# Analysis of Variance and Design of Experiments

## General Linear Hypothesis and Analysis of Variance

:::

## Lecture 4

## ANOVA models and Least Squares Estimation of Parameters


Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur


Slides can be downloaded from http://home.iitk.ac.in/~shalab/sp

## 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**

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

**where $\beta_1, \beta_2, \ldots, \beta_p$ are the unknown parameters and $x_{ij}$'s are the known values of independent covariates $X_1, X_2, \ldots, X_p$, $\varepsilon_i$'s are i.i.d. random error component with**

$$E(\varepsilon_i) = 0\ , \quad Var(\varepsilon_i) = \sigma^2, \quad Cov(\varepsilon_i, \varepsilon_j) = 0 (i \neq j).$$

## Relationship between the regression and ANOVA models:

**Now consider the case of one-way model and try to understand its interpretation in terms of the 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 $X_1, X_2, \ldots, X_p$ denotes these *p* values in the following way.**

# Relationship between the regression and ANOVA models:

**The linear model now can be expressed as**

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

**by defining**

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

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

$$\vdots$$

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

## Relationship between the regression and ANOVA models:

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

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

**If the effect of 2 Kg. of fertilizer is present then**

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

**If the 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.**

## Relationship between the regression and ANOVA models:

**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**

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

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

$$\vdots$$

$$Y_{1n_1} = \beta_0 + \beta_1.1 + \beta_2.0 + \ldots + \beta_p.0 + \varepsilon_{1n1}$$

## Relationship between the regression and ANOVA models:

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

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

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

$$\vdots$$

$$Y_{2n_2} = \beta_0 + \beta_1.0 + \beta_2.1 + \ldots + \beta_p.0 + \varepsilon_{2n_2}$$

## Relationship between the regression and ANOVA models:

**The experiment is continued and if $X_p = 1$ is repeated $n_p$ times, then on the same lines**

$$Y_{p1} = \beta_0 + \beta_1.0 + \beta_2.0 + \ldots + \beta_p.1 + \varepsilon_{P1}$$

$$Y_{p2} = \beta_0 + \beta_1.0 + \beta_2.0 + \ldots + \beta_p.1 + \varepsilon_{P2}$$

$$\vdots$$

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

# Relationship between the regression and ANOVA models:

**All these $n_1, n_2, .., n_p$ observations can be represented as**

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

**or**

$$Y = X\beta + \varepsilon.$$

## Relationship between the regression and ANOVA models:

**In the two-way analysis of variance model, there are two covariates and the linear model is expressible as**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \gamma_1 Z_1 + \gamma_2 Z_2 + \ldots + \gamma_q Z_q + \varepsilon$$

**where** $X_1, X_2, \ldots, X_p$ **denotes, e.g., the *p* levels of the quantity of fertilizer, say 1 Kg., 2 Kg.,...,*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.,...100 Cms. etc. The levels** $X_1, X_2, \ldots, X_p$ $Z_1, Z_2, \ldots, Z_q$ **are the counter variable indicating the presence or absence of the effect as in the earlier case.**

## Relationship between the regression and ANOVA models:

**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**

$$Y = \beta_0 + \beta_1.1 + \beta_2.0 + \ldots + \beta_p.0 + \gamma_1.1 + \gamma_2.0 + \ldots + \gamma_p.0 + \varepsilon$$
$$= \beta_0 + \beta_1 + \gamma_1 + \varepsilon.$$

**If $X_2 = 1$ and $Z_2 = 1$ is used, then the model is**

$$Y = \beta_0 + \beta_2 + \gamma_2 + \varepsilon.$$

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

**In the three-way analysis of variance model**

$$Y = \alpha + \beta_1 X_1 + \ldots + \beta_p X_p + \gamma_1 Z_1 + \ldots + \gamma_q Z_q + \delta_1 W_1 + \ldots + \delta_r W_r + \varepsilon$$

## Relationship between the regression and ANOVA models:

**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 effect 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 an additive constant. In this case, $\beta_j$ occurs with every observation and so it is also called a general mean effect.**

## Relationship between the regression and ANOVA models:

- If all $\beta's$ are observable random variables except the additive constant, then the linear model is termed as random effect model, model II or variance components model.

The objective, in this case, is to make inferences about the variances of $\beta's$, i.e., $\sigma^2_{\beta_1}, \sigma^2_{\beta_2}, \ldots, \sigma^2_{\beta_p}$ and error variance and/or certain functions of them.

# Relationship between the regression and ANOVA models:

- **If some parameters are fixed and some are random variables, then the model is called a mixed effect model or model III. In the mixed effect model, at least one $\beta_j$ is constant and at least one $\beta_j$ is a random variable.**

**The objective is to make inference about the fixed effect parameters, variance of random effects and error variance $\sigma^2$.**

## Analysis of variance:

**Analysis of variance is a body of statistical methods of analyzing the measurements assumed to be structured as**

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i,\ i = 1, 2, \ldots, n$$

**where $x_{ij}$ are integers, generally 0 or 1 indicating usually the absence or presence of effects $\beta_j$; and $\varepsilon_i$'s are assumed to be identically and independently distributed with mean 0 and variance $\sigma^2$.**

**It may be noted that the $\varepsilon_i$'s can be assumed additionally to follow a normal distribution $N(0, \sigma^2)$.**

## Analysis of variance:

**It is needed for the maximum likelihood estimation of parameters from the beginning of the analysis, but in the least-squares estimation, it is needed only when conducting the tests of hypothesis and the confidence interval estimation of parameters.**

**The least-squares method does not require any knowledge of distribution like normal up to the stage of estimation of parameters.**

**We need some basic concepts to develop tools.**

## Least squares estimate of $\beta$ :

**Let $y_1, y_2, \ldots, y_n$ be a sample of observations on $Y_1, Y_2, \ldots, Y_n$.**

**The least-squares estimate of $\beta$ is the values $\hat{\beta}$ of $\beta$ for which the sum of squares due to errors, i.e.,**

$$S^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \varepsilon' \varepsilon = (y - X\beta)'(y - X\beta)$$
$$= y'y - 2X'y + \beta' X'X\beta$$

**is minimum where $y = (y_1, y_2, \ldots, y_n)'$.**

## Least squares estimate of $\beta$ :

**Differentiating $S^2$ with respect to $\beta$ and substituting it to be zero, the normal equations are obtained as**

$$\frac{dS^2}{d\beta} = 2X'X\beta - 2X'y = 0$$

**or** $X'X\beta = X'y$

**If *X* has full rank *p*, then $(X'X)$ has a unique inverse and the unique least squares estimate (LSE) of $\beta$ is**

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

## Least squares estimate of $\beta$ :

**$\hat{\beta}$ (LSE) is the best linear unbiased estimator of $\beta$ in the sense of having minimum variance in the class of linear and unbiased estimator.**

**If the rank of *X* is not full, then generalized inverse is used for finding the inverse of (*X'X*).**

**If $L'\beta$ is a linear parametric function where $L = (\ell_1, \ell_2, \ldots, \ell_p)'$ is a non-null vector, then the least-squares estimate of $L'\beta$ is $L'\hat{\beta}$.**

## Least squares estimate of $\beta$ :

**A question arises that what are the conditions under which a linear parametric function $L'\beta$ admits a unique least-squares estimate in the general case.**

**The concept of estimable function is needed to find such conditions.**