Demography India
Vol. 22, No.l (1993), pp. 77-88

Narender Kumar*

Regression Diagnostics with Aggregate Data

Introduction

AGGREGATE data requires additional diagnostic checks as compared to individual data in development of regression models. Aggregate data consists of variables that describe groups of individuals rather than the individuals themselves. Such variables are also termed as ecological variables (Morgenstern 1982). Describing the typology of observational study designs, Kleinbaum *el al.* (1982) write that "An ecological study, also called on aggregate or description study, is one in which the unit of analysis, is a group, most often defined geographically. Ecological analysis may involve incidence, prevalence, or mortality data, but the latter is most common because of widespread availability of such data".

The regression model could become invalid in two ways. First, the random component may not be normally distributed. Secondly, the underlying relation between dependent and independent variables may be curvilinear rather than linear. Outliers can also distort the results. A study of residuals themselves throws light on all these aspects. Therefore, the study of residuals often suffices for diagnostics in regression models with individual data. Sometimes normal probability plot is also studied which offers a method of checking the assumption of normality and makes it possible to identify those parts of data set, or individual outlying data values, which appear inconsilent with the normal distribution model.

It has been shown that the correlation between two aggregate variables is often markedly different from the corresponding individual correlation within the same population (Robinson 1950). Kleinbaum *el* a/.(1982) also report that the magnitude of association between two socio-demographic variables tends to be substantially higher in aggregate analysis (also termed as ecologic analysis) than it is in individual analysis. This problem of malticollinearity requires specific diagnostic checks in modelling with aggregate data.

When units of observation are states or countries in an epidemiological study, each unit may have its own specific inner structure. It has been illustrated (Behnken and Draper 1972) that the estimated variances of predicted values (or, equivalently, the estimated variances of the residuals) contain relevant information beyond that furnished by residual plots or studentized residuals. Cook (1977) developed a composite measure combining studentized residuals and estimated variances of residuals to help isolate critical data values that may significantly influence regression co-efficients, the parameter of interest in regression modelling. If a potentially critical observation is detected, then the examination of the effects of deleting the observation would be a natural next step. These are additional diagnostic

* Assistant Director General, Div. of N.C.D., I.C.M.R., Ansari Nagar, New Dethi-110029.

checks required in regression models with aggregate data. The present paper has developed a regression model for under five mortality, and the significance of diagnostic checks for aggregate data has been brought out.

Material and Methods

The source of data for present study is Grant (1988). The data included in this publication has been drawn/estimated for a large number of sources. Data for following basic indicators was drawn for purpose of present study on regression diagnostics with aggregate data and developing statistical model for under five mortality:

X_1	GNP per capita (1985)
X_2	Adult male literacy percentage(1985)
X_3	Adult female literacy percentage (1985)
	Percentage of infants with low birth weight (1982-85)
	Average index of food production per capita (1983-85)
X_6	Daily per capita calorie supply (1985)
X_7	Percentage of population with access to drinking water (1983-86)
	Population annual growth rate (1980-85)
	Under-five mortality rate (1986)

The data on immunisation was available only for a small number of countries and, therefore, could not be included in the study. The years (according to English calender) to which the data pertain are shown in parenthesis. Data on all the above nine variables was available for 92 countries.

Statistical Methods

The first consideration was selection of variables and to decide that in what form should the variables be used—'raw' or mathematically transformed. The 8 independent variables described in the preceding para were included in the study as data for a substantial number of countries was available only for these variables that could possibly be associated with mortality rates in young children below 5 years age. Correlation matrix of dependent *(Y)* and independent (X,) variables was constructed to get a preliminary idea of relationship between independent variables and the dependent variable, and the extent of inter-correlations among independent variables.

Often the theory or background knowledge may also suggest the existence of a statisitical relationship, but it provides no guidance as to the function form: whether it is linear, logarithmic, quadratic and so on. The log-linear model is often adapted partly because of its proportionate effect interpretation and partly because it tends to remove skewness. In the present paper the normal linear model and log-linear model (making log transformations for all variables) were compared on the basis of residual plots. The study of residual plots showed that transformation of data was not needed.

Step up multiple regression analysis was used to select variables for inclusion in the model.

The explanatory variables selected as above were further examined for multicollinearity. This was studied by regressing in turn each explanatory variable on all the rest, a high R^2 indicates a warning sign. Large values of Variance Indflation Factor *(VIF)* indicate the particular variable which is mullicollinear. The relation between *VlF(j)* and Rj^2 is given by

$$VIF\ (j) = \frac{1}{(1 - R_j{}^2)}$$

The actual influence of each observation (country) on the linear regression model was studied by computing Cook's distance *(Di)* for each country. Cook (1977) developed a measure based on confidence ellipsoids for judging the contribution of each data point to the determination of the least squares estimate of regression co-efficients. A large value of *Di* indicates that the associated data point has strong influence on the estimates of regression co-efficients. The significance of *Di* can be determined by comparing *Di* to the probability points of the central F-distribution with *k* and *n* - *k* degrees of freedom, where *k* is the number of predictor variables and *n* is the number of observations (countries).

Cook's distance *(Di)* is a composite measure of two components, namely, *ilh* studcntized residual and the ratio of variance of *ilh* predicted value to that of *ilh* residual. *Di* can be calculated as follows:

$$Di = \frac{ti^2}{k} \frac{V(\hat{Y}i)}{V(Ri)}$$

where ti is the ith studentized residual,

k is the number of predictor variables,

$V(\hat{Y}i)$ is the variance of ith predicted value $(\hat{Y}i)$,

and $V(Ri)$ is the variance of ith residual $(Ri = Yi - \hat{Y}i)$

The differences between regression co-efficients with and without the observation with largest Cook's distance were studied, and the influential observation was deleted to arrive at the model based on data containing no influential observation (that could distort the results).

The regression co-efficients of the three explanatory variables (finally included in the model) calculated from data set contanining varying number of variables and observations (countries) have been compared to assess the stability/consistency of co-efficients.

Results

The correlation matrix of dependent *(Y)* and independent *(Xi)* variables is presented in Table 1. The correlation between X_2 (male literacy) and *Xi* (female literacy) is very high (0.958). Some other correlations also seem to be high. Therefore, it was expected that step-wise multiple regression analysis would help in selecting smaller number of independent variables, but it was also important to decide that in what form the variables should be used—'raw' or mathematically transformed before proceeding with further analysis.

TABLE 1 : CORRELATION MATRIX OF DEPENDENT *(Y)* AND INDEPENDENT (Xi) VARIABLES

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	*Y*
X1	1.000	0.583	0.610	-0.503	-0.059	0.538	0.691	-0.574	-0.644
X2		1.000	0.958	-0.588	-0.032	0.723	0.788	-0.530	-0.805
X3			1.000	-0.600	-0.096	0.709	0.780	-0.571	-0.837
X4				1.000	0.032	-0.599	-0.592	0.368	0.516
X5					1.000	-0.006	-0.060	0.017	0.108
X6						1.000	0.744	-0.607	-0.678
X7							1.000	-0,574	-0.786
X8								1.000	0.614
y									1.000

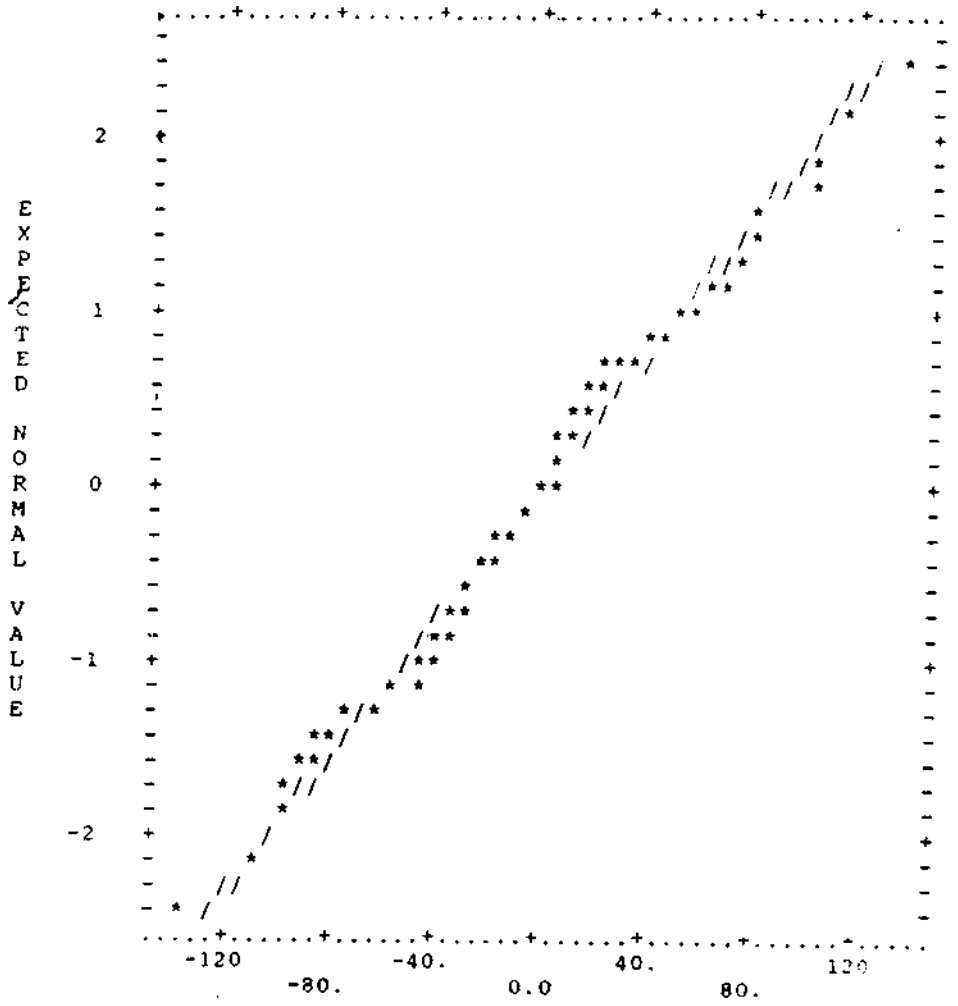

Fig. 1. Normal Probability Plot of Residuals

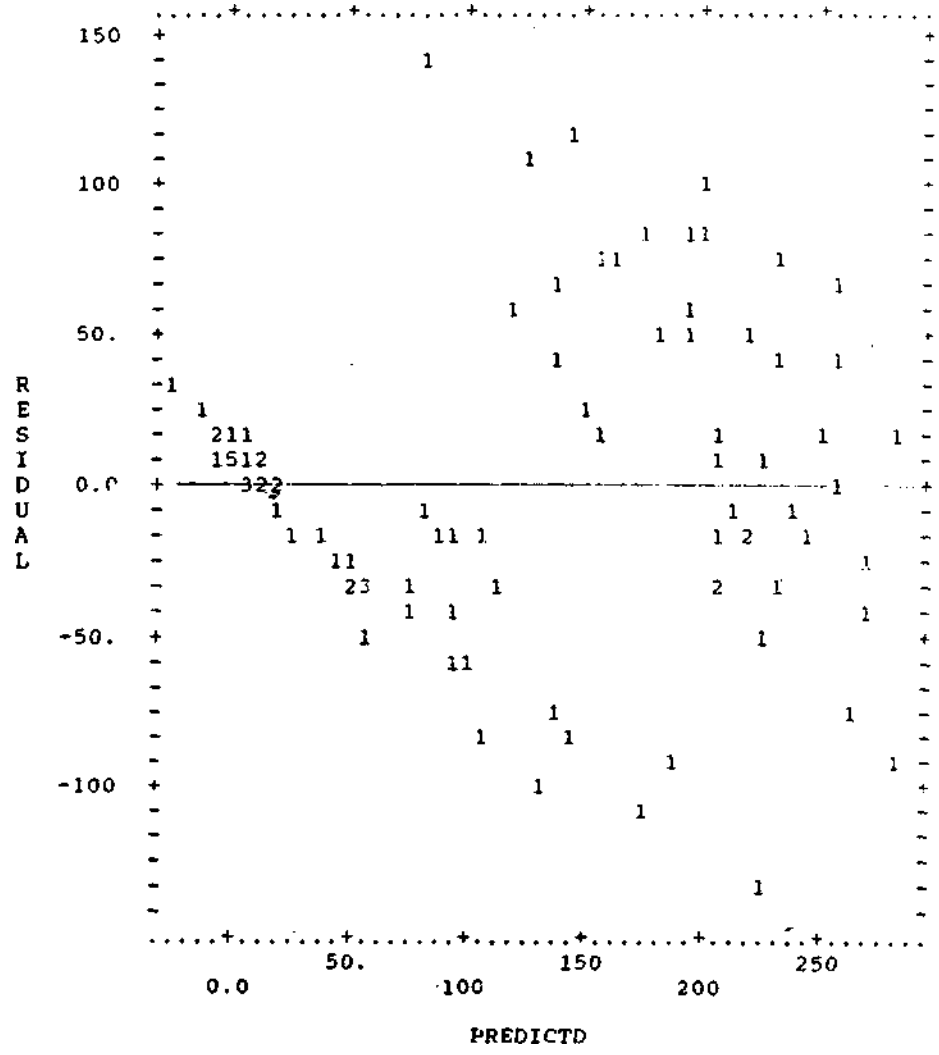

Fig. 2. Residuals Plotted against Predicted Values

In order to ascertain the form of distribution and to assess whether log transformations of variables will make the data more close to normal distribution, the residual plots and normal probability plots of models with log transformation of data and without any transformation of data were compared.

Figures 1 and 2 are for the model based on 'raw* data without any transformation. Fig. 1 is normal probability plot with each data value as the x-coordinate and the corresponding standard normal value as the y-coordinate. The points in Fig. 1 lie on a straight line which

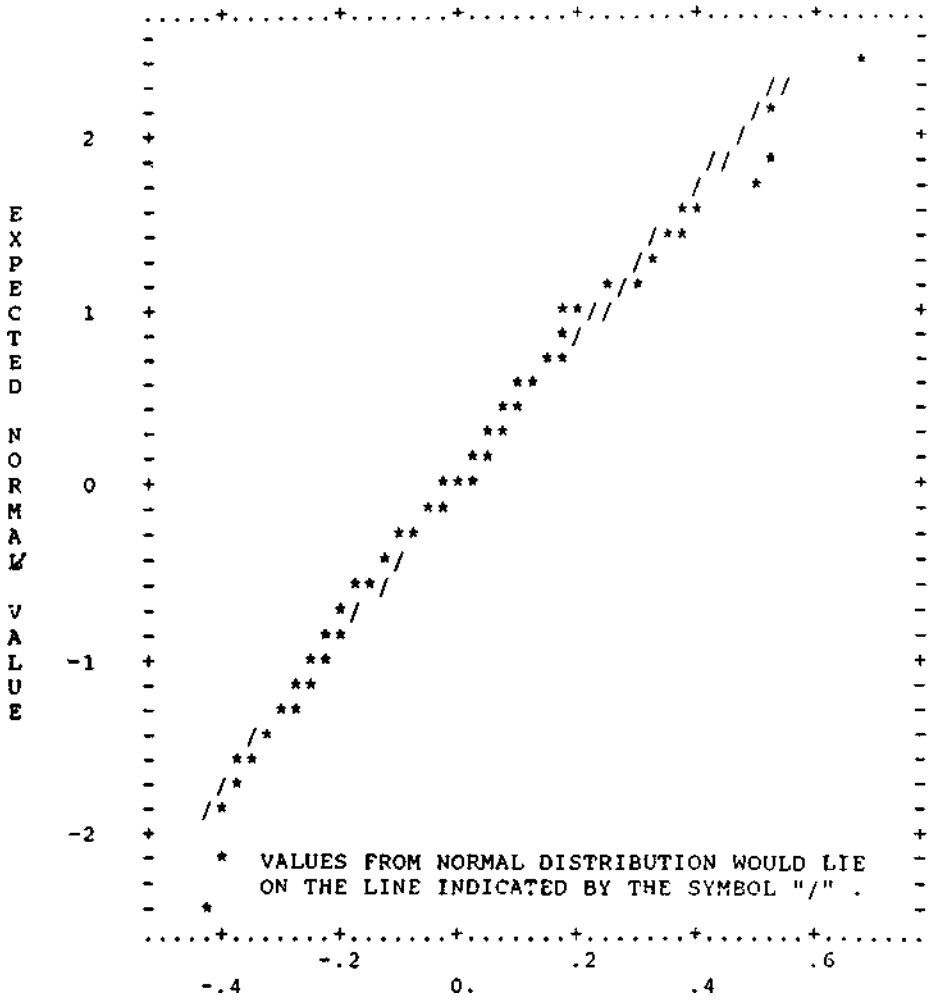

Fig. 3. Normal Probability Plot of Residuals (Plots after log transformation of data)

shows that the *x* and *y* coordinates are approximately equal and it is a good fit. Fig. 2 shows residuals plotted against predicted values. Now if the linear model is valid, the residuals should also have a normal probability distribution. It can be seen from Fig. 2 that the residuals exhibit a fairly symmetrical pattern above and below the horizontal line $e = 0$, that is, about their mean of zero.

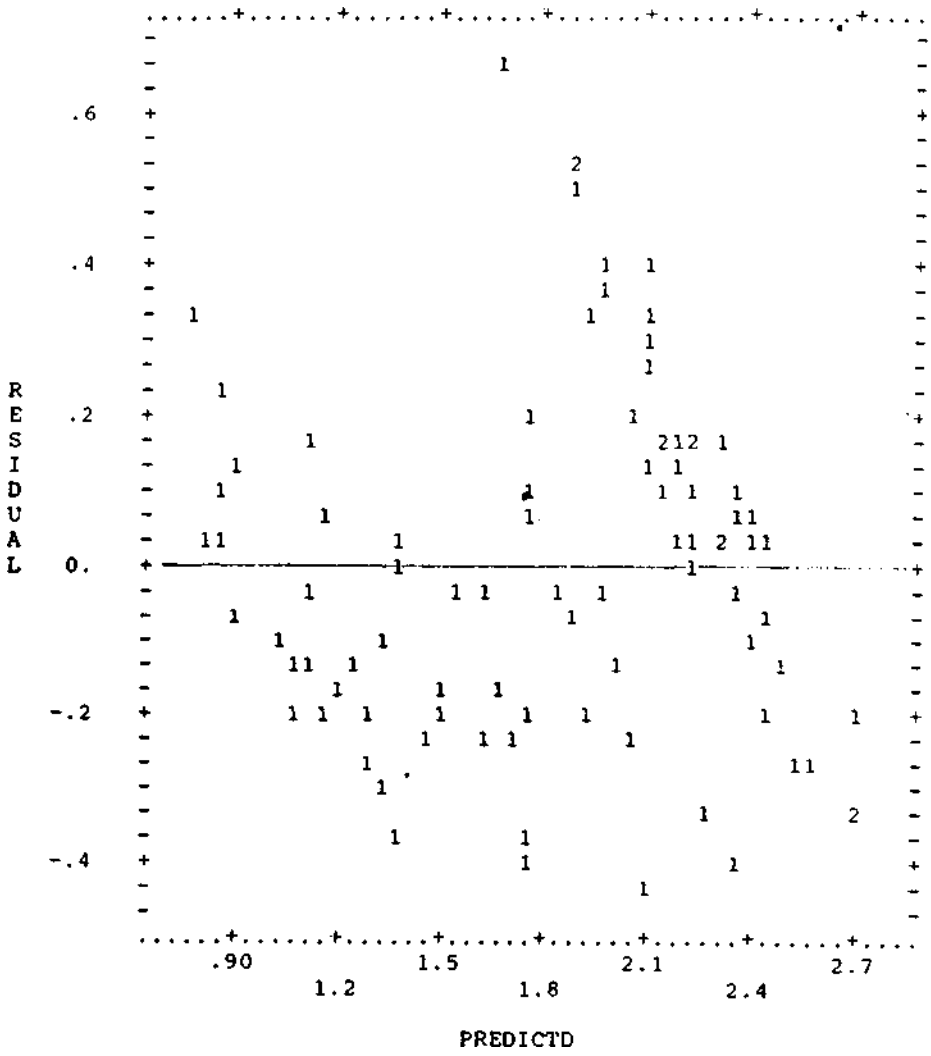

Fig. 4. Residuals Plotted against Predicted Values (Plots after log transformation of data)

Figs. 3 and 4 are for the model based on transformed data (making logarithmic transformation of all variables). Fig. 3 is normal probability curve. The points in Fig. 3 are roughly on a straight line, but it is no improvement over the distribution shown in Fig. 1 for 'raw' data. Fig. 4 shows residuals plotted against predicted values for transformed data and it does not show a symmetrical distribution about their mean of zero.

The above results indicate that linear normal model without transformation of data is suitable for data used in the present paper.

The measures of divergence from normality (skewness and kurtosis) for each variable were also studied to find whether any particular variable/s only showed significant deviations from normal distribution. This would help to know whether transformation of data on some particular variables was needed. The value of measure of skewness, g_1, should be nearly zero for normal distribution and large negative and positive values constitute the critical region for rejecting the hypothesis of normality. The critical values for skewness and kurtosis may be found in Sncdecor and Cochran (1967).

The 2% two-tailed value for measure of skewness is 0.596 and the critical value for kurtosis at 1% level is 4.39 for sample size 90 to 100. The calculated values of skewness and kurtosis for all study variables are given in given in Table 2. It is seen from Table 2 that only two variables—X_4 (percentage of infants with low birth weight) and X_1 (GNP) per capita) show significant divergence from normality. The logarithmic transformation of X_1 and X_4 was able to remove the skewness.

TABLE 2 : MEASURES OF DIVERGENCE FROM NORMALITY FOR EACH STUDY VARIABLE

Variable	*Skewness*	*Kurlosis*
X1	1.53	1.20
X2.	-6.75	-0.64
X3	-0.42 .	-1.35
X4	2.27	8.12
X5	0.18	-0.35
X6	0.11	-0.91
X7	-0.21	-1.47
X8	-0.19	-1.16
Y	0.29	-1.59

Step up multiple regression analysis was used to select variables for inclusion in the model. This analysis was carried out after making logarithmic transformation of X_1 and X_4 and also without any transformation of data. The results were same in both situations— same three variables were selected (X_1 and X_4 were not selected irrespective of transformation or no transformation). The following three independent variables were selected:

X_3 (adult female literacy percentage)

X_8 (population annual growth rate)

X_7 (percentage of population with access to drinking water)

Data on selected three explanatory variables and dependent variables was available for 107 countries.

The explanatory variables selected as above were further examined for multicollinearity by regressing in turn each explanatory variable on all the rest. The values of Rj^2 , *VIF* (j) are given in Table 3. Marquetart (1970) has recommended *VIF* (j) > 5 as an indicator of harmful multicollinearity (as multicollinearity is a problem of degree rather than kind). None of the three selected explanatory variables shows high multicollinearity (Table 3).

TABLE 3 : VARIANCE INFLATION FACTOR AND Rj^2 FOR EACH EXPLANATORY VARIABLE

Variable		Rj^2	*VIF (j)*
(X_3)	Adult female literacy percentage	0.5825	2.3941
(X_8)	Population annual growth rate	0.3463	1.5297
(X_7)	Percentage of population with access to drinking water	0.4939	1.9759

Checks for Influential Observations

Multiple regression analysis was carried out using observations of 107 countries for which data for 3 explanatory variables and dependent variables was available. Values of standardised residual and Cook's distance were computed for each observation (country). Standardized residual was found to be highest (2.99) for U.A.E. The critical values of standardised residuals for identifying outlier in linear models have been tabulated by Richard (1975). The critical value for *n* and *k* (sample size and variables) of present study is 3.77 at x = 0.01 and 3.43 at x = 0.05. Therefore, no observation was found to be outlier in this analysis. Even in the absence of outliers, an observation may be influential if important features of the analysis are changed substantially when the observation is deleted. Cook's distance provides a measure based on confidence ellipsoids for judging the contribution of each data point to the determination of the least squares estimate of *B*. The Cook's distance was also found to be highest (0.45) for U.A.E. If this country were omitted, the regression co-efficients would move from their values to the edge of a 28.44 per cent confidence ellipsoid. Therefore, U.A.E. is an influential observation and deletion of this one observation (country) from analysis changes regression coefficients significantly.

U.A.E was deleted and multiple regression analysis was carried out on data of 106 countries. Values of standardised residual and Cook's distance were computed for each observation (country). The highest value of standardised residual (2.60) was for Papua New Guinea and it was not significant. The highest value of Cook's distance (0.15) was found to be for Kuwait. If this country were omitted, the regression co-efficients would move from their values to the edge of a 7.04 per cent confidence ellipsoid. If change is less than 18 per cent in confidence ellipsoid by omission of an observation, then deletion of that observation is generally not recommended. Therefore, the model developed on data of 106 countries is proposed as the model for under five mortality from this study. The values of intercept and regression co-efficients for this model (study sample C) are presented in Table 5.

Stability of Regression Co-efficient

If diagnostic checks have not been carried out for multicollinearity, especially with aggregate data, then a substantial change in the co-efficient of an individual explanatory variable may occur as other variables are added to or subtracted from the model. It would be difficult to make any clear cut interpretation of the regression coefficients as measures of marginal effects in such a situation. Moreover, the co-efficients can also be unstable in the sense that small data revisions could have a disproportionate effect on the calculated co-efficients. The stability of regression co-efficients has been examined in present study by varying number of variables and number of cases (countries).

The regression coefficients and r-ratios for three models fitted by using different number of explanatory variables and cases (countries) are shown in Table 5. The first model *(A)* is based on data of 8 explanatory variables and dependent variable for 92 countries. The second model *(B)* is based on data of 3 explanatory variables and dependent variable for 107 countries. The third model *(C)* is based on data of 3 explanatory variables and dependent variables for 106 countries. The selection of variables and the sequence and need of fitting these models has already been explained in the section on methods and foregoing write up. Models *B* and *C* have already been compared in the preceding section while checking for influential observations. Here it is proposed to examine that how the regression co-efficients of three selected explanatory variables and their associated /-ratios have changed from model *A* to'model C, which have quite different number of total explanatory variables and number

TABLE 4 : OBSERVATIONS WITH LARGEST VALUES OF CASE STATISTICS— •STANDARDISED RESIDUAL AND COOK'S DISTANCE

Study sample	*Observation No. (country)*	*Standardized residual in absolute value/(critical value)*	*Cook's distance/ (influence)*
A (n = 92,k=8)	62 (Iraq)	*2.1* \ / (3.74)	
	100 (Kuwait)		0.18 / (0.73%)
B (n=107,k = 3)	85 (U.A.E.)	2.99 / (3.77)	0.45 / (28.44%)
C (n=106,k = 3)	67 (Papua New Guinea)	2.60 /(3.T7)	
	100 (Kuwait)		0.15 / (7.04%)

A : Data on dependent variable and 8 independent variables was available for 92 countries and constitutes study sample *A.* This study sample did not contain observation No.85 (U. A.E.) *B :* Data on dependent variable and 3 independent variables (selected through step up multiple regression analysis) was available for 107 countries and constitutes study sample *B.* This includes U.A.E. also. C : As U.A.E. was found to be an influential observation, it was deleted from sample *B* to constitute study sample *C* with data of 106 countries.

of observations (countries). Out of the three selected explanatory variables (Xs, *XT,* Xg), the changes in value of regression co-efficient for variables *Xj,* and X? are negligible. The third variable Xg was not significant when all 8 explanatory variables were used (model *A)* but it has become highly significant when only three explanatory variables are used (model Q. The analysis with 3 explanatory variables (model *C)* finds all the three variables as significant predictors, while analysis with 8 explanatory variables (model *A)* finds only 2 variables as significant predictors of dependent variable. This may be due to the fact that the presence of multicollinearity in 8 explanatory variables created difficulty in distangling separate effects on dependent variable with precision in model *A.* The inclusion of larger number of explanatory variables and a substantial reduction in number of observations (countries) (in model *A)* has no other variations or difference from model *C.*

TABLE 5 : CONSISTENCY/CHANGES IN REGRESSION CO-EFFICIENTS AND /-RATIOS WITH VARIATION IN NUMBER OF VARIABLES AND NUMBER OF CASES (COUNTRIES) IN THE STUDY

Variables	*Study sample A*			*Study sample B*			*Study sample C*		
	Keg. Coef.	t	*P*	*Reg. Coef.*	t	*P*	*Keg. Coef.*	t	*P*
Intercept	249.28	2.23	0.028	287.48	11.83	0.00	266.84	11.00	0.00
(X_1)	-0.002	-1.13	0.263						
X2)	0.21	0.03	0.821						
(X3)	-1.93	-2.82	0.006	-1.88	-6.80	0.00	-1.91	-7.20	0.00
(X4)	-0.89	-0.83	0.409						
(X5)	0.55	0.62	0.538						
(X6)	-0.18	-0.34	0.731						
(X7)	-0.90	-2.50	0.014	-0.94	-3.70	0.00	-0.75	-2.99	0.004
(X)	10.08	1.55	0.124	11.62	2.15	0.034	17.53	3.18	0.002
		$R^2 = 0.766$			$R^2 = 0.719$		$R^2 = 0.741$		

Note : Description of study samples *A, B* and *C* is given *in* Table 4.

Discussion

Several studies have been carried out on factors associated with infant mortality rate in our country, but there is hardly any study on factors associated with mortality rates in young children under five years of age. As our country is quite heterogeneous with respect to socio-demographic data, a study on factors associated with mortality rates in young children from data of states and districts may also be quite revealing. As the focus of present paper is on regression diagnostic methods, some further discussion on methodological aspects may be more pertinent to the objective of this communication.

Most of the demographic studies use aggregate data as the units of observations are usually geographically defined areas. The regression diagnostic checks as discussed in this

paper would help in developing stable models from demographic data. The present paper indicates that transformation of data may not always be necessary even if data consists of rates and percentages. The use of step up multiple regression analysis could help select explanatory variables in present study which did not have high multicollinearity. However, if a significant degree of multicollinearity still remains, then more efforts will be required. One possibility is to produce one or more derived aggregative variables based upon the given explanatory variables. An extension of this possibility is the multivariate procedure of applying principal components analysis to the explanatory variables. This leads to a sequence of principal component variables each of which is a weighted combination of the explanatory variables.

It is important to look for influential observations in modelling with aggregate data. Cook (1979) states that in the least squares analysis of data based on full linear regression model, an observaion may be influential if important features of the analysis are changed substantially when the observation in deleted. Cook's distance *(Di)* provides a useful measure to identify influential observations. The *Di* can be seen as a composite measure of two components: (i) the j'th studentized residual (standardized residual) to test that the ith observation is not an outlier, and (ii) the ratio of the variance of the *ilh* predicted value to that residual which reflects characteristics of location of the /th point. In the present study, the inner structure of data of U. A.E. was different (which reflects different inter-relationships amoijg development indices used as explanatory variables in the present study) and it was influencing important features of analysis (regression coefficients). Therefore, it was deleted from the model. This diagnostic check should always be carried out in developing regression model using aggregate data.

References

Behnken, D. W. and Draper, N. R., 1972, Residuals and iheir variance patterns, *Technometrics,* 14.
Cook, R. D., 1977, Detection of influential observation in linear regression, *Technometrics,* 19(1).
Cook, R. D., 1979, Influential observations in linear regression, *Journal of the American Statistical Association,* 74,365.
Grant, J. P., 1988, *The Stale of the World's Children 1988,* (UN1CEF), Oxford University Press. Kleinbaum, D. G., Kupper, L. L. and Morgenstem, H., 1982, *Epidemiological Research: Principles and Quantitative Methods,* Van Nostrand Reinhold, New York. Marquardt, D. W., 1970, Generalised inverse ridge regression—Biased linear and non-linear estimation, *Technometrics,* 12 Morgenstem, H., 1982, Uses of ecological analysis in epidemiological research, *American Journal of Public Health,* 72 (12).
Richard, E. L., 1975, Tables for an approximate test for outliers in linear models, *Technometrics,* 17 (4). Robinson, W. S., 1950, Ecological correlations and the behaviour of individuals, *American Sociological Review,* 15. Snedecor, G. W. and Cochran, W. G., 1967, *Statistical Methods,* The Iowa State University Press, Iowa, U.S.A.