6. DATA ANALYSIS

6.1 Introduction :

In the previous chapters, theory associated with various models and the different techniques of model checking are discussed. The initiative of this chapter is to illustrate some of these techniques with the help of two numerical examples. The major problem which occures at the initial stage of data analysis through model fitting is how to select a model from the general class of models defined in chapter 1, so that conclusion drawn after analysing the data are not far away from the truth. This problem occures because, as demonstrated at the end of chapter 2, classical linear models are not appropriate in many situations. Hence, though it is easy to fit classical linear model to the data, it is not a good technique to fit it blindly. Thus selecting a suitable model becomes a very crucial part of analysing the data through model fitting.

This chapter may help the reader to set a guideline to analyse the by fitting a model to it. Here we demonstrate a part of the model checking procedure with the help of some real life data sets and one artificial data set.

6.2 Analysis corresponding to the data set-I :

Consider example (3.1) discussed in chapter 3. From fig (3.1) it is clear that given data requires some type of variance stabilising transformation. Thus before fitting the model itself, we can conclude that classical linear model is not suitable for untransformed data.

If one fits blindly the classical linear model

$$E(Y) = \beta_0 + \beta_1 X \tag{1}$$

to the data, the scatter plots given below in the fig (1) and fig

(2) indicate that variance of the response variate is not stable.

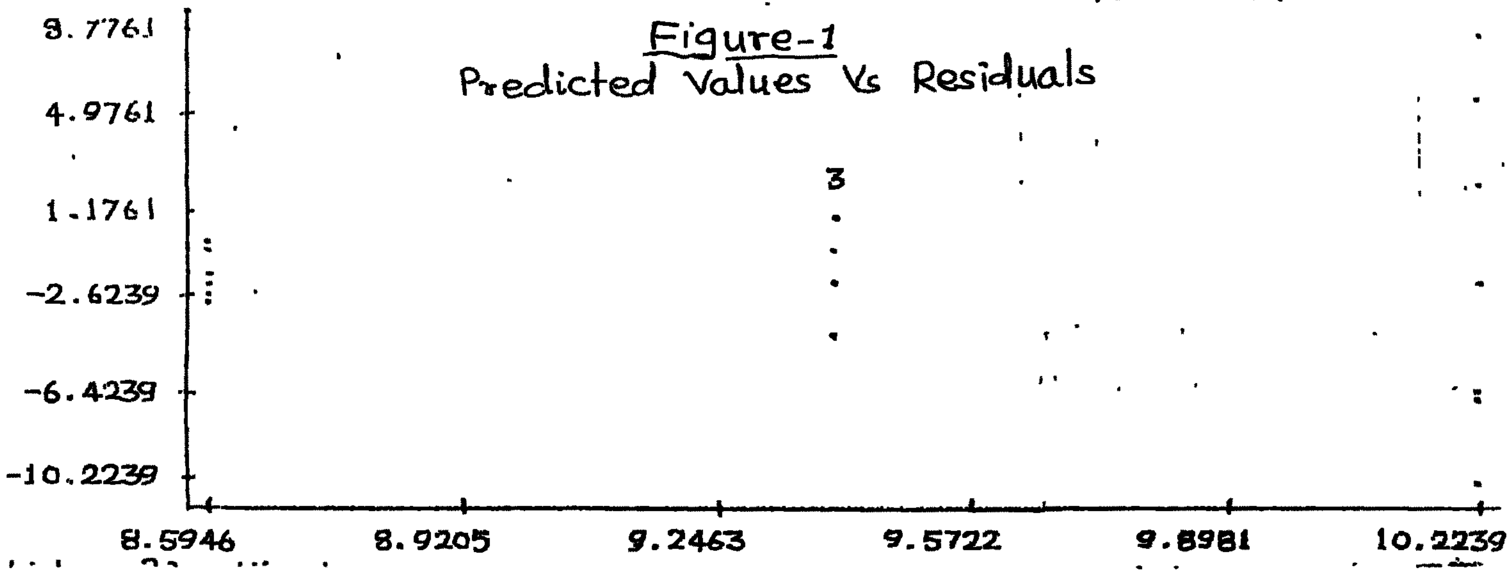


Figure-1
Predicted Values Vs Residuals

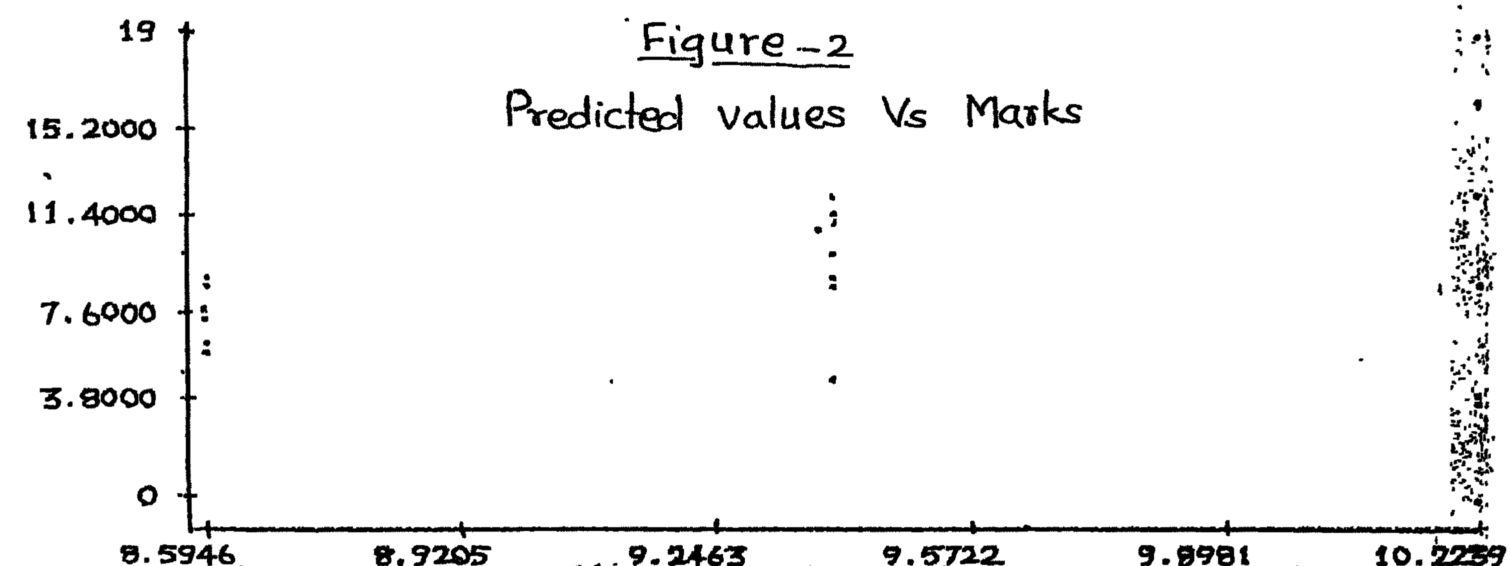


Figure-2
Predicted values Vs Marks

Hence, the model fitted is not suitable. This fact can be seen using usual pearson's chi square statistic. If the value is computed by fitting model (1), the value turns out to be 348.3545 with 18 degrees of freedom. This indicates that the fit not good. If we check the assumption of normality by using the well known chi square test of goodness of fit, it also will indicate that the data may not be from normal population. Thus classical linear model is not proper for this data.

If we observe the data carefully, it can be seen that B(20,p) distribution may fit well to the data. So we try for binomial fitting. If the linear logistic model is fitted, the value of Pearson's chi square statistic comes out to be 3,4213 with 18 degrees of freedom. This indicates that fitting of linear logistic model is far meaningful and suitable for the given data than classical linear model. One can also check for complementory log-log model, log-log model as well as probit model. Once the model is selected conclusions can be drawn easily by fitting the selected model.

6.3 Analysis corresponding to the data set-II :

Consider example (2.2) discussed at the end of second chapter. In chapter 2, it has been demonstrated that, to this data classical linear model is not suitable. This fact might have been observed before fitting the model by drawing a scatter plot of number of deaths (say, Y) against the dose of serum (say, X). This plot is givwn below in figure (3).

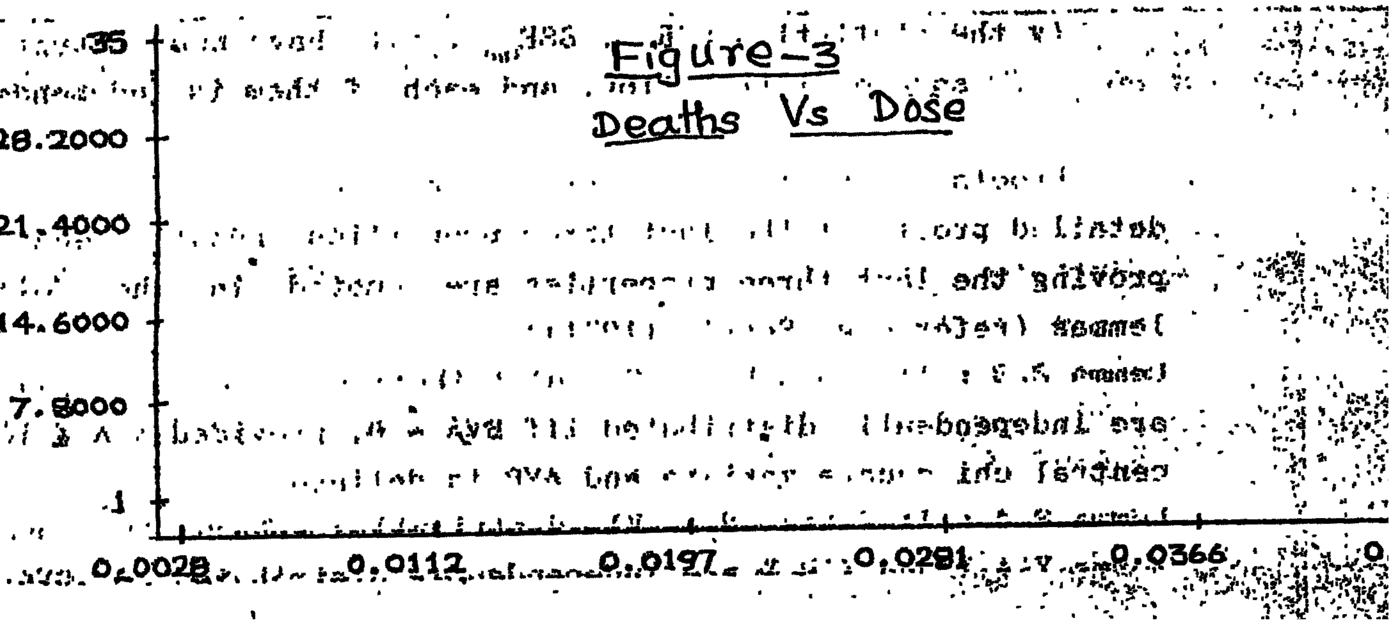
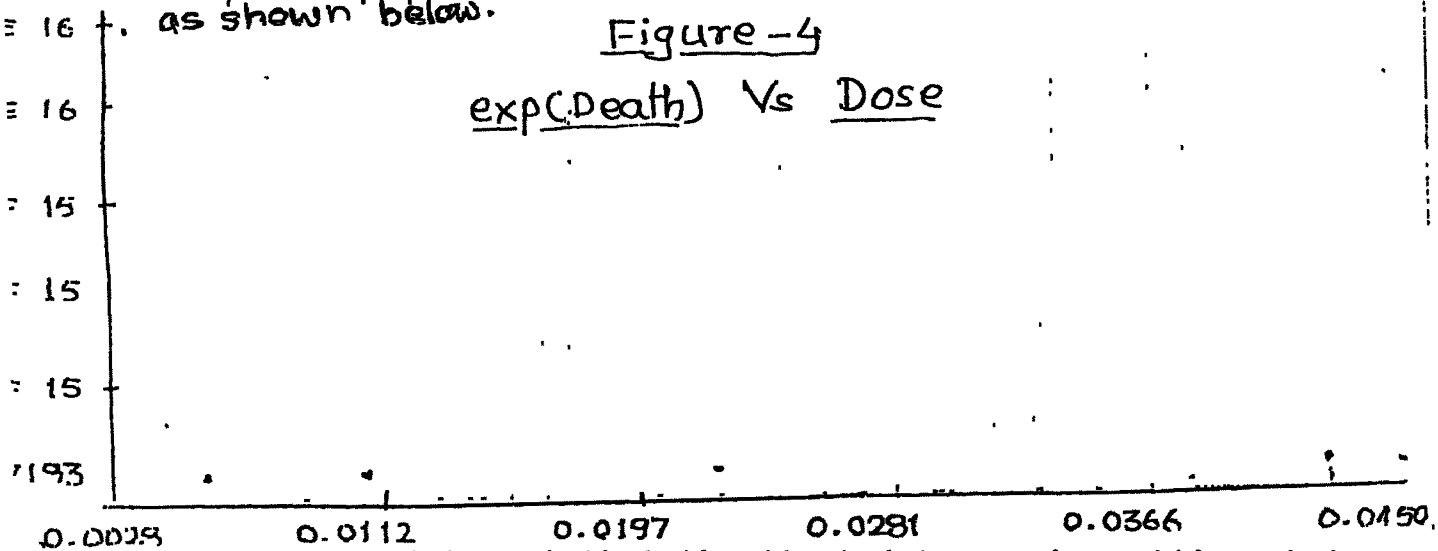


Figure-3
Deaths Vs Dose

Hence it is clear from the table (2.3) thatt either the quadratic term in dose is needed or some transformation on the response variate is needed. The graph itself in fig (3) shows that exponential transformation may be suitable. By considering thetransformation W = exp(Y), the scatter plot of W against X is as shown below.



This figure points out that the first data may be outlier, but we should test it. McCullagh & Nelder (1989), Collett (1991) and many others the techniques for identifying outliers in the data set. As we have not studied those methods one may propose the classical linear model,

$$E(W) = \beta_0 + \beta_1 X \qquad (2)$$

after checking the assumption of normality. It can be seen that the normality assumption is invalid classical linear model is not suitable. Again by looking the data carefully, it can be observed that B(40,p) distribution is appropriate for the data.

In this way one can proceed to select the model.