Kuwait J. Sci. **45 (3)** pp 29-38, 2018

A simulation-based evidence on the improved performance of a new modified leverage adjusted heteroskedastic consistent covariance matrix estimator in the linear regression model

Nuzhat Aftab[1], Sohail Chand[2,*]
[1,2]*College of Statistical and Actuarial Sciences, University of the Punjab, Lahore*
**Corresponding author: sohail.stat@gmail.com*

Abstract

In this paper, we present a new heteroskedastic consistent *(HC*) covariance matrix estimator which considers the effect of leverage observations and which has a better approximation of its true asymptotic distribution. We point out that the basic motivation behind this new modified *HC* estimator is to provide an estimator which does not require any user specified values. In terms of bias and mean squared error (MSE), a Monte Carlo simulation study provided evidence that this new estimator has better small sample properties over some existing estimators. A real-life example also evaluated the finite sample behavior in comparison to those existing estimators.

Keywords: HCs; heteroskedasticity; high Leverage points; linear regression; quasi-t test

1. Introduction

Linear regression has a central role in statistics, economics and other applied fields for the purpose of modeling because it allows for modeling the of cross-sectional data, (see for instance Al-Humoud and Al-Ghusain (2003). Under heteroskedasticity, the problem of inconsistency and inefficiency of the covariance matrix leads to biased inference which becomes more serious with an increase in the level of heteroskedasticity, see, e.g.,(Hausman & Palmer (2012).

To resolve the problem of biased inference, Eicker (1963) and White (1980) suggested the heteroskedastic consistent *(HC*) covariance matrix estimator. Although, the HC_0 (White, 1980) under certain scenario provides valid inference for infinitely large sample size, it but does not perform well for finite samples. MacKinnon and White (1985) suggested three modified versions of *HC* estimator, namely HC_1, HC_2 and HC_3, mainly for large samples and recommended their use for n ≥ 250. Several authors studied e.g., Cribari-Neto and Zarkos (1999, 2001); Cribari-Neto (2004), Cribari-Neto and Galv˜ao (2003), studied the finite sample performance of these *HC* estimators in terms of asymptotic distribution of quasi-t test statistic and concluded that HC_3 estimator is comparatively a better estimator see. (Cribari-Neto & Zarkos; 1999, 2001; Cribari-Neto, 2004; Cribari-Neto & Galvão, 2003) Furthermore, the Sstudies on these estimators include Davidson and MacKinnon (1993), Hodoshima and Ando (2006) and Cribari-Neto and Da Silva (2011). They evaluated the asymptotic approximation and relative bias in *HC* based tests. They considered the fixed and stochastic heteroskedastic linear regression models with the presence of leverage observations.

Cribari-Neto and Zarkos (2004) showed that the effect of high leverage observation is more critical for small samples and leads to imprecise inference. They suggested an HC_4 estimator to deal with this scenario. Later studies, see e.g., by Cribari-Neto *et al*. (2005), Cribari-Neto *et al*. (2007), Cribari-Neto and Da Silva (2011), for example, showed that the test using HC_4 has very poor approximation of its asymptotic distribution. Thus for this reason, they suggested a modified version, HC_{4m}, and showed its better approximation as compared to HC_{4m}. However, application of this the modified estimator requires the user defined values. In this article we suggest a new *HC* estimator which, unlike HC_{4m}, does not require any user specified values. Moreover, it leads to a better asymptotic approximation under homoskedasticity as well as under heteroskedasticity even for small samples.

The rest of the paper is organized as follows: Section 2 describes the model and covariance matrix estimators. Section 3 proposes a new estimator, . Numerical results of simulation study are given in Section 4. Real-life data are studied in Section 5. Finally, the Section 6 gives concluding remarks.

2. Materials and Methods

The linear regression model considered is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \grave{\mathbf{o}}, \tag{1}$$

where $\mathbf{X}_{n\times k}$ is the matrix of independent variables, n is the sample size, and k is the number of parameters. $Y_{n\times 1}$ is a vector of a dependent variable, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ is an $n \times 1$ vector of error term, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{k-1})$ is the vector of unknown parameters. Generally, we assume $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$ and $\varepsilon_i \sim N(0, \sigma^2)$. The first condition ensures uncorrelated errors while the second condition implies the homoskedastic errors. When the variance of $\boldsymbol{\varepsilon}$ does not remain constant, the situation $\varepsilon_i \sim N(0, \sigma_i^2), (0 < \sigma_i^2 < \infty)$, known as heteroskedasticity, occurs. Thus, $var(\boldsymbol{\varepsilon}) = \boldsymbol{\Omega}$ where,

$$\boldsymbol{\Omega}_{n\times n} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2), \tag{2}$$

The OLS estimator of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{3}$$

,such that, $E(\hat{\boldsymbol{\beta}}) == \boldsymbol{\beta}$ and $var(\hat{\boldsymbol{\beta}}) = \boldsymbol{\Psi} = \mathbf{P}\boldsymbol{\Omega}\mathbf{P}^T$ where,

$$\mathbf{P}_{k\times n} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T, \tag{4}$$

and $\boldsymbol{\Omega}$ is as defined in (2).

When the assumption of homoskedasticity is satisfied, the variance of $\boldsymbol{\beta}$ is simplified as $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$ and it is estimated as $\hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}$, where $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}\hat{\boldsymbol{\varepsilon}}^T / (n-k)$, and $\hat{\boldsymbol{\varepsilon}}$ is the vector of OLS residuals given as

$$\hat{\boldsymbol{\varepsilon}} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\mathbf{Y}, \tag{5}$$

where, $\mathbf{I}_n$ is the identity matrix of order n.

The HC_0 suggested by White (1980) to resolve the problem of estimation and inference in the presence of heteroskedasticity is given as

$$HC_0 = \mathbf{P}\hat{\Omega}\mathbf{P}^T, \tag{6}$$

and

$$\hat{\Omega} = diag\{\hat{\varepsilon}_1^2, \ldots, \hat{\varepsilon}_n^2\} \tag{7}$$

where $\hat{\varepsilon}_i^2$ is the i^th^ diagonal element of the matrix $\hat{\varepsilon}\hat{\varepsilon}^T$. Although HC_0 is consistent in both homoskedasticity and heteros-kedasticity, it is typically biased. Moreover, it tends to under-estimate the true variance of $\boldsymbol{\beta}$ in the case of small samples with leverage observations (see Long & Ervin 2000; Cribari-Neto & Zarkos 2001); Cribari-Neto and Zarkos; 2004). Some alternatives of HC_0 are proposed in literature, among which HC_{4m} and HC_5 to some extent may incorporate the effect of leverage observations.

The HC_5 estimator, suggested by Cribari-Neto *et al*. (2007), is given as

$$HC_5 = \mathbf{PE}_5\hat{\Omega}\mathbf{P}^T, \tag{8}$$

where $\mathbf{E}_5 = diag\{1/\sqrt{(1-h_{ii})^{\delta_i}}\}, h_{ii}$ is the i^th^ diagonal element of the projection matrix $H = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. For, HC_5 $\delta_i = min\{(nh_{ii})/k, max\{4, (nch_{max})/k\}\}$, where $h_{max} = max\{h_{11}, \ldots, h_{nn}\}$, and c is some fixed value in [0 1] interval (see Cribari-Neto *et al*. 2007).

In a following paper, Cribari-Neto and Da Silva (2011) suggested a modified version denoted by HC_{4m}, which was proven to have good asymptotic approximation.

The modified estimator HC_{4m} is given as

$$HC_{4m} = \mathbf{PE}_{4m}\hat{\Omega}\mathbf{P}^T, \tag{9}$$

where,

$$\mathbf{E}_{4m} = diag\{1/(1-h_{ii})^{\delta_i}\}$$

and

$$\delta_i = min\{\gamma_1, (nh_{ii})/k\} + min\{\gamma_2, (nh_{ii})/k\};$$

$i = 1, 2, \ldots, n$. The values of γ_1 and γ_2 are chosen by a user in a fashion to reduce the effect of high leverage on the estimation of the covariance matrix. Cribari-Neto and Da Silva (2011) suggested $\gamma_1 = 1$ and $\gamma_2 = 1.5$. The same values are used for this investigation.

Both HC_{4m} and HC_5 estimators depend heavily on the choices of user specified values. In the next section, we propose a new HC estimator which does not require any user specified information and which can outperform other estimators, especially in the striking case of small samples with high leverage observations.

3. A New HC estimator

The problems with the suggested HC estimators discussed in the previous section are the poor asymptotic approximation of HC_4 and the search of appropriate values of γ_1 and γ_2 in HC_{4m} and c in HC_5. To address these theoretical and practical issues, we propose a new HC estimator, denoted by HC_6, given as

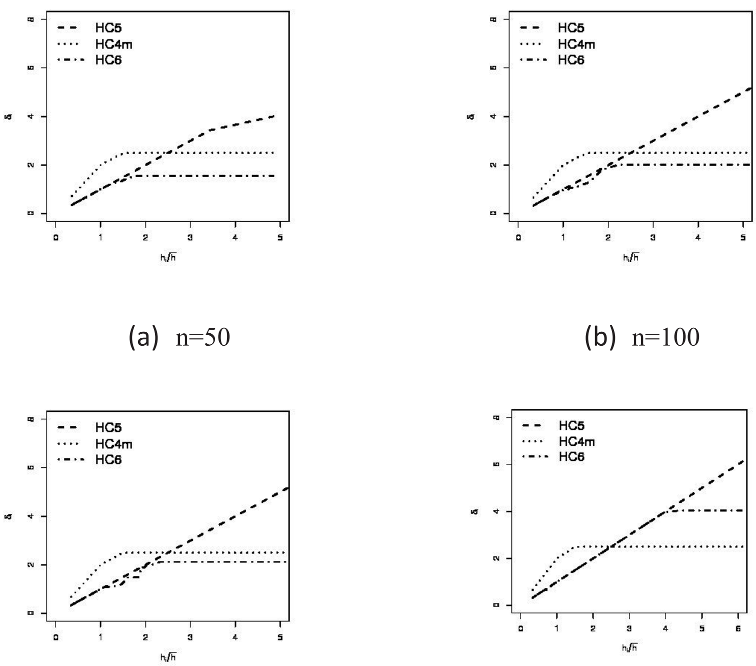
$$HC_6 = \mathbf{PE}_6\hat{\boldsymbol{\Omega}}\mathbf{P}^T, \tag{10}$$

where,

$$\mathbf{E}_6 = diag\{1/(1-h_{ii})^{\delta_i}\}$$

and,

$\delta_i = min\{\frac{nh_{ii}}{k}, \sqrt{\frac{nh_{max}}{2k}}\}$ $i = 1, \ldots, n$. Note, k is the number of parameters, h_{ii} is the leverage measure, and $h_{max} = max\{h_{ii}, i = 1, 2, \ldots, n\}$ is the maximum leverage value among h_{ii}. In equation (10), $\mathbf{P}$ and $\hat{\boldsymbol{\Omega}}$ are defined by (4) and (7) respectively. We define δ_i as a function of two values among which one is, $nh_{ii}/k = nh_{ii}/\sum_{i=1}^{n} h_{ii} = h_{ii}/\bar{h}$. The second value is the square root of the ratio of maximal leverage h_{max} and $2k/n$, where $2k/n$ is known as the bound above which an observation is considered as a



(a) n=50 (b) n=100

(c) n=150 (d) n=200

Fig 1. δ_i plotted against the ratio between individual h_{ii} leverages and the mean leverage $\overline{h}$

leverage observation (see Montgomery *et al*. 2001, p. 207). The quantity nh_{ii} / k is generally used in the HC estimators. Although this quantity provides a valid discount rate for low leverage observations, it puts a very heavy discount for the high leverage observations. The second quantity in δ_i puts a cap on the discount rate to prevent an unreasonable heavy discount.

To study the behavior of the discount rate, we have plotted δ_i of three considered HC covariance estimators which are HC_{4m} HC_5 and HC_6 against $h_{ii} / \overline{h}$.

It can be seen from Figure 1, that the δ_i of HC_5 results in a low discount rate for low-leverage observation, but there is an unreasonably high discount for high leverage observation. In contrast, δ_i of HC_{4m} puts an unreasonable heavy discount on low-leverage observations. Our suggested δ_i possesses the good properties of both competing estimators, as shown in Figure 1. Moreover, in the new δ_i the maximum discounting is not as intense as in other estimators. Unlike HC_{4m} and HC_5, our suggested estimator does not require any user specified information.

In the next section, we will evaluate the asymptotic approximation of our suggested estimator. We will also compare its performance with other estimators considered in this study.

4. Results and Discussion

In this section, we use extensive simulations to compare the relative probability discrepancy (RPD) in the quasi t-test. We also compare the bias and MSE of the HC estimators using bootstrap procedures. R code for implementing the simulations is available from the authors upon request.

Example 1: *Evaluation of relative probability discrepancy for homoskedastic and heteroskedastic cases.*

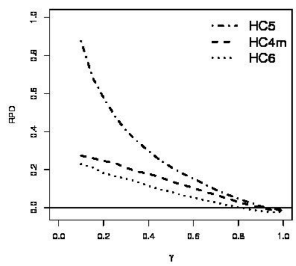
In this example, we compute the relative probability discrepancy (RPD) of quasi t-test based on our new suggested estimator HC_6 and compare it with HC_5 and HC_{4m}. We use the same study design as considered by Cribari-Neto and Da Silva (2011). The heteroskedastic linear regression model given in (1) can be written as

$$Y_i = \beta_0 + \sum_{j=1}^{k-1} \beta_j \mathbf{X}_{ij} + \epsilon_i, \; i = 1, 2, \ldots, n, \qquad (11)$$

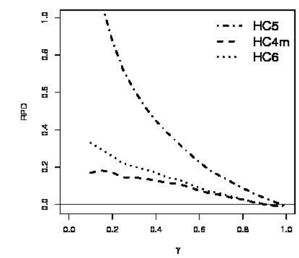
where $\epsilon_i \sim N(0, \sigma_i^2)$ The error variance σ_i^2 is defined as,

$$\sigma_i^2 = \exp(\sum_{j=1}^{k-1} \alpha_j X_{ij}) \text{ for } i = 1, 2, \ldots, n \text{ and } j = 1, 2, \ldots, k,$$

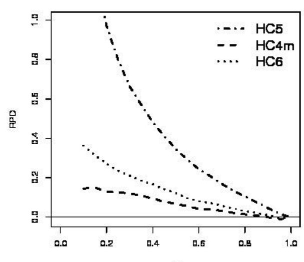
where k is the number of parameters in the model and α_j being the real scalar used to control the level of heteroskedasticity in the data. The covariates are generated from the standard lognormal distribution. The level of heteroskedasticity can be measured using $\lambda = max(\sigma_i^2) / min(\sigma_i^2)$. $\lambda = 1$ corresponds to homoskedasticity while $\lambda > 1$ implies the presence of heteroskedasticity in the data. The greater the value of λ, the more severe the level of heteroskedasticity becomes. We studied the model (11) for k=3 and *k=5*, with normal and non-normal errors under both homoskedasticity and heteroskedasticity. We set $\alpha = 0$ to obtain $\lambda = 1$, i.e., homoskedasticity, and $\alpha = 0.206$ to obtain λ approximately equal to 100. Several choices of the sample size have been studied (n = 50, 100, 150, 200).



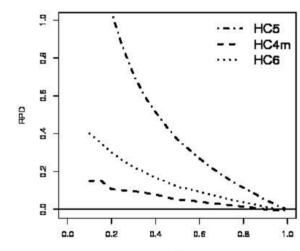
(a) n=50



(b) n=100

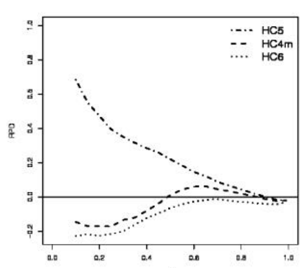


(c) n=150

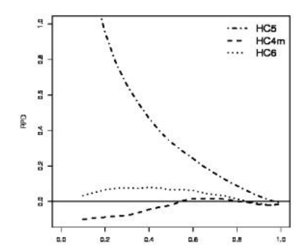


(d) n=200

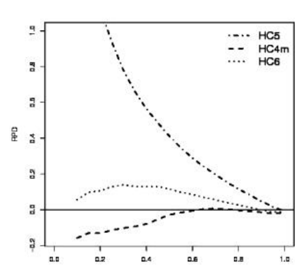
Fig 2. (*Homoskedasticity*)RPD vs asymptotic probabilities(γ); With k = 3 and α = 0



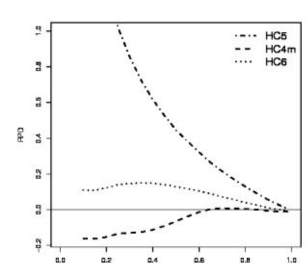
(a) n=50



(b) n=100



(c) n=150



(d) n=200

Fig 3. (*Heteroskedasticity*) RPD vs asymptotic probabilities(γ); With k = 3 and α = 0.206

In the present study, the interest lies in testing the hypothesis, $H_0 : \beta_j = 0 \quad j = 1,2,\ldots,k,$ against the two-sided alternative hypothesis, $H_a : \beta_j \neq 0$. The square of quasi t-test used is

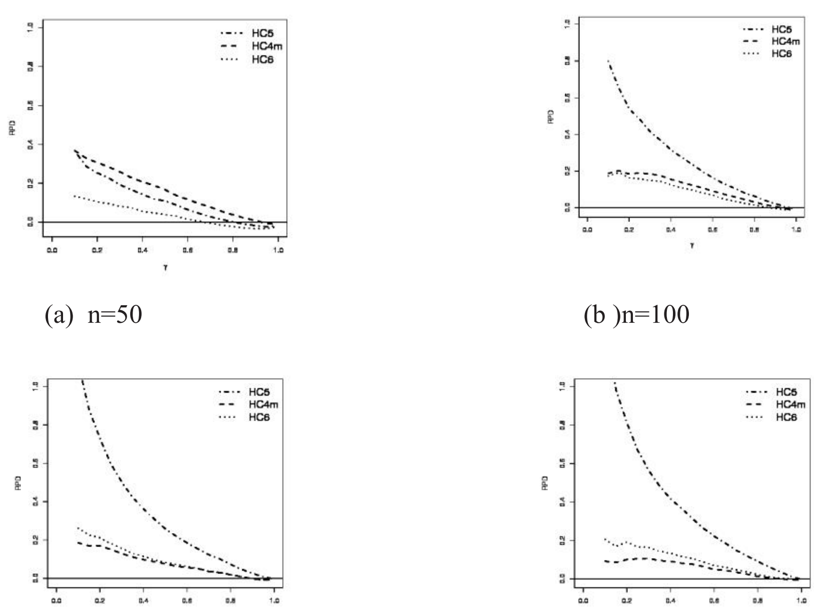
$$\tau^2 = \frac{\hat{\beta}_j^2}{var(\hat{\beta}_j)} \sim \chi_1^2, \tag{12}$$

where $\hat{\beta}_j$ denotes the OLS estimate of β_j and $var(\hat{\beta}_j)$ is based on HC_5, HC_{4m} and HC_6 estimators. Without the loss of generality, we can consider testing $H_0 : \beta_1 = 0$. Simulation results are based on 10,000 Monte Carlo runs. All the simulations are performed using the R language (R Development Core Team 2011). To assess the approximation of asymptotic distribution of the t-test, we used the measure relative probability discrepancy (RPD) defined as

$$RPD = \frac{\#(\tau^2 < \chi^2_{1,\gamma})/N - \gamma}{\gamma}, \tag{13}$$

where γ is the cumulative probability of asymptotic distribution, N is the number of Monte Carlo runs and #(condition) is the number of cases satisfying the given condition. The results of RPD are shown in Figures 2 and 3.

Case-I Homoskedasticity: In the case of homoskedastic errors (Figure 2) with $k = 3$, the asymptotic distribution approximation of quasi t-test based on HC_5 is very poor for all the considered choices of sample size. But in fairness, the approximation is better for the tests associated with HC_6 and HC_{4m} as compared to HC_5. For the small sample, $n = 50$, the approximation is better for HC_6-based test. However, the situation reverses for larger sample size choices, i.e n>100. Interestingly, the approximation of asymptotic distribution in the right tail for both HC_6-based and HC_{4m}-based tests is very close. This is the region that plays a critical role in hypothesis testing. For the homoskedastic case, the results for $k = 5$ are similar to $k = 3$ (Figure 4).



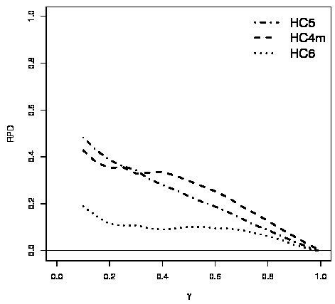
(a) n=50

(b)n=100

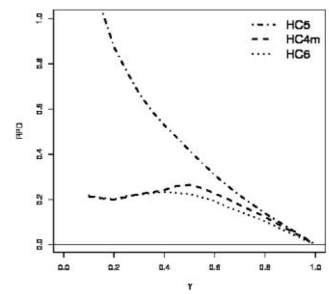
(c) n=150

(d) n=200

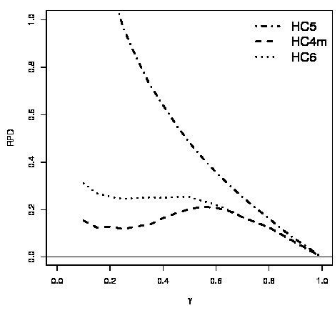
Fig 4. (*Homoskedasticity*) RPD vs asymptotic probabilities(γ); With k = 5 and α = 0



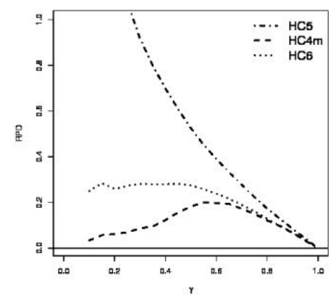
(a) n=50



(b) n=100



(c) n=150



(d) n=200

Fig 5. (*Heteroskedasticity*) RPD vs asymptotic probabilities(γ); With k = 5 and α = 0.206

Case-II Heteroskedasticity: Recall that the main objective of HC estimators is to provide valid estimates for the variance of regression parameters in the presence of heteroskedasticity. Figure 3 shows the results of RPD for $k=3$. Our suggested estimator HC_6 outperforms the other considered estimators. Although, for large choices of sample size, HC_6- based test shows positive RPD while for HC_{4m}- based test the RPD is generally negative but the approximation of asymptotic distribution for HC_6- based test is generally better than HC_{4m} especially in the right tail. When $n=50$ and $k=5$ the approximation of HC_6 based test is superior to other tests. While for $n>50$, the HC_6 and HC_{4m} based test shows same approximation, especially in the right tail of the asymptotic distribution. Thus, empirical size for the tests that employ HC_6 and HC_{4m} estimators have a close approximation for $\gamma > 0.8$.

The results of the RPD show that the test based on our suggested HC_6 estimator can provide a valid inference about the regression parameter in the case of heteroskedasticity. Now we evaluate the amount of bias and the MSE in the HC estimators under study. The wild bootstrap procedure used is given as below:

1. Generate the data $\mathbf{X}, \mathbf{Y}$ under model (1).

2. Fit model (1) and obtain the OLS residuals $\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \mathbf{X}\boldsymbol{\beta}$, where $\hat{\beta}$ is the least square estimate of $\boldsymbol{\beta}$.

3. Resample the residuals $\hat{\boldsymbol{\epsilon}}$ and obtain the wild bootstrap $\boldsymbol{\epsilon}^*$ using Liu (1988) transformation.

4. Bootstrap Y, say $\mathbf{Y}^*$ using the wild bootstrap residuals $\boldsymbol{\epsilon}^*$ and model (1) i.e., $\mathbf{Y}^* = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$ and obtain $\hat{\boldsymbol{\epsilon}}^* = \mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^*$ is the OLS estimator for the bootstrap sample $(\mathbf{X}, \mathbf{Y}^*)$.

5. Compute the matrix defined in equation (7) using $\hat{\boldsymbol{\epsilon}}^*$ and thus estimate HC^*, i.e. the HC estimator for the bootstrap sample.

6. Repeat steps 3 through 5, B times. Thus we obtain $(\boldsymbol{\beta}^{*(1)}, \ldots, \boldsymbol{\beta}^{*(B)})$ and compute the variance, denoted as $\boldsymbol{\Psi}^* = diag(\psi_j^* : j = 1, \ldots, k)$, which provides a true value of $var(\boldsymbol{\beta})$. In addition, compute $(HC^{*(1)}, \ldots, HC^{*(B)})$ and thus its expected value i.e., $E(HC^*)$ and the $var(HC^*)$. Using the results obtained in Step 6, we can compute the bias and MSE given as follows:

$$Bias_{\psi_j^*}(\hat{\psi}_j) = E(\hat{\psi}_j) - \psi_j^* \qquad (15)$$

$$MSE_{\psi_j^*}(\hat{\psi}_j) = var(\hat{\psi}_j) + (Bias_{\psi_j^*}(\hat{\psi}_j))^2$$

where $E(\hat{\psi}_j)$ is the jth element of $E(HC^*)$ and $var(\hat{\psi}_j)$ is the j[th] diagonal elements of $var(HC^*)$.

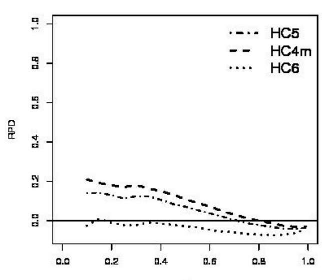
For bias and MSE, the linear regression model given in Equation (1) is used for $k=3$ and $k=5$. We have considered small samples only, since with large samples, the estimators show approximately similar results (see MacKinnon and White 1985 and Long and Ervin 2000). The predictors are simulated from standard lognormal distribution, while the level of heteroskedasticity is set at $\lambda = 100$.

The results for the average MSE and bias for $k = 3$ and $k = 5$ are given in Table 1 and Table 2, respectively. The figures in boldface are the smallest among the three considered estimators. The results present a clear dominance of efficiency in terms of minimum bias and MSE of HC_6 over HC_5 and HC_{4m} with few exceptions, as when HC_5 is efficient.

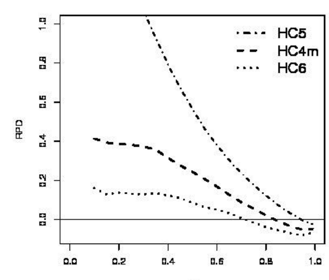
Now we will illustrate the use of our suggested estimator by applying it to real-life data. We will show that it is important to take the presence of heteroskedasticity into account and use the HC estimator-based quasi t-test when testing the significance of regression parameters.

5. Real data example

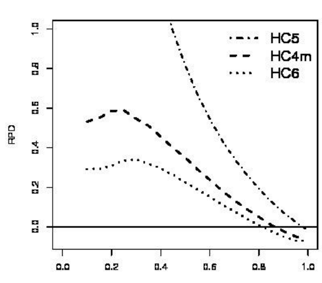
In this section, we have applied HC estimators to the education expenditure data taken from Chatterjee and Hadi (2015) as an illustration.



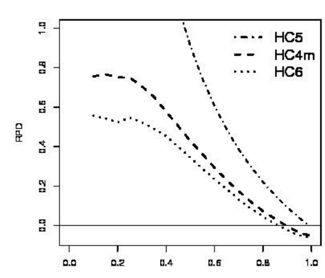
(a) $\phi = 2$



(b) $\phi = 3$



(c) $\phi = 4$



(d) $\phi = 5$

Fig 6. RPD vs asymptotic probabilities(γ); $X_{max} = \phi$ for n = 25

The relationship between expenditures on education and per capita income for 50 different regions has been studied. Figure 7 shows the scatter plot of data and the plot of residuals against the observed values. The scatter plots how a linear relation between two variables. Yet, due to the presence of one extreme observation, it tends to have a quadratic model.

We first test the data for possible heteroskedasticity. The Breusch-Pagan test (LM = 9.1399, df =2, p.value = 0.010) is significant at a 5% level of significance, thus indicating the presence of heteroskedasticity. The same findings can be confirmed from the residuals plot shown in Figure 7.

Now, we fit the following regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i, \quad i = 1, 2, \ldots, 50 \qquad (16)$$

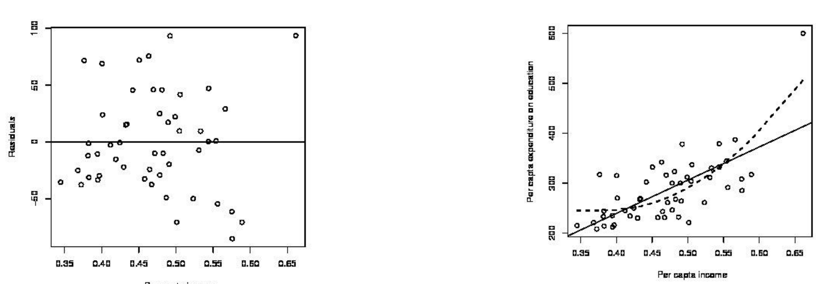
The values of the OLS estimates are $\hat{\beta}_0 = 748.3$, $\hat{\beta}_1 = -2618.4$ and $\hat{\beta}_2 = 3406.3$. The results given in Table 3 show that the OLS is influenced by the extreme value which suggests a quadratic model. On the other hand, the HC_6 estimator, along with the other HC estimators, clearly suggests a linear model for the education expenditure regressed on per capita income.

Table 1. Average Bias and MSE of HCs under heteroskedasticity using 500 Monte Carlo runs and 500 Bootstrap samples

		$n=25$		$n=50$		$n=100$	
	Estimator	β_1	β_2	β_1	β_2	β_1	β_2
MSE	HC_5	0.2778	0.0064	0.0091	**0.0025**	0.0048	0.0066
	HC_{4m}	0.5782	0.0144	0.0067	0.0037	0.0018	0.003
	HC_6	**0.1402**	**0.0047**	**0.0034**	0.0028	**0.0016**	**0.0027**
Bias	HC_5	0.1135	0.0121	0.0323	**0.0042**	0.0158	0.0178
	HC_{4m}	0.2415	0.0344	0.0268	0.0129	0.007	0.0086
	HC_6	**0.0437**	**0.0075**	**0.0145**	0.0072	**0.006**	**0.0073**

Table 2. Average bias and MSE of HCs under heteroskedasticity using 500 Monte Carlo runs and 500 Bootstrap samples (k = 5)

		$n=25$				$n=50$				$n=100$			
	Estimators	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4	β_1	β_2	β_3	β_4
Bias	HC_5	0.067	0.174	0.192	**0.097**	0.022	0.179	0.030	0.179	**0.009**	**0.015**	**0.006**	**0.010**
	HC_{4m}	0.195	0.492	0.516	0.356	0.055	0.148	0.064	0.148	0.010	0.013	0.006	0.010
	HC_6	**0.066**	**0.154**	**0.095**	0.124	0.026	0.062	0.038	0.060	0.007	0.008	0.003	0.007
MSE	HC_5	0.011	0.058	0.089	**0.034**	**0.018**	0.162	**0.024**	0.175	0.001	0.002	0.000	0.001
	HC_{4m}	0.074	0.414	0.569	0.233	0.029	0.116	0.047	0.118	0.001	0.002	0.000	0.001
	HC_6	**0.010**	**0.046**	**0.025**	0.045	0.020	**0.031**	0.028	**0.025**	**0.000**	**0.001**	**0.000**	**0.001**



(a) (b)

Fig7. Plots of education expenditure data for 1970 (a) X vs Y. (b) Residuals vs X.

6. Conclusions

In this article, a new HC covariance estimator HC_6 is suggested, which has a minimum bias and MSE. Furthrmore, our suggested estimator, unlike other competing estimators, i.e. HC_{4m} and HC_5, does not require any user specified information. The use of a quasi t-test based on the HC_6 estimator is recommended for the inference of linear regression model parameters when heteroskedasticity is present but no collinearity among predictors exists.

Table 3. Testing of $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ for Education Expenditure Data

Testing of $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$			
Test	S.E	t	p-value
OLS	1064.7	3.199	0.002
HC_5	7864.3	0.433	0.667
HC_{4m}	4845.0	0.703	0.485
HC_6	4323.7	0.787	0.434

Our suggested estimator outperforms the competing estimators, especially when there are small samples with high leverage observations.

For future research, we will develop estimators which can provide valid inferences for heteroskedastic regression models with collinearity among the predictors.

Acknowledgements

We thank the Editor and the referees for their helpful comments and suggestions that greatly improved the previous version of this paper. This work was supported by the Higher Education Commission of Pakistan under the Indigenous PhD Fellowships Phase-II Program.

References

Al-Humoud, J. & Al-Ghusain, I. (2003). Household demand for water: A case study in kuwait. *Kuwait Journal of Science and Engineering*, **30**(1): 197–212.

Chatterjee, S. & Hadi, A.S. (2015). Regression analysis by example. John Wiley & Sons, NJ, USA. pp 144 .

Cribari-Neto, F. (2004). Asymptotic inference under heteroskedasticity of unknown form. Computational Statistics & Data Analysis, **45**(2): 215–233.

Cribari-Neto, F. & Da Silva, W. (2011). A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. AStA Advances in Statistical Analysis, **95**(2): 129–146.

Cribari-Neto, F., Ferrari, S., & Oliveira, W.

(2005). Numerical evaluation of tests based on

different heteroskedasticity-consistent covariance matrix estimators. Journal of Statistical Computation and Simulation, **75**(8): 611–628.

Cribari-Neto, F. & Galvão, N. (2003). A class of improved heteroskedasticity-consistent covariance matrix estimators. Communications in Statistics-Theory and Methods, **32**(10): 1951–1980.

Cribari-Neto, F., Souza, T., & Vasconcellos, K. (2007). Inference under heteroskedasticity and leveraged data. Communications in Statistics-Theory and Methods, **36**(10): 1877–1888.

Cribari-Neto, F. & Zarkos, S. (1999). Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing. Econometric Reviews, **18**(2): 211–228.

Cribari-Neto, F. and Zarkos, S. (2001). Heteroskedasticity consistent covariance matrix estimation: White's estimator and the bootstrap. Journal of Statistical Computation and Simulation, **68**(4): 391–411.

Cribari-Neto, F. & Zarkos, S. G. (2004). Leverage-adjusted heteroskedastic bootstrap methods. Journal of Statistical Computation and Simulation, **74**(3): 215–232.

Davidson, R. & MacKinnon, J. (1993). Estimation and inference in econometrics. Oxford University Press, USA. pp 631.

Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions. The Annals of Mathematical Statistics, **34**(2): 447–456.

Hausman, J. & Palmer, C. (2012). Heteroskedasticity-robust inference in finite samples. Economics Letters, **116**(2): 232–235.

Hodoshima, J. & Ando, M. (2006). The effect of non-independence of explanatory variables and error term and heteroskedasticity in stochastic regression models. Communications in Statistics-Simulation and Computation, **35**(2): 361–405.

Liu, R. (1988). Bootstrap procedures under some non-iid models. The Annals of Statistics,

16(4): 1696–1708.

Long, J. & Ervin, L. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. The American Statistician, **54**(3): 217–224.

MacKinnon, J. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. Journal of Econometrics, **29**(3): 305–325.

Montgomery, D., Peck, E., Vining, G., & Vining, J. (2001). Introduction to linear regression analysis, Volume 3. Wiley: NJ, USA. p207.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. RFoundation for Statistical Computing, Vienna, Austria.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. Econometrica. Journal of the Econometric Society, **48**(4): 817–838.

Submitted: 13/12/2015
Revised : 06/11/2016
Accepted : 29/04/2017

تقدير جديد متسق وغير متجانس لمصفوفة التغاير في الانحدار الخطي يأخذ في الاعتبار المشاهدات المؤثرة

[1] نزهة افتاب، [2,*] سهيل شاند
[1,2,*] كلية العلوم الإحصائية والاكتوارية، جامعة البنجاب، لاهور
*sohail.stat@gmail.com

الملخص

في هذا البحث نقدم تقدير جديد غير متجانس لمصفوفة التغاير (HC) والذي يأخذ في الاعتبار المشاهدات المؤثرة والذي لديه تقريب أفضل لتوزيعه التقاربي. ونشير إلى أن الدافع الرئيسي وراء هذا التقدير هو الحصول على تقدير لا يتطلب أي قيم محددة من قبل المستخدم. قدمت دراسة مونت كارلو للمحاكاة دليلاً على أن هذا التقدير الجديد لديه خصائص أفضل من بعض التقديرات الموجودة في العينات الصغيرة من حيث مقدار التحيز ومقدار متوسط مربعات الأخطاء. وتم تقييم التقدير الجديد بالمقارنة مع التقديرات الموجودة باستخدام مثال من الحياة الواقعية.