8. The process of Machine Learning and using Overfitting to evaluate Linear Regression Model and Non-linear Regression .
   - Please compare the following two Regression Models to see which one has more serious overfitting issue.
     - Linear Regression Model 1
     - Non-Linear Regression Model 2
   - 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

   - 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 * x²
```

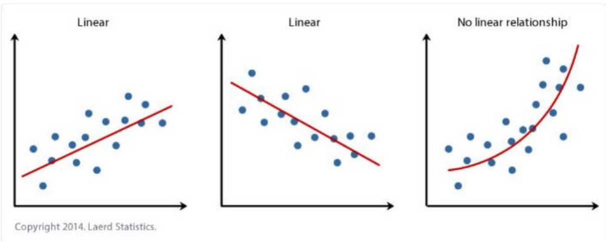
     - After the formulas are determined, you can use the formulas to calculate the ŷ values in the following phases:
       - Training Phase
       - Validation Phase
       - Test Phase
     - Note: The values of "x" in "ŷ=a1 + b1 * x" and "ŷ=a2 + b2 * x^2" are the same as the "x" list on the "Real Data 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

```
Machine Learning
  - Model Selection
    + Use Overfitting To Evaluate Different Models
```

   - Submit
     a. The URLs of the Google Slides and GitHub web pages related to this project.
     b. A PDF file of your Google Slides
- References
  - Thangarani Prabhu - TA, Summer 2018
  - Use Overfitting to evaluate different Models
  - R and Linear/Non-Linear Regression - Lena Lee, Fall 2015

Answer:

Non-linear regression has more serious overfitting issue. Linear regression is less prone to overfitting than non-linear regression because it has a simpler model structure.



Training Phase

X and Y values are from Real **Data Set 1**, 50% of the collected data.

As shown in the table below, values for X^2, Y^2, X*Y, X*X*Y, P*P and summation of each are calculated to be used in the equations.

Training Phase

	X	Y	X^2	Y^2	XY	XXY	P*P
	1	1.8	1	3.24	1.8	1.8	1
	2	2.4	4	5.76	4.8	9.6	16
	3.3	2.3	10.89	5.29	7.59	25.047	118.5921
	4.3	3.8	18.49	14.44	16.34	70.262	341.8801
	5.3	5.3	28.09	28.09	28.09	148.877	789.0481
	1.4	1.5	1.96	2.25	2.1	2.94	3.8416
	2.5	2.2	6.25	4.84	5.5	13.75	39.0625
	2.8	3.8	7.84	14.44	10.64	29.792	61.4656
	4.1	4	16.81	16	16.4	67.24	282.5761
	5.1	5.4	26.01	29.16	27.54	140.454	676.5201
Sum	31.8	32.5	121.34	123.51	120.8	509.762	2329.9862

Then we will calculate Slope(b) and Intercept(a) values for both **Model 1 (Linear Regression)** and **Model 2 (Non-Linear Regression)**.

Equations for Linear Regression

$Slope(b1) = (N\Sigma XY - (\Sigma X)(\Sigma Y)) / (N\Sigma X2 - (\Sigma X)2)$

$Intercept(a1) = (\Sigma Y - b(\Sigma X)) / N$

Equations for Non-Linear Equation

$Slope(b2) = (N\Sigma PY - (\Sigma P)(\Sigma Y)) / (N\Sigma P2 - (\Sigma P)2)$

$Intercept(a2) = (\Sigma Y - b(\Sigma P)) / N$

Where $P = X * X$

The values are as follows:

Slope(b1)	0.8631777
Intercept(a1)	0.505095

Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - $(ΣX)^2$)
Intercept(a) = (ΣY - b(ΣX)) / N

Slope(b2)	0.1345624
Intercept(a2)	1.617249

Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP^2 - $(ΣP)^2$)
Intercept(a) = (ΣY - b(ΣP)) / N
Where P = X * X

* The value of N is 10 for Training Phase.

By using the values calculated above, we can calculate ŷ values for both **Model 1** and **Model 2**.

ŷ=a1 + b1 * x	**ŷ=a2 + b2 * x^2**
1.36826	1.751809
2.23143	2.155489
3.353551	3.0826074
4.216721	4.1052634
5.079891	5.3970394
1.713528	1.8809866
2.663015	2.458249
2.921966	2.6721994
4.044087	3.8792026
4.907257	5.1171546

Validation Phase

X and Y values are from **Real Data Set 2**, 25% of the collected data.

As shown in the table below, values for X^2, Y^2, X*Y, X*X*Y, P*P and summation of each are calculated to be used in the equations.

		Validation Phase					
	X	Y	X^2	Y^2	XY	XXY	P*P
	1.5	1.7	2.25	2.89	2.55	3.825	5.0625
	2.9	2.7	8.41	7.29	7.83	22.707	70.7281
	3.7	2.5	13.69	6.25	9.25	34.225	187.416
	4.7	2.8	22.09	7.84	13.16	61.852	487.968
	5.1	5.5	26.01	30.25	28.05	143.055	676.52
			0	0	0	0	0
			0	0	0	0	0
			0	0	0	0	0
			0	0	0	0	0
			0	0	0	0	0
Sum	17.9	15.2	72.45	54.52	60.84	265.664	1427.69

Then we will calculate Slope(b) and Intercept(a) values for both **Model 1 (Linear Regression)** and **Model 2 (Non-Linear Regression)**.

Equations for Linear Regression

Slope(b1) = (NΣXY - (ΣX)(ΣY)) / (NΣX2 - (ΣX)2)

Intercept(a1) = (ΣY - b(ΣX)) / N

Equations for Non-Linear Equation

Slope(b2) = (NΣPY - (ΣP)(ΣY)) / (NΣP2 - (ΣP)2)

Intercept(a2) = (ΣY - b(ΣP)) / N

Where P = X * X

The values are as follows:

Slope(b1)	0.83229
Intercept(a1)	0.0302

Slope(b) = (NΣXY - (ΣX)(ΣY)) / (NΣX^2 - (ΣX)^2)
Intercept(a) = (ΣY - b(ΣX)) / N

Slope(b2)	0.17229
Intercept(a2)	0.54511

Slope(b) = (NΣPY - (ΣP)(ΣY)) / (NΣP^2 - (ΣP)^2)
Intercept(a) = (ΣY - b(ΣP)) / N
Where P = X * X

* The value of N is 5 for Validation Phase.

By using the values calculated above, we can calculate ŷ values for both **Model 1** and **Model 2**.

ŷ=a1 + b1 * x	**ŷ=a2 + b2 * x^2**
1.799845	1.920009
3.008283	2.7488986
3.698819	3.4593754
4.561989	4.5896794
4.907257	5.1171546

Test Phase

X values are from Real **Data Set 3**, 25% of the collected data.

We will calculate mean squared error (MSE) for both **Model 1** and **Model 2** by using the following equation.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

y - ŷ Model 1	y - ŷ Model 2
0.186399428	0.002322372
0.028415845	0.059785629
1.10996971	0.612474343
0.173656392	0.093185743
0.048447972	0.009416645
0.045594207	0.145150789
0.21438289	0.066692546
0.770943705	1.271934193
0.001943664	0.014592012
0.242795664	0.08000152
2.822549476	2.355555794
0.282254948	0.235555579

Training Set

y - ŷ Model 1	y - ŷ Model 2
0.009969024	0.04840396
0.095038408	0.002391073
1.437166995	0.920401158
3.104605236	3.202952355
0.351344264	0.1465706
4.998123927	4.320719146
0.999624785	0.864143829

Validation Set

Then we will decide which Model to choose by using the following equation.

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

Model 1

Training_Set_MSE = 0.28225

Validation_Set_MSE = 0.99962

0.99962/0.28225 = **3.54**

Model 2

Training_Set_MSE = 0.23555

Validation_Set_MSE = 0.86414

0.86414/0.23555 = 3.668

According to the calculation, **Model 1** is the better model. So, we will use the values of b and a from **Model 1**.

Slope(b1) = 0.8631777

Intercept(a1) = 0.505095

Test Phase	
	Model 1
X	ŷ=a1 + b1 * x
1.4	1.71354378
2.5	2.66303925
3.6	3.61253472
4.5	4.38939465
5.4	5.16625458

Final Answer:

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
1	**1.8**	1.36826	1.751809	**1.5**	**1.7**	1.799845	1.920009	**1.4**	1.71354378
2	**2.4**	2.23143	2.155489	**2.9**	**2.7**	3.008283	2.7488986	**2.5**	2.66303925
3.3	**2.3**	3.353551	3.0826074	**3.7**	**2.5**	3.698819	3.4593754	**3.6**	3.61253472
4.3	**3.8**	4.216721	4.1052634	**4.7**	**2.8**	4.561989	4.5896794	**4.5**	4.38939465
5.3	**5.3**	5.079891	5.3970394	**5.1**	**5.5**	4.907257	5.1171546	**5.4**	5.16625458
1.4	**1.5**	1.713528	1.8809866	**X**	**X**	**X**	**X**	**X**	**X**
2.5	**2.2**	2.663015	2.458249	**X**	**X**	**X**	**X**	**X**	**X**
2.8	**3.8**	2.921966	2.6721994	**X**	**X**	**X**	**X**	**X**	**X**
4.1	**4**	4.044087	3.8792026	**X**	**X**	**X**	**X**	**X**	**X**
5.1	**5.4**	4.907257	5.1171546	**X**	**X**	**X**	**X**	**X**	**X**