Advanced Analytical Theory and Methods: Regression Analysis

Prof. Dr. Shamim Akhter

Professor, Dept. of CSE

Ahsanullah University of Science and Technology

Simple Regression

- **Regression** model is a mathematical equation
  - describes the relationships between variables.
- **Simple regression** model includes only two variables: one independent(X) and one dependent(Y).

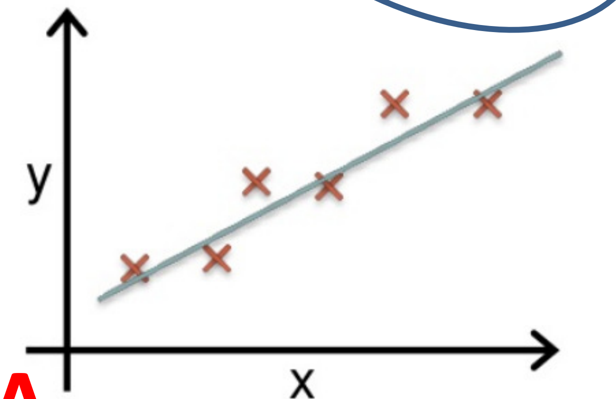
Explains the variation in Y

One being explained

Input to output define by

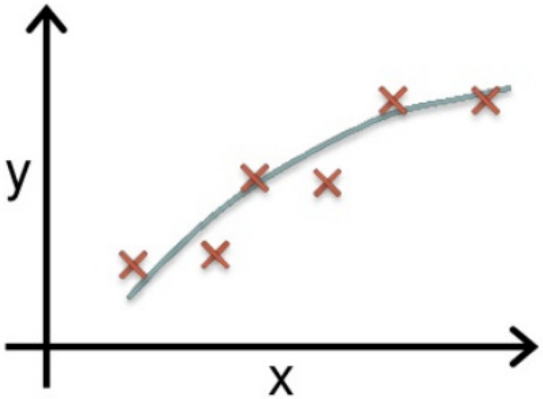
- a linear function (A)

- a non-linear function(B)



A

Linear regression

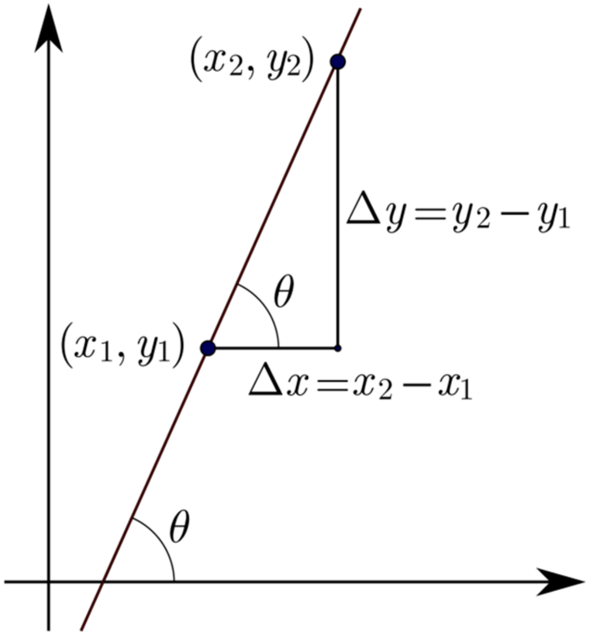


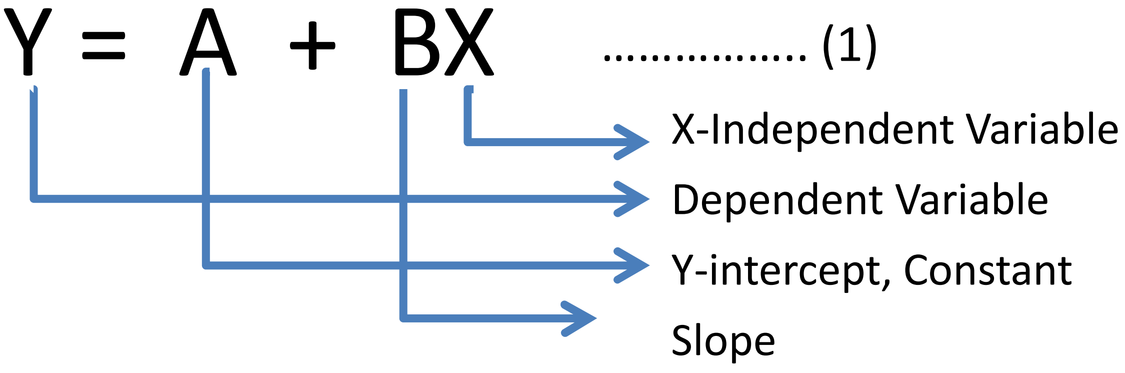
B

Nonlinear regression

Linear Regression

- Gives a straight-line (linear) relationship between two variables is called a linear regression model





$$Y = A + BX \quad \text{................ (1)}$$

X-Independent Variable

Dependent Variable

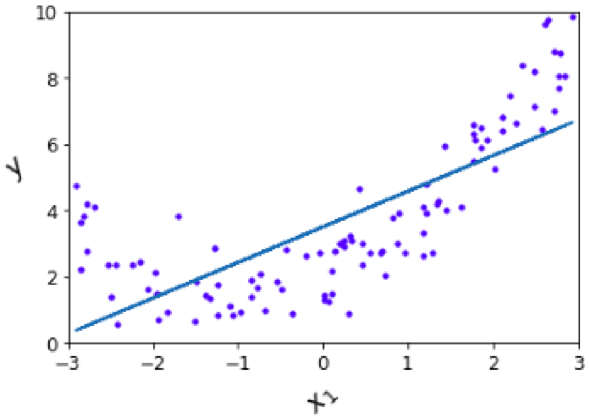
Y-intercept, Constant

Slope

Slope = tanΘ = changes in Y/changes in X

This model (1) is deterministic

○ gives exact relation ship between X and Y

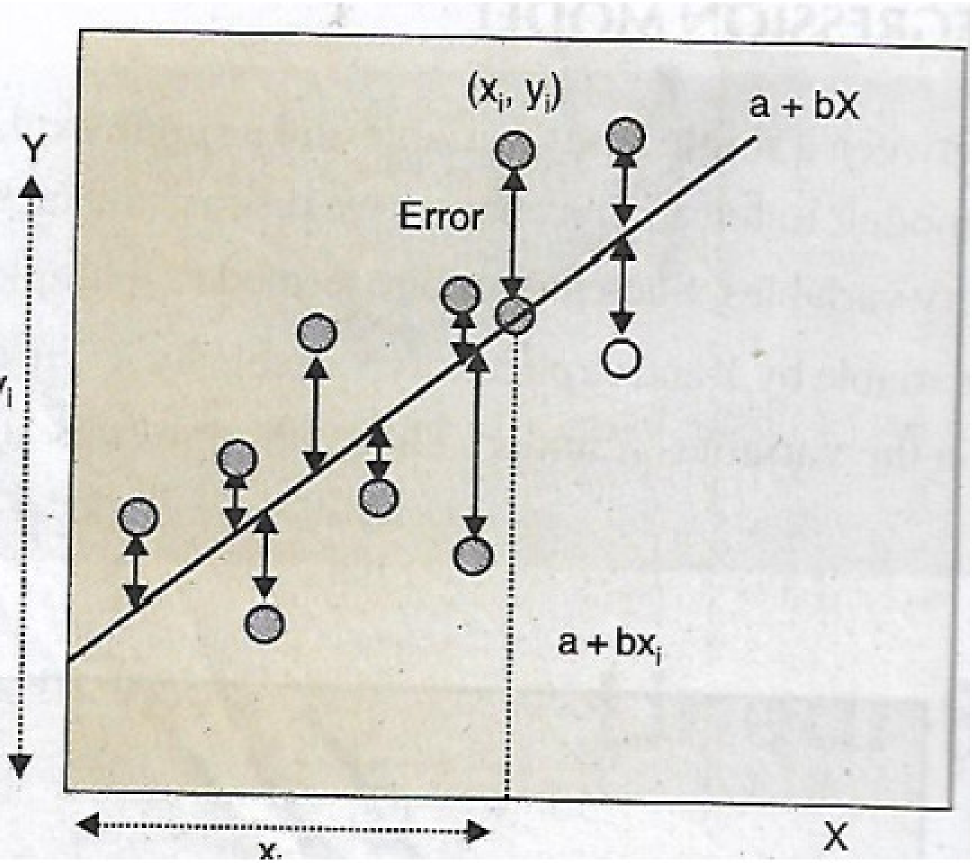


How about for (plotting) scatter outputs and requires a good fitting line ?

○ Introduces some random error

○ Objective : reduce the overall error

Regression with Random Error



$$Y = A + BX + e$$

Value of the error e

-positive if point positioned above the regression line.

-negative if point below the regression line.

e added two phenomena:

- ○ Included missing or omitted variables
- ○ Support Random variation in Y value

Regression is a learning process??

Learning refers to find a mapping that reduce the error on a set of training data.

Task: Most Suitable A & B Need to find out

-To minimize error

Direct regression : Least Squares Error

$$S(A, B) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - A - BX_i)^2$$

- Partial Derivatives with respect to A

$$\frac{\partial S(A, B)}{\partial A} = -2\sum_{i=1}^{n}(Y_i - A - BX_i)$$

- Partial Derivatives with respect to B

$$\frac{\partial S(A, B)}{\partial B} = -2\sum_{i=1}^{n}(Y_i - A - BX_i)Xi$$

- The solutions of A and B are obtained by setting

$$\frac{\partial S(A, B)}{\partial A} = 0 \quad \text{and} \quad \frac{\partial S(A, B)}{\partial B} = 0$$

The solutions of the two equations are called the **direct regression estimators,** or usually called as the **ordinary least squares (OLS)** estimators of A and B.

$$A = \bar{Y} - B\bar{X}$$

$$B = \frac{SS_{XY}}{SS_{XX}}$$

X bar and Y bar are the mean

$$SS_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$SS_{XX} = \sum (X_i - \bar{X})^2$$

Covariance: Covariance is a measure of how much two random variables vary together. It's similar to variance, but **co** variance tells how **two** variables vary together.

Variance : measures how far a set of (random) numbers are spread out from their average value.

Example 14.1: Watching television also reduces the amount of physical exercise, causing weigh gains. A sample of fifteen 10-year old children was taken. The number of pounds each child wa overweight was recorded (a negative number indicates the child is underweight). Additionally, th number of hours of television viewing per weeks was also recorded. These data are listed here.

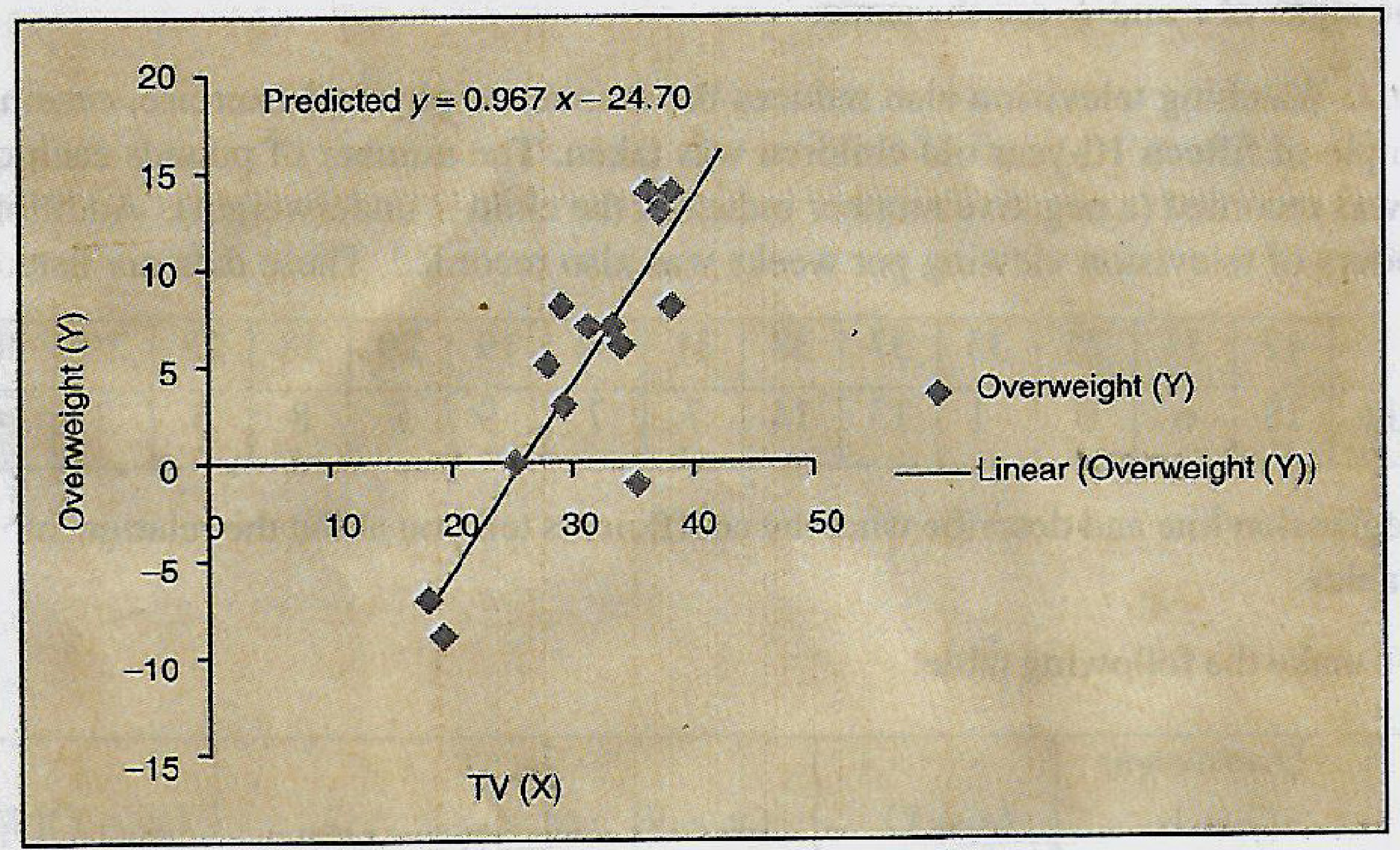
TV	42	34	25	35	37	38	31	33	19	29	38	28	29	36	18
Overweight	18	6	0	–1	13	14	7	7	–9	8	8	5	3	14	–7

Fit the regression line and describe what the coefficients tell you about the relationship betwee the two variables.

Solution: We make the following table:

TV (x_i)	*Overweight* (y_i)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
42	18	10.5333	12.2667	110.9511	129.2089
34	6	2.5333	0.2667	6.4178	0.6756
25	0	– 6.4667	– 5.7333	41.8178	37.0756
35	– 1	3.5333	– 6.7333	12.4844	– 23.7911
37	13	5.5333	7.2667	30.6178	40.2089
38	14	6.5333	8.2667	42.6844	54.0089
31	7	– 0.4667	1.2667	0.2178	– 0.5911
33	7	1.5333	1.2667	2.3511	1.9422
19	– 9	–12.4667	– 14.7333	155.4178	183.6756
29	8	– 2.4667	2.2667	6.0844	– 5.5911
38	8	6.5333	2.2667	42.6844	14.8089
28	5	– 3.4667	– 0.7333	12.0178	2.5422
29	3	– 2.4667	– 2.7333	6.0844	6.7422
36	14	4.5333	8.2667	20.5511	37.4756
18	–7	– 13.4667	– 12.7333	181.3511	171.4756
$\bar{x}$ = 31.4667	$\bar{y}$ = 5.7333	← Mean	Total →	*SSX* = 671.7333	*SSXY* = 649.8667

Using the above calculation, we obtain $\hat{b} = \frac{SSXY}{SSX} = \frac{649.8667}{671.7333} = 0.9674$ and $\hat{a} = \bar{y} - \hat{b}\bar{x} = 5.733 - 0.9674 \times 31.4667 = -24.709$. Therefore, the fitted simple linear regression model is $\hat{y} = -24.709 + 0.9674\,x$



Goodness of the Fitting Model

- Residuals of a fitting model
  - can be utilized to tell us the goodness of the model
- Once we finalized- “how the Y values are changing”
  - we can predict the value of Y values as well
  - Variance of Y values is $\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$
  - We can partition it as

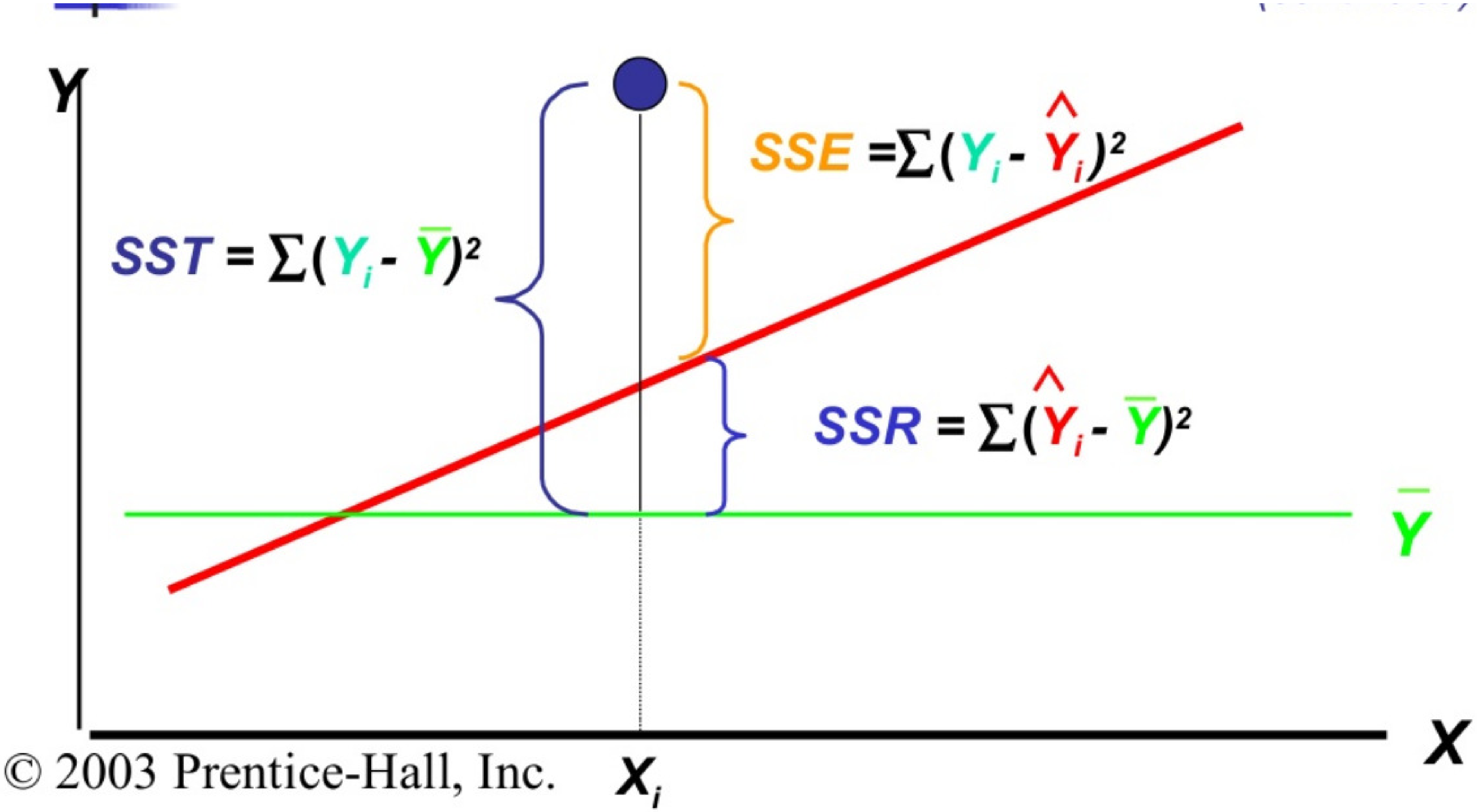
$$\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \frac{1}{N}\sum_{1}^{N}\left(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y}\right)^2$$

$$= \frac{1}{N}\sum_{1}^{N}\left(Y_i - \hat{Y}_i\right)^2 + \frac{1}{N}\sum_{1}^{N}\left(\hat{Y}_i - \bar{Y}\right)^2$$

$$\mathbf{SST = SSE + SSR}$$

Total Sample variability Error in Y is Sum of squares due to **error-unexplained variability** and sum of squares due to regression – **explained variability**

Measures of Variation: The Sum of Square

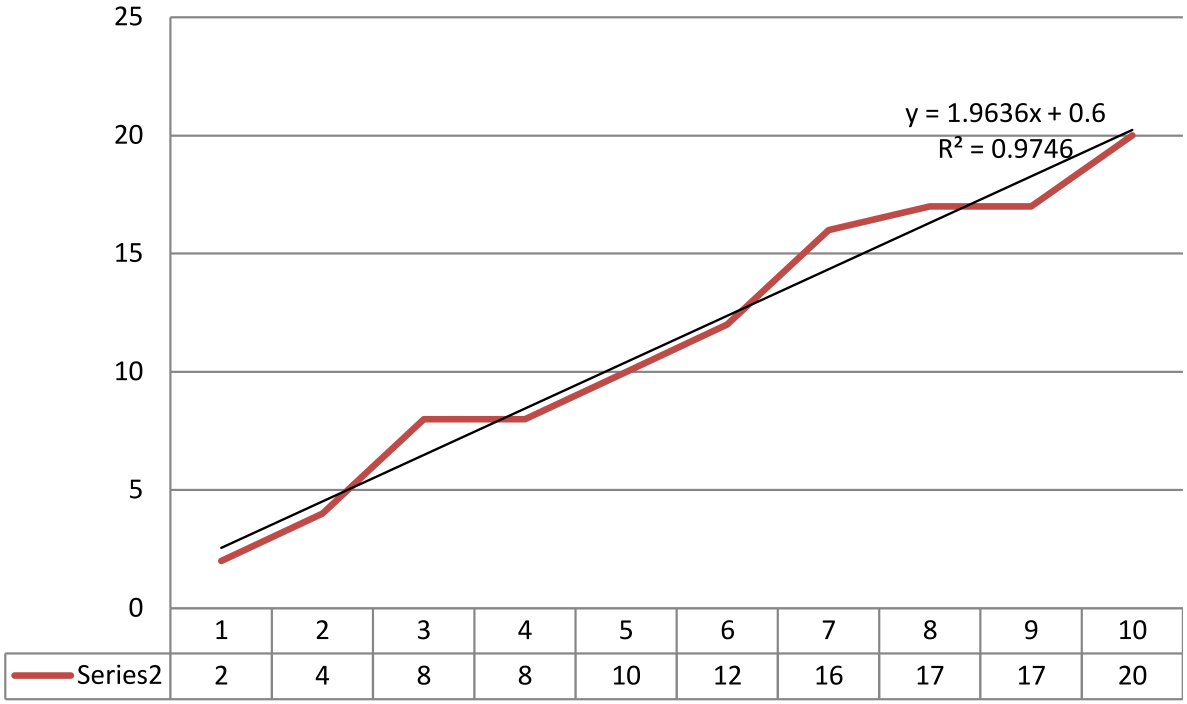


Y
$SSE = \sum(Y_i - \hat{Y}_i)^2$
$SST = \sum(Y_i - \bar{Y})^2$
$SSR = \sum(\hat{Y}_i - \bar{Y})^2$
$\bar{Y}$
X
X_i

© 2003 Prentice-Hall, Inc.

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

It is clear that 0<= R^2 <= 1. When SSR is closed to SST, R^2 will be close to 1. This means that regression explains most of the variability in Y and the fitted model is good. When SSE is closed to SST R^2 is closed to 0. This means the fitted model is not good.



	1	2	3	4	5	6	7	8	9	10
Series2	2	4	8	8	10	12	16	17	17	20

Add Trendeline Tool
to show R^2 in MSExcel

Multiple Linear Regression Model

- Sometimes, the dependent variable Y may depend on more than one independent variable (X)
  - The salary of a person may depend on education and experience.
  - Cost of an item depends on labor cost, electricity cost, and raw material cost.
  - Response variable (Y) depends on K explanatory variables denoted as X_1, X_2, X_3, …….X_k and the linear relationship of Y as a function of X_1, X_2, X_3, …….X_k can be written as

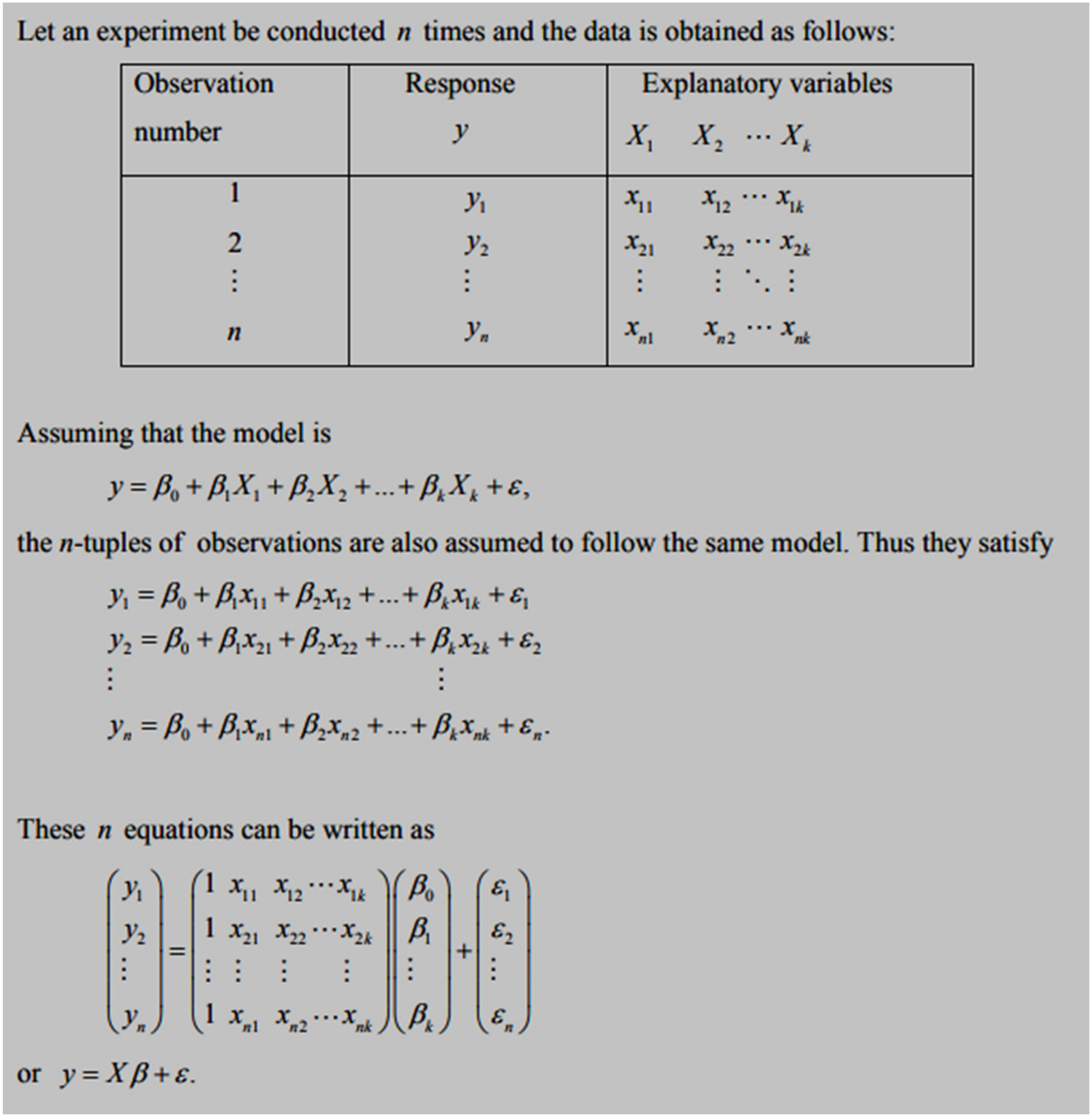
$$\mathbf{Y = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k + e}$$

A is the intercept or constant term.

B_1, B_2, B_3…B_k are called as regression slopes or regression coefficients.

Need to estimate A, B_1, B_2, B_3…B_k using least square method

Multiple regression in linear algebra notation



Let an experiment be conducted n times and the data is obtained as follows:

Observation number	Response y	Explanatory variables $X_1 \quad X_2 \cdots X_k$
1	y_1	$x_{11} \quad x_{12} \cdots x_{1k}$
2	y_2	$x_{21} \quad x_{22} \cdots x_{2k}$
$\vdots$	$\vdots$	$\vdots \quad \vdots \ddots \vdots$
n	y_n	$x_{n1} \quad x_{n2} \cdots x_{nk}$

Assuming that the model is

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon,$$

the n-tuples of observations are also assumed to follow the same model. Thus they satisfy

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + ... + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + ... + \beta_k x_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + ... + \beta_k x_{nk} + \varepsilon_n. \end{aligned}$$

These n equations can be written as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} \cdots x_{1k} \\ 1 & x_{21} & x_{22} \cdots x_{2k} \\ \vdots & \vdots & \vdots \quad\quad \vdots \\ 1 & x_{n1} & x_{n2} \cdots x_{nk} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

or $y = X\beta + \varepsilon.$

Least Squares Estimation

Multiple regression in linear algebra notation

- We can pack all response values for all observations into a n-dimensional vector called the response vector:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y1} \\ \mathbf{Y2} \\ \mathbf{Y3} \\ \cdots \\ \mathbf{Yn} \end{pmatrix}$$

- We can pack all predictors into **a n × p + 1 matrix** called the **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & X_{13} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & X_{23} & \cdots & X_{2p} \\ 1 & X_{31} & X_{32} & X_{33} & \cdots & X_{3p} \\ & \cdots & \cdots & \cdots & \cdots & \\ 1 & X_{n1} & X_{n2} & X_{n3} & \cdots & X_{np} \end{pmatrix}$$

- Note the initial column of 1’s. The reason for this will become clear shortly.

Least Squares Estimation

- We can pack the intercepts and slopes into a p+1-dimensional vector called the slope vector, denoted B:

$$\mathbf{B} = \begin{pmatrix} \mathbf{A} \\ \mathbf{B1} \\ \mathbf{B2} \\ \mathbf{\dots} \\ \mathbf{Bn} \end{pmatrix}$$

- Finally, we can pack all the errors terms into a n-dimensional vector called the error vector:

$$\in = \begin{pmatrix} \in \mathbf{1} \\ \in \mathbf{2} \\ \in \mathbf{3} \\ \mathbf{\dots} \\ \in \mathbf{n} \end{pmatrix}$$

- Using linear algebra notation, the model

$$\mathbf{Y_i = A + B_1 X_{i,1} + B_2 X_{i,2} + \cdots + B_p X_{i,p}} + \in_{\mathbf{i}}$$

Least Squares Estimation

- Previous model can compactly written:

$$\mathbf{Y} \quad = \mathbf{A} + \mathbf{BX} + \in$$

- where BX is the matrix-vector product.
- In order to estimate B, we take a least squares approach that is analogous to what we did in the simple linear regression case. That is, we want to minimize

$$\sum_{\mathbf{i}} \left(\mathbf{Y_i} - \mathbf{A} - \mathbf{B_1 X_{i,1}} - \mathbf{B_2 X_{i,2}} + \cdots - \mathbf{B_p X_{i,p}}\right)^{\mathbf{2}}$$

- over all possible values of the intercept and slopes.

Principle of Ordinary Least Squares (OLS)

- The object is to find a vector B that minimizes the sum of squared deviations of ϵ^2

$$S(B) = \sum_{i}^{n} \epsilon_i^2 \quad = \epsilon' \epsilon = (Y - XB)' \, (Y - XB)$$

$$= \sum_{i}^{n} Y'Y + B'X'BX - 2B'X'Y$$

- Differentiate S(B)with respect to B

$$\frac{\partial S(B)}{\partial B} = 2X'XB - 2X'Y = 0$$

$$\Rightarrow X'XB = X'Y \;\; \Rightarrow \; B = (X'X)^{-1}X'Y$$

X’X and (X’X) $^{-1}$ are p + 1 × p + 1 symmetric matrices.

X’Y is a p + 1 dimensional vector.

Principle of Ordinary Least Squares (OLS)

- If $\widehat{\mathrm{B}}$ is any estimator of B for the model Y = XB+ ε , then the fitted values are defined as $\widehat{\mathrm{Y}} = \widehat{\mathrm{B}}\mathrm{X}$ where $\widehat{\mathrm{B}}$ is any estimator of B .

The fitted values are

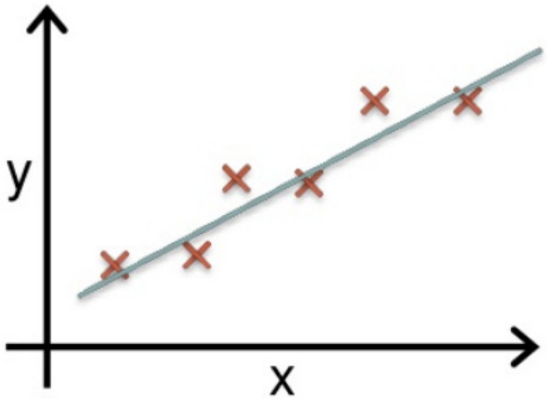
$$\widehat{\mathrm{Y}} = \widehat{\mathrm{B}}\mathrm{X} = \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}'\mathrm{Y}$$

and the residuals are

$$\mathrm{R} = \mathrm{Y} - \widehat{\mathrm{Y}} = (1 - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}')\mathrm{Y}$$

Polynomial Linear Regression

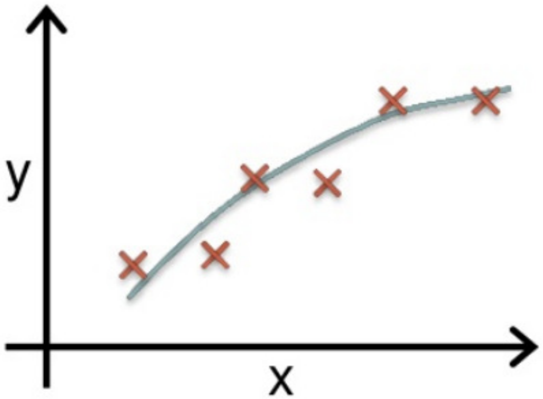
$$Y \quad = A + \underline{B_1}X_1 + \underline{B_2}X_1^2 + \cdots + \underline{B_n}X_1^n$$



Linear regression

$$Y \quad = A + B_1X_1$$

(Gradient and Interception)



Nonlinear regression

$$Y \quad = A + B_1X_1 + B_2X_1^2$$

Complicated Model (Polynomial)

Over fitting Problem

WHY Linear??

- Goal to find the coefficient

Special case of Multiple Linear Regression model

Polynomial Linear Regression

$$\mathbf{Y} \quad = \mathbf{A} + \mathbf{B}_1\mathbf{X}_1 + \mathbf{B}_2\mathbf{X}_1^2 + \cdots + \mathbf{B}_n\mathbf{X}_1^n$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \ldots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_1 & X_1{}^2 & X_1{}^3 & \ldots & X_1{}^k \\ 1 & X_2 & X_2{}^2 & X_2{}^3 & \ldots & X_2{}^k \\ 1 & X_3 & X_3{}^2 & X_3{}^3 & \ldots & X_3{}^k \\ & & \ldots\ \ldots & \ldots\ \ldots & & \\ 1 & X_n & X_n{}^2 & X_n{}^3 & \ldots & X_n{}^k \end{pmatrix} \begin{pmatrix} A \\ B1 \\ B2 \\ \ldots \\ Bn \end{pmatrix}$$

$$\mathbf{Y} = \mathbf{XW}$$

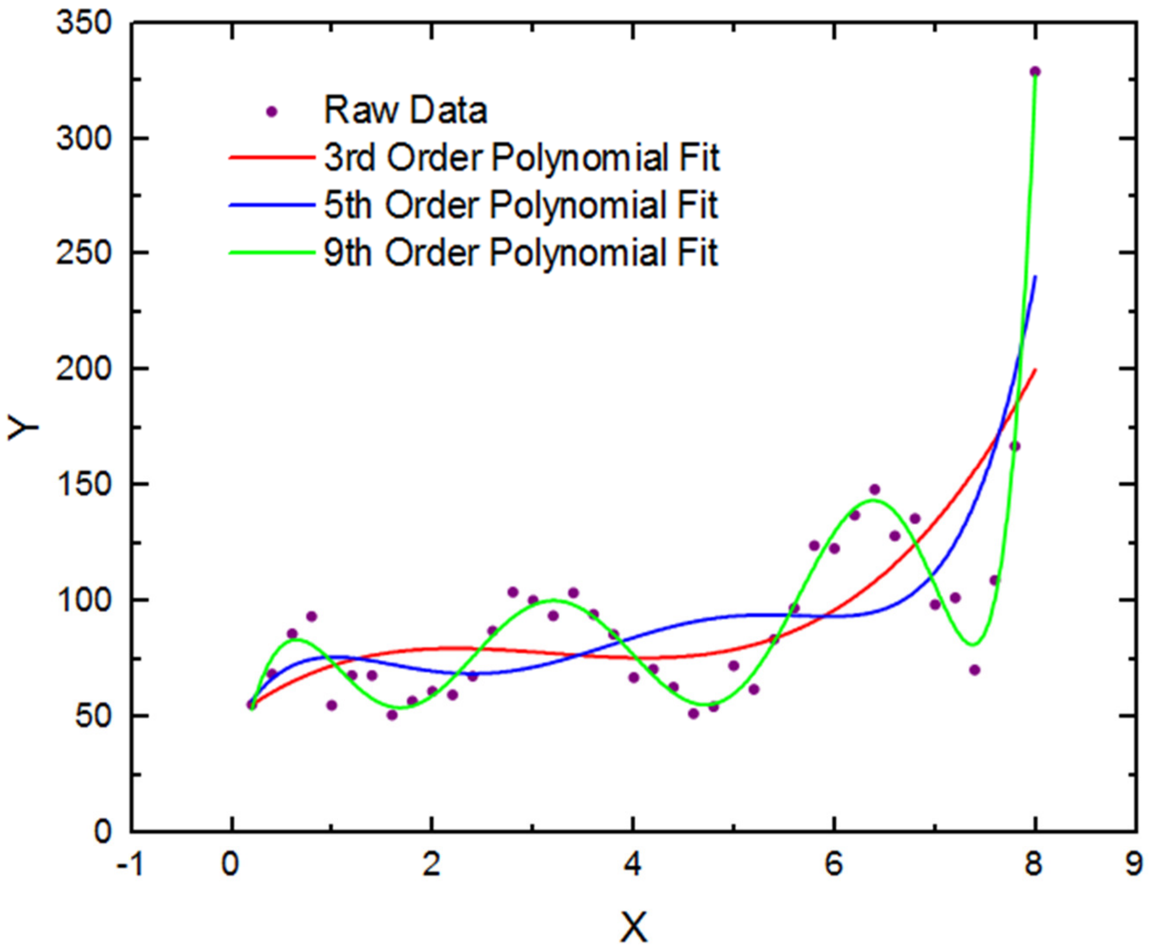
$$\Rightarrow \quad X^T X W = X^T Y$$

$$\Rightarrow \quad (X^T X)^{-1} X^T X W = (X^T X)^{-1} X^T Y$$

$$\Rightarrow \quad W = (X^T X)^{-1} X^T Y$$

Least Square Method

Over fitting and Polynomial Order



- Higher order polynomial reflects higher fitting accuracy
  - However, the overall result may not be the best always
  - Its overfitts and looses prediction power

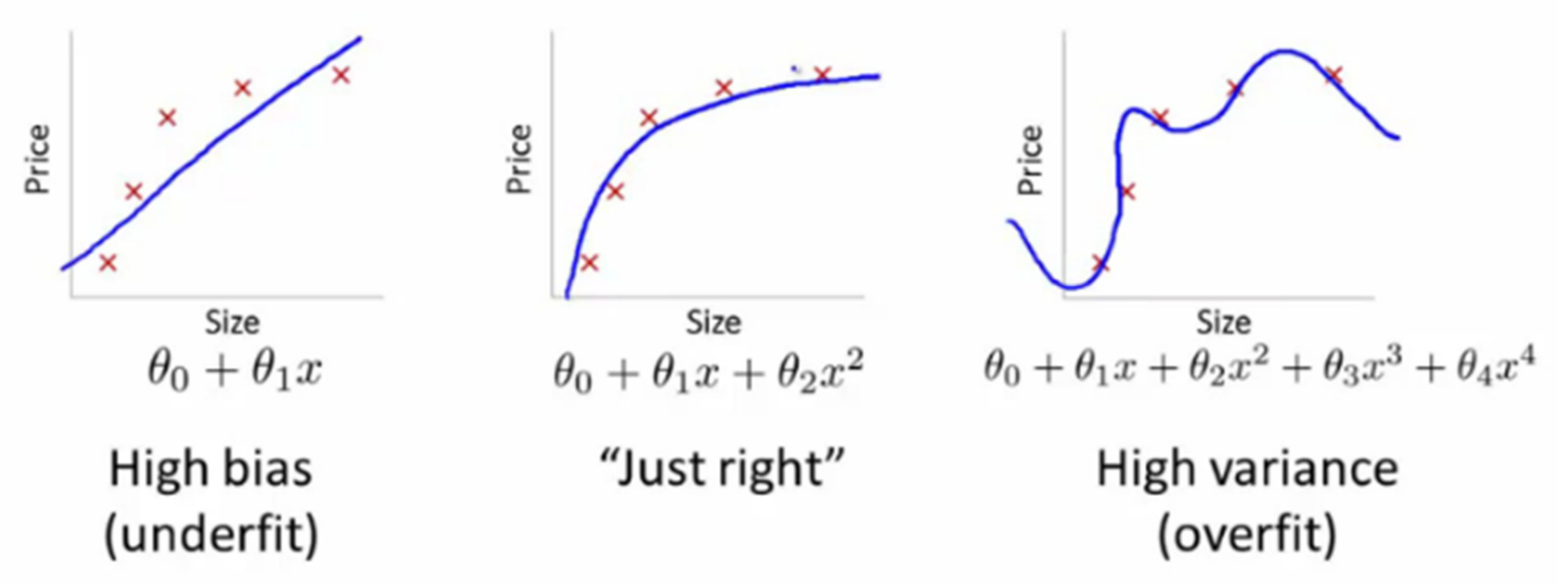
How does we decide most appropriate order of polynomial order?

Most appropriate polynomial order

- Choose the polynomial order based on
  - sum of the squares of the residuals, Sr is minimum?
  - we can always get Sr=0
    - if the polynomial order chosen is one less than the number of data points.
    - In fact, it would be an exact match.

Hands Note

Linear and Logistic Regression Problem



$\theta_0 + \theta_1 x$

High bias (underfit)

$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

High variance (overfit)

- Overfitting: If we have too many features, the learned model may fit the training set very well $S(A,B) = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - A - BX_i)^2$ ≈0 but fail to generalize the new examples.

Solution

1. **Reduce the number of features:** Factor analysis

2. **Regularization:**
   - Keep all the features, but reduce the magnitude/values of parameters θ_j
   - Works well when we have a lot of features, each of which contributes a bit to predicting Y.

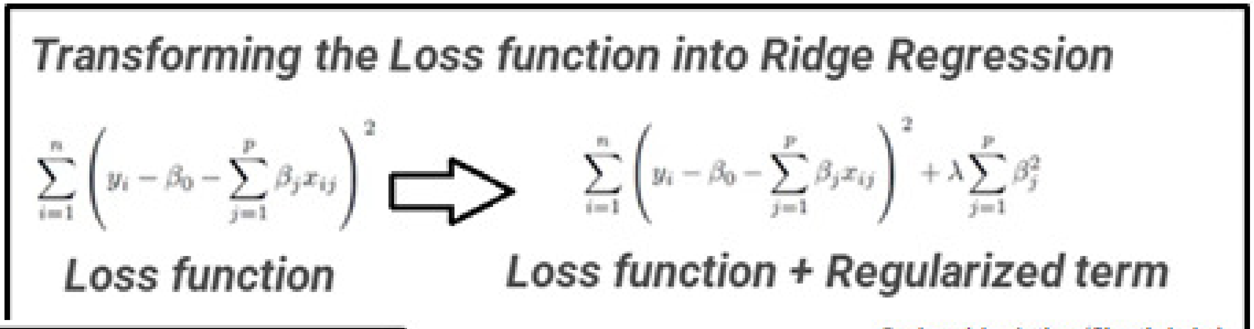
Regularization Techniques: Mitigate Multicollinearity

- Regularization methods like Ridge Regression (L2 regularization) and Lasso Regression (L1 regularization) can also help mitigate multicollinearity. These techniques add penalty terms to the regression model, reducing the impact of highly correlated features.

- **Loss= abs(Y_pred – Y_actual)**

- These update the general cost function by adding another term known as the regularization term.

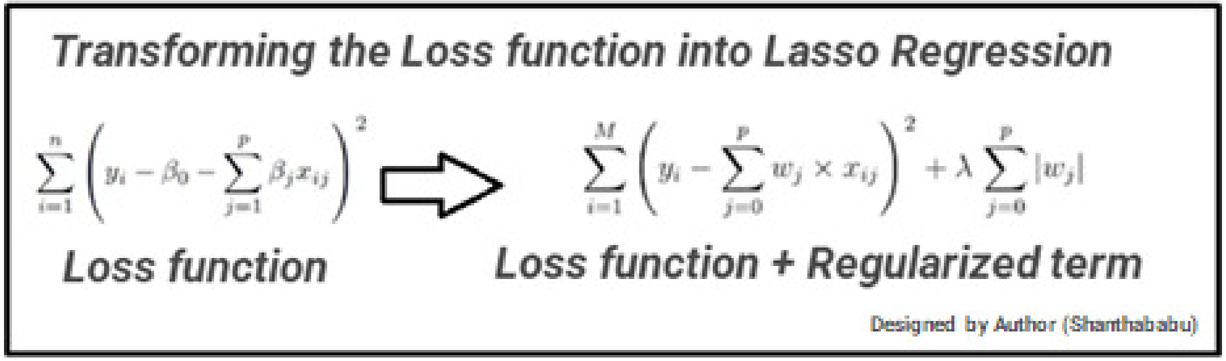
- ***Cost function = Loss + Regularization term***



Transforming the Loss function into Ridge Regression

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Loss function}} \Rightarrow \underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2}_{\text{Loss function + Regularized term}}$$

Designed by Author (Shanthababu)



Transforming the Loss function into Lasso Regression

$$\underbrace{\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2}_{\text{Loss function}} \Rightarrow \underbrace{\sum_{i=1}^{M}\left(y_i-\sum_{j=0}^{p}w_j\times x_{ij}\right)^2+\lambda\sum_{j=0}^{p}|w_j|}_{\text{Loss function + Regularized term}}$$

Designed by Author (Shanthababu)