2. CLASSICAL LINEAR MODEL

2.1 Introduction :

As discussed in the first chapter, there are different methods of analysing the numerical data. One such method is fitting an appropriate model to the data and then using fitted model for predicting future values. 'classical linear model' (CLM) is one such model. This is the oldest and simplest model. This model is useful in the situations where the response variable (Y) is thought to be linearly related to one or more covariates X_j ($j=1,2,\ldots,k$). The purpose of fitting classical linear model is to use data for estimation of the form of this linear relationship approximately.

Gauss and Legendre began the study of classical linear model, and applied the model to astronomical data where the variables are continuous. In their astronomical study a significant part of variation in the observations is due to measurement errors. Gauss suggested normal distribution to describe the distribution of the errors. With this assumption Gauss developed the theory of classical linear models. This theory has been discussed by many authors in literature. Some of them are Graybill (1961), Searle (1971), Plackett (1972), Rao (1973), Draper & Smith (1981) and Stigler (1981).

In this chapter we discuss the following points.

(1) Description of a classical linear model with k independent variates ;

(2) fitting of classical linear model;

(3) testing of hypotheses ;

(4) residual analysis ;

(5) limitations of the theory of classical linear model.

2.2 Classical linear model with k independent variates :

Definition-1 : Classical linear model : Let Y be the response variate having $N(\mu,\sigma^2)$ distribution and $X_j (j=1,2,...,k)$ be k stimulus variates. Then classical linear model can be written as,

$$Y = \underline{X}'\beta + e \tag{1}$$

or equivalently,

$$E(Y) = \underline{X}'\beta \tag{2}$$

where

(i) $\underline{X}' = (1\ \underline{X}_1')$;

(ii) $\underline{X}_1' = (X_1, X_2, ..., X_k)$ is the vector of stimulus variates ;

(iii) $\beta' = (\beta_0, \beta_1, \ ..., \beta_k)$ is the vector of parameters associated with the model ;

(iv) e is the error component corresponding to Y having $N(0,\sigma^2)$ distribution.

Note :- 1. Classical linear model do not require any sort of distributional assumption for estimating parameters in the model. It only requires the assumption of constant variance and the linear relationship between systematic effects and the expected value of response variate. However the assumption of normality is required for testing different hypotheses about the model parameters. Therefore we have included the assumption of normality in the definition itself.

2. The classical linear model in equation (2) can be viewed as a special case of 'general class of models' given in equation (1.1) under the assumptions $\tau = \underline{X}'\beta$ and the $N(\mu,\sigma^2)$ distribution for the response variate Y.

Usually, classical linear models are divided into two groups, based on the value of β_0. These are

(a) Intercept classical linear model;

(b) No-intercept classical linear model.

Definition-2 : Intercept classical linear model : A classical linear model with β_0 non zero, is known as 'intercept' classical linear model. In other words in intercept classical linear model, though the values of all independent variables are equal to zero, value of the response variate is expected (theoretically) to be non zero.

As an illustration, we present the following example.

Example 2.1: The example is taken from Kalbfleisch (1985). Data were collected to investigate how the amount of fuel oil required to heat a home depends upon the outdoor air temperature and wind velocity. Table (2.1) gives the observations for 10 winter days.

We expect that the fuel consumption should increase as wind velocity increases and it should decrease as temperature increases.

TABLE 2.1

Day (i)	Fuel Consumption (Y)	Temperature (X_1)	Wind Velocity (X_2)
1	14.96	-3.0	15.3
2	14.10	-1.8	16.4
3	23.76	-10.0	41.2
4	13.20	0.7	9.7
5	18.60	-5.1	19.3
6	16.79	-6.3	11.4
7	21.83	-15.5	5.9
8	16.25	-4.2	24.3
9	20.98	-8.8	14.7
10	16.88	-2.3	16.1

Here the response variate Y is fuel consumption required to heat home and two independent variates X_1 and X_2 are respectively temperature and wind velocity. If we assume that effects are linearly related, the model becomes

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + e_i \quad , \; i=1,2,\ldots,10. \qquad (3)$$

In this example we expect a positive fuel consumption on a day in the winter season when temperature and wind velocity are both zero. In most of the practical situations intercept classical linear models are appropriate.

<u>Definition-3</u> : <u>No-intercept</u> <u>classical</u> <u>linear</u> <u>model</u> : A classical linear model with $\beta_0 = 0$ is known as 'no-intercept' classical linear model. In other words in no-intercept classical linear model, if the values of all stimulus variates are equal to zero, then the value of response variate is expected

(theoretically) to be zero. For example let Y be the income of a person and the independent variate X be the age of that person. If we assume that effects are linearly related, the model becomes

$$Y_i = \beta_1 X_{i1} + e_i \quad , \ i=1,2,\ldots,n. \qquad (4)$$

In this example if the age of a person is zero, his income is naturally zero.

In classical linear models over parameterisation occurs many times. In other words, number of parameters (p, say) exceeds the number of independent equations in $E(\underline{Y}) = X\beta$. Thus the matrix X has rank less than p. Hence (X'X) becomes singular.

Once the model for analysing the data is decided, the next step is to fit a selected model to the data. Fitting a model to the data means to estimate unknown parameters in the model. Two well known methods of fitting the classical linear model are discussed in the next section.

2.3 Fitting of classical linear model :

Consider an intercept classical linear model with k stimulus variates X_j (j=1,2,...,k) and response variate Y. Suppose there are n observations on each variate. Then the model is given by

$$\underline{Y} = X\beta + \underline{e} \ , \qquad (1)$$

where

(i) $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)'$ is the vector of observations on the response variate Y ;

(ii) $X = ((X_{ij}; \ j=0,1,\ldots,k), \ i=1,2,\ldots,n)$ is the matrix of observations on the stimulus variates X_j (j=1,2,...,k) with $X_{i0} = 1$ for i=1,2,...,n ;

(iii) $\underline{e} = (e_1, e_2, \ldots, e_n)'$ is the vector of error components, with e_i is the error component corresponding to Y_i.

Here we discuss two methods of fitting classical linear model. These are

(a) least square method;

and

(b) maximum likelihood method.

2.3.1 :<u>Least square method</u> :

Before discussing the least square method, we present below the historical developement of this technique.

This method is very important and useful even in modern statistical analysis. But there has been some confussion about the discovery of this famous method. Legendre published this method in 1805, Adrain published it in 1808 and Gauss published the same in 1809. Hence it looks as Legendre discovered this method in 1805, but in 1809 Gauss had claimed that he was using this method since 1799. Therefore, there seems to be a little confussion about to whom the credit should go?

As Legendre published this method first time in 1805, his claim about the discovery is straightforward. On the other hand Gauss had not published the method till 1809. His claim is based on indirect evidence. Plackett (1972) presented some evidence to decide the unique discoverer of this method. One of the Gauss's claim was discussed by Plackett (1972) that Gauss told other astronomers about this method before 1805. Some of them are Olbers, Lendenau and Von Zoch. Stigler (1981) gave two new evidence and tried to see the truth behind Gauss's claim. These two new evidences are discussed below.

During the period 1800-1813, Von Zoch was the editor of an astronomical periodical, which generally consisted of reviews and letters. Lindenau assisted his work. Stigler (1981) found the new evidence in a review article in geodesy, dated August 1806.

This article contains detailed reference of the method of least squares. Though this artical described the method as Legendre's method, it is not the sufficient evidence to argue that Legendre was the first.

Another important evidence was the results of Gauss's calculations using 'Meine methode'. These results were published in 1799. The question is whether Gauss derived these results by using the method of least squares. Stigler (1981) goes for it, but unfortunately Gauss's results differs significantly from the results obtained by using least square method. This gives rise to two possibilities. One is Gauss applied the least square method and made error either due to rounding off or in arithmetic calculations. Second possibility is Gauss applied the method other than least square method. As Gauss was very much perfect in his calculations, the first possibility should not be considered. Stigler (1981) also verified that the required accuracy is attainable with the help of Valacq's 1794 table of logarithms, which Gauss might have used. Hence the difference between Gauss's results and those obtained from least square method can not be assigned completely to rouding off error. Thus only the last possibility is remaining.

According to Stigler (1981), since the first order approximation in computation of meredian quadrant not giving results with desired accuracy, Gauss might have gone for second order approximation. This approximation was used later by Bowditch (1832) and by Bessel (1837). Stigler (1981) showed that with Bessel's approch, non linear least square results are coinsiding with Gauss's results. This suports Gauss's claim about the discovery of method of least squares. With this background, now we describe below the method of least squares.

Least square Method : To fit a classical linear model by this method, assumption about the distribution of response variate (Y) is not required. Consider a classical linear model given in equation (1). In least square method of estimation, estimates of the unknown parameters are obtained so as to minimise the error sum of squares (E, say) under the assumption that error components are independently distributed with mean zero and variance σ^2(, say).

Here we have,

$$E = \underline{e}'\underline{e}$$

$$= \underline{Y}'\underline{Y} - 2\beta'X'\underline{Y} + \beta'(X'X)\beta. \qquad (2)$$

Taking partial differentiation w.r.t. β on both the sides of the equation (2) and equating the differential to zero, we get

$$X'\underline{Y} - (X'X)\hat{\beta} = \underline{0} \qquad (3)$$

This gives least square estimator of β as,

$$\hat{\beta} = \begin{cases} (X'X)^{-1}X'\underline{Y} \;; \text{ if } (X'X) \text{ is non-singular,} \\ (X'X)^{-}X'\underline{Y} \;; \text{ if } (X'X) \text{ is singular.} \end{cases} \qquad (4)$$

where, $(X'X)^{-}$ is a generalised inverse (g-inverse) of $(X'X)$. Note that g-inverse is not unique. Hence in the over parameterisation case there is no unique solution of the normal equations (3). In this case β is not estimable or identifiable. In order to obtain a particular solution, more equations are to be used so that $\hat{\beta}$ is the solution of the equations

$$-(X'X)\hat{\beta} + X'\underline{Y} = \underline{0},$$

and

$$C\,\hat{\beta} = \underline{0}.$$

As in over parameterisation case, extra equations are needed, generally classical linear model itself includes the constraint equations. A major advantage of least square method is that calculations are simple and straightforward as compared to the other methods of estimation. Further, following are some important properties of least square estimates. These methods are explicitly covered by Birkes (1993).

1. Least square estimates are best linear unbiased estimates (BLUEs).
2. As the distribution of population errors is normal, then least square estimates are uniformaly minimum variance unbiased estimates (UMVUEs), and are same as maximum likelihood estimates (MLEs).

Now we discuss below the second well known method, namely, method of maximum likelihood.

2.3.2 : Maximum likelihood method of estimation :

From the definition of classical linear model, we have e_i $(i=1,2,\ldots,n)$ as unobservable independently and identically distributed normal variates having $N(0,\sigma^2)$ distribution, so that we can write the log likelihood $l(\beta;\underline{e})$ based on n observations as

$$l(\beta;\underline{e}) = -\{(n/2)\ln(2\Pi)\} - \{(n/2)\ln(\sigma^2)\} - \{(\underline{Y}-X\beta)'(\underline{Y}-X\beta)/(2\sigma^2)\}. \tag{5}$$

Taking partial derivatives of both the sides of equation (5) w.r.t. β and equating it to zero we get the normal equations as

$$-(X'X)\hat{\beta} + X'\underline{Y} = \underline{0}. \tag{6}$$

This implies

$$\hat{\beta} = \begin{cases} (X'X)^{-1}X'\underline{Y} \text{ ; if } (X'X) \text{ is non-singular,} \\ (X'X)^{-}X'\underline{Y} \text{ ; if } (X'X) \text{ is singular.} \end{cases} \tag{7}$$

From the equations (4) and (7) it is observed that the maximum likelihood estimates and least square estimators of β are same.

For the sake of completeness and to update the list of the methods of fitting regression models, some of the important alternative methods of regression are mentioned below.

1. Weighted least squares (WLS) method.
2. Generalised least squares (GLS) method.
3. Least absolute deviations (LAD) method.
4. Bayesian regression.
5. Non parametric regression.
6. Ridge regression.
7. Principal component method.

Methods of obtaining M-estimates, R-estimates, L-estimates, Shrinkage estimates and High breakdown point estimates are also useful to estimate the model parameters.

Of the above mentioned methods, least absolute deviation method is the oldest method of regression. This method was introduced by Boscovich in 1757. In the recent literature this method is receiving great attention from the researchers due to the availability of the computational facilities. In this method, estimates of the model parameters β are obtained so as to minimise the sum of absolute values of the residuals; i.e. to minimise the sum,

$$\sum_i |Y_i - \sum_j X_{ij}\beta_j|,$$

with $X_{io}=1$, for all $i=1,2,\ldots,n$.

Due to the complexity in calculations and limitations of computation facilities available, this method could not become popular. Today, vast computation facilities are available, so that one can estimates the model parameters by using any of the above method. Later, in the chapter on quasi likelihood, we discuss in detail, least absolute deviation method for estimating model parameters in quasi likelihood model.

Below we summarise the situations where the alternative methods performs better than least square method.

Main disadvantage of using least square method is that, when population errors are having some non normal distribution, least square estimates are less efficient. In such situations least absolute deviation method, M-regression and non parametric regression gives more efficient estimates. These three methods give more accurate estimates as compared to least square estimates when some outliers are present in the data set. This is because, these three methods resist in much better way to the influence of outliers, as compared to least square method.

When we have the previous knowledge about the type of data to be analysed, Bayesian method gives better estimates than least square estimates. Further, when the population errors are having normal distribution, ridge estimates are more accurate than least square estimates.

Here, more discussion on the alternative regression methods may be a divertion from the purpose of the chapter. Hence we stop this discussion. One can refer Birkes (1993) for further details.

Below we illustrate with the help of numerical example, how the model parameters β in case of classical linear model can be estimated.

Example 2.1 (cont.):- In this example we have,

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -3.0 & -1.8 & -10.0 & 0.7 & -5.1 & -6.3 & -15.5 & -4.2 & -8.8 & -2.3 \\ 15.3 & 16.4 & 41.2 & 9.7 & 19.3 & 11.4 & 5.9 & 24.3 & 14.7 & 16.1 \end{bmatrix},$$

$\underline{Y}$=[14.96 14.10 23.76 13.20 16.60 16.79 21.63 16.25 20.98 16.66] ',

$$(X'X)^{-1} = \begin{bmatrix} 0.5794 & 0.0250 & -0.0194 \\ 0.0250 & 0.0050 & 0.0002 \\ 0.0050 & 0.0002 & 0.0012 \end{bmatrix}.$$

Hence

$$\hat{\beta} = (X'X)^{-1}X'\underline{Y} = (11.9339 \quad -0.6265 \quad 0.1298)'. \qquad (8)$$

In further discussion we assume that $(X'X)$ is non singular. Now to obtain the estimator of intercept β_0 and the parameters $\beta^* = (\beta_1, \beta_2, \ldots, \beta_k)'$ separately, we state the following lemma. (see e.g. Searle (1971).

Lemma 2.1 : If M is a non-singular matrix of order n (say), such that

$$M = \begin{bmatrix} X' \\ Z' \end{bmatrix} [\; X' \quad Z' \;] = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z \end{bmatrix}$$

$$= \begin{bmatrix} A & B \\ B' & D \end{bmatrix}, \text{ (say).}$$

Then the inverse of the matrix M can be written as

$$M^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B\,R\,B'A^{-1} & -A^{-1}B\,R \\ -R'B'A^{-1} & R \end{bmatrix}, \qquad (9)$$

provided the matrices A and $R = (D - B'A^{-1}B)^{-1}$ are non singular.

Theorem 2.2 : In classical linear models the least square estimators $\hat{\beta}$ of β can be given in the following form.

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix} = \begin{bmatrix} \bar{Y} - \bar{\underline{X}}'\underline{\hat{\beta}}^* \\ (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\underline{y} \end{bmatrix}, \qquad (10)$$

where

(i) $\bar{\underline{X}}' = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_K)$,

(ii) $\bar{X}_j = (1/n)\sum_i (X_{ij})$,

(iii) $\bar{Y} = (1/n)\sum_i (Y_i)$,

(iv) $y_i = Y_i - \bar{Y}$,

(v) $\mathcal{X} = X_1 - E_{n1}\bar{\underline{X}}'$

(vi) E_{mn} is the matrix of order $(m \times n)$ having all elements equal to unity.

Proof : We know that,

$$\hat{\beta} = (X'X)^{-1}X'\underline{Y},$$

$$\hat{\beta} = \left\{ \begin{bmatrix} E_{1n} \\ X_1' \end{bmatrix} [\; E_{n1} \quad X_1 \;] \right\} \begin{bmatrix} E_{1n} \\ X_1' \end{bmatrix} \underline{Y}$$

$$= \begin{bmatrix} n & n\bar{\underline{X}} \\ n\bar{\underline{X}}' & X_1'X_1 \end{bmatrix}^{-1} \begin{bmatrix} n\bar{Y} \\ X_1'\underline{Y} \end{bmatrix}. \qquad (11)$$

Now, substitute $A = n$, $B = n\bar{\underline{X}}'$ and $D = X_1'X_1$ in *lemma* (2.1). Now we compute different matrices in the expression of M^{-1}.

$$R^{-1} = D - B'A^{-1}B,$$

$$= X_1'X_1 - n \cdot \bar{\underline{X}}\, \bar{\underline{X}}'.$$

Therefore,

$$R = (X_1'X_1 - n\, \bar{\underline{X}}\, \bar{\underline{X}}')^{-1}. \qquad (12)$$

Consider,

$$A^{-1}B\; R\; B'A^{-1} = [\, \bar{\underline{X}}'(X_1'X_1 - n\, \bar{\underline{X}}\, \bar{\underline{X}}')^{-1}\bar{\underline{X}} \,]^{-1},$$

$$= \bar{\underline{X}}'R\, \bar{\underline{X}}\,, \qquad (13)$$

and

$$A^{-1}B\; R = \bar{\underline{X}}'R. \qquad (14)$$

Hence by using *lemma* (2.1) and the equations (12), (13), (14), the equation (11) can be written as

$$\hat{\beta} = \begin{bmatrix} (1/n)+\bar{\underline{X}}'R\,\bar{\underline{X}} & -\bar{\underline{X}}'R \\ -R\,\bar{\underline{X}} & R \end{bmatrix} \begin{bmatrix} n\bar{Y} \\ X_1'\underline{Y} \end{bmatrix},$$

$$\hat{\beta} = \begin{bmatrix} \bar{Y} - \bar{\underline{X}}'R(X_1'\underline{Y} - n\,\bar{\underline{X}}\,\bar{Y}) \\ R(X_1'\underline{Y} - n\,\bar{\underline{X}}\,\bar{Y}) \end{bmatrix}. \qquad (15)$$

Note that,

$$\mathcal{X}'\underline{y} = (X_1'\underline{Y} - n\,\bar{\underline{X}}\,\bar{Y}).$$

Thus the equation (15) can be written as

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}^* \end{bmatrix} = \begin{bmatrix} \bar{Y} - \bar{\underline{X}}'\hat{\beta}^* \\ (\mathcal{X}'\mathcal{X})^{-1}\mathcal{X}'\underline{y} \end{bmatrix}. \qquad (16)$$

Remark:- The form of $\hat{\beta}$ as in equation (16) does not apply for the no-intercept classical linear model, because in that case β can not be partioned as above. Hence in case of no-intercept model, we should use the equation (7) to estimate the parameters in the model. Again we continue with the example (2.1) to illustrate how the above computations can be carried out.

Example 2.1 (cont.):- For the example (2.1) we have

$$X_1' = \begin{bmatrix} -3.0 & -1.6 & -10.0 & 0.7 & -5.1 & -6.3 & -15.5 & -4.2 & -8.6 & -2.3 \\ 15.3 & 16.4 & 41.2 & 9.7 & 19.3 & 11.4 & 5.9 & 24.3 & 14.7 & 16.1 \end{bmatrix},$$

and

$$(X'X)^{-1} = \begin{bmatrix} 0.0050 & 0.0002 \\ 0.0002 & 0.0012 \end{bmatrix}.$$

By applying theorem (2.2) we have

$$\hat{\beta}^* = (X'X)^{-1}X'\underline{Y}$$

$$= (-0.6285 \quad 0.1298)', \qquad (17)$$

and

$$\hat{\beta}_0 = \bar{Y} - \bar{X}'\hat{\beta}^*$$

$$= 17.735 - [-5.63 \quad 17.43] \begin{bmatrix} -0.6285 \\ 0.1298 \end{bmatrix}, \qquad (18)$$

$$= 11.9339.$$

Hence

$$\hat{\beta} = (11.9339 \quad -0.6285 \quad 0.1298)'. \qquad (19)$$

Now before going to the model checking, we discuss the important properties of the estimators of model parameters.

2.3.3 <u>Properties of the estimators</u> :

Following are some important properties of the estimators. These properties can be verified easily.

(i) $\hat{\beta}$ is an unbiased estimator of β;

(ii) $\hat{\beta}$ is the best linear unbiased estimator (BLUE) as well as m.l.e.of β;

(iii) $\hat{\beta}$ is also least square estimator of β, and due to first property it is minimum variance unbiased estimator (MVUE) of β;

(iv) variance of $\hat{\beta}$ is

$$Var\ (\hat{\beta}) = (X'X)^{-1}\sigma^2;$$

(v) In case of no-intercept classical linear model variance of $\hat{\beta}$ can be written in the form

$$Var\ (\hat{\beta}) = \begin{bmatrix} (1/n)+\bar{\underline{X}}'R\ \bar{\underline{X}} & -\bar{\underline{X}}'R \\ -R\ \bar{\underline{X}} & R \end{bmatrix}\sigma^2;$$

(vi) $Cov\ (\hat{\beta}_o,\ \hat{\beta}^*) = -\ \{\ Var\ (\hat{\beta}^*)\ \}.$

Further we know,

(i) the residual sum of squares is given by

$$S_E^2 = SSE = \underline{Y}'\underline{Y} - \hat{\beta}'X'\underline{Y}\ ;$$

(ii) the residual mean sum of squares (s_E^2) is given by

$$s_E^2 = (SSE)/(n-k-1)$$ and is an unbiased estimator of the error variance σ^2.

After fitting classical linear model to the data, we try to get a simpler classical linear model by deleting those covariates in the fitted model indicating insignificant effect on the response variate. For this purpose it is necessary to test different hypotheses about the parameters β in the model.

2.4 Testing of hypotheses :

For testing different hypotheses about the parameters $\hat{\beta}$ or about the linear combinations of β, it is necessary to break up total sum of squares (TSS) into the sum of squares (SS) due to different components of the systematic effects and the SS due to the random effects. Then by comparing separately the effect due to each component of the systematic effects with the random effects one can test the different hypotheses.

2.4.1 Partitioning of the TSS :

The raw TSS is defined as

$$TSS = \underline{Y}'\underline{Y}, \tag{1}$$

and the SSE is given by

$$SSE = \underline{Y}'\underline{Y} - \hat{\beta}'X'\underline{Y}. \tag{2}$$

The quantity $\hat{\beta}'X'\underline{Y}$ is known as regression sum of squares (SSR).

Suppose that the model has no independent variate. This model can be written as

$$\underline{Y} = \beta_0 E_{n1} + \underline{e} . \tag{3}$$

Then $\hat{\beta}_0 = \bar{\underline{Y}}$ and SSR (S_R^2) = n $\bar{\underline{Y}}^2$. The SSR when there is no stimulus variate is known as the correction factor (c.f.). Let $TSS_{(m)}$ and $SSR_{(m)}$ be the corrected TSS and the corrected SSR respectively. Therefore,

$$TSS_{(m)} = TSS - c.f. \tag{4}$$

and

$$SSR_{(m)} = SSR - c.f. \tag{5}$$

2.4.2 Distributional properties of the model components :

Following are some important distributional properties of various quantities related to the classical linear model.

(i) the vector of estimators $\hat{\beta}$ has $N_{(k+1)}(\beta, (X'X)^{-1}\sigma^2)$ distribution ;

(ii) $\underline{y} = (\underline{Y} - E_{1n}\bar{Y})$ has $N_n(\,0,\sigma^2 I_n)$ distribution ;

(iii) the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are independently distributed;

(iv) (SSE/σ^2) has central chi-square distribution with $(n-k-1)$ degrees of freedom (d.f.) ;

(v the statistics SSR , $SSR_{(m)}$, c.f. have non-central chi square distribution, and each of them is independent of SSE.

Proofs of the first two properties are obvious. Now we give detailed proofs of the last three properties. Results needed for proving the last three properties are quoted in the following lemmas (refer e.g. Searle (1971)).

Lemma 2.3 : If $\underline{X}$ has $N_n(\,\mu, V\,)$ distribution, then $\underline{X}'A\,\underline{X}$ and $B\ \underline{X}$ are independently distributed iff $BVA = 0$, provided $\underline{X}'A\,\underline{X}$ is non central chi square variate and AVB is defined.

Lemma 2.4 : If $\underline{X}$ has $N_n(\mu,V)$ distribution, then the quadratic forms $\underline{X}'A\,\underline{X}$ and $\underline{X}'B\,\underline{X}$ are independently distributed iff $BVA = 0$ and $AVB = 0$.

Lemma 2.5 : If $\underline{X}$ has $N_n(\mu,V)$ distribution, then $\underline{X}'A\ \underline{X}$ has non central chi square distribution with r d.f. and non centrality parameter $\{\,(1/2)\ (\mu'A\,\mu)\,\}$ iff AV is idempotent. Here r is the rank of the matrix A.

Theorem 2.6 : The estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are independently distributed.

Proof : We Know,

$$\hat{\beta} = (X'X)^{-1}X'\underline{Y}, \tag{6}$$

and note that $\hat{\sigma}^2$ can be written as

$$\hat{\sigma}^2 = \underline{Y}'\{(I - X(X'X)^{-1}X')/(n-k-1)\}\underline{Y}. \tag{7}$$

As $\underline{Y}$ has $N_n(\,X\beta,\sigma^2 I_n)$ distribution, by *lemma* (2.3), the estimators $\hat{\beta}$ and $\hat{\sigma}^2$ are independently distributed iff

$$[(X'X)^{-1}X'][(I - X(X'X)^{-1}X')] = 0. \qquad (8)$$

Consider

$$[(X'X)^{-1}X'][(I - X(X'X)^{-1}X')] = (X'X)^{-1}X' - (X'X)^{-1}X'$$
$$= 0.$$

This shows that condition (8) is valid. This completes the proof of the theorem.

Theorem 2.7 : (SSE/σ^2) has a central chi squqre distribution with $(n-k-1)$ d.f.

Proof : We know,

$$SSE = \underline{Y}'\{(I - X(X'X)^{-1}X')\}\underline{Y}. \qquad (9)$$

Since $(I - X(X'X)^{-1}X')$ is idempotent matrix and $\underline{Y}$ has $N_n(X\beta, \sigma^2 I_n)$ distribution, by *lemma* (2.5), (SSE/σ^2) has a non central chi square distribution with $\{(1/2\sigma^2)\beta'X'(I - X(X'X)^{-1}X')X\beta\}$ as a non centrality parameter and $(n-k-1)$ d.f.

Further since $(I - X(X'X)^{-1}X')X = 0$ the non centrality parameter becomes zero. Thus (SSE/σ^2) has central chi squqre distribution with $(n-k-1)$ d.f.

Theorem 2.8 : (SSR/σ^2) and SSE are independently distributed and (SSR/σ^2) has non central chi square distribution.

Proof : We know,

$$SSR = \hat{\beta}'X'\underline{Y}, \qquad (10)$$

$$= \underline{Y}'X(X'X)^{-1}X'\underline{Y}. \qquad (11)$$

Note that the matrix $X(X'X)^{-1}X'$ is idempotent matrix and $\underline{Y}$ has $N_n(X\beta, \sigma^2 I_n)$ distribution. Hence by *lemma* (2.5), (SSR/σ^2) has non central chi square distribution with (k+1) d.f. and non centrality parameter $\{(1/2\sigma^2)\beta' X' X\beta\}$.

Further, from equations (9), (11) and lemma (2.4) it is clear that SSR and SSE are independently distributed.

Remarks : As in theorem (2.8) it can be shown that,

(i) (SSM/σ^2) has non central chi squqre distribution with 1 d.f. and $[(E_{1n}X\underline{\beta})^2/2n\sigma^2]$ as non centrality parameter, and it is independent of SSE.

(ii) $(SSR_{(m)}/\sigma^2)$ has non central chi squqre distribution with k d.f. and $[(\beta^{*}{}'(\mathcal{X}'\mathcal{X})\underline{\beta}^{*})/2\sigma^2]$ as non centrality parameter, and it is independent of SSE.

(iii) $(TSS_{(m)}/\sigma^2)$ has non central chi squqre distribution with (n-1) d.f. and $[(\underline{\beta}'(X'X)\underline{\beta})/2\sigma^2]$ as non centrality parameter, and it is independent of SSE.

The theorems (2.7), (2.8) and the remarks are useful in developing test statistics for testing hypotheses about the parameters β or β^*.

In classical linear models, some of the interesting hypothesis testing problems are,

(i) H_{01}: $\underline{\beta} = \underline{0}$ against H_{A1}: $\underline{\beta} \neq \underline{0}$, (12)

(ii) H_{02}: $\underline{\beta} = \underline{m}$ (specified) against H_{A2}: $\underline{\beta} \neq \underline{m}$, (13)

(iii) H_{03}: $\underline{\lambda}'\underline{\beta} = c$ (specified) against H_{A3}: $\underline{\lambda}'\underline{\beta} \neq c$, (14)

(iv) H_{04}: $\underline{\beta}_q = \underline{0}$ against H_{A4}: $\underline{\beta}_q \neq \underline{0}$, $(q < (k+1))$. (15)

Though the test stastics for these hypothesis testing problems are different, all these are special cases of the general hypothesis testing problem,

$$H_o: D'\beta = \underline{m} \text{ against } H_A: D'\beta \neq \underline{m}, \tag{16}$$

where

(i) D is a matrix of order [(k+1) X (s)] with full column rank ;

(ii) $\underline{m}$ is a column vector of known constants of order s ;

Hence first we obtain the test statistic for the general hypothesis testing problem described in equation (16) and then we show that the hypothesis testing problems in the equations (12) to (15) are special cases of the general hypothesis testing problem.

2.4.3.: <u>The general hypothesis testing problem</u> :

Here we develope a test for the hypothesis testing problem defined in (16).

As $\hat{\beta}$ has $N_{(k+1)}(\beta, (X'X)^{-1}\sigma^2)$ distribution, it is clear that,

$$(D'\hat{\beta} - \underline{m}) \text{ has } N_{(s)}[(D'\hat{\beta} - \underline{m}), D'(X'X)^{-1}D\sigma^2]. \tag{17}$$

Let

$$(S/\sigma^2) = (D'\hat{\beta} - \underline{m})'[D'(X'X)^{-1}D]^{-1}(D'\hat{\beta} - \underline{m})]/\sigma^2. \tag{18}$$

Since D has full column rank, from *lemma* (2.5) it is clear that(S/σ^2) has non central chi square distribution with s d.f. and non centrality parameter $(D'\hat{\beta}-\underline{m})'[D'(X'X)^{-1}D]^{-1}(D'\hat{\beta}-\underline{m})]/2\sigma^2$. By substituting the expression for $\hat{\beta}$ from equation (6), S can be written as

$$S = [\underline{Y}-XD(D'D)^{-1}\underline{m}]'K\ [\underline{Y}-XD(D'D)^{-1}\underline{m}], \tag{19}$$

where

$$K = \{X(X'X)^{-1}D[D'(X'X)^{-1}D]^{-1}D'(X'X)^{-1}D. \tag{20}$$

Now since

$$X'(I - X(X'X)^{-1}X') = 0, \tag{21}$$

and

$$(I - X(X'X)^{-1}X')X' = 0, \tag{22}$$

the SSE given in equation (9) can be written as

$$SSE = [\underline{Y}-XD(D'D)^{-1}\underline{m}]'\{(I - X(X'X)^{-1}X')\}[\underline{Y}-XD(D'D)^{-1}\underline{m}]. \tag{23}$$

The expressions (19) and (23) gives S and SSE as the quadratic forms in the same vector $[\underline{Y}-XD(D'D)^{-1}\underline{m}]$. Note that,

$$\mathrm{Var}[\underline{Y}-XD(D'D)^{-1}\underline{m}] = \mathrm{Var}(\underline{Y}) = \sigma^2 I_n, \tag{24}$$

and

$$K(\sigma^2 I_n)[I-X(X'X)^{-1}X'] = [I-X(X'X)^{-1}X'](\sigma^2 I_n)K = 0. \tag{25}$$

Hence according to *lemma* (2.4) S and SSE are independently distributed. Now by definition of non central F statistic we have the following result.

$$F(H) = \left\{\frac{(S/s)}{[SSE/(n-\rho(X))]}\right\} \text{ has } F'[s, n-\rho(X), (1/2)(D'\beta\ -\underline{m})'(D'\beta\ -\underline{m})]$$

distribution.

Under H_0 as defined in equation (16), the non centrality parameter becomes zero. Hence for the hypothesis testing problem (16), test statistic F(H) has central F distribution with $(s, n-\rho(X))$ d.f. The test procedure is same as in the usual F test.

The tests for four hypothesis testing problems can be derived by using the test obtained for the general hypothesis testing problem. These tests are as mentioned in the following table.

TABLE 2.2

Hypothesis testing problem	Replacement for D & $\underline{m}$	Expression for S	d.f. (S/σ^2)
12	$D = I_{(k+1)}$ $\underline{m} = \underline{0}$	S_1	$s_1 = (k+1)$
13	$D = I_{(k+1)}$.	S_2	$s_2 = (k+1)$
14	$D = \underline{\lambda}$	S_3	$s_3 = 1$, provided $\underline{\lambda} \neq \underline{0}$
15	$D' = [I_q \;\; 0]$ $\underline{m} = \underline{0}$	S_4	$s_4 = q$

where,

(i) $S_1 = \underline{Y}'\{X(X'X)^{-1}X'\}\underline{Y}$,

(ii) $S_2 = [\underline{Y}-X\underline{m}]'\{X(X'X)^{-1}X')\}[\underline{Y}-X\underline{m}]$,

(iii) $S_3 = \left\{\underline{Y}-\left[\dfrac{C}{(\sum_i \lambda_i^2)}\right]X\underline{\lambda}\right\}' \left[\dfrac{X(X'X)^{-1}\lambda\lambda'(X'X)^{-1}X'}{(\lambda'(X'X)^{-1}\lambda)}\right]\left\{\underline{Y}-\left[\dfrac{C}{(\sum_i \lambda_i^2)}\right]X\underline{\lambda}\right\}$,

(iv) $S_4 = \underline{Y}'X \begin{bmatrix} T_{qq} & T_{qp} \\ T_{pq} & T_{pq}\, T_{qq}^{-1}\, T_{qp} \end{bmatrix} X'\underline{Y}$,

(v) $(X'X)^{-1} = \begin{bmatrix} T_{qq} & T_{qp} \\ T_{pq} & T_{pp} \end{bmatrix}$.

As in the general hypothesis testing problem, test statistics for testing H_{0i} against H_{Ai} defined in the equations (12) to (15) are

$$F_i(H) = \frac{(S_i/s_i)}{(SSE/[n-\rho(X)])}, \quad i=1,2,3,4,$$

and has central F distribution with $(s_i, n-\rho(X))$ d.f. The test procedure is same as in usual F test.

Sometimes it may happen that the model fits well to the data, but the assumptions made, turn out to be invalid. Hence the final stage in analysing data by fitting the model is to check appropriateness of the fitted model. We discuss below the part of model checking based on the residual analysis.

2.5 : Residual analysis :

The residual analysis is necessary in every model fitting problem. While fitting the model different assumptions are made about error components in the model. for example, in classical linear models the assumption is that the error components are independently identically distributed $N(0,\sigma^2)$ variates. If the particular model fits well to the data, residuals must indicate that the assumptions made are not invalid. Hence after examining residuals we must be able to conclude that the assumptions made are either invalid or not necessarily invalid.

Draper & Smith (1981) explained different methods of residual analysis for checking appropriateness of the fitted model, some of them are,

(I) graphical metod;

(II) statistical method;

(III) by studing correlation among the residuals;

(IV) outliers;

(V) serial correlation in residuals.

2.5.1 : Graphical method :

This is the easiest method and if the fitted model is not proper, it will reveal invalidity of the assumptions. Different ways of plotting the residuals are

(i) overall,

(ii) in time sequence (if the order is known),

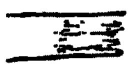
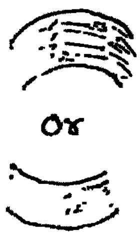
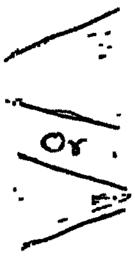
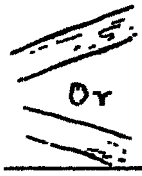
(iii) against the fitted values (Y_i),

(iv) against the values of stimulus variates.

Overall plot : This graph is plotted with residuals on the horizontal axis. In classical linear model if the fitted model is perfect, values of the residuals should make the impression that they have come from $N(0,\sigma^2)$ distribution. To see this the normal density curve is plotted and is partitioned into n equal parts. If each partition has one plotted point, the fitting is perfect.

In the remaining three plots residuals are taken along vertical axis, and the other factor along horizontal axis. Then plotting of residuals outputs different types of bands. These bands along with the conclusions to be drawn are tabulated below.

TABLE 2.3

Sr No	Band	Time sequence plot	Plot against fitted values	Plot against values of stimulus variate
1.		Time effect is not affecting the data and fitted model may be correct.	Fitted model may be correct.	Fitted model may be correct.
2.	 Or	Linear& quadratic terms in time should have been included in the model.	The fitted model is inadequate & either extra term is needed in the model or some transformation on the response variate is needed.	Either the extra quadratic term in X_j is needed or transformation on Y is needed.
3.	 Or	Variance changes with time. The weighted least square analysis must be used.	The variance changes. Hence weighted least square method or variance stabilising transformation is necessary.	The variance is not constant & hence weighted least square method or variance stabilising transformation is needed.
4.	 Or	The linear term in time required in the model.	The systematic effects are not completely removed.	The linear effect of X_j has not removed copletely.

Remark: We can have bands which are the combinations of above types of bands. The interpretation can be given accordingly.

2.5.2 Statistical methods :

Graphical method is the visual technique for checking validity of the assumptions. Consider the plot against the fitted values. There are three types of descripancies (2) to (4) as in table (2.3). Each of these descipancies can be measured by a proper statistic as follows. Define,

$$T^{*}_{pq} = \sum_{i} e_i^{\,p}\, \hat{Y}_i^{\,q} \,. \qquad (1)$$

Then, measures for descipancies of the types (2), (3) and (4) are respectively T_{12}, T_{21} and T_{11}.

2.5.3 <u>Correlation among the residuals</u> :

While fitting the intercept classical linear model with k stimulus variates, we are estimating (k+1) parameters from n observations so that the residuals can not be independent. If the model is as given in the equation (2.3-2) then we have,

$$\underline{e} = \underline{Y}-\hat{\underline{Y}} = \underline{Y} - X\hat{\beta}$$

$$= (I - X(X'X)^{-1}X')\underline{Y},$$

and,

$$\mathrm{Var}(\underline{e}) = \sigma^2 (I - X(X'X)^{-1}X'),$$

Thus correlation coefficient between e_i and e_j depends only on X.

The important question is 'do these correlations indicate the failure of the assumption of independence ?' Ancombe & Tukey (1963) stated that with four or more rows and columns the effect of correlation between residuals on the graphical method is negligible. In general situation this effect should not be considered if [(n-k-1)/n] is quite small.

2.5.4 <u>Outliers</u> :

An outlier among the residuals is far away from the rest. It is a peculiarity and indicates a data point which is not 'similar' to the remaining data. If there are outliers, then a careful examination should be carried out to find the cause for its peculiarity.

Rules have been proposed for rejecting the outliers. It is not a good technique to reject the outliers always, because sometimes the outlier is providing information which the other data points can not, due to the fact that it arises from unusual combination of circumstances which may be of vital interest. In such situations further investigation is necessary. As a general rule, outlier should be rejected only if it has been found that it has occured due to error in recording the observation or in carrying out the experiment.

2.5.5 Serial correlation in residuals :

In classical linear models it is assumed that the residuals are pairwise independent, but it is not at all true. There are many ways in which the errors may be correlated. A common way is they may be serially correlated, i.e. the residuals which are apart by s steps are having same value of correlation coefficient. This type of serial correlation may be used for residual analysis.

As the residual analysis is a vital part of the model fitting problem, we have discussed it in short. (For more details one can refer Draper & Smith (1981)). We illustrate the residual analysis in the concluding chapter of the dissertation.

Until now we have discussed theory related to classical linear models. Due to the availability of the software packages many times classical linear models are fitted to the different types of data. The question that may arise is 'Is the fitting of classical linear model appropriate in the situations ?' Answer to this question is no. This is because classical linear model is not proper to the data with non constant variance. Below is the discussion explaining why classical linear model is not applicable to such type of data, followed by illustration.

2.6 Limitations of the theory of fitting classical linear models:

When the response variate is having some distribution with non constant variance, there are number of drawbacks of fitting a classical linear model to the data.

First is about variance of the response variate Y. For example, if we are dealing with binary data with the i^{th} observation Y_i on the response variate Y has distribution such that $Y_i^*=m_i^*Y_i$ has binomial distribution $B(m_i^*,p_i)$ so that variance of Y_i is $p_i(1-p_i)/n_i$. Thus variance depends on the number of successes in the i^{th} sample, though we assume sample sizes are equal. If m_i^*'s are approximately equal, the variance stabilising transformation $Sin^{-1}[(p_i)^{1/2}]$ can be used. This transformation is known as angular transformation.

Secondly, since the response variate is not normal, the distribution theory associated with fitting of classical linear model is not valid. For large sample sizes as most of the distributions tends to the normal distribution. This drawback is not much serious.

The final drawback is more serious and is about the fitted values of response variate. For example when Y_i has the distribution such that $Y_i^*=m_i^*Y_i$ has binomial distribution $B(m_i^*,p_i)$ then it is about the fitted values p_i. In classical linear model there are no restrictions on the estimated values of the parameters. Hence estimated values of the response variate corresponding to different combinations of the values of explanatory variates lie any where in the range $(-\infty,\infty)$. As the fitted values p_i of p_i are obtained from the expression $\underline{p} = X\hat{\beta}$, there is no gurantee that they should take values in (0,1). This fact can be illustrated with the help of example in better way.

Example 2.2 : This example is taken from Collett (1991). Smith (1932) studied the protective effect of a particular serum on pneumococcus. Pneumococcus is the bacterium causing the disease pneumonia. Each of forty mice was injected with a combination of infecting dose of pneumococcus, and one of the five doses of anti-pneumococcus serum. For all the mice which died during the seven day period after injection, a blood smear taken from the heart was examined. Thus the variate Y' is the death from pneumonia, within seven days after injuction. The following table gives the number of deaths from pneumonia, among the different samples of forty mice each, exposed to the five different doses of serum.

TABLE 2.4

Dose of Serum	0.0028	0.0056	0.0112	0.0225	0.0450
Number of deaths out of 40 mice	35	21	9	6	1

One may be interested in finding relationship between the probability of death p_i and dose of serum (d_i, say). Fitting classical linear model to the data with response variate Y gives

$$p_i = 0.64 - 16.08\ d_i . \qquad (1)$$

The equatin (1) gives the fitted probability for a mouse injucted 0.045 cc of the serum is -0.0836. Thus the classical linear model is not acceptable. So it necessary to fit some other type of model to the response variate having non constant variance.

In the above example it is shown that classical linear model is not suitable in many practical situations. This happens because the response variate on its original scale is not having constant variance and the systematic effects are not linearly related with mean of the response variate.

One of the possible way of analysing such type of data is to make transformations on the response variable. Box & Cox (1964) discussed an analysis based on transformations. By the term 'data transformation', we mean to change the original data set $\underline{Y}$ to the new data set $\psi(\underline{Y})$ (, say) through the functional form of $\psi(.)$. Now we discuss how to make a data transformation.

Suppose the response variate Y is not having normal distribution. Then find a monotonic function $\psi(.)$ such that $\psi(Y)$ is approximately normally distributed with mean $\mu^{*}=\underline{X}'\beta$ and constant variance σ^{*2}. If such type of function exists, the required data transformation is from Y to $\psi(Y)$. Hence the approximate density function of Y becomes

$$f(y;\mu^{*},\sigma^{*2}) = \left\{\frac{|\psi'(y)|}{(2\Pi\sigma^{*2})^{(1/2)}}\right\} \exp(-1/2\sigma^{*2})[\psi(Y)-\mu^{*}] \quad (2)$$

By using the approximate density of Y given in the above equation (2), one can obtain maximum likelihood estimates of β as usual. One such family of transformations suggested by Box & Cox (1964) is $\psi(Y) = Y^{\theta}$. One can use this method of transformation, but oftenly it happens that the trnsformation giving normality do not give linearity of the systematic effects with the mean of response variate. Thus we require two transformations, one for 'linearising' and the other for 'normalising'. Nelder (1966) has discussed these types of transformations.

Nelder & Pregibon (1987) pointed out several disadvantages of the data transformations. Following are the major disadvantages of analysing data by using response variable transformations.

(1) When the response variate is having discrete distribution, the range of Y is restricted. This causes range restriction on $\psi(Y)$. As the range of $\psi(Y)$ is restricted, normal approximation for $\psi(Y)$ is not suitable.

(2) It is very much difficult to find a monotonic function $\psi(.)$ giving both constant variance and linearity of systematic effects with the mean responses.

Because of these disadvantages we will not analyse the data making data transformations. An intersted person can look to Box & Cox (1964) for further details.

Obviously, an alternative approach is necessary and it is proposed by Nelder & Wedderburn (1972). They introduced a new class of models called 'generalised linear model'.

In the next chapter we discuss 'generalised linear models', which includes the models for the responses having the distribution as a member of 'one parameter natural exponential family'.