Regression Analysis and Learning

- ✓ **Regression analysis** is a set of statistical processes for estimating the relationships among variables.
- ✓ From biology: 'Heights of descendants of tall ancestors tend to regress down towards a normal average (regression toward the mean)'.
- ✓ Most commonly, estimates the average value of the dependent variable when the independent variables are fixed.
- ✓ Widely used in statistics for prediction and forecasting, and substantially overlap with the field of machine learning.

❖ **Linear Regression**:

- Simplest form of Regression
- Data are modeled using a straight line [Fitting a straight line]
- Bivariate Linear Regression (dependent and independent variables) is similar to Univariate function, $y = f(x) = ax + b$, where y-output, x-input(variable)

- Linear Regression Learning Problem:

  $Y = \alpha X + \beta$

  Y – random variable (response, dependent)

  X – random variable (predictor, independent)

  α, β - regression coefficients, that are **to be learned**

- To solve means to find estimated values of α and β that best describes the field data
- Methods of **least squares** can be used to find α and β minimizing error between the actual data and the estimate of the line.
- Traditionally, from Gauss, the squared loss function values, summed over all training examples are minimized, yielding

  $\alpha = \Sigma_{i=1:s} (x_i - x') (y_i - y') / \Sigma_{i=1:s} (x_i - x')^2$, $\beta = y' - \alpha x'$,

  where x' - average of $x_1, x_2, \ldots , x_s$, y' - of $y_1, y_2, \ldots , y_s$,

  given sample data points $(x_1, y_1), (x_2, y_2), \ldots, (x_s, y_s)$.

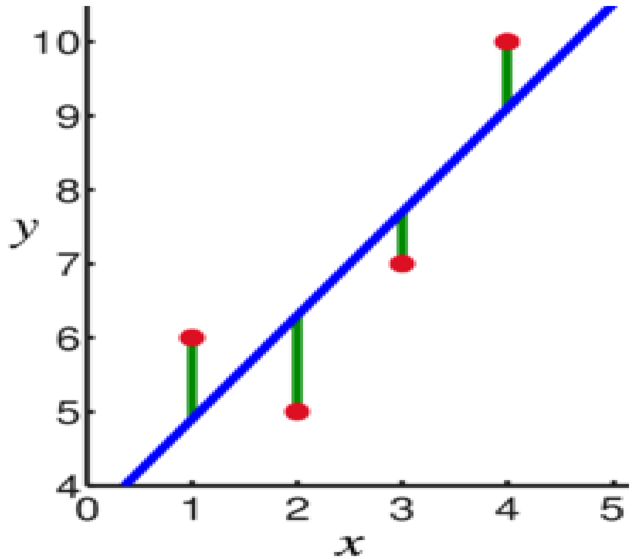
- The line thus obtained can be used to predict an appropriate value of y, given an unknown x.

- Mean Absolute Error (MAE, L1 loss) is sometimes used to assess performance of a model that does not consider the direction of the outliers.
- For a data point y_i and its predicted value $\hat{y}_i$, where n is the total number of data points in the dataset:

$$MAE = \Sigma_{i=1:n} |y_i - \hat{y}_i| / n$$

- Mean Squared Error (MSE, L2 loss) is also used which is computed as follows:

$$MSE = \Sigma_{i=1:n} (y_i - \hat{y}_i)^2 / n$$



[Source: Internet]

- **Example.** Sample data (Salary data)

Serial	X (Year of experience)	Y (Salary in 1000 Taka)
1	3	30
2	8	57
3	9	64
4	13	72
5	3	36
6	6	43
7	11	59
8	21	90
9	1	20
10	16	83

✓ We get, Y = 3.5X + 23.6, and from it predict 58.6K salary after 10 years of experience.

- We can think of multiple regression like the one below:

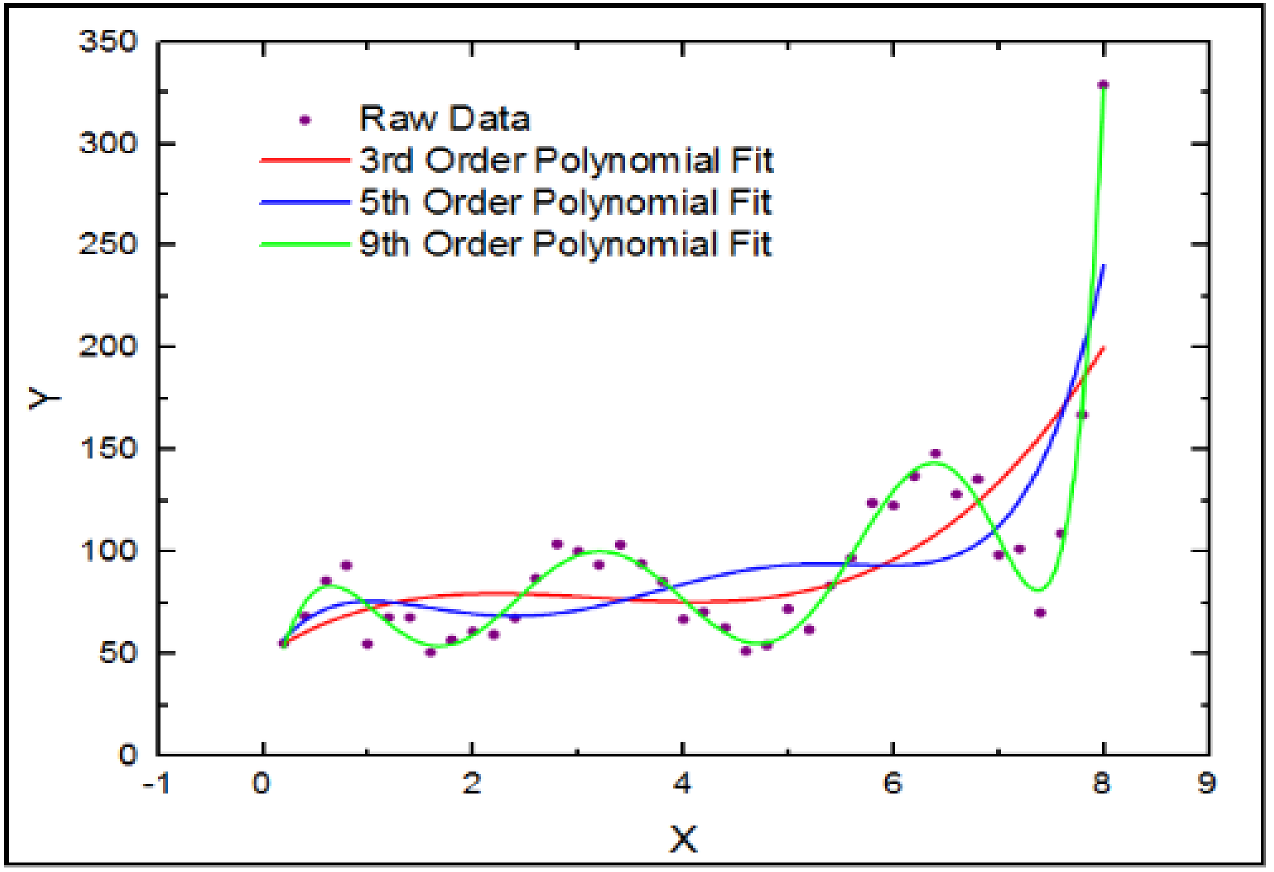
$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \beta,$$

which can also be solved using least squares method.

➢ And nonlinear regression (polynomial) like the one below:

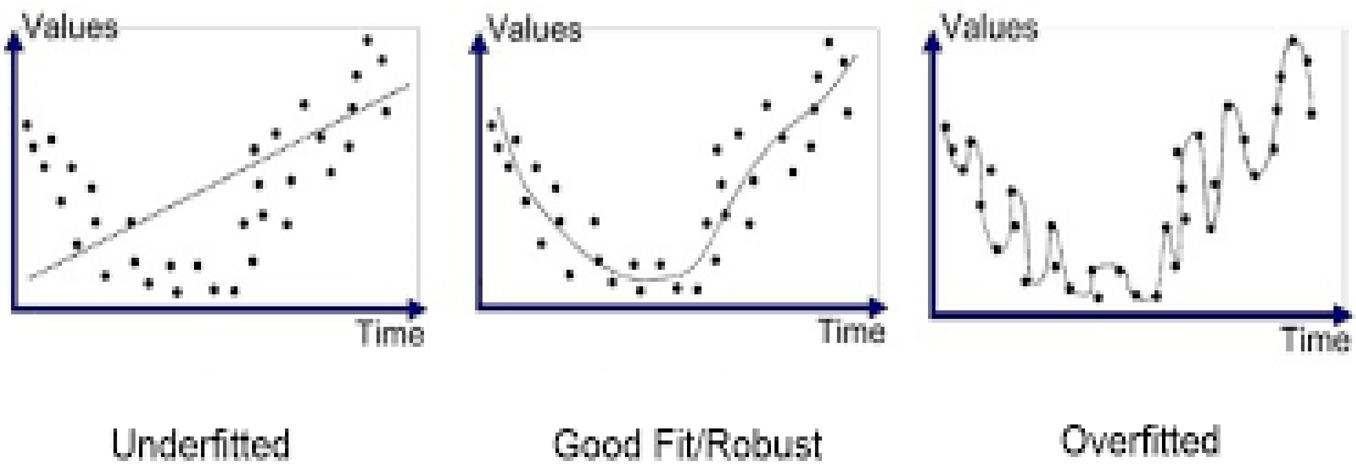
$$Y = \alpha_3 X^3 + \alpha_2 X^2 + \alpha_1 X + \beta,$$

transforming it, and applying least squares method.



[Source: Internet]

- Mind overfitting and underfitting models with data



[Source: Internet]