Research Paper Statistics

Inference in Nonlinear Regression Model With Heteroscedastic Errors

B. Mahaboob	Research Scholar, Department of Mathematics
Dr. M. Ramesh	Senior Data Scientist, Tech Mahindra, Hyderabad.
Dr. Sk. Khadar Babu	Assistant Professor (Senior), Department of Mathematics, VIT, Vellore, Tamilnadu.
P. Srivyshnavi	Academic Consultant, Department of CSE, SPMVV Engineering College, Tirupati.
Prof. P. B.Sireesha	Research Scholar, Department of Statistics, S.V.University, Tirupati.
Balasiddamuni	Rtd. Professor

ABSTRACT

An advent of modern computer science has made it possible for the applied mathematician to study the inferential aspects of an increasing number of nonlinear regression models in recent years.

The inferential aspects for nonlinear regression models and the error assumptions are usually analogous to those made for linear regression models. The tests for the hypotheses on parameters of nonlinear regression models are usually based on nonlinear least squares estimates and the normal assumptions of error variables.

A more common problem with data that are best fit by a nonlinear regression model than with data that can be adequately fit by the linear regression model is the problem of heteroscedasticity of errors. The problem of heteroscedasticity can be studied with reference to nonlinear regression models in a variety of ways.

In the present study, an attempt has been made by developing inferential method for heteroscedastic nonlinear regression model.

KEYWORDS : Intrinsically Nonlinear model; Violation of assumptions on Errors; Problem of Heteroscedasticity.

I. INTRODUCTION

In estimating the linear regression models, the least squares method of estimation is often applied to estimate the unknown parameters of the linear regression model and errors are usually assumed to be independent and identically distributed normal random variables with zero mean and unknown constant variance. The violations of these crucial assumptions on error variables can have several consequences on estimates of parameters and test statistics.

The inferential aspects of nonlinear regression models and the error assumptions are usually analogous to those made for linear regression models. The tests for the hypotheses on parameters of nonlinear regression models are usually based on nonlinear least squares estimates and the normal assumptions of error variables.

Among all assumptions, an important assumption of the regression model is that the errors need to have the constant or homoscedastic variance. Errors that do not have constant variances are known as heteroscedastic errors. The presence of heteroscedastic errors in the nonlinear regression model disturbs the optimal properties of the NLLS estimators of the parameters. These errors produce inefficient estimates of the parameters and invalid inferences concerning the true values of the parameters of the nonlinear regression model.

The various inferential problems on nonlinear regression models involving heteroscedastic errors have been studied by Gallant and Goebel (1976), Carroll and Ruppert (1982a, 1982b), Cook and Weisberg (1983), Baljet (1986), Beal and Sheiner (1988), Welsh, Carroll and Ruppert (1994), Smyth (2002), Fox and Wiesberg (2010), Potocky and Stehlik (2010) and others.

II. NONLINEAR STUDENTIZED RESIDUALS

Residuals have vital role in many testing procedures designed to examine various types of disagreement between data and an assumed nonlinear regression model. In testing nonlinear hypotheses and procedures for detecting the problem of heteroscedasticity, several transformations of the Nonlinear least squares (NLLS) residuals have been suggested to overcome partially some of their shortcomings.

Consider the standard nonlinear regression model with usual assumptions in vector notation as

$$Y_{n\times 1} = f_{n\times 1}(\beta) + \varepsilon_{n\times 1} \qquad ...(2.1)$$

and β is $(p\times 1)$ vector of unknown parameters. Suppose that $\hat{\beta}$ is the nonlinear least squares estimator of β.

For large samples, the NLLS residuals vector is given by

$$e = \left[Y - \hat{Y}\right] = \left[Y - f\left(\hat{\beta}\right)\right] \qquad \ldots(2.2)$$

where $\hat{\beta} \cong \beta + (F'F)^{-1} F'\varepsilon$ …(2.3)

and $$F = F(\beta) = \left(\frac{\partial}{\partial \beta_j} f(X_i, \beta)\right)_{n\times p} \qquad \ldots(2.4)$$

Here, $\frac{\partial}{\partial \beta_j} f(X_i, \beta)$ is the $(i,j)^{th}$ element of $(n\times p)$ matrix $F(\beta)$.

An approximate relationship between e and ε is given by

$$e \cong M\varepsilon, \qquad \text{where } M = \left[I - F(F'F)^{-1} F'\right]$$

or $e \cong [I - \nu]\varepsilon$, where $\nu = (\nu_{ij}) = F(F'F)^{-1} F'$ is symmetric idempotent matrix known as 'HAT' matrix.

or in scalar form,

$$e_i \cong \left[\varepsilon_i - \sum_{j=1}^{n} \nu_{ij}\varepsilon_j\right], \quad i=1,2,\ldots,n$$

…(2.5)

Since, ν is symmetric idempotent matrix, it follows that

$$\text{trace}(\nu) = \text{rank}(\nu) = p$$

and $$\sum_{j=1}^{n} \nu_{ij}^2 = \nu_{ii}, \quad i=1,2,\ldots,n$$

If ε follows $N(O, \sigma^2 I)$ then e follows a singular normal distribution with zero mean vector and variance $\sigma^2 I$. Here ν controls the variation in e.

Since, the variance of each e_i is a function of both σ^2 and ν_{ii}, i=1,2,…,n; the NLLS residuals have a probability distribution that is scale dependent. The nonlinear studentized residuals do not depend on either of these quantities and they have probability distribution that is free of the nuisance scale parameters.

One can make a further distinction between internal studentization and external studentization.

(A) INTERNALLY NONLINEAR STUDENTIZED RESIDUALS

In NLLS regression, the internally nonlinear studentized residuals are defined by,

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{(1-\nu_{ii})}}, \qquad i = 1,2,\ldots,n \qquad \ldots(2.6)$$

where $\hat{\sigma}^2 = \frac{e'e}{n-p} = \frac{\sum_{i=1}^{n} e_i^2}{n-p}$ …(2.7)

Here, $\left[\frac{e_i^{*2}}{n-p}\right] \overset{asy}{\sim}$ Beta distribution with parameters $\frac{1}{2}$ and $\frac{(n-p-1)}{2}$.

It follows that $E(e_i) = 0$ and $Var(e_i) = 1,\ \forall i = 1,2,...,n$

Also, $Cov(e_i^*, e_j^*) = \frac{-v_{ij}}{\left[(1-v_{ii})(1-v_{jj})\right]^{1/2}},\ \forall, i \neq j = 1,2...,n$

(B) EXTERNALLY NONLINEAR STUNDENTIZED RESIDUALS

The externally nonlinear studentized residuals are defined by,

$$e_i^{**} = \frac{e_i}{\hat{\sigma}_{(i)}(1-v_{ii})^{1/2}},\ \forall i = 1,2,...,n \quad \text{…(2.8)}$$

where $\hat{\sigma}^2_{(i)} = \frac{(n-p)\hat{\sigma}^2 - \left[e_i^2 / (1-v_{ii})\right]}{n-p-1}$

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

Under normality $\hat{\sigma}^2_{(i)}$ and e_i are independent.

Here, $e_i^{**} \sim$ Student's t- distribution with $(n-p-1)$ degrees of freedom.

A relationship between internally and externally nonlinear studentized residuals is given by

$$e_i^{**} = e_i^*\left[\frac{n-p-1}{n-p-e_i^{*2}}\right]^{1/2},\ i = 1,2,...,n \quad \text{…(2.9)}$$

Thus, e_i^{**} is a monotonic transformation of e_i^{**2}.

III. ESTIMATION OF PARAMETERS OF NONLINEAR REGRESSION MODEL WITH HETEROSCEDASTIC ERRORS BY USING NONLINEAR STUDENTIZED RESIDUALS

Consider the standard nonlinear regression model

$$Y_i = f(X_i, \beta) + \varepsilon_i, \qquad i = 1,2,...,n \quad \text{…(3.1)}$$

which may be written in matrix notation as

$$Y_{n\times 1} = f_{n\times 1}(\beta) + \varepsilon_{n\times 1} \quad \text{…(3.2)}$$

where $X_i = (X_{i1}, X_{i2}, ..., X_{ik})$ is a k- component vector denotes the i^{th} observation on known k-explanatory variables;

β is a $p\times 1$ vector of unknown parameters;

$f(\square)$ is a known twice continuously differentiable function of β;

The usual assumptions of the nonlinear regression model are:

i. $E\left[Y_i/X_i\right]=f\left(X_i,\beta\right),\quad i=1,2,....,n;$

ii. β is estimable or identified;

iii. $E\left[\varepsilon_i/f\left(X_i,\beta\right)\right]=0$

iv. $E\left[\varepsilon_i^2/f\left(X_j,\beta\right),\ j=1,2,..,n\right]=\sigma^2$ a finite constant and

$$E\left[\varepsilon_i\varepsilon_j/f\left(X_i,\beta\right),f\left(X_j,\beta\right),\ i,j=1,2,..,n\right]=0 \qquad \forall j\neq i$$

That is, the ε_i's are conditional homoscedastic and nonautocorrelated random error variables (or) $E\left(\varepsilon\varepsilon'\right)=\sigma^2 I_n$

v. ε_i's are normally distributed. i.e.,

$$\varepsilon_i \overset{i.i.d}{\sim} N\left(O,\sigma^2\right), \qquad i=1,2,...,n$$

By minimizing the residual sum of squares

$R\left(\hat{\beta}\right)=\left[Y-f\left(\hat{\beta}\right)\right]'\left[Y-f\left(\hat{\beta}\right)\right]$ with respect to $\hat{\beta}$, for large samples, under iterative process, an iterative nonlinear least squares (NLLS) estimator for β is given by

$$\hat{\beta}_{n+1}=\hat{\beta}_n+\left[F'\left(\hat{\beta}_n\right)F\left(\hat{\beta}_n\right)\right]^{-1}F'\left(\hat{\beta}_n\right)\left[Y-\lambda f\left(\hat{\beta}_n\right)\right] \qquad \ldots(3.3)$$

where $F\left(\hat{\beta}_n\right)=\left[\dfrac{\partial f}{\partial \beta'}\right]_{\hat{\beta}_n}$ as the regressor matrix.

Here, all the terms on the R.H.S of (3.3) are evaluated at $\hat{\beta}_n$ and $\left[Y-\lambda f\left(\hat{\beta}_n\right)\right]$ is the vector of nonlinear least squares residuals for an arbitrary value of λ.

By violating the assumption of homoscedastic errors in the nonlinear model (3.1) one may assume that

$$E\left[\varepsilon\varepsilon'\right]=\Phi=\sigma^2\psi \qquad \ldots(3.4)$$

where Φ or ψ is symmetric positive definite matrix

If the diagonal elements of dispersion matrix Φ are not all identical and ε is free from autocorrelation then Φ can be considered a diagonal matrix.

$$\Phi=\text{diag}\left(\sigma_1^2,\sigma_2^2,....,\sigma_n^2\right) \qquad \ldots(3.5)$$

and with i^{th} diagonal element is given by σ_i^2.

Define the proposed iterative NLLS residual vector based on $\hat{\beta}_n$ as $e_n=\left[Y-f\left(\hat{\beta}_n\right)\right]$.

Also, Iterative Nonlinear Internally Studentized Residuals are defined by

$$e_{ni}^*=\frac{e_{ni}}{\hat{\sigma}\sqrt{\left(1-v_{nii}\right)}},\ i=1,2,...,n \qquad \ldots(3.6)$$

where $\nu_n = \left(\left(\nu_{nij}\right)\right) = F\left(\hat{\beta}_n\right)\left[F'\left(\hat{\beta}_n\right)F\left(\hat{\beta}_n\right)\right]^{-1}F'\left(\hat{\beta}_n\right)$

…(3.7)

is symmetric idempotent matrix known as 'HAT' matrix.

$$\hat{\sigma^2} = \left[\frac{e_n' e_n}{n-p}\right] = \frac{\sum_{i=1}^{n} e_{ni}^2}{n-p} \qquad \ldots(3.8)$$

Here, $\left[\frac{e_{ni}^{*2}}{n-p}\right] \overset{asy}{\sim}$ Beta distribution with parameters ½ and $(n-p-1)/2$,

It follows that, $E\left[e_{ni}\right] = 0$ and $Var\left(e_{ni}\right) = 1, \quad \forall i = 1,2,...,n$

Also, $\operatorname{cov}\left(e_{ni}^*, e_{nj}^*\right) = \frac{-\nu_{nij}}{\left[\left(1-\nu_{nii}\right)\left(1-\nu_{njj}\right)\right]^{1/2}}, \quad \forall i \neq j = 1,2,....,n \qquad \ldots(3.9)$

Consider an estimator for Φ

$$\hat{\Phi}_n^* = \operatorname{diag}\left[e_{n1}^{*2}, e_{n2}^{*2}, ...e_{nn}^{*2}\right] \qquad \ldots(3.10)$$

Now, an Iterative Estimated Nonlinear Generalized Least Squares (IENLGLS) estimator for β is given by

$$\tilde{\beta}_{n+1}^* = \tilde{\beta}_n^* + \left[F'\left(\tilde{\beta}_n^*\right)\hat{\Phi}_n^{*-1}F\left(\tilde{\beta}_n^*\right)\right]^{-1}F'\left(\tilde{\beta}_n^*\right)\left[Y - \lambda f\left(\tilde{\beta}_n^*\right)\right] \qquad \ldots(3.11)$$

Here $F\left(\tilde{\beta}_n^*\right) = \left[\frac{\partial f}{\partial \beta'}\right]_{\tilde{\beta}_n^*}$ as the regressor matrix and $\left[Y - \lambda f\left(\tilde{\beta}_n^*\right)\right]$ is the vector of IENLGLS residuals for an arbitrary value of λ.

$$\text{Further } Var\left(\tilde{\beta}_n^*\right) = \left[F'\left(\tilde{\beta}_n^*\right)\hat{\Phi}_n^{*-1}F\left(\tilde{\beta}_n^*\right)\right]^{-1} \qquad \ldots(3.12)$$

IV. A TEST FOR HETEROSCEDASTICITY IN NONLINEAR REGRESSION MODEL BY USING ITERATIVE NLLS INTERNALLY STUDENTIZED RESIDUALS

One of the crucial assumptions of the nonlinear regression model is that the error observations have equal variances. But, in practice, it has been observed that errors are heteroscedastic. If errors are heteroscedastic, the Nonlinear Least Squares (NLLS) estimates of the parameters are inefficient and usual method of inference may produce misleading conclusions. Thus, there is need for testing the existence of the problem of heteroscedasticity in the nonlinear regression model. A wide number of tests have been developed, with a quickening of interest in the last two decades.

With usual notation, consider the nonlinear regression model

$$Y_i = f\left(X_i, \beta\right) + \varepsilon_i, \qquad i = 1,2,...,n$$

where $X_i = \left(X_{i1}, X_{i2}, ..., X_{ik}\right), i = 1,2,..,n$ is a k-componnent vector denoting the ith observation on known independent variables;

β is $(p\times 1)$ vector of unknown parameters. $\varepsilon_i, i=1,2,...,n$ are i.i.d. error random variables with mean zero and unknown unequal variances i.e., the errors are heteroscedsatic errors.

In testing the null hypothesis of homoscedsticity of errors, the following procedure may be applied:

Step (1): The observations on dependent variable Y are arranged according to the ascending order of observations on independent variable X, with which the hereroscedasticity might be associated.

Step (2): Divide the arranged data into k groups of sizes $n_1, n_2, ..., n_k$ respectively. Here, $n_1, n_2, ..., n_k$ should be approximately equal. One may choose k such that the size of each group is reasonably small and it is greater than the number of parameters in the nonlinear regression model. For instance, for a sample of 30 observations, k may be chosen as 3 such that n_1, n_2 and n_3 may be equal to 10.

Step (3): Run separate nonlinear regression models on these k groups of observations and obtain the Iterative Nonlinear Least Squares Internally Studentized Residual Sum of Squares (INLLSISRSS) for each nonlinear regression model and pooled them as $(RSS)_I$ with degrees of freedom $(n_1-p)+(n_2-p)+.....+(n_k-p)$.

Step (4): Obtain the INLLSISRSS for the combined data as (RSS) by estimating a single nonlinear regression model with degrees of freedom $\left(\sum_{j=1}^{k} n_j - p\right)$.

Step (5): Compute the F-test statistic for testing the null hypothesis of homoscedastic errors as $F = \dfrac{[RSS - RSS_I]/p}{RSS_I \Big/ \left(\sum_{j=1}^{k} n_j - kp\right)} \sim F_{\left[p, \left(\sum_{j=1}^{k} n_j - kp\right)\right]}$ and compare the calculated value of F-test statistic with its critical value and draw the inference accordingly.

V. CONCLUSIONS

Most of the mathematical statisticians have studied various inferential aspects of regression models under heteroscedasticity by using Ordinary Least Squares (OLS) residuals, Best Linear Unbiased Scalar (BLUS) residuals and Recursive residuals. Since, shortcomings of OLS residuals arise due to heteroscedastic errors, researchers have suggested several transformations of the OLS residuals to overcome partially some of their shortcomings.

Nonlinear studentized residuals have been defined to study inferential aspects of nonlinear regression models with heteroscedastic errors.

A general structure for heteroscedastic errors has been specified and estimation method has been developed to estimate the parameters of the nonlinear regression model under heteroscedastic error structure.

A new test for the problem of heteroscedasticity in nonlinear regression model has been derived by using iterative NLLS internally studentized residuals.

work.

REFERENCES

[1] Baljet, M. (1986), Heteroscedasticity in Nonlinear Regression, technical report, University of Leiden, the Netherlands, Dept. of medical statistics.

[2] Beal, S.L. and Sheiner, L.B. (1988), Heteroscedastic Nonlinear Regression, Technometrics, 30, 327-337.

[3] Carroll, R.J. and Ruppert, D. (1982a), A Comparison Between Maximum Likelihood and Generalized Least Squares in a Heteroscedastic Linear Model, J. Amer. Statist. Assoc. 77, 878-882.

[4] Carroll, R.J. and Ruppert, D. (1982b), Robust Estimation in Heteroscedastic Linear Models, The Ann.Statist. 10, 429-441.

[5] Cook, R.D and Weisberg.S.(1983), Diagnostics for Heteroscedasticity in Regression, Biometrika, 70, 1-10.

[6] Gallant, A.R. and Goebel, J.J. (1976), Nonlinear Regression with Autocorrelated Errors, Journal of the American Statistical Association, 71, 961-967.

[7] Gordon K. Smyth (2002) Nonlinear regression, Volume 3, pp 1405-1411 in Encyclopedia of Environmetrics.

[8] John Fox & Sanford Weisberg (2010), Nonlinear Regression and Nonlinear least squares in R (An appendix to An R companion to Applied Regression second edition).

[9] Rastislav Potocky and Milan Stehlik (2010), Nonlinear Regression Models with Applications in Insurance, The Open Statistics & Probability Journal, 2010, 2, 9-14.

[10] Welsh, A.H., Carroll, R.J. and Ruppert, D. (1994), Fitting Heteroscedastic Regression Models, J. Amer. Statist. Assoc. 89, 100-116.