Suppose we collect a set of sample data and distribute the sample data by

```
Training phase: 50%
Validation phase: 25%
Test phase: 25%
```

Training Phase			Validation Phase			Test Phase	
Real Data Set 1 50% of the collcted data	**Model 1: Linear Regression**	**Model 2: Non-Linear Regression**	**Real Data Set 2 25% of the collcted data**	**Model 1: Linear Regression**	**Model 2: Non-Linear Regression**	**Real Data Set 3 25% of the collcted data**	**The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ**

- **After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase.**
- **Only ŷ values are changed with the new Real Data Sets.**

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8			1.5	1.7			1.4	
2	2.4			2.9	2.7			2.5	
3.3	2.3			3.7	2.5			3.6	
4.3	3.8			4.7	2.8			4.5	
5.3	5.3			5.1	5.5			5.4	
1.4	1.5			X	X	X	X	X	X
2.5	2.2			X	X	X	X	X	X
2.8	3.8			X	X	X	X	X	X
4.1	4.0			X	X	X	X	X	X
5.1	5.4			X	X	X	X	X	X

Note:

- Real Data Set 1 can be used to determine the formulas for Model 1: Linear Regression and Model 1: Linear Regression. That is, to determine the valuese of a1, b1, a2, and b2 in the following formulas:
  -
  - ŷ=a1 + b1 * x
  - ŷ=a2 + b2 * x^2

    - After the formulas are determined, you can use the formulas to calculate the ŷ values in the following phases:
        - Training Phase
        - Validation Phase
        - Test Phase
    - Note: The values of "x" in "$\hat{y}=a1 + b1 * x$" and "$\hat{y}=a2 + b2 * x^2$" are the same as the "x" list on the "Real Data Set".
- Optional: You may want to implement the following 3 programs:
    - Program 1: To implement Linear Regression Model 1
      Note:
        - This program is to use RealData Set 1 to determine a1 and b1 based on Model 1.
        - The program can be used to fill part of the blank spaces in above table.
    - Program 2: Non-Linear Regression Model 2
      Note:
        - This program is to use RealData Set 1 to determine a2 and b2 based on Model 2.
        - The program can be used to fill part of the blank spaces in above table.
    - Program 3: Calculate MSE
- Adding the project to your portofolio
    a. Please use Google Slides to document the project
    b. Please link your presentation on GitHub using this structure

Answer:

Training phase

Linear Regression

N=10

Find x*y , x * x

x	y	x * y	x * x
1.00	1.80	1.80	1.80
2.00	2.40	4.80	4.80
3.30	2.30	7.59	7.59
4.30	3.80	16.34	16.34
5.30	5.30	28.09	28.09
1.40	1.50	2.10	2.10
2.50	2.20	5.50	5.50
2.80	3.80	10.64	10.64
4.10	4.00	16.40	16.40
5.10	5.40	27.54	27.54

Find ΣX, ΣY, ΣXY, ΣXX

ΣX	31.80
ΣY	32.50
ΣXY	120.80
ΣXX	121.34

Use slope formula

$Slope(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

b1 = (10 * 120.80 – 31.80 * 32.5) / (10 * 121.34^2 – 31.80^2)

= 174.5 / 202.16

= 0.86

Use intercept formula

$Intercept(a) = (\Sigma Y - b(\Sigma X)) / N$

a1 = (32.50 – 0.86 * 31.80) / 10

= 5.15 /10

= 0.52

$Regression\ Equation(y) = a + bx$

y = 0.52 + 0.86x

Non-Linear Regression

create X from X

A	B	C
x	x	y
1.00	1.00	1.80
2.00	4.00	2.40
3.30	10.89	2.30
4.30	18.49	3.80
5.30	28.09	5.30
1.40	1.96	1.50
2.50	6.25	2.20
2.80	7.84	3.80
4.10	16.81	4.00
5.10	26.01	5.40

N = 10

Find X * Y, X * X

x	y	x * y	x * x
1.00	1.80	1.80	1.00
4.00	2.40	9.60	16.00
10.89	2.30	25.05	118.59
18.49	3.80	70.26	341.88
28.09	5.30	148.88	789.05
1.96	1.50	2.94	3.84
6.25	2.20	13.75	39.06
7.84	3.80	29.79	61.47
16.81	4.00	67.24	282.58
26.01	5.40	140.45	676.52

Find ΣX, ΣY, ΣXY, ΣX^2.

ΣX		121.34
ΣY		32.50
ΣXY		509.76
ΣXX		2329.99

slope formula

$Slope(b) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X^2 - (\Sigma X)^2)$

b2 = (10 * 509.76 – 121.34 * 32.50) / (10 * 2329.99 – 121.34^2)

= 1154.05 / 8576.50

= 0.13

intercept formula

$Intercept(a) = (\Sigma Y - b(\Sigma X)) / N$

a2 = (32.50 – 0.13 * 121.34) / 10

= 16.73 / 10

= 1.67

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

y = 1.67 + 0.13 x^2

1	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2
2	1.00	1.80	1.38	1.80
3	2.00	2.40	2.24	2.19
4	3.30	2.30	3.36	3.09
5	4.30	3.80	4.22	4.07
6	5.30	5.30	5.08	5.32
7	1.40	1.50	1.72	1.92
8	2.50	2.20	2.67	2.48
9	2.80	3.80	2.93	2.69
10	4.10	4.00	4.05	3.86
11	5.10	5.40	4.91	5.05

- **After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase.**

- **Only ŷ values are changed with the new Real Data Sets.**

Validation phase

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x2
1.50	1.70	1.81	1.96
2.90	2.70	3.01	2.76
3.70	2.50	3.70	3.45
4.70	2.80	4.56	4.54
5.10	5.50	4.91	5.05

Calculate the MSE

Training:

Model 1:

((1.38-1.80)^2+(2.24-2.40)^2+(3.36-2.30)^2+(4.22-3.80)^2+(5.08-5.30)^2+(1.72-1.50)^2+(2.67-2.20)^2+(2.93-3.80)^2+(4.05-4.00)^2+(4.91-5.40))/10 = 0.21

Model 2:

((1.80-1.80)^2+(2.19-2.40)^2+(3.09-2.30)^2+(4.07-3.80)^2+(5.32-5.30)^2+(1.92-1.50)^2+(2.48-2.20)^2+(2.69-3.80)^2+(3.86-4.00)^2+(5.05-5.40))/10 = 0.19

Validation:

Model 1:

((1.70-1.81)^2+(2.70-3.01)^2+(2.50-3.70)^2+(2.80-4.56)^2+(5.50-4.91)^2)/5 = 1.00

Model 2:

((1.70-1.96)^2+(2.70-2.76)^2+(2.50-3.45)^2+(2.80-4.54)^2+(5.50-5.05)^2)/5 = 0.84

MSE = max(Training_Set_MSE, Validation_Set_MSE) / min(Training_Set_MSE, Validation_Set_MSE)

Model 1: 1.00/0.21 = 4.76

Model 2: 0.84/0.19 = 4.42

Model 2 is smaller, is better.

Test phase

Use Model 2

x	$\hat{y}$=a2 + b2 * x2
1.40	1.92
2.50	2.48
3.60	3.35
4.50	4.30
5.40	5.46

Finally

Training Phase			Validation Phase			Test Phase	
Real Data Set 1 50% of the collcted data	Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 2 25% of the collcted data	Model 1: Linear Regression	Model 2: Non-Linear Regression	Real Data Set 3 25% of the collcted data	The better model (Model 1 or Model 2) selected from the Validation Phase based on the analysis of overfitting will be used to calculate ŷ

- After calculating a1, b1, a2, b2 in Training Phase, the values are not changed with the new Real Data Sets in Validation Phase and Test Phase.
- Only ŷ values are changed with the new Real Data Sets.

x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	y	ŷ=a1 + b1 * x	ŷ=a2 + b2 * x^2	x	ŷ=a1 + b1 * x or ŷ=a2 + b2 * x^2
1	1.8	1.38	1.80	1.5	1.7	1.81	1.96	1.4	1.92
2	2.4	2.24	2.19	2.9	2.7	3.01	2.76	2.5	2.48
3.3	2.3	3.36	3.09	3.7	2.5	3.70	3.45	3.6	3.35
4.3	3.8	4.22	4.07	4.7	2.8	4.56	4.54	4.5	4.30
5.3	5.3	5.08	5.32	5.1	5.5	4.91	5.05	5.4	5.46
1.4	1.5	1.72	1.92	X	X	X	X	X	X
2.5	2.2	2.67	2.48	X	X	X	X	X	X
2.8	3.8	2.93	2.69	X	X	X	X	X	X
4.1	4.0	4.05	3.86	X	X	X	X	X	X
5.1	5.4	4.91	5.05	X	X	X	X	X	X