Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 9 (22): 4013-4019, 2009
ISSN 1812-5654

Estimation of Parameters in Heteroscedastic Multiple Regression Model using Leverage Based Near-Neighbors

[1]H. Midi, [1]S. Rana and [2]A.H.M.R. Imon
[1]Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia
[2]Department of Mathematical Sciences, Ball State University, Muncie, IN 47306, USA

Abstract: In this study, we propose a Leverage Based Near-Neighbor (LBNN) method where prior information on the structure of the heteroscedastic error is not required. In the proposed LBNN method, weights are determined not from the near-neighbor values of the explanatory variables, but from their corresponding leverage values so that it can be readily applied to a multiple regression model. Both the empirical and Monte Carlo simulation results show that the LBNN method offers substantial improvement over the existing methods. The LBNN has significantly reduced the standard errors of the estimates and also the standard errors of residuals for both simple and multiple linear regression models. Hence, the LBNN can be established as one reliable alternative approach to other existing methods that deal with heteroscedastic errors when the form of heteroscedasticity is unknown.

Key words: Weighted least squares, near-neighbors, leverages, Monte Carlo simulation

INTRODUCTION

One of the crucial assumptions in the linear model is that the variance of the errors is constant. The Ordinary Least Squares (OLS) method is very popular with statistics practitioners as it provides efficient and unbiased estimates of the parameters when the assumptions, especially the assumption of homoscedastic error variances are met. But in many real applications, variances of the errors vary across observations. Since, homoscedasticity is often an unrealistic assumption, researchers should consider how the results are affected by heteroscedasticity. Eventhough the OLS estimates retain unbiasedness in the presence of heteroscedasticity, its estimates become inefficient. Heteroscedasticity yield hypothesis tests that fail to keep false rejections at the nominal level, or estimated standard errors as well as confidence intervals that are either too narrow or too large. As heteroscedasticity is a common problem in cross-sectional data analysis, it is very important to find a reliable method that can correct such problem for prudent data analysis.

A good number of works is now available in the literatures (Carroll and Ruppert, 1982, 1988; Robinson, 1987; Montgomery *et al*., 2001; Gujarati, 2003; Chatterjee and Hadi, 2006; Greene, 2008) for correcting the problem of heteroscedasticity. When the form and magnitude of heteroscedasticity are known, the correction for heteroscedasticity is very simple by means of the weighted least squares. The WLS is equivalent to performing OLS on transformed variables. Nevertheless, in real situation it is not easy to get a suitable transformation, especially for multiple regression, where some predictors may require transformations and others may not. If the form of heteroscedasticity involves a small number of unknown parameters, the variance of each residual can be estimated first and these estimates can be used as weights in a second step. Nevertheless, in practice, 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 variables (Welsberg, 1980) or by transforming both sides (Carroll and Ruppert, 1988). Although, these approaches can provide an efficient and elegant solution to the problems caused by heteroscedasticity, when the results need to be interpreted in the original scale of the variables, nonparametric methods may be necessary (Carroll and Ruppert, 1988). As noted by Emerson and Stoto (1983), re-expression moves us into a scale that is often less familiar. Furthermore, if there are theoretical reasons to believe that errors are heteroscedastic around the correct functional form, transforming the dependent variable is inappropriate.

Corresponding Author: Sohel Rana, Laboratory of Applied and Computational Statistics, Institute for Mathematical Research, University Putra Malaysia, 43400 Serdang, Selangor, Malaysia

Consider the general multiple linear regression model:

$$y_i = x_i^T\beta + \varepsilon_i \ , \quad i = 1,2,\cdots\cdots,n \tag{1}$$

where, y_i is the ith observed response, x_i is a p×1 vector of predictors including the intercept, β is a p×1 vector of unknown finite parameters and ε are uncorrelated random errors with mean 0 and variance σ^2. Writing $y = (y_1,y_2,...,y_n)^T$, $X = (x_1,x_2,...,x_n)^T$ and $\varepsilon = (\varepsilon_1,\varepsilon_2,...,\varepsilon_n)^T$, model (1) can be written as:

$$y = X\beta+\varepsilon \tag{2}$$

Inferences about β may be based on the fact that $(\hat{\beta}-\beta)$ has mean zero and covariance matrix:

$$V(\hat{\beta}) = (X^TX)^{-1}X^T\Omega X(X^TX)^{-1} \tag{3}$$

where, $E(\varepsilon\varepsilon^T)$, a positive definite matrix. If the errors are homoscedastic, that is, $\Omega = \sigma^2 I_n$, Eq. 3 simplifies to:

$$V(\hat{\beta}) = \sigma^2(X^TX)^{-1} \tag{4}$$

and $\hat{\beta} = (X^TX)^{-1}X^Ty$ is an unbiased and consistent estimator of β. If the errors are heteroscedastic, that is, $\Omega = \sigma^2 V$, Eq. 3 becomes:

$$V(\hat{\beta}) = \sigma^2(X^TX)^{-1}X^TVX(X^TX)^{-1} \tag{5}$$

The above problem can be solved by transforming the model to a new set of observations that satisfy the standard least squares assumptions. Then the OLS is applied on the transformed data. Since, σ^2 is the covariance matrix of the errors, V must be nonsingular and positive definite and:

$$\hat{\beta}_{GLS} = (X^TV^{-1}X)^{-1}X^TV^{-1}y \tag{6}$$

is the Generalized Least Squares (GLS) estimates of β. When the errors ε are uncorrelated but have unequal variances, the covariance matrix of ε is written as:

$$\sigma^2 V = \text{Diag}\,[1/w_i],\ i = 1, 2, \ldots, n$$

Consequently, the GLS is the solution of the heteroscedastic model. If we define $W = V^{-1}$, W becomes a diagonal matrix with diagonal elements or weights $w_1,w_2,...,w_n$. From Eq. 6, the weighted least squares estimator is $\hat{\beta}_{WLS} = (X^TWX)^{-1}X^TWy$ and $V(\hat{\beta}_{WLS}) = \sigma^2_{WLS}(X^TWX)^{-1}$.

Where:

$$\sigma^2_{WLS} = \frac{\sum w_i\hat{\varepsilon}_i^2}{n-p}$$

If the heteroscedastic error structure of the regression model is known, it is easy to compute the weights of W matrix and consequently the WLS would be a good solution of heteroscedastic regression model.

Unfortunately, in practice, the structure of the heteroscedastic error is unknown. In the presence of heteroscedasticity, as there is no biasness problem of the OLS estimator, a Heteroscedasticity Consistent Covariance Matrix (HCCM) proposed by White (1980), is used to solve the consistency problem of the estimator. Defining the residuals, $\hat{\varepsilon}_i = y_i - x_i\hat{\beta}$, where, x_i is the ith row of X, the OLS covariance matrix (OLSCM) of estimates of the regression parameters is:

$$\text{OLSCM} = \frac{\sum\hat{\varepsilon}_i^2}{n-p}(X^TX)^{-1} \tag{7}$$

The OLSCM is appropriate for hypothesis testing and also for computing confidence intervals when the standard assumptions of the regression model, including homoscedasticity, hold. When there is heteroscedasticity, tests based on the OLSCM are likely to be misleading since the Eq. 4 will not generally be equal to the Eq. 3. Equation 3 can be used to correct for heteroscedasticity, if the errors are heteroscedastic and Ω is known. In many applications, the form of the heteroscedasticity is unknown and the HCCM may be an alternative method.

However, there is no general agreement among statisticians about which of the four estimators of the HCCM (HC0, HC1, HC2, HC3) should be used (Judge *et al*., 1988; MacKinnon and White, 1985; Davidson and MacKinnon, 1993; Long and Ervin, 2000; Gujarati, 2003; Greene, 2008).

MATERIALS AND METHODS

The leverage based near-neighbor method: Several attempts have been made in the literature other than the HCCM to obtain consistent estimators when the form of heteroscedasticity is unknown. A good review of these methods is available in Montgomery *et al*. (2001) and Chatterjee and Hadi (2006). Moreover, Montgomery *et al*. (2001) and Chatterjee and Hadi (2006) proposed several WLS methods to solve this problem by developing weighting techniques that are later used for estimating the parameters of a heteroscedastic model. Instead of fitting regression with all the data, Montgomery *et al*. (2001) suggested finding several near-neighbor groups in the explanatory variable. We refer to this

method as the Montgomery, Peck and Vining (MPV) method. The group means would now represent the explanatory variables (X). The groups in the response variable Y is formed in accordance with the groups formed in X. The sample variance of each groups of Y and the mean of each group of X are then computed. These group variances in Y are then regressed on the corresponding group mean of X. In the presence of heteroscedasticity we expect variations in errors among these groups. Hence the inverse of the fitted response can be used as weights for stabilizing variance heterogeneity. The values of X are first sorted to form near neighbor groups. The main limitation of this method is that it cannot be applied to multiple linear regression model as we are not able to sort the values of X for more than one explanatory variables and we are unable to make group means and group variances for a multiple linear regression model. In this situation, it is quite impossible to find near neighbors of the X's. Chatterjee and Hadi (2006) propose a two step estimation technique that can be used for multiple regression model. We refer to this method as the Chatterjee and Hadi (CH) method. In this method we need some known data that are grouped in a natural way. It is assume that there is a unique residual variance associated with each of the group. Let us consider we have g known groups and the variances are denoted as $(c_j\sigma)^2$, j = 1,2,3,... ...,g, where the residual standard error, σ is the common part and c_j's are unique to the groups. According to the principle of weighted least squares, the regression coefficients should be determined by:

$$\text{Minimizing } S_w = \sum_{j=1}^{g} S_j$$

where:

$$S_j = \sum_{i=1}^{n_j} \frac{1}{c_j^2}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \ldots \ldots - \beta_k x_{ik}) ; j = 1, 2, \ldots, g \quad (8)$$

The factors $1/c_j^2$ are the weights that determine how much influence each observation has in estimating the regression coefficients. The estimate of c_j^2 is define as:

$$\hat{c}_j^2 = \frac{\hat{\sigma}_j^2}{n^{-1}\sum_{i=1}^{n} e_i^2}$$

The main advantage of this method is that it can be employed to multiple regression model. But its main limitation is that it can be applied only in situations where we definitely know the heteroscedastic groups. Nevertheless, this method requires prior information about the heteroscedastic structures. Without this information, it is not possible to construct the near neighbor groups of the X's.

In this study we refer the works of Montgomery *et al*. (2001) and Chatterjee and Hadi (2006) as MPV and CH methods, respectively. However, their works have motivated us to develop a new method in multiple linear regression to obtain consistent estimators when the form of heteroscedasticity is unknown. As already been mentioned the MPV can only be applied to simple linear regression when nothing is known about the form of heteroscedasticity. On the other hand, the CH can be applied to the multiple linear regression but this method requires prior information on the structures of the heteroscedastic errors. An attempt has been made to make a compromise between these two approaches. Hence, in this study a Leverage Based Near-Neighbor (LBNN) method similar to that of MPV method is proposed. We extend the idea of Montgomery *et al*. (2001) in multiple linear regression when the structure of heteroscedasticity is unknown. We propose using the leverage values of the explanatory variables as a basis of finding the "near-neighbor" groups. Specifically, the weights of the LBNN are determined not from the near-neighbor values of the explanatory variables, but from their corresponding leverage values so that it can be readily applied to a multiple regression model. We observe that the ith leverage value, denoted as $h_{ii} = x_i(X^TX)^{-1}x^T_i$, i = 1, 2, …, n, is the standardized measure of the distance of the ith observation from the centre (or centroid) of the x-space. Thus, we can form the near neighbor groups of X by using the leverage values and this method can be used not only for a single explanatory variable as proposed by Montgomery *et al.* (2001), but also for multiple explanatory variables. The main attraction of this method is that it can be implemented regardless of whether the prior information about the structure of heteroscedasticity is known or not. The proposed LBNN algorithm consists of the following steps:

Step 1: Finding near-neighbor groups

- Compute the diagonal elements h_{ii} of the hat matrix. Corresponding to each h_{ii}, there are y_i and x_{ij}, where i = 1, 2, …, n; j = 1, 2, …, p
- Arrange h_{ii}'s in increasing order, carrying along the y_i and the x_{ij}. Then determine the near-neighbor groups of the explanatory variables X's according to the sorted diagonal elements of the hat matrix H. A number of clustering methods is available in the literature for finding near-neighbor groups. We suggest using the clustering method proposed by Struyf *et al*. (1997) because it is very popular and effective technique and as a result is used as a default in S-PLUS. If our data set contains

g groups of leverages, we obtain $Y_{(i)k}$ and $X_{(i)jk}$ where, (i) = 1, 2, …, g, j = 1, 2, …, p, where, p is the number of parameters and k = 1, 2, …,n_i, where n_i is the number of observations in each near neighbor group

Step 2: Determining weights for near-neighbor groups

- Compute variances of all groups of the response variable, denoted as var($y_{(i)k}$), corresponding to the near-neighbor groups of X. Also compute the mean of each g groups for each explanatory variables, denoted as $\bar{X}_{(i)j}$
- Regress var ($y_{(i)k}$) on the mean ($\bar{X}_{(i)j}$), I = 1, 2, …, g; j = 1, 2, …, p, by the OLS method and obtain the regression coefficients from this fitting
- Obtain the linear regression line of y on x's by using the parameter estimates computed in step 4. Calculate the fitted values of y based on the values of the variables X's
- The inverse of these absolute fitted values denoted by w_i will be reasonable estimates of the weights

Step 3: Heteroscedasticity correction

- Finally perform a WLS regression using weights w_i. The regression coefficients obtained from this WLS regression are the desired estimate of the heteroscedastic model

RESULTS

Numerical examples: In order to evaluate performance of the LBNN method, two real data sets are considered.

Restaurant food sales data: Our first example presents the average monthly income (y) corresponding to one predictor variables of 30 restaurants taken from Montgomery *et al*. (2001). Here we wanted to show that our proposed method can also be applied to a single predictor variable. We apply the proposed LBNN methods to this data and compare its performance with the traditionally used OLS and recently developed MPV methods. To apply the MPV to the data, we first need to determine the near neighbor groups. By examining the Restaurant Food Sales Data, we observe that there are several sets of x values that are near-neighbors, that is, observations which have approximate repeat points of x. We will assume that these near neighbors are close enough to be considered as repeat points. Once the near neighbors are identified, the variance of the response at those considered repeat points are computed and then we observe how var(y)

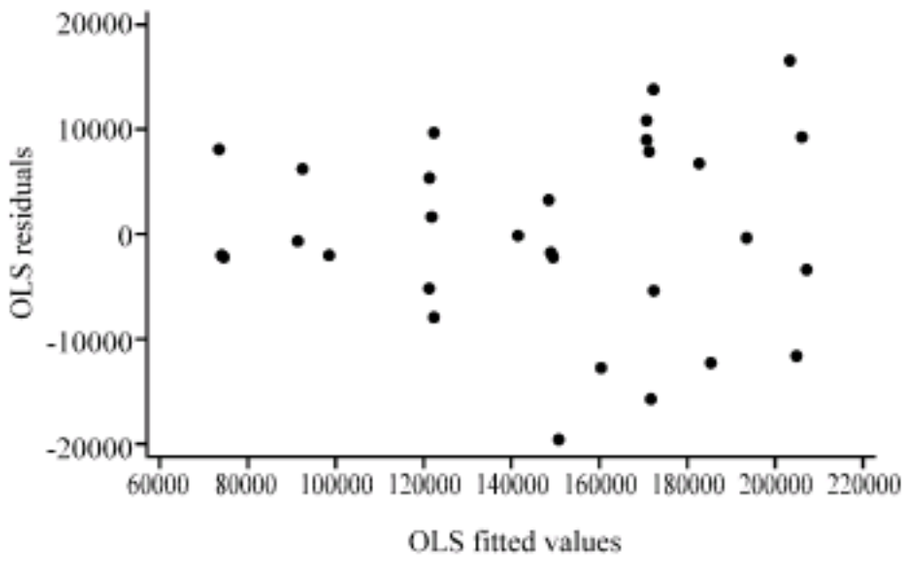

Fig. 1: Plot of OLS residuals versus fitted values for the restaurant food sales data

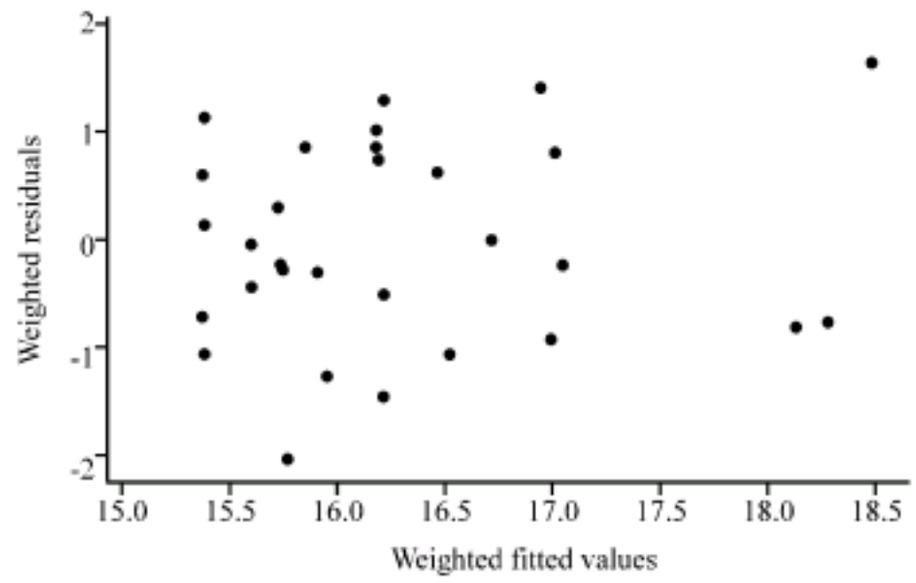

Fig. 2: Plot of MPV residuals versus fitted values for the restaurant food sales data

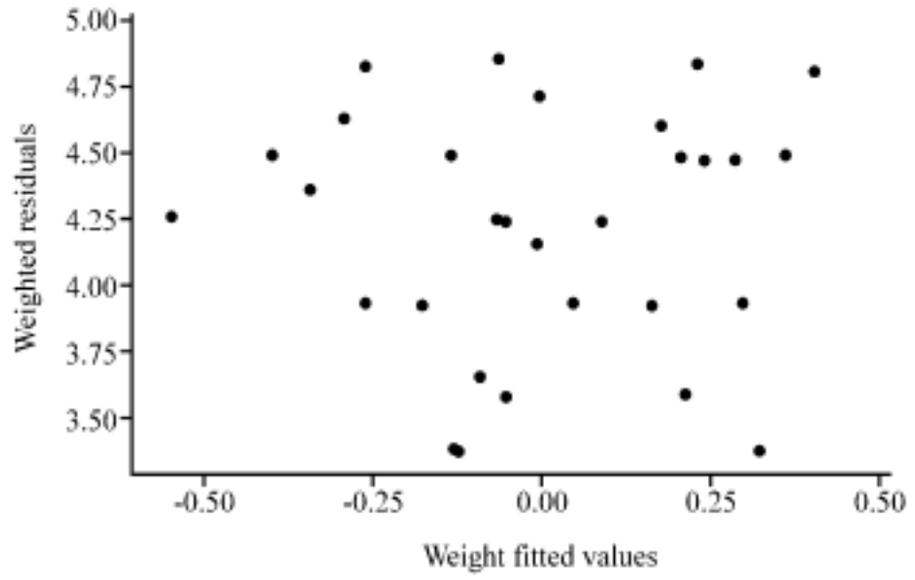

Fig. 3: Plot of LBNN residuals versus fitted values for the restaurant food sales data

changes with x. The weights, w_i are obtained according to the MPV algorithm. The residuals from the OLS fit are plotted against $\hat{y}$ in Fig. 1.

The MPV residuals are plotted against the weighted fitted values of $\hat{y}$ in Fig. 2.

Figure 3 displays the residuals of the LBNN estimates against the LBNN fitted values of $\hat{y}$.

Table 1: Summary statistics for the restaurant food sales data

Method	Coefficients	Estimates	SE (Estimates)	SE (Res)
OLS	α	49504.3900	4277.8193	8975.8500
	β	8.0489	0.3257	
MPV	α	50106.5986	3630.6911	0.5924
	β	7.9992	0.2981	
LBNN	α	50515.7586	3130.1862	0.2489
	β	7.9655	0.2767	

Table 2: Summary statistics for the education expenditure data

		Coefficients				
Method	Statistical analysis	β_0	β_1	β_2	β_3	SE (Res)
OLS	Value	-556.560	0.07200	1.5520	-0.0040	40.470
	SE	123.200	0.01200	0.3150	0.0510	
CH	Value	-496.676	0.07510	1.3431	-0.0126	38.506
	SE	101.549	0.00970	0.2642	0.0448	
HC0	SE	172.670	0.01570	0.4242	0.0559	
HC1	SE	179.864	0.01630	0.4419	0.0582	
HC2	SE	222.483	0.01990	0.5415	0.0673	
HC3	SE	290.586	0.02570	0.7025	0.0834	
LBNN	Value	-453.830	0.06900	1.2580	0.0109	0.799
	SE	109.378	0.00109	0.2590	0.0444	

The standard errors of the parameters estimates and residual standard errors for the three methods are exhibited in Table 1.

Education expenditure data: This data is taken from Chatterjee and Hadi (2006). It represents the relationship between response variable and three independent variables for all 30 states in USA. It is worth mentioning here that the structure of the heteroscedasticity of this data is known since the states are grouped according to geographical regions. The variables names are as follows:

Y: Per capita income on education projected for 1975
X_1: Per capita income in 1973
X_2: Number of residents per thousand under 18 years of age in 1974
X_3: Number of residents per thousand under 18 years of age in 1974

The states are grouped according to geographical regions based on the pre-assumption that there exists a sense of regional homogeneity. The four geographic regions, (1) Northeast, (2) North centre, (3) South and (4) West, are used to define the groups. It is important to point out that since this data set contains three explanatory variables, we cannot apply the MPV method here. Figure 4 presents the OLS residual versus fitted plot for this data.

Figure 5 and 6 show the residuals versus the fitted plots for the CH and the LBNN methods, respectively.

Table 2 presents the summary statistics of different estimation techniques for the education expenditure data.

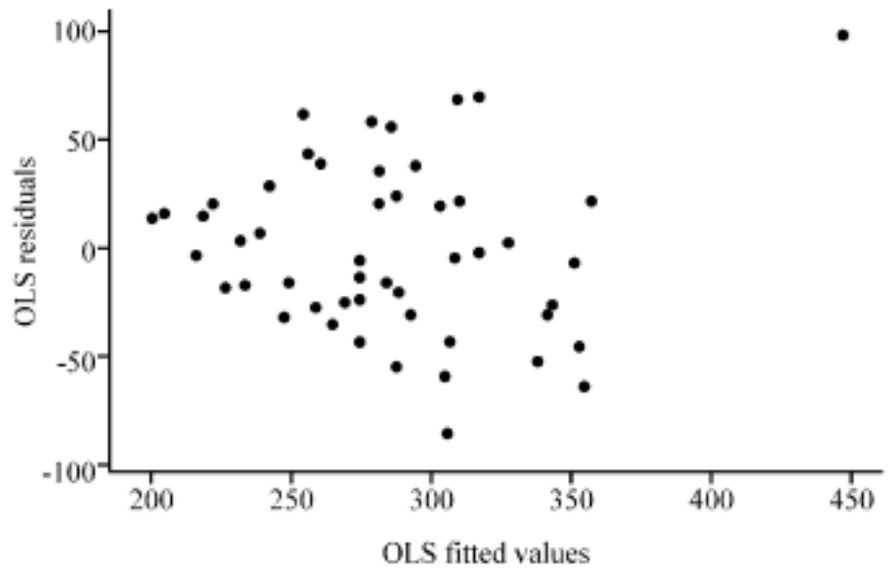

Fig. 4: Plot of OLS residuals versus fitted values for the education expenditure data

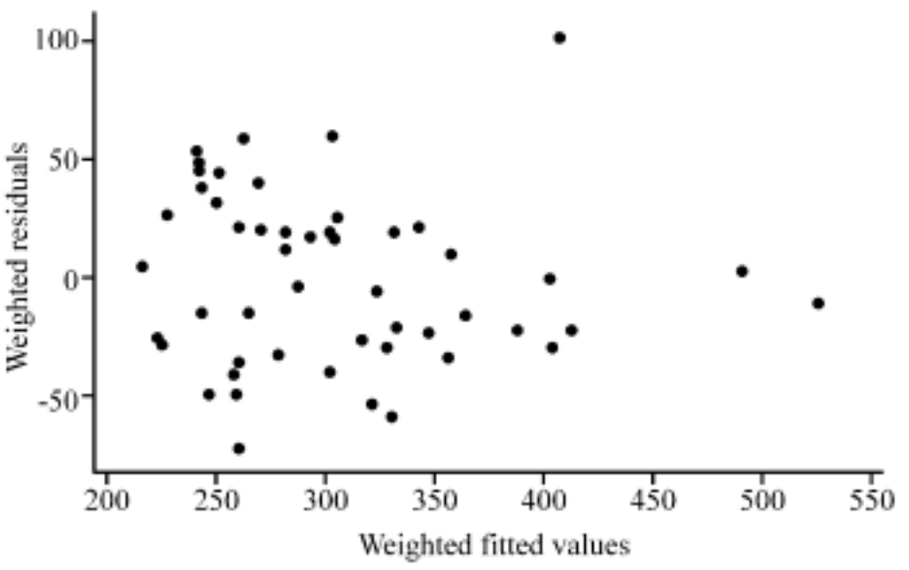

Fig. 5: Plot of CH residuals versus fitted values for the education expenditure data

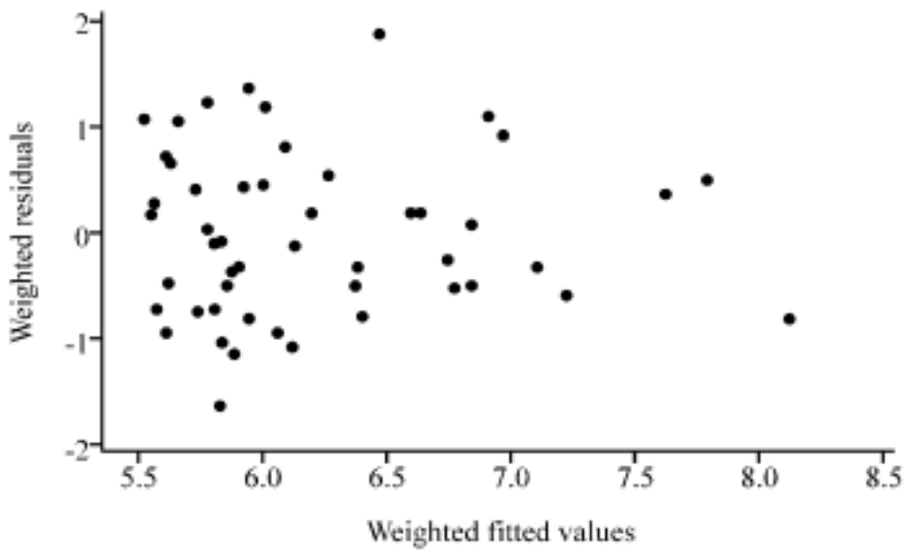

Fig. 6: Plot of LBNN residuals versus fitted values for the education expenditure data

Monte carlo simulation: Here, we report a Monte Carlo simulation study that is designed to investigate different heteroscedasticity correction methods under a variety of situations. In this experiment we consider a true multiple regression model:

$$y_i = 10 + 2X_{1i} + 2.5X_{2i} + 3X_{3i} + \varepsilon_i \quad (9)$$

Table 3: Simulated summary statistics for coefficient β_0 (True value = 10)

		Sample size				
Methods	Statistical analysis	30	60	100	250	400
OLS	Estimate	10.0031	9.9977	10.0049	10.0312	9.9841
	SE	1.1168	1.1736	1.3538	2.0813	2.4836
HC0	SE	1.0391	1.2435	1.3419	2.0543	2.4731
HC1	SE	1.1989	1.3323	1.3978	2.0877	2.4980
HC2	SE	1.1219	1.3017	1.3748	2.0743	2.4869
HC3	SE	1.2166	1.3644	1.4093	2.0948	2.5009
LBNN	Estimate	9.8860	10.0294	10.0353	10.0477	9.9896
	SE	0.7620	0.6526	0.6962	0.9079	0.7526

Table 4: Simulated summary statistics for coefficient β_1 (True value = 2)

		Sample size				
Methods	Statistical analysis	30	60	100	250	400
OLS	Estimate	1.9930	2.0002	2.0025	1.9491	2.0071
	SE	1.6938	1.7064	2.1369	3.3717	4.0312
HC0	SE	1.5684	1.7203	2.1091	3.3488	4.1002
HC1	SE	1.8097	1.8432	2.1970	3.4032	4.1416
HC2	SE	1.7015	1.8013	2.1573	3.3823	4.1233
HC3	SE	1.8493	1.8898	2.2070	3.4166	4.1466
LBNN	Estimate	2.1573	1.9907	1.9707	1.9808	2.0189
	SE	1.2623	0.8938	1.0802	1.1440	1.2058

Table 5: Simulated summary statistics for coefficient β_2 (True value = 2.5)

		Sample size				
Methods	Statistical analysis	30	60	100	250	400
OLS	Estimate	2.4959	2.5006	2.4845	2.5036	2.5180
	SE	0.4091	0.4965	0.5620	0.8689	1.1544
HC0	SE	0.3744	0.5152	0.5608	0.8475	1.1159
HC1	SE	0.4320	0.5520	0.5841	0.8613	1.1272
HC2	SE	0.4136	0.5500	0.5757	0.8582	1.1239
HC3	SE	0.4601	0.5886	0.5911	0.8691	1.1319
LBNN	Estimate	2.4818	2.4815	2.4893	2.4827	2.5310
	SE	0.3532	0.2734	0.2739	0.3426	0.4237

where, X_{1i} is uniformly distributed on [0,1], X_{2i} is also normally distributed on [0,1] and X_{3i} is distributes as X^2_{1df}.

In order to generate heteroscedastic errors, the random errors ε_i's are drawn from normal distribution, i.e., $\varepsilon_{ij} \sim N(0, \sigma^2_j$, I = 1, 2, …, n and j=1,2,…, g where, g is the number of error groups in each sample. Each error group consists of 10 random errors. The value of σ^2_j in each error group corresponds to j. Here we consider five different sample sizes 30, 60, 100, 250 and 400 which are kept fixed over repeated samples. We assign different random errors for different sample sizes. For example, for generating 30 random errors we take the first 10 random errors from N(0,1), the second 10 from N(0,2) and the third 10 random errors from N(0,3). For these 30 random errors, it is obvious that although the mean of the errors are zero but their variances are not constant. Similarly, for generating 60 random errors, the first 10 random errors are generated from N(0,1), the second 10 from N(0,2), the third 10 from N(0,3), … and the last 10 random errors from N(0,6). The random errors for other sample sizes are generated in the manner describe earlier. The OLS, HC0, HC1, HC2, HC3 and LBNN are then applied to these data. In each simulation run, there were 10,000 replications. Table 3-6 illustrate the average measures of the regression coefficients and their corresponding standard errors while Table 7 displays the simulated residuals standard errors.

Table 6: Simulated summary statistics for coefficient β_3 (True value = 3)

		Sample size				
Methods	Statistical analysis	30	60	100	250	400
OLS	Estimate	2.9972	2.9994	2.9961	3.0047	3.0048
	SE	0.3697	0.4346	0.4669	0.7576	0.7754
HC0	SE	0.2576	0.4496	0.3786	0.6869	0.71
HC1	SE	0.2972	0.4818	0.3944	0.6981	0.7171
HC2	SE	0.3321	0.4868	0.4044	0.7055	0.7207
HC3	SE	0.4602	0.529	0.4336	0.7262	0.7319
LBNN	Estimate	3.0199	2.9928	2.9814	2.9815	2.991
	SE	0.3919	0.2404	0.3044	0.3032	0.2319

Table 7: Simulated residuals standard error

	Sample size				
Methods	30	60	100	250	400
OLS	2.11768	3.8469	6.1716	14.8331	23.4968
LBNN	0.99270	0.8170	1.3053	1.5243	1.3765

DISCUSSION

At first we discuss results of the numerical examples. The plot as shown in Fig. 1 for the restaurant food sales data signifies that the OLS fit is inappropriate as there is a clear indication of heterogeneous error variances. It can be observed from Fig. 2 that the MPV residuals plot versus the fitted values indicates much improvement when compared to the OLS fit. We can see from Fig. 3 that, the LBNN transformation helps in producing a constant error variance. The summary statistics exemplified in this table show that the LBNN method does a superb job. The LBNN method outperforms the OLS and the MPV by possessing the lowest residual standard errors and lowest standard errors of the parameter estimates.

The OLS residual plot for the education expenditure data as displayed in Fig. 4 shows a funnel shape indicating heteroscedasticity. By looking at Fig. 5, we observe that all points are reasonably randomly distributed around zero, indicating that the CH has become successful in solving the problem of heteroscedasticity. For this particular example, the four geographical regions are used to define the near neighbor groups in order to compute the weights of the CH method. However, in practice we seldom get this kind of information. We employ the LBNN method for this data as if we did not know these four geographical regions. Figure 6 shows that the LBNN corrections work very well and we get a picture very similar to that produced by the CH method. We observe from Table 2 that the CH and the LBNN produce standard errors of estimates which are

much less than those obtained by the OLS and the HCCM. Even without any prior information about the heteroscedastic structure, the proposed LBNN method performs very well for this data.

Finally we discuss the results obtained from the simulations. Several interesting points emerge from these results as reported in Table 3-7. We observe that the LBNN and the OLS give close estimates in terms of the true value of the parameters. These results do not report any biasness problem of the regression parameters and also suggest that these estimates get even closer to the true value as the sample sizes get larger. It confirms the standard theory assumption that the presence of heteroscedasticity retains the unbiasedness property of the OLS estimates. The main interest here is to investigate the standard errors of the regression coefficients in the presence of heteroscedasticity. The results presented in Table 3-6 clearly show that the LBNN possesses the least value of the standard errors in comparison with other estimators. The values of the standard errors of the LBNN estimator are consistently the smallest for all sample sizes. It is also interesting to note that the residual standard errors based on the LBNN as presented in Table 7 are substantially smaller than the OLS. The results are consistent for each sample sizes.

The present findings contradicted by White (1980), Long and Ervin (2000), Montgomery *et al.* (2001) and Chatterjee and Hadi (2006), since their standard error of the parameters are relatively larger than our proposed LBNN method. It is important to point out; the proposed LBNN method can be used when the form of heteroscedasticity is unknown. Thus, it can be concluded that the LBNN method provides the best estimates in the presence of heteroscedasticity.

CONCLUSION

The main focus of this study was to develop a reliable alternative approach for correcting the problem of heteroscedastic errors with unknown structures. We have considered several estimators in this regard. The OLS based HCCM is not efficient at all and the other versions are not any better either. The MPV is a good method but it can be applied only to simple linear regression model. The performance of the CH is good and it can be applied to more than one predictor variables but this technique requires prior information on the structures of heteroscedastic errors. In this study we proposed a leverage-based near neighbor method where we do not require any prior information on the structures of heteroscedastic errors and it can be easily applied to multiple regression models. The empirical studies and simulation experiments show that the LBNN method offers a substantial improvement over the other existing methods. The LBNN can significantly reduce standard errors of estimates and standard error of residuals for both simple and multiple regression models. Thus we can consider the LBNN as a better estimation method and strongly recommend using this method especially when heteroscedasticity occurs in the data.

REFERENCES

Carroll, R.J. and D. Ruppert, 1982. Robust estimation in Heteroscedastic Linear models. Annal. Stat., 10: 429-441.

Carroll, R.J. and D. Ruppert, 1988. Transformation and Weighting in Regression. 2nd Edn., Chapman and Hall, New York, ISBN: 0412014211..

Chatterjee, S. and A.S. Hadi, 2006. Regression Analysis by Example. 4th Edn., Wiley, New York.

Davidson, R. and J.G. MacKinnon, 1993. Estimation and Inference in Econometrics. 1st Edn., Oxford University Press, New York.

Emerson, J.D. and M.A. Stoto, 1983. Transforming Data. In: Under-standing Robust and Explanatory Data Analysis, Hoaglin, D.C., F. Mosteller and J.W. Tukey (Eds.). Wiley, New York.

Greene, W., 2008. Econometric Analysis. 6th Edn., Prentice Hall, New Jersey, ISBN: 0135132452.

Gujarati, D.N., 2003. Basic Econometrics. 4th Edn., McGraw-Hill, New York, ISBN: 0-07-233542-4, pp: 202-247.

Judge, G.G., R.C. Hill and W.E. Griffiths, 1988. Introduction to the Theory and Practice of Econometrics. 2nd Edn., John Wiley and Sons, Inc., USA.

Long, J.S. and L. Ervin, 2000. Using heteroscedasticity-consistent standard errors in the linear regression model. Amer. Statis., 54: 217-224.

MacKinnon, J.G. and H. White, 1985. Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties. J. Economet., 29: 305-325.

Montgomery, D., E. Peck and G.G. Vining, 2001. Introduction to Linear Regression Analysis. 3rd Edn., Jon Wiley and Sons, New York.

Robinson, P.M., 1987. Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form. Econometrica, 55: 875-891.

Struyf, A., M. Hubert and P.J. Rousseeuw, 1997. Integrating robust clustering techniques in S-PLUS. Computat. Statis. Data Analy., 26: 17-37.

Welsberg, S.D., 1980. Applied Linear Regression. John Wiley, New York, pp: 174-204.

White, H., 1980. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. Econometrica, 48: 817-838.