LECTURE - 35

<u>Agenda:</u>

1. The multivariate linear model
2. Least squares estimates for the multivariate linear model

We now consider a more general situation where we want to study the relationship between a characteristic Y and another group of characteristics $X_1, X_2, ---, X_p$ associated with the population/phenomenon under study. For example, Y is the income and we would like to understand its relationship with other characteristics like level of education, field of study, race, gender etc.

<u>Data:</u> We have n independently chosen vectors of observation, namely

$$(Y_1, X_{11}, X_{12}, ---, X_{1p})$$
$$(Y_2, X_{21}, X_{22}, ---, X_{2p})$$
$$\vdots$$
$$(Y_n, X_{n1}, X_{n2}, ---, X_{np})$$

Here, x_{ij} denotes the value of the characteristic X_j for the i^{th} individual / object.

Model: The linear model in this situation says that the value of the Y characteristic is a linear combination of the characteristics $X_1, X_2, \ldots, X_p$ plus some random error, i.e.,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for every $i = 1, 2, \ldots, n$,
where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent errors with a common distribution and $E[\varepsilon] = 0$, $V(\varepsilon) = \sigma^2$.

Estimation: We need to find estimates for the parameters $\beta_0, \beta_1, \ldots, \beta_p$ and σ^2. We will look at the estimate for σ^2 next time, but let us first derive the estimates for $\beta_0, \beta_1, \ldots, \beta_p$. Let us introduce some notation to facilitate our derivation. Let

$$\underline{\beta} \hat{=} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad X \hat{=} \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

and

$$\underline{Y} \triangleq \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

We want to estimate the parameter vector $\underline{\beta}$. Consider the error sum of squares

$$S(\underline{\beta}) = \sum_{i=1}^{n} \Big(Y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \Big)^2$$

Observed Y value for i^{th} individual / object

Expected Y value under the model for the i^{th} individual / object.

It is natural to choose parameter estimates which minimize the error sum of squares. Hence, the "least squares" estimate $\hat{\underline{\beta}}$ of the parameter vector $\underline{\beta}$ is defined by

$$\hat{\underline{\beta}} = \arg\min_{\underline{\beta}} S(\underline{\beta}).$$

Note that

$$
\begin{aligned}
S(\underline{\beta}) &= (\underline{Y} - X\underline{\beta})^T(\underline{Y} - X\underline{\beta}) \\
&= \underline{Y}^T\underline{Y} - (X\underline{\beta})^T\underline{Y} - \underline{Y}^T X\underline{\beta} + (X\underline{\beta})^T X\underline{\beta} \\
&= \underline{Y}^T\underline{Y} - 2\underline{Y}^T X\underline{\beta} + \underline{\beta}^T X^T X\underline{\beta}.
\end{aligned}
$$

($\because$ Using $(AB)^T = B^T A^T$ and $\underline{a}^T\underline{b} = \underline{b}^T\underline{a}$)

Hence, any minimizer of $S(\underline{\beta})$ satisfies the gradient condition, i.e.,

$$\nabla_{\underline{\beta}} S(\underline{\beta}) = 0.$$

$$\Leftrightarrow \nabla_{\underline{\beta}}\left\{ \underline{Y}^T\underline{Y} - 2\underline{Y}^T X\underline{\beta} + \underline{\beta}^T X^T X \underline{\beta} \right\} = 0$$

$$\Leftrightarrow \nabla_{\underline{\beta}}(\underline{Y}^T\underline{Y}) - \nabla_{\underline{\beta}}(2\underline{Y}^T X\underline{\beta}) + \nabla_{\underline{\beta}}(\underline{\beta}^T X^T X\underline{\beta}) = 0$$

$$\Leftrightarrow 0 - 2X^T\underline{Y} + 2X^T X\underline{\beta} = 0$$

($\because$ $\nabla_{\underline{\beta}}\, \underline{a}^T\underline{\beta} = \underline{a}$ and $\nabla_{\underline{\beta}}\, \underline{\beta}^T A\underline{\beta} = 2A\underline{\beta}$)

$$\Leftrightarrow X^T X\underline{\beta} = X^T\underline{Y}$$ $\longrightarrow$ These are often known as the "normal equations".

$$\Leftrightarrow \underline{\beta} = (X^T X)^{-1} X^T \underline{Y}$$

We are inherently assuming that $(X^T X)^{-1}$ exists, i.e; $\text{Rank}(X) = p+1$. This is not a very stringent condition. One can look at the Hessian of $S(\underset{\sim}{\beta})$ at $\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{Y}$ and verify that

$$\boxed{\hat{\underset{\sim}{\beta}} = (X^T X)^{-1} X^T \underset{\sim}{Y}}$$

is indeed the minimizer of $S(\underset{\sim}{\beta})$.