IMPORTANCE OF STUDENTIZED AND PRESS RESIDUALS FOR NONLINEAR MULTIVARIATE REGRESSION MODELS

Dr. Kesavulu Poola[1], Prof. M. Bhupathi Naidu[2].

[1] Assistant Professor, Emeralds Advanced Institute of Management Studies, Tirupati.

[2] Professor, Department of Statistics, S.V.University, Tirupati.

ABSTRACT: ***The problem of outliers is very common in nonlinear models and identification of these outliers also complicated. . In this article we propose several outlier detection techniques for nonlinear regression. The main idea is to use the linear approximation of a nonlinear model and consider the gradient as the design matrix. This paper builds the algorithm to compute the Studentized and Predicted residual sum of squares (PRESS) when obtaining nonlinear equations. PRESS is a well-known "leave-one-out" (LOO) cross-validation method. This method is more significant in regression analysis to decide, how well the model predicts for new observations. This paper develops a method to approximate cross-validation statistics for nonlinear regression. The main objective of is to explain the importance of using the Predicted residual sum of squares (PRESS). This paper advocates the concept of cross-validation and recommends using PRESS for cost analysis. And several business statistical packages assume that, PRESS is for linear and log-linear models. But even though we can calculate PRESS directly by definition for a nonlinear equation, we should avoid running nonlinear regression multiple times.***

KEY WORDS: Studentized, PRESS, Outliers, Cross-Validation.

I. INTRODUCTION

In present era, Regression Analysis has several advances in detecting the outliers. The identification of outliers is very crucial because it is accountable for producing huge interpretative problem in linear as well as in nonlinear regression models. Enormous research has been accomplished on the identification of outlier in linear regression models, but not in nonlinear regression models. To overcome this area of problems Statisticians are using Studentized and PRESS residuals in order to identify the outliers.

According to Habshah et al., Belsley et al., Anscombe, Hadi , Bartlett, Draper, N. R., Tukey, Cook and Weisberg not much work has been explored in the formulation of the outlier's identification method in nonlinear regression. Few researches have been developed in locating the outlier in nonlinear regression models.

.

II. STUDENTIZED FOR MULTIVARIATE NONLINEAR MODELS

So far, we have gone through diverse measures in order to detect the high leverage 'X' variable and unusual outliers 'Y'. But when we try to locate the outlier, the vital problem is, the potential outlier influences the regression model, in a way the estimated value /function dragged towards the potential outlier, in order that it isn't flagged as an outlier using the standardized residual criterion. To deal with this subject, **studentized residuals** propose an unconventional criterion for locating the outliers. The fundamental idea is to delete the observations one at a time, each time refitting the regression model on the remaining *n*–1 observations. Then, we evaluate the observed response values to their fitted values based on the models with the i^{th} observation deleted. .

Consider the general multivariate nonlinear regression model

$$Y_{\alpha t} = f_{\alpha}\left(X_t, \theta_{\alpha}^{0}\right) + E_{\alpha t} \qquad \ldots(2.1)$$

Here θ_{α}^{0} is p-dimensional vector matrix

Suppose $\hat{\theta}$ is the nonlinear least square estimator of θ for large sample, nonlinear least square residual vector.

$$e = \left(Y_i - \hat{Y}_t\right)$$

$$= \left(Y - f\left(\hat{\theta}\right)\right) \qquad \ldots(2.2)$$

Here $\hat{\theta} \square \theta + \left(F'F\right)^{-1} F'\varepsilon$ …(2.3)

And $$F = F\left(\hat{\theta}\right) = \left[\frac{\partial}{\partial \theta_j} f\left(X_i, \theta\right)\right]_{n\times p} \qquad \ldots(2.4)$$

Here $\frac{\partial}{\partial\theta_j} f(X_i,\theta)$ is the $(i,j)^{th}$ element of $(n\times p)$ matrix $F(\theta)$ then the general relationship between 'e' and ' ε ' is

$$e \approx M\varepsilon$$

Here $M=\left[I-F(F'F)^{-1}F'\right]$

or $e \approx [I-H]\varepsilon$ where $M=(H_{ij})=F(F'F)^{-1}F'$ is a symmetric idempotent matrix (or) HAT matrix

In scalar form

$$e_i \approx \left[\varepsilon_i-\sum_{j=1}^{n}H_{ij}\varepsilon_j\right], \qquad j=1,2,...,n \qquad \ldots(2.5)$$

as H is HAT Matrix

Trace (H) = Rank (H) = P

and $\sum_{i=j}^{n}H_{ij}^2=H_{ij}$

Here ε follows $N_0(0,\sigma^2 I)$, so ε following normal distribution with zero mean and variance is $\sigma^2 I$. Here H controls the e.

As we know, variance of each e_i is a function of both σ^2 and $H_{ij}, i=1,2,...n.$

The nonlinear least square residuals have a probability distribution that is scalar dependant. So, the nonlinear studentized residuals do not depend on either of these quantities and they have probability distribution and we have both internally nonlinear studentized residuals and externally nonlinear studentized residuals.

a) INTERNALLY NONLINEAR STUDENTIZED RESIDUALS

In nonlinear regression models, internally nonlinear studentized residuals are define by

$$e_i^*=\frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ij}}}\sim N(0,1) \qquad i=1,2,...n \qquad \ldots(2.6)$$

Here $\hat{\sigma}^2=\frac{e'e}{n-p}$

$$=\frac{\sum_{i=1}^{n}e_i^2}{n-p} \qquad \ldots(2.7)$$

Here $\left[\frac{e_i^{*2}}{n-p}\right]\sim\beta$ - distribution with parameters $\frac{1}{2}$ and $(n-p-1)/2$

It follows, $E(e_i^{*'})=0$ & $Var(e_i)=1 \qquad \forall i=1,2,...,n$

$$Cov(e_i^*e_j^*)=\frac{-h_{ij}}{\sqrt{(1-h_{ii})(1-h_{jj})}} \qquad \forall i\neq j=1,2,...,n$$

Here h_{ij} add up to the trace of the hat matrix = P. Average 'h' is p/n which should be small, so usually $\sqrt{1-h_{ii}}$.

b) EXTERNALLY NONLINEAR STUDENTIZED RESIDUALS:

The externally nonlinear studentized residuals are define by

$$e_i^{**}=\frac{\hat{\varepsilon}_i/(1-h_{ij})}{\sqrt{MSE_{(i)}/(1-h_{ij})}} \qquad \ldots(2.8)$$

$$=\frac{\hat{\varepsilon}_i}{\sqrt{MSE_{(i)}(1-h_{ij})}} \qquad \ldots(2.9)$$

Here $MSE_{(i)}$ = estimate of σ^2 not baring data point 1.

i.e. $$e_i^{**} = \frac{\varepsilon_i}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ij})}} \qquad \forall i = 1,2,...,n \qquad \ldots(2.10)$$

Based on the Normal distribution $\sigma^2_{(i)}$ and ε_i are

i.e., M.S.E (or) $$\hat{\sigma}^2 = \frac{(n-p-1)\sigma_i^2 + \hat{\varepsilon}_i^2/(1-h_{ii})}{n-p}$$

(or) $$\sigma^2_{(i)} = \hat{\sigma}^2\left[\frac{n-p-e_i^*}{n-p-1}\right]$$

So, the relationship between internally and externally nonlinear studentized residuals is given by

$$e_i^{**} = e_i^*\left[\frac{n-p-1}{n-p-e_i^*}\right], \quad i = 1,2,...,n \qquad \ldots(2.11)$$

III. PRESS RESIDUALS FOR MULTIVARIATE NONLINEAR MODELS

PRESS statistic is the sum of the squares of all the residuals such that the predicted value is calculated for the omitted observation in each refitted regression model (see Allen, 1974). PRESS is also known as "leave-one-out" (LOO) statistic, is commonly used in regression analysis for cross-validation. In general, in non linear least squares, and studentized residuals fitting is dependents on all the variables in the data. But in predicted residuals for nonlinear model i.e. i^{th} nonlinear predicted residual is depends on the fit to the data, where i^{th} case is excluded. PRESS evaluates the model in three steps, such as systematically removing each observation from the data set, Refitting the equation, and computing the square of the residual for the removed data point

Suppose $\hat{\theta}$ is the nonlinear least square estimate of θ based on the full data, and $\hat{\theta}_{(i)}$ be the respective estimate where the i^{th} case is excluded.

Now the i^{th} nonlinear predicted residuals i.e.,

$$e_{(i)} = \left[Y_i - f_i\left(\hat{\theta}_{(i)}\right)\right] \qquad i = 1,2,...,n \qquad \ldots(3.1)$$

Here $e_{(i)}$ is the prediction error

$\therefore$ the nonlinear PRESS defined by

$$NLPRESS = \sum_{i=1}^{N} e_{(i)}^2 \qquad \ldots(3.2)$$

So, finally, the relationship between nonlinear predicted residuals and nonlinear studentized residual are given by

(i) $$e_i^* = \frac{e_{(i)}}{\hat{\sigma}/\sqrt{(1-h_{ij})}} \qquad \ldots(3.3)$$

(ii) $$e_i^{**} = \frac{e_{(i)}}{\hat{\sigma}_{(i)}\sqrt{(1-h_{ij})}} \qquad \ldots(3.4)$$

and $$e_{(i)} = \frac{e_i}{(1-h_{ij})} \quad \forall i = 1,2,...,n$$

IV. CONCLUSION

In general phenomenon Regression models need an extensive approach in order to test the significance. Especially in the case of nonlinear regression models common ordinary least square (OLS) method is not appropriate. The OLS estimation is unsuccessful parameters in nonlinear. However Studentized and PRESS residual methods can be applied to estimate parameters of this model. In the above research we have discussed analytical model of the Studentized and PRESS residuals.

REFERENCES

[1] Allen, D. M., "The Relationship Between Variable Selection and Data Augmentation and a Method for Prediction," Technometrics, 16, (1974), 125–127.

[2] Atkinson, A.C., (1981), Two graphical displays for outlying and influential observations in regression, Biometrika, 68, 1, 13-20.

[3] Atkinson, A.C., (1982), Regression Diagnostics, Transformations and Constructed Variables, Journal od Royal Statistical Society, B, 44, 1, 1- 36.

[4] Bates, D.M. Watts, D.G., (1980). Relative curvature measures of nonlinearity, J. R. statist. Ser. B 42, 1-25.

[5] Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), Regression Diagnostics, John Wiley & Sons, New York.

[6] Cook, R. D. and S. Weisberg, "Residuals and Influence in Regression," Chapman and Hall, 1982.
[7] Draper, N. R. and H. Smith, "Applied Regression Analysis (2nd edition)," New York: John Wiley & Sons, Inc., 1981.
[8] Fox, T., Hinkley, D. and Larntz, K., (1980), Jackknifing in nonlinear regression. Technometrics, 22, 29-33.
[9] Hadi, A.H. (1992). A new measure of overall potential influence in linear regression, Computational Statistics and Data Analysis 14 (1992) 1-27.
[10] Montgomery D. C. and E. A. Peck, "Introduction to Linear Regression Analysis (2nd edition)," Wiley-Interscience, 1992.
[11] Morrison, D. F., "Applied Linear Statistical Methods", New Jersey: Prentice-Hall, Inc. 1983.
[12] Morrison, D. F., "Multivariate Statistical Methods, 2nd ed.", New York: McGraw-Hill Book Company, 1976.
[13] Seber, G. A. F. and C. J. Wild, "Nonlinear Regression," New York: John Wiley & Sons, 1989, pages 37, 46, 86-88.
[14] Weisberg, S., "Applied Linear Regression, 2nd Edition," New York: John Wiley & Sons, 1985, pages 87-88.