Analysis of Variance and Design of Experiment-I

MODULE – II

LECTURE - 4

GENERAL LINEAR HYPOTHESIS
AND ANALYSIS OF VARIANCE

Dr. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur
Regression model for the general linear hypothesis

Let \( Y_1, Y_2, \ldots, Y_n \) be a sequence of \( n \) independent random variables associated with responses. Then we can write it as

\[
E(Y_i) = \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, p
\]

\[
Var(Y_i) = \sigma^2.
\]

This is the linear model in the expectation form where \( \beta_1, \beta_2, \ldots, \beta_p \) are the unknown parameters and \( x_{ij} \)'s are the known values of independent covariates \( X_1, X_2, \ldots, X_p \).

Alternatively, the linear model can be expressed as

\[
Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, p
\]

where \( \varepsilon_i \)'s are identically and independently distributed random error component with mean 0 and variance \( \sigma^2 \), i.e.,

\[
E(\varepsilon_i) = 0, \quad Var(\varepsilon_i) = \sigma^2 \quad \text{and} \quad Cov(\varepsilon_i, \varepsilon_j) = 0 (i \neq j).
\]

In matrix notations, the linear model can be expressed as

\[
Y = X \beta + \varepsilon
\]

where

- \( Y = (Y_1, Y_2, \ldots, Y_n)' \) is \( n \times 1 \) vector of observations on response variable,

- the matrix \( X = \begin{pmatrix}
X_{11} & X_{12} & \cdots & X_{1p} \\
X_{21} & X_{22} & \cdots & X_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
x_{n1} & x_{n2} & \cdots & x_{np}
\end{pmatrix} \) is \( n \times p \) matrix of \( n \) observations on \( p \) independent covariates \( X_1, X_2, \ldots, X_p \),
• \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)' \) is a \( p \times 1 \) vector of unknown regression parameters (or regression coefficients) \( \beta_1, \beta_2, \ldots, \beta_p \) associated with \( X_1, X_2, \ldots, X_p \), respectively and

• \( \varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)' \) is a \( n \times 1 \) vector of random errors or disturbances.

• We assume that \( E(\varepsilon) = 0 \), covariance matrix \( V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_p \), \( \text{rank}(X) = p \).

In the context of analysis of variance and design of experiments,

• the matrix \( X \) is termed as **design matrix**;

• unknown \( \beta_1, \beta_2, \ldots, \beta_p \) are termed as **effects**;

• the covariates \( X_1, X_2, \ldots, X_p \) are **counter variables** or **indicator variables** where \( x_{ij} \) counts the number of times the effect \( \beta_j \) occurs in the \( i^{th} \) observation \( x_i \).

• \( x_{ij} \) mostly takes the values 1 or 0 but not always.

• The value \( x_{ij} = 1 \) indicates the presence of effect \( \beta_j \) in \( X_i \) and \( x_{ij} = 0 \) indicates the absence of effect \( \beta_j \) in \( X_i \).

Note that in the **linear regression model**, the covariates are usually continuous variables.

When some of the covariates are counter variables and rest are continuous variables, then the model is called as **mixed model** and is used in the analysis of covariance.
The same linear model is used in the linear regression analysis as well as in the analysis of variance. So it is important to understand the role of linear model in the context of linear regression analysis and analysis of variance.

Consider the multiple linear model

\[ Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + \ldots + X_p\beta_p + \varepsilon. \]

In the case of analysis of variance model,

- the one-way classification considers only one covariate,
- two-way classification model considers two covariates,
- three-way classification model considers three covariates and so on.

If \( \beta, \gamma \) and \( \delta \) denote the effects associated with the covariates \( X, Z \) and \( W \) which are counter variables, then in

One-way model: \( Y = \alpha + X\beta + \varepsilon \)

Two-way model: \( Y = \alpha + X\beta + Z\gamma + \varepsilon \)

Three-way model: \( Y = \alpha + X\beta + Z\gamma + W\delta + \varepsilon \) and so on.

Consider an example of agricultural yield. The study variable denotes the yield which depends on various covariates \( X_1, X_2, \ldots, X_p \). In case of regression analysis, the covariates \( X_1, X_2, \ldots, X_p \) are the different variables like temperature, quantity of fertilizer, amount of irrigation, etc.
Now consider the case of one way model and try to understand its interpretation in terms of multiple regression model. The covariate $X$ is now measured at different levels, e.g., if $X$ is the quantity of fertilizer then suppose there are $p$ possible values, say 1 Kg., 2 Kg., ..., $p$ Kg. then $X_1, X_2, ..., X_p$ denotes these $p$ values in the following way.

The linear model now can be expressed as

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

by defining

$X_1 = \begin{cases} 1 & \text{if effect of 1 Kg. fertilizer is present} \\ 0 & \text{if effect of 1 Kg. fertilizer is absent} \end{cases}$

$X_2 = \begin{cases} 1 & \text{if effect of 2 Kg. fertilizer is present} \\ 0 & \text{if effect of 2 Kg. fertilizer is absent} \end{cases}$

$\vdots$

$X_p = \begin{cases} 1 & \text{if effect of } p \text{ Kg. fertilizer is present} \\ 0 & \text{if effect of } p \text{ Kg. fertilizer is absent.} \end{cases}$

If effect of 1 Kg. of fertilizer is present, then other effects will obviously be absent and the linear model is expressible as

$$Y = \beta_0 + \beta_1 (X_1 = 1) + \beta_2 (X_2 = 0) + \cdots + \beta_p (X_p = 0) + \varepsilon$$

$$= \beta_0 + \beta_1 + \varepsilon.$$

If effect of 2 Kg. of fertilizer is present then

$$Y = \beta_0 + \beta_1 (X_1 = 0) + \beta_2 (X_2 = 1) + \cdots + \beta_p (X_p = 0) + \varepsilon$$

$$= \beta_0 + \beta_2 + \varepsilon.$$
If effect of $p$ Kg. of fertilizer is present then

$$Y = \beta_0 + \beta_1 (X_1 = 0) + \beta_2 (X_2 = 0) + \ldots + \beta_p (X_p = 1) + \varepsilon$$

$$= \beta_0 + \beta_p + \varepsilon$$

and so on.

If the experiment with 1 Kg. of fertilizer is repeated $n_1$ number of times then $n_1$ observation on response variables are recorded which can be represented as

$$Y_{11} = \beta_0 + \beta_1.1 + \beta_2.0 + \ldots + \beta_p.0 + \varepsilon_{11}$$

$$Y_{12} = \beta_0 + \beta_1.1 + \beta_2.0 + \ldots + \beta_p.0 + \varepsilon_{12}$$

$$\vdots$$

$$Y_{1n_1} = \beta_0 + \beta_1.1 + \beta_2.0 + \ldots + \beta_p.0 + \varepsilon_{1n_1}.$$

If $X_2 = 1$ is repeated $n_2$ times, then on the same lines $n_2$ number of times then $n_1$ observation on response variables are recorded which can be represented as

$$Y_{21} = \beta_0 + \beta_1.0 + \beta_2.1 + \ldots + \beta_p.0 + \varepsilon_{21}$$

$$Y_{22} = \beta_0 + \beta_1.0 + \beta_2.1 + \ldots + \beta_p.0 + \varepsilon_{22}$$

$$\vdots$$

$$Y_{2n_2} = \beta_0 + \beta_1.0 + \beta_2.1 + \ldots + \beta_p.0 + \varepsilon_{2n_2}.$$
The experiment is continued and if $X_p = 1$ is repeated $n_p$ times, then on the same lines

$$
Y_{p1} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \ldots + \beta_p \cdot 1 + \varepsilon_{p1}
$$
$$
Y_{p2} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \ldots + \beta_p \cdot 1 + \varepsilon_{p2}
$$
$$
\vdots
$$
$$
Y_{pn_p} = \beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0 + \ldots + \beta_p \cdot 1 + \varepsilon_{pn_p}.
$$

All these $n_1, n_2, \ldots, n_p$ observations can be represented as

$$
\begin{pmatrix}
  y_{11} \\
  y_{12} \\
  \vdots \\
  y_{1n_1} \\
  y_{21} \\
  y_{22} \\
  \vdots \\
  y_{2n_2} \\
  \vdots \\
  y_{p1} \\
  y_{p2} \\
  \vdots \\
  y_{pn_p}
\end{pmatrix}
= \begin{pmatrix}
  1 & 1 & 0 & 0 & \ldots & 0 & 0 \\
  1 & 1 & 0 & 0 & \ldots & 0 & 0 \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  1 & 1 & 0 & 0 & \ldots & 0 & 0 \\
  1 & 0 & 1 & 0 & \ldots & 0 & 0 \\
  1 & 0 & 1 & 0 & \ldots & 0 & 0 \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  1 & 0 & 1 & 0 & \ldots & 0 & 0 \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  1 & 0 & 0 & 0 & \ldots & 0 & 1 \\
  1 & 0 & 0 & 0 & \ldots & 0 & 1 \\
  \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
  1 & 0 & 0 & 0 & \ldots & 0 & 1
\end{pmatrix}
\begin{pmatrix}
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_{n_1} \\
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_{n_2} \\
  \vdots \\
  \beta_0 \\
  \beta_1 \\
  \vdots \\
  \beta_{n_p}
\end{pmatrix}
+ \begin{pmatrix}
  \varepsilon_{11} \\
  \varepsilon_{12} \\
  \vdots \\
  \varepsilon_{1n_1} \\
  \varepsilon_{21} \\
  \varepsilon_{22} \\
  \vdots \\
  \varepsilon_{2n_2} \\
  \vdots \\
  \varepsilon_{p1} \\
  \varepsilon_{p2} \\
  \vdots \\
  \varepsilon_{pn_p}
\end{pmatrix}
$$

or

$$
Y = X \beta + \varepsilon.
$$
In the two way analysis of variance model, there are two covariates and the linear model is expressible as

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \gamma_1 Z_1 + \gamma_2 Z_2 + \ldots + \gamma_q Z_q + \epsilon \]

where \( X_1, X_2, \ldots, X_p \) denotes, e.g., the \( p \) levels of quantity of fertilizer, say 1 Kg., 2 Kg., \ldots, \( p \) Kg. and \( Z_1, Z_2, \ldots, Z_q \) denotes, e.g., the \( q \) levels of level of irrigation, say 10 Cms., 20 Cms., \ldots, 10q Cms. etc. The levels \( X_1, X_2, \ldots, X_p \), \( Z_1, Z_2, \ldots, Z_q \) are defined as counter variable indicating the presence or absence of the effect as in the earlier case. If the effect of \( X_1 \) and \( Z_1 \) are present, i.e., 1 Kg of fertilizer and 10 Cms. of irrigation is used then the linear model is written as

\[ Y = \beta_0 + \beta_1 1 + \beta_2 0 + \ldots + \beta_p 0 + \gamma_1 1 + \gamma_2 0 + \ldots + \gamma_p 0 + \epsilon = \beta_0 + \beta_1 + \gamma_1 + \epsilon. \]

If \( X_2 = 1 \) and \( Z_2 = 1 \) is used, then the model is \( Y = \beta_0 + \beta_2 + \gamma_2 + \epsilon. \)

The design matrix can be written accordingly as in the one way analysis of variance case.

In the three way analysis of variance model

\[ Y = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p + \gamma_1 Z_1 + \ldots + \gamma_q Z_q + \delta_1 W_1 + \ldots + \delta_r W_r + \epsilon. \]
• The regression parameters $\beta$'s can be fixed or random.

• If all $\beta$'s are unknown constants, they are called as **parameters** of the model and the model is called as a **fixed-effects model** or **model I**. The objective in this case is to make inferences about the parameters and the error variance $\sigma^2$.

• If for some $j$, $x_{ij} = 1$ for all $i = 1, 2, \ldots, n$ then $\beta_j$ is termed as **additive constant**. In this case, $\beta_j$ occurs with every observation and so it is also called as **general mean effect**.

• If all $\beta$'s are observable random variables except the additive constant, then the linear model is termed as **random-effects model**, **model II** or **variance components model**. The objective in this case is to make inferences about the variances of $\beta$'s, i.e., $\sigma_{\beta_1}^2, \sigma_{\beta_2}^2, \ldots, \sigma_{\beta_p}^2$ and error variance $\sigma^2$ and/or certain functions of them.

• If some parameters are fixed and some are random variables, then the model is called as **mixed-effects model** or **model III**. In mixed effect model, at least one $\beta_j$ is constant and at least one $\beta_j$ is random variable. The objective is to make inference about the fixed effect parameters, variance of random effects and error variance $\sigma^2$. 