Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2014) 22: 412 – 424
© TÜBİTAK
doi:10.3906/elk-1207-23

How do HCCMEs perform in small samples?

Mehmet ORHAN[1,*], **Esra ŞİMŞEK**[2]
[1]Department of Economics, Fatih University, İstanbul, Turkey
[2]Department of International Trade and Finance, Süleyman Şah University, İstanbul, Turkey

Received: 09.07.2012 • **Accepted:** 17.09.2012 • **Published Online:** 17.01.2014 • **Printed:** 14.02.2014

Abstract: The purpose of this paper is to explore the performances of prominent and popular heteroscedasticity-consistent covariance matrix estimators (HCCMEs) in small samples. The HCCMEs are selected from the literature and their ability to estimate the true covariance matrix of the coefficient's vector is evaluated through simulation runs. We calculate the percentage difference between the expected value of the HCCME and the true covariance matrix to set a convenient stage to make the comparisons under several different regression settings. The main contribution of the paper is the inclusion of the HCCMEs that have been introduced into the literature recently. We report the performances of the HCCMEs under different settings of the covariates and error term variances. We let the covariates follow uniform, normal, Student's t, and Cauchy distributions and tailor the error term variances to increase gradually.

Key words: Heteroscedasticity-consistent covariance matrix estimator, minimum covariance determinant, Monte Carlo simulation

1. Introduction

Regression analyses are very extensively used in disciplines ranging from chemistry to biology and electronics to computer engineering, including applications of multilayer perceptrons and wavelet networks [1]. The use of regression in such disciplines is growing rapidly due to the main functionality of regression analysis in working out the nature of the relation between dependent and independent variables, especially when the data available are insufficient and the relation among the variables is ambiguous. To cite a few, Jeng et al. [1], Moghram and Rahman [2], Cordón et al. [3], Kerr et al. [4], Maffeis et al. [5], Peeters et al. [6], Faraway [7], Breese and Hill [8], and Heinemann et al. [9] have made use of regression analyses to work out the unknown relation between variables. One of the most widely used estimation methods in regression analysis is ordinary least squares (OLS). Following the introduction of [2], in a stylized regression setting, we present the control variable as a function of the exogenous variables as:

$$y\left(t\right) = a_0 + a_1 x_1\left(t\right) + \ldots + a_n x_n\left(t\right) + a(t),$$

where y(t) is a dependent variable, such as the electrical load; $x_1\left(t\right), \ldots, x_n\left(t\right)$ are the explanatory variables correlated with y(t); a(t) is a random variable with zero mean and constant variance; and $a_0, a_1, \ldots, a_n$ are the regression coefficients.

One standard assumption of the OLS, as stated in the Gauss–Markov theorem, is the equality of the error term variances known as homoscedasticity. However, many data sets in real life prove to have different

*Correspondence: morhan@fatih.edu.tr

variances defined as heteroscedasticity. The covariance matrix estimators of the OLS coefficient estimates under heteroscedasticity are biased in small samples, and therefore inference based on heteroscedastic error terms is of limited use [10]. The bias indicates the difference between the expected value of the estimator and the true value of the parameter. There are several attempts to find estimators of this covariance matrix with lower biases in small samples. These are called the heteroscedasticity-consistent covariance matrix estimators (HCCMEs). The most prominent and widely used HCCME is that of White [11]. In that study, White used the earlier work of Eicker [12,13] and introduced the consistent estimator of this covariance matrix, where he had promoted the test for heteroscedasticity, as well. This was a cornerstone in the study of heteroscedastic regression and has since been cited many times.

There were several attempts to alleviate the bias, which is a function of the leverage. An observation (x_i, y_i) is said to be a leverage point when it lies outside of the majority of the regressors. The term 'leverage' comes from the mechanics, because such a point pulls the regression line towards itself. Horn et al. [14], Hinkley [15], and Efron [16] weighted the OLS squared residuals by different factors. One competitive estimator came from Efron's jackknifing [16]. Indeed, the anonymous referee to MacKinnon and White [17] made the formula of the so-called one-delete-jackknife estimator available, which enabled later studies to make use of this formula. There were many other efforts to find better HCCMEs based on different criteria. Orhan and Zaman [18] included more than 10 such HCCMEs, including the bootstrap estimator and minimum norm quadratic unbiased estimator (MINQUE), in a comparison study based on Monte Carlo simulation. More recent work on this issue resulted in the estimators of Cribari-Neto [19] and Cribari-Neto et al. [20]. The problem with the bias and variance calculation for the HCCMEs is that it is very sophisticated and comparison studies so far rely on computer simulations. However, the Monte Carlo simulation assumes some certain setups of the design matrix and the variance pattern of the error terms and can thus be questioned since other setting designs may lead to divergent conclusions. In similar studies, Andrews [21] used the heteroscedasticity and autocorrelation consistent estimation of covariance matrices, Mackinnon [22] compared bootstrap methods, Bera et al. [23] reconsidered the MINQUE, and Hayes and Cai [24] reported Type I error rate simulation results.

This paper has the main contribution of comparing the HCCMEs, including the most recent ones, under different settings. This is a very critical and important contribution, since many econometric applications make use of the entries belonging to HCCMEs. Moreover, many statistical analyses make use of the variances and standard errors of the coefficient estimates. Our conclusions will hopefully have vast and wide applications ranging from significance tests of partial regression coefficients to their confidence intervals.

Section 1 of this article is a brief introduction. Section 2 explains the model and HCCMEs. Section 3 explains the settings used for comparisons and the performances of the HCCMEs. Finally, Section 4 concludes.

2. Heteroscedasticity consistent covariance matrix estimators

We assume $y = X\beta + \varepsilon$ of a simple regression, where y is the $\mathrm{T} \times 1$ vector of the response variable, X is the matrix of the regressors with the first column allocated to the intercept, and ε is the $\mathrm{T} \times 1$ vector of the disturbance terms. Here, T is the number of observations. The disturbance terms are assumed to have flexible variances to allow heteroscedasticity, i.e. $\varepsilon \sim N(0, \Sigma)$ where $\Sigma = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_T^2)$. Note that the disturbance terms are pair-wise uncorrelated.

The OLS estimator of the coefficients' vector is $\hat{\beta} = (X'X)^{-1}X'y$. The main point of interest is the covariance matrix of $\hat{\beta}$, $\Omega = Cov(\hat{\beta}) = Cov[(X'X)^{-1}X'y] = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$. The most critical term in $Cov(\hat{\beta})$ is Σ, since X and all of its products are known.

Indeed, there were several attempts to estimate this term and these attempts led to HCCMEs named HC0, HC1, etc., as we define the most prominent and popular ones below with their references. Indeed, there were some other HCCMEs suggested in the literature, just like the MINQUE of Bera et al. [23] and the bootstrap estimator explained by Wu [25], but they are not adopted here due their failure and low estimation performance. Earlier studies in the literature help us choose the 6 estimators as follows:

$\mathrm{HC0} = diag\left(e_1^2, e_2^2, \ldots, e_T^2\right)$, White [11]

$\mathrm{HC1} = \frac{T}{T-k} diag(e_1^2, e_2^3, \ldots, e_t^2)$, Hinkley [15]

$\mathrm{HC2} = diag\left(\frac{e_1^2}{1-H_{11}}, \frac{e_2^2}{1-H_{22}}, \ldots, \frac{e_T^2}{1-H_{TT}}\right)$, Horn et al. [14]

$\mathrm{HC3} = diag\left(\left(\frac{e_1}{1-H_{11}}\right)^2, \left(\frac{e_2}{1-H_{22}}\right)^2, \ldots, \left(\frac{e_T}{1-H_{TT}}\right)^2\right)$, Efron [16]

$\mathrm{HC4} = diag\left(\frac{e_1^2}{(1-H_{11})^{\delta_1}}, \frac{e_2^2}{(1-H_{22})^{\delta_2}}, \ldots, \frac{e_T^2}{(1-H_{TT})^{\delta_T}}\right)$, Cribari-Neto [19]

$\mathrm{HC5} = diag\left(\frac{e_1^2}{(1-H_{11})^{\alpha_1}}, \frac{e_2^2}{(1-H_{22})^{\alpha_2}}, \ldots, \frac{e_T^2}{(1-H_{TT})^{\alpha_T}}\right)$, Cribari-Neto et al. [20]

In these formulations, e_i^2 are the squared OLS residuals; T is the number of observations; k is the number of independent variables, including the intercept term; and H is the hat matrix, $H = X(X'X)^{-1}X$, $\delta_i = \min(4, \frac{H_{ii}}{\bar{H}})$, where $\bar{H} = \frac{1}{T}\sum_i H_{ii}$ in HC4 and $\alpha_i = \min\left(\frac{H_{ii}}{\bar{H}}, \max(4, \frac{kH_{\max}}{\bar{H}})\right)$ in HC5. HC3 is an approximation to the jackknife estimator. The original estimator by MacKinnon and White [17] is slightly changed to get HC3.

3. Simulation runs

We run simulations under different settings of regressors and the variances of disturbance terms to work out the performances of the HCCMEs. We reuse the settings inspired by earlier studies in the literature that scan over different levels of error term variances and leverages of covariates. Namely, the covariates follow uniform (Case 1), normal (Case 2), Student t with 3 degrees of freedom (Case 3), and Cauchy (Case 4) distributions. See the works of Cragg [26], Flachaire [27], and Lima et al. [28] for such designs. The Monte Carlo simulation is based on a simple regression model of:

$$y_i = \beta_0 + \beta_1 x_{xi} + \sigma_i e_i, \quad e_i \sim N(0,1).$$

The regression parameters are set at $\beta_0 = \beta_1 = 1$ and we make use of the degree of heteroscedasticity defined as $\lambda = \max(\sigma_i^2)/\min(\sigma_i^2)$, which returns 1 under homoscedasticity; λ becomes larger when heteroscedasticity is more intensive. The simulation program is coded in GAUSS 7 and we set the Monte Carlo sample size to 10,000 replications. Regarding the error term variances, we use the same formulation and play with the parameters to adjust the level of heteroscedasticity. More specifically, we set $\sigma_i^2 = c_0 + c_1 x_i + c_2 x_i^2$ to manage the variance of the ith observation's error term. In Case A '$c_0 = 1, c_1 = 0, c_2 = 0$' and in Case B '$c_0 = 1, c_1 = 0.5, c_2 = 0.5$', whereas in Case C '$c_0 = 1, c_1 = 2, c_2 = 2$' and in Case D '$c_0 = 1, c_1 = 4, c_2 = 4$'. Needless to say, Case A formulates homoscedasticity, where the variances of the error terms are all the same and equal to 1. To the contrary, Case D is the setup that allows higher degrees of heteroscedasticity, especially when the covariates have higher leverage observations. The degree of heteroscedasticity increases from Case B to Case D since the coefficients multiplying the covariates are getting larger.

The program we coded in GAUSS first generates the covariates and then starts fixing the error terms from predetermined distributions in each iteration. The dependent variable values are returned accordingly. We estimate the population regression function with the OLS and use the residuals as well as the covariates to compute the HCCMEs. Next, we calculate the percentage difference with the true values of the variance-covariance matrix entries, Ω_{ii}, i = 0 for the intercept and 1 for the slope parameter. Namely, the percentage difference (PD) for the intercept and slope parameter variances are: $PD_{\beta_i} = 100 \times \frac{\hat{\Omega}_{ii}^{HC_j} - \Omega_{ii}}{\Omega_{ii}}$, $i = 0, 1$ and $j = 0, 1, ..., 5$. Note that j is running over the 6 HCCMEs. We included the related part of the GAUSS code algorithm in the Appendix of the paper.

Articles written on such comparisons use the quasi-t statistic as a benchmark, which makes sense since this is the test statistic corresponding to the significances of partial regression coefficients and the intercept parameter. Indeed, the percentage differences we report make similar sense and, at the same time, are very closely related to the quasi-t statistic. Because the t-statistic is nothing other than the unbiased coefficient estimate divided by the standard errors, we use this same number in the numerator of the percentage difference. This means that the statistics we are reporting and the quasi-t figures return parallel figures. We prefer to report the percentage differences because these differences make more sense of the HCCME achievements. We also compute the symmetric, entropy, and quadratic losses but prefer not to report them since these losses are in line with the percentage differences that we table and do not provide any new and further information about the comparisons. Still, we are ready to provide these losses to interested readers.

In Table 1, we present the percentage differences belonging to the covariates generated from the uniform distribution in order to curb the covariates with high leverages. For β_0, when T = 30, HC2 is the top performer, followed by HC1, whereas HC0 and HC3 are the worst HCCMEs for Case A, homoscedasticity. HC4 and HC5 return the same percentage loss, which is better than that of HC0 and HC3. The same is true for Case B. Cases C and D are nearly the same, with the ordering of HC1 and HC2 different. For β_1, the order is exactly the same for Cases A and B. Indeed, the HC2 estimates are very successful, in that the percentage difference is less than 1. For this parameter, the HC3 percentage differences are around 10. When the sample size is increased to 50, all of the HCCMEs, except for HC2, decrease their percentage differences to more than half, and the ordering does not change; still, HC2 is the best performer for Cases A and B and the ordering alters between HC2 and HC1 for Cases C and D. Coming to the variance estimation of β_1, HC2 is always the best performer, followed by HC1, HC4-HC5, HC0, and HC3. The same comments are true when the sample size is 100, 200, 300, and 500.

Note that the HCCMEs perform commendably better when the sample size is as high as 200. At this sample size, the percentage differences are around 1% or 2%, and when the number of observations is 300 or more, no estimator yields more than a 1% difference. However, one has to keep in mind that the covariates are generated from the uniform distribution, which makes the task of the HCCMEs very easy since there are no high leverage observations.

In Table 2 (Case 2), the covariates are generated from the standard normal distribution to have moderate level of leverages. For the intercept term's variance, HC2 is the best performer, followed by HC1, HC0, and HC3 in Case A. The worst performers are HC4 and HC5. The same ordering is valid for Case B, as well. For Case C, HC4 and HC5 rank third after HC1, but for Case D, they are the worst performers with almost a 16% difference from the true values. Regarding the slope parameter, β_1, HC2 performs far better than all of the rest, and HC1 follows with almost a 6% difference. The worst performers are HC4 and HC5. For Case B, the 2 best performers are the same, but the last performer is now White's estimator. For Case D, the best performing HC2 is this time followed by HC3, and then come HC1 and HC0.

Table 1. HCCME performances in percentage differences, Case 1, covariates generated from uniform distribution.

T		β_0				β_1			
		A	B	C	D	A	B	C	D
30	HC0	–9.06	–9.42	–6.66	–7.38	–9.56	–10.53	–8.47	–8.79
	HC1	–2.57	–2.95	0.01	–0.76	–3.10	–4.14	–1.93	–2.27
	HC2	–0.03	0.56	1.32	1.51	–0.05	–0.24	–0.01	–0.10
	HC3	10.04	11.83	10.07	11.35	10.57	11.39	9.29	9.48
	HC4	5.44	8.78	4.02	5.81	5.99	8.09	3.19	3.70
	HC5	5.44	8.78	4.02	5.81	5.99	8.09	3.19	3.70
50	HC0	–5.25	–4.74	–4.90	–3.91	–5.62	–5.14	–5.05	–5.35
	HC1	–1.30	–0.77	–0.94	0.09	–1.69	–1.18	–1.09	–1.41
	HC2	–0.02	0.36	1.41	1.34	0.05	0.06	0.47	0.06
	HC3	5.55	5.75	8.24	6.92	6.10	5.56	6.38	5.79
	HC4	2.82	2.28	7.62	3.92	3.39	2.09	4.41	2.59
	HC5	2.82	2.28	7.62	3.92	3.39	2.09	4.41	2.59
100	HC0	–2.34	–2.16	–2.22	–1.95	–2.85	–2.75	–3.01	–2.86
	HC1	–0.34	–0.17	–0.22	0.05	–0.87	–0.76	–1.03	–0.87
	HC2	–0.03	0.11	0.23	0.62	0.03	–0.06	–0.21	–0.06
	HC3	2.35	2.45	2.75	3.25	3.00	2.71	2.67	2.81
	HC4	0.68	0.72	1.14	1.76	1.76	1.18	1.22	1.34
	HC5	0.68	0.72	1.14	1.76	1.76	1.18	1.22	1.34
200	HC0	–1.27	–1.36	–1.09	–0.93	–1.38	–1.56	–1.35	–1.42
	HC1	–0.28	–0.37	–0.09	0.07	–0.38	–0.57	–0.35	–0.43
	HC2	0.01	–0.03	0.19	0.29	0.01	–0.12	0.06	0.00
	HC3	1.30	1.32	1.50	1.52	1.41	1.34	1.48	1.45
	HC4	0.53	0.65	0.77	0.77	0.68	0.70	0.78	0.76
	HC5	0.53	0.65	0.77	0.77	0.68	0.70	0.78	0.76
300	HC0	–0.88	–0.81	–0.76	–0.73	–0.94	–0.93	–1.06	–1.03
	HC1	–0.22	–0.14	–0.09	–0.06	–0.27	–0.27	–0.39	–0.36
	HC2	–0.04	0.05	0.10	0.14	–0.03	0.00	–0.02	–0.05
	HC3	0.81	0.92	0.97	1.01	0.89	0.95	1.02	0.94
	HC4	0.32	0.43	0.52	0.56	0.42	0.47	0.69	0.54
	HC5	0.32	0.43	0.52	0.56	0.42	0.47	0.69	0.54
500	HC0	–0.55	–0.51	–0.48	–0.38	–0.56	–0.62	–0.60	–0.53
	HC1	–0.15	–0.11	–0.08	0.02	–0.16	–0.23	–0.20	–0.13
	HC2	–0.04	0.00	0.04	0.13	–0.02	–0.06	–0.03	0.02
	HC3	0.47	0.51	0.56	0.64	0.53	0.51	0.54	0.56
	HC4	0.17	0.22	0.29	0.35	0.24	0.25	0.28	0.26
	HC5	0.17	0.22	0.29	0.35	0.24	0.25	0.28	0.26

Again, the lowest performers are HC4 and HC5. The other point that deserves attention is the equal difference reported for both HC4 and HC5, which is definitely the consequence of the same estimates by Cribari-Neto [19] and Cribari-Neto et al. [20]. There is a significant improvement in the performances of all HCCMEs but HC4 and HC5 when the sample size increases to 50. HC2 almost hits the target with percentage differences of less than 1% for Cases A, B, C, and D, followed by HC1, which is doing quite well. HC4 and HC5 are sometimes sharing the 3rd and 4th rankings and are sometimes the worst performers. Similar comments are true for the slope parameter variances, as well. The ordering is almost the same when the sample size increases to 100, 200, 300, and 500, but this time, the performances of the HCCMEs progress significantly.

Table 2. HCCME performances in percentage differences, Case 2, covariates generated from standard normal distribution.

T		β_0				β_1			
		A	B	C	D	A	B	C	D
30	HC0	–8.15	–10.69	–7.93	–12.08	–12.21	–14.44	–15.78	–19.64
	HC1	–1.59	–4.32	–1.36	–5.80	–5.94	–8.32	–9.77	–13.90
	HC2	0.21	–1.20	0.45	–0.52	0.10	–2.76	–3.90	–5.40
	HC3	9.88	9.68	9.89	12.97	14.63	10.78	9.90	11.60
	HC4	11.75	9.31	7.01	15.95	21.77	12.51	11.73	20.72
	HC5	11.75	9.31	7.01	15.95	21.77	12.51	11.73	20.72
50	HC0	–4.01	–5.00	–5.12	–6.06	–6.49	–11.35	–9.86	–15.69
	HC1	–0.01	–1.04	–1.16	–2.15	–2.59	–7.66	–6.10	–12.18
	HC2	0.00	–0.07	0.27	0.84	0.02	–2.45	–2.34	–4.80
	HC3	4.23	5.27	6.07	8.51	7.06	7.65	5.86	7.76
	HC4	1.45	5.40	4.74	13.55	6.12	18.82	6.54	28.18
	HC5	1.45	5.40	4.74	13.55	6.12	18.82	6.54	28.18
100	HC0	–1.97	–2.55	–3.05	–3.33	–3.99	–5.33	–5.94	–6.94
	HC1	0.03	–0.56	–1.07	–1.36	–2.03	–3.40	–4.02	–5.04
	HC2	0.02	–0.08	0.00	–0.09	0.04	–1.05	–1.44	–2.03
	HC3	2.07	2.48	3.17	3.29	4.32	3.45	3.31	3.19
	HC4	1.13	1.86	3.35	3.54	7.73	6.20	6.59	8.22
	HC5	1.13	1.86	3.35	3.54	7.73	6.20	6.59	8.22
200	HC0	–0.95	–1.39	–1.20	–1.84	–1.82	–3.15	–3.65	–4.41
	HC1	0.05	–0.39	–0.20	–0.84	–0.83	–2.17	–2.68	–3.45
	HC2	0.05	–0.07	0.21	–0.01	0.06	–0.72	–1.11	–1.33
	HC3	1.07	1.27	1.64	1.87	1.98	1.78	1.52	1.87
	HC4	0.51	1.24	1.69	2.84	2.64	4.07	4.49	6.28
	HC5	0.51	1.24	1.69	2.85	2.64	4.07	4.49	6.33
300	HC0	–0.66	–0.88	–1.17	–1.14	–1.31	–2.00	–2.48	–2.30
	HC1	0.01	–0.21	–0.50	–0.48	–0.65	–1.34	–1.82	–1.64
	HC2	0.00	0.00	–0.07	–0.04	0.06	–0.43	–0.74	–0.65
	HC3	0.67	0.89	1.04	1.08	1.44	1.16	1.03	1.03
	HC4	0.35	0.84	1.35	1.27	2.29	2.53	2.85	2.57
	HC5	0.35	0.84	1.35	1.27	2.29	2.53	2.85	2.57
500	HC0	–0.41	–0.53	–0.68	–0.72	–0.81	–1.23	–1.55	–1.60
	HC1	–0.01	–0.13	–0.29	–0.32	–0.41	–0.84	–1.16	–1.20
	HC2	–0.01	0.00	–0.01	–0.02	–0.01	–0.21	–0.42	–0.47
	HC3	0.39	0.54	0.66	0.69	0.80	0.82	0.73	0.67
	HC4	0.18	0.52	0.82	0.94	1.25	1.85	2.00	1.95
	HC5	0.18	0.52	0.86	0.94	1.25	1.85	2.23	1.96

For the third case (Table 3), the covariates are generated from the Student t distribution, which is known to have thick tails to allow for a greater possibility of high leverage observations. When the sample size is 30, HC2 is the best performer, followed by HC1 and HC4-HC5, and then the worst performers are HC0 and HC3. This same ordering is true for the remaining cases. Regarding the slope parameter variances, HC4-HC5 steps into the second ranking for Cases A, B, and C, and for Case D, the HC4-HC5 pair is the best performer, followed by HC2. The same ordering prevails when T is incremented to 50 for the intercept parameter in Cases

A, B, and C. However, for Case D, the best performer is HC1, followed by HC2, with the percentage differences very close at 0.02% versus 0.14%.

Table 3. HCCME performances in percentage differences, Case 3, covariates generated from Student t distribution with 3 degrees of freedom.

T		β_0				β_1			
		A	B	C	D	A	B	C	D
30	HC0	–7.06	–7.43	–7.34	–7.41	–8.59	–10.12	–8.69	–9.80
	HC1	–0.43	–0.82	–0.72	–0.80	–2.06	–3.70	–2.17	–3.36
	HC2	–0.05	–0.18	0.18	0.13	0.02	–0.93	–1.02	–1.59
	HC3	7.60	7.75	8.42	8.35	9.51	9.24	7.37	7.42
	HC4	1.74	1.85	2.79	2.03	3.69	3.54	1.41	1.20
	HC5	1.74	1.85	2.79	2.03	3.69	3.54	1.41	1.20
50	HC0	–4.10	–4.14	–4.74	–4.02	–4.77	–5.34	–6.96	–5.52
	HC1	–0.10	–0.14	–0.77	–0.02	–0.80	–1.39	–3.08	–1.58
	HC2	–0.06	0.00	0.07	0.14	–0.06	–0.24	–1.27	–0.77
	HC3	4.16	4.34	5.17	4.49	4.89	5.14	4.78	4.23
	HC4	0.38	0.72	1.98	0.67	1.07	1.51	1.88	0.35
	HC5	0.38	0.72	1.98	0.67	1.07	1.51	1.88	0.35
100	HC0	–2.01	–2.24	–2.43	–2.36	–2.39	–2.78	–2.88	–2.94
	HC1	–0.01	–0.25	–0.43	–0.37	–0.40	–0.80	–0.89	–0.96
	HC2	0.00	–0.07	0.07	–0.07	0.01	–0.19	–0.33	–0.47
	HC3	2.05	2.16	2.63	2.27	2.48	2.48	2.29	2.06
	HC4	0.21	0.42	0.96	0.39	0.61	0.74	0.60	0.19
	HC5	0.21	0.42	0.96	0.39	0.61	0.74	0.60	0.19
200	HC0	–1.05	–1.12	–1.30	–1.04	–1.22	–1.52	–1.66	–1.38
	HC1	–0.05	–0.12	–0.30	–0.04	–0.22	–0.52	–0.66	–0.39
	HC2	–0.03	–0.02	–0.10	0.09	0.01	–0.17	–0.27	–0.14
	HC3	1.01	1.10	1.12	1.23	1.24	1.19	1.14	1.12
	HC4	0.12	0.25	0.30	0.33	0.35	0.38	0.36	0.22
	HC5	0.12	0.25	0.30	0.33	0.35	0.38	0.36	0.22
300	HC0	–0.64	–0.82	–0.78	–0.78	–0.83	–1.06	–0.98	–0.99
	HC1	0.02	–0.15	–0.12	–0.11	–0.17	–0.39	–0.31	–0.33
	HC2	0.02	–0.01	–0.01	0.02	0.01	–0.17	–0.10	–0.13
	HC3	0.70	0.80	0.76	0.84	0.87	0.72	0.79	0.74
	HC4	0.12	0.21	0.18	0.24	0.29	0.15	0.21	0.16
	HC5	0.12	0.21	0.18	0.24	0.29	0.15	0.21	0.16
500	HC0	–0.43	–0.43	–0.46	–0.45	–0.52	–0.55	–0.58	–0.58
	HC1	–0.03	–0.03	–0.06	–0.05	–0.12	–0.15	–0.18	–0.18
	HC2	–0.03	0.02	0.00	0.01	–0.01	–0.02	–0.04	–0.07
	HC3	0.38	0.48	0.46	0.48	0.50	0.51	0.49	0.45
	HC4	0.03	0.13	0.12	0.12	0.15	0.17	0.16	0.09
	HC5	0.03	0.13	0.12	0.12	0.15	0.17	0.16	0.09

The ordering for Cases A and B in β_1 are the same: HC2 followed by HC1 and HC4-HC5. Coming to Case C, the HC5-HC5 pair becomes the second-best performer after HC2. Finally, in Case D, the HC4-HC5 pair outperforms all of the rest, surprisingly. When the sample size increases to 100, HC2 is the best performer, followed by HC1 and the HC4-HC5 pair, and then come HC0 and HC3 as the worst performers. For the slope parameter of β_1, HC2 is followed by HC1 and then by the HC4-HC5 pair as the best performers, and then come HC3 and HC0 for Case A. The ordering changes by the replacement of HC4-HC5, stepping in as the second-best performer in Cases B and C, and then again, surprisingly, the HC4-HC5 pair is the best performer in Case D, followed by HC2 and HC1. Similar rankings are valid for T = 200, 300, and 500, and when the sample size increases beyond 300, all of the HCCMEs perform very well with percentage differences of less than 1%. Indeed, the rankings do not matter much since all of the HCCMEs do perform very well when there are at least 300 observations.

The last case that we include in this simulation study is that of the covariates following the ratio of 2 standard normally distributed random variables; that is, the covariates follow the Cauchy distribution. We intentionally selected this distribution to make the regression have observations with very high leverages. Indeed, it is interesting to see how the HCCMEs perform in such a difficult situation, and maybe this will be the case to sort the HCCMEs in such a difficult setting. We tabled the percentage differences of the HCCMEs in Table 4. A first superficial look at the table gives the first impression of relatively poor performance for all of the HCCMEs. We note percentage differences as high as 10,000%. Indeed, there are much higher figures, but we prefer to report them as '>1000%' or '>100,000%'. These are the cases where one has to seriously question using the HCCMEs with these dramatically poor performances or discarding them.

For Case A of the intercept parameter, HC2 is the best performer, followed by HC1 and HC0, whereas the worst performers are HC4 and HC5. One other point that deserves attention is the poorer performance of HC5 as compared to HC4, even though HC5 is the more recent HCCME introduced in the literature. The ordering of the HCCMEs does not change for Case B, but the performances are much worse in Cases C and D. The picture is similar for the slope parameter. The performances are relatively good when T is increased to 50, where the ordering of the HCCMEs is still the same. The other surprising outcome is the best performance of HC1 in Case C and of HC0 in Case D. However, for β_1, HC2 is always the best performer (except for in Case C), followed by HC3 and HC1 in different situations. The other point to pay attention is the extremely high percentage differences. All of the HCCMEs, except for HC4 and HC5, increase their performances as the sample size increases to more than 100. At this level, the best performer alternates among HC2, HC1, and HC3 for Cases B, C, and D. For β_0, all of the HCCMEs, except for HC5, perform very well when the sample size is increased to 500, and for Case A, the percentage differences are less than 1%. However, for HC5, the percentage difference is greater than 100,000%.

For Case B, HC2 has a percentage difference of –0.92%, but HC0 and HC1 have percentage differences of around 48%. HC4 has a percentage difference of 723% and HC5 is even worse. For Case C, HC0, HC1, and HC2 have about 80% differences, whereas HC3 has about 90% difference, and again, HC4 and HC5 are far worse. For Case D, HC0, HC1, and HC2 have about 16% differences and both HC4 and HC5 have differences of more than 100,000%.

The situation is similar for the slope parameter variances. HC2 is the best performer, followed by HC1, HC0, and HC3. HC4 is much worse and HC5 is tremendously bad. Surprisingly, HC3 performs the best for Case C, followed by HC2, HC1, and HC0. The percentage differences are close to 100%, and there is a similar situation for Case D.

The worst performers are HC4 and HC5, with percentage differences of more than 100,000%. We care about the sample size of 500 since, if the HCCME cannot perform reasonably well even when there are 500 observations, one has to question using it or not. Another point to pay attention is the definition of the error term variances. Note that these variances are defined as functions of the covariates, and when the covariates have higher leverages, the error term variances will have more variation; this makes the estimation more difficult, which is why the HCCMEs are poor in Case D.

Table 4. HCCME performances in percentage differences, Case 4, covariates generated from Cauchy distribution.

T		β_0				β_1			
		A	B	C	D	A	B	C	D
30	HC0	–8.56	–26.69	–61.15	–38.12	–30.87	–62.83	–72.85	–78.94
	HC1	–2.02	–21.46	–58.37	–33.70	–25.93	–60.17	–70.91	–77.44
	HC2	–0.10	–7.82	–35.39	–3.53	0.01	–35.14	–44.35	–45.11
	HC3	13.08	25.72	26.69	82.78	59.20	21.66	24.30	72.54
	HC4	63.08	198.66	557.67	$> 10^3$	398.75	385.67	611.53	$> 10^3$
	HC5	200.28	735.46	$> 10^3$	$> 10^5$	$> 10^3$	$> 10^3$	$> 10^3$	$> 10^5$
50	HC0	–4.14	–9.32	–5.67	6.38	–24.65	–24.38	–29.33	–54.42
	HC1	–0.15	–5.54	–1.74	10.81	–21.51	–21.23	–26.38	–52.53
	HC2	0.04	–0.17	3.14	22.04	0.18	–8.02	–11.96	–14.23
	HC3	5.29	10.72	13.08	47.77	39.10	12.60	11.04	65.65
	HC4	13.19	30.28	20.95	163.42	196.13	66.91	75.63	559.98
	HC5	83.65	34.39	22.66	$> 10^3$	$> 10^3$	78.49	120.00	$> 10^5$
100	HC0	–2.23	–0.60	–35.45	–37.58	–28.17	–51.15	–50.23	–68.61
	HC1	–0.24	1.43	–34.14	–36.31	–26.71	–50.16	–49.22	–67.97
	HC2	–0.01	7.62	–13.55	–19.02	0.20	–25.17	–23.23	–42.80
	HC3	2.98	20.08	23.05	17.55	50.81	18.97	22.11	10.71
	HC4	10.86	67.11	190.71	242.14	317.42	226.36	232.08	355.38
	HC5	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$
200	HC0	–0.94	–14.75	–34.74	–36.43	–34.13	–27.11	–100.00	–99.92
	HC1	0.06	–13.89	–34.08	–35.79	–33.47	–26.37	–100.00	–99.92
	HC2	–0.04	–0.85	–33.93	–33.93	–0.34	–8.22	–99.40	–97.31
	HC3	1.05	16.64	59.06	51.08	69.90	16.23	38.56	28.89
	HC4	3.04	66.22	$> 10^5$	$> 10^5$	545.68	89.21	$> 10^5$	$> 10^5$
	HC5	$> 10^5$	$> 10^3$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^3$	$> 10^5$	$> 10^5$
300	HC0	–0.53	22.72	–9.80	–16.95	–20.36	–56.80	–22.01	–20.19
	HC1	0.14	23.54	–9.20	–16.40	–19.82	–56.51	–21.49	–19.65
	HC2	–0.01	31.58	0.98	–6.08	–0.39	–19.53	–6.31	–8.00
	HC3	0.54	44.60	13.72	7.07	30.26	55.71	13.02	6.59
	HC4	0.44	95.97	46.56	41.87	154.56	538.82	66.39	45.14
	HC5	$> 10^5$	$> 10^5$	$> 10^3$	$> 10^3$	$> 10^5$	$> 10^5$	$> 10^3$	$> 10^3$
500	HC0	–0.39	–47.75	79.60	–17.08	–20.34	–54.02	–99.66	–99.93
	HC1	0.01	–47.54	80.32	–16.74	–20.02	–53.84	–99.66	–99.93
	HC2	0.01	–0.92	81.45	–14.95	–0.12	–6.53	–94.37	–97.47
	HC3	0.46	94.68	88.88	65.41	28.56	94.85	14.18	41.05
	HC4	0.92	723.35	$> 10^3$	$> 10^5$	129.74	805.72	$> 10^5$	$> 10^5$
	HC5	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$	$> 10^5$

4. Concluding remarks

The winner of the game is HC2, by Horn et al. [14]; although it is occasionally beaten by other HCCMEs, its percentage difference is always at tolerable levels. The maximum percentage difference is less than 100%. One may think that this is a high difference, but it is reasonably preferable, especially when compared to the others.

One main shortcoming of comparing the HCCMEs with the help of the simulation is that the covariates and error term variances generated do not give full insight for a complete analysis. That is why we have selected several different patterns to produce covariates and error term variances, to have a more detailed understanding of the comparisons. However, although we shed some light on the settings where some HCCMEs outperform others, there may be other settings with different conclusions.

Needless to say, OLS is the best performer under homoscedasticity. There is no need to make use of the HCCMEs under homoscedasticity and there is no significant improvement in using the so-called HCCMEs. For the remaining cases, we scanned over the salient patterns of the covariates and error term variances in the literature and came up with 4 cases of covariates and error term variances.

The purpose of this paper is the comparison of the prominent HCCMEs of the related literature. As such, we included the 2 HCCMEs introduced recently. This is the only study with these HCCMEs, by Cribari-Neto [19] and Cribari-Neto et al. [20], to the best of our knowledge. Unfortunately, these estimators performed worse than others, especially when the high leverage observations were not removed. Moreover, the percentage differences of these observations were dramatically high, where HC4 was slightly better than HC5. Surprisingly, these HCCMEs sometimes performed second-best, but usually they were extremely bad.

According to our comparisons, the HCCME by Horn et al. [14] is the best performer under almost all of the settings, with and without the high leverage points. This estimator is followed by Hinkley's estimator [15]. Efron's jackknife estimator [16] appears as sometimes the 2nd and sometimes the 3rd best performer, depending on the setting. Regarding the underestimation and overestimation of the HCCMEs, the percentage differences that we report for White's HCCME are always negative, which suggests that HC0 underestimates the true covariance matrix. The same is true for HC1, in spite of a few exceptional cases, whereas HC2 is negative for the majority of the cases. To the contrary, HC3, HC4, and HC5 are almost always positive; they overestimate the true covariance. These comments may help hypotheses testers have a better idea about the true test statistic and evaluate the bias being upward or downward. Today, many packages make use of White's estimator to alleviate the bias of the test statistic. Knowing that this bias is almost always positive provides a basis to better assess the test statistics. The other point to pay attention is the close link between the HCCME performances and the settings used. Going through Tables 1–4 reveals that the HCCME performances get worse as the standard deviation of the covariates and the error terms get larger and larger.

Our findings are in line with the findings of MacKinnon and White [17], as well as the other few existing papers in the literature, but are in contradiction to those of Cribari-Neto [19] and Cribari-Neto et al. [20]. A very surprising outcome of our study is the extremely poor performances of HC4 and HC5. Indeed, these estimators are claimed to be superior to the existing ones in favorable journal articles. Most probably, the initiators of HC4 and HC5 used settings in which they managed to outperform others, but one has to be tolerable enough to use vast assumptions to arrange the settings. Similar findings could be reported, had the authors used settings like ours.

In order to let any interested reader repeat our results, we intended to provide the covariates and the error term variances in the Appendix; however, these vectors are so lengthy that we cannot present them in this paper. Instead, we can send them to any interested reader, should they be requested. Furthermore, we tried to run the GAUSS code with 1000 observations, i.e. T = 1000, and returned the outcomes with this very high sample size of the regression. We think that we could have increased T more, but the results that we got were similar and we decided to stop and report results for the maximum T value of 500.

Indeed, one can easily note that the HCCME performances are closely related to the setting used. The interested reader can look over Tables 1–4 with increasing numbers of high leverage observations. It is well known that the HCCMEs perform better in cases of no or limited high leverage observations. This directly leads us to the idea of removing the outliers to improve the HCCME performances. Indeed, there are convenient methods to work out the observations with high leverages that somehow prevent the masking effect. One can employ the minimum covariance determinant distances by Rousseeuw and Van Driessen [29] to safely figure out the observations with high leverages and these observations can be eliminated to improve the HCCME performances. This idea may lead to a new study, and possibly a new paper, which leaves the door open for further research. Indeed, Orhan and Şimşek [30] demonstrated that this idea is viable to some extent.

A very crucial point is the reason behind the good or bad performances of the estimators. Indeed, all that matters is the alleviation of the leverage. That is why massive effort is concentrated on the diminishing of the leverage. The HCCME treatments to the leverage represented by the hat matrix entries, i.e. H_{ii} terms, are indicative in the performances of the estimators. Indeed, there are some ideas to decrease the bias by simply estimating the bias terms with OLS and subtracting them from the original HCCMEs. This idea was implemented by Cribari-Neto et al. [31]. The idea worked well, as illustrated by the application in the paper, but the computation is really very difficult, which makes the use of the technique overwhelmingly difficult.

Finally, as mentioned, one has to keep in mind that simulation studies rely on some certain types of Xs and Σs, and are therefore of limited use in generalizations. One can change Xs and Σs completely and get much better or worse performances of the HCCMEs. The better way to make an analysis is with proof-type studies, where there is more insight into the subject matter. However, this area of study does not yet include such proof types and simulations can be used to have comparisons. This is the second research proposal that we suggest as the extension of our study.

Acknowledgments

The authors would like to thank TÜBİTAK for its valuable support for this project with reference number 110K439. This paper was presented at the 13th International Conference on Econometrics, Operations Research, and Statistics, 24–26 May 2012.

Appendix

GAUSS program algorithm

Input the parameters of the setting
Generate the covariates according to cases
Generate the error term variances
Compute all terms outside the while loop
Do while Monte Carlo sample is not attained
    Generate the error terms and dependent variable values
    Estimate the residuals
    Compute the HCCMEs
    Save the HCCMEs in a separate matrix
Compare the HCCMEs' biases
Compute the HCCMEs' loss functions

References

[1] J.T. Jeng, C.C. Chuang, S.F. Su, "Support vector interval regression networks for interval regression analysis", Fuzzy Sets and Systems, Vol. 138, pp. 283–300, 2003.

[2] I. Moghram, S. Rahman, "Analysis and evaluation of five short-term load forecasting techniques", IEEE Trans Power Systems, Vol. 4, pp. 1484–1491, 1989.

[3] O. Cordón, F. Herrera, L. Sánchez, "Solving electrical distribution problems using hybrid evolutionary data analysis techniques", Applied Intelligence, Vol. 10, pp. 5–24, 1999.

[4] M.K. Kerr, M. Martin, G.A. Churchill, "Analysis of variance for gene expression microarray data", Journal of Computational Biology, Vol. 7, pp. 819–837, 2000.

[5] C. Maffeis, G. Talamini, L. Tat, "Influence of diet, physical activity and parents' obesity on children's adiposity: a four year longitudinal study", International Journal of Obesity, Vol. 22, pp. 758–764, 1998.

[6] B. Peeters, J. Maeck, G. De Roeck, "Vibration-based damage detection in civil engineering excitation sources and temperature effects", Smart Materials and Structures, Vol. 10, pp. 518–527, 2001.

[7] J. Faraway, "Modeling reach motions using functional regression analysis", SAE Technical Paper 2000-01-2175, 2000.

[8] E.L. Breese, J. Hill, "Regression analysis of interactions between competing species", Heredity, Vol. 31, pp. 181–200, 1973.

[9] G. Heinemann, D. Nordman, E. Plant, "The relationship between summer weather and summer loads – a regression analysis", IEEE Transactions on Power Apparatus and Systems, Vol. 85, pp. 1144–1154, 1966.

[10] A. Chesher, I. Jewitt, "The bias of a heteroskedasticity-consistent covariance matrix estimator", Econometrica, Vol. 55, pp. 1217–1222, 1987.

[11] H. White, "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity", Econometrica, Vol. 48, pp. 817–838, 1980.

[12] F. Eicker, "Asymptotic normality and consistency of the least squares estimators for families of linear regressions", Annals of Mathematical Statistics, Vol. 34, pp. 447–456, 1963.

[13] F. Eicker, "Limit theorems for regressions with unequal and dependent errors", Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, pp. 59–82, 1967.

[14] S.D. Horn, R.A. Horn, D.B. Duncan, "Estimating heteroscedastic variances in linear models", Econometrica, Vol. 60, pp. 539–547, 1975.

[15] D.V. Hinkley, "Jackknifing in unbalanced situations", Technometrics, Vol. 19, pp. 285–292, 1977.

[16] B. Efron, The Jackknife, the Bootstrap, and Other Resampling Plans, Society for Industrial and Applied Mathematics, Philadelphia, 1982.

[17] J.G. MacKinnon, H. White, "Some heteroskedasticity consistent covariance matrix estimators with improved finite sample properties", Journal of Econometrics, Vol. 29, pp. 305–325, 1985.

[18] M. Orhan, A. Zaman, "Comparison of several heteroskedasticity consistent covariance matrix estimators", Middle East Technical University Economic Research Center Conference, 2000.

[19] F. Cribari-Neto, "Asymptotic inference under heteroskedasticity of unknown form", Computational Statistics and Data Analysis, Vol. 45, pp. 215–233, 2004.

[20] F. Cribari-Neto, T.C. Souza, K.L.P. Vasconcellos, "Inference under heteroskedasticity and leveraged data", Communications in Statistics—Theory and Methods, Vol. 36, pp. 1877–1888, 2007.

[21] D.W.K. Andrews, "Heteroskedasticity and autocorrelation consistent covariance matrix estimation", Econometrica, Vol. 59, pp. 817–858, 1991.

[22] J.G. MacKinnon, "Thirty years of heteroskedasticity-robust inference", in: X. Chen, N.R. Swanson, editors, Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr., Springer, Berlin, pp. 437–461, 2013.

[23] A.K. Bera, T. Suprayitno, G. Premaratne, "On some heteroskedasticity-robust estimators of variance-covariance matrix of the least-squares estimators", Journal of Statistical Planning & Inference, Vol. 108, pp. 121–136, 2002.

[24] A.F. Hayes, L. Cai, "Using heteroskedasticity-consistent standard error estimators in OLS regression: an introduction and software implementation", Behavior Research Methods, Vol. 39, pp. 709–722, 2008.

[25] C.F.J. Wu, "Jackknife, bootstrap and other resampling methods in regression analysis", Annals of Statistics, Vol. 14, pp. 1261–1295, 1986.

[26] J.G. Cragg, "More efficient estimation in the presence of heteroscedasticity of unknown form", Econometrica, Vol. 51, pp. 751–763, 1983.

[27] E. Flachaire, "Bootstrapping heteroscedastic regression models: wild bootstrap vs. pairs bootstrap", Computational Statistics and Data Analysis, Vol. 49, pp . 361–376, 2005.

[28] M.C. Lima, T.C. Souza, F. Cribari-Neto, G.B. Fernandes, "Heteroscedasticity-robust inference in linear regressions", Communications in Statistics—Simulation and Computation, Vol. 39, pp. 194–206, 2010.

[29] P.J. Rousseeuw, K. Van Driessen, "A fast algorithm for the minimum covariance determinant estimator", Technometrics, Vol. 41, pp. 212–223, 1999.

[30] M. Orhan, E. Şimşek "How to improve HCCME performances in small samples?", 13th International Conference on Econometrics, Operations Research and Statistics, 2012.

[31] F. Cribari-Neto, S.L.P. Ferrari, G.M. Cordeiro, "Improved heteroscedasticity-consistent covariance matrix estimators", Biometrika, Vol. 87, pp. 907–918, 2000.