REGRESSION

By:
Muhammad Shaheryar

USING LINEAR REGRESSION TO PREDICT CONTINUOUS VARIABLE

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

DEPENDENT VS INDEPENDENT VARIABLES

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

X: Independent variable

Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

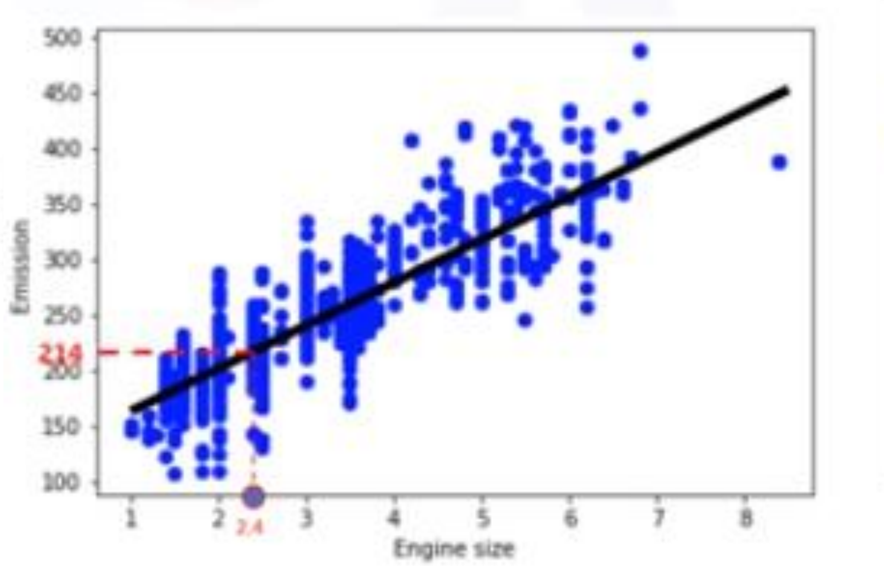
Continuous Values

REGRESSION TYPE

- Simple Linear Regression
  - Predict co2emission vs EngineSize of all cars
    - Independent variable (x): EngineSize
    - Dependent variable (y): co2emission
- Multiple Linear Regression
  - Predict co2emission vs EngineSize and Cylinders of all cars
    - Independent variable (x): EngineSize, Cylinders, etc
    - Dependent variable (y): co2emission

SIMPLE LINEAR REGRESSION

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?



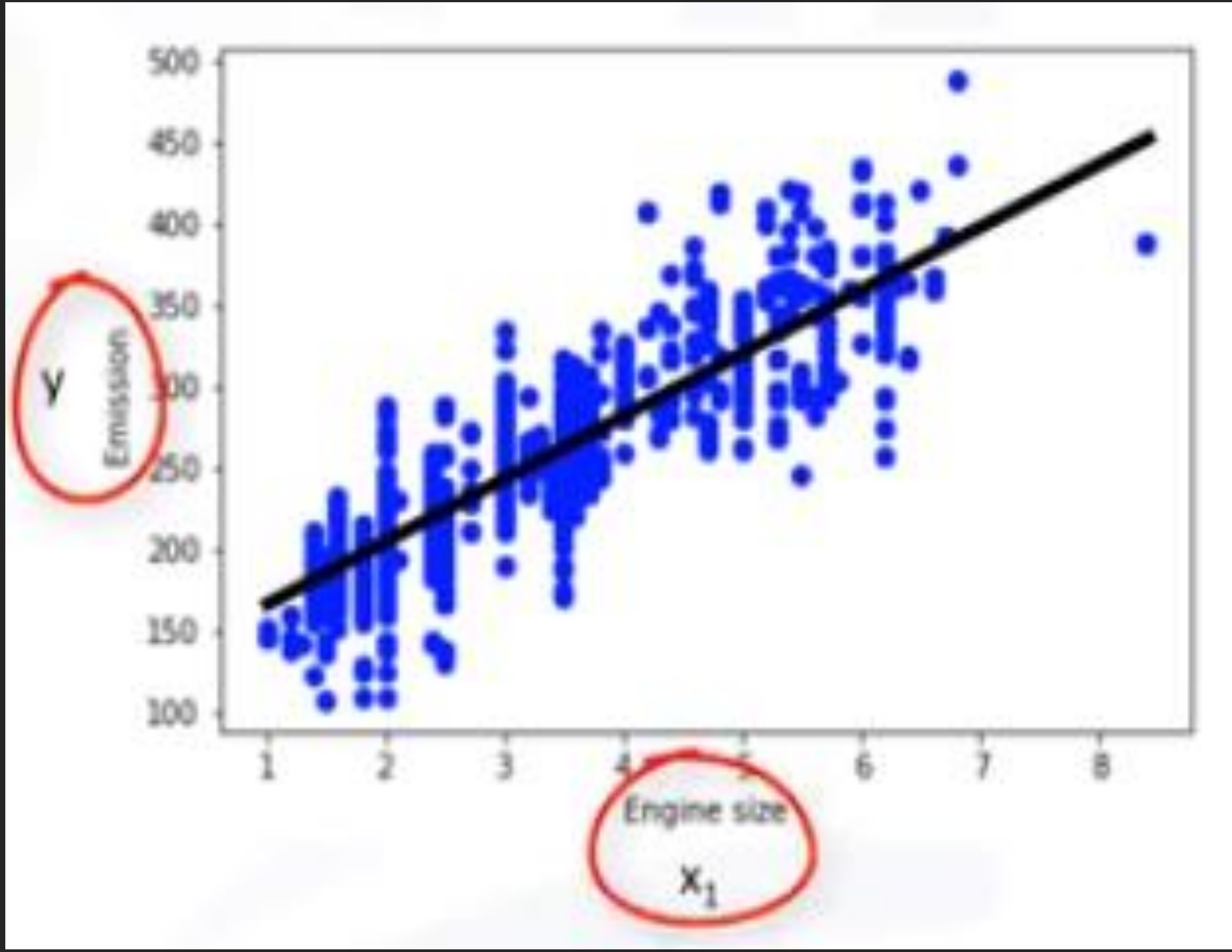
FITTING LINE

$$\hat{y} = \theta_0 + \theta_1 x_1$$

a single predictor

response variable

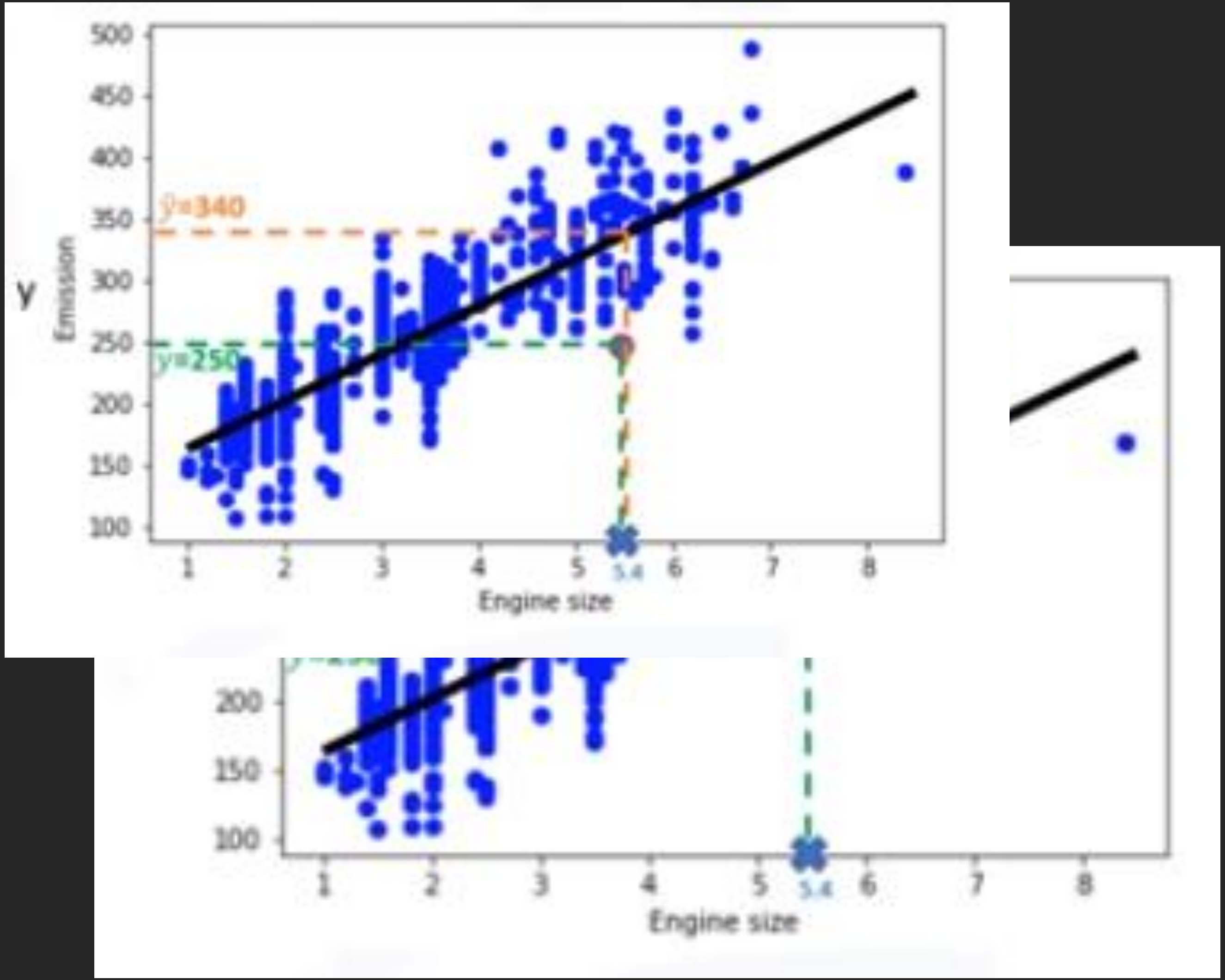
$$\hat{y} = \theta_0 + \theta_1 x_1$$



θ0 and θ1 are also called the coefficients of the linear equation. Slope and intercept respectively

HOW TO FIND BEST FIT?

$$\text{Error} = y - \hat{y}$$
$$= 250 - 340$$
$$= -90$$



This means our prediction line is not accurate.
This error is also called the residual error.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

OBJECTIVE

- The objective of linear regression is to minimize **this MSE equation, and to minimize it, we should find the best parameters, ϴ0 and ϴ1**

- Actually, we have two options here:
- Option 1 - We can use a mathematic approach.
- Option 2 - We can use an optimization approach.

ESTIMATING THE PARAMETERS

$$\hat{y} = \theta_0 + \theta_1 x_1$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{s}(x_i-\bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.34$$

$$\bar{y} = (196 + 221 + 136 + \dots)/9 = 256$$

$$\boxed{\theta_1 = \frac{(2.0-3.34)(196-256)+(2.4-3.34)(221-256)+\dots}{(2.0-3.34)^2+(2.4-3.34)^2+\dots}}$$

$$\boxed{\theta_1 = 39}$$

$$\theta_0 = 256 - 39 * 3.34$$

$$\theta_0 = 125.74$$

PREDICTING WITH LINEAR REGRESSION

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$$\hat{y} = \theta_0 + \theta_1 x_1$$

$$\boxed{Co2Emission = \theta_0 + \theta_1\ EngineSize}$$

$$Co2Emission = 125 + 39\ EngineSize$$

$$Co2Emission = 125 + 39 \times 2.4$$

$$Co2Emission = 218.6$$

REGRESSION TYPE

- Predict co2emission vs EngineSize and Cylinders of all cars
  - Independent variable (x): EngineSize, Cylinders, etc
  - Dependent variable (y): co2emission

MULTIPLE LINEAR REGRESSION

EXAMPLES OF MLR

- Independent variables effectiveness on prediction
  - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
- Predicting impacts of changes
  - How much does blood pressure go up (or down) for every unit increase (or decrease) in the BMI of a patient?

PREDICTING WITH MLR

$$Co2\ Em = \theta_0 + \theta_1 Engine\ size + \theta_2 Cylinders + \ \dots$$

$$\boxed{\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ \dots + \theta_n x_n}$$

$$\boxed{\hat{y} = \theta^T X}$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, \dots]$$

$$\boxed{X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \dots \end{bmatrix}}$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

- How to estimate θ?
    - Ordinary Least Squares
        - Linear algebra operations
            - Takes a long time for large datasets (10K+ rows)
- An optimization algorithm
    - Gradient Descent
        - Proper approach if you have a very large dataset

PREDICTING VALUES

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	13
3	3.5	6	11.1	25
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

$\theta^T = [125, 6.2, 14, ...]$

$$\boxed{\hat{y} = 125 + 6.2x_1 + 14x_2 +}$$

$$Co2Em = 125 + 6.2EngSize + 14\,Cylinders + ...$$

$$Co2Em = 125 + 6.2 \times 2.4 + 14 \times 4 + ...$$

$$Co2Em = 214.1$$

Q & A

- How to determine whether to use simple or multiple linear regression?

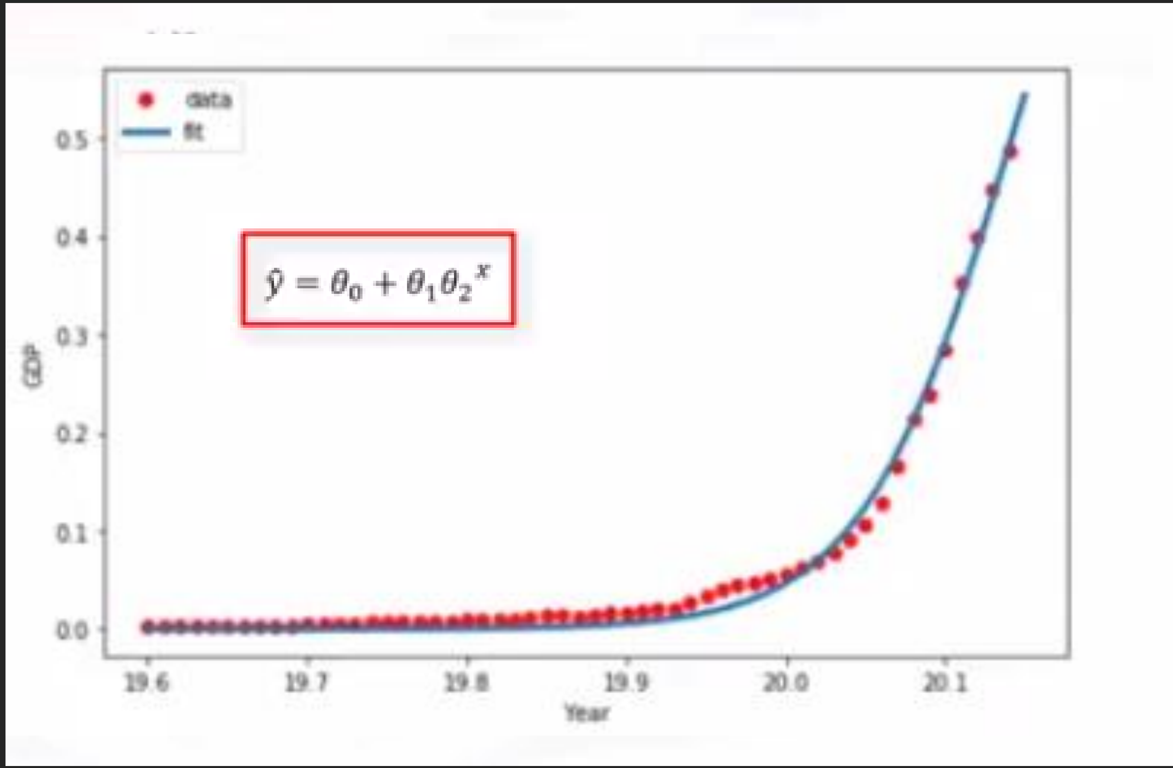
- How many independent variables should you use?

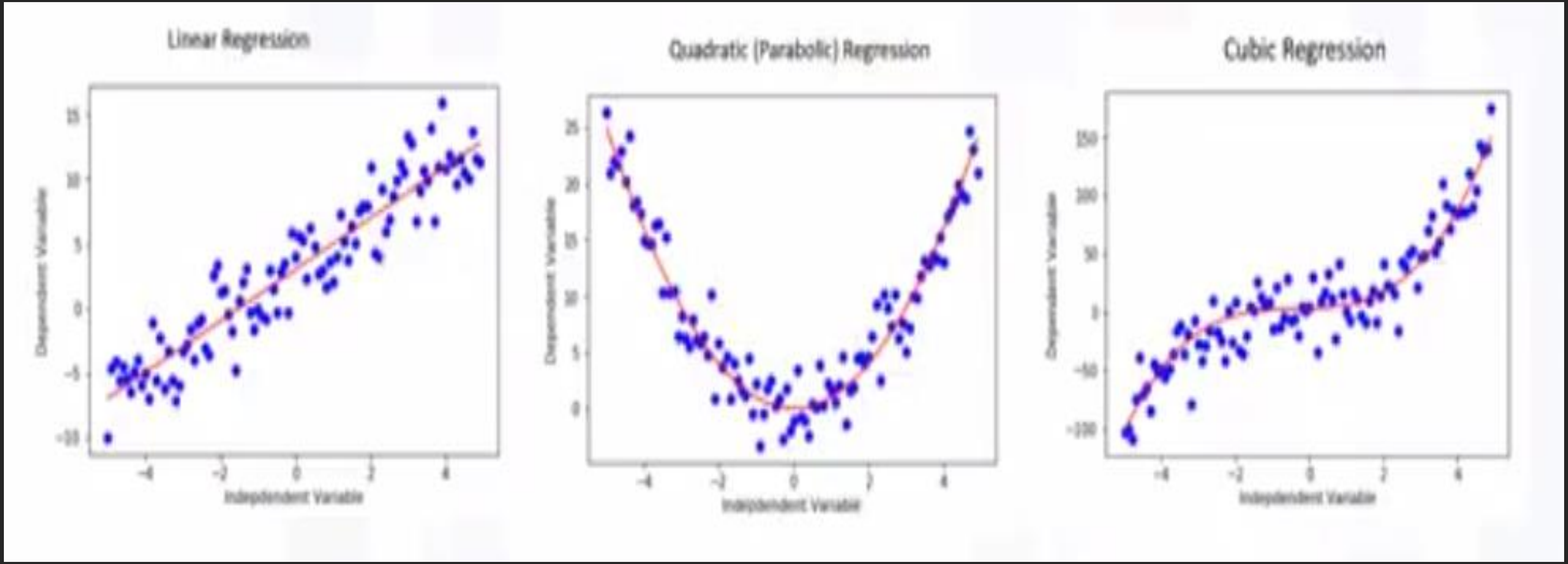
- Should the independent variable be continuous?

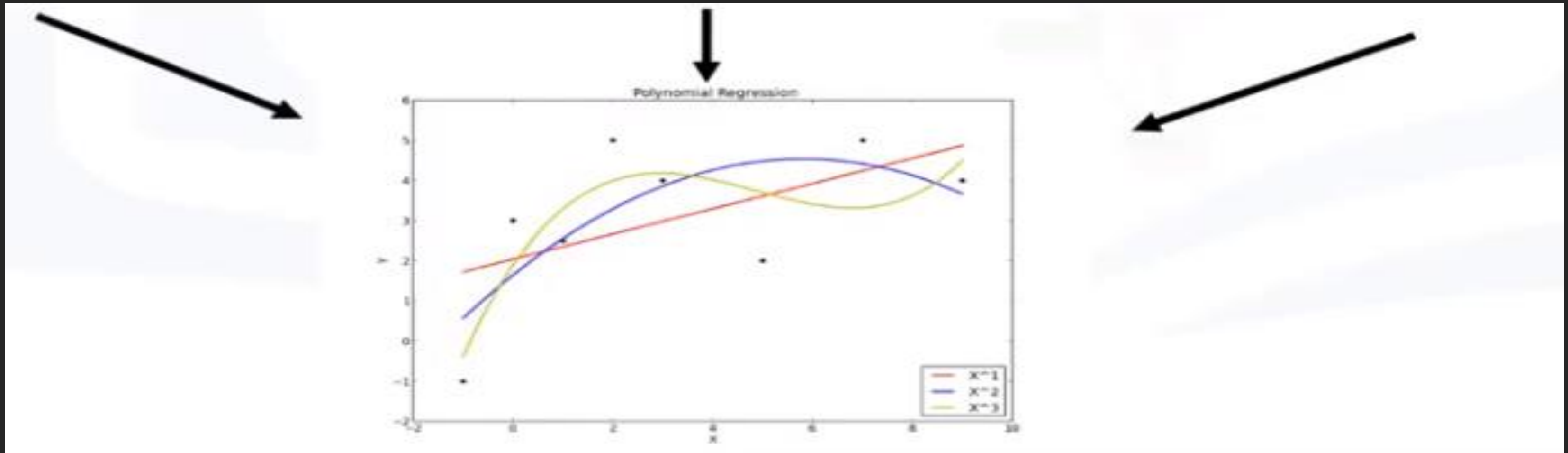
- What are the linear relationships between the dependent variable and the independent variables?

NON-LINEAR REGRESSION

	Year	Value
0	1960	5.918412e+10
1	1961	4.955705e+10
2	1962	4.668518e+10
3	1963	5.009730e+10
4	1964	5.906225e+10
5	1965	6.970915e+10
6	1966	7.587943e+10
7	1967	7.205703e+10
8	1968	6.999350e+10
9	1969	7.871882e+10
...	.....	....




Linear Regression
Quadratic (Parabolic) Regression
Cubic Regression
Dependent Variable
Indepdendent Variable
Independent Variable


Polynomial Regression
X^1
X^2
X^3

NON-LINEAR REGRESSION

- To model non-linear relationship between the dependent variable and a set of independent variables
- $\hat{y}$ must be a non-linear function of the parameters θ, not necessarily the features x

$$\hat{y} = \theta_0 + {\theta_2}^2 x$$

$$\hat{y} = \theta_0 + \theta_1 {\theta_2}^x$$

$$\hat{y} = \log(\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3)$$

$$\hat{y} = \frac{\theta_0}{1 + {\theta_1}^{(x-\theta_2)}}$$

LINEAR VS NON LINEAR REGRESSION

- How can I know if a problem is linear or non-linear in an easy way?
  - Inspect visually

  It's best to plot bivariate plots of output variables with each input variable.
  Also, you can calculate the correlation coefficient between independent and dependent variables, <u>and if for all variables it is 0.7 or higher there is a linear tendency</u>

- How should I model my data, if it displays non-linear on a scatter plot?
  - Polynomial regression
  - Non-linear regression model
  - Transform your data

POLYNOMIAL VS NON-LINEAR REGRESSION

- POLYNOMIAL REGRESSION FITS A CURVE LINE TO YOUR DATA. A SIMPLE EXAMPLE OF A POLYNOMIAL WITH A DEGEE OF 3 CAN BE SHOWN AS:

$$y - hat = b_0 + b_1 x^1 + b_2 x^2 + b_3 x^3$$

3rd degree polynomial regression

WHERE B0 IS THE INTERCEPT OR BIAS UNIT AND B1 TO B3 ARE THE SLOPES OF EACH INDEPENDENT VALUE OF VARIABLE X.

IT SURE LOOKS LIKE A FEATURE SET FOR A MULTIPLE LINEAR REGRESSION RIGHT? JUST LIKE THE ONE BELOW, YES, IT DOES. INDEED A POLYNOMIAL REGRESSION IS A SPECIAL CASE OF MULTIPLE LINEAR REGRESSION, WITH THE MAIN IDEA OF 'HOW DO YOU SELECT YOUR FEATURES?'.

$$y - hat = b_0 + b_1 x_1^1 + b_2 x_2^2 + b_3 x_3^3$$

3rd degree multiple linear regression 3 variables

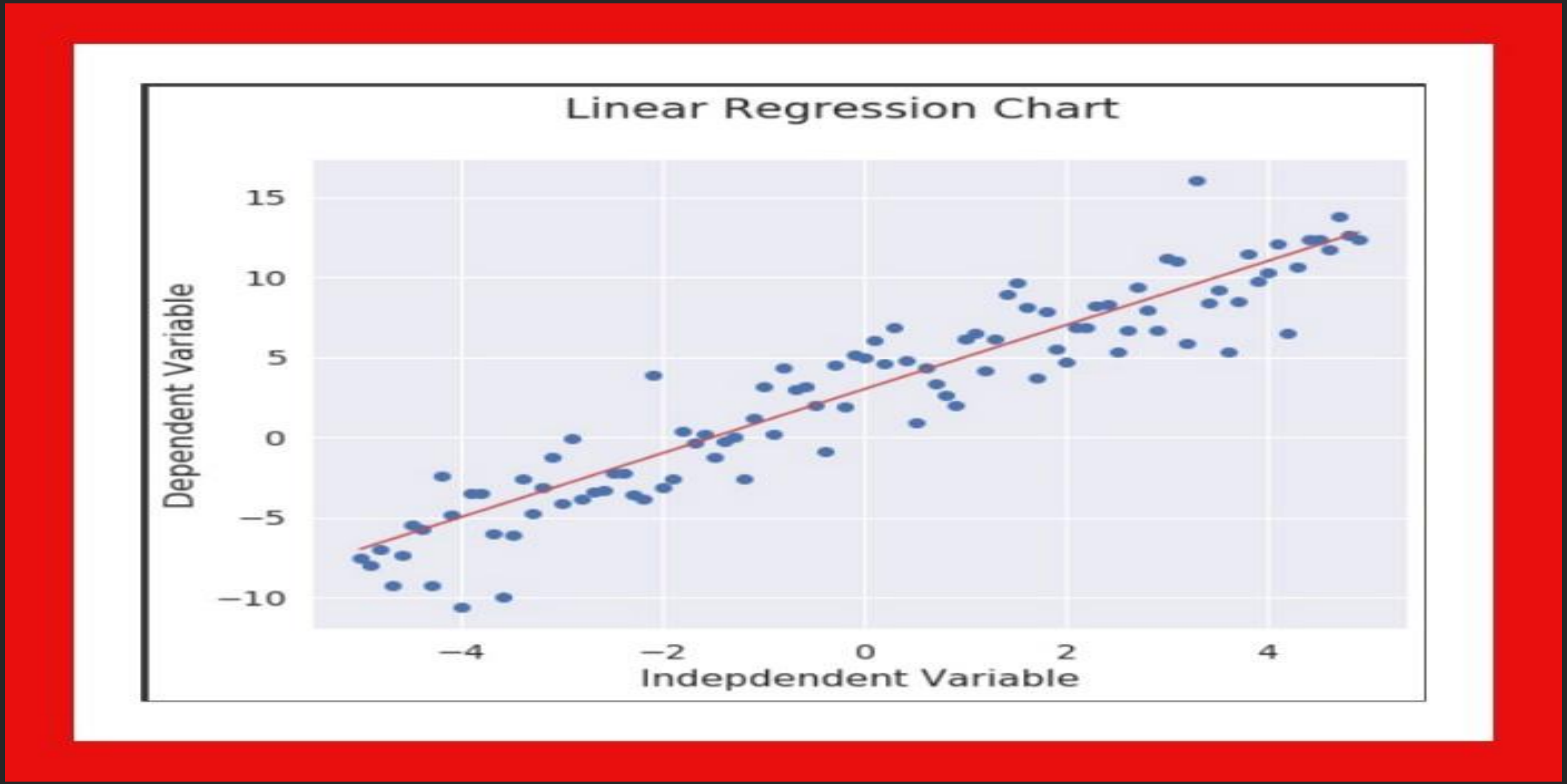
WHERE B0 IS THE INTERCEPT OR BIAS UNIT AND B1 TO B3 ARE THE SLOPES OF EACH INDEPENDENT VARIABLE X1 TO X3

COMMON TYPES OF NON-LINEAR REGRESSION

- BEFORE WE GO ON, LET'S BRIEFLY LOOK AT LINEAR REGRESSION. IT IS OF THE EQUATION:

$$Y = B0 + B1X1$$

LINEAR REGRESSION MODELS A RELATIONSHIP BETWEEN A DEPENDENT VARIABLE ***Y*** AND THE INDEPENDENT VARIABLE ***X***. THIS RELATIONSHIP HAS A DEGREE OF 1.



1. CUBIC

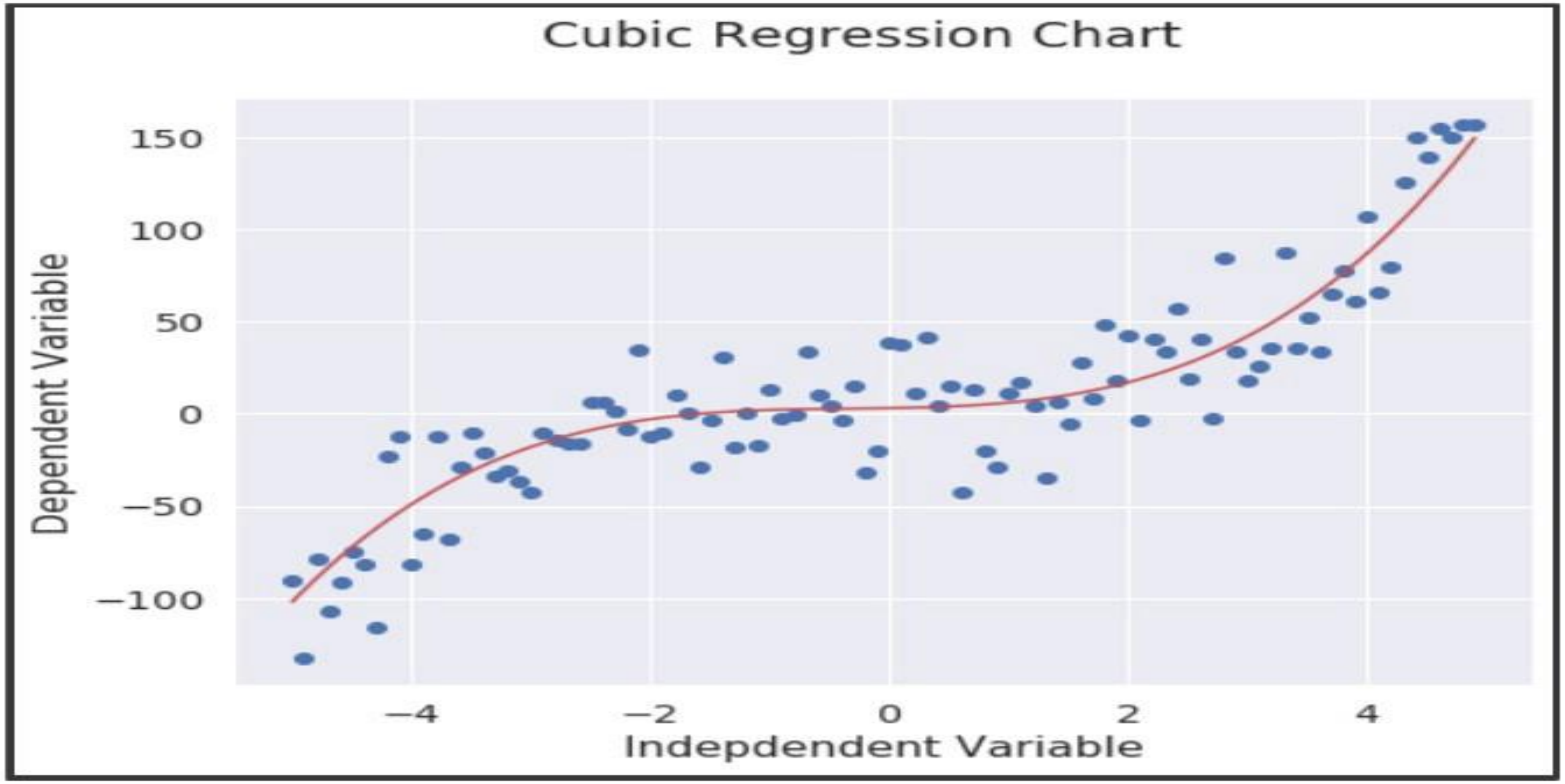
- A CUBIC FUNCTION IS OF THE FORM: *Y_HAT* IS EQUAL TO *INTERCEPT* PLUS VARIABLE *X* RAISED TO THE THIRD POWER PLUS *X* RAISED TO THE SECOND POWER AND SO ON. IT COULD ALSO BE IN REVERSE FROM 1ST POWER TO 3RD POWER
- THE GRAPH OF THIS FUNCTION IS NOT A STRAIGHT LINE OVER THE 2D PLANE.
- LET'S PLOT ONE, BUT FIRST, TAKE A LOOK AT THE CUBIC EQUATION BELOW.

$$y - hat = b_0 + 1(x^3) + 1(x^2) + 1(x^1)$$

Cubic Regression Equation

y_hat = intercept + x raised to power 3 + x raised to power 2 + x …

SAMPLE CUBIC REGRESSION CHART



2. QUADRATIC

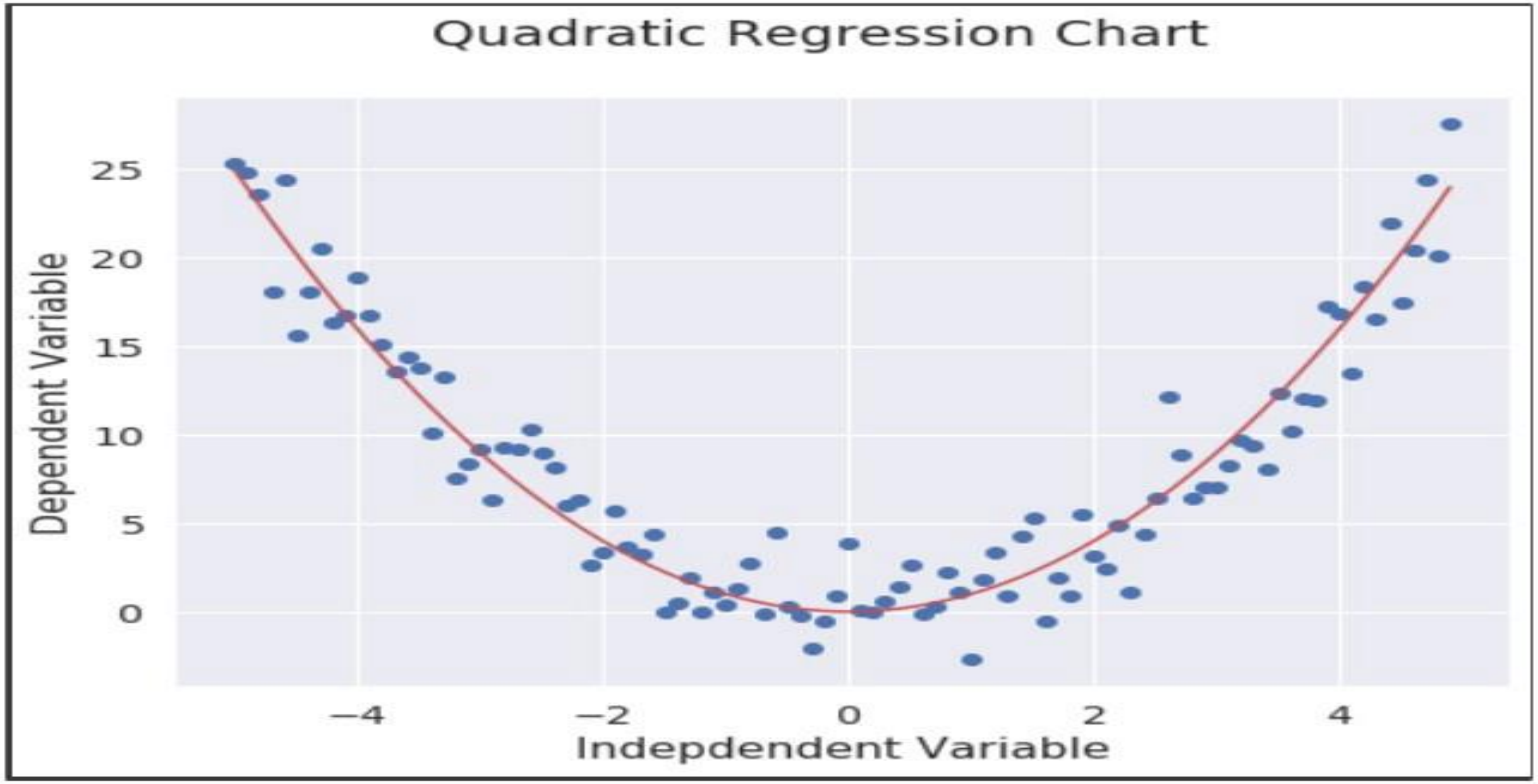
- A QUADRATIC FUNCTION IS OF THE EQUATION: *Y_HAT* IS EQUAL TO VARIABLE *X* MULTIPLIED BY VARIABLE *X* OR RAISED TO THE POWER OF 2.

$$Y - hat = X^2$$

Quadratic Regression Equation

y_hat = X squared

SAMPLE QUADRATIC REGRESSION CHART



3. EXPONENTIAL

- An exponential function with base *c* is defined as *y-hat* is equal to *intercept* plus *slope* multiplied by a constant(*c*) which is raised to the power of variable *X*. See expression below.

$$Y = a + bc^X$$

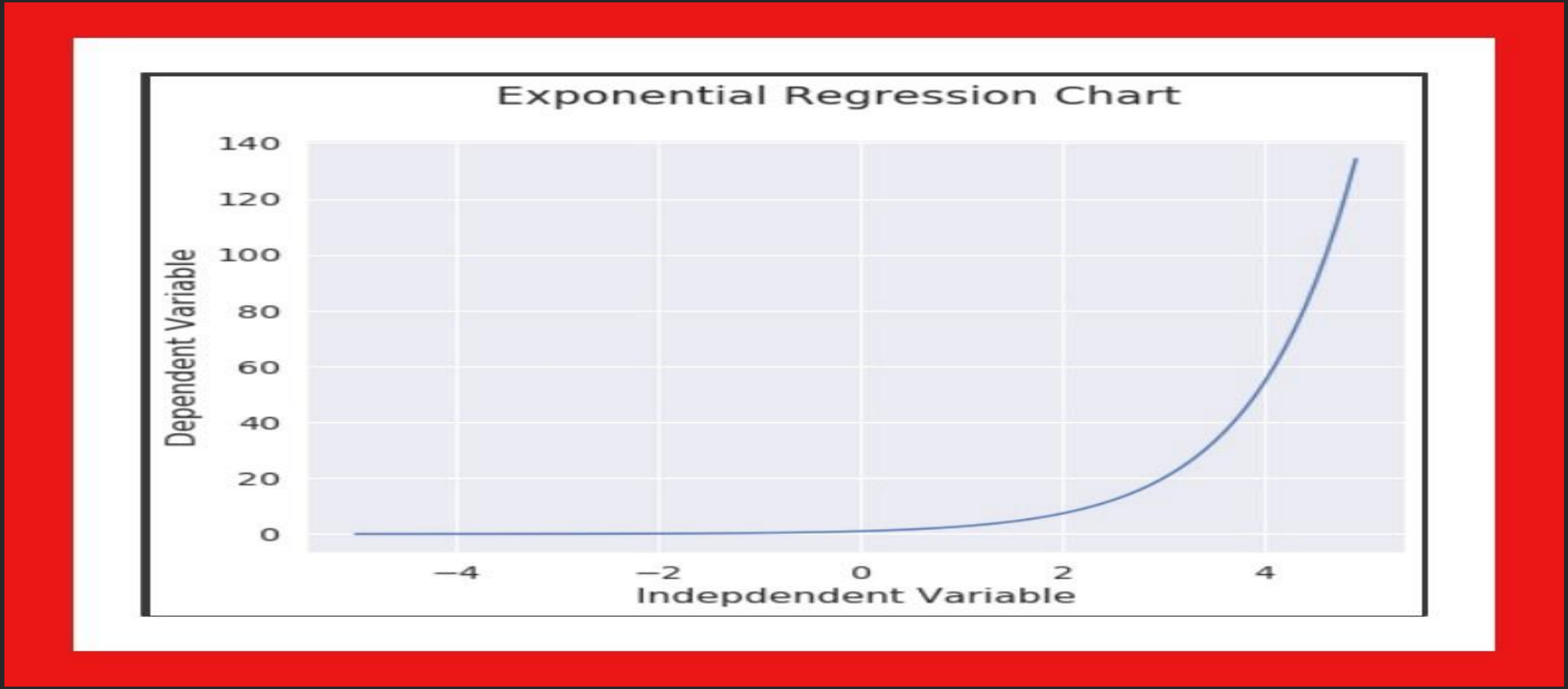
Exponential Regression Function

where b != 0, c > 0 != 1, x is a variable and a real number and c is also a constant.

Exponential might seem a bit confusing, but plotting it is pretty straight forward.

SAMPLE EXPONENTIAL REGRESSION CHART

- SIMPLY APPLY THE *NUMPY.EXP()* FUNCTION AND PASS VARIABLE *X* AS ITS ARGUMENT IN THIS FORM: *Y_HAT = NP.EXP(X).*
- THEN PLOT VARIABLE *X* ON THE X-AXIS AND VARIABLE *Y ON THE* Y-AXIS.



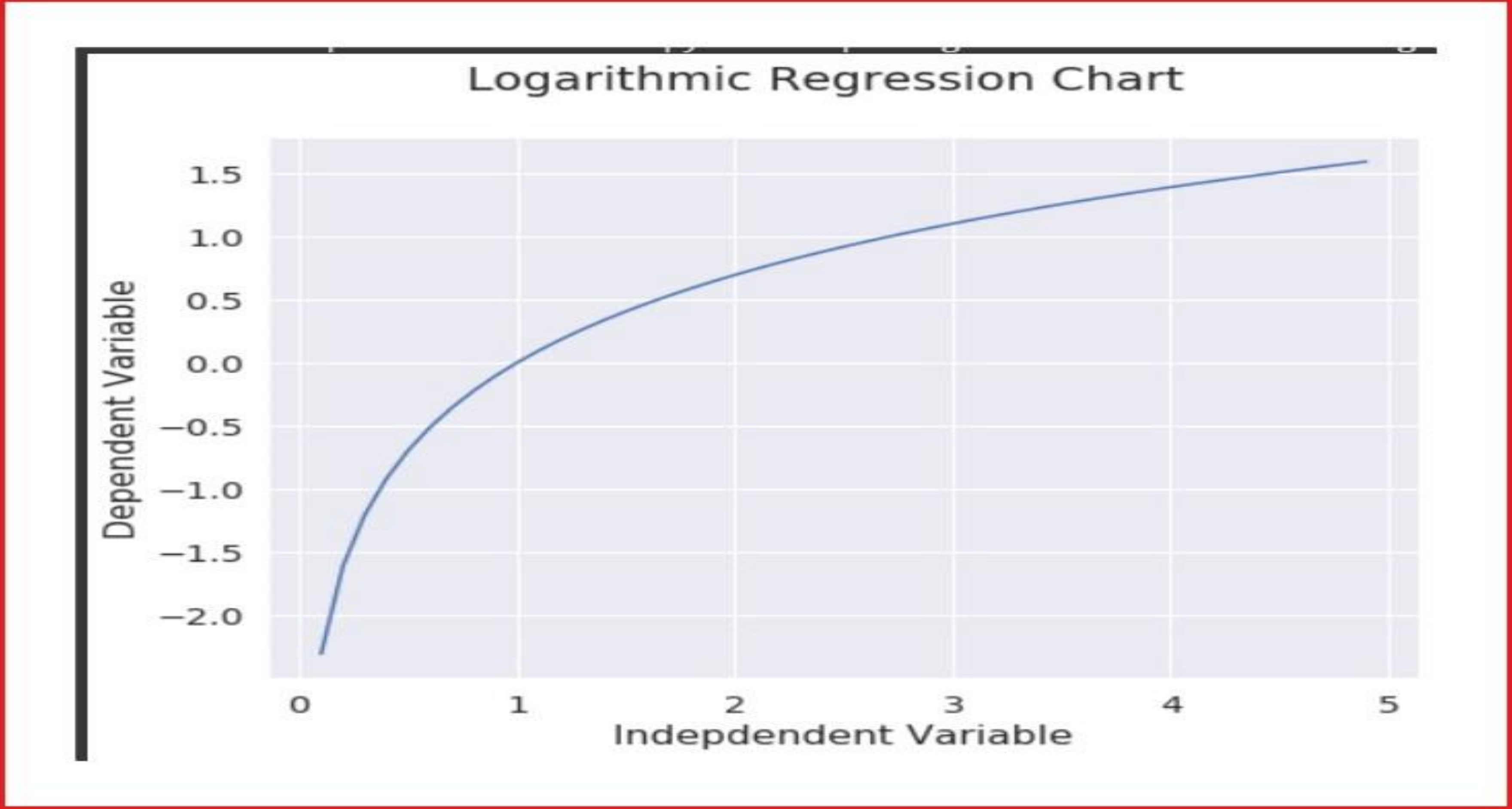
4. LOGARITHMIC

- In logarithmic function, *y_hat* is a result of applying a logarithmic map on variable ***X***.
- It is one of the simplest expressions of a logarithmic function.

$$y - hat = \log(X)$$

Logarithmic Regression Equation

SAMPLE LOGARITHMIC REGRESSION CHART



5. SIGMOIDAL / LOGISTIC

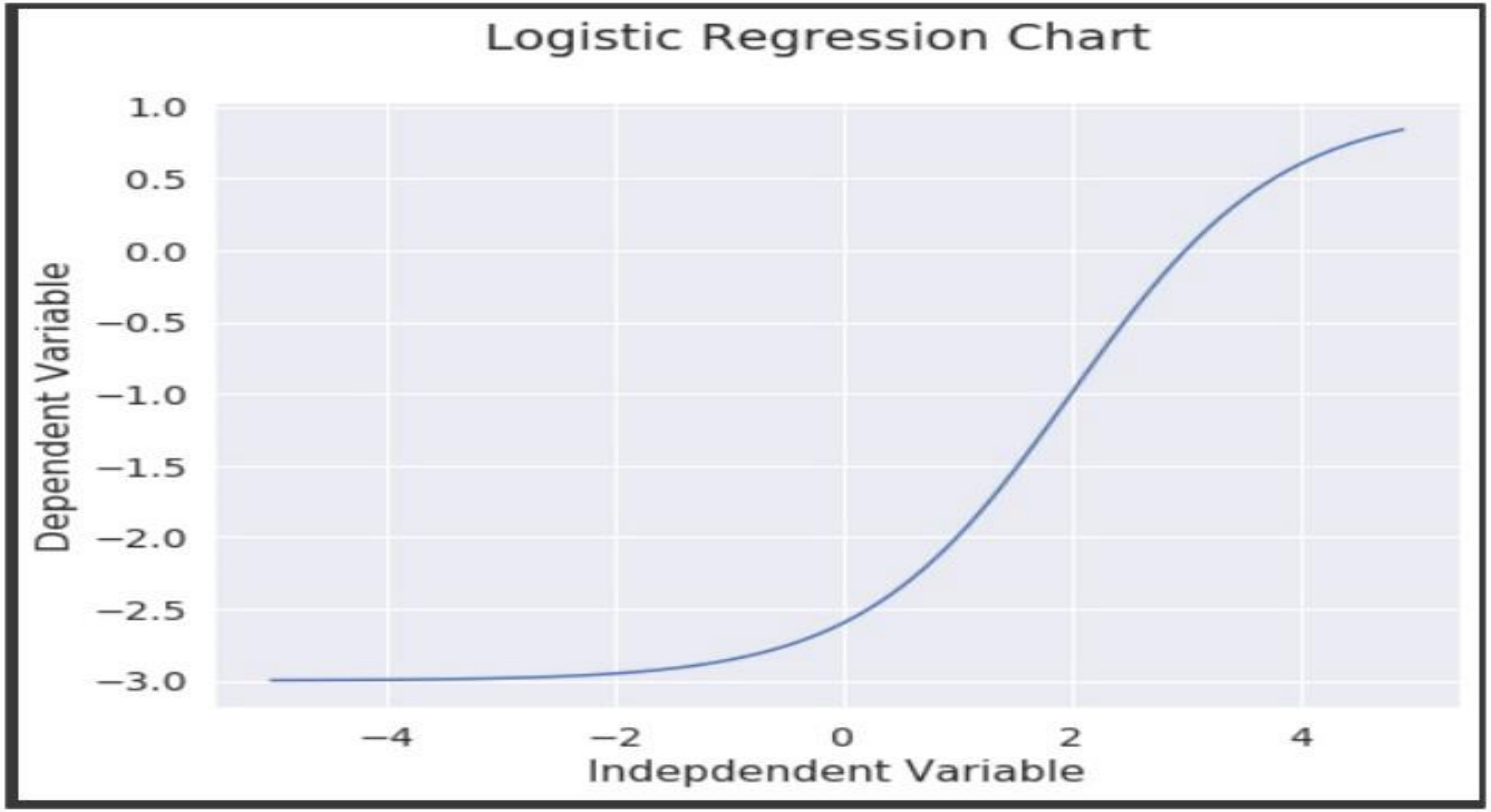
- LOGISTIC REGRESSION IS A VARIATION OF LINEAR REGRESSION, USEFUL WHEN THE OBSERVED DEPENDENT VARIABLE *Y*, IS A CATEGORICAL VARIABLE.
- IT FITS A SPECIAL S-SHAPED CURVE BY TAKING THE LINEAR REGRESSION AND TRANSFORMING THE NUMERIC ESTIMATES INTO A PROBABILITY SCORE, USING THE SIGMOID FUNCTION.

$$\hat{Y} = \frac{1}{1 + e^{\beta_1(X - \beta_2)}}$$

Logistic Regression Equation

β1 controls the curves steepness, **β2** controls the curve on the x-axis.

SAMPLE LOGISTIC REGRESSION CHART



Remember,

it is important to pick a regression model that fits the data set the best.