Indian Journal of Agricultural Sciences **82** (5): 422–5, May 2012

Genetic algorithm optimization technique for linear regression models with heteroscedastic errors

M A IQUEBAL[1], PRAJNESHU[2] and HIMADRI GHOSH[3]

Indian Agricultural Statistics Research Institute, Library Avenue, New Delhi 110 012

Received: 28 July 2011; Revised accepted: 23 February 2012

ABSTRACT

Most widely used statistical technique for estimating cause-effect relationships is the Linear regression methodology. Ordinary least squares (OLS) method, which is valid under certain assumptions, is generally used to estimate the underlying parameters. If the errors are not homoscedastic, OLS estimates lead to incorrect inferences. In this article, use of the powerful stochastic optimization technique of Genetic algorithm (GA) is advocated for estimation of regression parameters and variance parameter simultaneously even when nothing is known about the form of heteroscedasticity. Parametric bootstrap methodology is employed to obtain standard errors of the estimates. The methodology is illustrated by applying it to a dataset.

Keywords: Genetic algorithm, Heteroscedasticity, Linear regression model, White's general heteroscedasticity test

Linear regression methodology is useful in explaining cause-effect relationship between a dependent (endogenous or output) variable, usually denoted by Y, and one or more independent (exogenous or input) variables, usually denoted by X_1, X_2,… More technically, linear regression is a method of estimating the conditional expected value of one variable Y given the values of some other variable or variables X. A key feature of all regression models is the error term, which is included to capture sources of error that are not captured by other variables. Linear regression models have been rigorously studied, and are very well understood. They are only appropriate under certain assumptions, and are often misused, even in published journal articles.

It is well known that under certain assumptions, the Ordinary least square (OLS) provides efficient and unbiased estimates of the parameters. One important assumption is the homoscedasticity of errors. Unfortunately, the usual practice is not to make any effort to examine it. Under heteroscedasticity, it is not advisable to use the OLS estimator. If we persist with using the usual testing procedures despite heteroscedasticity, the conclusions or inferences would be misleading. The proper approach, which is the focus of our article, is to use the powerful optimization technique of Genetic algorithm (GA) for estimation of regression parameters and variance parameter simultaneously even when nothing is known about the form of heteroscedasticity. The GA is a target-oriented parallel search technique, mainly applied to optimization process searching for universal or nearly universal extreme values. It processes a population of individuals, which presents search space solution, employing three operators, viz. selection, crossover and mutation. A heartening aspect of GA is that it is capable of obtaining global optimum solution of parameter estimates.

In the present paper, a brief description of linear regression models, definition of heteroscedasticity problem, testing of heteroscedasticity, brief discussion of GA methodology used to tackle the problem efficiently an illustration of the methodology for data on expenditure on total food and total expenditure of 55 rural households from India are attempted.

[1]Scientist (e mail: asif@iasri.res.in), Division of Biometrics and Statistical Modelling, [2]Principal Scientist and Head (e mail: prajneshu@yahoo.co.in), Division of Biometrics and Statistical Modelling, [3]Senior Scientist (e mail: hghosh@gmail.com), Division of Biometrics and Statistical Modelling

MATERIALS AND METHODS

Linear regression model and heteroscedasticity

The linear regression model, in which response variable **(y)** is modelled as a linear systematic component plus error, is typically stated as follows:

$$\mathbf{y} = X\boldsymbol{\beta} + \varepsilon \qquad \ldots(1)$$

where y is an $n \times 1$ vector of observations, X is a fixed matrix of dimension $n \times p$ with full column rank (rank (X) = p<n)

containing explanatory variables, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)$ is a p-vector of unknown linear parameters, and $\boldsymbol{\varepsilon}$ is an *n*-vector of errors, having mean zero and variance σ_i^2. We denote the covariance matrix of $\boldsymbol{\varepsilon}$ as $\Sigma = \text{diag}\ \{\sigma_1^2, \sigma_2^2, ..., \sigma_n^2\}$. When the errors are homoscedastic, $\sigma_i^2 = \sigma^2 > 0$, i.e. $\Sigma = \sigma^2 I_n$, where I_n is an identity matrix of order n. The ordinary least squares (OLS) estimate of $\hat{\boldsymbol{\beta}}$ is given by $\hat{\beta} = (X'X)^{-1}X'\mathbf{y}$, which has mean $\hat{\boldsymbol{\beta}}$ (i.e. it is unbiased) and variance structure $(\hat{\boldsymbol{\beta}}) = \psi$, with $\psi = \sigma^2(X'X)^{-1}$.

It is well known that when the assumptions of the linear regression model are not violated, the OLS provides efficient and unbiased estimates of the parameters. Heteroscedasticity occurs when the variance of the errors varies across observations. When the errors are heteroscedastic, the OLS estimator remains unbiased, but becomes inefficient. In addition, the standard errors are biased when heteroscedasticity is present. This, in turn, leads to bias in test statistics and confidence-intervals. More importantly, usual procedures for hypothesis testing are no longer appropriate. Given that heteroscedasticity is common in cross-sectional data, methods that correct for heteroscedasticity are essential for prudent data analysis.

When the form and magnitude of heteroscedasticity are known, Generalized least squares technique could be employed to correct for heteroscedasticity. If the form of heteroscedasticity involves a small number of unknown parameters, variance of each residual can be estimated first and these estimates can be used as weights in the second step. In many cases, however, the form of heteroscedasticity is unknown, which makes the weighting approach impractical. When heteroscedasticity is caused by an incorrect functional form, it can be corrected by making variance-stabilizing transformations of the dependent variable (Weisberg 2005) or by transforming both sides (Carroll and Ruppert 1988).

White's general heteroscedasticity test

Unlike the Goldfeld-Quandt test, which requires recording the observations with respect to the X variable that supposedly cause heteroscedasticity, or the Breusch-Pagan-Godfrey (BPG) test (designed to detect any linear form of heteroscedasticity), which is sensitive to the normality assumption, the general test for heteroscedasticity proposed by H. White does not rely on the normality assumption and is easy to implement. The White test (Gujarati 2003) proceeds as follows:

Step 1. Given the data, obtain the residuals $\hat{u}_i$.

Step 2. The squared residuals from the original regression are regressed on the original X variables or regressors, their squared values, and the cross product(s). It is necessary to add a constant term in this equation even though the original regression may or may not contain it. Obtain the R^2 from this (auxiliary) regression.

Step 3. Under the null hypothesis that there is no heteroscedasticity, it can be shown that the sample size (n) times R^2 obtained from the auxiliary regression asymptotically follows the chi-square distribution with degrees of freedom equal to the number of regressors (excluding the constant term) in the auxiliary regression. That is,

$$n R^2 \underset{\text{asy}}{\sim} \chi^2_{\text{df}}$$

Step 4. If the chi-square value obtained in Step 3 exceeds the critical chi-square value at the chosen level of significance, the conclusion is that there is heteroscedasticity.

Suggested procedure using Genetic algorithm approach

Fortunately, a very powerful and versatile optimization technique of Genetic algorithm (GA), motivated by the principles of genetics and natural selection, has recently been developed (Goldberg 1989). It is a combination of Charles Darwin's principle of 'natural selection' and 'survival of the fittest' with computer-constructed evolution mechanism to select better offspring from the parent population. Then the information is exchanged randomly among parents, expecting a superior offspring. Besides, in order to avoid missing some good species and becoming a local optimization, several mutations must be processed. In this methodology, some fundamental ideas of genetics are borrowed and used artificially to construct search algorithms that are robust and require minimal problem information. The three operators, viz. selection, crossover and mutation, make GA an important tool for optimization. When a string (parameter solution) is created by GA, it is evaluated in terms of its fitness (objective function), which is taken to be the Residual sum of squares (RSS).

Selection operator of GA is performed to identify good solutions, to make its multiple copies and to eliminate bad solutions from the population. The most common method of selection that has better convergence and computational time efficiency is the Tournament selection (Deb 2002). Since selection operator cannot create a new solution, the crossover and mutation operators are used to create a new population so that global minima may be achieved in succeeding generations. There exists a number of crossover operators due to which different string pairs are expected to have some good bit representation. Such an operator should not be allowed to use all strings in a population to preserve some good strings selected during the selection operator. If a crossover probability p_c is used, $100p_c\%$ strings in the population are used in the crossover and rest of the population is simply copied to the new population.

In real-coded GA, a pair of real-parameter decision variable vector is used to create a new pair of offspring vectors by applying crossover operator. Here, an important operator of this type, called Simulated binary crossover operator, was developed by Deb and Agrawal (1995). Two offspring $x_i^{(1, t+1)}$ and $x_i^{(2, t+1)}$ are produced from two parent solutions $x_i^{(1, t)}$ and $x_i^{(2, t)}$, where $x_i^{(j, t)}$ is the value of i^{th}

variable of j^{th} parent in t^{th} generation, i = 1, 2, . . ., p; j = 1, 2; and $t \geq 1$. To this end, after drawing the spread factor β_i from the probability distribution with mode unity (Deb 2002), the offspring are calculated as:

$$x_i^{(1,\,t+1)} = 0.5\ \{(1 + \beta_{qi})\ x_i^{(1,t)} + (1 - \beta_{qi})\ x_i^{(2,t)}\}$$
$$x_i^{(2,\,t+1)} = 0.5\ \{(1 - \beta_{qi})\ x_i^{(1,t)} + (1 + \beta_{qi}) x_i^{(2,t)}\} \quad \ldots(2)$$

These offspring are symmetrically distributed and the points of symmetry are equispaced from the mid-point of parent solutions. Thus, biasedness towards any particular parent solution is avoided. Another important aspect of this crossover operator is that for a fixed ηc' the offspring have a spread which is proportional to that of the parent solution, i e

$$x_i^{(2,\,t+1)} - x_i^{(1,\,t+1)} = \beta_{qi}\ (x_i^{(2,t)} - x_i^{(1,t)}) \quad \ldots(3)$$

The mutation operator is used to ensure diversity in the population. It alters a string locally to create a better string (parameter solution). If certain genes (digits) of all chromosomes (strings) in the population are identical, their values will never change after selection and crossover. This will reduce the chance for some new chromosomes to enter this population, thereby lending the GA-procedure into the trap of local optima. To avoid this situation, mutation with very small probability, say 0.01 is required. In real-coded GA, mutation operator does the same task as performed by real parameter crossover operator. The polynomial mutation operator has the advantage that the probability distribution does not change with generations, thereby avoiding local optima and so this mutation operator is used in present study. It mutates the i^{th} variable $x_i^{(1,\,t+1)}$ to $y_i^{(1,\,t+1)}$ by the transformation:

$$y_i^{(1,\,t+1)} - x_i^{(1,\,t+1)} + \left\{x_i^{(U)} - x_i^{(L)}\right\}\bar{\delta}_i \quad \ldots(4)$$

where $\bar{\delta}_i$ follows polynomial probability distribution.

The advantage of GA over other optimization methods is that it works with a population of solutions instead of a single solution. GA uses probabilistic rules and an initial random population to guide the search, unlike other traditional methods which employ fixed transition rules to move from one solution to another. Another advantage of GA is that it reduces the overall computational time substantially. The exploitation and exploration aspects of GAs can be controlled almost independently. This provides a lot of flexibility in designing a GA.

RESULTS AND DISCUSSION

As an illustration, data on expenditure on total food (dependent) and total expenditure (independent), measured in Rupees, for a sample of 55 rural households from India, as given in Gujarati (2003), is considered to test presence of heteroscedasticity, White's general heteroscedasticity test is applied to the data. The calculated value of chi-square is obtained as 7.29, which exceeds the tabulated value of 5.99 at 5% level of significance, implying presence of heteroscedasticity. Hence, it is not advisable to apply OLS for estimation of parameters. Attempts are then made to fit one variable linear regression equation with heteroscedastic error of the following form using above described GA methodology:

$$\sigma_i^2 = \alpha_0 + \alpha_1 X_i^2$$

To this end, the fitness (objective) function to be minimized may be written as:

$$\sum_{t=1}^{T}\left[(Y_i - \beta_0 - \beta_1 X_i)\Big/\mathrm{sqrt}\left(\alpha_0 + \alpha_1 X_i^2\right)\right]^2 \quad \ldots(5)$$

Relevant computer programmes are developed using C-language in Microsoft visual C++ compiler and could be obtained from first author on request. The GA parameters, viz. population size, crossover probability and mutation probability for minimization of equation (5), are respectively computed as 500, 0.9, 0.01 with number of generations as 100. Using above parameter set up, GA has generally terminated with accuracy level ($\varepsilon = 10^{-3}$) in 99 out of 100 runs. This confirms that in presence of heteroscedasticity in linear regression model, the regression coefficients as well as variance function are successfully estimated simultaneously using GA-optimization technique. The standard errors for parameter estimates are obtained using bootstrap technique. The minimum value of the fitness function is found as 0.055. The root mean squares error (RMSE) value is computed as 0.032, which is much lower than the RMSE values (65.97) obtained through OLS procedure. The parameter estimates along with estimated bias and standard errors obtained through GA are reported in Table 1. It may be noted that the percentage standard errors throughout is generally quite low, indicating thereby that the parameters are estimated efficiently.

Utility of genetic algorithm optimization methodology for fitting of linear regression model with heteroscedastic errors is highlighted. The proposed procedure is successfully demonstrated through a dataset. There is no hard and fast rule for optimum population size for applying this technique effectively in real life research problems but it is advisable that there should be at least 30 data points. The significance

Table 1 Parameter estimates with standard errors and estimated bias

Parameter		Estimated value	Estimated bias	Standard error
Regression	β_0	135.74	–0.05	8.56
	β_1	0.35	–0.01	0.09
Variance	α_0	24.92	–0.74	5.78
	α_1	10.34	–0.16	2.93

of this work is that this methodology is applicable even in those cases in which we do not know the form of heteroscedasticity and Least squares methodology is not applicable. The results obtained using this technique may be used for efficient prediction of response variable for a given value of explanatory variable. Possibility of modifying the methodology when errors are not independent but follow autocorrelated AR(1) errors is currently being explored and the same shall be reported separately in due course of time. As future research work, methodology needs to be developed for application of GA to the situation when functional relation between dependent and independent variables is nonlinear.

REFERENCES

Carroll R J and Ruppert D. 1988. *Transformation and Weighting in Regression*. Chapman & Hall, London.

Deb K. 2002. *Multi-Objective Optimization using Evolutionary Algorithms*. John Wiley, Singapore.

Deb K and Agrawal R B. 1995. Simulated binary crossover for continuous search space. *Complex Systems* **9**: 115–48.

Goldberg D E. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Pub. Co., USA.

Gujarati D N. 2003. *Basic Econometrics*, edn 4. McGraw-Hill, New York.

Weisberg S. 2005. *Applied Linear Regression*. edn 3. John Wiley, USA.