Pak Jour. Agri. Sci.
22 (4), 245, 1985

DEVELOPMENT AND VALIDATION OF PREDICTION MODEL IN MILCH ANIMALS

Muhammad Ateeq Qureshi* and Muhammad Idrees Ahmad**

Multiple Linear Regression Model for predicting lactation yield on the basis of initial milk yield, peak milk yield and persistency of production was developed for the data on 201 Tharparkar cows maintained at Rakh Ghulaman Farm. The prediction model was tested for inadequacy and other deficiencies. Eign-system analysis revealed that there was no serious problem of multicollinearity amongst exogenous variables. Residuals were computed and various scatter charts of e_i v/s endogenous and exogenous variables were prepared. Funnel shape and double bow pattern were symptotic of model defects. 66 outliers were detected which were deleted. Fresh regression analysis was performed. The magnitude ofPRESS STATISTIC in the presence of outliers was 6455.5 and it was 2335.4 after removing the influential observations. This substantial decrease of 59% in the PRESS-STATISTIC is indicative of the fact that the new model was more reliable.

INTRODUCTION

Linear Regression models are extensively used by the researchers in various fields for prediction and parameter estimation. Som assumptions are generally laid down by the model developers but there is a rare case of testing validation for the linear regression model. Since the models are to be used by the applied specialists, it is the moral duty of model developers to test its validity and ensure whether or not the model will be a sussessful predictor in that environment. Model validation requires that prediction performance should work both forward and backward i. e. estimation in the past and in future should be done with high accuracy.

* Lecturer, Deptt. of Statistics, Government Islamia College, Faisalabad.
** Deptt. of Statistics, University of Agriculture, Faisalabad.

The object of the present study was :-

(i) to make inference about the unknown parameters of the linear regression model for predicting lactation yield on the basis of initial milk yield, peak yield and persistency of production,

(ii) to diagnose multicollinearity,

(iii) to detect outliers,

(iv) to assess validation of the model using PRESS STATISTIC.

MATERIALS AND METHODS

The data pertaining to monthly milk records of lactations from 201 Tharparkar Cows maintained at the Livestock Experiment Station Rakh Ghulaman during 1972-76, were taken from the Faculty of Animal Husbandry. The following variable-terminology were utilised.

X_1— peak milk yield, X_2— initial milk yield

X_3— persistency, Y — lactation milk yield

OLS method was used to fit the general linear regression model $Y = X\beta + \varepsilon$ and for the estimation of $\hat{\beta} = (XX)^{-1} XY$ and Var $(\beta) = (XX)^{-1} \sigma^2$ (Montgomery, 1982).

F-test was used to test the hypothesis about the parameters of the regression model.

MULTICOLLINEARITY DIAGNOSTICS

The problem of multicollinearity arises due to the presence of high correlation amongst the exogeneous variables. It can seriously disturb the OLS fit and in some situations may render the model quite useless. Multicollinearity is possible even if simple correlations are low but R^2 is high and partial correlation are low. Eigen values of $\acute{X}X$ can be used to measure the magnitude of multicollinearity. The presence of multicollinearity implies that some of the eigenvalues of $\acute{X}X$ are small. Multicollinearity may also be digonosed by examining the condition number

$$K = \frac{\lambda \text{ max.}}{\lambda \text{ max.}} \quad \text{where } \lambda \text{ is an eigne-valve of } \acute{X}X$$

There is no serious problem with multicollinearity if $K < 100$. Condition numbers between 100 and 1000 imply moderate to strong multicollinearity

while K > 1000 indicates severe multicollincarity. (Montgomery, 1982).

RESIDUAL ANALYSIS

Analysis of the residul $e_i = Y_i - \hat{Y}_i$ is an effective method of discovering several types of model deficiencies. For further development of the model, the most sensitive assumptions about the residuals should be fulfilled. If not, indentification of influential observations may diagonose the several types of model deficiencies.

Plots of residuals are used detect the departure from normality, outliers, inequality of variances and wrong functional specification for the regressors. (Barnett, & Lewis 1978, Daniel & Wood 1980. Ganaanadeskan 1977). Residuals may be used to construct tolerance limits.

$$Y - 2\,S_{n-2} < \mu < Y + 2\,S_{n-2} \quad \text{where } S^2_{n-2} = \frac{\Sigma e_i^2}{n-2}$$

for the detection/rejection of outliers. (Anscombe, 1960, Anscombe & Tukey, 1963 and Rosner, 1975).

VALIDATION AND PRESS-STATITIC

Regression models are generally used for prediction or estimation of mean response. Selections of regressors are therefore, made in such a manner that mean sqeare of prediction is reduced i. e exogeneous variables with small effects should be deleted. For model validation, a procedure is to split the data for estimation and prediction. This may be done in a variety of ways such as using cross-validation or PRESS-STATISTIC. (Mosteller & Tukey 1968 and Stone 1974). For PRESS-STATISTIC, select an observation say j-th and the regression model fitted to the remaining n-1 observations. This equation is used for prediction. The prediction error for the point j, also known as deleted residual, will be $e_{(j)} = Y_{(j)} - \hat{Y}_{(j)}$. The values $e_{(j)}$ for j=1,2,3..... n are obtained by repeating procedure for each observation.

PRESS-STATISTIC $\Sigma[Y_j - Y_{(j)}]^2 = e^2_{(j)}$ is used to compare the alternate models. A model with smaller value of PRESS-STATISTIC is considered more reliable.

RESULTS AND DISCUSSION

The data pertaining to monthly milk record for 201 Tharparkar Cows

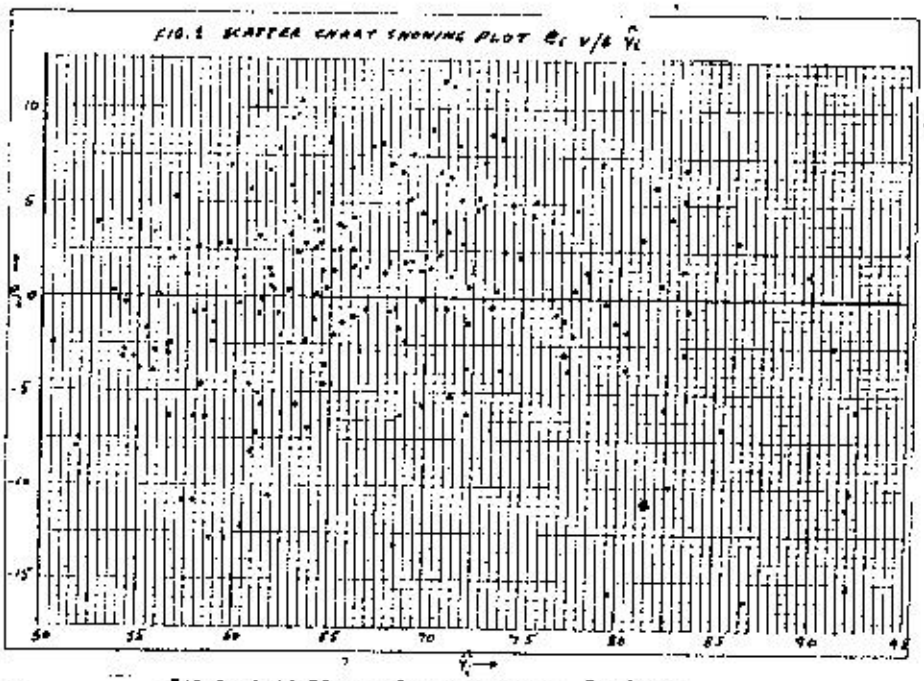


FIG. 1 SCATTER CHART SHOWING PLOT e_i V/s $\hat{Y}_i$

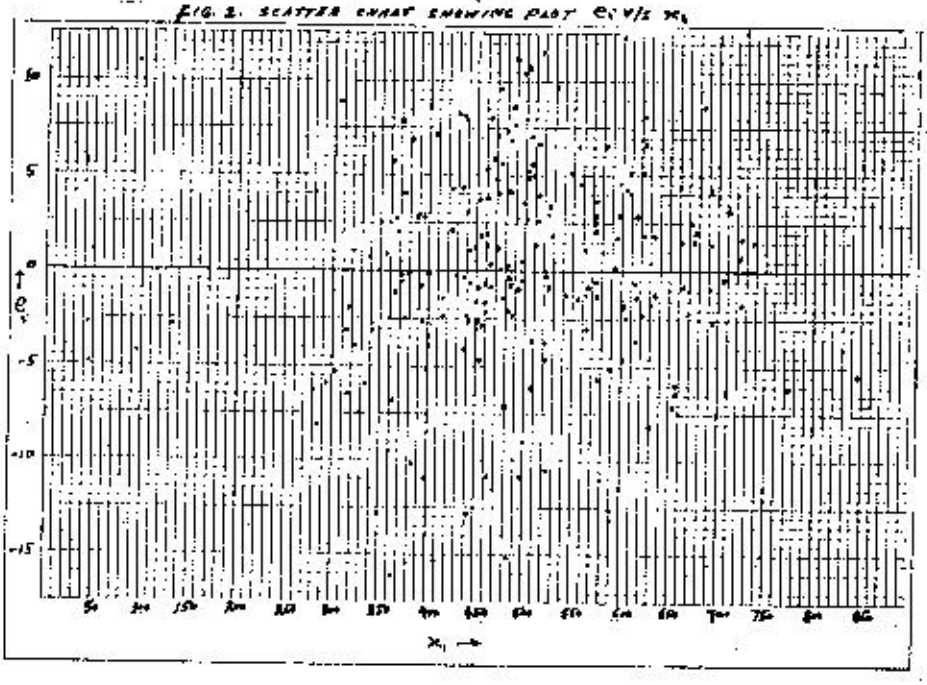


FIG. 2. SCATTER CHART SHOWING PLOT e_i V/s x_1

FIG. 3. SHOWING SCATTER CHART BETWEEN e_1 & x_2

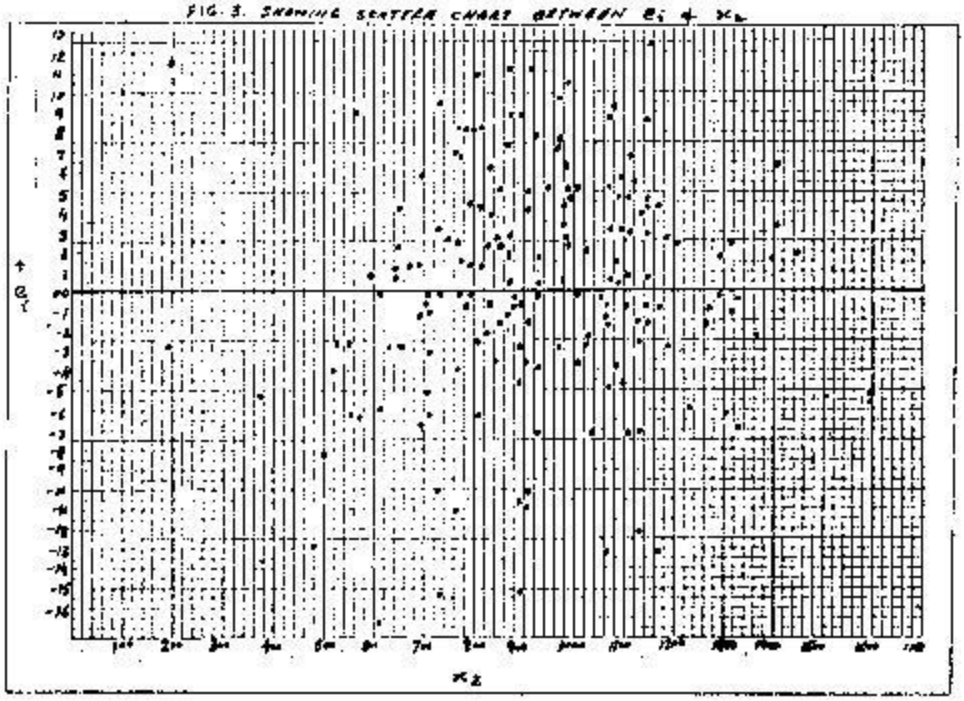
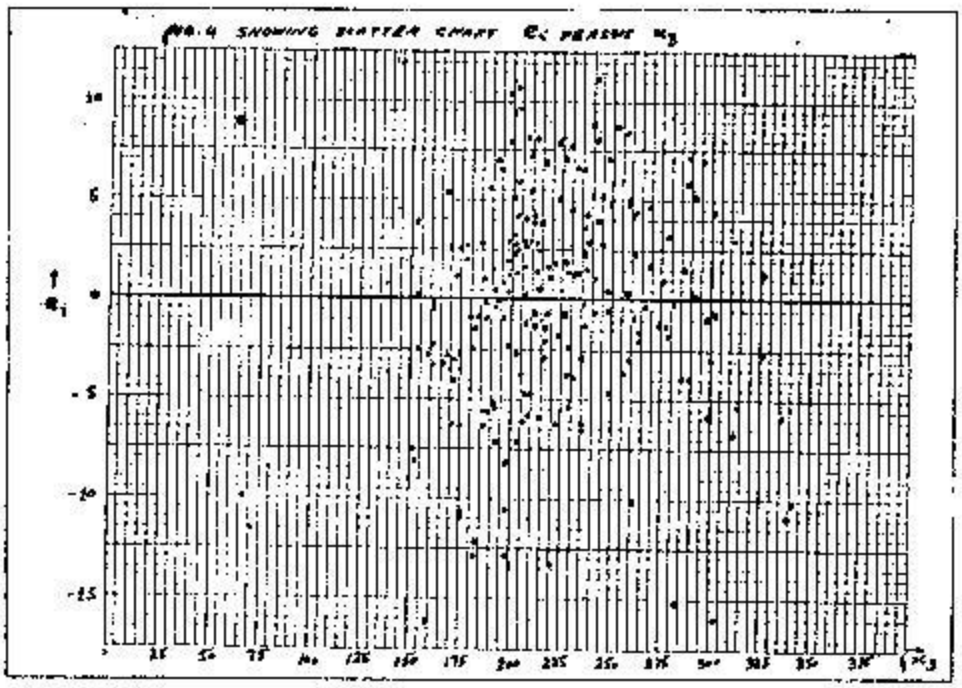


FIG. 4 SHOWING SCATTER CHART e_1 VERSUS x_3



were utilized to develop the multiple develop the multiple linear regression model:

$$\hat{Y} = 18.0234 - 0.015264X_1 + 0.011396X_2 + 0.209589X_3$$

for prediction performance. F-Statistic 179 25 was significant at $P < .01$. The Eigen-system analysis revealed the values of $\lambda_1 = 2$, $\lambda_2 = 2$, $\lambda_3 = 0.0513$ and the condition number $K = 38.99 < 100$ is indicative of the fact that there was no serious problem with multicollinearity.

The residules $e_i = Y_i - \hat{Y}_i$ were computed. Scatter charts of endogenous and exogenous variables v/s residual e_i were prepared. Figures 1, e_i v/s $\hat{Y}_i$ resembles a funnel shape. The other three figures e_i v/s X_1, X_2 and X_3 followed a double bow pattern. However, all the four figures were sympotic of model deficiencies i.e., there were obvious model defects. Average lactation yied was 62.7, $S^2_{n-2} = 32.4397$ and the Tolerance limits were : $56.3 < \mu < 79.1$. There were 66 outliers which were beyond these limits and hence deleted. After eliminating these 66 influential observation, fresh regression analysis was performed. The new multiple linear regression model was :

$$\hat{Y} = 23.6036 + 0.17929738X_1 - 0.00908324X_2 + 0.160223323X_3$$

The magnitude of PRESS-STATISTIC in the presence of bad-values was 6455.5 and it was 2635.4 after eliminating the influential observations.

CONCLUSIONS

Comparison of the alternate models i. e. before and after eliminating outliers revealed that error mean square and F-statistic decreased by 36.6% and 72.13% respectively and there is a substantial reduction of 59% in PRESS STATISTIC. These are all indicative of the fact that the new model after eliminating influential observations has become more reliable.

ACKNOWLEDGEMENTS

Special mention is made of Dr. Manzur-ud-Din Ahmad, Dean, Faculty of Animal Husbandry, University of Agriculture, Faisalabad for providing the basic data and his expert advice.

REFERENCES

Anscombe, F.J. 1960. Rejection of outliers; Technometrics, 2, 123-167.

Anscombe, F.J. and J.W. Tukey. 1963. The Examination and Analysis of Residuals : Technometrics, 5, 141-160.

Barnett, V. and T. Lewis 1978 Outliers in Statistical Data, Wiley, New York.

Daniel, C. and F. S. Wood 1980 Fitting Equations to Data, 2nd Ed., Wiley, New York.

Gnanadesikan, R. 1977 Methods of Statistical Analysis of Multivariate Data, Wiley, New York.

Montgomery, D. C. 1982 Introduction to Linear Regression Analysis. 109, 146, 147-167, Wiley, New York.

Rosner, B. 1975 On the Detection of many outliers, Technometrics, 17, 221-229.