M.Sc (Statistics)

CORE COURSE - VI

SUB. CODE : 18KP2S06

LINEAR MODEL AND DESIGN OF EXPERIMENTS

UNIT- I

Linear Models: Definition - Functionally Related Models - Mean Related Model - Regression Model - Experimental Design Model - Components of Variance Model - Point Estimation - Estimation of β and σ^2 under normal theory - Gauss Mark off theorem.

UNIT- II

RBD - Missing observations in RBD - Analysis of RBD with one and two missing values - Analysis of LSD with one and two missing values - Orthogonal Latin Squares - Graeco LSD.

UNIT- III

Factorial Experiment: Main effects, Interaction effects, Orthogonal Contrasts - Design for 2^n, 3^n experiments. Total and Partial confounding in 2^3 experiments - Analysis of Split Plot Design.

UNIT- IV :

Incomplete Block Design - Balanced Incomplete Block Design - Relationship between the parameters. Constructions of BIBD - Recovery of Inter Block Information - Resolvable Designs.

UNIT- V :

Analysis of PBIBD - Two Associate classes and Response surface design.

LINEAR MODEL :-

A Linear Statistical model is generally used to determine the value Y from a knowledge of X. The form of this model is $Y = \mu(x) + \epsilon$ where Y and ϵ are random variables.

X is non-random variables

$\mu(x)$ is a function of X.

Here,

Y is called dependent variables (or) response variables.

X is called Independent variables (or) predictor variables.

The difference $Y - \mu(x)$ is denoted by a letter ϵ and is called a error.

GENERAL LINEAR MODEL :-

In general Linear model the random variable ϵ is non-observable let us assume that the dependent variable Y. The function of (k-1) Independent variable is denoted as

$$Y_i = B_1 X_1 + B_2 X_2 + \cdots + B_n X_n + \mu \; ; \; (i = 1, 2, \ldots n) \quad \text{(or)}$$

$$Y = XB + M$$

The above general Linear model is subject to the following is assumptions

(i) V_i is normally distributed $u_i \; Y_i \sim N(0,1)$

(ii) $E(u_i) = 0$

(iii) $E(u_i)^2 = \sigma_u^2$.

(iv) $E(U_i u_j) = 0 \quad \forall i \neq j$.

LINEAR MODEL : Definition :

Linear model is an equation that involves random variable and mathematical, variables and parameter and that is linear in the

for example : if $\gamma_0, \gamma_1, \gamma_2$ are unknown parameters, then $\gamma_0 + \gamma_1 x + \gamma_2 y = 0$ is a linear model however $\gamma_0^2 + x \sin \gamma_1 + \gamma_2 = 0$ is not a linear model of the parameter γ_0, γ_1 and γ_2.

CLASSIFICATION OF LINEAR MODEL :-

They are many types of Linear model that the scientist may want use in getting information that will help predict quantityes in the real world there are five types of linear models will be described they are

(i) Functionally related models
(ii) Mean related models
(iii) Regression models
(iv) Experimental design models
(v) Comparence of variance models.

FUNCTIONAL RELATED MODELS :-

The models are characterised by a functional relationship among mathematical variables. Some of which can be observed owing to errors of measurement. Suppose two quantities Y^* and x statisfied the equation $Y^* = \alpha_0 + \alpha_1 x$.

Implied be can observed x but not Y^* we observed y (where $y = Y^* + \epsilon$. Suppose ϵ is a random error with mean zero) then.

$$y = y^* + e$$

$$E(y) = E(y^*) + E(e) \qquad E(e) = 0$$

$$= E(y^*)$$

$$E(y) = y^*$$

we can write the model is

$$y = y^* + e$$

$$y - e = y^*$$

$$y - e = \alpha_0 + \alpha_1 x$$

$$y = \alpha_0 + \alpha_1 x + e$$

where y and e are r.v's the y and x are observed.

EXAMPLE :-

In electricity ohm's laws states that the voltage v^* in a circuite is equal to the product to the resistid ρ of the wire and the current I. the equation is $v^* = \rho I$

Suppose an experimental wants to find the ρ of a circute suppose he observed v.

$$v = v^* e$$

$$v = \rho I + e$$

here v and e are r.v's I is a mathematical variable, ρ is unknown parameter to be determine.

2. MEAN RELATED MODEL :-

This model are characterised by equ error suppose that the functional relationships given by

$$f(y; x_1, x_2, \ldots x_n) = 0.$$

suppose that

$f(y; x_1, x_2, \ldots x_n)$ can be written if the term $g(x_3, \ldots x_n)$ is droped its clear that y is not exactly determined for fixed values of x_1, x. Since, the remaining variable may take on various value every though the first and variable are fixed.

we can write

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \epsilon$$

It should be pointed of that the x_i can be function of other variables.

Example : $x_1 = \log t$; $x_2 = 3^t$ than $y = \alpha_0 + \alpha_1 \log t + \alpha_2 3^t + e$.

(or) $x_1 = t$, $x_2 = t^2$ then we have curve linear model

$$y = \alpha_0 + \alpha_1 t + \alpha_0 t^2 + e$$

In this model the dependent variable y is r.v's where as the Independent variable α_1 and α_2 are not r.v's but are pre determine mathematical variable also the expected value of the dependent variable y does not give a functional relationship x_i and x_2

$$E(y) = Y.$$

The functional relationship is

$$Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + y(x_3 + \cdots + x_n)$$

$$E(y) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2.$$

which is equal to (the average value of y) average is mean.

3. REGRESSION MODEL

This model are characterised by that fact that the variables enterining into there r.v let $y = f(Y, x)$ be the joint distribution of y and x. $E(Y/x = x) = \alpha_0 + \alpha_1 X$

$$Y_x = \alpha_0 + \alpha_1 x + e$$

It may be that two inobserwable r.v Y^* and X^* have the joint distribution given above if be observed $Y = Y^* + e$

$$x = x^* + d$$

We then have a regression model with measurement error.

FOR EXAMPLE:-

Suppose its desided to product the temperature in a certain locality by using only humality x. We shall assume that Y and x froma bivariate frequency distribution. We can use regression model and predict the avg temparature for a given value of humality. If there are error in measuring either temperature (or) humality than we can use the Alternative model with measurement error.

4. EXPERIMENT DESIGN MODEL:-

The model concided upto now have been such that they permitted as to say something about Y (or) E(Y) (or) the parameter in the model. $Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + e$ where X_i valid in some interval

The E.D model is some quads e in this models x_i takes only the value 0 and 1.

FOR EXAMPLE :-

Suppose we want to example in characteristic of two different kinds of points.

The point are each subjected to a fireen and the time the take for this missing two were the point is recorded we should like a formula. For predicting time for each point for where away point 1 and 0 bear known Numerical relationship to each other. The predistion equation can be as

$Y = \alpha_1 x_1 + \alpha_2 x_2$, where α_1 and α_2 take the value 0 and 2 α_i = time It takes point I to where away thus $Y = \alpha_1 x_1 + \alpha_2 x_2$.

x_i per i to were away $Y = \alpha_1$ where $x_1 = 1$, $x_2 = 0$ and $Y = \alpha_2$ where $x_2 = 1$, $x_1 = 0$, Y_i we the time we takes for point for where away we can write the model $y = \alpha_1 x_1 + \alpha_2 x_2$ as $\mu_i = \alpha_i$. Y_i's the function of α_i but since, $x_i = 0$ (or) 1. than $Y_i = \alpha_i + e$, $Y = \alpha_1 x_1 + \alpha_2 x_2 - e$.

5. COMPONENT'S OF VARIANCE MODEL :-

This model is similar to the enperimental design model in that x variables against take the variable 0 and 1.

The model can be written as $Y = a_1 x_1 + a_2 x_2 + e$.

however a_1 and a_2 are non parameters but are unobserved r.v's from distribution with variance σ_1^2 and σ_2^2 from respectively. The object in this models if to observed values of y and estimate σ_1^2 and σ_2^2. σ^2 is the variance of error components.

FOR EXAMPLE :-

Y_{ij} denoted the j^{th} measurement of the tract and e_{ij} of r.v from a distribution from a σ^2 the model can be written as $Y_{ij} = \mu + \alpha_i + e_{ij}$ $j = 1, 2, \dots m$, $i = 1, 2, \dots n$ where μ is the constant which is the object value of Y_{ij} the estimate σ_1^2 and σ_2^2 from the observed values of Y_{ij}. The kmo from $e_{ij} = \sigma_1^2 + \sigma_2^2$ that its a linear function of the variance.

GAUSS MARK OF THEOREM :-

If the general linear hypothesis model of full rank. $Y = XB + e$ is such that the following two condition on the random vector are met:

(i) $E(e) = 0$ (ii) $E(ee') = \sigma^2 I$.

The best (minimum variance) linear unbiased estimated of β is given by least square $\hat{\beta} = S^{-1}X'Y$ is the (Blue) Best linear unbiased estimator of B where $S = X'X$.

PROOF:

A be any $P \times n$ constant matrix and let $B^* = AY$. B^* is the greatest general linear function of Y which we shall take as an estimate the β we must specified the element of A. So, that β^* is the best unbiased estimator of β. then Let that $A = S^{-1}X' + B$ Since $S^{-1}X'$ we have known, we must find B in order to be able to be specific A. for unbiasness we have $E(B^*) = E(AY) = E[(S^{-1}X' + B)Y] = (S^{-1}X' + B)X\beta$ $= \beta + BX\beta$ but to be unbiased $E(\beta^*)$ must equal β and this implies that $BX\beta = 0$ for all β. Thus unbiasedness specified that $BX = 0$ for the properties of best we must find the matrix B that minimized the variance of β^*_I where $I = 1, 2, \cdots P$. subject to the restriction $BX = 0$. Consider the co-variance of $Cov(\beta^*) = E[(\beta^* - \beta)(\beta^* - \beta)']$

$= E[(S^{-1}X' + B)Y - \beta][(S^{-1}X'^{-1} + B)(Y - \beta)']$

Substituting

$XB + e$ for Y and using $BX = 0$ we get.

$= [(S^{-1}X' + B)(XB + e) + B]$

$Cov(\beta^*) = E[S^{-1}X'ee'XS^{-1} + Bee B' + S^{-1}X'ee' + B' + Be'eXS^{-1}]$

$= \sigma^2(S^{-1} + BB')$

Let $BB' = G = (g_{ij})$ then $Cov(\beta^*) = \sigma^2(S^{-1} + G)$

The diagonal elements of $cov(\beta^*)$ are the respective variance of the β_i^*.
(To minimize each $var(\beta_i^*)$ minimize each diagonal element of $cov(\beta^*)$. Since σ^2 and S^{-1} are constant find a matrix G such that each diagonal elements of G is minimum. But $G = BB'$ is positive semi definite hence $g_{ij} \geq 0$.

Thus, the diagonal elements of $cov(\beta^*)$ will attain their minimum when $g_{ii} = 0$ for $(i = 1, 2, \ldots p)$. But if $B = (b_{ij})$ then $g_{ii} = \sum_{j=1}^{n} b_{ij}^2$.

There fore if g_{ii} is equal 0 for all i, it must be true that $b_{ij} = 0$ for all i and all j. This implies that $B = 0$. The condition $B = 0$ is compatable with the condition of unbiasedness $BX = 0$. Therefor $A = S^{-1}x$ and $\beta^* = \hat{\beta}$.

UNIT- II

RANDOMISED BLOCK DESIGN (R.B.D).

In field experimentation, if the whole of the experimental area is not homogenous and the fertility gradient is only is one direction, the simple method of controlling the variability of the experimental material consists in stratifying or grouping the whole are into relatively homogeneous strata or sub-groups (or blocks or replicates as they are called) perpendicular to the direction of the fertility gradient. Now, if the treatments are applied at random to relatively homogenous units within each strata or block and replicated over all the blocks the design a (RBD.) Randomised Block Design.

ESTIMATE OF MISSING VALUE in R.B.D :-

Let the observation $y_{ij} = x$ (says) in the j^{th} block and receiving the i^{th} treatment be missing.

Treatments

		1	2	....	i	...	t	TOTAL
Blocks	1	y_{11}	y_{21}	....	y_{i}	...	y_{t1}	$y_{.1}$
	2	y_{12}	y_{22}	....	y_{i2}	...	y_{t2}	$y_{.2}$
	3	y_{31}	y_{23}	....	y_{i3}	...	y_{t3}	$y_{.3}$
	⋮	⋮	⋮		⋮		⋮	⋮
	j	y_{1j}	y_{2j}	....	y_{ij}	...	y_{tj}	$y_{.j}+x$
	⋮	⋮	⋮		⋮		⋮	⋮
	r	y_{1r}	y_{2r}		y_{ir}		y_{tr}	$y_{..}+x$
		$y_{1.}$	$y_{2.}$	....	y_{i+x}	...	y_{t}	

$y_{i.}$ is total of known observation getting i^{th} treatment

y_{ij}' is the total of known observations in j^{th} block and i^{th} treatment.

$y_{..}'$ is total of all the known observation total $S.S = \Sigma\Sigma y_{ij}^2 - C.f = x^2 +$ constant w.r.t $x - C.f$.

$$S.S.T = \frac{1}{r}\left[(y_{i}' + x^2) + \text{constant w.r.t } x\right] \; ; \; S.S.B = \frac{1}{t}\left[(y_{.j}' + x^2) + \text{constant w.r.t } x\right] - C.f$$

where $C.F = (y_{..}' + x)^2 / rt$, E = residual , $S.S = T.S.S - S.S.E - S.S.T$

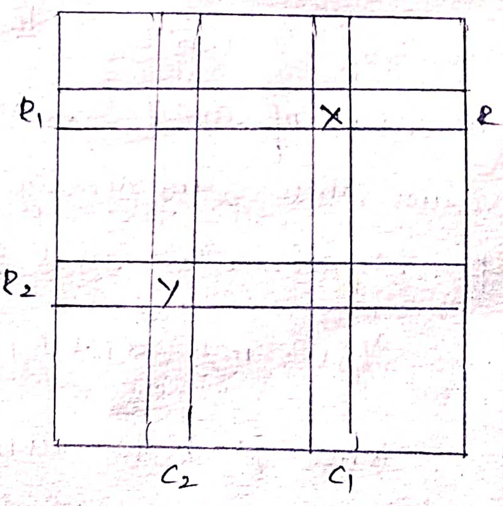
$$= x^2 - \frac{1}{t}(y_{.j}' + x)^2 + \frac{1}{r}(y_{i.}' + x)^2 + \left(\frac{y_{..}' + x}{rt}\right)^2 + \text{constant terms independent of } x.$$

$$\frac{\partial E}{\partial x} = 0 = 2x - 2/t\,(y_{.j}' + x) - 2/r\,(y_{i.}' - Ax) + \frac{2}{rt}(y_{..}' + x)$$

$$x = \frac{r y_{ij}' + t y_{i.}' - y_{..}'}{(r-t)(t-1)}$$

TWO MISSING OBSERVATION :

For two missing values, say x and y. Let R_1 & R_2 be the totals of knowns observations in the rows containing x and y respectively and C_1 & C_2 be the total of known observations in the columns containing x & y respectively.



$$E = x^2 + y^2 - 1/t\left[(R_1 + x)^2 + (R_2 + y)^2\right] - 1/r\left[(C_1 + x)^2 + (C_2 + y)^2\right] + 1/rt\,(S + x + y)^2 + \text{Terms independent of } x \,\&\, y.$$

For a minima of E subject to variations in x and y we must have,

$$\frac{\partial E}{\partial y} = 0 = y - 1/t\,(R_2 + y) - 1/r\,(C_2 + y) + 1/rt\,(S + x + y)$$

$$\frac{\partial E}{\partial x} = 0 = x - 1/t\,(R_1 + x) - 1/r\,(C_1 + x) + 1/rt\,(S + x + y).$$

$\Rightarrow (r-1)(t-1)x = rR_1 + tC_1 - S - y$ & $(r-1)(t-1)y = rR_2 + tC_2 - S - x$.

Solving these equations simultaneously we get the estimates of x & y

LATIN SQUARE DESIGN :-

In RBD whole of the experimental area is divided into relatively homogenous groups and treatments are allocated as random to units within each blocks, ie, randomisation was subjected to one restriction ie, within blocks. A useful method of eliminating facility variations consists in an experimental layout which will control variation in two perpendicular directions such a layout is a Latin Square Design (LSD).

STATISTICAL ANALYSIS OF m×m LSD FOR ONE OBSERVATION :-

Let y_{ijk} $(i,j,k = 1,2,\ldots m)$ denote response from the unit plot in field experimental in the i^{th} row j^{th} treatment column and receiving k^{th} treatment the triplet (i,j,k) assumes only m^2 of the possible m^3 values of an L.S selected by the experiment. If S represent the set of m^2 values then simbolically (i,j,k) belongs to S. Single observation is made per experimental unit then linear additive model is

$$y_{ijk} = m\mu + \alpha_i + \beta_j + \tau_k + \epsilon_{ijk}, \quad (i,j,k) \in S$$

where μ is the constant effect α_i, β_j and τ_k are the effect due to i^{th} row j^{th} column and k^{th} treatment respectively and ϵ_{ijk} is error effect due to random component assumed to be normally distributed with mean zero and variance σ_e^2 ie $\epsilon_{ij} \sim N(0, \sigma_e^2)$.

If, we write $G = Y_{\ldots}$ = Total of all the m^3 observation, $R_i = Y_{i..}$ = Total of the m^{th} observation time in the i^{th} row, $C_j = y_{.j.}$ = Total of the m observation time in the j^{th} column, $T_k = Y_{..k}$ = Total of the m observation time from k^{th} treatment.

$$\sum_{i,j,k\in S}(y_{ijk}-\bar{y}_{...})^2 = \sum_{i,j,k\in S}\left[(\bar{y}_{i..}-\bar{y}_{...})+(\bar{y}_{.j.}-\bar{y}_{...})+(\bar{y}_{..k}-\bar{y}_{...})\right]\{y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+\bar{y}_{...})\}$$

$$= m\sum_i(\bar{y}_{i..}-\bar{y}_{...})^2+m\sum_j(\bar{y}_{.j.}-\bar{y}_{...})^2+m\sum_k(\bar{y}_{..k}-\bar{y}_{...})^2+\sum_{ijk\in S}(y_{ijk}-\bar{y}_{i..}-\bar{y}_{.j.}-\bar{y}_{..k}+2\bar{y}_{...})^2 .$$

$\because$ T.S.S = S.S.R + S.S.C + S.S.T + S.S.E

Where T.S.S is the total sum of squares and S.S.R, S.S.C, S.S.T and S.S.E represent. Sum of square due to rows, columns, treatment error respectively given by,

$T.S.S = \sum_{ijk\in S}(y_{ijk}-\bar{y}_{...})^2$; $S.S.R = S_R^2 = m\sum_i(\bar{y}_{i..}-\bar{y}_{...})^2$. $S.S.C = S_C^2 = m\sum_j(\bar{y}_{.j.}-\bar{y}_{...})^2$

$S.S.T = S_T^2 = m\sum_k(\bar{y}_{..k}-\bar{y}_{...})^2$ & $S.S.E = S_E^2 = T.S.S - S.S.R - S.S.C - S.S.T$.

ANOVA TABLE :-

SOURCE OF VARIATION	d.f	S.S	M.S.S	Variation Ratio F
Rows	$m-1$	S_R^2	$s_R^2 = S_R^2/m-1$	$F_R = s_R^2/s_E^2$
Column	$m-1$	S_C^2	$s_C^2 = S_C^2/m-1$	$F_C = s_C^2/s_E^2$
Treatments	$m-1$	S_T^2	$s_T^2 = S_T^2/m-1$	$F_T = s_T^2/s_E^2$
Error	$(m-1)(m-2)$	S_E^2	$s_E^2 = S_E^2/(m-1)(m-2)$	
Total	m^2-1			

Let us setup the null hypothesis for row effects H_α, $\alpha_1=\alpha_2 \dots \alpha_m=0$ column effect $H_\beta: \beta_1=\beta_2=\dots=\beta_m=0$ and for treatment efforts $H_T: T_1=T_2=\dots T_m=0$. The variance rations F_R, F_C and F_T follow (Central) F distribution with $(m-1)(m-1)(m-2)$ d.f under of null hypothesis H_α, H_β & H_T respectively.

Let $F_\alpha = F_\alpha\{(m-1), (m-1), (m-2)\}$ be the tabulated value of F for (m-1) (m-1) (m-2) d.f as the level of significance 'α'. Thus if $F_R > F_\alpha$ similar we can test for H_β and H_α.

ESTIMATION OF MISSING VALUES:-

Let us suppose that in m×m Latin square, the observation occuring in the ith row and jth column and receiving the kth treatment is missing. Let us assume that its value is x i.e $y_{ijk} = x$. R. To tel of known observations of the ith row. ie, the row containing 'x' column. ie, the re C- Total of known observations receiving kth treatment. i.e total of all known, treatment values containing 'x'.

S = Total of known observation. Then

$$T.S.S = x^2 + \text{constants w.r.t } x \; \frac{(S+x)^2}{m^2}$$

$$S.S.R = \frac{(R+x)^2}{m} + \text{constant w.r.t } x - \frac{(S+x)^2}{m^2}$$

$$S.S.C = \frac{(C+x)^2}{m} + \text{constant w.r.t } x - \frac{(S+x)^2}{m^2}$$

$$S.S.T = \frac{(T+x)^2}{m} + \text{constant w.r.t } x - \frac{(S+x)^2}{m^2}$$

E = Residual sum of square (S.S.E) = T.S.S − S.S.R − S.S.C − S.S.T.

$$= x^2 - \frac{1}{m}\left[(R+x)^2 + (C+x)^2 + (T+x)^2\right] + 2\,\frac{(S+x)^2}{m^2}.$$

$$\frac{\partial E}{\partial x} = 0 \Rightarrow 2x - 2/m\left(R+C+T+3x\right] + 4\,\frac{(S+x)}{m^2} \quad \Rightarrow (m^2-3m+2)x = m(R+C+T) - 2S$$

$$x = \frac{m(R+C+T) - 2S}{(m-1)(m-2)}$$

ORTHOGONAL LATIN SQUARE :-

In combinatorics, two Latin squares of the same size are said to be orthogonal if when superimposed the ordered paired entries in the positions are all distinct. A set of Latin squares, all the same order, all pairs of which are orthogonal is called a set of Mutually - Orthogonal Latin Squares

GRAECO-LATIN SQUARE :-

A Graeco-Latin Square or Euler square or pair of orthogonal Latin square of order n over two sets S and T (which may be the same), each consisting of n symbols is an n×n arrangement of cells, each cell containing an ordered pair (s,t), where s is in S and t is in T, such that every row and every column contains each element of S and each element of T exactly once and that no two cells contain the same ordered pair.