ChE 3240

MATLAB

Statistics: Regression Analysis

Lecture-02

Least Squares of Linear Regression

For points $(x_1,y_1),(x2,y2)$, ……., (xn,yn) the least square regression can be given by-

$$f(x) = b + mx$$

$$\text{Error, } ei = yi - f(xi)$$

$$\text{Sum of squared error, SSE} = \sum(yi - f(xi))^2$$

Therefore,

$$y1 = (b + mx1) + e1$$

$$y2 = (b + mx2) + e2$$

•

•

$$yn = (b + mx_n) + en$$

Now let's set up an matrix equation. Let

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix} \qquad X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \cdots & \cdots \\ 1 & x_n \end{bmatrix} \qquad A = \begin{bmatrix} b \\ m \end{bmatrix} \qquad E = \begin{bmatrix} e_1 \\ e_2 \\ \cdots \\ e_n \end{bmatrix}$$

matrix equation: $\boldsymbol{Y = XA + E}$.

We now just need to solve this for **A**.

The solution to least squares regression equation $\boldsymbol{Y = XA + E}$ is:

$$A = (X^T X)^{-1} X^T Y$$

The sum of the squared errors is:

$$SSE = E^T E$$

$$X^TX = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & x_n \\ 1 & \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}$$

$$X^TY = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}$$

Quadratic Regression, in General

- In general, the design matrix and the normal equations for a ***quadratic model*** based on data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, are:

Design Matrix

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

Normal Equations

$$X^T X\mathbf{b} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \end{bmatrix} \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ x_1^2 & x_2^2 & \cdots & x_n^2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = X^T\mathbf{y}$$

Normal Equations

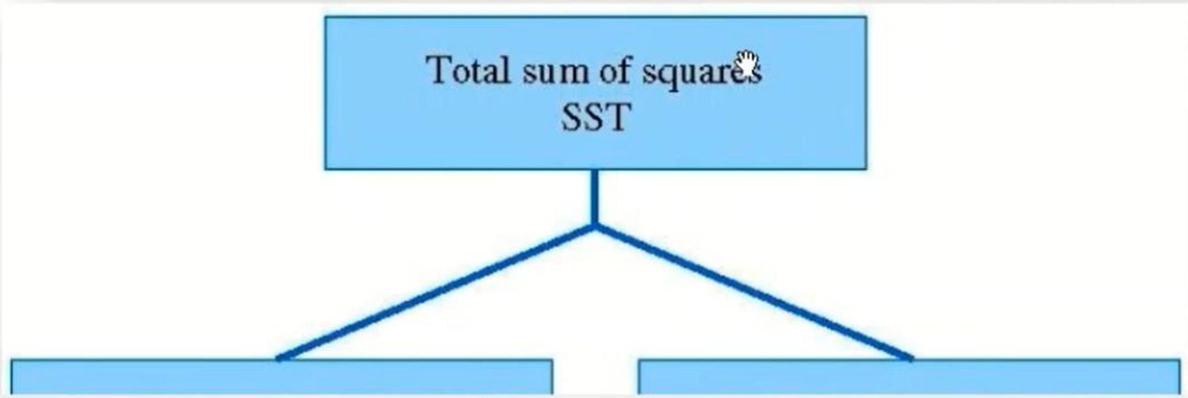
$$X^T X\mathbf{b} = \begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix} = X^T\mathbf{y}$$

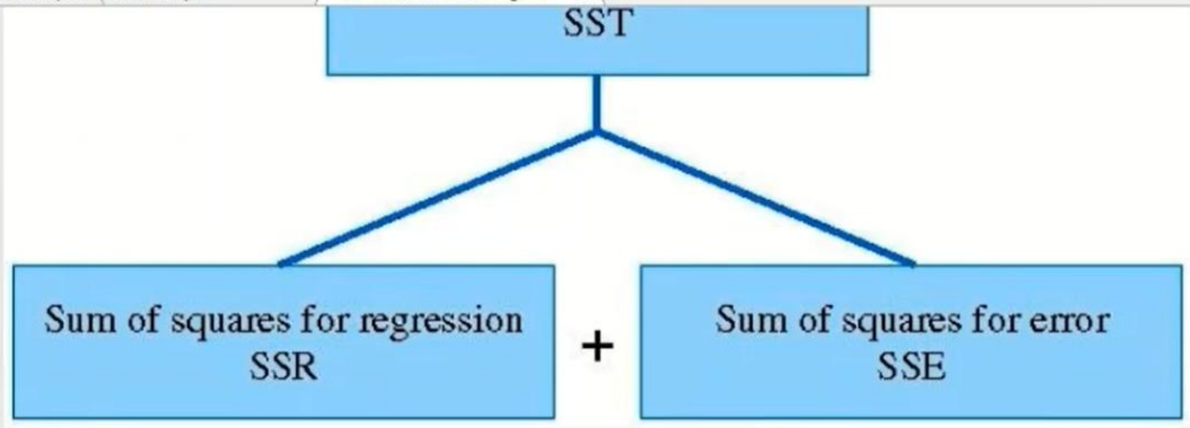
Solution Vector

$$\mathbf{b} = \left(X^T X\right)^{-1}\left(X^T\mathbf{y}\right)$$

Goodness of Fit in Regression and Coefficient of Determination:

$Y_i - \overline{Y}$ is called the total deviation and corresponding $\sum(Y_i - \overline{Y})^2$ is called total sum of squares (SST), $\widehat{Y}_i - \overline{Y}$ is called explained deviation and corresponding $\sum\left(\widehat{Y}_i - \overline{Y}\right)^2$ is called sum of squares for regression (SSR) and $Y_i - \widehat{Y}_i$ is called unexplained deviation and corresponding $\sum\left(Y_i - \widehat{Y}_i\right)^2$ is called sum of squares for error (SSE). The relation among SSE, SST, SSR is





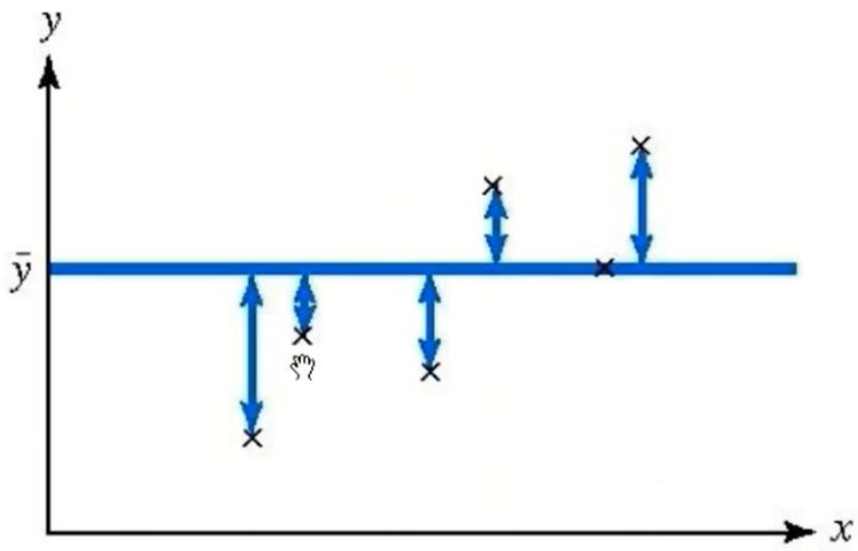
$$SST = SSE + SSR$$

Symbolically

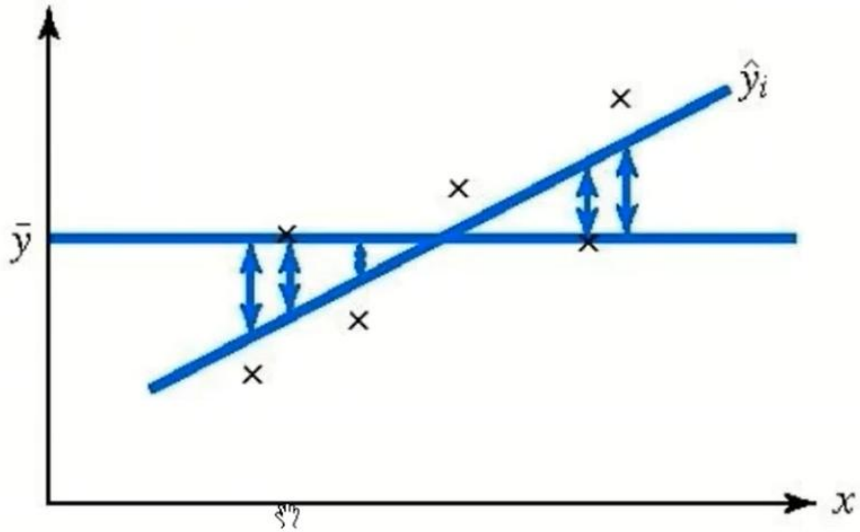
$$\sum(Y_i - \bar{Y})^2 = \sum(Y_i - \hat{Y}_i)^2 + \sum(\hat{Y}_i - \bar{Y})^2$$

In deterministic relationship SSE = 0 i.e. for a perfect fitting estimation line SST = SSR and hence SSR/SST=1. For the worst case of data SSR = 0 i.e. SSE = SST and hence SSR/SST=0

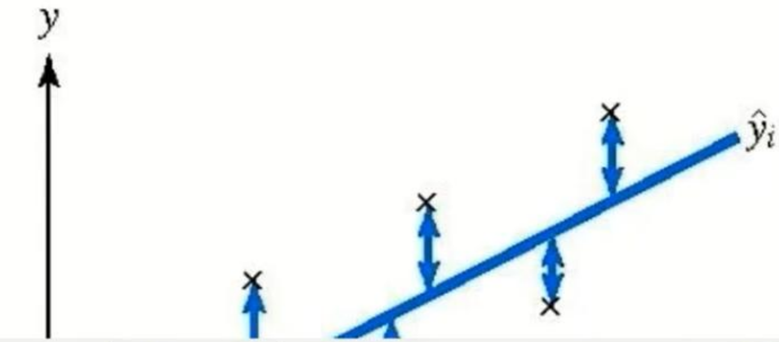
$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$



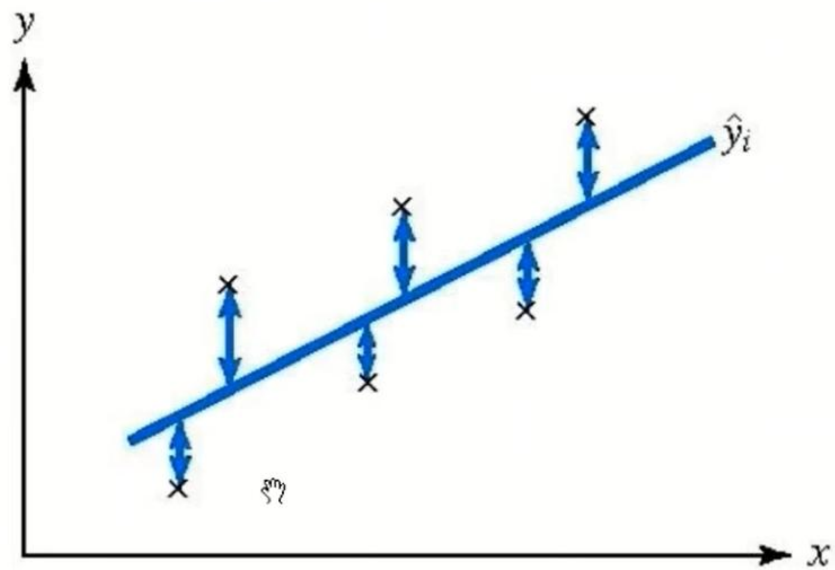
$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$



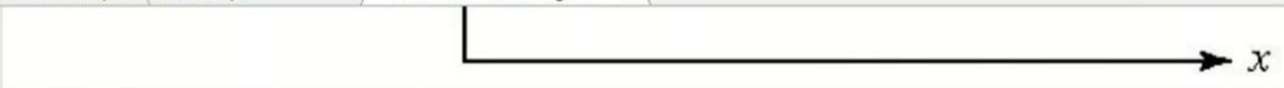
$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$



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



So the ratio SSR/SST evaluate how good the estimated regression line is, values of this ratio closer to 1 would imply better fitting estimated line. Thus the ratio SSR/SST is known as the coefficient of determination.



So the ratio SSR/SST evaluate how good the estimated regression line is, values of this ratio closer to 1 would imply better fitting estimated line. Thus the ratio SSR/SST is known as the coefficient of determination.

$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

$$= 1 - \frac{SSE/n}{SST/n} = var$$

r^2 is a non-negative value and it's limit are $0 \le r^2 \le 1$. Verbally, r^2 measures the percentage of the total variation in the dependent variable explained by the regression model.