	X

CONT'D:



Calculating for Test phase:-
The better model is selected from V:P based on analysis of overfitting. will be used to $\hat{y}$

Test phase.

x	$\hat{y} = a_1 + b_1 * x$ (04) $\hat{y} = a_2 + b_2 * x$
1.4	0.51 + 0.86(1.4) = 1.714
2.5	0.51 + 0.86(2.5) = 2.66
3.6	0.51 + 0.86(3.6) = 3.606
4.5	0.51 + 0.86(4.5) = 4.38
5.4	0.51 + 0.86(5.4) = 5.154
x	x
x	x
x	x
x	x
x	x

WHICH MODEL IS BETTER?

For Better model to choose :-

⇒ How to choose which model is better?

The Mean Squared Error (MSE) is a measure of how close a fitted line is to data points.

- The smaller the MSE, the closer the fit is to the data.

Training set :-

Model 1

$$MSE = \frac{[(1.37-1.8)^2 + (2.74-2.4)^2 + (3.348-2.3)^2 + (4.208-3.8)^2 + (5.068-5.3)^2 + (1.714-1.5)^2 + (2.66-2.2)^2 + (2.9183-3.8)^2 + (4.036-4.0)^2 + (4.896-5.4)^2]}{10}$$

$$= 0.290972$$

Model 2 =

$$MSE = \frac{[(1.8-1.8)^2 + (2.19-2.4)^2 + (3.0857-2.3)^2 + (4.0737-3.8)^2 + (5.3217-5.3)^2 + (1.9248-1.5)^2 + (2.4825-2.2)^2 + (2.6892-3.8)^2 + (3.8553-4.0)^2 + (5.0513-5.4)^2]}{10}$$

$$= 0.237344$$

Validation phase :-

Model 1

$$MSE = \frac{[(1.883-1.7)^2 + (3.004-2.7)^2 + (3.692-2.5)^2 + (4.552-2.8)^2 + (4.896-5.5)^2]}{5}$$

$$= 0.99152$$

Model 2

$$MSE = \frac{[(1.9625-1.7)^2 + (2.7633-2.7)^2 + (3.4497-2.5)^2 + (4.5417-2.8)^2 + (5.0513-5.5)^2]}{5}$$

$$= 0.841939$$

Calculating for Model 1

$$MSE = \frac{2.90972}{0.99152} = 0.293461$$

Calculating for Model 2

$$MSE = \frac{0.237344}{0.841939} = 0.281902$$

Conclusion : Model 1 is better.

KEY TAKEAWAYS:

- Both linear and nonlinear regression predict Y responses from an X variable (or variables).
- Nonlinear regression is a curved function of an X variable (or variables) that is used to predict a Y variable
- Nonlinear regression can show a prediction of population growth over time.

BIBLIOGRAPHY:

https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/linear_regression_example.html

https://ml-cheatsheet.readthedocs.io/en/latest/linear_regression.html#:~:text=Linear%20Regression%20is%20a%20supervised,(e.g.%20cat%2C%20dog).

https://npu85.npu.edu/~henry/npu/classes/data_science/algorithm/slide/overfit.html

https://www.mathworks.com/discovery/nonlinear-regression.html