Analysis of Variance and Design of Experiments

Experimental Design Models

:::

Lecture 12

Basics for ANOVA in Experimental Design Models

Shalabh

Department of Mathematics and Statistics

Indian Institute of Technology Kanpur

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

Experimental Design Models:

We consider the models which are used in designing an experiment.

The experimental conditions, experimental setup and the objective of the study primarily determine that what type of design is to be used and hence which type of design model can be used for the further statistical analysis to conclude about the decisions.

These models are based on one-way classification, two-way classifications (with or without interactions), etc.

Experimental Design Models:

We discuss them now in detail in a few setups which can be extended further to any order of classification.

We discuss them now under the set up of one-way and two-way classifications.

It may be noted that it has already been described how to develop the likelihood ratio tests for the testing the hypothesis of equality of more than two means from normal distributions and now we will concentrate more on deriving the same tests through the least-squares principle under the setup of the linear regression model.

One way classification:

Let *p* random samples from *p* normal populations with the same variances but different means and different sample sizes have been independently drawn.

Let the observations Y_{ij} follow the linear regression model setup and

Y_{ij} denotes the j^{th} observation of dependent variable *Y* when the effect of i^{th} level of the factor is present.

The design matrix is assumed to be not necessarily of full rank and consists of 0's and 1's only.

One way classification:

Then Y_{ij} are independently normally distributed with

$$E(Y_{ij}) = \mu + \alpha_i, i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, n_i$$

$$V(Y_{ij}) = \sigma^2$$

where

μ **is the general mean effect.**

μ **is fixed. It gives an idea about the general conditions of the experimental units and treatments.**

α_i **is the effect of i^{th} level of the factor. It can be fixed or random.**

One way classification: Example

Consider a medicine experiment in which there are three different dosages of drugs - 2 mg., 5 mg., 10 mg. which are given to patients for controlling the fever.

These are the 3 levels of drugs, and so denote

$$\alpha_1 = 2 \text{ mg.}, \qquad \alpha_2 = 5 \text{ mg.}, \qquad \alpha_3 = 10 \text{ mg.}$$

Let *Y* denotes the time taken by the medicine to reduce the body temperature from high to normal.

One way classification: Example

Suppose two patients have been given 2 mg. of dosage, so Y_{11} and Y_{12} will denote their responses.

So we can write that when $\alpha_1 = 2\ mg.$ is given to the two patients, then

$$E(Y_{1j}) = \mu + \alpha_1;\ j = 1, 2.$$

Here μ denotes the general mean effect which may be thought as follows: The human body has a tendency to fight against the fever, so the time taken by the medicine to bring down the temperature depends on many factors like body weight, height, general health condition etc. of the patient.

One way classification: Example

So μ denotes the general effect of all those factors which are present in all the observations.

In the terminology of the linear regression model, μ denotes the intercept term which is the average value of the response variable when all the independent variables are set to take value zero.

In experimental designs, the models with intercept term are more commonly used and so generally we consider these types of models.

One way classification: Example

Similarly, if $\alpha_2 = 5$ mg. and $\alpha_3 = 10$ mg. of dosages are given to 4 and 7 patients respectively then the responses follow the model

$$E(Y_{2j}) = \mu + \alpha_2;\ j = 1,2,3,4$$

$$E(Y_{3j}) = \mu + \alpha_3;\ j = 1,2,3,4,5,6,7.$$

One way classification:

Also, we can express

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij};\ i = 1, 2, \ldots, p,\ j = 1, 2, \ldots, n_i$$

where ε_{ij}**‘ s the random error component in** Y_{ij} **.**

It indicates the variations due to uncontrolled causes which can influence the observations.

We assume that ε_{ij}**'s are identically and independently distributed as** $N(0, \sigma^2)$ **with** $E(\varepsilon_{ij}) = 0, Var(\varepsilon_{ij}) = \sigma^2.$

One way classification:

Note that the general linear model considered is $E(Y) = X\beta$

for which Y_{ij} **can be written as** $E(Y_{ij}) = \beta_i$

When all the entries in *X* are 0's or 1's, then this model can also be re-expressed in the form of

$$E(Y_{ij}) = \mu + \alpha_i .$$

This gives rise to some more issues.

Consider and rewrite

$$E(Y_{ij}) = \beta_i = \overline{\beta} + (\beta_i - \overline{\beta})$$
$$= \mu + \alpha_i$$

where

$$\mu \equiv \overline{\beta} = \frac{1}{p}\sum_{i=1}^{p} \beta_i$$

$$\alpha_i \equiv \beta_i - \overline{\beta}.$$

One way classification:

Now let us see the changes in the structure of the design matrix and the vector of regression coefficients.

The model $E(Y_{ij}) = \beta_i = \mu + \alpha_i$ **can now be rewritten as**

$$E(Y) = X^* \beta^*$$

$$Cov(Y) = \sigma^2 I$$

where $\beta^* = (\mu,\ \alpha_1,\ \alpha_2, \ldots,\ \alpha_p)'$ **is a (*p*+1) x 1 vector and** $X^* = \begin{bmatrix} 1 & \\ 1 & \\ \vdots & X \\ 1 & \end{bmatrix}$

is a $n \times (p+1)$ **matrix, and *X* denotes the earlier defined design matrix in which**

- first n_1 rows as (1,0,0,…,0),

- second n_2 rows as (0,1,0,…,0) …, and

- last n_p rows as (0,0,0,…,1).

One way classification:

We earlier assumed that $rank(X) = p$ **but can we also say that** $rank(X^*)$ **is also *p* in the present case?**

Since the first column of *X is the vector sum of all its remaining *p* columns, so** $rank(X^*) = p$**.**

It is thus apparent that all the linear parametric functions of $\alpha_1, \alpha_2, \ldots, \alpha_p$ **are not estimable.**

The question now arises is what kind of linear parametric functions are estimable?

One way classification:

Consider any linear estimator $L = \sum_{i=1}^{p} \sum_{j=1}^{n_i} a_{ij} Y_{ij}$ **with** $C_i = \sum_{j=1}^{n_i} a_{ij}.$

Now $E(L) = \sum_{i=1}^{p} \sum_{j=1}^{n_i} a_{ij} \; E(Y_{ij})$

$$= \sum_{i=1}^{p} \sum_{j=1}^{n_i} a_{ij} \; (\mu + \alpha_i)$$

$$= \mu \sum_{i=1}^{p} \sum_{j=1}^{n_i} a_{ij} + \sum_{i=1}^{p} \sum_{j=1}^{n_i} a_{ij} \alpha_i$$

$$= \mu (\sum_{i=1}^{p} C_i) + \sum_{i=1}^{p} C_i \alpha_i .$$

Thus $\sum_{i=1}^{p} C_i \alpha_i$ **is estimable if and only if** $\sum_{i=1}^{p} C_i = 0,$ **i.e.,** $\sum_{i=1}^{p} C_i \alpha_i$ **is a contrast.**

Thus, in general, neither $\sum_{i=1}^{p} \alpha_i$ **nor any** $\mu, \alpha_1, \alpha_2, \ldots, \alpha_p$ **is estimable.**

One way classification:

Thus, in general, neither $\sum_{i=1}^{p} \alpha_i$ nor any $\mu, \alpha_1, \alpha_2, \ldots, \alpha_p$ is estimable.

If it is a contrast, then it is estimable.

This effect and outcome can also be seen from the following explanation based on the estimation of parameters $\mu, \alpha_1, \alpha_2, \ldots, \alpha_p$.

One way classification:

Consider the least-squares estimation $\hat{\mu}, \hat{\alpha}_1, \hat{\alpha}_2, \ldots, \hat{\alpha}_p$ of $\mu_1, \alpha_1, \alpha_2, \ldots, \alpha_p$ respectively.

Minimize the sum of squares due to ε_{ij} 's.

$$S = \sum_{i=1}^{p} \sum_{j=1}^{n_i} \varepsilon_{ij}^2 = \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i)^2$$

to obtain $\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_p$.

(a) $$\frac{\partial S}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{p} \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) = 0$$

(b) $$\frac{\partial S}{\partial \alpha_i} = 0 \Rightarrow \sum_{j=1}^{n_i} (y_{ij} - \mu - \alpha_i) = 0, \; i = 1, 2, \ldots, p.$$

Note that (a) can be obtained from (b) or vice versa.

So (a) and (b) are linearly dependent in the sense that there are (p + 1) unknowns and p linearly independent equations.

One way classification:

Consequently, $\hat{\mu}, \hat{\alpha}_1, \ldots, \hat{\alpha}_p$ **do not have a unique solution.**

Same applies to the maximum likelihood estimation of $\mu, \alpha_1, \ldots \alpha_p$.

If a side condition that $\sum_{i=1}^{p} n_i \hat{\alpha}_i = 0$ **or** $\sum_{i=1}^{p} n_i \alpha_i = 0$ **is imposed then (a) and (b) have a unique solution as**

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{p} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{oo},$$

$$\hat{\alpha}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} - \hat{\mu} = \bar{y}_{io} - \bar{y}_{oo}$$

where $n = \sum_{i=1}^{p} n_i$.

One way classification:

In case, all the sample sizes are the same, then the condition

$\sum_{i=1}^{p} n_i \hat{\alpha}_i = 0$ **or** $\sum_{i=1}^{p} n_i \alpha_i = 0$ **reduces to** $\sum_{i=1}^{p} \hat{\alpha}_i = 0$ **or** $\sum_{i=1}^{p} \alpha_i = 0$ **.**

So the model $y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$ **needs to be rewritten so that all the parameters can be uniquely estimated. Thus**

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij} = (\mu + \bar{\alpha}) + (\alpha_i - \bar{\alpha}) + \varepsilon_{ij}$$
$$= \mu^* + \alpha_i^* + \varepsilon_{ij}$$

where $\mu^* = \mu + \bar{\alpha}, \quad \alpha_i^* = \alpha_i - \bar{\alpha}, \quad \bar{\alpha} = \frac{1}{p}\sum_{i=1}^{p} \alpha_i$ **and** $\sum_{i=1}^{p} \alpha_i^* = 0$

This is a reparameterized form of the linear model.

One way classification:

Thus in a linear model when *X* is not of full rank, then the parameters do not have unique estimates.

In such conditions, a restriction $\sum_{i=1}^{p} \alpha_i = 0$

(or equivalently $\sum_{i=1}^{p} n_i \alpha_i = 0$ **in case all n_i's are not the same) can be added and then the least squares (or maximum likelihood) estimators obtained are unique.**

The model

$$E(Y_{ij}) = \mu^* + \alpha_i^*; \ \sum_{i=1}^{p} \alpha_i^* = 0$$

is called a reparametrization of the original linear model.

One way classification:

Let us now consider the analysis of variance with an additional constraint. Let

$$Y_{ij} = \beta_i + \varepsilon_{ij},\ i = 1, 2, \ldots, p;\ j = 1, 2, \ldots, n_i$$
$$= \bar{\beta} + (\beta_i - \bar{\beta}) + \varepsilon_{ij}$$
$$= \mu + \alpha_i + \varepsilon_{ij}$$

with

$$\mu = \bar{\beta} = \frac{1}{p}\sum_{i=1}^{p}\beta_i, \quad \alpha_i = \beta_i - \bar{\beta}, \quad \sum_{i=1}^{p} n_i\alpha_i = 0, \quad n = \sum_{i=1}^{p} n_i,$$

and ε_{ij} 's are identically and independently distributed with mean 0 and variance σ^2.