LECTURE - (31)

Agenda:

(1) The linear model
(2) The method of least squares

The linear model is the most popular statistical model because of its simplicity and ease of interpreatation. Let us begin with the simplest linear model. Suppose we have a characteristic Y that we want to study and understand, and we have information about another characteristic X. More precisely, we want to understand how Y is related to X. For example, Y could be the height and X could be the weight, or Y could be the stopping distance of an automobile and X could be its speed. The LINEAR MODEL is a probabilistic model, which represents the relationship between Y and X as

$$\underbrace{Y}_{\substack{\text{Dependent variable} \\ \text{or Variable of Interest}}} = \beta_0 + \beta_1 \underbrace{X}_{\substack{\text{Independent variable} \\ \text{or covariate}}} + \varepsilon \longrightarrow \text{Error (Random)}$$

β_0 and β_1 are the unknown parameters in this model, which we wish to estimate from observed data.

Typically, the observed data will be a collection of n sets of observations of the two characteristics, namely, $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. For example we will randomly sample n individuals from the population and note down their weight and height. Using the linear model, we have

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \quad i = 1, 2, \ldots, n$$

where we assume that the errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent and identically distributed with mean 0 and unknown variance σ^2.

Today, we will tackle the problem of finding estimates for β_0 and β_1 based on the observed data. Consider the quantity

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

$S(\beta_0, \beta_1)$ is the "error sum of squares". Note that $\beta_0 + \beta_1 X_i$ is the expected value of Y_i, since $E[Y_i] = E[\beta_0 + \beta_1 X_i + \varepsilon_i] = \beta_0 + \beta_1 X_i$. Hence $(Y_i - \beta_0 - \beta_1 X_i)^2$ is the squared deviation between the observed and expected values of Y_i. A very intuitive criterion would be to find values of β_0 and β_1, which minimize the "error sum of squares".

Hence,

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{(\beta_0,\beta_1)} S(\beta_0,\beta_1),$$

i.e., $\hat{\beta}_0$ and $\hat{\beta}_1$ are the values of β_0, β_1 that minimize $S(\beta_0,\beta_1)$. Using calculus, we can solve for $\hat{\beta}_0$ and $\hat{\beta}_1$ by setting both the derivatives of $S(\beta_0,\beta_1)$ w.r.t. β_0 and β_1 to be zero.

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = 0 \iff \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

$$\iff -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\iff n\bar{Y} - n\beta_0 - n\beta_1 \bar{X} = 0$$

$$\iff \bar{Y} = \beta_0 + \beta_1 \bar{X}.$$

$$\frac{\partial}{\partial \beta_1} S(\beta_0,\beta_1) = 0 \iff \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = 0$$

$$\iff -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\iff \sum_{i=1}^{n} X_i Y_i - n\bar{X}\beta_0 - \beta_1 \sum_{i=1}^{n} X_i^2 = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} x_i y_i - n\bar{x}(\bar{y} - \beta_1 \bar{x}) - \beta_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$(\because \quad \bar{Y} = \beta_0 + \beta_1 \bar{X})$$

$$\Leftrightarrow \quad \beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

(Why?!)

Hence, the LEAST SQUARES ESTIMATORS for (β_0, β_1) are given by

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum (x_i - \bar{x})^2} \quad \begin{matrix} \rightarrow \text{Sample covariance of } x \text{ and } y \\ \rightarrow \text{Sample variance of } x \end{matrix}$$

and $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.