Analysis of Variance and Design of Experiments

General Linear Hypothesis and Analysis of Variance

:::

Lecture 3

Regression and Analysis of Variance Models

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

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

$$Var(Y_i) = \sigma^2.$$

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 $X_1, X_2, \ldots, X_p$.

Regression model for the general linear hypothesis:

Alternatively, the linear model can be expressed as

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

where ε_i**'s are identically and independently distributed random error component with mean 0 and variance** σ^2**, i.e.,**

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

Regression model for the general linear hypothesis:

In matrix notations, the linear model can be expressed as

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

where

- $Y = (Y_1, Y_2, ..., Y_n)'$ **is a *n* x 1 vector of observations on the response variable,**
- **the matrix** $X = \begin{pmatrix} X_{11} & X_{12} \ldots X_{1p} \\ X_{21} & X_{22} \ldots X_{2p} \\ \vdots & \vdots \quad \ddots \quad \vdots \\ X_{n1} & X_{n2} \ldots X_{np} \end{pmatrix}$ **is a *n* x *p* matrix of *n* observations on *p* independent covariates** $X_1, X_2, ..., X_p$,

Regression model for the general linear hypothesis:

- $\beta = (\beta_1, \beta_2, ..., \beta_p)'$ **is a *p* x 1 vector of unknown regression parameters (or regression coefficients)** $\beta_1, \beta_2, ..., \beta_p$ **associated with** $X_1, X_2, ..., X_p,$ **respectively and**
- $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ **is a *n* x 1 vector of random errors or disturbances.**
- **We assume that** $E(\varepsilon) = 0,$ **the covariance matrix**

$$V(\varepsilon) = E(\varepsilon\varepsilon') = \sigma^2 I_p,\ rank(X) = p$$

Regression model for the general linear hypothesis:

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

- the matrix X is termed as the **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 β_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 β_j in x_i and $x_{ij} = 0$ indicates the absence of effect β_j in x_i.

Regression model for the general linear hypothesis:

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 a mixed model and is used in the analysis of covariance.

Relationship between the regression and ANOVA models:

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 a linear model in the context of linear regression analysis and analysis of variance.

Consider the multiple linear model

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

Relationship between the regression and ANOVA models:

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.**

Relationship between the regression and ANOVA models:

If β, γ and δ denote the effects associated with the covariates *X, Z* and *W* which are the 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.**

Relationship between the regression and ANOVA models:

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 the case of regression analysis, the covariates $X_1, X_2, \ldots, X_p$ are the different variables like temperature, the quantity of fertilizer, amount of irrigation etc.

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 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.