Multiple Linear Regression

I- Definition

→ Multiple linear regression is a statistical method that extends simple linear regression to handle multiple independent variables. It models the relationship between a dependent variable (Y) and two or more independent variables ($X_1, X_2, \ldots, X_p$)

II- Regression Hyperplane

dependent variable

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

coefficients for the independent variables

intercept

error

→ The goal here is to estimate the best hyperplane using least square method. (minimize $\sum e_i^2$)

→ Finding coefficients $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$

→ We simplifies the representation of the multiple regression model by using matrix notation

→ The multiple regression model can be expressed in matrix as follows:

→ $Y = X\beta + \varepsilon$

vector of coefficients (including the intercept)

vector of errors

vector of observed values for D.V.

matrix of observed values for all I.V. (including a column of ones for the intercept)

→ The matrix representation would be:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

→ The vector of coefficients β_j $(j: 0 \to p)$ can be estimated using the following equation

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

(X^T: transpose; y: vector of observed values for the D.V.)

→ **SST / SSR / SSE**

Like in simple linear regression we have:

SSR (Sum of Squares due regression)

$$= \sum (\hat{y}_i - \bar{y})^2$$

SSE (Sum of Squares due errors)

$$= \sum (y_i - \hat{y}_i)^2$$

SST (Total Sum of Squares)

$$= SSR + SSE$$

$$= \sum (y_i - \bar{y}_i)^2$$

III- Regression Result

→ Error variance Estimator : $S^2 = \frac{SSE}{df} = \frac{SSE}{n-p-1}$ (nb of I.V.)

→ Standard error of estimate : $S = \sqrt{\frac{SSE}{df}} = \sqrt{\frac{SSE}{n-p-1}}$

→ Model Significance Test

↳ Step1 : Hypothesis

$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$

$H_1 : \exists j \in \{1, \dots, p\} / \beta_j \neq 0$

↳ Step2 : Test Statistic :

we use F-test

$$F = \frac{SSR / p}{SSE / (n-p-1)}$$

↳ Step 3 : Decision Rule

if $F > F_{critical}$ => reject H_0

↳ ANOVA Table :

Source of Variation	Sum of Squares	Degrees of freedom	Means of Squares	F
Regression	SSR	p	MSR = SSR / p	MSR / MSE
Residuals	SSE	n - (p+1)	MSE = SSE / n-p-1	
Total	SST	n - 1		

Significance test of a parameter β_j

- The objective is to test the absence of a linear connection between X_j and Y. So we must test the nullity of β_j for $j = 1 \dots p$

- **Step1: Hypothesis**

  $H_0 : \beta_j = 0$

  $H_1 : \beta_j \neq 0$

- **Step2 : Test Statistic**

  $$T = \frac{\hat{\beta}_j - \beta_j}{S\hat{\beta}_j} \quad ; \quad S\hat{\beta}_j = \frac{S}{\sqrt{\sum (x_i - \bar{x})^2}}$$

- **Step3 : Critical value**

  t_table : $df = n - p - 1$

- **Decision Rule**

  if $|t| >$ t_critical $\Rightarrow$ reject H_0

Confidence Interval for β_j

$$CI(\beta_j) = [\hat{\beta}_j \pm t\_critical \times S\hat{\beta}_j$$

- we reject H_0 if 0 does not belong to this interval