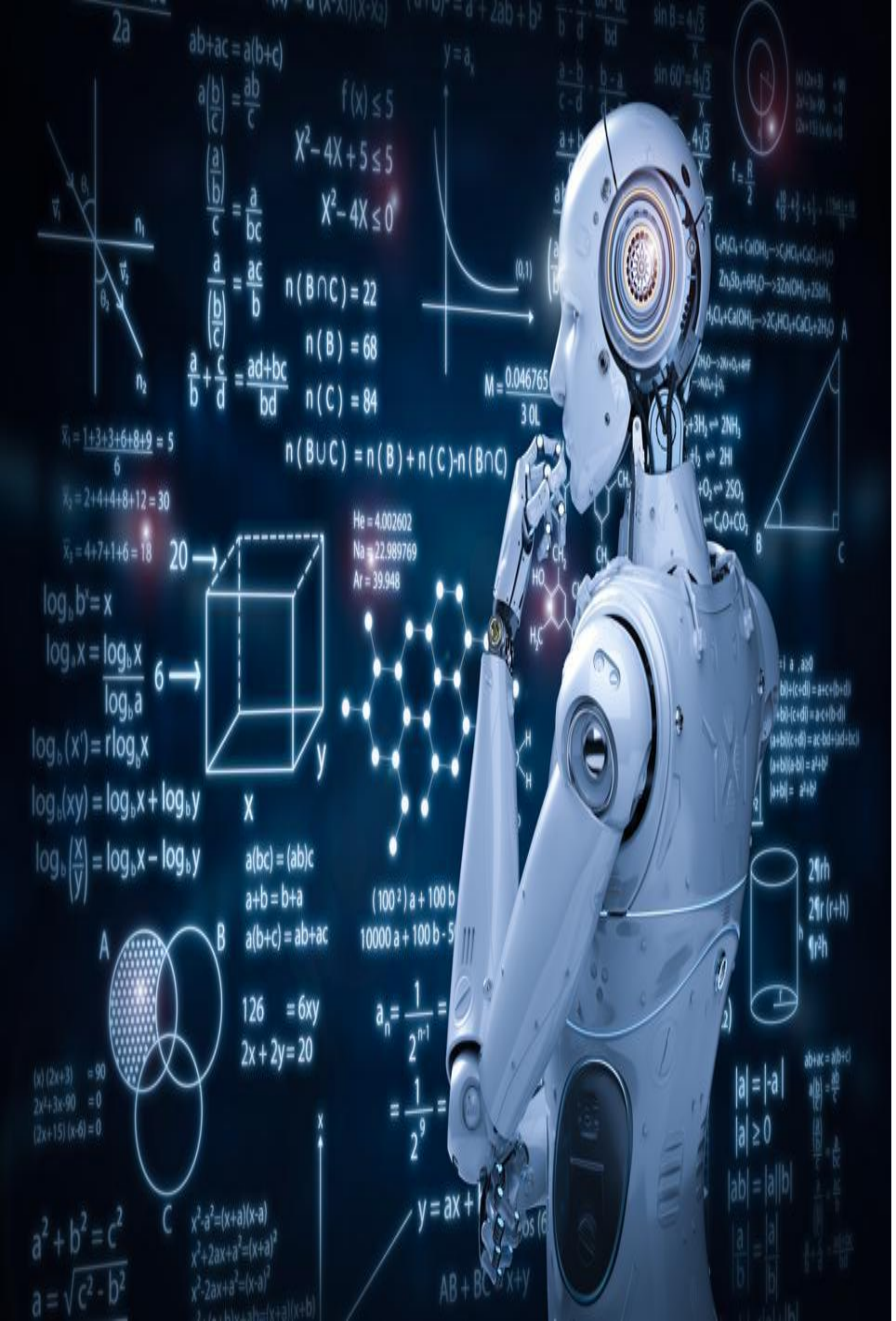


Practical Machine Learning

Day 5: SEP23 DBDA

Kiran Waghmare

Agenda

- Regression
- Types of Regression

Iris dataset

- Many exploratory data techniques are nicely illustrated with the iris dataset.
  - Dataset created by famous statistician Ronald Fisher
  - 150 samples of three species in genus *Iris* (50 each)
    - *Iris setosa*
    - *Iris versicolor*
    - *Iris virginica*
  - Four attributes
    - sepal width
    - sepal length
    - petal width
    - petal length
  - Species is class label

Iris virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

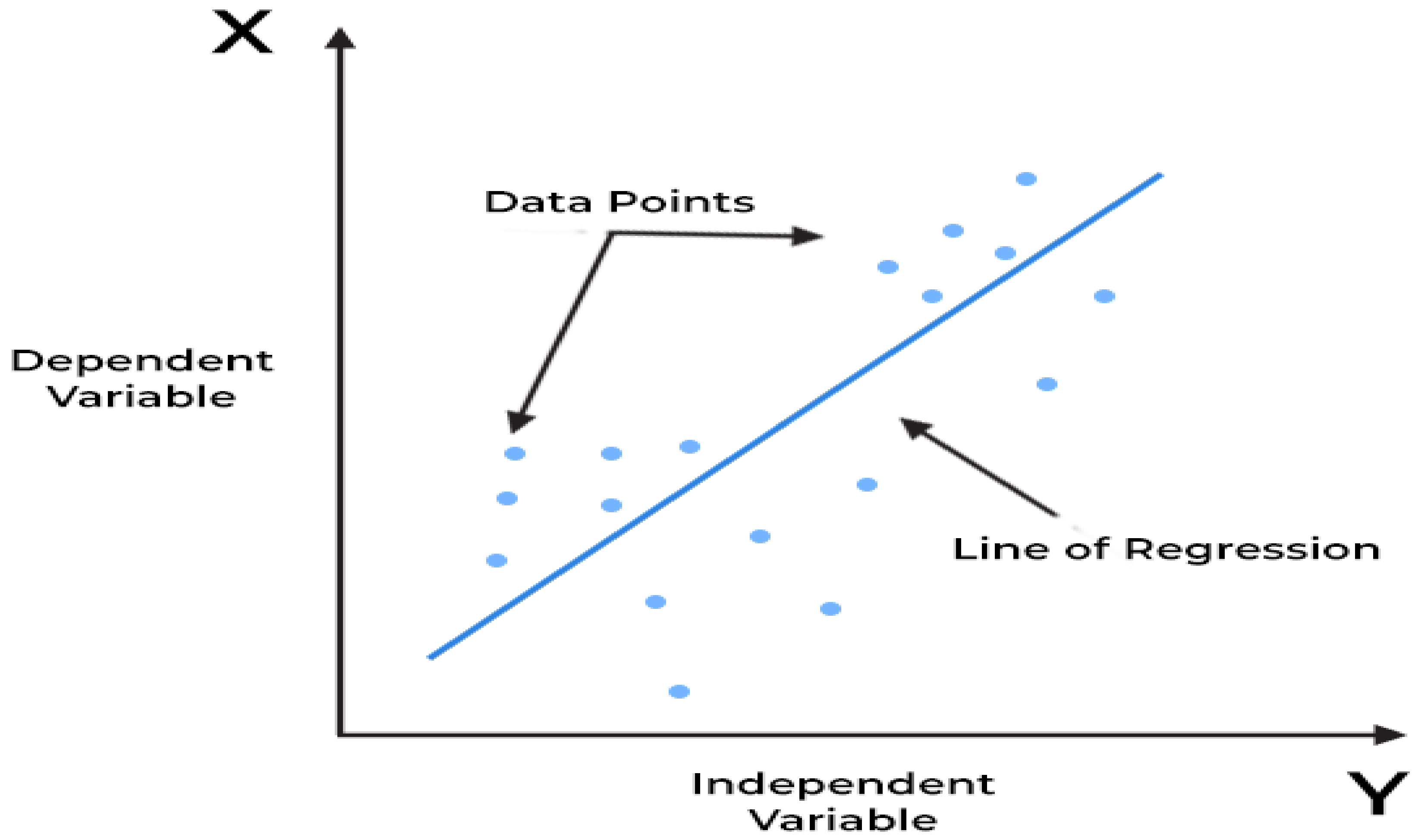
Linear model

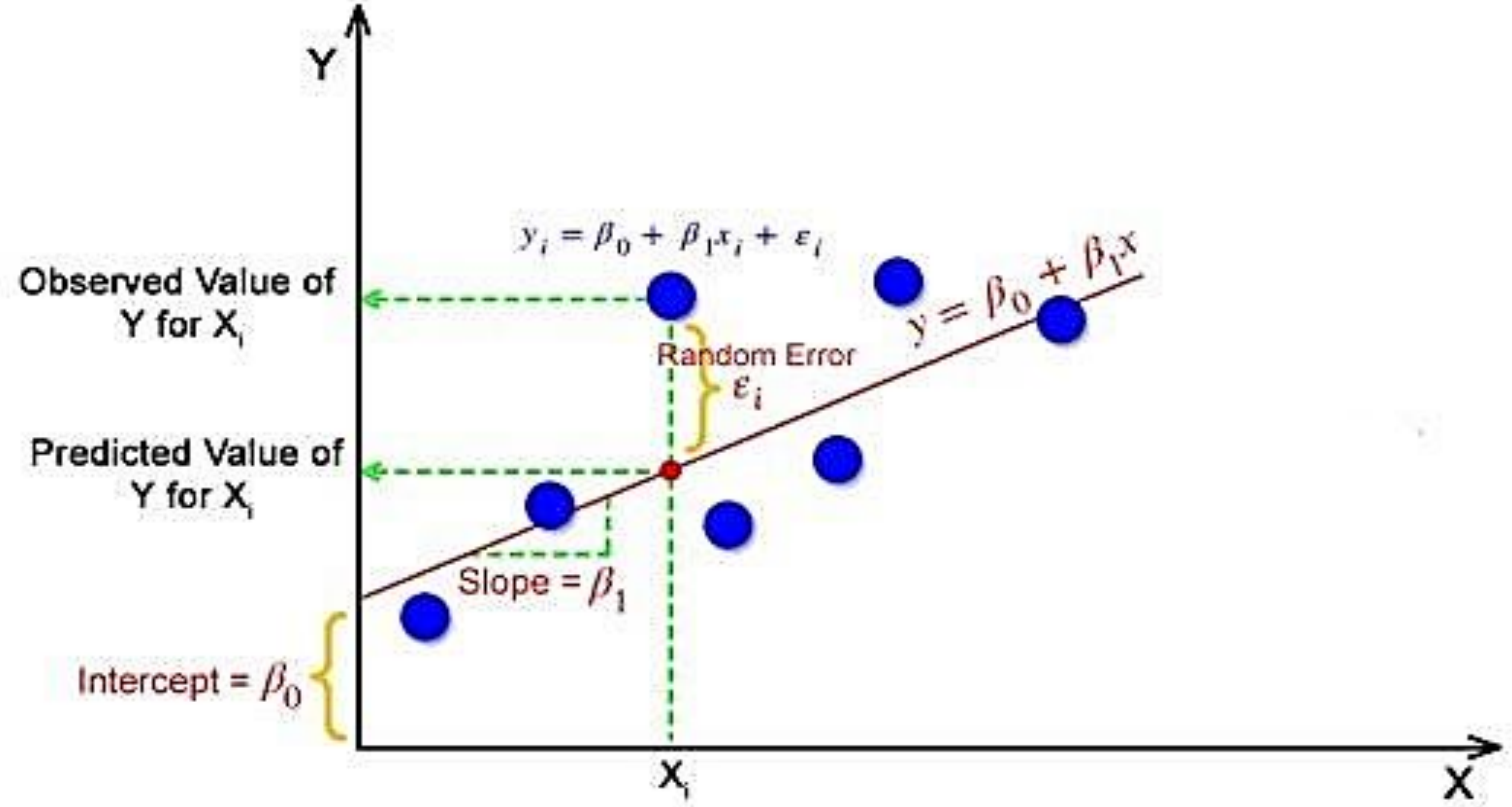
In regression, the relationship between Y and X is modelled in the following form:

$$Y = a + b * X + E$$

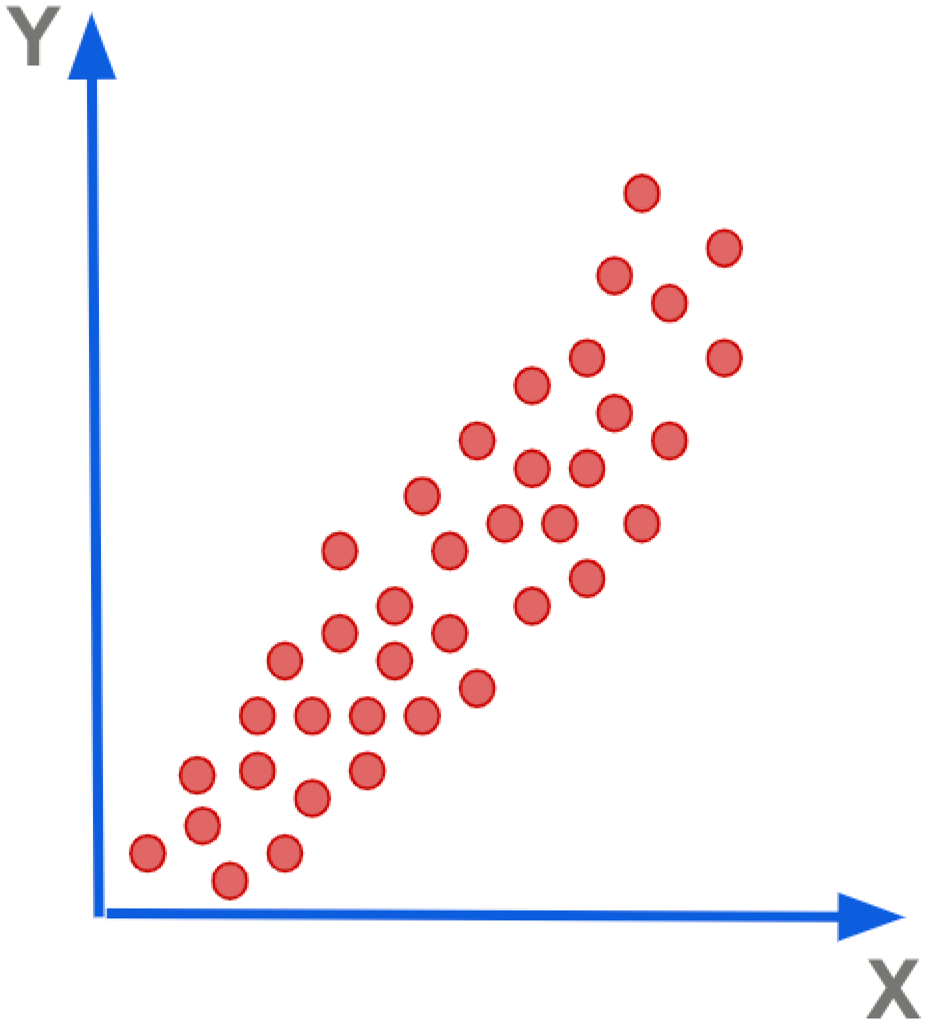
where:

- **Y** is the dependent variable (Income in the example)
- **X** is the independent variable (IQ in the example)
- **a** is an intercept
- **b** is the coefficient
- **E** is an error term for each observation (since there is additional variation not explained by income)

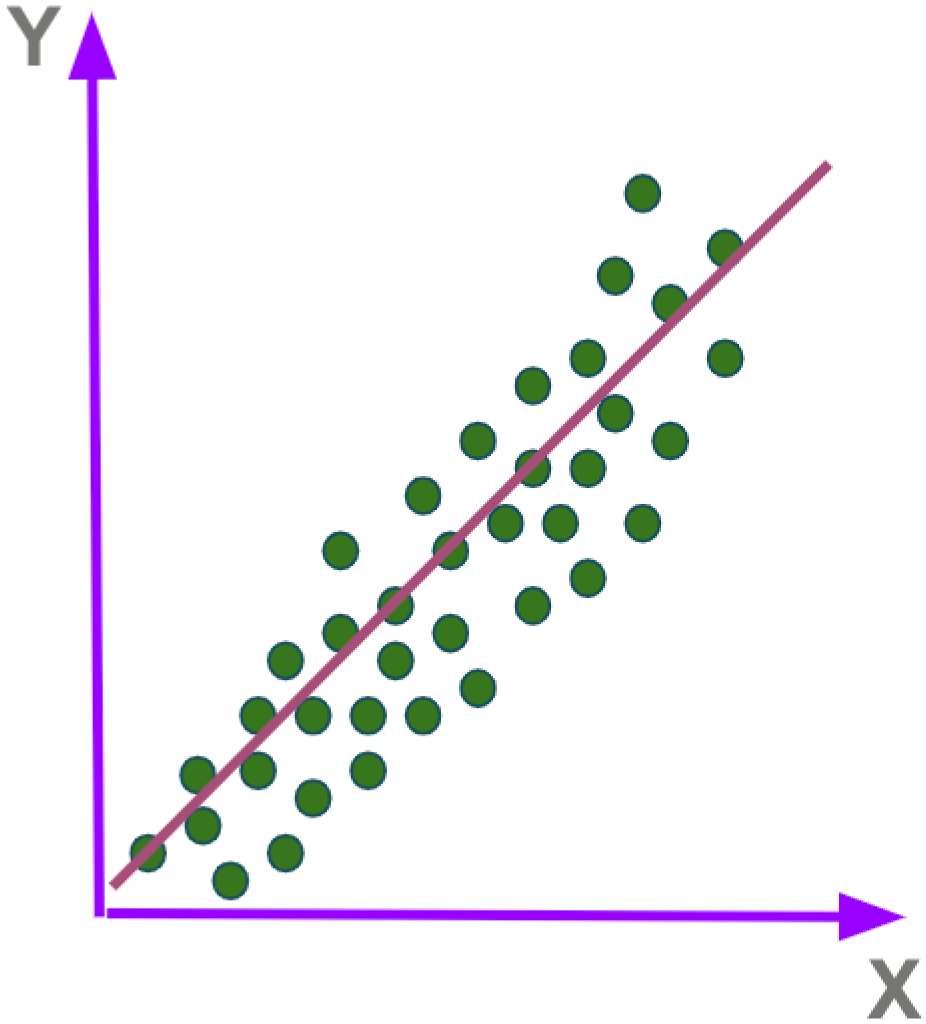

X
Dependent
Variable
Data Points
Line of Regression
Independent
Variable
Y


Y
$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Observed Value of Y for X_i
$y = \beta_0 + \beta_1 x$
Random Error ε_i
Predicted Value of Y for X_i
Slope = β_1
Intercept = β_0
X_i
X

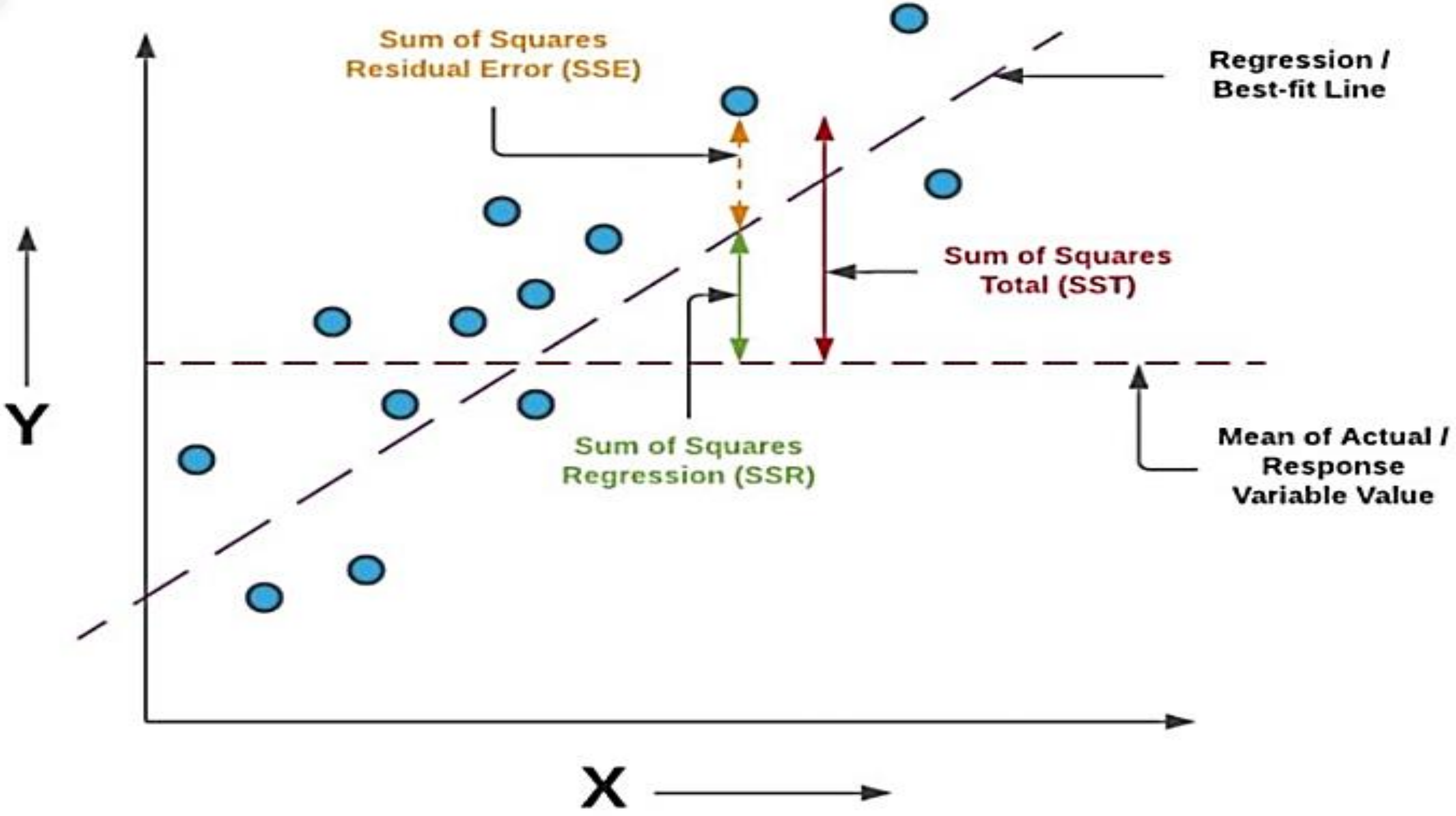
Correlation



Linear Regression



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$


Sum of Squares
Residual Error (SSE)
Regression /
Best-fit Line
Sum of Squares
Total (SST)
Y
Sum of Squares
Regression (SSR)
Mean of Actual /
Response
Variable Value
X

What is multiple linear regression (MLR)? Visual model

Linear Regression

Single predictor $X \longrightarrow Y$

Multiple Linear Regression

Multiple predictors

X_1
X_2
X_3
X_4
X_5

Y

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

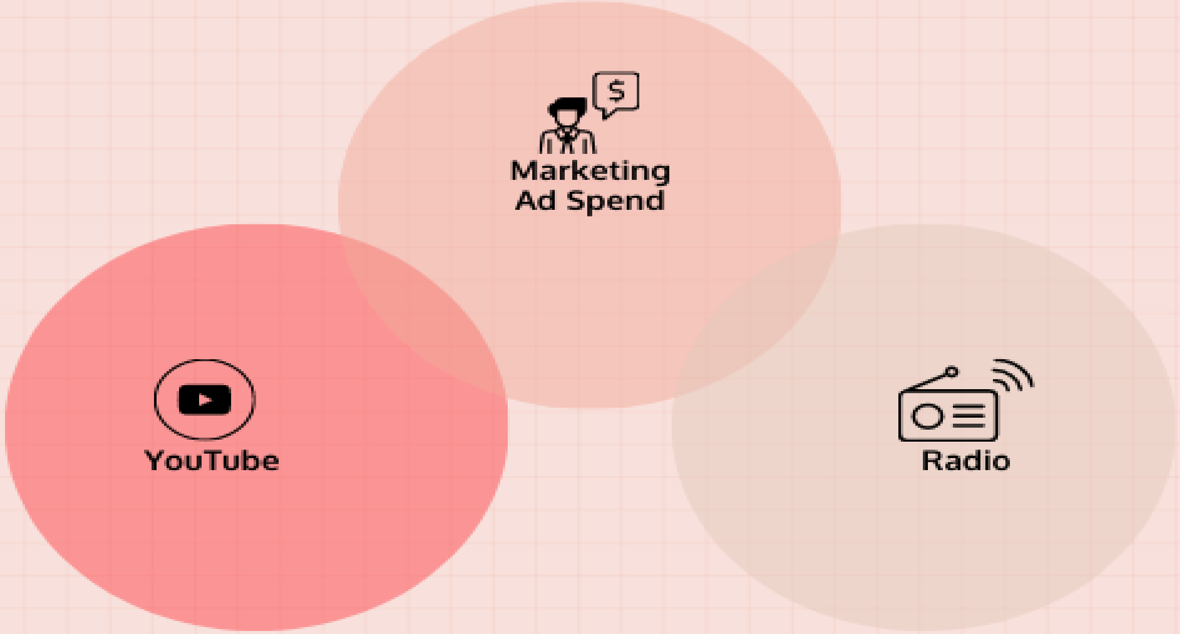
Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Polynomial Linear Regression

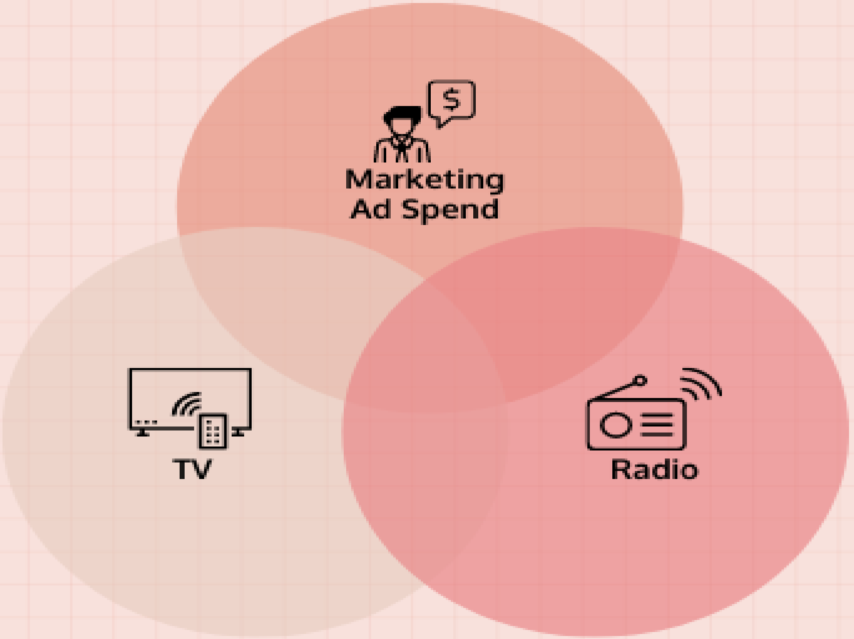
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

The concept of Multicollinearity



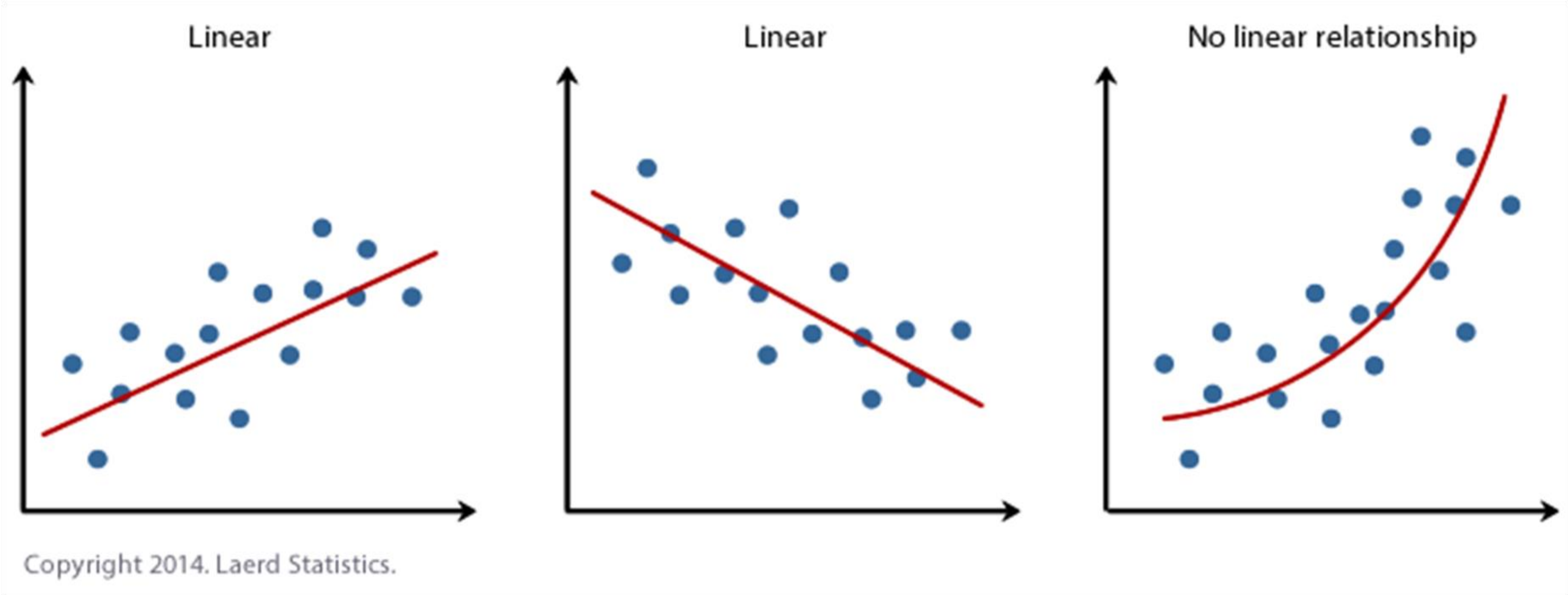
No Multicollinearity

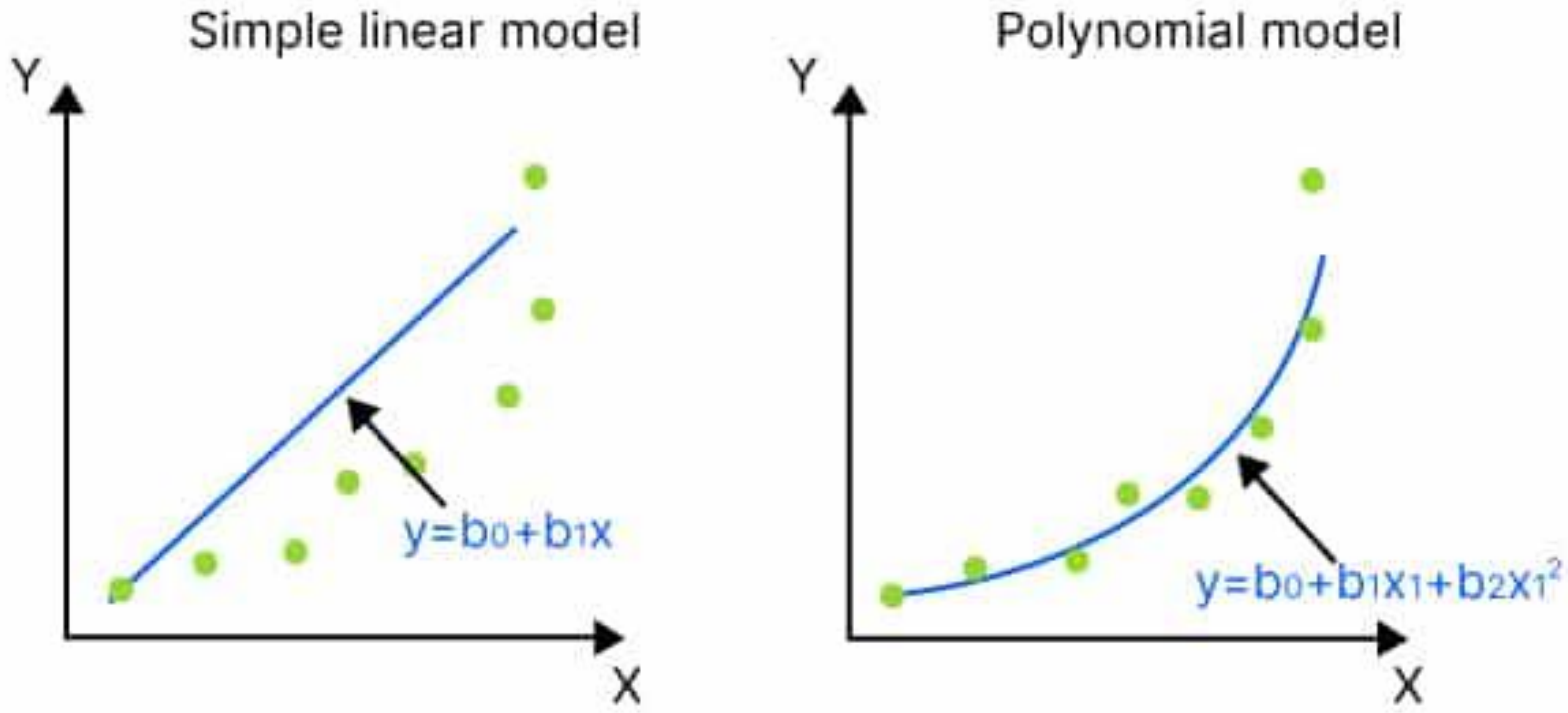
Not correlated - any statistical model will be able to pick apart which one is driving the effects

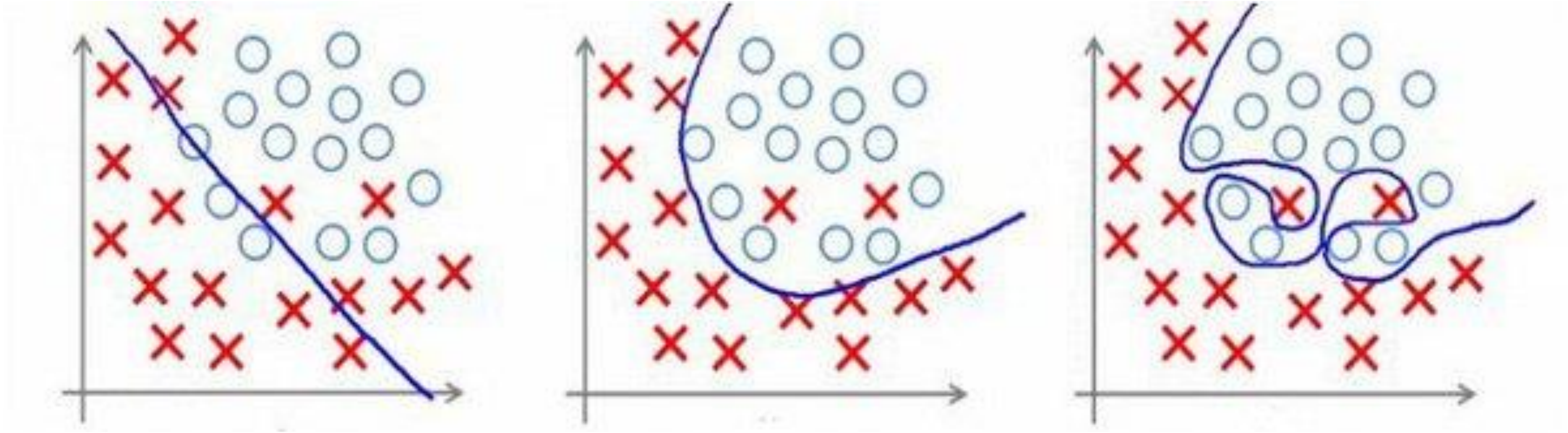


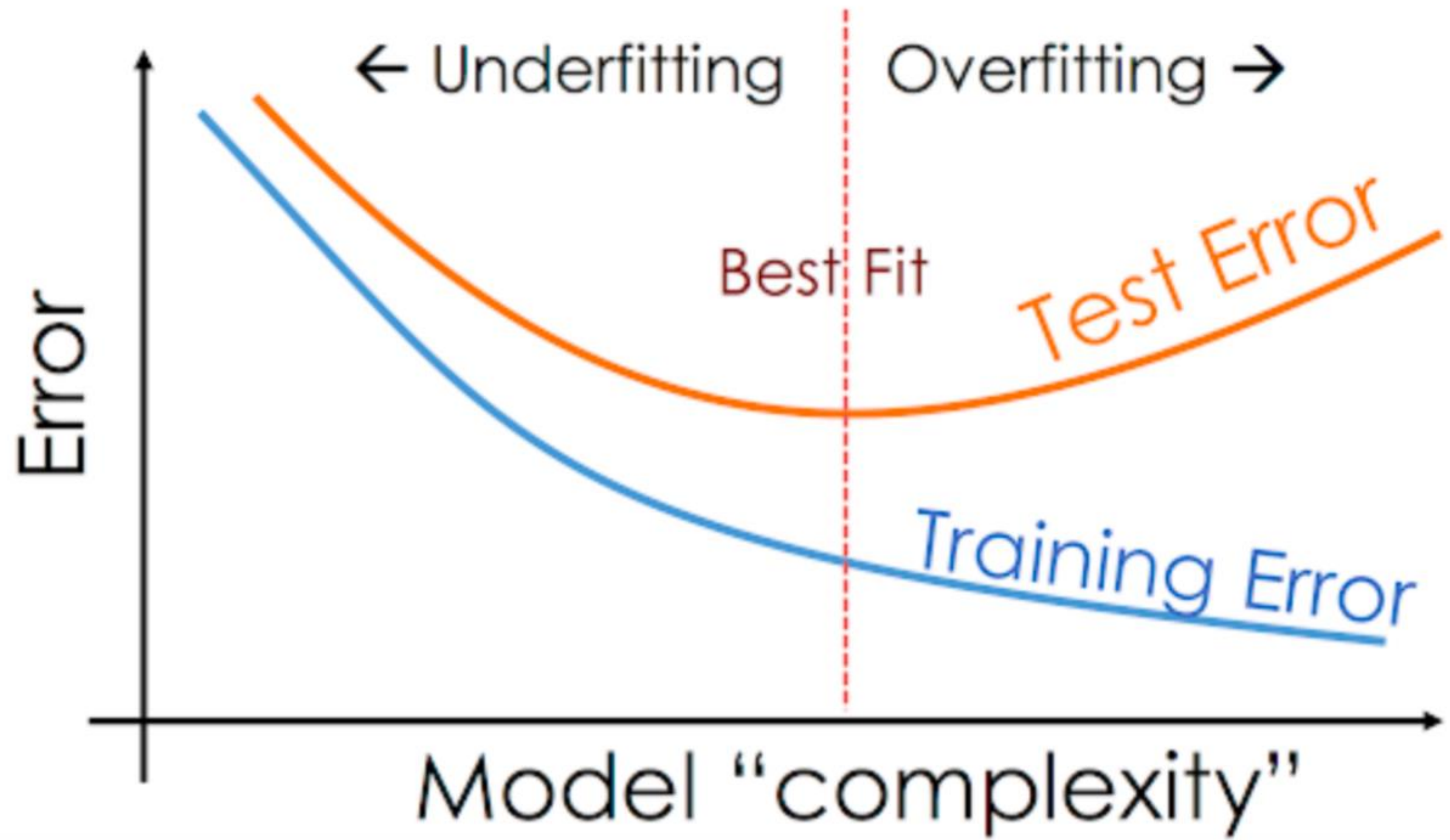
Multicollinearity

Very correlated - no statistical model will be able to pick apart which one is driving the effects

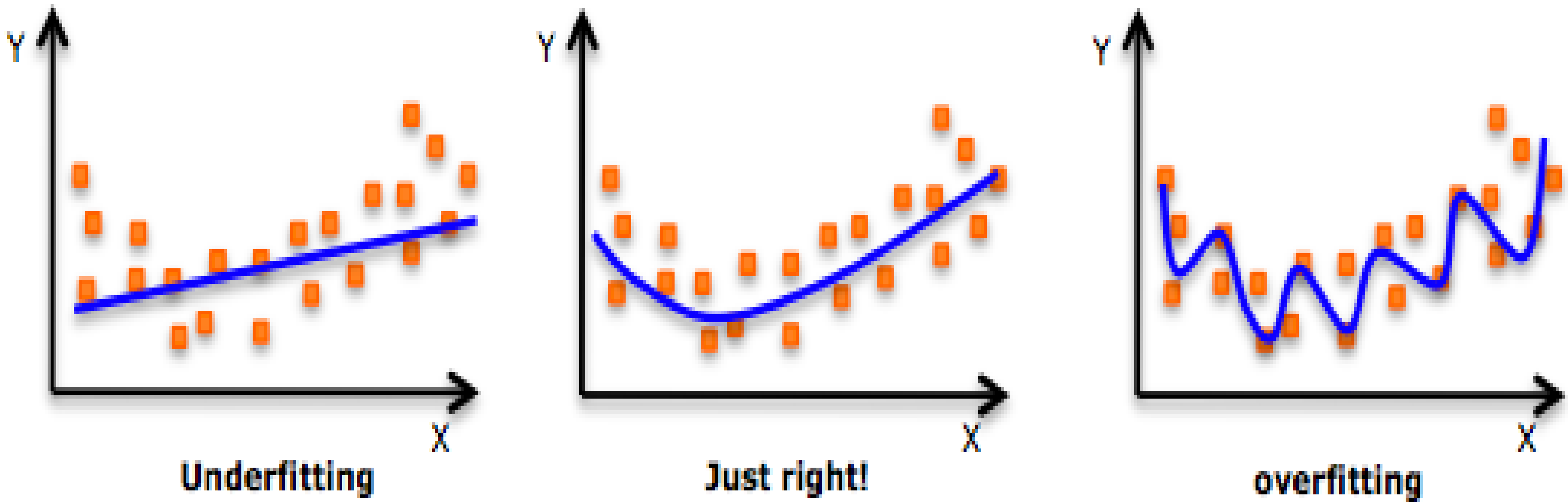

Linear
Linear
No linear relationship
Copyright 2014. Laerd Statistics.

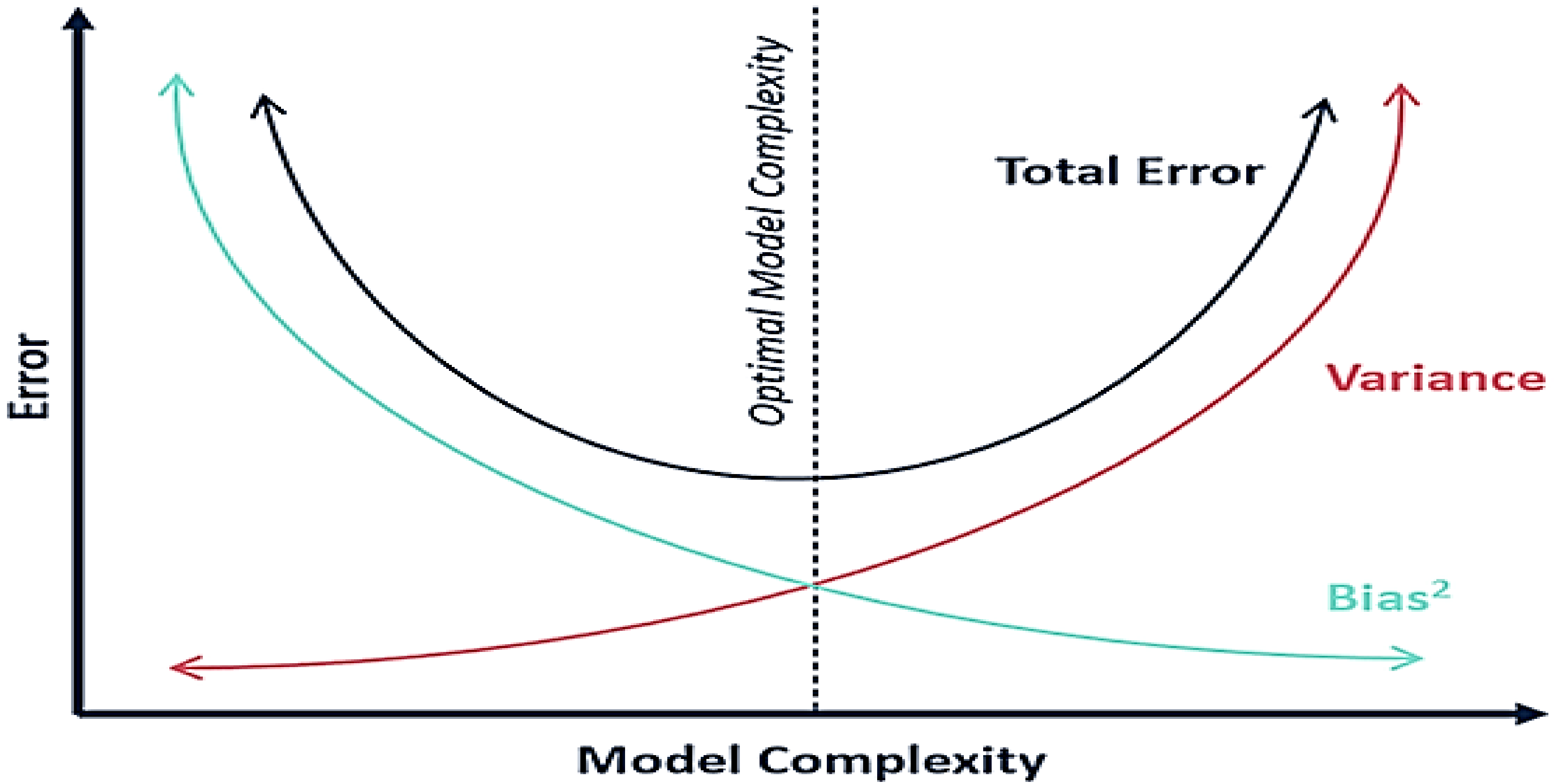

Simple linear model
Y
X
y=b0+b1x
Polynomial model
Y
X
y=b0+b1x1+b2x1²

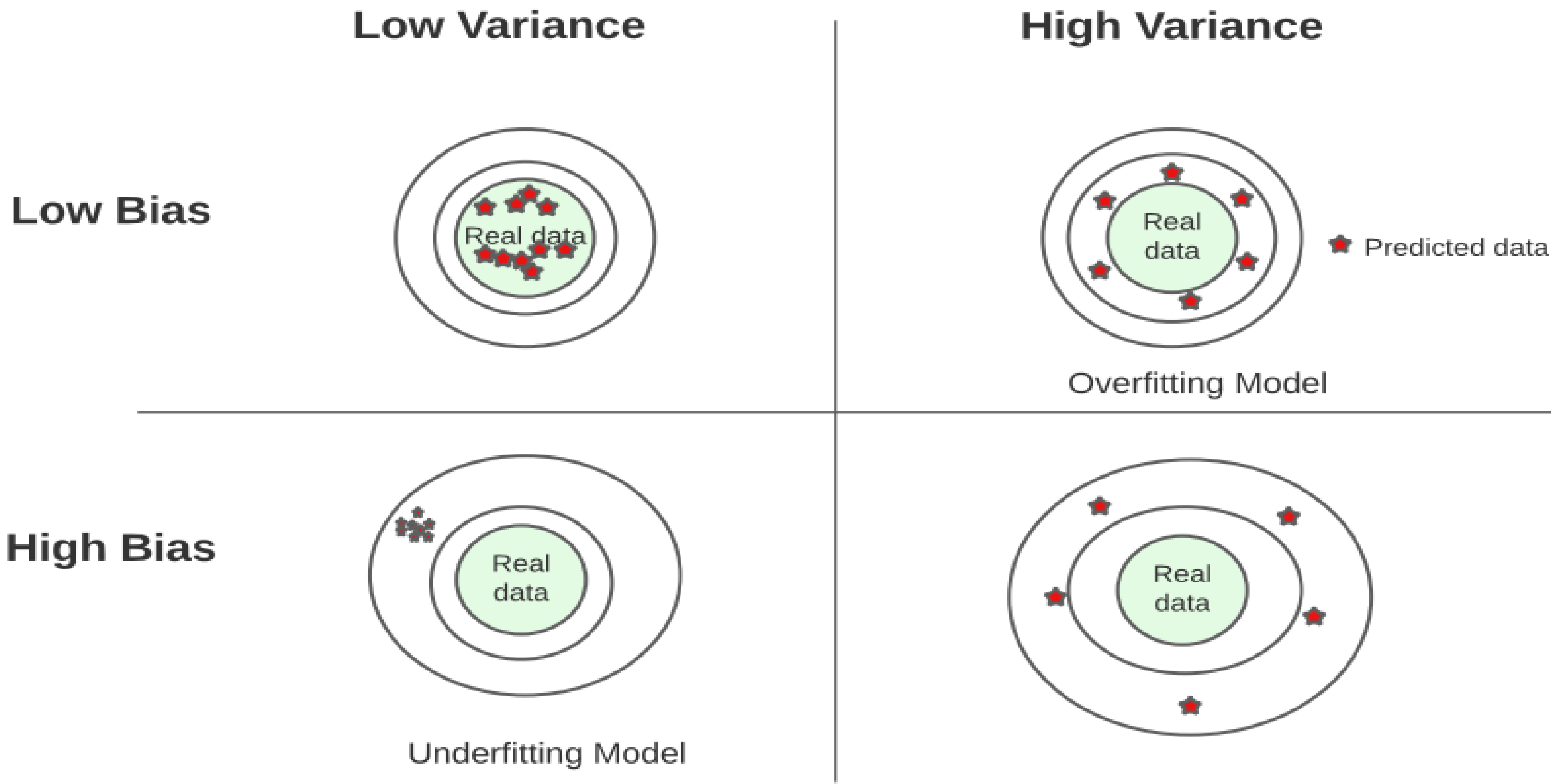


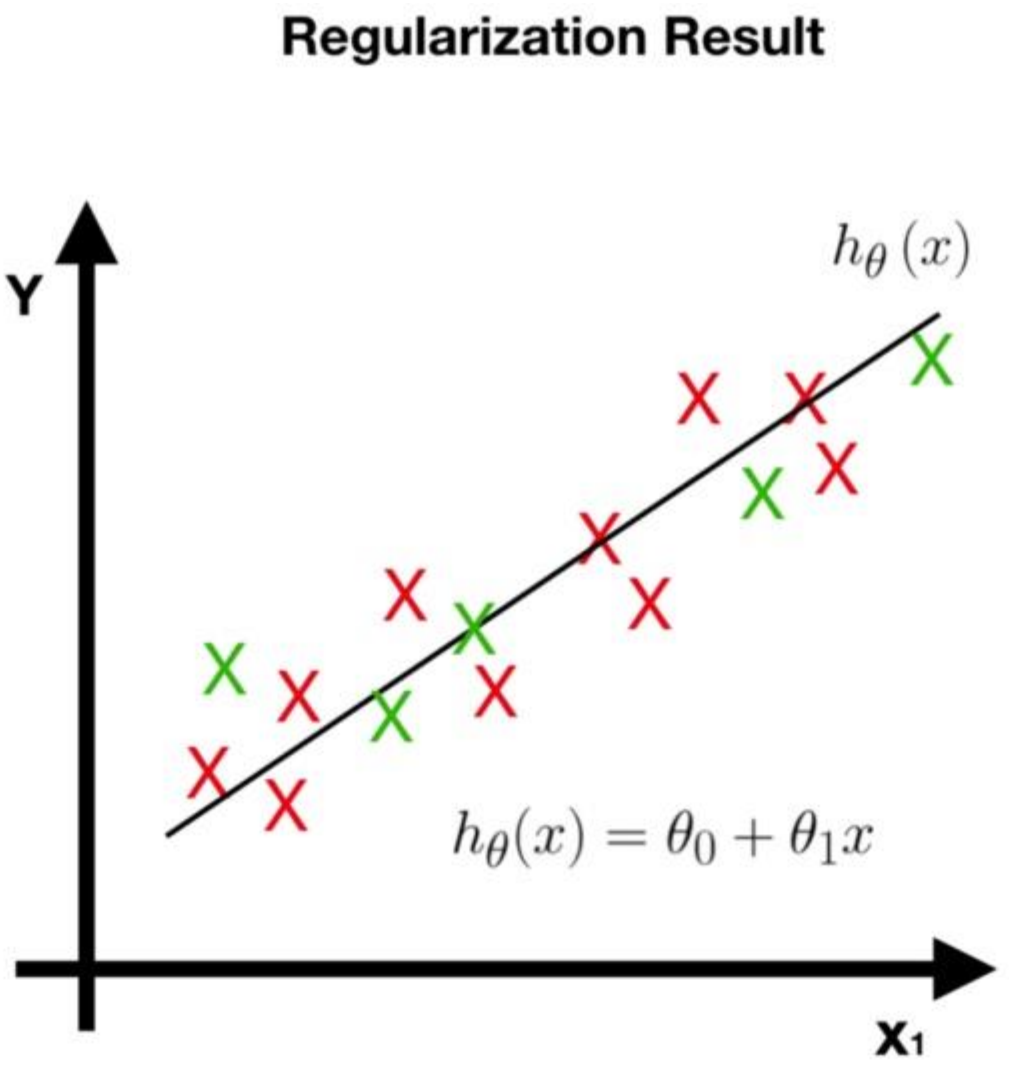


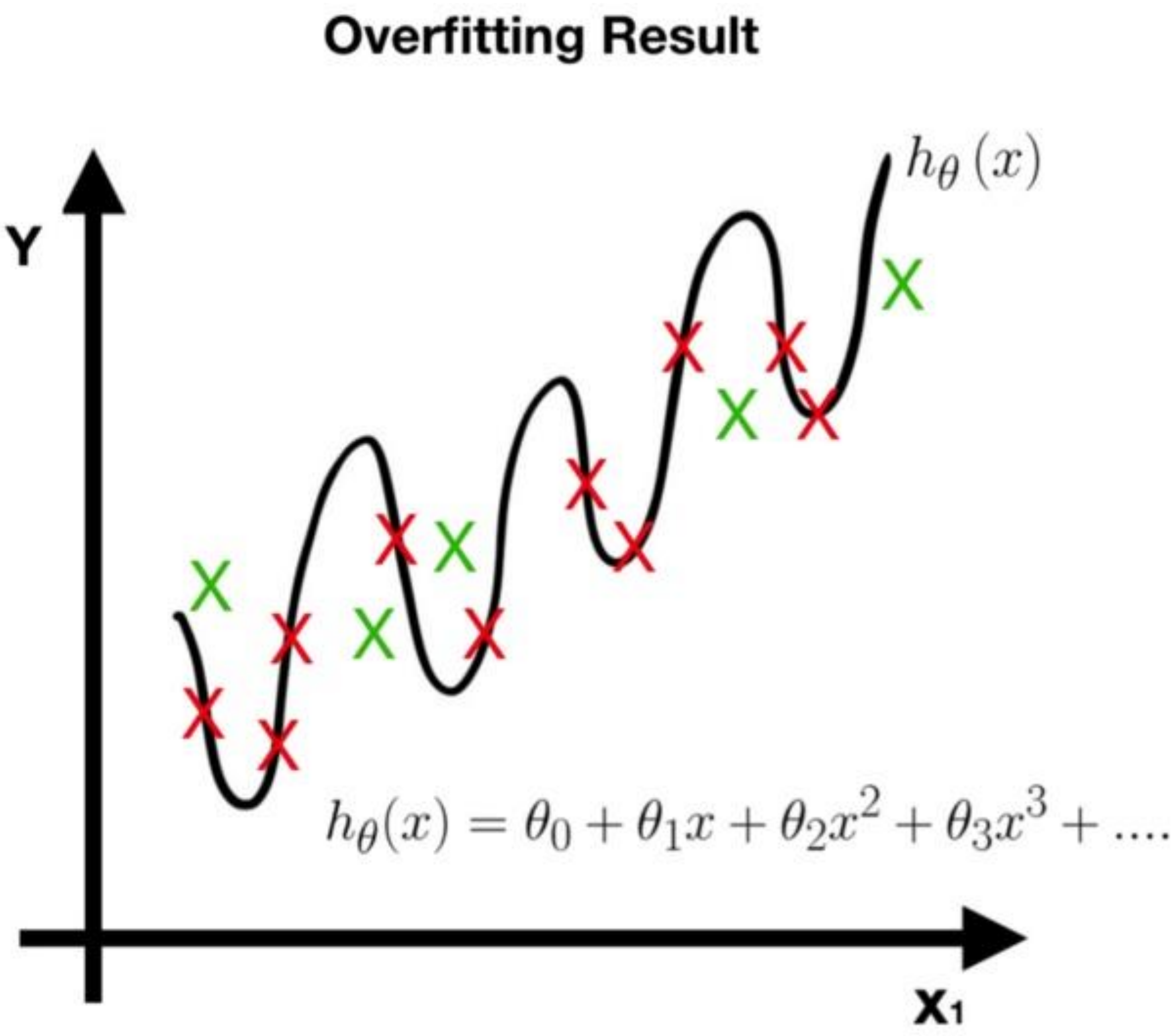
← Underfitting
Overfitting →
Best Fit
Test Error
Training Error
Error
Model "complexity"

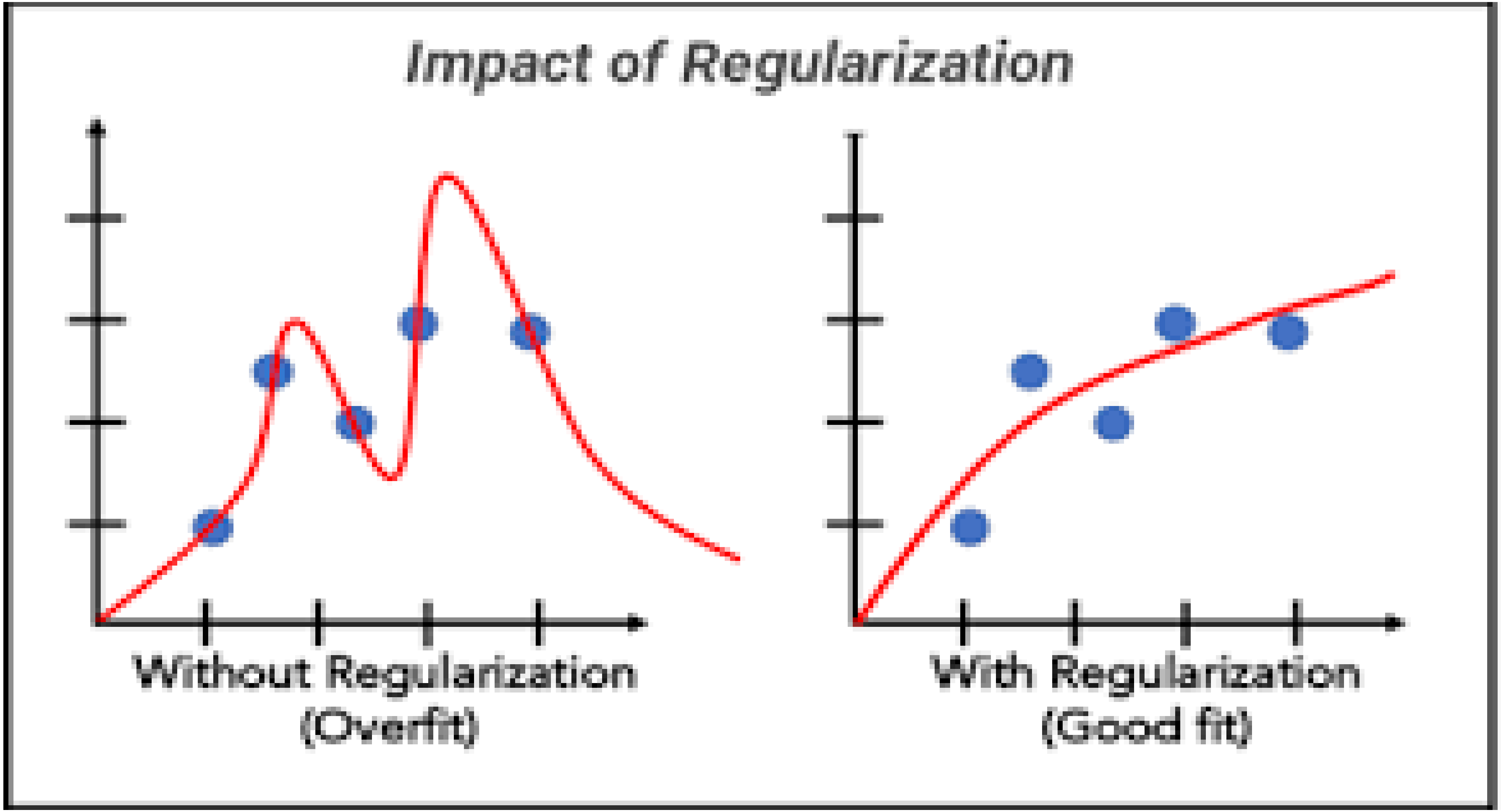

Y
X
Underfitting
Y
X
Just right!
Y
X
overfitting


Error
Optimal Model Complexity
Total Error
Variance
Bias²
Model Complexity


Low Variance
High Variance
Low Bias
High Bias
Real data
Real
data
Predicted data
Overfitting Model
Real
data
Real
data
Underfitting Model


Regularization Result
Y
$h_\theta(x)$
$h_\theta(x) = \theta_0 + \theta_1 x$
X1


Overfitting Result
Y
$h_\theta(x)$
$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + ...$
X1

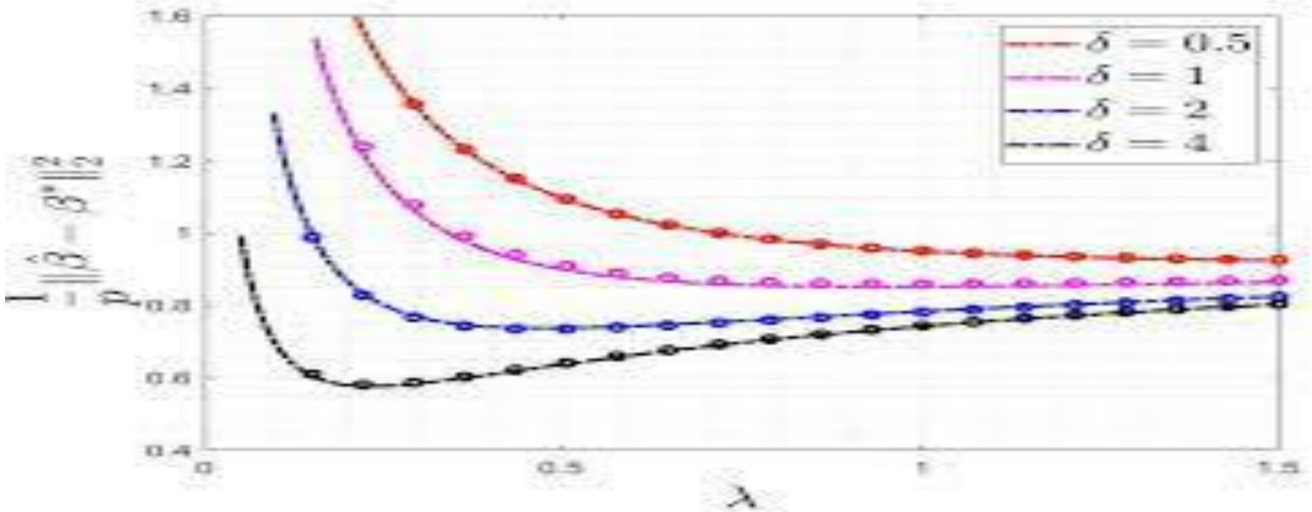

Impact of Regularization
Without Regularization
(Overfit)
With Regularization
(Good fit)

Transforming the Loss function into Lasso Regression

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \Rightarrow \quad \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{p} w_j \times x_{ij}\right)^2 + \lambda \sum_{j=0}^{p}|w_j|$$

Loss function $\qquad$ ***Loss function + Regularized term***

Designed by Author (Shanthababu)



Regularization

- The loss function is the function that computes the distance between the current output of the algorithm and the expected output .
- It's a method to evaluate how your algorithm models the data.
- A loss function specifies a penalty for an incorrect estimate from a statistical model.

Regularization Term

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 + \boxed{\boxed{\lambda}\sum_{j=1}^{n}{\theta_j}^{2}}$$

start at θ_1

Regularization Parameter

Lasso Regression

Lasso regression uses the same mean squared error loss function and this applies L1 Regularization and will repeat the same steps as Ridge. The cost function of Lasso Regression *J(θ)* is given as

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y - \hat{y})^2 + \lambda \sum_{j=1}^{n}|w_j|$$

where

$\lambda \sum_{j=1}^{n}|w_j|$ is the penalty (L1 Regularization).

$$Loss\ Function = \frac{1}{m}\sum_{i=1}^{n}(y - \hat{y})^2$$

$$Regularization\ (L1) = \frac{1}{m}\sum_{i=1}^{n}(y - \hat{y})^2 + \lambda \sum_{j=1}^{m}|w_i|$$

Ridge Regression

Ridge regression uses the mean squared error loss function and applies L2 Regularization. Its cost function *J(θ)* is given as

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}(y-\hat{y})^2 + \lambda\sum_{j=1}^{n} w_j^2$$

where,

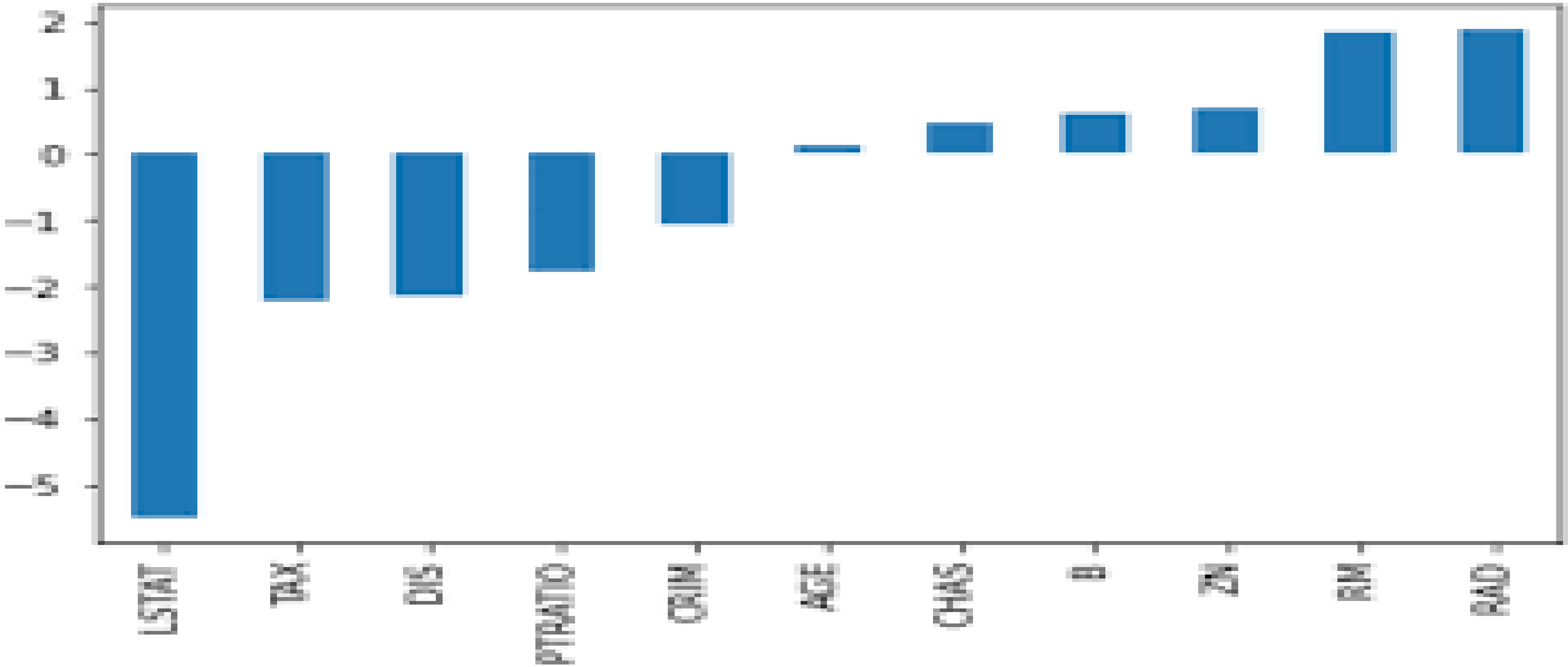
$\frac{1}{m}\sum_{i=1}^{m}(y-\hat{y})^2$ is the Mean Squared error (loss function)

$\lambda\sum_{j=1}^{n} w_j^2$ is the penalty (L2 Regularization)

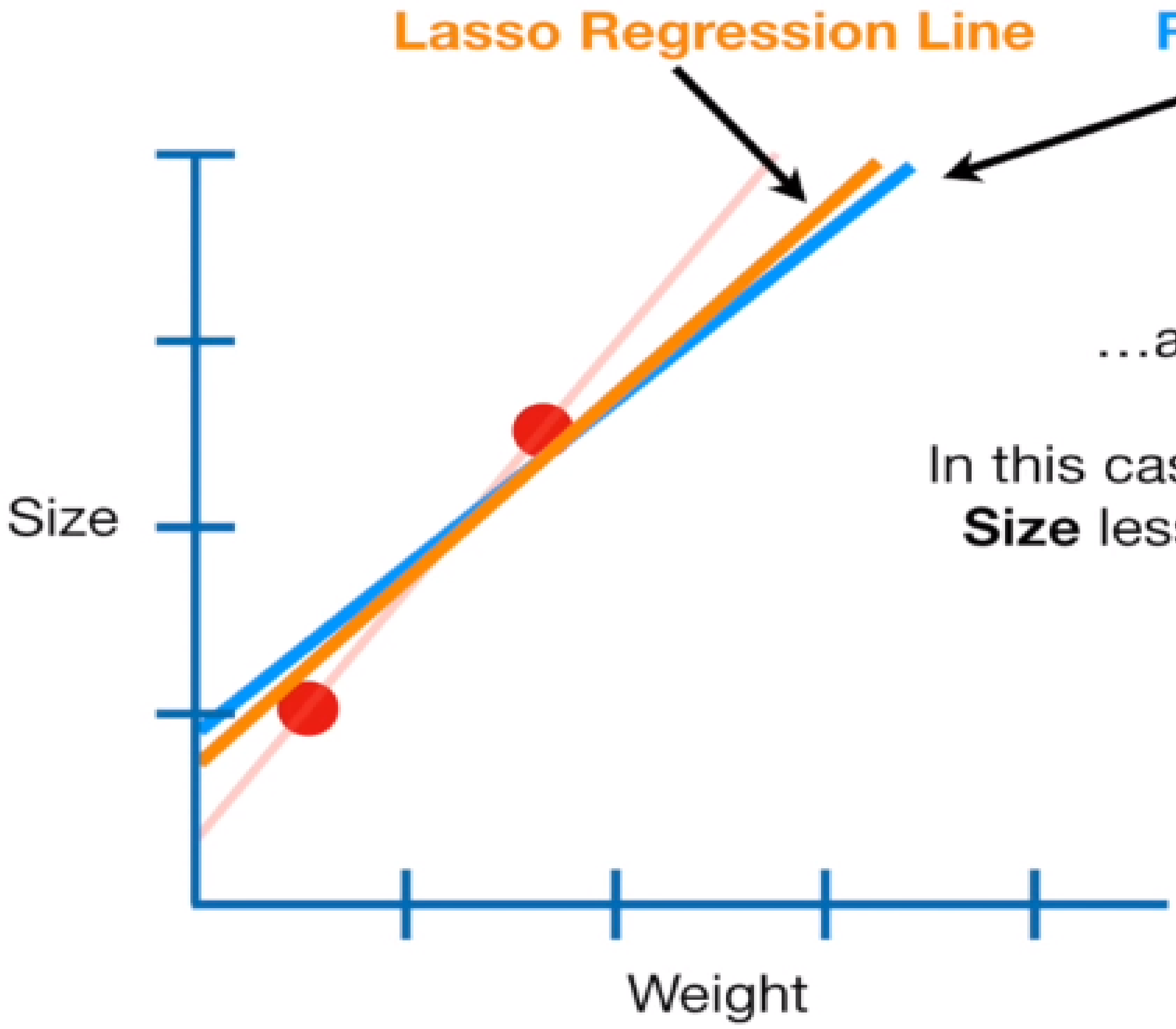
Now, substitute $\hat{y}$ as $wx_i + b$.

Ridge Regression

$$Regularization\ (L2) = Loss\ Function + \lambda \sum_{i=1}^{m} w_i^2$$

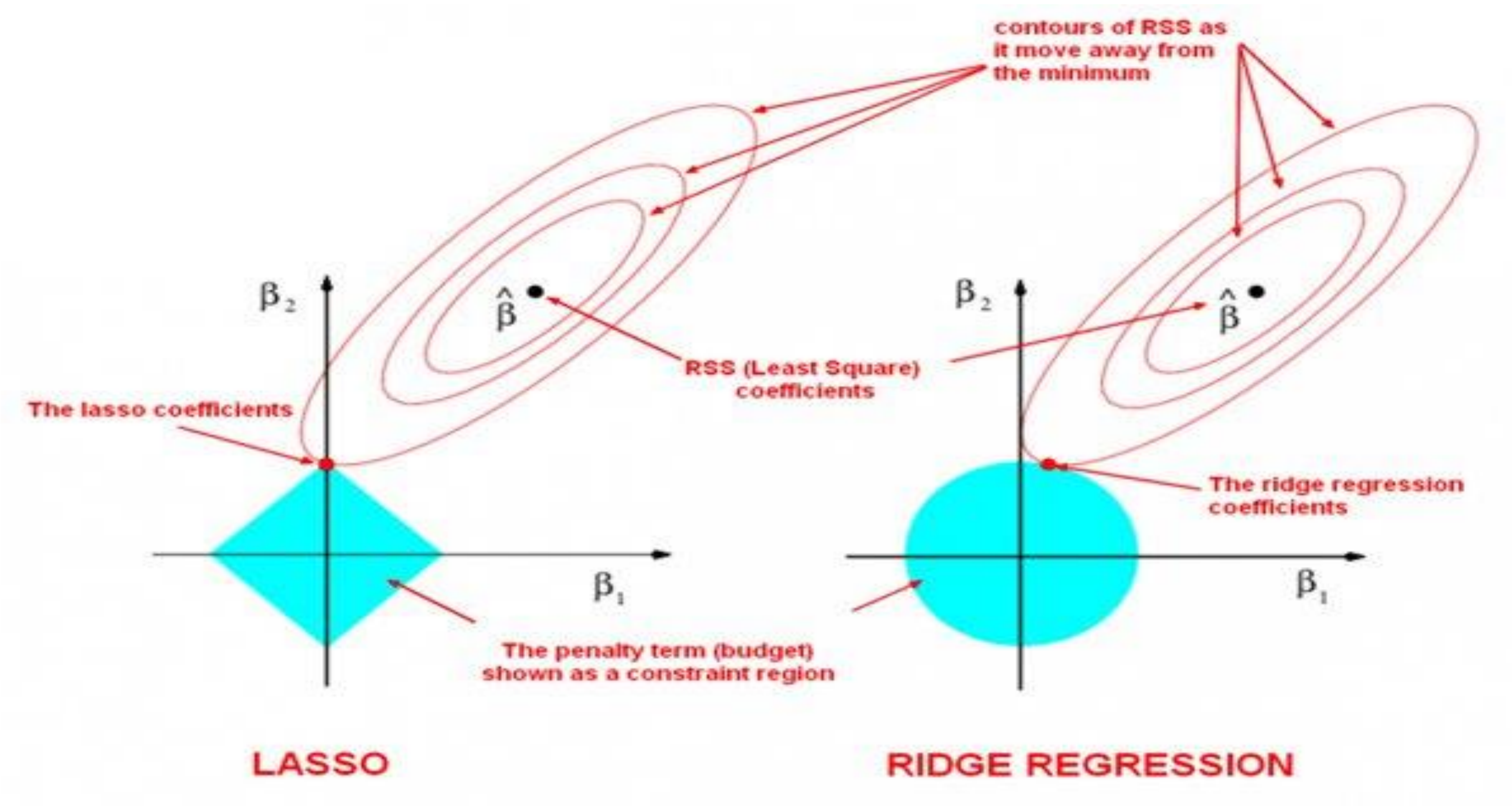


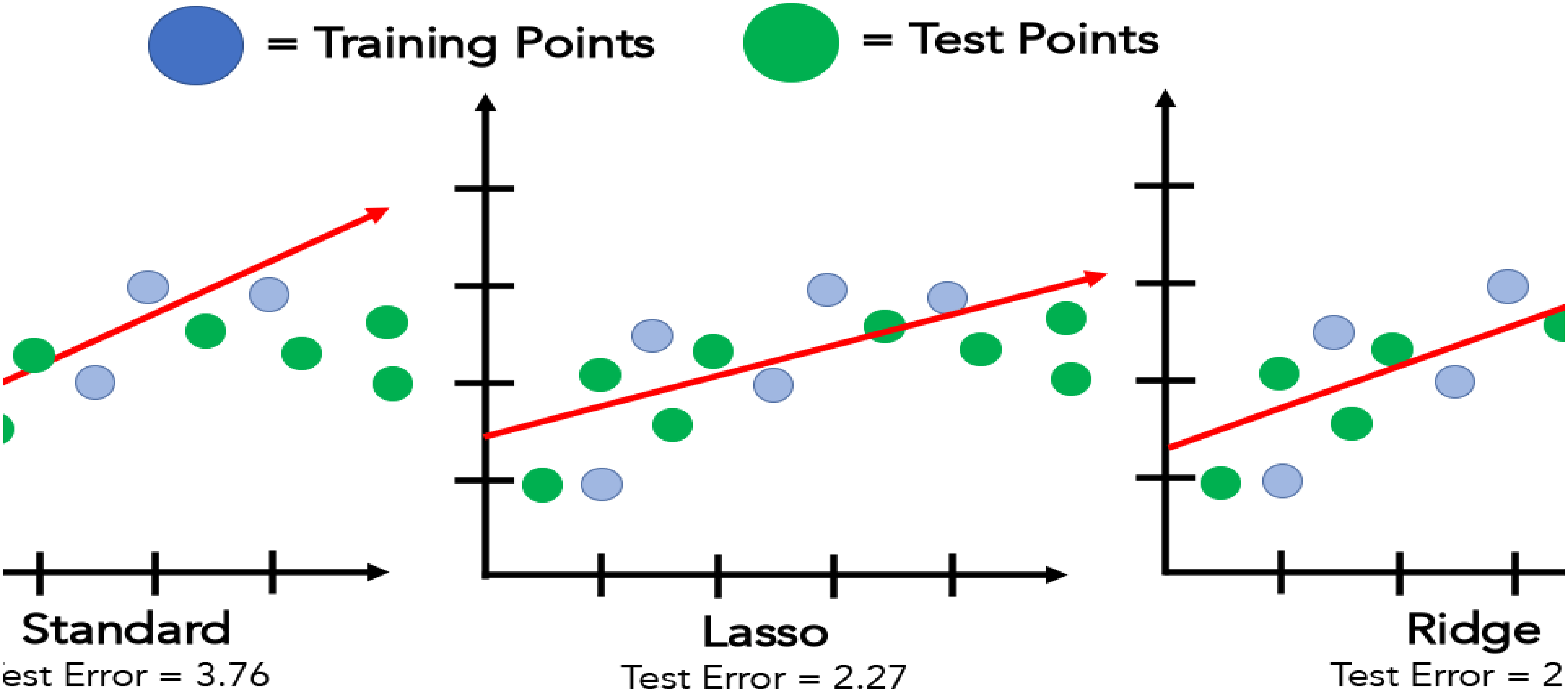
2
1
0
−1
−2
−3
−4
−5
LSTAT
TAX
DIS
PTRATIO
CRIM
AGE
CHAS
B
ZN
RM
RAD



...and they do similar things.

In this case, they make our predictions of **Size** less sensitive to this tiny **Training Dataset**.


contours of RSS as it move away from the minimum
β_2
$\hat{\beta}$
RSS (Least Square) coefficients
The lasso coefficients
β_1
The penalty term (budget) shown as a constraint region
β_2
$\hat{\beta}$
The ridge regression coefficients
β_1
LASSO
RIDGE REGRESSION



**Our Data was actually "parabolic" but we couldn't tell from the small training sample.*