Jour. Ind. Soc. Ag. Statistics
Vol. XLI, No. 3 (1989), pp. 271-278

MODEL SPECIFICATION FOR REGRESSION ANALYSIS IN THE PRESENCE OF OUTLIERS

SURUCHI JENA
OUAT, Bhubaneswar

(Received : August, 1987)

SUMMARY

This study demonstrates the role of residuals in specification of model for regression analysis in the presence of outliers. Outlier. present in the groundnut productivity data in the state of Orissa reduces the regression coefficient and also the coefficient of determination. By using Weighted Least Square regression mothod the regression coefficient and the coefficient of determination can be improved without deleting any outlier or information but regression analysis by OLS or WLS method would be completed after examining the validation of basic assumptions based on residuals. The study on plot of studentized residual against time for time series data has important role in the regression analysis.

Keywords : Residuals, Studentized residuals, Outliers, Coefficient of determination (R^2), Weighted Least Square regression (WLS).

Introduction

Most widely used statistical tool for analysing multifactorial data is the regression analysis. In regression analysis, generally parameters are estimated by OLS method and statistics like t, F and R^2 are used to evaluate the fit of the regression equation, but there is a simple and effective method i.e. examination of residuals, which has much importance in regression analysis. Residual simply means the algebraic difference between the fitted and observed value. The study on corresponding studentized residuals is also equally important before adopting further analysis.

Main purpose of the paper is to study the residuals and the studentized residuals obtained by OLS and WLS methods based on the Productivity

Index data of groundnut crop in the state of Orissa. Keeping in view the main objective, the following aspects are studied :

—To detect the number of outliers in the observed data,

—To test the significance of outliers,

—To estimate the regression coefficients by OLS method in the presence of outliers (full data set) and in the absence of outliers (reduced data set),

—To obtain the residuals (e_t) and the corresponding studentized residuals (e_{it}),

—To examine the plot of studentized residuals against the fitted values (OLS),

—To re-estimate the regression coefficient by WLS method using full data set,

—To obtain the studentized residuals for WLS method and to plot the same against the fitted values,

—To compare the estimates and the plots obtained by OLS and WLS methods,

—To detect the presence of heteroscedastic error in above cases,

—Lastly, to find the evidence of autocorrelation from studentized residuals.

Secondary data on productivity index number of groundnut crop for 24 years (i.e. from 1960-61 to 1983-84) have been collected from Agricultural Index Number, Bureau of Statistics and Economics, Orissa, Bhubaneswar.

Model

An exponential model like $Y_{It} = \alpha\,\beta^t$ has been fitted to the observed data. It is expressed in terms of logarithms as

$$\log Y_{It} = \log \alpha + (\log \beta)\, t + U_t$$

i.e. $$\log Y_{It} = A + \beta \cdot t + U_t$$

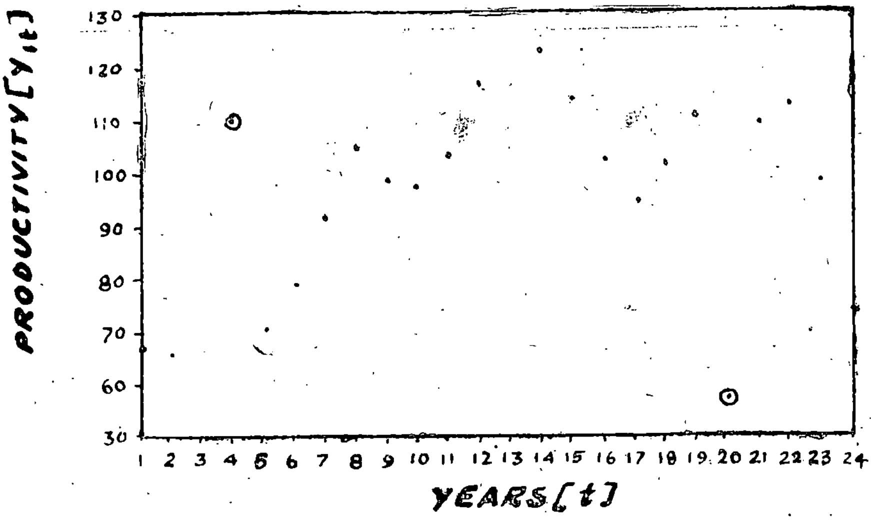
where, $\log \alpha = A$, $\log \beta = B$,

U_t = the random error, assuming normally distributed with zero mean and unit variance,

a and b = the estimated values of A and B respectively by OLS method,
a' and b' = the estimated values of A and B respectively by WLS method, and
$\log Y_{it}$ = the fitted values.

Observation

The plot of the groundnut productivity index (Y_{it}) against the independent variable (t) indicates a nonlinear relationship between two variables and presence of two outliers (circled in Fig. 1).



CIRCLED POINTS INDICATE OUTLIERS

Fig. 1 : Plot of Y_{it} versus t.

By applying Dixon's test it is found that one outlier has significant effect on the data. Then the values of regression coefficient, its standard error, coefficient of determination, residual sum of squares, are estimated by OLS method for full data set and are presented in Table 1.

Table 1 indicates that regression coefficient obtained by OLS method for $n = 24$ is 0.00606 and is significant at five per cent level while the value of R^2 indicates the poor relationship between dependent and independent variables. The residual sum of squares is found to be 0.1594.

TABLE 1—SUMMARY OF REGRESSION RESULTS (OLS) FOR FULL DATA

Items	*Values*
b	0.00606 (2.41)*
$SE(b)$	0.00251
a	1.91226 (56.76)**
$SE(a)$	0.03369
R^2	0.2093
Σe_t^2	0.1594
n	24

t — statistics in the parentheses,

* — significant at five per cent level,

** — significant at one per cent level.

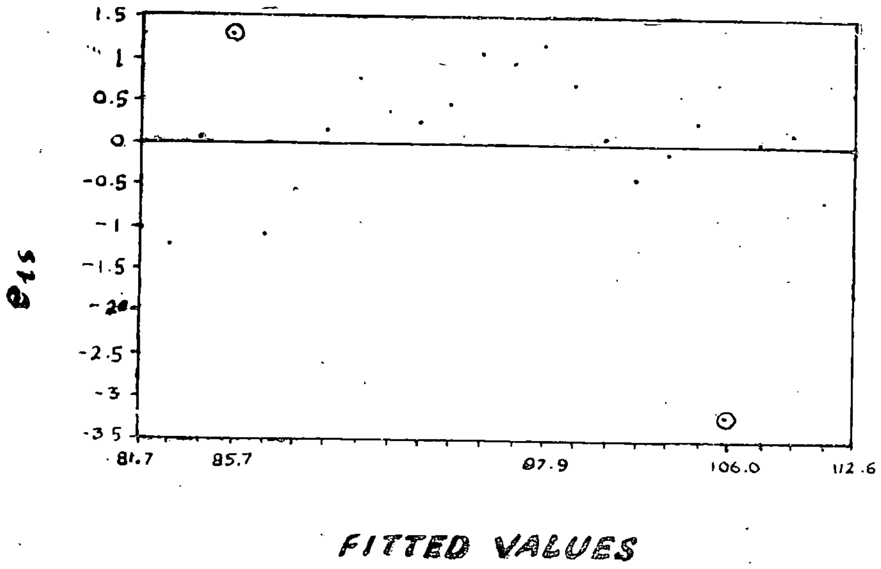


Fig. 2 : Plot of e_{ts} versus fitted values by OLS with outliers.

Then studentized residuals are plotted against the fitted values (Fig. 2) and found two distinct outliers (circled), one is upper outlier (at $t = 4$) with positive sign and the other is lower outlier (at $t = 19$) with negative sign. Again these two outliers are tested against the critical values for testing outliers as presented in Table 2.

TABLE 2—ESTIMATED r_{22} AND THEIR CRITICAL VALUES FOR TESTING OUTLIERS

Outlier	e_t	e_{ts}	r_{22}	*Critical values at 5 % level*
$t = 4$	0.1126	1.32	0.1468	0 413
$t = 19$	−0.2700	−3.17	0.4863**	0.413

** — stgnificance at one per cent level.

It is clear that the residual obtained at $t = 19$ has significant effect to the groundnut productivity data. The distribution of studentized residuals are not within ± 2 indicating the lack of fit of the model by OLS method for full data set. Moreover distribution indicates the presence of heterocedastic error and evidence of autocorrelation.

An attempt has been made to find out the regression coefficient and other statistics in the absence of significant outlier i.e. after deleting the observation at $t = 19$ and results obtained are given in Table 3.

TABLE 3—SUMMARY OF REGRESSION RESULTS (OLS) FOR THE REDUCED DATA SET

Item	*Values*
b	0.00840 (4.40)**
$SE(b)$	0.00191
a	1.89926 (77.36)**
$SE(a)$	0.02455
R^2	0.4792
Σe_t^2	0.0776
n	23

t — statistics in the parentheses,
°° — significance at one per cent level.

Table 3 illustrates that after deleting the significant outlier, the relation between dependent and independent variables has been doubled, whereas the residual sum of squares has been reduced to half (in comparison with Table 1). Moreover, the estimated value of b i.e. 0.00840 is highly significant and indicates that the significant outlier is also responsible for deflating the regression coefficient in the full data set.

The studentized residuals for reduced data set are plotted against the fitted values (Fig. 3) to study the distribution pattern and it is found that all the studentized residuals are within the prescribed range, indicating better fit of the model by OLS method for reduced data set. Moreover, the distribution indicates the absence of heteroscedastic error but there is evidence of autocorrelation.

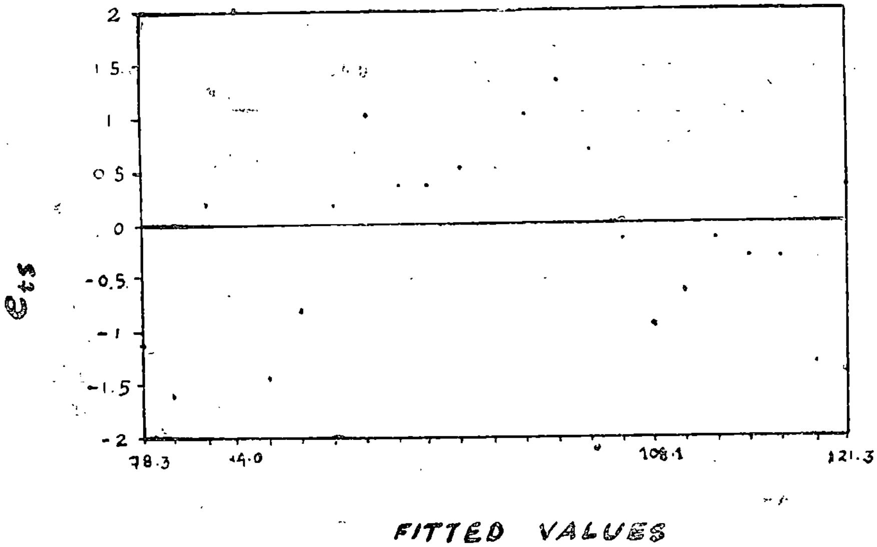


Fig. 3 : Plot of e_{ts} versus Fitted Values by OLS without Outliers

As it is not always advisable to delete the data which may provide more valuable information to the analysis, so in the presence of significant outlier, another method i.e. Weighted Least Square (WLS) regression technique has been applied instead of OLS to obtain the better estimate of the parameters. In this case, weights applied are inversely proportional to the variance of errors so as to minimise the weighted error sum of squares. Let

$$C_t = W_t / \overline{W} \text{ (relative weights)}$$

where, $W_t = 1/s_t^2$

$\overline{W}$ = average of the weighting factors.

The regression coefficient and other relevant statistics are re-estimated by using C_t weights and results are placed in Table 4.

TABLE 4—SUMMARY OF REGRESSION RESULTS (WLS) FOR THE FULL DATA SET

Items	*Values*
b'	0.00609 (16.92)**
$SE(b)$	0.00036
a'	1.91701 (380.36)**
$SE(a)$	0.00504
R^2	0.9286
Σe_t^2	0.00432
n	24

t — statistics in parentheses,

** — significant at one per cent level.

The estimated value of regression coefficient obtaineb by WLS method using weights as C_t is quite different from that obtained by OLS without outlier but nearer to that obtained by OLS method in the presence of outlier. Moreover values of R^2 and Σe_t^2 indicate the resultant effect of weights. To examine the fitness of the model by WLS method, the studentized residuals obtained are plotted against the fitted values (Fig. 4).

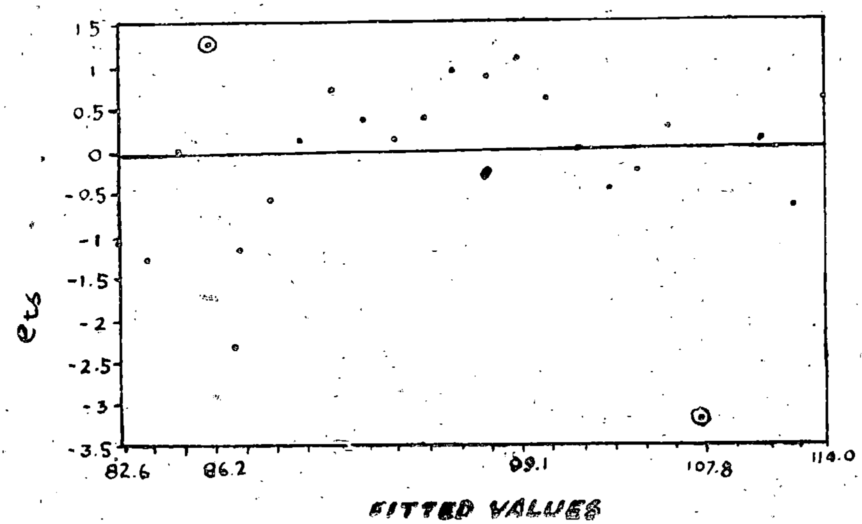


Fig. 4 : Plot of *ets* Versus Fitted Values by WLS with Outliers

In Fig. 4, the distinct pattern of distribution of the standardized residuals indicates the violation in the model specification. Since the deviations with positive sign are more, it indicates predictions by the model to be lower than the observed value. Even though the distribution indicates the absence of heteroscedastic error the cluster residuals indicate the evidence of autocorrelation. So the fitted equation is not well fitted to the data under study.

Conclusion

Although, the regression coefficient is highly rignificant and value of R^2 is on the high side, the regression analysis would not be completed until the basic regression assumptions based on residuals are valid. For analysis of modern trends based on time, the most meaningful analysis is the plot of residuals against time. It serves as the starting point for checking the deficiency of model assumption, lack of constant variance, presence of outliers, and lastly evidence of autocorrelation.

REFERENCES

[1] Croxton, F. E. and Cowden, D. J. (1982) : *Correlation of Time Series, Applied General Statistics*, Third Edition, Prentice Hall, New Delhi, 480 p.

[2] Daniel, C. and Wood, P. S. (1980) : *Fitting Equations to Data, Computer Analysis of Multifactor Data*, 2nd Edition, John Wiley, New York.

[3] Dixon, W. J. (1964) : Query 4 : Rejection of Outlying Values, *Technometrics*, 6 : 238.

[4] Owen, L. Davies and Peter, Goldsmith (1977) : *Weighted Linear Regression, Statistical Melhods in Research and Production*, 4th revised edition, Longmans Inc, New York, pp 202-206.